# Hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory

Subjects: Others

Contributor: Sunusi Bala Abdullahi , Zakariyya Abdullahi Bature , Lubna A. Gabralla , Haruna Chiroma

Recognition of lying is a more complex cognitive process than truth-telling because of the presence of involuntary cognitive cues that are useful to lie recognition. Researchers have proposed different approaches in the literature to solve the problem of lie recognition from either handcrafted and/or automatic lie features during court trials and police interrogations.

artificial intelligence bidirectional long short-term memory convolutional neural network

behavioral biometrics deception detection

## 1. Introduction

On average, every person tells lies at least twice a day [1]. More aggravating is lies presented against others during court trials, police interrogations, interviews, etc., which change the outcome of relevant facts and may lead to wrong judgments or convictions. These problems have inspired the development of computer engineering systems, such as electroencephalography (EEG). Despite the benefits of computer engineering systems for lie recognition, some restrictions exist, such as being cumbersome, which allows a liar to understand that they are being monitored, thus resulting in the presence of deliberate behavioral attitude that can confuse the interviewers. Such deliberate behavioral attitude affects involuntary cues, which mislead the actual results. These involuntary cues comprise facial expression, body language, eye motion, and hand motion, as shown in **Figure 1**. Each subfigure contains a scene from a court trial video. The scene contains a label in the white box corresponding to the number from the video clip of the original court trial video data set. The scenes from the top left corner to the right show the behavioral attitudes of lying people, while the scenes from the bottom left corner to the right show the behavioral attitudes of truth-telling people. Addressing the problem of learning human involuntary cues, recent research in the field of image processing/CV and machine learning reshapes computer engineering systems into machine learning-based (ML) systems [2][3][4]. ML-based systems can learn tiny facial marks [4] and behaviors in connection with body motion, as well as hand gestures [5], therefore making lie recognition suitable via CV and ML techniques. However, a combination of two or more human involuntary actions (known as cognitive cues) provides good results at some expenses [6]. Therefore, deep learning with CV features, such as bidirectional long short-term memory (BLSTM), has advanced with appreciable performance, although only a few examples have appeared in the literature [1]. However, the weights of BLSTM do not highlight the key information in the context, which leads to information redundancy when learning long video sequences [7], as well as insufficient recognition accuracy and model instability [1].

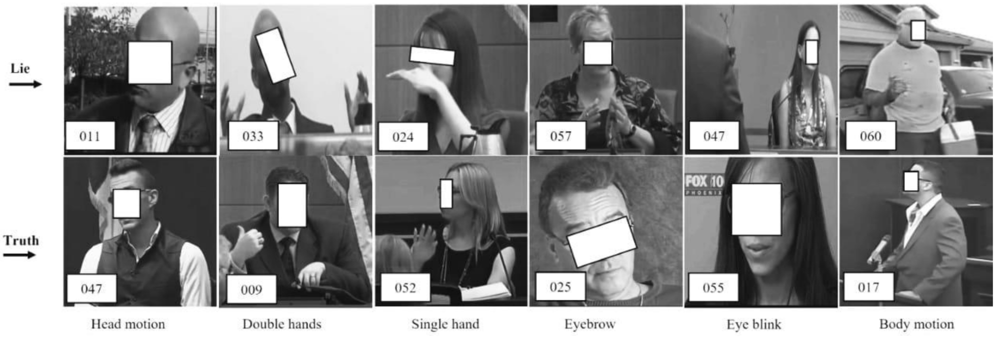

**Figure 1.** Sample of real-life court trial video data set with involuntary cognitive cues. From the top left corner: (1) lying during the court trial with a forward head motion, (2) lying with double-hand motions, (3) lying with a single hand motion, (4) lying with an eyebrow, (5) lying with an eye blink, and (6) lying with a body motion. From the left bottom corner to the right: (7) truth-telling during the court trial with a forward head motion, (8) truth-telling with double-hand motions, (9) truth-telling with a single hand motion, (10) truth-telling with an eyebrow, (11) truth-telling with an eye blink, and (12) truth-telling with a body motion [1].

The recognition accuracy of these methods [1][4] is low because some multi-modal involuntary cues and their complementary information are missing; thus, their accuracy needs to be improved since involuntary cues, such as those shown in **Figure 1**, are significant factors in determining people's behaviors while giving testimony during court trials or investigations. These involuntary cues are difficult to capture by using classical technique. Thus, deep learning methods are the suitable choice. However, deep learning methods provide a huge amount of information that is sometimes irrelevant to lie recognition. Uncertainty about the type of multi-modal information to be used for lying recognition remains a key factor. Thus, researchers improve this process by highlighting the key information of multi-modal features by proposing multi-modal spatial–temporal state transition patterns (STSTP). It is found that the highlighted multi-modal STSTP information provides a sound basis for lie recognition under real-life court trial videos and paves the way for the development of explainable and principled tools. Inspired by these results, researchers propose spatial–temporal state transition patterns based on involuntary actions of lying and truth-telling persons.

## 2. Eye Blinking Approach

Eye blinking is an involuntary cue during lying or truth-telling actions; however, it is a valuable index to enhance effective recognition. The eye-blinking cues of a lie are hard to learn during a cross-examination or a court trial. Although complex techniques are in use to record eye blinking, such as eye trackers, these techniques need a biomarker and complex data interpretation. Therefore, RGB videos from computer vision (CV) provide a flexible data set for the recognition of lies. CV allows an algorithm to be built without the need for a biomarker and/or complex data interpretation support. Eye-gaze lie systems, such as that of Bhaskaran et al. [8], propose eye-gaze features based on dynamic Bayesian learning. This method was reported to achieve an accuracy of 82.5% in learning distinct features between deceit and non-deceit cues. The major limitation of this work includes failure to reflect real-life scenarios, such as a suspect or witness wearing glasses or showing flicking an eyebrow motion. Proudfoot et al. [9] proposed eye pupil diameter using a latent growth curve modeling technique to capture changes in the eyes of the suspect and complainant, while George et al. [10] evaluated the number of eyeblink counts and their duration among lying and truth-telling persons. The former study finds that significant changes occur when a person is telling lies, while the latter study can conclude when a lying person is pressurized. The advantage of the work by Avola et al. [4] is that it highlights the benefits of extracting macro- and micro-expressions (MME) during police interrogation, cross-examination, and court trials. Macro- and micro-expressions of the face are built in an ensemble fashion. Therefore, it can be observed that a truth-telling or lie-telling person employs various body cues (multi-modal cues) to express themselves, as shown in **Figure 1**; thus, single-body cues are not sufficient to discriminate lies from facts.

## 3. Multi-Modal Cue Approaches

An automated multi-modal lie recognition system can allow the building of a system with potential behavioral cues to distinguish a lie from the truth [11]. The work by PrezRosas et al. [12] exploited verbal and non-verbal indices to detect court verdicts with decision trees and random forests. Abouelenien et al. [13] demonstrated the performance of cross-referencing physiological information with a decision tree and majority voting strategy, while Karimi et al. [14] exploited visual and acoustic cues using large margin nearest neighbor learning. Wu et al. [15] considered visual, audio, and text information in unison to compare and select the best classifier among decision trees, random forests, and linear SVM. Rill-Garcia et al. [16] jointly combined visual, acoustical, and textual indices using SVM to evaluate the effectiveness of the combined information. Krishnamurthy et al. [17] utilized a 3D CNN for feature extraction, and classification was conducted using a multi-layer perceptron.

Furthermore, hand features are very stable cues for identifying human actions and intentions, as reported in the literature [3][18][19]. Lu et al. [20] extracted hand and facial features using color 3-D LUT, which are further utilized with blob analysis to track head and hand motions (behavioral state). Their method needs to be improved to avoid complex segmentation and long processing time. Meservy et al. [11] extracted hand and facial features using color analysis, eigenspace-based shape segmentation, and Kalman filters. The major limitation of this method is user invariability. Avola et al. [1] extracted hand features from RGB videos using OpenPose. In their method, the hand is represented using 21 finger joint coordinates per frame along with acceleration and velocity. In addition, their method calculates hand elasticity and openness to observe hand behavior while lying or speaking the truth. Mut Sen et al. [5] proposed visual, acoustic, and linguistic modalities. This method designs automatic and manually annotated features using a random seed, and the features are validated using different classifiers in semi-automatic and automatic modes. The best results are obtained from the semi-automatic system with artificial neural network classifiers. The work in [5] proposes a multi-feature approach based on subject-level analysis. The features are detected manually, which affects the performance. Most of the current best works achieve the best result via deep learning methods. However, methods that utilize eyebrow, eye blinking, and optical flow of involuntary information have not been addressed by the current challenges.

## 4. The Hybrid CNN-BiLSTM Architecture: Mechanisms and Advantages

The architectural gap identified in prior methods, specifically, the inability to jointly model salient spatial features and their long-range temporal dependencies in involuntary cue data, is addressed by hybrid Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) models. This paradigm operates on a principled division of labor: the CNN backbone acts as a hierarchical spatial feature extractor, transforming raw video frames or optical flow fields into compact, discriminative representations of local patterns (e.g., micro-expressions, hand contours). These sequential feature vectors are then processed by the BiLSTM layer, which models their temporal evolution by learning contextual dependencies in both forward and backward directions. This bidirectional context is critical for interpreting cues like the progression of a gesture or the dynamics of a facial action unit. The comparative advantage of this hybrid approach lies in its end-to-end learning capability, which supersedes the need for manual feature engineering. It directly optimizes the integration of spatial and temporal information, thereby mitigating information redundancy and enhancing model stability for the complex, variable-length sequences characteristic of real-world behavioral data.

## 5. Conclusion and Future Research Trajectories

In conclusion, the application of deep learning, particularly hybrid spatial-temporal models like CNN-BiLSTM, represents a significant advance in automated lie recognition from visual cues. These systems move beyond isolated cue analysis toward a more holistic integration of multi-modal behavioral signals. Future research must navigate several key challenges to transition from constrained experimental settings to robust, real-world deployment. These include: (i) the curation of large-scale, ecologically valid video datasets that reflect diverse populations, lighting conditions, and cultural nuances in nonverbal behavior; (ii) the development of explainable AI (XAI) frameworks to render model decisions interpretable to forensic experts and legal professionals, ensuring adherence to evidentiary standards; and (iii) the exploration of efficient, lightweight architectures suitable for potential real-time analysis scenarios. Addressing these challenges will be pivotal in developing principled, reliable tools for assistive forensic analysis.

### References

1. Avola, D.; Cinque, L.; Maria, D.; Alessio, F.; Foresti, G. LieToMe: Preliminary study on hand gestures for deception detection via Fisher-LSTM. Pattern Recognit. Lett. 2020, 138, 455–461.
2. Al-jarrah, O.; Halawan, A. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artif. Intell. 2001, 133, 117–138.
3. Abdullahi, S.B.; Khunpanuk, C.; Bature, Z.A.; Chroma, H.; Pakkaranang, N.; Abubakar, A.B.; Ibrahm, A.H. Biometric Information Recognition Using Artificial Intelligence Algorithms: A Performance Comparison. IEEE Access 2022, 10, 49167–49183.
4. Avola, D.; Cascio, M.; Cinque, L.; Fagioli, A.; Foresti, G. LieToMe: An Ensemble Approach for Deception Detection from Facial Cues. Int. J. Neural Syst. 2021, 31, 2050068.

5. Sen, U.; Perez, V.; Yanikoglu, B.; Abouelenien, M.; Burzo, M.; Mihalcea, R. Multimodal deception detection using real-life trial data. IEEE Trans. Affect. Comput. 2022, 2022, 2050068.
6. Ding, M.; Zhao, A.; Lu, Z.; Xiang, T.; Wen, J. Face-focused cross-stream network for deception detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019.
7. Abdullahi, S.B.; Chamnongthai, K. American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach. IEEE Access 2022, 10, 15911–15923.
8. Bhaskaran, N.; Nwogu, I.; Frank, M.G.; Govindaraju, V. Lie to me: Deceit detection via online behavioral learning. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–23 March 2011; IEEE: Piscataway, NJ, USA, 2011.
9. Jeffry, G.; Jenkins, J.L.; Burgoon, J.K.; Judee, K.; Nunamaker, J.F. Deception is in the eye of the communicator: Investigating pupil diameter variations in automated deception detection interviews. In Proceedings of the 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), Baltimore, MD, USA, 27–29 May 2015; IEEE: Piscataway, NJ, USA, 2015.
10. Thakar, M.K.; Kaur, P.; Sharma, T. Validation studies on gender determination from fingerprints with special emphasis on ridge characteristics. Egypt. J. Forensic Sci. 2022, 8, 20.
11. Meservy, T.O.; Jensen, M.L.; Kruse, J.; Burgoon, J.K.; Nunamaker, J.F.; Twitchell, D.P.; Tsechpenakis, G.; Metaxas, D.N. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. IEEE Intell. Syst. 2005, 20, 36–43.
12. Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Burzo, M. Deception detection using real-life trial data. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, DC, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015.
13. Abouelenien, M.; Pérez-Rosas, V.; Mihalcea, R.; Burzo, M. Detecting deceptive behavior via integration of discriminative features from multiple modalities. IEEE Trans. Inf. Forensics Secur. 2017, 5, 1042–1055.
14. Karimi, H.; Tang, J.; Li, Y. Toward end-to-end deception detection in videos. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: Piscataway, NJ, USA, 2018.
15. Wu, Z.; Singh, B.; Davis, L.; Subrahmanian, V. Deception detection in videos. In Proceedings of the AAAI conference on artificial intelligence, New Orleans, LA, USA, 2–7 February 2018; AAAI: Washington, DC, USA, 2018.
16. Rill-García, R.; Jair, E.H.; Villasenor-Pineda, L.; Reyes-Meza, V. High-level features for multimodal deception detection in videos. In Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; IEEE: Piscataway, NJ, USA, 2019.

17. Krishnamurthy, G.; Majumder, N.; Poria, S.; Cambria, E. A deep learning approach for multimodal deception detection. arXiv 2018, arXiv:1803.00344.
18. Abdullahi, S.B.; Ibrahim, A.H.; Abubakar, A.B.; Kambheera, A. Optimizing Hammerstein-Wiener Model for Forecasting Confirmed Cases of COVID-19. IAENG Int. J. Appl. Math. 2022, 52, 101–115.
19. Abdullahi, S.B.; Muangchoo, K. Semantic parsing for automatic retail food image recognition. Int. J. Adv. Trends Comput. Sci. Eng. 2020, 53, 7808–7816.
20. Lu, S.; Tsechpenakis, G.; Metaxas, D.N.; Jensen, M.L.; Kruse, J. Blob analysis of the head and hands: A method for deception detection. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 3–6 January 2005; IEEE: Piscataway, NJ, USA, 2005.

Retrieved from https://encyclopedia.pub/entry/history/show/132094