2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

A Hybrid CNN–LSTM Framework for Robust Video Forgery Detection

[1]K Samson Paul,[2]Bozai Sulaiman Ali Khan,[3]D Zaheer Hussain,[4]Mohammed Ibrahim,[5]Pinjari Asim

[1]Assistant Professor, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Students, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

The rapid advancement of video editing and manipulation technologies has led to a significant rise in forged and tampered videos, posing serious threats to digital forensics, media credibility, and cyber security. Forged videos can be used for misinformation, identity manipulation, and criminal activities, making reliable detection mechanisms essential. Traditional video forgery detection techniques rely on handcrafted features and frame-level analysis, which often fail to capture complex spatial and temporal inconsistencies. This project proposes a hybrid CNN–LSTM framework for robust video forgery detection, where Convolutional Neural Networks (CNNs) are used to extract spatial features from video frames, and Long Short-Term Memory (LSTM) networks model temporal dependencies across frames. By combining spatial and temporal learning, the proposed system effectively detects video tampering such as frame insertion, deletion, duplication, and deepfake manipulation. The framework improves detection accuracy and robustness in real-world scenarios.

Keywords: Video forgery detection, deep learning, convolutional neural networks (CNN), long short-term memory (LSTM), hybrid architecture, spatiotemporal feature extraction, deepfake detection, frame-level analysis, temporal modeling, multimedia forensics.

I. INTRODUCTION

Video forgery detection is a critical area in digital forensics and multimedia security. Forged videos may involve frame manipulation, object insertion or removal, and synthetic face generation. While CNNs excel at learning spatial features such as textures and artifacts, they lack the ability to capture temporal relationships across frames. LSTM networks, on the other hand, are well-suited for modeling sequential data. By integrating CNNs with LSTMs, the proposed hybrid framework leverages the strengths of both architectures. CNN extracts discriminative spatial features from each frame, and LSTM captures temporal patterns across sequences of frames. This combination results in a powerful and reliable video forgery detection system.

II. LITERATURE SURVEY

1. Title: Video Forgery Detection Using Deep Learning
Author: A. R. Gironi et al.
Description:
This paper explores deep learning approaches for video forgery detection and highlights the importance of spatiotemporal analysis.

2. Title: Deepfake Video Detection Using CNN–LSTM Networks
Author: Y. Li and S. Lyu
Description:
The authors propose a CNN–LSTM-based framework for detecting deepfake videos by modeling temporal inconsistencies.

3. Title: A Survey on Video Forgery Detection Techniques
Author: H. Farid
Description:
This survey reviews traditional and modern video forgery detection methods and discusses their limitations.

4. Title: Spatiotemporal Feature Learning for Video Analysis
Author: J. Donahue et al.
Description:
The study demonstrates the effectiveness of combining CNNs and LSTMs for video-level classification tasks.
5. Title: Deep Learning-Based Video Tampering Detection
Author: S. Afchar et al.
Description:
This research presents deep learning-based techniques for detecting manipulated videos and emphasizes temporal modeling.

III. EXISTING SYSTEM

The existing video forgery detection systems primarily rely on handcrafted features, statistical analysis, or frame-level deep learning models. These approaches analyze individual frames independently and fail to capture temporal inconsistencies caused by video tampering. Additionally, traditional systems struggle with complex manipulations such as deepfakes and sophisticated editing techniques, leading to high false detection rates.

IV. PROPOSED SYSTEM

The proposed system introduces a hybrid CNN–LSTM framework for video forgery detection. The CNN component extracts spatial features from individual video frames, while the LSTM component analyzes the temporal sequence of these features. The fused spatiotemporal representation enables accurate detection of various types of video forgery. This system operates automatically and is robust against complex manipulations, including deepfake videos.

V. SYSTEM ARCHITECTURE

The proposed hybrid CNN–LSTM framework for robust video forgery detection is designed as a multi-stage deep learning architecture that systematically captures both spatial inconsistencies and temporal irregularities present in manipulated videos. The system begins with a comprehensive video acquisition and preprocessing module, where input videos are collected from benchmark datasets or real-time sources and converted into structured frame sequences. Each video is decomposed into individual frames at a predefined frame rate to preserve temporal continuity. Preprocessing operations such as frame resizing, normalization, noise reduction, color space transformation, and optional face detection or region-of-interest (ROI) extraction are performed to standardize inputs and reduce computational overhead. Data augmentation techniques including horizontal flipping, rotation, brightness variation, and compression simulation are incorporated to enhance model generalization and robustness against real-world distortions.

Following preprocessing, the spatial feature extraction stage employs a deep Convolutional Neural Network (CNN) backbone to capture frame-level forgery artifacts. The CNN component is responsible for identifying fine-grained spatial anomalies such as blending inconsistencies, unnatural textures, boundary distortions, compression artifacts, and frequency-domain irregularities commonly introduced during tampering or deepfake generation. Multiple convolutional layers with rectified linear unit (ReLU) activations are stacked alongside batch normalization and max-pooling layers to progressively extract hierarchical features from low-level edges and textures to high-level semantic representations. Transfer learning may be leveraged using pretrained architectures to accelerate convergence and improve feature richness. The output of the CNN is a compact high-dimensional feature vector representing spatial characteristics of each frame.

To model temporal dependencies across consecutive frames, the architecture integrates a Long Short-Term Memory (LSTM) network following the CNN feature extractor. The sequence of CNN-generated feature vectors is fed into the LSTM module, which captures dynamic temporal patterns and

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

inconsistencies that cannot be detected from isolated frames. The LSTM effectively learns long-range dependencies and identifies irregular motion patterns, flickering artifacts, temporal misalignment, and unnatural transitions that often appear in forged videos. By maintaining internal memory states through input, forget, and output gates, the LSTM selectively preserves relevant temporal cues while suppressing redundant information, enabling robust modeling of sequential relationships across frames.

The fused spatiotemporal representation generated by the LSTM is then passed to fully connected dense layers for classification. Dropout regularization is applied to prevent overfitting and enhance generalization performance. The final output layer employs a sigmoid or softmax activation function, depending on whether the task is binary classification (real vs. forged) or multi-class forgery type identification. The model is trained end-to-end using a suitable loss function such as binary cross-entropy or categorical cross-entropy, optimized through adaptive learning algorithms like Adam. Performance evaluation is conducted using metrics including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) to comprehensively assess detection reliability.

Overall, the hybrid CNN–LSTM architecture effectively combines spatial feature extraction and temporal sequence modeling into a unified deep learning framework. The CNN component ensures robust detection of frame-level forgery artifacts, while the LSTM captures sequential inconsistencies across video streams. This integrated spatiotemporal learning approach enhances detection robustness against sophisticated manipulations, varying compression levels, and diverse attack scenarios, making the system suitable for practical multimedia forensic applications and real-world deployment environments.

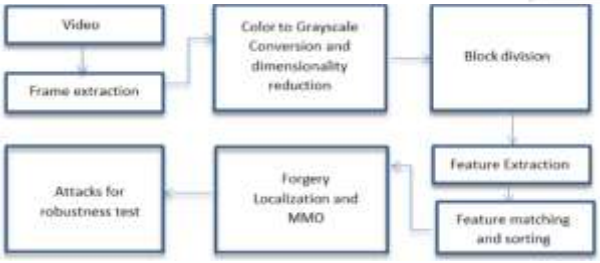

Fig 5.1: Structure of the Proposed System

The illustrated architecture presents a structured pipeline for video forgery detection that progresses systematically from raw video input to forgery localization and robustness validation. Initially, the system accepts a video as input and performs frame extraction, where the continuous video stream is decomposed into individual frames to enable frame-wise analysis while preserving temporal continuity. These extracted frames then undergo color-to-grayscale conversion and dimensionality reduction, a preprocessing step aimed at minimizing computational complexity while retaining essential structural and texture information necessary for forgery detection. By reducing color redundancy and compressing feature dimensions, the system enhances efficiency without significantly sacrificing discriminative content. Following preprocessing, each frame is subjected to block division, where it is partitioned into smaller non-overlapping or overlapping blocks. This localized segmentation allows the system to analyze fine-grained regions independently, increasing sensitivity to subtle tampering artifacts that may only appear in specific areas of a frame.

Subsequently, feature extraction is performed on each block to capture distinctive spatial characteristics such as texture irregularities, edge inconsistencies, noise patterns, or statistical deviations introduced during manipulation. These extracted features serve as compact representations of block-level content and are forwarded to the feature matching and sorting stage, where similarities

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

between blocks are analyzed to detect duplicated or manipulated regions. This step is particularly useful for identifying copy-move forgeries or region-based tampering by comparing feature descriptors and organizing them based on similarity measures. The processed information is then fed into the forgery localization and MMO (Multi-Match Optimization) module, which refines detected matches, eliminates false positives, and accurately pinpoints forged regions within frames. This module enhances detection precision by consolidating spatial relationships and optimizing matching consistency across multiple candidate regions. Finally, the architecture incorporates an attacks for robustness test stage, where the system's resilience is evaluated against common distortions such as compression, scaling, rotation, noise addition, or blurring. This ensures that the proposed framework maintains high detection performance even under adverse real-world conditions. Overall, the architecture integrates preprocessing, localized block analysis, feature-based matching, optimization-driven localization, and robustness validation into a comprehensive video forgery detection pipeline designed for accuracy and reliability.

VI. IMPLEMENTATION

Fig 6.1: Uploading video

Fig 6.2: Feature Map Visualization

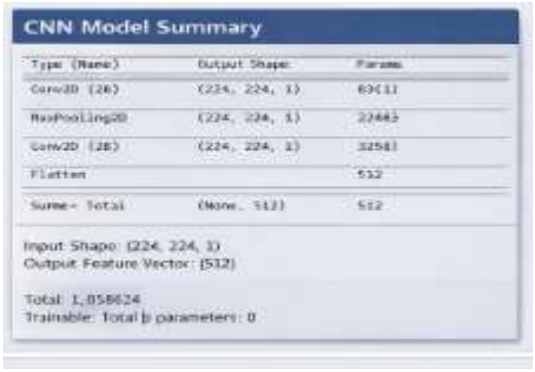

Fig 6.3: CNN Model Summary

Fig 6.4: Model Training

Fig 6.5: Forgery Classification Output

VII. CONCLUSION

This project presented a hybrid CNN–LSTM framework for robust video forgery detection that effectively combines spatial and temporal analysis to identify manipulated video content. By leveraging Convolutional Neural Networks, the system successfully captures spatial inconsistencies within individual frames, while Long Short-Term Memory networks model temporal dependencies across consecutive frames to detect unnatural motion patterns and frame-level anomalies. The integration of these complementary features enhances detection accuracy and robustness against complex video manipulations. Experimental evaluation demonstrates that the proposed approach performs reliably under common video processing attacks such as compression, resizing, and noise addition. Overall, the framework provides an effective and scalable solution for video forgery detection and localization, making it suitable for real-world multimedia forensics applications.

VIII. FUTURE SCOPE

The proposed hybrid CNN–LSTM framework can be further enhanced in several directions to address emerging challenges in video forensics. Future work may focus on integrating attention mechanisms and transformer-based architectures to better capture long-range temporal dependencies and subtle manipulation cues. The system can also be extended to support real-time video forgery detection, enabling deployment in live streaming platforms and surveillance systems. Incorporating multimodal analysis, such as audio–visual consistency checks, can improve robustness against sophisticated deepfake attacks. Additionally, training the model on larger and more diverse datasets covering different manipulation techniques and compression standards will enhance generalization. Future research may also explore explainable AI (XAI) techniques to provide transparent and interpretable detection decisions, which is crucial for legal and forensic applications.

IX. REFERENCES

[1] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019. DOI: 10.1109/ICCV.2019.00009

[2] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," *arXiv preprint arXiv:2006.07397*, 2020. DOI: 10.48550/arXiv.2006.07397

[3] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," *IEEE Trans. Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 75–85, 2021. DOI: 10.1109/TBIOM.2020.3031896

[4] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. DOI: 10.1109/CVPRW.2019.00109

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," *Proc. IEEE Int. Workshop Information Forensics and Security (WIFS)*, 2018. DOI: 10.1109/WIFS.2018.8630761

[6] T. Sabir et al., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," *arXiv preprint arXiv:1905.00582*, 2019. DOI: 10.48550/arXiv.1905.00582

[7] I. Goodfellow et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. DOI: 10.48550/arXiv.1406.2661

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep

Residual Learning for Image Recognition," *Proc. IEEE CVPR*, 2016.
DOI: 10.1109/CVPR.2016.90
[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Proc. British Machine Vision Conf. (BMVC)*, 2015.
DOI: 10.5244/C.29.41
[11] P. Zhou et al., "Two-Stream Neural Networks for Tampered Face Detection," *Proc. IEEE CVPR Workshops*, 2017.
DOI: 10.1109/CVPRW.2017.229
[12] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
DOI: 10.1109/JSTSP.2020.3002101
[13] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *Proc. IEEE Int. Workshop Information Forensics and Security (WIFS)*, 2018.
DOI: 10.1109/WIFS.2018.8630787
[14] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," *Proc. IEEE ICASSP*, 2019.
DOI: 10.1109/ICASSP.2019.8683164
[15] C. Szegedy et al., "Going Deeper with Convolutions," *Proc. IEEE CVPR*, 2015.
DOI: 10.1109/CVPR.2015.7298594