An Automated Workflow for Deepfake Detection

Anirudh Joshi and Chandrashekar Chavan

October 8, 2023

An Automated Workflow For Deepfake Detection*

1st Anirudh Joshi
CSE
PESU
Bangalore, India
joanirudh@outlook.com

2nd Dr. Chandrashekar Chavan
CSE
PESU
Bangalore, India
cpchavan@pes.edu

***Abstract*—The lightning development of Artificial Intelligence (AI) has brought significant changes to modern society, including the emergence of AI-generated art and image enhancement techniques. However, one of the most alarming consequences of AI advancement is the creation of deepfake images and videos through the use of General Adversarial Networks (GANs). As these deepfakes become increasingly convincing and widespread, there is an urgent need for measures to detect and prevent their dissemination, especially on independent social-media websites. The proposed approach results in accuracy scores comparable to and surpassing several SOTA(State-of-the-art) approaches on three benchmark datasets, while consuming considerably lesser computational overhead, and containing over 100x lesser trainable parameters which was achieved using the extraction and manipulation of geometrical features.**

***Index Terms*—Deep-fake detection, Lucas-Kande, Autonmation, Bi-LSTM, Computer Vision, Celeb-DF, UADFV, FF++, Facial Landmarks**

I. Introduction

The problem of deepfakes continues to grow as the power of Deep Learning methods increases over time. With the proliferation of deepfake videos on social media, there is a growing concern for the spread of misinformation, defamation of individuals, especially women, due to the widespread availability of deepfaked adult content on the internet. It is essential to develop a technique that can robustly detect deepfaked videos before they are uploaded on social media sites to prevent their vast distribution.

Previous research has heavily focused on Convolutional Neural Networks (CNN) [1]–[6]to detect deepfake images and videos due to their ability to compress and extract useful features using pooling techniques. However, as the focus shifted to deepfake videos, there was a switch to architectures such as Long Short-Term Memory (LSTM), which can process a video as a sequence of image frames, making it easier to observe correlation between events across the frames.

Several approaches were later developed to track the movement of relevant facial landmarks and identify possible manipulation of facial features, such as eye blinking, to find dissimilarities in the temporal aspect of the video. However, these methods can be computationally intensive and may overload social media servers.

To address this, the proposed approach presents a lightweight technique that extracts the optical flow of facial landmarks over time and feeds them to a Bi-directional LSTM to detect deepfake videos. This technique allows for quicker processing of data without overloading the servers of social media sites. By leveraging the temporal information in the optical flow, this approach can accurately detect deepfaked videos with high precision, thus reducing the risk of the spread of misinformation and defamation on social media.

إسرائيل تفرج عن الناشط صدقي المقت بعد 27 عاما في سجون الاحتلال
14:01 غرينتش

Fig. 1. A deepfaked TV Anchor

This paper contributes a computationally cheap workflow to the ends of detecting deepfaked video without much compromise in its precision. When compared to existing SOTA(State-of-the-art) models, the proposed approach is almost comparable in performance while using significantly less paramters, this was achieved through the extraction and manipulation of geometric features such as facial landmarks using optical flow techniques. The utility of this approach was further proven through its successful implementation as a workflow in the form of a web application.

II. Related Work

As the use of manipulated facial images and videos has become more prevalent, there has been a growing need to develop robust techniques for detecting such forgeries. In recent years, several approaches have been proposed to address this challenge. In this section, we discuss some of the related works in this area.

Two-Stream neural networks for tampered face detection [2], proposed by Zhou et al. (2017), utilizes two-stream neural networks for detecting tampered faces. One stream extracts features from the face region, while the other extracts features from the optical flow of the face region.

Meso-Net [1], introduced by Afchar et al. (2018), is a compact facial video forgery detection network that utilizes a series of meso-scale convolutional neural networks (CNNs) to identify manipulated videos.

Xception [3], proposed by Rossler et al. (2019), utilizes a modified version of the Xception CNN architecture to detect manipulated facial images. The network is trained on a large-scale dataset of real and manipulated images, using an adversarial training strategy to improve its robustness to various types of manipulations.

Capsule network-based approach [4], proposed by Nguyen et al. (2019), utilizes capsule networks to detect fake images and videos. The network is trained on a large-scale dataset of manipulated images, using a combination of reconstruction and classification loss functions to improve its performance.

FWA [5], proposed by Li and Lyu (2019), detects manipulated videos by detecting face warping artifacts. The approach utilizes a deep CNN to extract features from the face region, which are then used to identify regions with significant distortions.

DSP-FWA [6], proposed by He et al. (2015), uses spatial pyramid pooling in deep convolutional networks for visual recognition. The approach utilizes a hierarchical pooling strategy to extract features from different scales of the input image, which are then used to detect face warping artifacts.

Facial Landmarks Based approaches [7]: The work by Li et al. (2018) introduces a facial landmark-based approach called ”ictu oculi” for detecting AI-generated fake face videos. The term ”ictu oculi” is derived from Latin and translates to ”in the blink of an eye.” The approach outlined in the paper focuses on the eye region as a key source of information. It extracts relevant features from the eyes and utilizes them to detect irregular blinking patterns, which are indicative of AI-generated fake face videos. By analyzing and comparing the blinking patterns observed in the videos, the algorithm can identify discrepancies and irregularities that suggest manipulation.

Face Speaking expressions approach [8], proposed by Agarwal et al. (2019), detects manipulated videos by analyzing the speaking expressions of the subject in the video. The approach utilizes a deep CNN to extract features from the face region, which are then used to identify irregular patterns of facial movement that are characteristic of manipulated videos.

Optical Flow-based CNN approach [9], proposed by Amerini et al. (2019), detects deepfake videos using optical flow-based CNNs. The approach extracts features from the optical flow of the input video and uses them to detect inconsistencies in the motion patterns of the face region.

Lucas Kanade Algorithm [10] proposed by Simon Baker and Iain Matthews presents a comprehensive overview and analysis of the Lucas-Kanade algorithm, a widely used optical flow estimation technique. The authors discuss the evolution of the Lucas-Kanade algorithm over the course of 20 years, highlighting its strengths, limitations, and various extensions. The paper provides a unified framework that brings together different variations and improvements of the algorithm, shedding light on its underlying principles and offering insights into its applications and advancements in the field of computer vision.

III. Proposed Methodology

This paper aimed to develop a computationally efficient method for accurately detecting tampered videos. The proposed approach involves several steps. Firstly, the input video is divided into individual image frames. Next, a facial recognition classifier based on DLIB’s HOG+SVM algorithm is employed to identify the face in each frame. The classifier marks the 68 facial landmarks in each frame and records their corresponding positions, which provide crucial spatial information about the face.

Subsequently, to capture the motion patterns within the video, the optical flow of each landmark is computed using the Lucas Kanade equation, which calculates the displacement of pixels between consecutive frames. These optical flow vectors capture the dynamic changes in facial landmarks over time, enabling the detection of subtle movements.

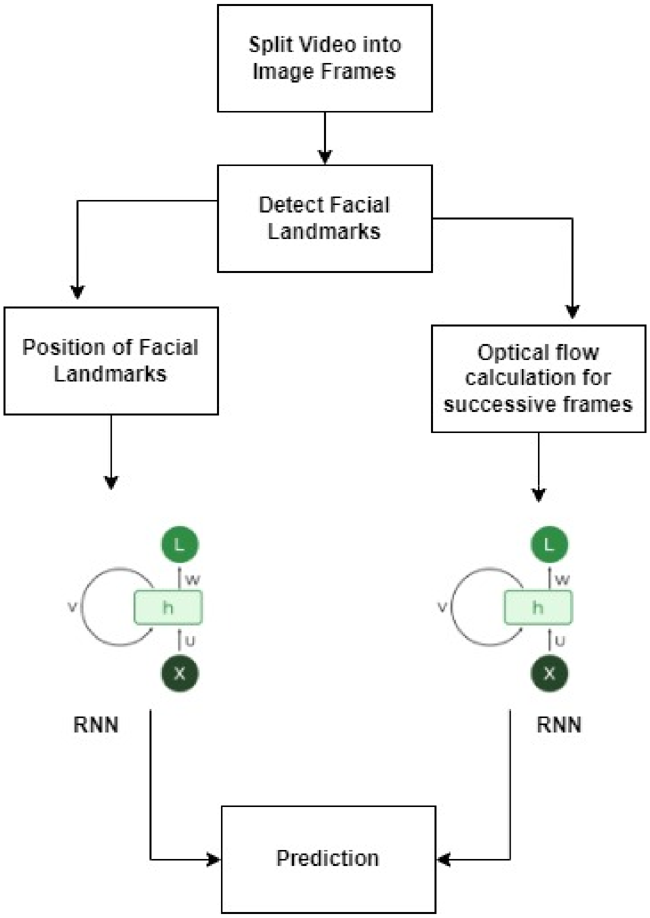

Split Video into Image Frames
Detect Facial Landmarks
Position of Facial Landmarks
Optical flow calculation for successive frames
L
w
v
h
u
X
RNN
L
w
v
h
u
X
RNN
Prediction

Fig. 2. Flow of Proposed approach

To effectively model the temporal dependencies and patterns, two bi-directional RNN models are employed. The choice of GRU (Gated Recurrent Unit) as the RNN architecture ensures better handling of long-term dependencies and mitigates the vanishing gradient problem. The first RNN receives the landmark positions at each timestep, allowing it to capture the spatial relationships of the facial landmarks over time. The second RNN takes the optical flow data from timesteps t and t+1, incorporating motion information between consecutive frames.

Finally, the outputs from both RNN models are concatenated, combining the spatial and temporal information, to make a prediction regarding whether the input video is a deepfake or not. By leveraging the strengths of both RNNs, the method achieves a more comprehensive analysis of the facial landmarks, enhancing the accuracy of deepfake detection.

IV. Results

The proposed approach utilized two Bi-directional GRU's with a total sum close to 250k trainable paramters. The GRU'S architecture consisted of multiple droput units where the probability was set to 0.25. The models were trained for 300 epochs with the help of the ADAM optimizer.

TABLE I
TEST DATA PERFORMANCE COMPARISON WITH OTHER METHODS

Method	Parameters	Dataset			Data Augmentation
		UADFV	FF++	Celeb-DF	
MesoNet [1]	0.03M	84.3%	84.7%	54.8%	✗
Two-stream [2]	-	85.1%	-	53.8%	✗
FWA [5]	26M	97.4%	80.1%	56.9%	✓
Capsule [4]	15M	61.3%	96.6%	57.5%	✗
Xception [3]	21M	80.4%	**99.7**%	48.2%	✗
DSP-FWA [6]	28M	**97.7**%	93.0%	**64.6**%	✓
Proposed Method	0.25M	96.2%	99.3%	56.4%	✗

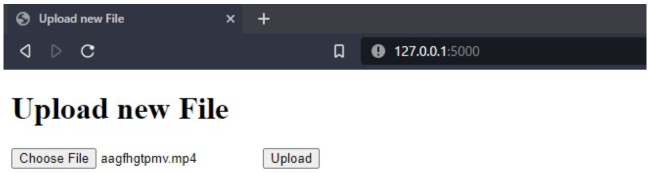
After comparison with past approaches it's evident that the proposed approach performs comparably to other state-of-the-art approaches [6] [3] while being significantly cheaper to train as displayed by the number of trainable parameters used. While the MesoNet model [1] uses lesser parameters it's performance is well below the required standard to be considered as a reliable approach.

Its evident that this approach contains 112x, 104x, 84x lesser trainable paramters than DSP-FWA [6], FWA [5] and Xception Net [3] respectively.

Thus, it's safe to say that the proposed approach is the perfect balance between Model performance and Computation cost.

V. Discussion & Future Work

The proposed workflow was successfully implemented in the form of a web app, making it more accessible and usable for general users who now have the capability of detecting deepfake videos by just uploading a video. This application is lightweight enough to be run on a basic computer without the worry of needing a GPU.

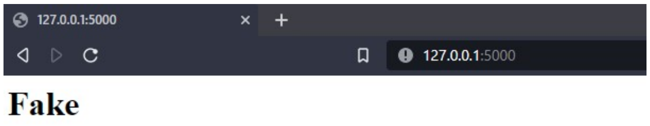


Upload new File

aagfhgtpmv.mp4

Fig. 3. Workflow implementation taking input

This work focuses on leveraging geometric features, specifically facial landmarks, for deepfake detection. However, an inherent limitation of the approach lies in its dependence on the accuracy of facial landmark detection. It is essential to acknowledge that misclassifications in identifying facial landmarks in each image can occur. Therefore, future research should conduct an in-depth study to evaluate the impact of landmark detectors on the performance of approaches that rely on the movement of facial landmarks to make predictions. Understanding the influence of landmark detection accuracy will enable researchers to identify potential sources of error and devise strategies to mitigate them.



Fake

Fig. 4. Workflow implementation displaying prediction

Moreover, incorporating additional denoising methods represents a promising avenue for improving the accuracy and robustness of the proposed approach. By integrating denoising techniques into the preprocessing pipeline, the system could effectively reduce noise and artifacts present in the video frames, thus enhancing the quality and reliability of the input data. Investigating various denoising methods, such as image filtering or deep learning-based denoising algorithms, could yield significant improvements in the overall performance of the proposed approach.

Furthermore, it is worth considering the exploration of alternative or complementary feature representations that can further enhance the discrimination between genuine and manipulated videos. While the use of facial landmarks provides valuable spatial and motion information, incorporating additional features or exploring alternative representations, such as texture descriptors or frequency-based features, could potentially improve the detection accuracy and bolster the system's resilience against sophisticated deepfake manipulation techniques.

Future research should delve deeper into understanding the impact of landmark detectors on the overall performance. Additionally, incorporating denoising methods and exploring alternative feature representations offer promising avenues to enhance the accuracy and robustness of deepfake detection systems.

Exploiting facial landmarks for denoising images and accurately tracking them in videos can enhance deepfake detection. Additionally, evolutionary algorithms [11] hold promise in optimizing neural network architectures, guiding evolution towards superior techniques or discovering simpler, less computationally intensive models with acceptable performance trade-offs [12] [13] [14].

VI. Conclusion

In conclusion, the proposed approach offers a solution that significantly limited the computation cost and training time while performing at the level of SOTA(State-of-the-art) Approaches and guaranteeing a robust and reliable solution. While also proving the effectiveness of Geometric features in boosting the performance of deep learning models for deepfake detection. In reducing computational cost it was seen that feature extraction and manipulation of facial landmarks was key, while using a bi-directional GRU helped in detecting abnormal movement of facial landmarks hinting that the video had been tampered which allowed the models to learn significantly useful information. It's also to be noted that this approach is heavily dependent on the accuracy of detecting these facial landmarks before optical flow is computed.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[2] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Chris- ¨ tian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In ICCV, 2019.

[4] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467, 2019.

[5] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence (TPAMI), 2015.

[7] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. CoRR, abs/1806.02877, 2018.

[8] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[9] Amerini, Irene, et al. "Deepfake video detection through optical flow based cnn." Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019.

[10] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision, 56(3):221–255, 2004

[11] A. Prem, A. Joshi, H. Madana, J. J and A. Arya, "Attention Based Evolutionary Approach for Image Classification," 2023 15th International Conference on Computer and Automation Engineering (ICCAE), Sydney, Australia, 2023, pp. 237-243, doi: 10.1109/ICCAE56788.2023.10111236.

[12] T. Vignesh, S. Reddy, S. Kumar, A. Chourey and C. P. Chavan, "Malware Detection Using Ensemble Learning and File Monitoring," 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2023, pp. 1-6, doi: 10.1109/ICSTSN57873.2023.10151567.

[13] V. L, N. J C, H. Prasad H R, J. K. A and C. Pomu Chavan, "Smart Farm Android Application Using IoT and Machine Learning," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-4, doi: 10.1109/I2CT57861.2023.10126447.

[14] A. Khubchandani, A. Ray, S. Shenoy, C. R. Cardoza and C. Pomu Chavan, "Emergency Reporting System for Animals," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9825241.