A Survey on Identification of Deepfake Videos using Artificial Intelligence

Shilpa B[1], Abhishek B K[2], Anush Kamath[3], Hemanth Bhat[4], Sathwik A M[5]

Assistant Professor, Department of Information Science & Engineering[1]
Students, Department of Information Science & Engineering[2,3,4,5]
Canara Engineering College, Mangalore, Karnataka, India

Abstract: *In the world of ever-expanding social media platforms, deepfakes are seen as the biggest threat posed by AI. There are many scenarios where realistic deepfakes with face swaps are used to create political pranks, and fake terrorist incidents and people intimidation are easy to imagine. Examples include Brad Pitt and Morgan Freeman fake videos. Advances in computing power have made deep-learning algorithm so powerful that it has become so easy to create human-synthesized indistinguishable videos, commonly known as deepfakes. It's easy to imagine scenarios where this realistic face could be traded for a fake video to do political pain, fake terrorist attacks and intimidation. This work describes a survey on novel deep learning- based techniques that are effective in telling apart phoney and authentic videos produced by AI.*

Keywords: Deepfakes, Neural Networks, long short-term memory, Convolutional Neural Networks

I. INTRODUCTION

Deepfake technology is a technology that processes videos using high-performance computers and deep learning. The result is a highly realistic video of an event that never happened. A deepfake is a digitally forged image or video of a person to make it look like someone else. This is next level fake content creation powered by artificial intelligence (AI). Impersonators can be celebrities, politicians, business owners, and other celebrities targeted by disinformation campaigns. However, deepfakes can also be used to spread false information about someone. Deepfake videos use two models of machine learning (ML).

One model creates fakes from a dataset of sample videos, and the other model tries to identify if a video is actually fraudulent. Deepfakes are probably credible enough for human viewers if the second model can no longer determine if the video is fake, this technique is called Generative Adversarial Network (GAN). GANs work well when larger the data set they can operate on. For this reason, many early deepfake videos tended to show politicians and celebrities in the entertainment industry, they have many videos that GANs can use to create highly realistic deepfakes. Most of the academic research on deepfakes focuses on detecting deepfake videos. One of the approaches to deepfake detection is to use algorithms to recognize patterns and track subtle discrepancies that appear in deepfake videos. A team from the University of Buffalo has unveiled a technique that uses the glow in the eyes of subjects to detect deepfakes with a high success rate, without the use of AI detection tools at least for now.

Deep fakes are increasingly being spread over social media platforms, which results in spamming and the spread of false information. Think of a profound fake of our prime leader threatening war on nearby nations or of a well-known star assaulting their supporters, deepfakes of this nature are nasty and threaten and mislead the general public. Deep fake detection is crucial in order to get out of this dilemma. We therefore require an innovative deep learning-based approach that is effective at separating synthetically produced misleading videos (also known as deep Fake Videos) from actual videos In order to identify deep fakes and stop them from spreading across the internet, it is crucial to develop technology that can detect fakes.

II. LITERATURE SURVEY

In the method proposed in the paper [1], this deepfake identification approach compared a specific convolutional neural network model with a source region of the face and its surrounding regions, there were two types of facial abnormalities in this study. This plan was entirely based on the idea that, deepfake classifiers which are currently existing in market

DOI: 10.48175/IJARSCT-7755

today, can produce videos of a certain resolution only and needs to be modified further to match the face areas which are to be replaced in the real video. This method did not take into account frame temporal analysis, which is an important factor in identifying deepfakes.

In contrast to the detection system mentioned above in paper [1], the method suggested in the paper [2] discusses an innovative method that uses the eye blinking as a differentiating factor between authentic and deep-fake films. The Long-term Recurrent Convolution Network(LRCN) was used to do temporal analysis on the clipped eye blinking frames. Because deepfake producing algorithms are so effective right now, the lack of eye blinking cannot be the only clue that something is fake. Additional variables like dental enchantment, face wrinkles, wrong brow alignment, etc. should be considered for the recognition of profound fakes.

In The method suggested in the paper [3], In a number of circumstances like machine-generated video identification and replay attack identification, we may utilise a capsule network to locate fraudulent, changed images and videos. Here, random noise during the training phase is ignored. Using random noise in the approach's training phase might not have been the greatest choice for this strategy. Even if the model performed well on their dataset, real-time data may not perform as well for it due to training-stage noise. It is usually better to use real-time, noise-free datasets while training models.

The approach recommended in the paper [4] is utilising RNN with consecutive frame processing in addition to the already trained ImageNet model. This method uses a software called PY-torch to train the model, which internally uses CUDA architecture. As of right now, the CUDA architecture only performs well with NVidia GPUs. Thus, it might not function on other GPUs.

Paper [5] gives us the most effective method of identifying deepfakes till date using Celebrity-deepfake and other new deepfakes detection dataset, by considering the largest number of fake detection algorithms and datasets for detection. In this evaluation, first try to use the average performance of detection as a challenge-level benchmark for various deepfake datasets, and then compare Celeb-DF with the latest deepfake datasets, achieving two goals. To do this we also test the effectiveness of present existing deepfake identification methods on various deepfakes, especially his Celeb-DF premium videos.

On pristine and deepfake portrait video pairings, the approach of paper [6] is to extract biological signals from facial regions. Feature vector and photoplethysmography (PPG) maps can be used to collect the signal features. The spatial coherence and temporal consistency can then be calculated using transformations, and the results can be improved by further training a probabilistic Support Vector Machine (SVM) and a Convolutional Neural Network. The average of genuine certainty is then used to classify the video as deepfake or faultless. Fraudulent Catcher accurately distinguishes between phoney and real content regardless of the source, content, resolution, and video quality. Since there is no discriminator, it is difficult to develop a differentiateble lossy function that stick to the recommended signal process techniques, which leads in the loss in their findings to retain biological information.

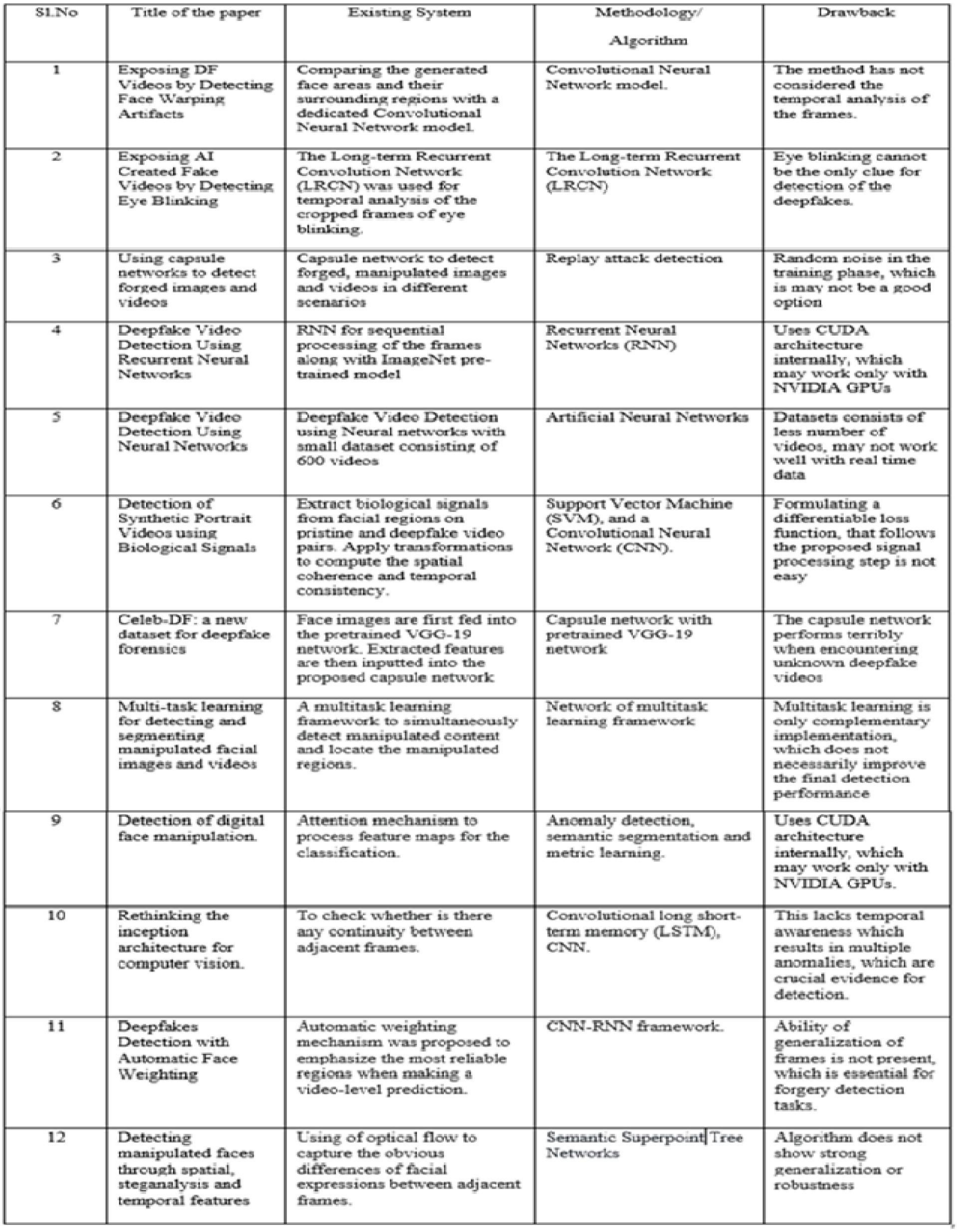
In technology used in paper [7], the facial photos are initially sent into the pretrained VGG (Visual geometry group) network in this manner, proposed capsule network which consists of a number of main capsules and two output capsules, is fed with the extracted features. A dynamic routing algorithm dynamically calculates the agreement between the features retrieved by the primary capsules and the appropriate output capsule is then sent the findings. Capsule networks and a dynamic routing strategy function well together to detect tampering. Capsule networks still need to be improved in order to detect high-fidelity films, as evidenced by how poorly they performed when confronted with unknown deepfake videos.

In work [8], this method is a semi-supervised learning approach for simultaneously detecting changed information and identifying changed regions. This approach develops a multi-task learning framework. Supervised multitask learning is only a complementary implementation of such schemes, but it may or may not improve the final recognition accuracy. The use of attentional processes has facilitated further development.

The foundation of the approach in paper [9] is a conventional frequency domain analysis followed by a simple classifier. The technique demonstrated very strong results using only a few annotated training samples and even obtained good accuracies in wholly unsupervised circumstances, in contrast to earlier systems that require enormous amounts of labelled data to be supplied. This method created a new benchmark called Faces-HQ by combining multiple available datasets of real and artificial faces for the evaluation on high resolution face photos. With such high-resolution

DOI: 10.48175/IJARSCT-7755

images, this method may train on as few as 20 annotated examples and achieve a perfect classification accuracy of 100%. In a second trial, this technique evaluates medium-resolution photos from the CelebA dataset with 100% supervised accuracy and 96% unsupervised accuracy.



Sl.No	Title of the paper	Existing System	Methodology/ Algorithm	Drawback
1	Exposing DF Videos by Detecting Face Warping Artifacts	Comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network model.	Convolutional Neural Network model.	The method has not considered the temporal analysis of the frames.
2	Exposing AI Created Fake Videos by Detecting Eye Blinking	The Long-term Recurrent Convolution Network (LRCN) was used for temporal analysis of the cropped frames of eye blinking.	The Long-term Recurrent Convolution Network (LRCN)	Eye blinking cannot be the only clue for detection of the deepfakes.
3	Using capsule networks to detect forged images and videos	Capsule network to detect forged, manipulated images and videos in different scenarios	Replay attack detection	Random noise in the training phase, which is may not be a good option
4	Deepfake Video Detection Using Recurrent Neural Networks	RNN for sequential processing of the frames along with ImageNet pre-trained model	Recurrent Neural Networks (RNN)	Uses CUDA architecture internally, which may work only with NVIDIA GPUs
5	Deepfake Video Detection Using Neural Networks	Deepfake Video Detection using Neural networks with small dataset consisting of 600 videos	Artificial Neural Networks	Datasets consists of less number of videos, may not work well with real time data
6	Detection of Synthetic Portrait Videos using Biological Signals	Extract biological signals from facial regions on pristine and deepfake video pairs. Apply transformations to compute the spatial coherence and temporal consistency.	Support Vector Machine (SVM), and a Convolutional Neural Network (CNN).	Formulating a differentiable loss function, that follows the proposed signal processing step is not easy
7	Celeb-DF: a new dataset for deepfake forensics	Face images are first fed into the pretrained VGG-19 network. Extracted features are then inputted into the proposed capsule network	Capsule network with pretrained VGG-19 network	The capsule network performs terribly when encountering unknown deepfake videos
8	Multi-task learning for detecting and segmenting manipulated facial images and videos	A multitask learning framework to simultaneously detect manipulated content and locate the manipulated regions.	Network of multitask learning framework	Multitask learning is only complementary implementation, which does not necessarily improve the final detection performance
9	Detection of digital face manipulation.	Attention mechanism to process feature maps for the classification.	Anomaly detection, semantic segmentation and metric learning.	Uses CUDA architecture internally, which may work only with NVIDIA GPUs.
10	Rethinking the inception architecture for computer vision.	To check whether is there any continuity between adjacent frames.	Convolutional long short-term memory (LSTM), CNN.	This lacks temporal awareness which results in multiple anomalies, which are crucial evidence for detection.
11	Deepfakes Detection with Automatic Face Weighting	Automatic weighting mechanism was proposed to emphasize the most reliable regions when making a video-level prediction.	CNN-RNN framework.	Ability of generalization of frames is not present, which is essential for forgery detection tasks.
12	Detecting manipulated faces through spatial, steganalysis and temporal features	Using of optical flow to capture the obvious differences of facial expressions between adjacent frames.	Semantic Superpoint Tree Networks	Algorithm does not show strong generalization or robustness

Figure 1: Analysis Table

DOI: 10.48175/IJARSCT-7755

The idea of using RNNs to find fake videos in a research paper [10] revealed that when faces are generated frame by frame, the autoencoder cannot recognize previously generated faces at all. This temporary lack of awareness leads to various anomalies that are key evidences of deepfake detection. A trainable iterative deepfake video detection system is presented to verify the continuity between adjacent frames. The convolutional LSTM structure is a key feature of the proposed system for analysing frame sequences and it ccts as a building block.

In publication [11], it was proposed to use an autonomous weighting technique within the CNN-RNN architecture. Because the quality of faces in a given frame is not very good, an algorithmic weighting method was recommended that stresses the most trustworthy areas when giving predictions of videoframe-level features. Experiments using the DFDC dataset showed that the combination of CNN and RNN yields high detection accuracy. When it comes to the problem of counterfeit detection, the ability to generalize is as important as the algorithm's resilience.

In paper [12], obvious variations in facial emotions across subsequent frames are captured via optical flow. These investigations, however, did not demonstrate robustness or great generalisation. In order to address this issue, the author presented the SSTNet manipulation detection framework, which takes advantage of both low-level artefacts and temporal inconsistencies.

III. PROPOSED SYSTEM

After analysing most of the different available methods, we propose to develop a machine-learning-based approach to classify videos as synthetic or concrete, along with the sureness level of the method which is proposed. It can be done by building a model for processing 1 second of video (10 frames per second), by doing this the algorithm can predict the outcome with a high degree of accuracy. Building a model by combining a pre-trained ResNext convolution neural network model for extracting frame-level features and a long short-term memory for time series processing that helps identify changes between t and t-1 frames. Our system analyses the video frame by frame, looking for unexpected frames to determine if the video is real or artificial.

IV. CONCLUSION

There is a need to design and develop deep learning algorithms to classify videos into deepfakes or originals. Using artificial intelligence tools to create deep fakes has become straightforward and very easy process, but identifying these deep fakes is a real difficult challenge. There are already many examples in history where deepfakes being used as powerful tools to create political tensions, fake terrorist events, and extortion of the public. So, it becomes very important build a powerful algorithm to detect these deepfake and avoid the percolation of deepfake through social media.

REFERENCES

[1]. Yuezun Li, Siwei Lyu, “Exposing DF Videos By Detecting Face Warping Artifacts” in arXiv:1811.00656v3.

[2]. Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arXiv:1806.02877v2.

[3]. Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ” in arXiv:1810.11215.

[4]. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks" 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[5]. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. "Deepfake Video Detection Using Neural Networks" Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008. Anchorage.

[6]. Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.

[7]. Li, Y., et al.: “ Celeb-DF: a new dataset for deepfake forensics” arXiv preprint arXiv:1909.12962. (2019).

[8]. Nguyen, H.H., et al.: “multi-task learning for detecting and segmenting manipulated facial images and videos” arXiv preprint arXiv:1906.06876. (2019)

[9]. Dang, H., et al.:“ On the detection of digital face manipulation” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, pp. 5781–5790 (2020).

[10]. Szegedy, C., et al.: “Rethinking the inception architecture for computer vision” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016).

[11]. Montserrat, D.M., et al.: “Deepfakes Detection with Automatic Face Weighting” In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2851–2859. (2020).

[12]. Wu, X., et al.: SSTNet: “Detecting manipulated faces through spatial, steganalysis and temporal features” In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2952–2956. IEEE (2020).