iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44110

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

DeepFake Detection: A survey of countering malicious Deep-Fakes

Kimaya Kulkarni[1], Sahil Khanolkar[2], Yash Walke[3], Rahul Sonkamble[4]
[1, 2, 3]*Student* [4]*Professor, Department of Computer Science, MIT ADT University Pune, Loni Campus, India*

Abstract: ***The free access to large-scale public databases, together with the fast progress of deep learning techniques, in particular with the Generative Advertising Networks, has led to the creation of very realistic fake content with its corresponding society in this time of false or fake news. This survey provides a thorough review of techniques to detect DeepFake manipulations.***

Keywords: ***DeepFake Detection, Computer Visions, Deep Learning***

I. INTRODUCTION

Deepfake (coming from "deep learning" and "fake") is a method that will superimpose face images of a target person onto a video of a source person to create a video of the target person doing or saying things the source person does. Fake images and videos, including facial feature information, are generated by digital manipulation. They have become a greater public concern than ever before, particularly with the DeepFake methods.

This term was coined by a Reddit user named "deepfakes" in late 2017 to have developed a machine-learning algorithm that helped him to transpose celebrity faces into porn videos. Deepfakes are being used to swap the faces of celebrities or targeted politicians with bodies in pornographic images and videos.

Deepfakes can thus be used to incite political or religious tensions between countries, as well as to deceive the public and affect results in election campaigns, or create chaos in financial markets by creating fake news. It may be even used to generate fake satellite images of the Earth and make it contain objects that do not exist in the real world to confuse military analysts, e.g., like creating a fake bridge across a river, since there is no such a bridge in the real world. This will mislead a force of troops who have been guided to cross the bridge in a real battle.

There are also positive uses of deepfakes, such as creating voices for those who have lost theirs or updating episodes of movies without reshooting them. However, the number of malicious applications of this deepfake outnumbers the positive ones. The method that creates those manipulated images and videos has become much simpler today as it needs as little as an identity photo or a short video of a target individual. Less and less effort is required to produce stunningly convincing tempered footage. Recent advancements can even create a deepfake from a still image. That's why Deepfakes are a threat affecting not only public figures but also ordinary people at this time. This survey includes techniques for detecting deepfakes. The paper examines various methods for detecting deepfakes. Deepfake detection is typically regarded as a binary classification problem in which classifiers are used to distinguish between authentic and tampered videos.

1) *Temporal Features across Video Frames:* Observations that temporal coherence is not effectively enforced in the synthesis process of deepfakes, we leveraged the use of spatio-temporal features of video streams to detect deepfakes. On the other hand, the use of a physiological signal, eye blinking, to detect deepfakes was proposed, based on the observation that a person in deepfakes blinks much less frequently than a person in untampered videos. Deepfake algorithms cannot generate fake faces that blink normally without access to images of people blinking. In other words, the blinking rates in deepfakes are much lower than those in normal videos. After a few steps of preprocessing such as aligning faces and extracting and scaling the bounding boxes of eye landmark points to create new sequences of frames, these cropped eye area sequences are distributed into long-term recurrent convolutional networks (LRCN) for dynamic state prediction. Eye blinking shows strong temporal dependencies, and thus the implementation of LSTM helps to capture these temporal patterns effectively. The blinking rate is calculated based on the prediction results where a blink is defined as a peak above the threshold of 0.5 with a duration of fewer than 7 frames. This method is tested using a data set compiled from the internet 49 interview and presentation videos, as well as their corresponding spoof videos, were generated by the deep-fake algorithms. The experimental results indicate the promising performance of the proposed method in detecting fake videos, which can be further improved by considering the dynamic pattern of blinking, e.g., highly frequent blinking may also be a sign of tampering.

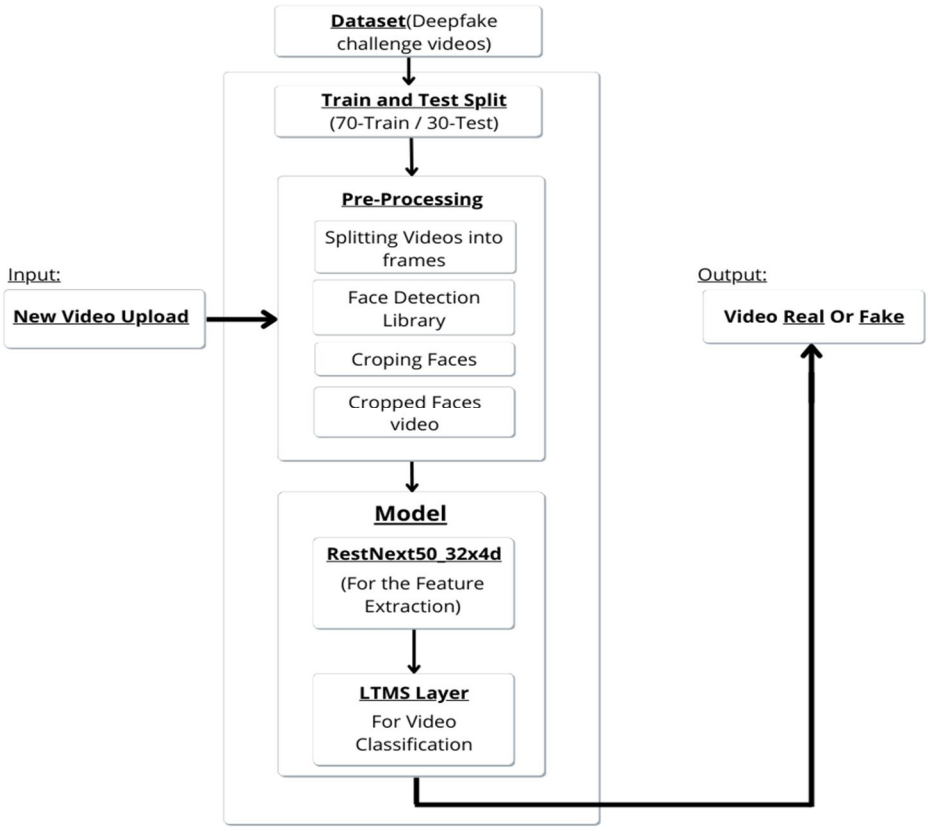
2) *Visual Artifacts within Video Frame:* As can be noticed in the previous subsection, the methods using temporal patterns across video frames are mostly based on deep recurrent network models to detect deepfake videos. This subsection investigates the other approach that normally decomposes videos into frames and explores visual artefacts within single frames to obtain discriminant features. To distinguish between fake and real videos, the features are distributed in either a deep or shallow classifier. We thus group methods in this subsection based on the types of classifiers, i..,. either deep or shallow..
3) *Deep Classifiers:* Deepfake videos are normally created with limited resolutions, which require an affine face warping approach to match the configuration of the original ones. Because of the resolution inconsistency that is between the warped face area and the surrounding context, this process leaves artefacts that can be detected by CNN models such as VGG16, ResNet50, ResNet101, and ResNet152.
4) *Shallow Classifiers:* Deepfake detection is primarily based on artefacts or inconsistencies in intrinsic features between fake and real images or videos. Yang et al. proposed a detection method based on differences in 3D head poses, which include head orientation and position and are estimated using 68 facial landmarks in the central face region.
5) *Optical Flow-based CNN:* Optical flow [4, 3] is a vector field calculated on two consecutive frames f(t) and f(t + 1) to extract apparent motion between the observer and the scene. We specifically hypothesise that the optical flow can exploit motion discrepancies between synthetically generated frames and those naturally generated by a video camera. It is to be more noticeable in optical flow matrices and the introduction of fake, as well as all the unusual movements of the entire face.
6) *Fake spoter:* Wang et al. hypothesised that monitoring neuron behaviour could also be used to detect fake faces because layer-by-layer neuron activation patterns could capture more subtle features important for facial manipulation detection systems. It was extracted from deep face recognition systems as the features neuron coverage behaviours of real and fake faces. Using the FaceNet model, this model had an overall accuracy of 84.7 percent in detecting fakes.
7) *Pixel Co-occurrence:* Fake detection systems based on steganalysis were also investigated. Nataraj et al. proposed a detection system based on pixel co-occurrence matrices and Convolutional Neural Networks (CNN). The authors conducted an interesting analysis to see the robustness of the proposed approach against fake images generated by different GAN architectures (CycleGAN vs. StarGAN), with good generalisation results. This detection approach was later implemented using images from the 100K-Faces database, achieving an EER of 12.3 percent for the best fake detection performance at the time.

II. LITERATURE REVIEW

Motivated by the ongoing success of digital face manipulations, particularly DeepFakes, this survey offers various detection techniques over time. In general, most current face manipulations appear to be easy to detect in controlled scenarios, i.e., when fake detectors are evaluated in the same conditions for which they were trained. It has been demonstrated that the majority of the benchmarks included in this survey achieve very low error rates in manipulation detection. This scenario, however, may not be very realistic because fake images and videos are commonly shared on large social networks and suffer from a wide range of variations such as compression level and resizing, noise, and so on. On the other hand, current detection methods are primarily focused on the drawbacks of deepfake generation pipelines, i.e. identifying competitors' weaknesses in order to attack them. This type of information and knowledge is not always available in the advertising environment, where attackers primarily try not to expose such deepfake creation technologies. Furthermore, facial manipulation techniques are constantly being refined. These factors motivate further research into the fake detectors' ability to generalise against unknown conditions. Another research direction could be to incorporate detection methods into distribution platforms such as social media to increase its overall effectiveness in dealing with the widespread impact of deep-fake. On these platforms, a screening or filtering mechanism based on effective detection methods can be implemented to aid in the detection of deepfakes. Videos and photographs have been widely used as evidence in police investigations and legal proceedings. Digital media forensics experts with a background in computer or law enforcement and experience collecting, examining, and analysing digital information may present them as evidence in a court of law. This approach can be used by intelligence services attempting to influence decisions made by influential figures such as politicians who are at the forefront of national and international security threats. Detecting the deepfake alarming issue, The research community has concentrated on developing deepfake detection algorithms, with numerous results published. Using detection methods to detect deepfakes is critical, but understanding the true intent of those who publish deepfakes is even more critical. This necessitates user judgement based on the social context in which deepfake is discovered, for example, who distributed it and what they said about it. A study on the social context of deepfakes to assist users in making such decisions is thus worthwhile. Machine learning and AI algorithms were used to help determine the authenticity of digital media and produced accurate and reliable results.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

III. PROPOSED METHODOLOGY



IV. DATASET

We are using Kaggle's Deepfake challenge dataset [3], which contains 3000 videos from randomly collected sources. Our dataset is divided into 70 percent train dataset and 30 percent test dataset.

V. PREPROCESSING

Dataset preprocessing includes splitting videos into frames and ten, followed by face detection, cropping of the detected frame, and creating a new face cropped dataset. The remaining frames will be ignored during preprocessing

VI. MODEL

The model is made up of resnext50 32x4d and an LSTM layer. The Data Loader loads and divides preprocessed face cropped videos into train and test sets. The frames from the processed videos are then passed to the model in mini batches for training and testing.

VII. FEATURE EXTRACTION WITH RESTNEXT50

The ResNext50 is being used to extract the features and also accurately detecting the frame level of the features. The CNN Network will then be tuned by adding extra layers and selecting a reasonable learning rate to converge with the gradient. Following the last pooling layers, there are 2048-dimensional feature vectors that will be used for sequential LSTM input..

VIII. LTMS FOR SEQUENTIAL PROCESSING

Assume we take ResNext CNN feature vectors of input frames as input and train a 2-node neural network with a probability that the sequence is part of a deep fake video. The main challenge here is to design a model that can recursively process a sequence in a meaningful pattern. Now, we propose the use of a 2048 LSTM unit with a 0.4 chance of dropout, which is capable of achieving this goal. LSTM is used for sequentially processing frames in order to perform temporal analysis of a video by comparing the frame at "t"and second with the frame at "t-n" seconds. Where n is the number of frames preceding the t.

IX. PREDICTION

When a new video is uploaded, it will go through the same preprocessing step to obtain the cropped video with the face. The data will then be passed directly to the trained model, which will predict whether the video is real or fake..4

X. RESULTS

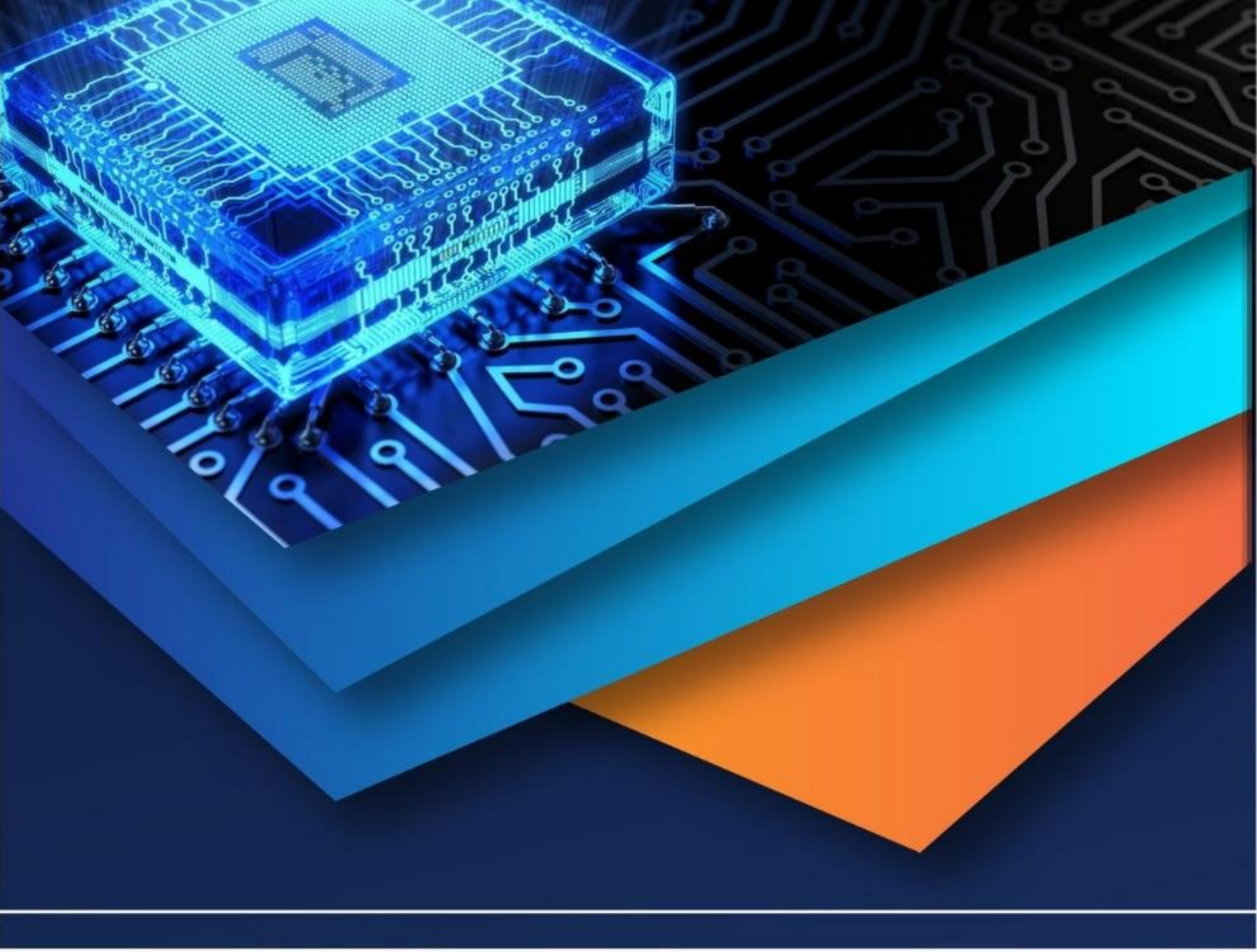
The model's output will indicate whether the video is real or fake based on the model's confidence.

XI. LIMITATIONS

Audio altered deepfakes are not detected in the current module, but this can be accomplished in the future.

REFERENCES

[1] Yuezun Li, Siwei Lyu, “ExposingDF 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.
[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
[4] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
[5] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.
[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
[7] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[9] An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnetand-its-variants-5281e2f56035
[10] Long Short-Term Memory: From Zero to Hero with Pytorch: https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/
[11] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_mod els_tutorial.html
[12] https://discuss.pytorch.org/t/confused-about-the-image-preprocessing-in-classification/3965
[13] https://www.kaggle.com/c/deepfake-detection-challenge/data
[14] https://www.researchgate.net/publication/336058980_Deep_Learning_for_Deepfakes_Creation_and_Detection_A_Survey
[15] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.
[16] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI.
[17] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, “Transferable deep-CNN features for detecting digital and print-scanned morphed face images,” in CVPRW. IEEE, 2017.
[18] Tiago de Freitas Pereira, Andr´e Anjos, Jos´e Mario De Martino, and S´ebastien Marcel, “Can face anti spoofing countermeasures work in a real world scenario?,”in ICB. IEEE, 2013.
[19] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “Distinguishing computer graphics from natural images using convolution neural networks,” in WIFS. IEEE, 2017.
[20] F. Song, X. Tan, X. Liu, and S. Chen, “Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients,” Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014. [21]D. E. King, “Dlib-ml: A machine learning toolkit,” JMLR, vol. 10, pp. 1755–1758, 2009



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY