International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Review of Deepfake Detection Techniques

Karthik P C
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy
Thrissur, India

Sanjana S
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy
Thrissur, India

M P Adithya Vijayan
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy
Thrissur, India

Thushara P
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy
Thrissur, India

Aswathy Wilson
Asst. Prof, Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy
Thrissur, India

***Abstract*:- Deep fake videos are AI-generated videos that look real but are fake. Deep fake videos are generally created by face-swapping techniques. It started as fun but like any technology, it is being misused. In the beginning, these videos could be identified by human eyes. But due to the development of machine learning, it became easier to create deep fake videos. It has almost become indistinguishable from real videos. Deep fake videos are usually created by using GANs (Generative Adversarial Network) and other deep learning technologies. The danger of this is that technology can be used to make people believe something is real when it is not. Smartphone desktop applications like FaceApp and Fake App are built on this process. These videos can affect a person's integrity. So identifying and categorizing these videos has become a necessity.This paper evaluate methods of deepfake detection and discuss how they can be combined or modified to get more accurate results. Hopefully, we will be able to make the internet a safer place.**

Index Terms:-Detection, Classification, Deepfake Video, Generative Adversarial Network, Artificial Neural Netwok, Machine Learning

I. INTRODUCTION

A growing disquiet as settled around the emerging deepfake that make it possible to create evidence of scenes that has never ever happened. Celebrieties and politcians are the ones who are considerably affected by this.Deepfake can optimally stitch anyone into a video or photo that they never have actually knowledge with.Nowadays since technologies are elevating widely the systems can synthesize images and videos more quickly. A creator would first train a neural network on many hours of real video footage to give it a realistic understanding of what he or she looks like on many angles or lighting inorder to create a deepfake video of someone.Then they would combine the trained network into graphics techniques to superimpose a copy of person into different one.

AI-Generated synthetic media, which is also known as deepfakes, ofcourse have many positive sides.Deepfakes en-ables clear benefits in areas such as education, accessibility, film production, criminal forensics, and artistic expression. It can ac-celerate the artistic quest into equity. Creative use of synthetic voice and video can enhance overall success and learning outcomes with scale and limited expenditure. Deepfakes can democratize VFX technology as a strong tool for independent story tellers. It could give individuals new tools for self-expression and amalgamation in online world. Deepfakes also has disadvantages which affects different groups of our society. It is being used to revenge porn to defame certain celebrities, creating fake news and propaganda etc...As soon as these fake videos goes viral people believe initially ,and keep on sharing with others makes the targeted person embarressed watching this fake stuff.

Until or unless an official statement of targeted one not comes,many keep on believing this stuff making their life difficult and are followed attack by the society in platforms like Facebook,Twitter via Instagram.

II. DISSCUSSION

Different types of deepfake detection methods are available today and each method has its own advantages and disadvan-tages. This paper tries to evaluate such methods from different papers and points out how these methods can be combined and modified in a new project in order to get more accurate results.

In the paper [1] "Deepfake Video Detection Using Recurrent Neural Network", David Guera and Edward J Delp propose a temporalaware pipeline to automatically detect deepfake videos. In order to detect deepfake videos, firstly we need to have a clear knowledge of how it is created, which helps us to understand the weak points of deepfake generation so that by exploiting those weak points, deepfake detection can be done. In the approach discussed in this paper, framelevel scene inconsistency is the first feature that is exploited. If the encoder is not aware of the skin or other scene information, there will be boundary effects due to a seamed fusion between the new face and the rest of the frame which is another weak point. The third major weakness that is exploited here is the source of multiple anomalies and leads to a flickering phenomenon in the face region. This flickering is common to most of the fake videos. Even though this is hard to find with our naked eye, it can be easily captured by a pixellevel CNN feature extraction. Dataset used here contains 300 videos from the HOHA dataset. Preprocessing steps are clearly described in this paper. Here the proposed system is composed of a convolution LSTM structure for processing frame sequences. CNN for frame feature extraction and LSTM for temporal sequence analysis are the 2 essential components in a convolutional LSTM. For an unseen test sequence,set of features for each frame are generated by CNN. After that features of multiple consecutive frames are concatenated and pass them to the LSTM for analysis which finally produces an estimated likelihood of the sequence being either a deepfake or nonmanipulated video. With less than 2 seconds this system could accurately predict if the fragment being analyzed comes from a deepfake video or not with an accuracy greater than 97 percentage.

In the paper [2] ”Effective and Fast Deepfake detection method based on Haarwavelet Transform” by Mohammed Akram Younus and Taha Mohammed Hasan describes another method to detect deepfake videos by haar wavelet transform. The method described here take the advantage of the fact that during deepfake video generation, deepfake algorithm could only generate fake faces with specific size and resolution. In order to match and fit the arrangement of the source’s face on original videos, a further blur function must be added to the synthesized faces. This transformation causes exclusive blur inconsistency between the generated face and its background outcome deepfake videos. The method detects such inconsis-tency by comparing the blurred synthesized areas ROI and the surrounding context with a dedicated Haar Wavelet transform function. The two main advantage of this Haar Wavelet transform function is that it first distinguishes different kinds of edges and the retrieves sharpness from the blurred image. It is very effective and fast since the uniform background of the faces in the images will have no effect and it does not need to reconstruct the blur matrix function. To estimate the blur extend, two methods such as direct and indirect can be used. Direct method can measure the blur function extent by testing some distinctive features in an image. Eg: edge feature. The indirect method depends on the blur reconstruction function when the H matrix is unknown (H matrix is blur’s estimation and blur identification). Dirac structure, Step structure, and Roof structure are the different types of edges present in an image. A blur extends is identified by taking the sharpness of roof structure and G step structure into account. The sharpness of the edge is indicated by the parameter (0¡¡ /2), if is larger means the edge is sharper. By comparing the blur extent of the ROI with the blur extend of the rest of the image, we can determine if the images(frames of video) have tampered or not. UADFV dataset which contains 49 unmanipulated and 49 manipulated videos is used here. Videos are divided into frames and from each frame, the face region is extracted and deepfake detection algorithm using haar wavelet transform is applied. This algorithm is clearly described in this paper. This proposed model contains an accuracy of 90.5 percentage.

In the paper [3],” OC Fake Dect: Classifying Deepfakes using OneClass Variational Autoencoder” by Hasam Khalid and Simon S. Woo, the proposed model needs only real images for training. As new methods for deepfake video creation are increasing today due to technology advancement, for a model to detect such videos, datasets containing fake videos are very scarce for training. It affects the model’s accuracy. But in the model proposed in this paper needs only real videos for training so that it can overcome data scarcity limitation.FaceForensic ++ is the dataset used here. It contains real images and 5 sets of fake images: FaceSwap dataset, Face2Face dataset (F2F), Deepfake dataset(DF), Neural Tex-tures dataset (NT), Deepfake detection Dataset(DFD), After collecting the video datasets, they are converted into frames and face detection and alignment is done using MTCNN. One class variational encoder is used here. It consists of an encoder and a decoder. At the encoder side, image is given as input, and scaling is done using convolutional layer and mean and variance is calculated and the result is given as input into decoder and the RMSE value is calculated which is low for real image and high for fake images. Two methods are discussed in this paper: OCFakeDect1 and OCFakeDect2. In OCFakeDect1 from input and output image itself, reconstruction score is computed directly and in OCFakeDect2 contains additional encoder structure which computes reconstruction score from input and output latent information. Eventhough it has 97.5 percentage accuracy, better performance is only on NT and DFD datasets.

In the paper [4] ”Deep Fake Source Detection via Interpreting Residuals with Biological Signals”, Umur Aybars Ciftci, Ilke Demir and Lijun Yin presented a deep fake source detection technique via interpreting residuals with biological signals. To their knowledge it is the first method to apply biological signals for the task of deep fake source detection. In addition to this they had experimentally validated this method through various ablation studies their experiments had achieved 93.39accuracy on FaceForensics++ dataset on source detection from four deep fake generators and real videos. Other than this they had demonstrated the adaptability of the approach to new generative models, keeping the accuracy unchanged. After studying biological signal analysis on deepfake videos, it is found that ground truth PPG data along side original and manipulated videos enabled new direction in research on deepfake analysis and detection. In the next stage of their work, . With ground truth PPG, they planned to create a new dataset with certain distribution variation as well as source variations. It is worth noting that these work looks for generator signatures in deep fakes, while the prevailing work reported by Ciftci et al. [23] looks for signatures in real videos. For detecting signatures on both real and fake videos, a holistic system combining these two perceptives can be developed. They posed this idea for their immediate future work.

In the paper [5] " Digital Forensics and Analysis of Deep-fake Videos" by Mousa Tayseer Jafar, Muhammed Ababneh, Muhammad Al-Zoube, Ammar Elhassan proposed a method detect deepfakes using mouth features.Nowadays deepfake videos can have an adverse effect on a society and these videos can challenge a person's integrity. Deepfake is a video that has been constructed to make a person appear to say or do something that they never said or did. Therefore there shows the increase in demand to detect methods to identify deepfakes. In this proposed model mouth features is used to detect deepfake video. A deepfake detection model with mouth features(DFT-MF),using deep learning approach to detect deepfake videos by isolating analysing and verifying lip/mouth movement is designed and implemented here. Here, dataset contains the combination of fake and real videos. Some preprocessing is done prior to performing analysis. Then the mouth area is been cropped from a face. There will be fixed coordinates for face. Working on a typical image frame facial landmark detector is used to estimate the location of 68 (X,Y)coordinates. In next step all face containing closed mouth is excluded and face with only open mouth is been tracked having teeth with reasonable clarity. CNN is used to classify videos into fake or real based on a threshold number of fake frames based on calculating three variable word per sentence, speech rate and frame rate. If the number of fake frames is greater than 50 the video is been classified as fake or else as real.

III. CONCLUSION

In this paper, we have presented a brief review of some papers which describes different methods to detect deepfake videos and images. Also how those methods can be modified or combined in our new project inorder to get more accurate results than prevailing methods. Hope we will succeed in our project.

REFERENCES

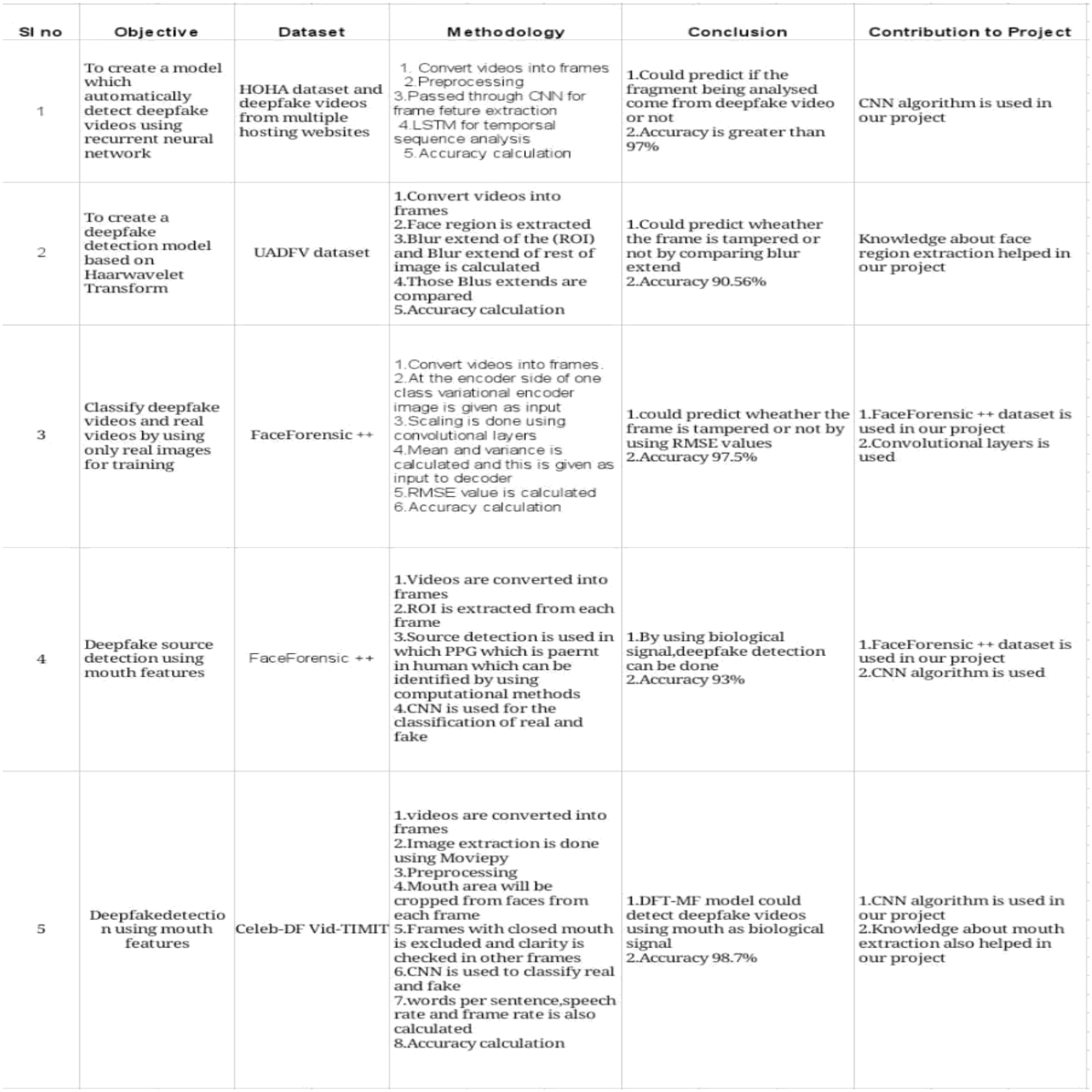
[1] D. Guera¨ and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6.

[2] M. A. Younus and T. M. Hasan, "Effective and fast deepfake detection method based on haar wavelet transform," in 2020 International Confer-ence on Computer Science and Software Engineering (CSASE), 2020, pp. 186–190.

[3] H. Khalid and S. S. Woo, "Oc-fakedect: Classifying deepfakes using one-class variational autoencoder," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 2794–2803.

[4] U. Ciftci, I. Demir, and L. Yin, "How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals," 08 2020.

[5] M. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," 04 2020, pp. 053–058.



Sl no	Objective	Dataset	Methodology	Conclusion	Contribution to Project
1	To create a model which automatically detect deepfake videos using recurrent neural network	HOHA dataset and deepfake videos from multiple hosting websites	1. Convert videos into frames 2.Preprocessing 3.Passed through CNN for frame feture extraction 4.LSTM for temporsal sequence analysis 5.Accuracy calculation	1.Could predict if the fragment being analysed come from deepfake video or not 2.Accuracy is greater than 97%	CNN algorithm is used in our project
2	To create a deepfake detection model based on Haarwavelet Transform	UADFV dataset	1.Convert videos into frames 2.Face region is extracted 3.Blur extend of the (ROI) and Blur extend of rest of image is calculated 4.Those Blus extends are compared 5.Accuracy calculation	1.Could predict wheather the frame is tampered or not by comparing blur extend 2.Accuracy 90.56%	Knowledge about face region extraction helped in our project
3	Classify deepfake videos and real videos by using only real images for training	FaceForensic ++	1.Convert videos into frames. 2.At the encoder side of one class variational encoder image is given as input 3.Scaling is done using convolutional layers 4.Mean and variance is calculated and this is given as input to decoder 5.RMSE value is calculated 6.Accuracy calculation	1.could predict wheather the frame is tampered or not by using RMSE values 2.Accuracy 97.5%	1.FaceForensic ++ dataset is used in our project 2.Convolutional layers is used
4	Deepfake source detection using mouth features	FaceForensic ++	1.Videos are converted into frames 2.ROI is extracted from each frame 3.Source detection is used in which PPG which is paernt in human which can be identified by using computational methods 4.CNN is used for the classification of real and fake	1.By using biological signal,deepfake detection can be done 2.Accuracy 93%	1.FaceForensic ++ dataset is used in our project 2.CNN algorithm is used
5	Deepfakedetection using mouth features	Celeb-DF Vid-TIMIT	1.videos are converted into frames 2.Image extraction is done using Moviepy 3.Preprocessing 4.Mouth area will be cropped from faces from each frame 5.Frames with closed mouth is excluded and clarity is checked in other frames 6.CNN is used to classify real and fake 7.words per sentence,speech rate and frame rate is also calculated 8.Accuracy calculation	1.DFT-MF model could detect deepfake videos using mouth as biological signal 2.Accuracy 98.7%	1.CNN algorithm is used in our project 2.Knowledge about mouth extraction also helped in our project