HBRP PUBLICATION

Advancement of Computer Technology and its Applications
Volume 4 Issue 2

A Survey on Deepfake Analysis and Recognition using Deep Learning

***Deepak N R*[1], *Mohamed Aweez Akram*[2*], *Mohammed Adnan K*[3], *Mohammed Shibil*[4], *Shameema Banu R*[5]**
[1]Professor, [2,3,4,5]Student,
Department of CSE, HKBKCE, Bangalore, India

****Coressponding Author***
E-mail Id:-*aweezakram1999@gmail.com*

ABSTRACT

The recent advancements in the field of artificial intelligence, deep learning and image processing, has led to the rise of a software called deepfake. It is a tool that can produce extreme transformations in human faces like aging, gender swap, etc., and can make someone say and do things which never happened in reality. The resulting content produced are utterly realistic, which can be highly dangerous and may have the potential of altering the truth and eroding trust by giving false reality. Deepfakes can have negative or positive implications on society. It can be used in different domains like advertising, creative arts, film production, video games, etc., to name a few. But it could also pose huge security threats like influence the public opinion during elections, perpetrate fraud or blackmail someone. Current solutions can be used to recognize only specific manipulation techniques like splicing, colouring, etc., and results provided have poor accuracy. Hence, there is a need for automated tools competent of detecting fake multimedia content.

Keywords:-*CNN, deepfakes, digital image forensics, image forgery detection, deep learning.*

INTRODUCTION

The recent advances in AI, deep learning and image processing have advanced the creation of manipulated synthetic media called deepfakes. Usually, photoshop or GIMP would be used to forge photographs in order to change their semantics, contents, potentially everything in an image. However, throughout recent years, investigation in this kind of manipulation has made it possible to develop commercial tools that can detect and describe the manipulation.

In the last few years, fake images have become a major problem after the emergence of deepfakes, i.e., images that have been manipulated using easy-to-use and powerful deep learning tool, especially with the development of Generative Adversarial Networks (GAN), has facilitated to the creation of highly refined techniques that are capable to attack digital data, change it or create its own content from scratch. These software can output lifelike results to create deepfake images.

The most primitive form of deepfake technology switches the faces between two pictures. The popular targets of deepfake attacks have been celebrities, politicians, journalists and many others. It can be used for malicious purposes like creating inappropriate images to blackmail people, or manipulating the public opinion during elections through fake-news campaigns. In the long run, this may reduce the trust on serious and reliable journalism which is very harmful to society. On the contrary, it may find use in computer games, augmented simulations and furthermore may before long be incorporated in film

creation.
Deepfakes are difficult to detect for humans, but recent works have revealed that it can be easily detected using Convolutional Neural Networks (CNN) that are specifically trained for the function. Nevertheless, CNN systems introduced so far are insufficient in terms of robustness, generalization capacity and comprehension.

The difficulty with current systems to detect fake images is that they only work on specific manipulation techniques like splicing, coloring, etc., and also the accuracy of the results that these systems provide is poor.

LITERATURE SURVEY

[1] The existing system has been developed for distinguishing fake faces from genuine ones. It accomplishes this by the use of colour texture analysis. The texture data from the brightness channel and the difference in brightness of images is used for drawing out complementary low-level feature descriptions from various colour spaces in color texture analysis. The feature histograms are computed over each image band individually. This system looked into different colour spaces and descriptors that can be utilized for marking out the intrinsic discrepancy in the colour texture information of authentic and non-authentic faces. A fusion study was also done to observe the complementarity of the various descriptors and various colour spaces. This system has been tested thoroughly on three difficult databases: Firstly, MSU Mobile Face Spoof Database, secondly CASIA Face Anti-Spoofing Database and finally the Replay-Attack Database. It was found that the performance of this system was excellent compared to other modern methods. Unlike other methods, this method was able to achieve a stable and consistent performance across all 3 benchmarks. A cross-database evaluation was done to verify the performance results, which also indicated that this method had superior performance. In conclusion, colour texture representation is more consistent at detection in unknown conditions than its gray-scale variant that does not exploit color quality features.

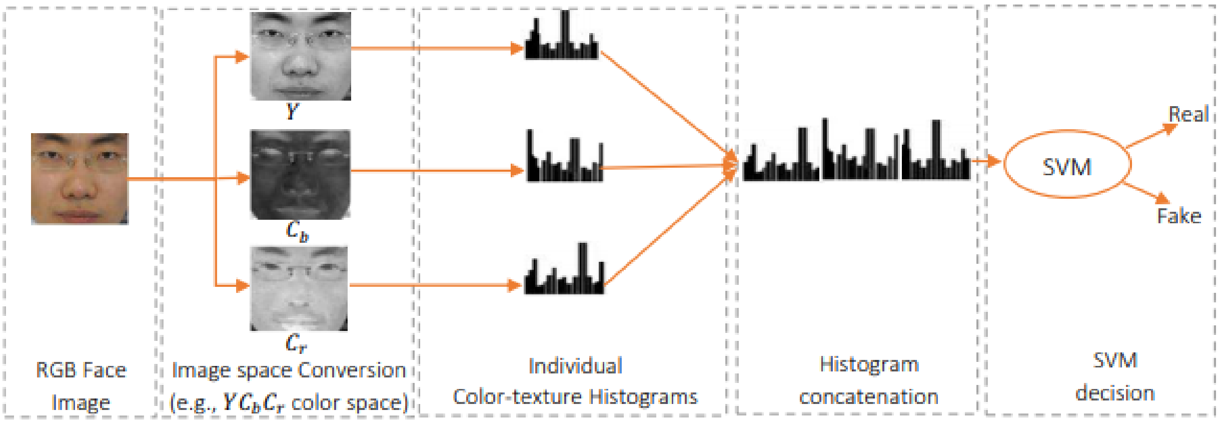

RGB Face Image
Y
C_b
C_r
Image space Conversion (e.g., YC_bC_r color space)
Individual Color-texture Histograms
Histogram concatenation
SVM
Real
Fake
SVM decision

Fig.1:-*Face anti-spoofing approach. [1]*

[2] Microblogs are popular platforms where news is reported. Nevertheless, there is an abundance of fake news propagating on these platforms which in turn reduces its credibility. For the purpose of automated news integrity authentication, this current framework makes use of digital images. In order to identify false news, patterns are discerned on the basis of numerous statistical and visual properties. Specifically, five visual properties are used: visual clustering,

clarity, diversity, similarity distribution histogram, and coherence score. These properties define distributed picture characteristics from various visual elements and therefore expose concealed distributed pattern of pictures that are found in the news. The pictures in legitimate news are more varied than the pictures in false news. This variance in picture distribution pattern of false news pictures and authentic news pictures, as depicted by Figure 2, is exploited for discerning false images from authentic images. The efficiency of the proposed image features in this system was validated by testing on a thorough multimedia dataset. It was found that the efficiency of this technique is superior than other standard techniques.

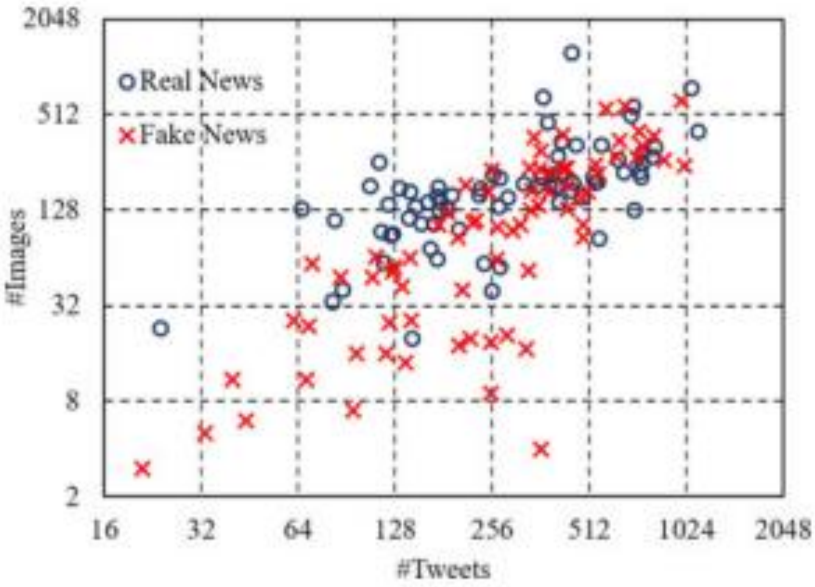


Fig.2:*-Plot of the number of tweets and the number of pictures in false and authentic news. [2]*

[3] To differentiate between manipulated and non-manipulated areas, the current methodology presented a manipulation localization architecture that makes use of resampling characteristics, such as encoder-decoder network and long-short term memory (LSTM) cells. The proposed system uses a hybrid CNN-LSTM model to successfully categorise non-manipulated and manipulated objects. To identify characteristics like shearing, upsampling, downsampling, rotation and loss of image quality, resampling properties are employed. An encoder network that outputs spatial feature maps of forged objects is designed using the CNN architecture. In LSTM network, the resampling properties of the patches are included to look at the contrast between manipulated and unmanipulated patches. A decoder network is utilized to comprehend the mapping from encoded feature maps to binary mask. A broad image splicing databank is also incorporated in the proposed method to guide the training process. Even at the pixel level, the machine is effective at restricting image manipulations. Thus, the proposed system concludes that its architecture outperforms other currently used advanced modus operandi by considerable amount on a given dataset. Along with that, it is capable of dividing efficiently a wide range of image manipulations that include splicing, copy-move and object deletion.

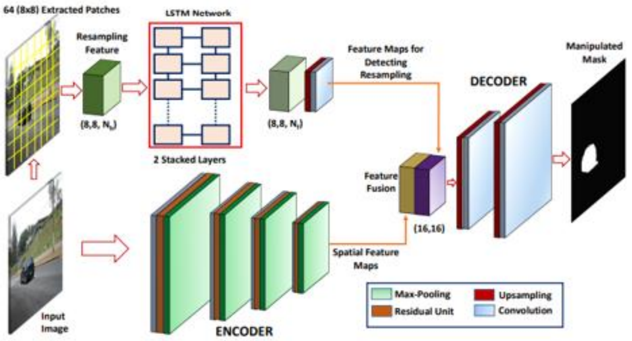


Fig.3:*-Skeleton for manipulated image areas localization. [3]*

[4] This paper's principal goal is to resolve a new issue in the area of counterfeit image detection: fake colourized image detection. Computer scientists are currently focusing their efforts on splicing identification, copy-move identification, and image retouching identification. Colorization is a new image processing procedure in which grayscale photographs are colourized with natural colours. This approach can also be used to deliberately confuse object recognition algorithms by applying it to specific images. The statistical variations between authentic images and their corresponding artificial colourized images in the properties of saturation, shade, dim, and bright channels are observed. This paper proposes two basic but efficient identification techniques for false colourized images based on possible indications in the hue, saturation, dim, and bright networks. They are: Histogram based Fake Colorized Image Detection (FCID-HIST) and Feature Encoding based Fake Colorized Image Detection (FCID-FE). FCIDHIST uses the image's existing statistical variations to spot bogus colourized images. The hue factor, the contrast feature, the dark channel, and the light channel are all used to catch counterfeits in FCID-HIST. To distinguish the false colourized images from the real images, the distinguishing features should, on the surface, show the most significant differences between the two types of images. However, since the distributions are reconstructed channel by channel in FCID-HIST, the properties do not completely use the statistical variations between the real and false colourized images. As a result, a new scheme, known as FCID-FE, is used to further manipulate statistical data by combining the modelling of the data distribution and leveraging the divergences within various intervals of the distribution. The findings show that FCID-HIST and FCID-FE work impressively when compared to various colorization methods, with FCID-FE providing more reliable and better results in most experiments than FCIDHIST.

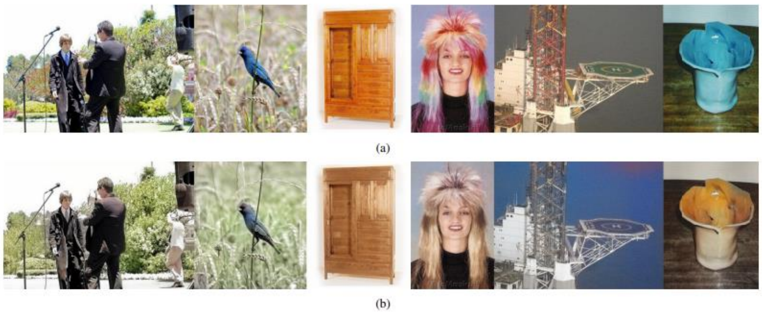
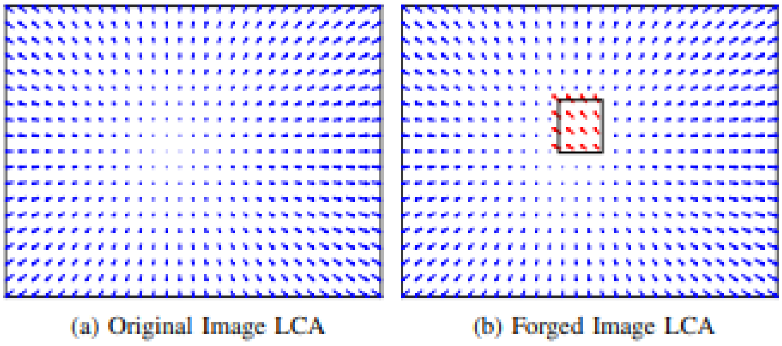


Fig.4:*-(a) Authentic images. (b) False colorized images. [4]*

[5] Centered on detecting localised LCA anomalies, the current framework proposes a technique for locating manipulated regions in a picture. This form of disparity happens when the material of one image is replicated and pasted into another, resulting in the emergence of lateral chromatic aberration (LCA), an imaging property. The system proposes an effective algorithm for correctly estimating local LCA irregularities by decreasing the amount of similarity measurements needed. Diamond search, a block matching algorithm, is used to accomplish this. As a result of using this algorithm, which was initially conceived for effective MPEG encoding, greatly redcued the computational expense and time associated with using lateral chromatic aberration as a property for identification of counterfeit recognition. The system was put through a number of assessments which tested the performance of the proposed technique and compared it against past benchmarks. As a result, the proposed approach in a simplified fabrication context improved accuracy by 51% when compared to prior studies at a false alarm estimate of 1%. The proposed approach increased the identification rate of forged photographs with solid LCA characteristics by 68% compared to past studies with a false alarm estimate of 1%.


(a) Original Image LCA

(b) Forged Image LCA

Fig.5:-*Comparison of Lateral chromatic aberration (LCA) dislocation areas in (a) a genuine image and (b) a bogus image. [5]*

[6] The core focus of this paper is to fetch the graph of relationships betwixt possibly related images using image provenance analysis. The task is to extract the collection of original images whose content matches the query image from a huge set of images and a query image. Image provenance analysis involves two steps:

- Provenance image filtering: This aids in the identification for photographs that are directly linked to the query picture in a possibly vast pool of pictures. The query will often be a picture that has been modified to a degree.
- Provenance graph construction: In this process, we learn about the connections between photographs as a result of provenance image filtering.

To achieve the best results, this paper provides an image indexing scheme that employs distributed interest point sorting iterative and filtering. It also proposes strategies for constructing provenance graphs that expand on traditional approaches. For graph enhancement, a novel clustering algorithm is employed. To produce detailed output outcomes, these approaches are analysed using the NIST nimble challenge and the multiple-parent phylogeny dataset.

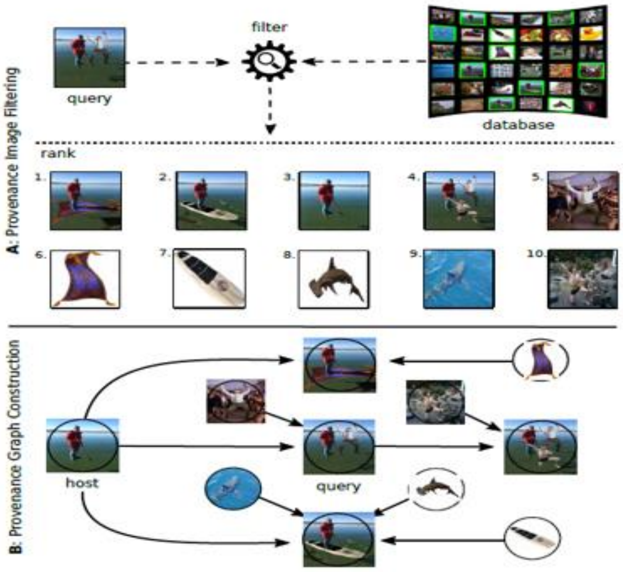

A: Provenance Image Filtering
query
filter
database
rank
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
B: Provenance Graph Construction
host
query

Fig.6:-*Image provenance analysis workflow [6]*

[7] Generative Adversarial Networks (GAN) can be used to obtain very realistic images. To combat these, improved image forgery detection techniques is in high need. An algorithm which can identify the structural defects in GAN generated images was designed, which makes use of the up-sampling procedure that is done by the Transposed Convolution process. Instead of relying on local data, the self-attention mechanism was adopted to broaden the methodology and understand more about the global context. The global information in the image remains to be incomplete because of the Transposed Convolution. The algorithm used here is designed with better comprehension of global information which can be considered as a drawback of the current existing methods. The proposed algorithm has been seen to be superior to current approaches after testing it against various types of different images and under different circumstances. This technique outperforms the others because of its ability to capture the special texture pattern which is a result of the overlapping checkerboard artifacts.

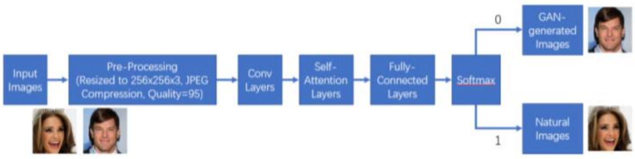


Fig.7:-*GAN-Generated Image Detection Framework [7]*

Table 1:-*Accuracy of existing methodologies*

Method	**Accuracy**
[1] Colour Texture Analysis	2.8% (HTER)
[2] Novel Visual and Statistical Image Features	83.6%
[3] Hybrid LSTM and Encoder-Decoder Architecture	94.80%
[4] Fake Colorized Image Detection	20.20% (HTER)
[5] Lateral Chromatic Aberration	79%
[6] Image Provenance Analysis	90.7%

CONCLUSION

This paper performs an extensive analysis of the various detection methods that have been used to detect forged images. Falsified multimedia has become a serious concern in recent years, particularly after the arrival of deepfakes, which are hyper-realistic images that apply artificial intelligence (AI) to depict someone. Through the various papers, it is found that there are different accuracies and speeds based on the datasets utilized by the detectors. The main obstacle with the current existing system is that they are effective against only specific tampering methods like colouring, splicing, etc. The performance of these systems reduces when deepfakes are brought into the equation. To overcome this, there is a need for a deep learning model that will be able to distinguish between real and superior fabricated images. It should be able to improve the efficiency of multimedia forensics in detecting high quality fakes. It should also be able to ensure information integrity with high accuracy and also be capable of detecting almost all kinds of tampering on images.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to our principal Dr. B.S Bhagyashekar, our HOD Dr. Loganathan R, Dept. of Computer Science and Engineering, and our guide Prof. Deepak N R, Dept. of Computer Science and Engineering, for their constant support and guidance in giving us valuable inputs and timely feedback.

REFERENCES

1. Boulkenafet, Z., Komulainen, J., & Hadid, A. (2016). Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1818-1830.
2. Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, *19*(3), 598-608.
3. Bappy, J. H., Simons, C., Nataraj, L., Manjunath, B. S., & Roy-Chowdhury, A. K. (2019). Hybrid LSTM and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, *28*(7), 3286-3300.
4. Guo, Y., Cao, X., Zhang, W., & Wang, R. (2018). Fake colorized image detection. *IEEE Transactions on Information Forensics and Security*, *13*(8), 1932-1944.
5. Mayer, O., & Stamm, M. C. (2018). Accurate and efficient image forgery detection using lateral chromatic aberration. *IEEE Transactions on information forensics and security*, *13*(7), 1762-1777.
6. Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., ... & Scheirer, W. J. (2018). Image provenance analysis at scale. *IEEE Transactions on Image Processing*, *27*(12), 6109-6123.
7. Mi, Z., Jiang, X., Sun, T., & Xu, K. (2020). GAN-Generated Image Detection With Self-Attention Mechanism Against GAN Generator Defect. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 969-981.
8. Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1038-1048.
9. Song, C., Zeng, P., Wang, Z., Li, T., Qiao, L., & Shen, L. (2019). Image forgery detection based on motion blur estimated using convolutional neural network. *IEEE Sensors Journal*, *19*(23), 11601-11611.

10. Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910-932.
11. Mayer, O., & Stamm, M. C. (2020). Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1049-1064.
12. He, P., Li, H., & Wang, H. (2019, September). Detection of fake images via the ensemble of deep representations from multi color spaces. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 2299-2303). IEEE.
13. Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019, November). Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 518-527). IEEE.
14. Jayan, T. J., & Aneesh, R. P. (2018, July). Image Quality Measures Based Face Spoofing Detection Algorithm for Online Social Media. In *2018 International CET Conference on Control, Communication, and Computing (IC4)* (pp. 245-249). IEEE.
15. Kanwal, N., Girdhar, A., Kaur, L., & Bhullar, J. S. (2019, April). Detection of digital image forgery using fast fourier transform and local features. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 262-267). IEEE.
16. William, Y., Safwat, S., & Salem, M. A. M. (2019, September). Robust image forgery detection using point feature analysis. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 373-380). IEEE.
17. Kaur, T., Girdhar, A., & Gupta, G. (2018, December). A Robust Algorithm for the Detection of Cloning Forgery. In *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-6). IEEE.
18. Suryawanshi, P., Padiya, P., & Mane, V. (2019, March). Detection of Contrast Enhancement Forgery in Previously and Post Compressed JPEG Images. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* (pp. 1-4). IEEE.
19. Kashyap, A., Parmar, R. S., Suresh, B., Agarwal, M., & Gupta, H. (2016, December). Detection of digital image forgery using wavelet decomposition and outline analysis. In *2016 International Conference on Signal Processing and Communication (ICSC)* (pp. 187-190). IEEE.
20. Chaitra, B., & Reddy, P. B. (2019, December). A Study on Digital Image Forgery Techniques and its Detection. In *2019 International Conference on contemporary Computing and Informatics (IC3I)* (pp. 127-130). IEEE.