Detecting Deepfakes with Deep Learning and Gabor Filters

Wildan J. Hadi[1], Suhad M. Kadhem[2], Ayad R. Abbas[2]

[1]Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq.

[2]Department of Computer Science, University of Technology, Baghdad, Iraq.

Abstract–**The proliferation of many editing programs based on artificial intelligence techniques has contributed to the emergence of deepfake technology. Deepfakes are committed to fabricating and falsifying facts by making a person do actions or say words that he never did or said. So that developing an algorithm for deepfakes detection is very important to discriminate real from fake media. Convolutional neural networks (CNNs) are among the most complex classifiers, but choosing the nature of the data fed to these networks is extremely important. For this reason, we capture fine texture details of input data frames using 16 Gabor filters in different directions and then feed them to a binary CNN classifier instead of using the red-green-blue color information. The purpose of this paper is to give the reader a deeper view of (1) enhancing the efficiency of distinguishing fake facial images from real facial images by developing a novel model based on deep learning and Gabor filters and (2) how deep learning (CNN) if combined with forensic tools (Gabor filters) contributed to the detection of deepfakes. Our experiment shows that the training accuracy reaches about 98.06% and 97.50% validation. Likened to the state-of-the-art methods, the proposed model has higher efficiency.**

***Index Terms*— Deepfake detection; deep learning; Gabor filter; VGG16.**

I. Introduction

With the emergence of deep learning algorithms in artificial intelligence and its breakthrough in many areas such as Ghafoor, et al., 2021, Al-Talabani, 2020, Abdul, 2019, as well as computers' development in simulating reality, it has become possible to rely on them to create fake characters with scenes and recognize them as realistic. These represent primary reasons that have effectively contributed to the emergence of the so-called "deepfake" (deep learning + fake) technology. To make the fake video, we need the target video to use as a basis for the deepfake and then a set of videos of the person, we want to include in the fake video (Johnson, 2021). An example of deepfakes is shown in Fig. 1; this snapshot from a fake video starring Amy Adams in the film Man of Steel (Moran, 2021). Deepfakes rely mainly on deep learning algorithms, which are a form of artificial intelligence. One of these forms is called autoencoders (Zhu, et al., 2018, Zhou and Shi, 2017). Another form of artificial intelligence that has contributed to the emergence of deepfakes is generative adversarial networks (Mao and Li, 2021, Wang, et al., 2018, Karras, Laine and Aila, 2019, Yu, et al., 2017). These networks study large amounts of data to build new examples that mimic the original character incredibly.

Today, social media sites have become a good way to spread and promote fake clips, as many political figures, heads of state, and celebrity actors have been targeted (Lim, 2020). The widespread of deepfakes in social media can also provide new ways for deepfakes to target non-celebrities. The possibility of such things happening causes the loss of credibility for all published videos and their failure to use them as evidence or conviction in the courts. In general, the deepfakes technique, when directed at celebrities, some methods save them from such fabricated videos. Still, the real danger is when such technology targets the ordinary person who does not have an instrument to defend himself (Moran, 2021). Therefore, it has become necessary to build robust models in detecting deepfakes and proving the reliability of the spread data.

Inspired by the promising results achieved using deep learning in detecting deepfakes, we built a novel model for detecting deepfakes based on deep learning. Convolution neural network (CNN) is chosen to see the ability of this network to extract features from the input texture data instead of red-green-blue (RGB) image contents and to see how effective this system is in detecting deepfakes. Furthermore, the Gabor filter will be used to extract the texture properties of the input data before feeding input data to binary CNN classifier.

This paper processes the problem of deepfakes and is organized as follows: Section II presents related works. In Section III, methodology of the proposed model is described. The results of the proposed model with comparison to previous works are given in Section IV. Finally, Section V gives the conclusion of this work.

ARO-The Scientific Journal of Koya University
Vol. X, No. 1 (2022), Article ID: ARO.10917, 5 pages
DOI: https://doi.org/10.14500/aro.10917
Received: 13 December 2021; Accepted: 05 March 2022
Regular research paper: Published: 18 March 2022
Corresponding author's email: wildanjh_comp@csw.uobaghdad.edu.iq

II. Related Work

This section describes previous work that was devoted to detecting manipulation, whether in images or videos. First, the methods based on multimedia forensic techniques will be discussed, and then, the methods based only on deep learning will be described:

A. Multimedia Forensic-based Approaches

Several approaches are used to detect manipulation in the hidden structure of the media content by exploiting compression parameters, frequency domain parameters, noise maps, etc. These methods were previously relied on before deep learning algorithms appeared and depended on in the detection of forgery. For example, the author of Puglisi, et al., 2013, utilizes how implementing successive quantization followed by dequantizations introduces some arranges in the sequence of zeros and non-zeros in the coefficient distributions of the histogram. This method allows for recapturing the coefficients of the first compression in a double JPEC compression. Another form of exploiting multimedia forensic tools in detection manipulation is utilizing the statistical distribution of discrete cosine transform coefficients. In Battiato and Messina, 2009, this approach is used for forgery detection. Lens distortion is also exploited as a forensic tool. Each camera is fitted with a composite optical system; it cannot focus light at different wavelengths. These distortion signals are used to detect image manipulation (Fu and Cao, 2012). Noise residuals map also exploited for forgery detection. In Mullan, et al., 2017, a statistical comparison between two video sequences is implemented. In each sequence, noise residues within and between frames are calculated, and a statistical model is built for the first sequence to compare it with the second sequence.

B. Deep Learning-based Approaches

After the emergence of deep learning algorithms, these algorithms entered into many areas, including the detection of manipulation and forgery. Many architectures based on deep learning have been used in detecting deepfakes as we have seen later. For example, two CNN architectures with a few layers and parameters are proposed in Afchar, et al., 2018, to detect deepfakes. The first network (Meso-4) consists of four layers of convolution and pooling fooled by one dense layer. A 5 × 5 kernel size is used in the convolution layer of the first network. The second network (MesoInception-4) is based on the inception module in its architecture. Rather than using 5 × 5 convolutions of the first network, the authors used 3 × 3 dilated convolutions combined with the inception module to avoid high semantic. Both networks are designed to utilize mesoscopic features. Montserrat, et al., 2020, combined two deep learning models, the CNN model and recurrent neural network (RNN), to detect facial manipulation in the video. The basic idea is to use the CNN network for extracting features from input frames and then feed them to RNN for temporal feature extraction. Another form of using deep learning is in Güera and Delp, 2018, first, CNN is used to learn frame-level features followed by RNN to classify if the input video is fake. The CNN model used is the inception V3 with removed fully connected layer to input the final feature vectors after final max-pooling layer to RNN network. A new attention mechanism has been proposed by Dang, et al., 2020, to enhance and process the feature maps of the classifier model. This attention-based layer concentrates the network direction to manipulated and discriminative regions only. The promising results are achieved using capsule networks paid researches to use this network in deepfakes detection. For example, Nguyen, Yamagishi and Echizen, 2019, used capsule network architectures for deepfakes detection and are obtained good results. Texture features give good results in image forgery detection. Based on this motivation, pixel gradient information combined with pixel intensity information to produce texture information is used in a Multi-scale Texture Difference model named as MTD-Net for robust face forgery detection (Deepfakes) (Yang, et al., 2021).

Fig. 1. Example of Deepfakes (Source YouTube).

III. Methodology

The dataset that was relied on in this research is Deepfake Detection dataset from Google-and-Jigsaw (Dufour, et al., 2019). It is a large dataset consisting of about 3000 manipulated videos created using 28 actors with different actions and positions. Fig. 2 shows a few examples of this dataset.

The first step is to extract frames from fake and real videos loaded from the dataset and store them in separate folders. Then, the front face detector in dlib (an open-source library) is used to extract the facial area (Region of Interest). All cropped face images with categories real and fake normalized to size 224 × 224. The next step is to extract texture maps from real and fake face images using Gabor filters, then feed them to VGG 16 to classify real image from fake ones. The steps of our method are described on the flowchart, as shown in Fig. 3.

A. Gabor Filters

In the field of image processing and computer vision, the Gabor filter has been widely used in texture analysis. When

http://dx.doi.org/10.14500/aro.10917

the Gabor filter convolves with an image, it gives a high response to areas where texture changes. Gabor filter depends on a certain number of parameters which are described in Equation (1).

$$g\left(x,y,\lambda,\theta,\psi,\sigma,\Upsilon\right)=\exp\left(-\frac{x^{\sim 2}+\gamma^{2}y^{\sim 2}}{2\sigma^{2}}\right)\exp\left(i\left(2\pi\frac{x^{\sim}}{\lambda}+\psi\right)\right) \tag{1}$$

with:

$$x^{\sim}=x\cos\theta+y\sin\theta$$

$$y^{\sim}=-x\sin\theta+y\cos\theta$$

Where, λ (lambda) represents a wavelength of the sinusoidal factor, θ (theta) is the tendency of the normal to the parallel stripes of the Gabor function, ψ (psi) refers to phase offset, σ (sigma) is the standard deviation of the Gaussian function used in the Gabor filter, and finally, Υ (gamma) refers to spatial aspect ratio (Dora, et al., 2017). The Gabor bank consists of 16 filters with different orientations; the values of the other parameters are kept unaltered because of the absence of scale variations in the cropped face image. The Gabor bank provides us with texture information that needs for the next step. Fig. 4 shows the workflow of processing cropped face images using Gabor bank filters.

Fig. 2. Examples of videos from deepfakes dataset (Dufour, et al., 2019).

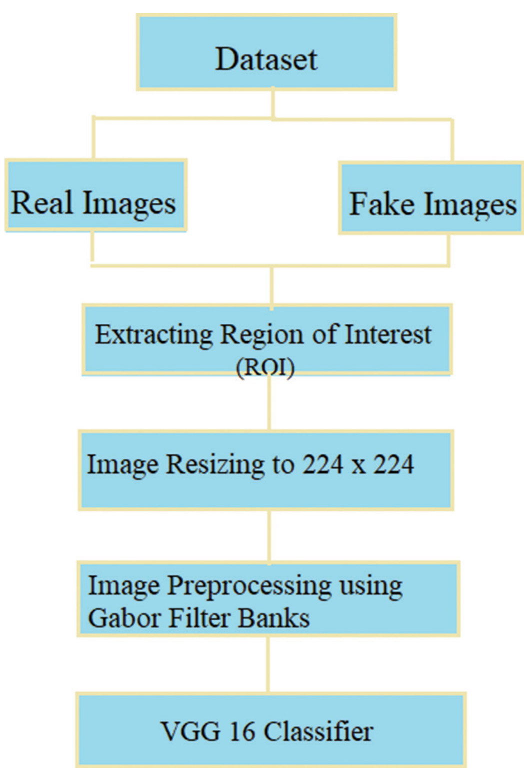

Fig. 3. The steps of proposed method.

B. CNN (VGG16 Classifier)

VGG 16 is a specified CNN intended for classification and localization. The VGG network model was first suggested by Simonyan and Zisserman, 2014. The architecture of the VGG16 network is shown in Fig. 5, which includes five blocks of convolution layer and max-pooling layer followed by three fully connected layers. Convolution layers depend on a 3 × 3 kernel with the value of 1 for both padding and stride to ensure that the resulting map has the same length of dimensions as the activation map in the previous layer. Also to guarantee the spatial dimension of the activation map from the previous layer is halved, the max-pooling layer uses a 2 × 2 kernel size, a stride of 2, and no padding. This differentiates the CNN when it is compared with the size of backpropagation networks of the same number of layers. At the end of each block of convolution and max pooling, a rectified linear unit activation is used to reduce the spatial dimension. Finally, three fully connected layers are used in VGG16 architecture for final classification.

The fully connected layers predict the real and fake classes based on the input layer. The softmax function is used to squash the outputs of each neuron between 0 and 1 by applying the following equation:

$$S(y)_i=\frac{\exp(y_i)}{\sum_{j=1}^{n}\exp(y_i)} \tag{2}$$

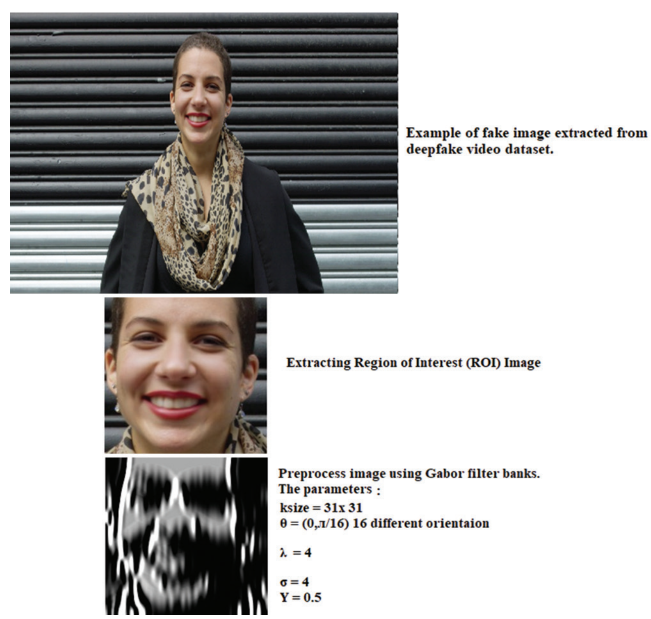

Fig. 4. The workflow of applying Gabor bank and its parameter description.

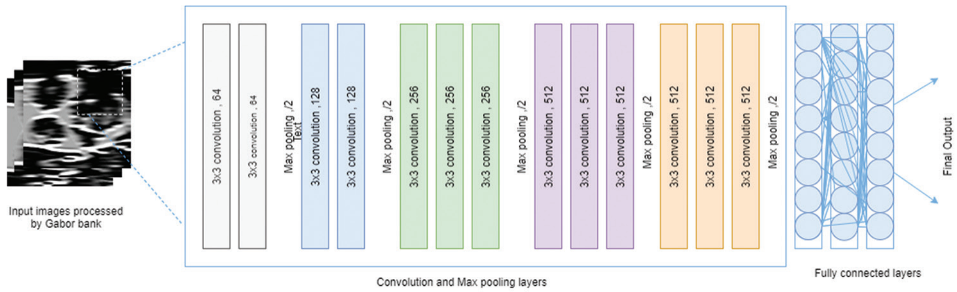

Fig. 5. The architecture of VGG 16 network.

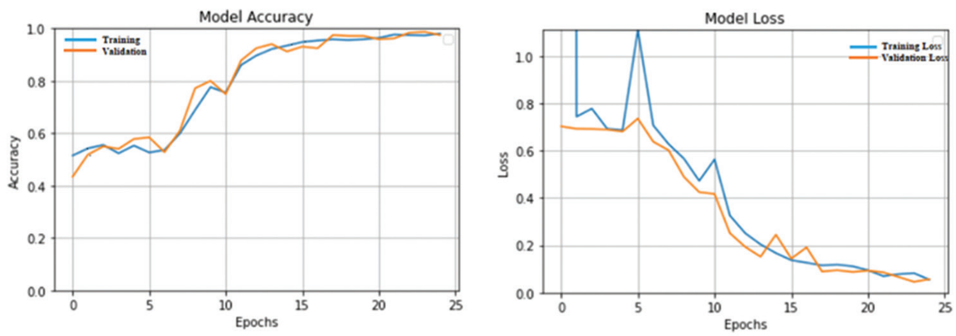

Fig. 6. The accuracy and loss of the proposed model.

Where, y_i is the i-th element of input vector (Sudiatmika and Rahman, 2019).

IV. Results and Discussion

In this section, the experimental results of the proposed deepfake detection algorithm will be described. The images extracted from the deepfake video dataset are divided using 80–20% for training and testing. As shown in Fig. 6, the accuracy of training ranges between 50 and 90, which means that the proposed method can learn this type of data.

From the results above, we find that the training accuracy of the model has reached 98% and for validation 97% employing 25 epochs. As a result, the use of the VGG16 model in learning texture features from the input data processed by the Gabor filter contributed to detecting deepfakes.

The execution of the proposed model compared with alternate methods such as MesoNet (Afchar, et al., 2018) and automatic face weighting (Montserrat, et al., 2020). The performance of these methods reached 84.3 and 91.88, respectively. The advantages of our system involve minimizing the required numbers of epochs for training also the high accuracy values for both training and testing. These advantages are satisfied due to using texture maps instead of RGB maps as input to the CNN classifier. However, the GPU environment is necessary for other methods, but our model can be implemented in the CPU environment. Furthermore, the principles on which other methods relied are not recognized, but our proposed model is based on texture maps in detecting deepfakes.

V. Conclusion

In this study, a novel method for deepfakes detection is proposed. It is shown that using Gabor bank filters to extract texture features in 16 different orientations and feeding them to the VGG model provide good results. This means that combining techniques from forensic tools such as Gabor filters and deep learning methods such as the VGG model affect deepfake detection accuracy. Therefore, the new oriented takes advantage of the methods used in both areas (multimedia forensics + deep learning) and uses them in deepfakes detection.

For the future work, some considerations can be taken into account such as creating a hybrid dataset by combining two or more deepfakes datasets to enable the proposed CNN model to be trained on it and then increase its ability to recognize unseen data. Furthermore, we can use Gabor transform (which is a 1-D transform used to processes 1d signals) for detecting deepfakes audio due to the promising results, we have achieved from using Gabor filters in this work.

References

Abdul, Z.K., 2019. Kurdish speaker identification based on one dimensional convolutional neural network. *Computational Methods for Differential Equations*, 7, pp.566-572.

Afchar, D., Nozick, V., Yamagishi, J. and Echizen, I., 2018 Mesonet: A compact facial video forgery detection network. In: *2018 IEEE International Workshop on Information Forensics and Security* (*WIFS*). IEEE, pp.1-7.

Al-Talabani, A.K., 2020. Automatic recognition of arabic poetry meter from speech signal using long short-term memory and support vector machine. *ARO-The Scientific Journal of Koya University*, 8(1), pp.50-54.

Battiato, S. and Messina, G., 2009. Digital forgery estimation into DCT domain:

A critical analysis. In: *Proceedings of the First ACM Workshop on Multimedia in Forensics*, pp.37-42.

Dang, H., Liu, F., Stehouwer, J., Liu, X. and Jain, A.K., 2020. On the Detection of Digital Face Manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pp.5781-5790.

Dora, L., Agrawal, S., Panda, R. and Abraham, A., 2017. An evolutionary single Gabor kernel based filter approach to face recognition. *Engineering Applications of Artificial Intelligence*, 62, pp.286-301.

Dufour, N., (2019) 'Deepfakes Detection Dataset'.

Fu, H. and Cao, X., 2012. Forgery authentication in extreme wide-angle lens using distortion cue and fake saliency map. *IEEE Transactions on Information Forensics and Security*, 7(4), pp.1301-1314.

Ghafoor, K.J., Rawf, K.M.M., Abdulrahman,A.O., Taher, S.H., 2021. Kurdish dialect recognition using 1D CNN. *ARO-The Scientific Journal of Koya University*, 9(2), pp.10-14.

Güera, D. and Delp, E.J., 2018. Deepfake video detection using recurrent neural networks. In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance* (*AVSS*). IEEE, pp.1-6.

Johnson, D., 2021. *What is a Deepfake? Everything you need to Know*. Available from: https://www.businessinsider.com/what-is-deepfake [Last accessed on 2021 Sep 07].

Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4401-4410.

Lim, H., 2021. *The Best* (*and worst*) *Deepfakes Developments in 2020, Lionbridge AI*. Available from: https://lionbridge.ai/articles/a-look-at-deepfakes-in-2020 [Last accessed on 2020 Apr 24].

Mao, X. and Li, Q., 2021. *Generative Adversarial Networks for Image Generation*. Springer Nature. Available from: https://books.google.com/books?hl=en&lr=&id=u9oeEAAAQBAJ&oi=fnd&pg=PR7&dq=generative+adversarial+networks+%2Bface2face%2Bimage+synthesis&ots=TwOl2AJx0K&sig=emvISr6kpl5okeSBjPDMcYT6iZM

Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F. and Delp, E.J., 2020. Deepfakes detection with automatic face weighting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.668-669.

Moran, T.B., 2021, *What is a deep Fake and how are they made?* Available from: https://www.smh.com.au/technology/what-is-the-difference-between-a-fake-and-a-deepfake-20200729-p55ghi.html [Last accessed on 2021 Nov 30].

Mullan, P., Cozzolino, D., Verdoliva, L. and Riess., C, 2017. Residual-based forensic comparison of video sequences. In: *2017 IEEE International Conference on Image Processing* (*ICIP*). IEEE, pp.1507-1511.

Nguyen, H.H., Yamagishi, J. and Echizen, I., 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, pp.2307-2311.

Puglisi, G., Bruna, A.R., Galvan, F. and Battiato, S., 2013. First JPEG quantization matrix estimation based on histogram analysis. In: *2013 IEEE International Conference on Image Processing*. IEEE, pp.4502-4506.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014, p.14091556.

Sudiatmika, I.B.K. and Rahman, F., 2019. Image forgery detection using error level analysis and deep learning. *Telkomnika*, 17(2), pp.653-659.

Wang, X., Li, W., Mu, G., Huang, D. and Wang, Y., 2018. Facial expression synthesis by u-net conditional generative adversarial networks. In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp.283-290.

Yang, J., Li, A., Xiao, S., Lu, W. and Gao, X., 2021. MTD-net: Learning to detect deepfakes images by multi-scale texture difference. *IEEE Transactions on Information Forensics and Security*, 16, pp.4234-4245.

Yu, Y., Gong, Z., Zhong, P. and Shan, J., 2017. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In: *International Conference on Image and Graphics*. Springer, Berlin. pp.97-108.

Zhou, Y. and Shi, B.E., 2017. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: *2017 7th International Conference on Affective Computing and Intelligent Interaction* (*ACII*). IEEE, pp.370-376.

Zhu, H., Zhou, Q., Zhang, J. and Wang, J.Z., 2018. Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression. *arXiv*, 2018, p.180402740.