Unmasking Deepfakes by Fusing Rich Features from Two-Stream CNN Model

Wildan Jameel Hadi[1], Suhad Malallah Kadhem[2], Ayad Rodhan Abbas[2]

1. Department of Computer Science, Baghdad University, Iraq

2. Department of Computer Science, AL-Technology University, Iraq

*Corresponding author E-mail: wildanjh_comp@csw.uobaghdad.edu.iq

Doi:10.29072/basjs.20220216

ARTICLE INFO

Keywords

Deep Learning; Gabor filter; RGB; Texture; Two-stream CNN

ABSTRACT

In contrast to the traditional object detection methods, image manipulation detection focuses on tampering artifacts instead of image content, indicating that more depth features must be learned to detect image manipulation. Deepfakes are one of these techniques that have appeared in recent times and need to learn a lot of the richer features to be detected. Deepfakes are a harmful application that affects all segments of society. It is meant to change the person's face and replace it with another person using deep learning techniques. In this paper, we contribute to finding a solution to detect the fakes. A new two-stream CNN model-based deep learning is developed, where two streams are combined, exploiting the fusion layer. Following the fusion layer, the data is classified using the classification layer. The first stream is a semantic stream to extract specific features from the RGB image input to identify manipulation artifacts such as blurring variation, the boundary of the face mask, and lighting difference. The second stream is a texture stream that exploits the texture features extracted from a Gabor bank filters layer. The proposed strategy significantly outperformed the previous methods that were in use. The measured performance metrics have an accuracy of more than 99.5%.

Received 5 July 2022; Received in revised form 8 Aug 2022; Accepted 22 Aug 2022, Published 31 Aug 2022

1. Introduction

Artificial intelligence [1,2] and deep learning [3,4] techniques have contributed to the establishment of many applications. One of these applications is the so-called deepfake, which is a type of manipulation implemented on the video to replace a person's face with another person's face to create a kind of illusion that the target person is the one who made these movements [5]. This type of manipulation was created using an artificial intelligence algorithm known as generative adversarial networks (GANs) [6,7,8], which helped create faces that the human eye cannot distinguish as fake. All digital representations are being targeted by deep-fake technology to create images, videos, or even sounds that act on the ground as real. Despite exploiting this technology in cinema and games, it does not prevent damage to society's security. For example, many celebrities are exploited as targets for deepfakes by including them in indecent clips (see Fig. 1) or changing the facial gestures of heads of state and well-known personalities due to the large availability of their images on the Internet. These methods may threaten the security of society and relations between countries as a result of spreading incorrect words and actions [9]. Another example is that it is possible to use forged digital evidence in the courts. Thus innocent people will be imprisoned due to tampering with the evidence presented to the court. Alternatively, publishing malicious advertisements aimed at harming a particular product or the reputation of a particular company [10,11]. These are examples of how harmful the deepfake technique can cause tampering. Therefore, to maintain the credibility of digital evidence, it is crucial to develop methods for unmasking deepfakes.

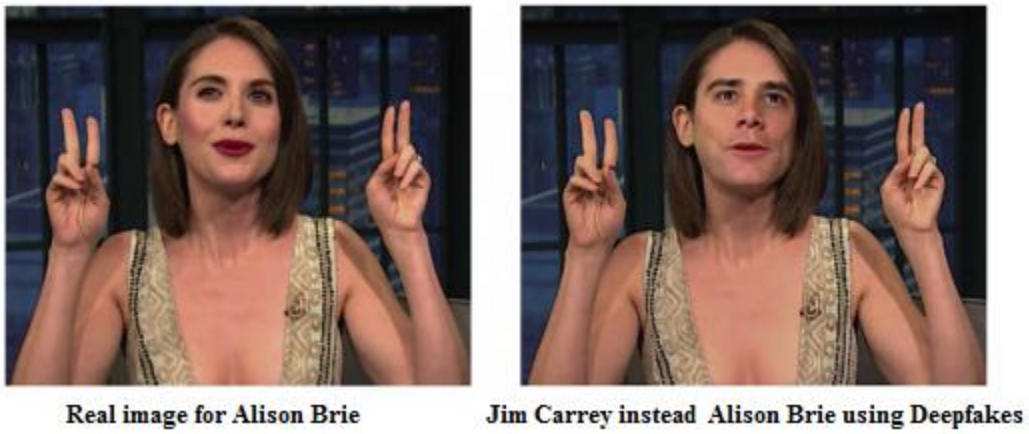


Figure 1: Deepfakes targeting celebrities, taken from [9].

Current deepfake detection methods have difficulty detecting data that has not been seen. To increase the robustness and generalization capability, we combined two deep CNN models, which extract features from two input data sources (RGB and texture) to increase the robustness and generalization capability. The first is a semantic stream that extracts particular features from the RGB image input to spot manipulation artifacts. The second stream is a texture stream that uses the texture features taken from the layer of the Gabor bank filters. This non-linear CNN model attempts to learn features not only from RGB information but also from texture information. The rest of the paper is organized as follows: related works are included in Section 2. The proposed model's technique is explained in Section 3. In Section 4, the results of the suggested model are presented. Finally, the conclusion of this work is presented in Section 5.

2. Related works

Exploiting the mesoscopic characteristics of images was one of the early attempts to identify deep-fake images. To find tampering, Meso-4 and MesoInception-4 architectures were suggested [12]. Meso-4 is made up of four layers of pooling and subsequent convolutions. A dense network with just one hidden layer follows these layers. The first two convolutional layers are swapped out with a modified version of the inception module to create MesoInception-4, based on Meso-4. MesoNets and their derivatives have demonstrated promising results in deepfake detection. The authors of [13] developed a capsule network. The proposed methodology uses three essential capsules with two outputs to distinguish between fake and real information. In the area of deepfake detection, transfer learning also provides a good clue. One example includes testing a combination of trained CNN models [14]. Focusing on specific types of manipulation evidence has had good results in detecting deepfakes. Therefore, a model has been built that integrates the two streams of deepfake detection by focusing on capturing the evidence of manipulation within the intended face, while the second is designed to capture the remaining evidence of noise residuals [15]. In general, deepfake technology works to create fake face images that have a fixed size and must undergo a set of post-processing operations to match the configurations of the source's face. The method in [16] exploits artifacts by comparing the resolution of the areas of the fake face with its surrounding areas based on the dedicated CNN model.To exploit the effectiveness of previous methods in analyzing manipulated images and videos that are based on forensic tools integrated with current deep learning models. Multimedia forensics methods detect the different types of manipulation depending on the fact that RGB channels are insufficient to handle all possible

manipulation scenarios. According to these facts, we proposed a two-stream CNN model that learns features from RGB and texture images.

3. Methodology

Usually, images resulting from applying deepfake algorithms often need more transformation to fit the area to be forged in the source video. Such transformations leave distinct defects. The new researchers are trying to use deep learning to exploit such defects to detect deepfake. CNN are among the most complex classifiers, but choosing the nature of the data fed to these networks is extremely important. So that developing an algorithm for deepfake detection is very important to discriminate real from fake media. The proposed system consists of a two-branch end-to-end CNN model, which extracts features from two input data sources (RGB and texture images). The semantic stream is supplied with RGB images, while texture images are fed to the texture stream (as shown in Figure 2). The Deepfake-Detection-dataset from Google-and-Jigsaw [17] was the dataset that was used in this work. It is a sizable dataset made up of about 3000 fake videos recorded with the help of 28 actors in various poses and motions. First, separate folders must be created to keep the frames extracted from the real and fake videos loaded from the dataset. The facial region is then extracted using the front face detector in dlib (an open-source package) to represent the region of interest (ROI). Finally, real and fake facial images that have been cropped are all normalized to 224x224 pixels.

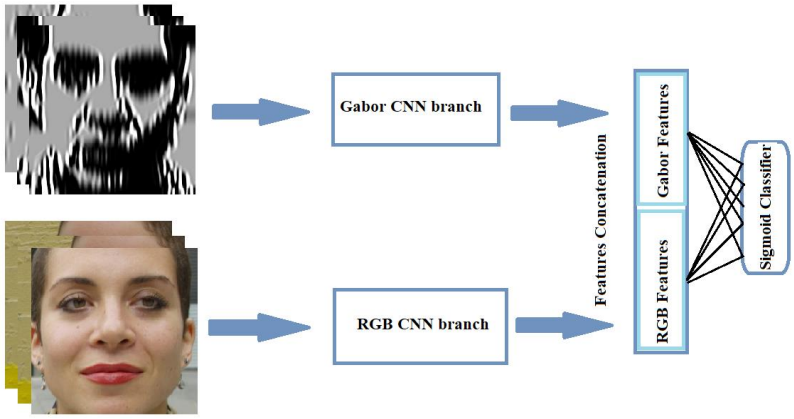


Figure 2: The new two-stream CNN network.

3.1 Semantic stream

When we talk about RGB images, we mean that every point in this image has three colors red, green, and blue. The ROI face images extracted from the used dataset are actually in the form of RGB format. The semantic stream in this model consists of one input layer, two

convolutional layers, and one max-pooling layer followed by dropout to avoid overfitting. To ensure that the resultant map has the same dimension's length as the activation map in the preceding layer, convolution layers use a 3 x 3 kernel with a value of 1 for both padding and stride. The max-pooling layer also utilizes a 2x2 kernel size, a stride of 2, and no padding to ensure that the spatial dimension of the activation map from the preceding layer is halved. The fourth convolutional layer with kernel 2x2 is utilized in the second block, followed by a max-pooling layer and dropout.

3.2 Texture stream

The RGB image information is not adequate to process all manipulation types. So, we suggest adding additional evidence that works side by side with RGB information. This addition includes utilizing the texture information to provide a new clue in detecting deepfakes with RGB information. Because of the promising results of using texture maps in deepfake detection [3], It is used here alongside semantic stream as another stream to increase generalizability. Gabor filters have been widely employed in texture analysis in image processing, so it is used to extract texture maps from ROI images to input them into the texture stream. When the Gabor filter is applied to an image, it responds strongly to regions where the texture varies. Some of the parameters that the Gabor filter depends on are given in (1):

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

With:

$$\tilde{x} = x\cos\theta + y\sin\theta, \quad (2)$$

$$\tilde{y} = -x\sin\theta + y\cos\theta, \quad (3)$$

Where λ (lambda) denotes the wavelength of the sinusoidal factor, θ (theta) the tendency of the Gabor function's normal to its parallel stripes, (psi) the phase offset, (sigma) the standard deviation of the Gabor filter's underlying Gaussian function, and (Y) the spatial aspect ratio [3]. The cropped face image lacks scale variations, thus the values of the other parameters are left unchanged. The Gabor bank consists of 16 filters with various orientations. We can get the texture data we need for the subsequent step from the Gabor bank.

3.3 Fusion features

Figure 3 shows that the fusion is implemented at the feature level. This non-linear CNN model attempt to learn feature not from RGB information but also exploited the texture information.

To learn generic characteristics, the first CNN model uses the original RGB ROI images. On the other hand, the second CNN model utilizes the information on the structure of images that are found in the texture. Each branch CNN model had one input layer, three convolutional layers, and two max pooling layers mixed with the dropout layer to avoid overfitting. We fed RGB maps to the first branch, while Gabor maps were fed to the second branch. After flattening all feature maps from two branches, the fusion is implemented at layer eight. A fully connected layer then processes the flattened fused feature maps. Finally, layer 10 was the output layer with two nodes (0 "fake" + 1 "real") and Sigmoid activation. Sigmoid function exits between (0 to 1) so that we use it to predict the probability and find output. The equation for the sigmoid function is [18]:

$$S(x) = \frac{1}{1+e^{-x}}, \qquad (4)$$

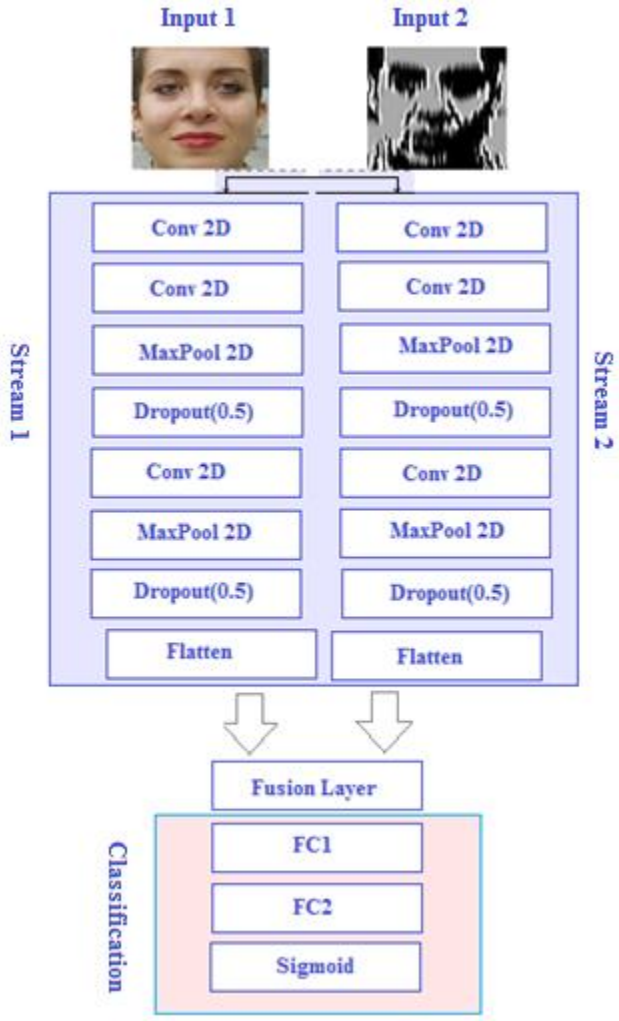


Figure 3: Flowchart of proposed two-stream CNN model.

4. Results and discussion

The experimental results of the suggested deepfake detection method will be discussed in this section. The ROI images extracted from the deepfake video dataset are separated into training and testing using an 80–20% rate. This separation is applied for both inputs of the two streams. The training epoch is 30, and the number of batches per epoch is 150. The number of epochs is controlled with early stopping using testing accuracy as a metric for monitoring. We used Adam optimizer and binary cross-entropy as loss functions, with a learning rate of 0.01. Tensor-flow and Keras are used to construct a two-stream CNN model in this work. Figure 4 shows the accuracy and loss for the third proposed model.

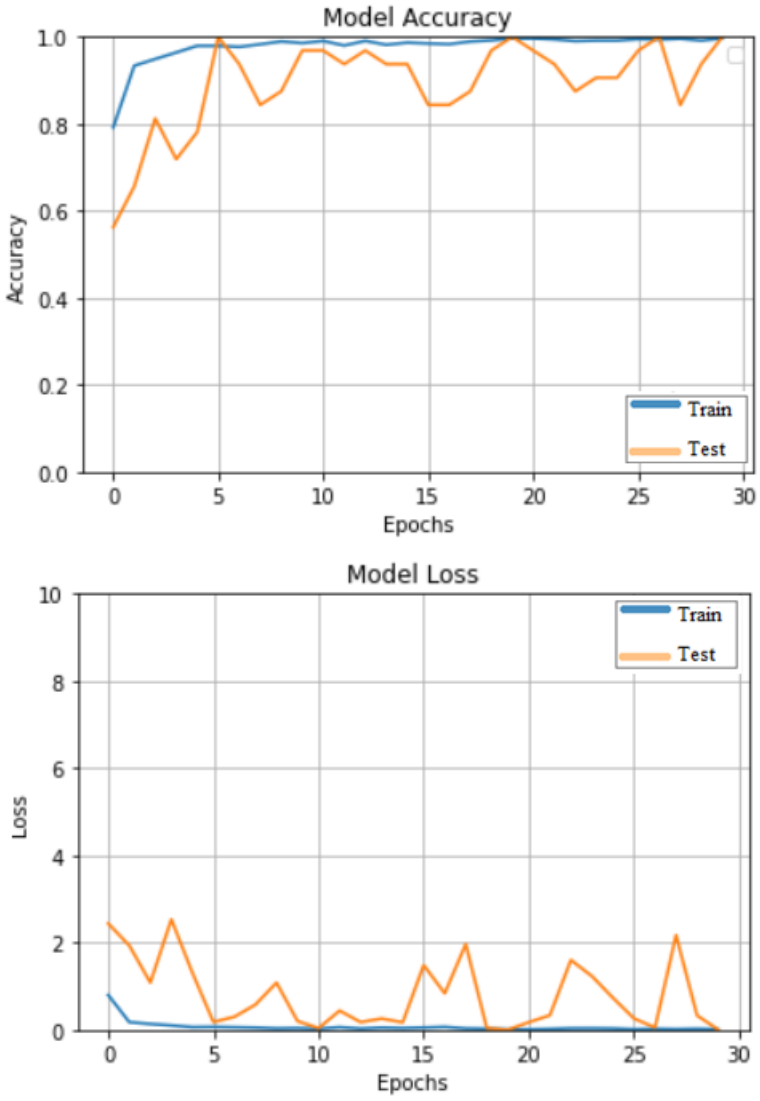


Figure 4: The accuracy and loss for the two-stream CNN model.

The proposed model is compared with related works, as shown in Table 1. All systems are executed using the Google Deepfakes Detection dataset. As shown in the intended table, the proposed system preceding the previous works in the following point:

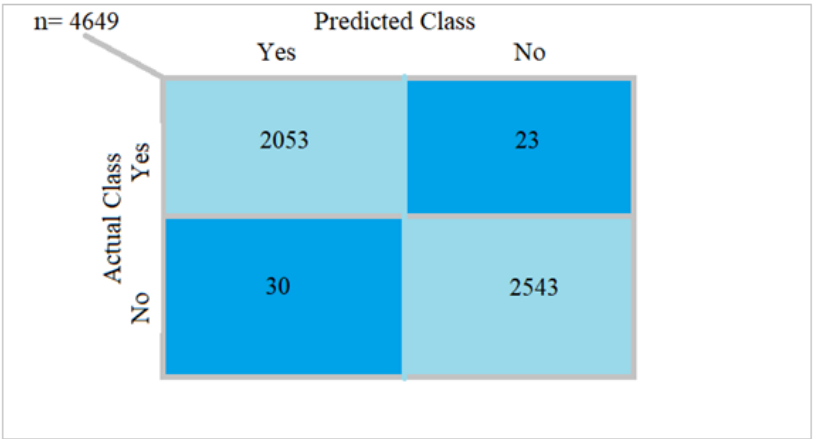
1. The obtained accuracy is enhanced by 7.5%.
2. The number of feature maps with different textures used in training is larger than in other works.
3. Learning rich features give good results in detecting deepfakes.
4. Reducing the number of epochs required for the proposed system because there are two branches for accepting data.

Table 1: Overall performance of the three proposed CNN models on a new dataset.

Model	*Accuracy*
Capsule-Forensics-Noise [13]	*92.17%*
Meso-1 [12]	*89.10%*
MesoInception-2 [12]	*91.70%*
Our Proposed model	*99.67%*

For more accuracy, about 5000 images are taken to see the result of the testing mode using a confusion matrix, as shown in Table 2.

Table 2: The confusion matrix of the proposed model.



5. Conclusion and future work

In this research, a novel method for fake face detection is suggested. A two-stream deep CNN methodology was built in the suggested method. A fusion layer was used to integrate the two suggested CNN models in the two-stream CNN architecture. RGB and textured images are the two-stream CNN model's inputs. Evaluation of the performance of the deepfake detection system with novel types of face manipulation. The experiment results indicate that the new method performed better than the traditional methods. In further work, some issues can be considered, such as the possibility of building a hybrid dataset by mixing two or more deepfakes datasets. This would allow the proposed two-stream CNN model to be trained on the hybrid dataset, which would then increase the model's ability to detect data that has not been seen before. Due to the encouraging results that we have obtained from utilizing Gabor filters in this work, we are also able to use the Gabor transform (which is a 1-D transform that processes 1-D signals) for the purpose of detecting deepfakes in audio.

References

[1] A. K. Abdul-Hassan, I. H. Hadi: A Proposed Authentication Approach Based on Voice and Fuzzy Logic, in Recent Trends in Intelligent Computing, Communication and Devices, Springer, (2020) 489–502, doi:10.1007/978-981-13-9406-5_60.

[2] M. M. Rahman, B. C. Desai, P. Bhattacharya, Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion, Comput. Med. Imaging Graph., 32(2008) 95–108, doi:10.1016/j.compmedimag.2007.10.001.

[3] W.J. Jameel, S. M. Kadhem, A. R. Abbas: Detecting Deepfakes with Deep Learning and Gabor Filters, ARO-THE Sci. J. KOYA Univ.,10(2022)18–22, doi: 10.14500/aro.10917.

[4] Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering, 19(2017)221, doi: 10.1146/annurev-bioeng-071516-044442.

[5] W. J. Hadi, S. M. Kadhem, A. R. Abbas: Fast discrimination of fake video manipulation., Int. J. Electr. Comput. Eng., 12,(2022)2582-2587, doi: 10.11591/ijece.v12i3.

[6] X. Yi, E. Walia, P. Babyn, "Generative adversarial network in medical imaging: A review," Med. Image Anal., 58(2019)101552, doi: 10.1016/j.media.2019.101552.

[7] Ma, Jiayi, Wei Yu, Pengwei Liang, Chang Li, Junjun Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Information fusion, 48 (2019)11-26, doi: 10.1016/j.inffus.2018.09.004.

[8] X. Mao and Q. Li, Generative Adversarial Networks for Image Generation, Gener. Advers. Networks Image Gener., (2021)1-7, doi: 10.1007/978-981-33-6048-8.

[9] J. Kietzmann, L. W. Lee, I. P. McCarthy, T. C. Kietzmann, Deepfakes: Trick or treat?, Bus.

Horiz., 63(2020)135–146, doi: 10.1016/j.bushor.2019.11.006.

[10] S. Yavuzkilic, A. Sengur, Z. Akhtar, K. Siddique, Spotting deepfakes and face manipulations by fusing features from multi-stream cnns models, Symmetry (Basel)., 13(2021)1352, doi: 10.3390/sym13081352.

[11] M. Koopman, A. M. Rodriguez, and Z. Geradts, Detection of deepfake video manipulation, in The 20th Irish machine vision and image processing conference (IMVIP), (2018)133–136, doi:10.1016/j.gltp.2018.04.017.

[12] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), (2018)1–7, doi: 10.48550/arXiv.1809.00888.

[13] H. H. Nguyen, J. Yamagishi, I. Echizen, 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), (2019)2307–2311, doi: 10.1109/ICASSP.2019.8682602.

[14] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, S. Tubaro, Video face manipulation detection through ensemble of cnns, in 2020 25th International Conference on Pattern Recognition (ICPR), (2021)5012–5019, doi: 10.48550/arXiv.2004.07676.

[15] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, "Two-stream neural networks for tampered face detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),(2017)1831–1839,doi: 10.48550/arXiv.1803.11276.

[16] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv Prepr. arXiv1811.00656, (2018), doi: 10.48550/arXiv.1811.00656.

[17] N. Dufour and A. Gully, "Contributing data to deepfake detection research," Google AI Blog,1(2019), doi: 10.1616/j.gltp.2019.04.017.

[18] S. Albawi, O. Bayat, S. Al-Azawi, O. N. Ucan: Social touch gesture recognition using convolutional neural network, Comput. Intell. Neurosci.,2018(2018)71-75,doi: 10.1155/2018/6973103.

كشف التزييف العميق بواسطة دمج ميزات غنية من خلال استخدام نموذج شبكة التلافيفية العصبية ثنائية المجرى

ولدان جميل هادي 1, سهاد مال الله كاظم 2, اياد روضان عباس 2

1. قسم علوم الحاسوب، جامعة بغداد
2. قسم علوم الحاسوب، الجامعة التكنولوجية

المستخلص

على عكس الطرق التقليدية المستخدمة للكشف عن الكائن، تركز طرق كشف التلاعب في الصور على العيوب التي يولدها التلاعب بدلا من محتويات الصورة. مما يشير الى الحاجة الى تعلم المزيد من الميزات العميقة للكشف عن التلاعب. يعتبر التزييف العميق أحد هذه التقنيات التي لها أثر سلبي يهدد جميع شرائح المجتمع. التزييف العميق هو تقنية من خلالها يتم تغير وجه شخص ما واستبداله بوجه شخص اخر بواسطة استخدام تقنيات التعلم العميق. في هذه الورقة البحثية ساهمنا بإيجاد حل يساهم في الكشف عن التزييف العميق. النموذج المقترح هو تصميم شبكة تلافيفيه عصبية تجمع بين تيارين نتيجة استغلال طبقة الدمج. بعد هذه الطبقة تأتي طبقة التصنيف التي تعمل على تصنيف البيانات المدخلة. التيار او المجرى الاول هو مجرى دلالي يعمل على استخراج ميزات معينة من الصور الملونة المدخلة مثل التباين، فرق في قيم الاضاءة وحدود قناع الوجه. التيار او المجرى الثاني هو مجرى النسيج الذي يعمل على استخراج ميزات من النسيج المدخل الذي تم استخراجه بواسطة مرشحات كابور. تفوقت الاستراتيجية المقترحة بشكل كبير على الأساليب السابقة التي كانت قيد الاستخدام. النظام المقترح له دقة تزيد عن 99.5 ٪.