FaceSwap based DeepFakes Detection

Article *in* The International Arab Journal of Information Technology · January 2022

DOI: 10.34028/iajit/19/6/6

4 authors:

Marriam Nawaz
University of Engineering and Technology Taxila
67 PUBLICATIONS **3,055** CITATIONS
SEE PROFILE

Momina Masood
35 PUBLICATIONS **2,262** CITATIONS
SEE PROFILE

Ali Javed
University of Engineering and Technology Taxila
186 PUBLICATIONS **4,595** CITATIONS
SEE PROFILE

Tahira Nazir
Riphah International University
61 PUBLICATIONS **2,283** CITATIONS
SEE PROFILE

FaceSwap based DeepFakes Detection

Marriam Nawaz
Department of Software Engineering, University of Engineering and Technology, Pakistan
marriam.nawaz@uettaxila.edu.pk

Momina Masood
Department of Computer Science, University of Engineering and Technology, Pakistan
momina.masood@uettaxila.edu.pk

Ali Javed
Department of Software Engineering, University of Engineering and Technology, Pakistan
ali.javed@uettaxila.edu.pk

Tahira Nazir
Faculty of Computing, Riphah International University, Islamabad, Pakistan
tahira.nazir@riphah.edu.pk

Abstract: *The progression of Machine Learning (ML) has introduced new trends in the area of image processing. Moreover, ML presents lightweight applications capable of running with minimum computational resources like Deepfakes, which generates widely manipulated multimedia data. Deepfakes introduce a serious danger to the confidentiality of humans and bring extensive religion, sect, and political anxiety. The FaceSwapp-based deepfakes are problematic to be identified by people due to their realism. Hence, the researchers are facing serious issues to detect visual manipulations. In the presented approach, we have proposed a novel technique for recognizing FaceSwap-based deepfakes. Initially, landmarks are computed from the input videos by employing Dlib-library. In the next step, the computed landmarks are used for training two classifiers namely Support Vector Machine (SVM) and Artificial Neural Network (ANN). The reported results demonstrate that SVM works well than ANN in classifying the manipulated samples due to its power to deal with over-fitted training data.*

Keywords: *Deepfakes, faceswap, ANN, SVM.*

Received February 15, 2021; accepted January 23, 2022
https://doi.org/10.34028/iajit/19/6/6

1. Introduction

With the advancements in the field of artificial intelligence, particularly the Generative Adversarial Networks (GAN) [12], the quality of synthesized images and videos has improved significantly, and differentiating between fake and real is subtle. These so-called AI-synthesized media known as "Deepfakes" are created to represent a person saying and doing whatever his creator wants [1]. At the same time, the availability of digital technologies and the internet such manipulated content can rapidly spread disinformation online through social media platforms. These fake videos could, potentially, be presented to cause extreme anxiety in the public due to their unregulated growth, the potential for fraud, and cybercrimes. Deepfakes have made these scenarios seriously threatening because of their highly realistic nature [4].

Deepfakes are categorized into three different types such as FaceSwap [11], puppet-master [15], and lip-syncing [14] where the current study is related to FaceSwap-based deepfakes detection only. Face-swap deepfakes are mostly used to create fake videos, where the facial identity of a person is substituted with the identity of another person. The recently launched ZAO [2] and REFACE [10] apps are prominent among others because of their robustness to create highly realistic visual manipulations. These apps have become popular as non-technical people can replace their faces with celebrities in renowned TV and movie series. Several online accessible applications of FaceSwap deepfakes employing deep neural networks are available like DeepFaceLab [11] and FaceSwapGAN [12] leading to an increased number of synthesized media clips. Recently, Deeptrace a cyber-security organization recently performed a survey [4] and identified 14,698 deepfakes videos on cyberspace over 7,964 manipulated videos six months back which shows an increase of 84%.

It is of crucial importance to identify the real content from AI-synthesized fake media. Multiple attempts have been made by the researchers to develop algorithms for deepfake video detection. The deepfakes detection approaches are generally divided into two types named ML and Deep Learning (DL) based techniques. Yang *et al*. [16] introduced a method to identify the visual alterations by computing the 3D head orientations from 2D face region positions. The calculated variance between the head postures was utilized as a keypoints vector for the Support Vector Machine (SVM) classification. The approach works well for deepfakes detection, however, not effective for blurred images. A framework was presented in [1] where a subject-oriented method for deepfakes identification was proposed. OpenFace2 [3] toolkit was used to capture the facial keypoints and head positions.

The calculated landmarks were used for the SVM training to categorize among the genuine and forges faces.

Now, researchers are focusing to explore the DL methods for deepfakes detection. Li *et al*. [6] presented a technique to identify the visual alterations by employing the reason that the altered content has no precise eye blinking in manipulated facial regions. Spatial and temporal features computation-based method was utilized to identify the abnormal eye pattern from the visual content to uncover the forensic changes. This method improves deepfakes recognition performance, however, it cannot detect the manipulations in videos having frequent eye blinking. In [7] the author introduced a framework to locate the manipulated samples by computing the pixel co-occurrence matrices at pixels channels of the input sample. After this, Convolutional Neural Network (CNN) was applied to detect the reliable keypoints from it. Sabir *et al*. [13], explored that while producing the forged content, manipulators often lack to impose time-based patterns in the forging procedures. Therefore, in [13] RNN framework was applied to analyze the time-based video behavior for locating the forgeries in the images. The works [7, 13] acquire enhanced identification accuracy, however, with images only.

Several works have been elaborated by the research community for the efficient detection of deepfakes, however, the techniques still need performance results improvement due to the increased realism of generated fake data. In the presented work, we have tried to overcome the existing challenges of face-swap detection by introducing an effective framework. Initially, we extract the landmarks features from the videos by employing the Dlib library. Based on which two classifiers namely SVM and ANN are trained to classify the original and manipulated face images. We evaluated the performance of our method on the deepfakes dataset [1] and the reported results show the efficacy of our approach for the classification of original and fake videos. Following are the main contributions of our work:

- A novel recognition of landmarks to recognize the biomedical facial pattern to detect the FaceSwap-based deepfakes detection.
- We evaluate the robustness of our technique over a dataset where train and test sets do not intersect to show its generalization power.
- Accurate and precise detection of FaceSwap-based deepfakes due to robust facial landmark features.
- Performance analysis of SVM and ANN in terms of entire and individual subjects.

2. Proposed Methodology

In this paper, we have presented an approach to detect and classify the Face-swap based deepfakes. The entire workflow of the proposed approach is illustrated in Algorithm (1) while the visual representation is given in Figure 1 while the detailed discussion can be found in subsequent sections.

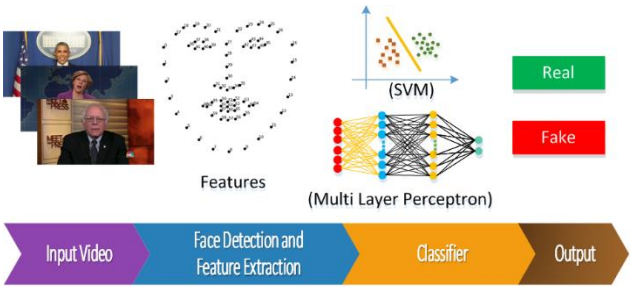

Figure 1. Flow diagram of proposed method.

2.1. Landmarks Computation

To detect landmarks that characterize facial features, we need to efficiently identify the face region in the video. We use Dlib [5] face recognition library to locate the face from each frame in the input video. Dlib calculated 68 landmark features for each detected face in the frame. The computed landmarks are employed as a feature vector in our proposed method.

2.2. Classification

To classify each input video as original or manipulated, we have utilized two classifiers namely SVM and ANN.

2.2.1. SVM

SVM is a mathematical technique that generates hyper-planes to divide data into the respective classes [9]. SVM can easily deal with the problem of high feature space as compared to other traditional methods e.g., Nave Bayes, K Nearest Neighbor (KNN) and reduces experimental error while preserving the mapping function complexity. Such behavior of the SVM classifier makes it appropriate for deepfakes detection.

We have used an SVM classifier with RBF kernel and trained it over 70% of training data. After computation of the landmark features, we utilized these keypoints for classifier training to recognize each video sample either as original or manipulated. The training samples contain N keypoints vectors elaborated as: $(x^{(i)}, y^{(i)})$, i=1,….N, where $y^{(i)} \in \{1, -1\}$ shows the original and manipulated visual classes. For all $x^{(i)}$, it draws a hyper-plane to perform binary classification as:

$$w^T.x^{(i)} + \beta \geq 1 \text{ if } y^{(i)} = +1 \quad (1)$$

$$w^T.x^{(i)} + \beta < 1 \text{ if } y^{(i)} = -1 \quad (2)$$

Here w and β are showing the weight vector and the bias. SVM is concerned to reduce the vector distance via decreasing the norm $||w||$ known as the quadratic optimization problem as mentioned in Equation (3):

$$\min ||w||, \text{ such that } y^{(i)}\left(w^T.x^{(i)} + \beta\right) \geq 1 \quad (3)$$

The output class either original or manipulated is estimated through the function $f(x)$= sign $(w^T. x^{(i)} + \beta)$ as follows:

$$\begin{cases} \text{original}, f(x^{(i)}) = +1, \\ \text{manipulated}, f(x^{(i)}) = -1 \end{cases} \quad (4)$$

Algorithm1: Steps for Face-Swap-based deepfakes detection.

START
INPUT: Video sample, Faces
OUTPUT:
Trained SVM, ANN: Trained over real and fake samples.
Classified samples: Classification using SVM and ANN.
VideoFrames← [x y]
// Face detection
α← *FaceEstimationWIthDlibLibrary (VideoFrames)*
[Tr, Ts*]* ← *partitioning of the database into* train *and* test set
// Model Training
SVM←TrainedModel(α*)*
ANN← TrainedModel*(* α*)*
// Deepfakes detection Training Unit
For each Videoframe i *in* →Tr
Compute Dlib-based keypoints →*n*m
End For
Training SVM and ANN over nm
// Deepfakes detection Testing Unit
For each sample I *in* → Ts
a) β←compute keypoints through Dlib
*b) [*output_class*] ←PredictSVM (β)*
*c) [*output_class*] ←PredictANN (β)*
d) Compute accuracy
End For
FINISH.

2.2.2. ANN

Artificial Neural Networks (ANNs) are mathematical frameworks inspired by the biological brain extensively used for signal processing, medical image analysis, robotics, and speech recognition. Multi-layer perceptron NN is the most recognized type of ANNs. MLP is a feed-forward neural network, in which the data can move in one way, from the input to the output layer. It comprises three layers:

a) Input layer.
b) Hidden layer.
c) Output layer.

The input layer takes the image features as input to the network. While the first hidden layer accepts the weighted values from the input layer and forwards data from the preceding layer to the next one. The inclusion of additional layers at the second level enables the MLP to solve complex classification problems. And the last layer which is the output layer holds the final classification output. Many techniques are employed for the learning step of MLP which is the most famous is back-propagation. So, the main architecture of MLP along with back-propagation is comprised of four stages which are as follows: weights initialization, feed-forward, error back-propagation and finally updating the value of weight. In our work, we have tested the MLP for deepfakes classification.

In MLP, the total nodes in the input layer are specified by the dimensions of the feature vector. While the number of output classes specifies the nodes in the output layer. The presented ANN framework is comprised of 68 input neurons, N1, N2, .., N68, and two output nodes, Y1, Y2, that show the original and manipulated visual classes. For original visual content, the output value is set to 0, while 1 for manipulated frames. The number of hidden layers is set to 8 in our approach.

3. Results and Discussion

In this part, a detailed analysis of the performance results attained by the introduced technique is elaborated. Moreover, the description of the employed database is discussed.

3.1. Dataset

In the proposed technique, we have used the deepfakes dataset given in [1]. The deepfakes dataset comprised of both original and manipulated audiovisual content of five subjects i.e., Barack Obama, Hillary Clinton, Bernie Sanders, Donald Trump, and Elizabeth Warren. However, we have utilized only the original and Face-swap based deepfakes manipulated videos in our approach. The video samples of all subjects are of varying lengths from 10sec and 2.5min. Furthermore, the videos are captured with 30 fps employing the mp4-format at a relatively high quality of 20.

3.2. Evaluation Metrics

We used several standard measures namely Precision (P), Recall (R), Accuracy (Acc), True Positive Rate (TPR), and F1-score to check the performance of our work. We calculated these metrics as follows:

$$P = \frac{tp}{tp + fp} \quad (5)$$

$$R = \frac{tp}{tp + fn} \quad (6)$$

$$Acc = \frac{tp + tn}{tp + fp + tn + fn} \quad (7)$$

$$TPR = \frac{tp}{p} \quad (8)$$

$$F1 = \frac{2PR}{P + R} \quad (9)$$

Where *tp*, *tn*, *fp*, and *fn* are showing the true positive, negative, and false positive, negative samples, respectively.

3.3. Proposed Method Performance

An experiment is designed to check the capability of the proposed solution for deepfakes recognition. Two classifiers namely SVM and ANN are employed over the computed keypoints of the deepfakes dataset. The class-wise deepfakes detection and classification performance of both classifiers, in form of precision, recall, F1-score, accuracy, and error rate are shown in Table 1. Our technique attained the average accuracy values of 98.5 and 95.35, and the average error rate of 1.64 and 3.9 for SVM and ANN classifiers respectively for both classes. The results clearly depict that for both classifiers, the presented work has acquired robust precision, recall, and F1-score values, with fewer error rates. The major cause for the enhanced deepfakes identification performance is due to the effectiveness of the employed keypoints extraction method that shows all classes in an effective way. Moreover, SVM shows better performance as compared to ANN classifiers because of its ability to tackle the over-tuned model data. Figure 2 shows the confusion matrix for SVM and ANN classifiers.

Table 1. Class wise performance of the introduced method.

Classes		P	R	F1	Acc	ER
Zero (real)	SVM	99	98.5	98.7	99.95	1.26
	ANN	97	96.6	96.7	98.4	3.21
One (deepfake)	SVM	99	97.0	97.9	98.77	2.02
	ANN	96	94.5	95.2	99.7	4.76

To show more rigorous experimentation of the presented model, we have trained both classifiers for all subjects. Table 2 shows the overall results of the proposed method in terms of precision, recall, F1-score, Error rate, and accuracy.

Table 2. Overall performance.

Proposed		P	R	F1	ER	Acc
deepfakes	SVM	98.77	98.3	98.53	1.47	99.87
	ANN	82.6	99.7	90.34	9.65	98.5

3.4. Comparative Analysis

In this section, we have evaluated the performance of our proposed model in terms of TPR and AUC with an existing method [1] over the deepfakes dataset [1] and comparative results are reported in Table 3. The results clearly show that our approach performs well for all subjects in comparison to the work in [1] for both evaluation metrics. The method in [1] attained the average True Positive Rate (TPR) of 0.88, while in the case of our method achieved a TPR of 0.92. Similarly, the method [1] attained AUC of 0.99, 0.95, 096, 090, and 0.98, while our technique achieved 1.0, 1.0, 0.99, 1.0, and 0.99 for the subjects namely Barack Obama, Hillary Clinton, Bernie Sanders, Donald Trump, and Elizabeth Warren respectively. More clearly, the work in [1] shows the average Area Under the Curve (AUC) and TPR values of 95.60% and 88.80% which are 99.20% and 88.80% for our case. Hence, for the AUC and TPR metrics, the introduced solution gives the average performance of 3.6% and 3.2% which is showing the show the robustness of the presented framework.

Furthermore, we have evaluated the proposed approach against state-of-the-art approaches namely VGG [8] and ResNet [8] in terms of TPR, and obtained results are shown in Figure 3. It is quite evident that our model has outperformed the other approaches. More specifically, the comparative methods show the average TPR value of 88.20% which is 99.20% for our method. Hence, we have provided an average performance gain of 3.8%. The major reason for the better performance of the proposed solution is because of the better face recognition ability of the Dlib model which assists in effectively detecting the real and manipulated faces.

Table 3. Subject-wise AUC of the existing and presented approach.

Subject	AUC		TPR	
	[1]	Proposed	[1]	Proposed
Barack Obama	0.99	0.98	0.97	0.99
Hillary Clinton	0.95	1	0.89	0.94
Bernie Sanders	0.96	1	0.92	0.93
Donald Trump	0.90	0.99	0.74	0.8
Elizabeth Warren	0.98	0.99	0.92	0.94

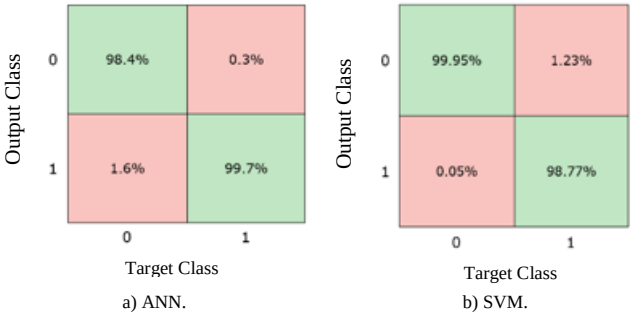

Figure 2. Confusion matrix.

4. Conclusions

In this paper, we presented a technique for precise and automated classification of real and fake samples in which facial features of source subjects are swapped with the target to manipulate the visual content. In the introduced method, Dlib facial bounding box library is employed to compute the landmark features from the input samples. Then, the calculated keypoints are used to train the SVM and ANN classifiers to distinguish the actual and manipulated content. The results exhibit that the proposed method accurately classifies the input samples and serves as a new automated tool for Face-swap based deepfakes detection. As a future direction, we aim to perform the classification of other types of deepfakes and to evaluate our approach over more complex deepfakes datasets.

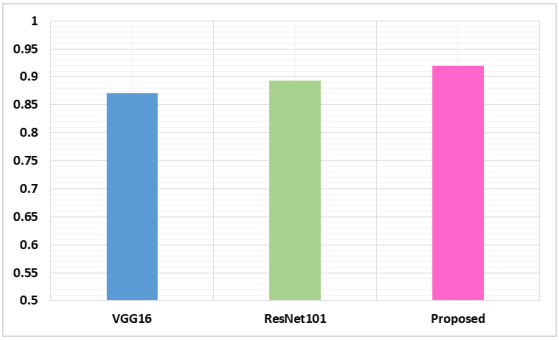

Figure 3. Comparison in terms of TPR.

Reference

[1] Agarwal S., Farid H., Gu Y., He M., Nagano K., and Li H., "Protecting World Leaders Against Deep Fakes," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, California, pp. 38-45 2019.

[2] Ballester P. and Araujo R., "On the Performance of GoogLeNet and AlexNet Applied to Sketches," *in Proceedings of the 3th AAAI Conference on Artificial Intelligence*, Phoenix, pp. 1124-1128, 2016.

[3] Baltrušaitis T., Robinson P., and Morency L., "Openface: An Open Source Facial Behavior Analysis Toolkit," *in Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Lake Placid, pp. 1-10, 2016.

[4] Hsu C. and Lin H., "Universal Forgery Attack on a Strong Designated Verifier Signature Scheme," *The International Arab Journal of Information Technology*, vol. 11, no. 5, pp. 425- 428, 2014.

[5] King D., "Dlib-ml: A Machine Learning Toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.

[6] Li Y., Chang M., and Lyu S., "In Ictu Oculi: Exposing ai Generated Fake Face Videos By Detecting Eye Blinking," *in Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, pp. 1-7, 2018.

[7] Nataraj L., Mohammed T., Manjunath B., Chandrasekaran S., Flenner A., Bappy J., and Roy-Chowdhury A., "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," *Electronic Imaging*, vol. 31, pp. 532-1-532-7, 2019.

[8] Nawaz M., Masood M., Javed A., Iqbal J., Nazir T., Mehmood A., and Ashraf R., "Melanoma Localization and Classification Through Faster Region-Based Convolutional Neural Network and SVM," *Multimedia Tools Applications*, vol. 80, no. 19, pp. 28953-28974, 2021.

[9] Nazir T., Javed A., Masood M., Rashid J., and Kanwal S., "Diabetic Retinopathy Detection based on Hybrid Feature Extraction and SVM," *in Proceedings of the 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics*, Karachi, 2019.

[10] Neocortext, *Reface App*. Available from: https://reface.app/, Last Visited, 2020.

[11] Petrov I., Gao D., Chervoniy N., Liu K., Marangonda S., Umé C., Jiang J., RP L., Zhang S., and Wu P., "DeepFaceLab: A Simple, Flexible And Extensible Face Swapping Framework," *arXiv preprint arXiv:2005.05535*, 2020.

[12] Pudaruth S., Soyjaudah S., and Gunputh R., "Classification of Legislations using Deep Learning," *The International Arab Journal of Information Technology*, vol. 18, no. 5, pp. 651-662, 2021.

[13] Sabir E., Cheng J., Jaiswal A., AbdAlmageed W., Masi I., and Natarajan P., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80-87, 2019.

[14] Suwajanakorn S., Seitz S., and Kemelmacher-Shlizerman I., "Synthesizing Obama: learning Lip Sync from Audio," *ACM Transactions on Graphics*., vol. 36, no. 4, pp. 1-13, 2017.

[15] Thies J., Zollhofer M., Stamminger M., Theobalt C., and Niessner M. "Face2face: Real-Time Face Capture and Reenactment of Rgb Videos," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2387-2395, 2016.

[16] Yang X., Li Y., and Lyu S., "Exposing Deep Fakes Using Inconsistent Head Poses," *in Proceedings of the ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, 2019.

Marriam Nawaz currently enrolled as Ph.D. student in Software Engineering Department, UET Taxila, Pakistan. Her research interest includes image/ video forensic analysis and digital image processing.

Momina Masood currently enrolled as Ph.D. student in Computer Science Department, UET Taxila, Pakistan. Her research interest includes image/ video forensic analysis and digital image processing.

Ali Javed associate professor in Computer Science Department, UET Taxila, Pakistan. His research interest includes multimedia signal processing, Machine learning and Computer Vision.

Tahira Nazir completed Ph.D. degree in Computer Science discipline form the Computer Science Department, UET Taxila, Pakistan. Her research interest includes Computer Vision and digital image processing.