Comparative Analysis of Internal and External Facial Features for Enhanced Deep Fake Detection

Fatimah Alanazi
School of Computing, Newcastle University, U.K.
College of Computer Science and Engineering, University of Hafr Albatin, Saudi Arabia

Keywords: Deepfakes Detection, Face Recognition, Image Analysis, Feature Fusion, Facial Features.

Abstract: In the burgeoning era of deepfake technologies, the authenticity of digital media is being perpetually challenged, raising pivotal concerns regarding its veracity and the potential malicious uses of manipulated content. This study embarks on a meticulous exploration of the effectiveness of both internal and external facial features in discerning deepfake content. By conducting a thorough comparative analysis, our research illuminates the criticality of facial features, particularly those situated beyond the face's center, in distinguishing between genuine and manipulated faces. The results elucidate that such features serve as potent indicators, thereby offering valuable insights for enhancing deepfake detection methodologies. Consequently, this research, therefore, not only underscores the paramount importance of these often-overlooked facial aspects but also contributes substantively to the domain of digital forensics, providing a nuanced understanding and innovative approaches towards advancing deepfake detection strategies. By bridging the gap between technological advancements and ethical digital media practices, this study stands as a beacon, advocating for the imperative need to safeguard the integrity of digital communications in our progressively digitized world.

1 INTRODUCTION

In the contemporary digital era, underscored by the omnipresence of digital media and the swift advancement of artificial intelligence (AI), the advent of deepfake technology has precipitated substantial apprehension amongst scholars, policymakers, and the general populace alike. The term "deepfake," a portmanteau of "deep learning" and "fake," signifies a category of manipulated media content that encompasses images, audio, and video, crafted meticulously through sophisticated machine learning algorithms(Dagar and Vishwakarma, 2022).

The burgeoning of deepfakes poses intricate challenges by eroding trust, disseminating misinformation, and potentially destabilizing foundational principles of truth and authenticity in digital media(Tolosana et al., 2020). Deepfakes refer to synthetic media where an individual's likeness in an existing image or video is substituted with another's, a technology that can be wielded for benign purposes such as creating memes or educational content, or malicious intents like promulgating misinformation and tarnishing reputations(Korshunov and Marcel, 2018).

The pursuit of deepfake detection, which entails the identification and verification of manipulated media, has become paramount in preserving digital integrity. Although various algorithms for deepfake detection have been devised, they are perpetually in a race against the escalating sophistication of deepfake creation methods(Tolosana et al., 2020).

In this paper, we endeavor to compare the efficacy of internal and external facial features in the realm of deepfake detection. Through the deployment of a deep learning model, trained to discern deepfakes utilizing both internal and external facial features, we unearth that the model attains augmented accuracy when deploying external facial features, such as the forehead and mouth, in its detection methodology.

Our research holds the potential to enhance the precision of deepfake detection algorithms, an imperative endeavor considering the malicious applications of deepfakes. By forging ahead in the development of more accurate deepfake detection algorithms, we contribute towards safeguarding individuals and entities from the detrimental repercussions of manipulated media.

Alanazi, F.
Comparative Analysis of Internal and External Facial Features for Enhanced Deep Fake Detection.
DOI: 10.5220/0012346100003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 308-314
ISBN: 978-989-758-680-4; ISSN: 2184-433X

2 RELATED WORK

The deepfake detection domain, characterized by a substantial surge in research and developmental activities, has witnessed a myriad of methodologies introduced with the primary objective of distinguishing between genuine and manipulative visual media. In this endeavor, a plethora of strategies, technologies, and approaches have been meticulously designed and deployed, each offering a unique perspective and solution to the challenges posed by deepfakes. In this section, we delve into a detailed exploration of various paradigms, ranging from biometric and physiological models to advanced neural network applications, aiming to draw insights and potentially identify gaps in existing methodologies.

2.1 Biometric and Physiological Approaches

In a pivotal exploration of biometric systems, (Menotti et al., 2015) ventured into uncharted territories, examining the potential of biometric systems in detecting a variety of spoofing attacks, which spanned across iris and fingerprint forgeries as well as facial recognition manipulations. Their methodology, a blend of a meticulously designed convolutional neural network (CNN) architecture and refined strategies such as fine-tuning network weights through backpropagation, presented a novel approach to biometric-based deepfake detection.This investigation into biometric systems underscored the paramount importance of physiological attributes in discerning authentic from manipulated content.

Moreover,the study(Xu et al., 2021) utilizes remote photoplethysmography (rPPG) technology for detecting deepfake videos. This method involves capturing periodic changes in skin color caused by the heartbeat cycle through sensors such as cameras. The focus is on leveraging deep learning-based methods to enhance the accuracy of rPPG algorithms. These advancements enable effective handling of challenges like lighting changes and motion artifacts in measurements. The application of rPPG in this study is particularly directed towards the detection of DeepFake videos, showcasing its potential in this emerging field. The paper discusses the principles of rPPG, its recent progress, and specifically, its application in DeepFake detection.

2.2 Facial Recognition and Mask Attacks

In a world increasingly relying on facial recognition technologies, researchers(Steiner et al., 2016) introduced a cross-modal approach aimed at thwarting facial mask attacks, a prevalent challenge in the field. This approach, which utilized multispectral short-wave infrared (SWIR) imaging, aimed to authenticate faces by mitigating errors and inaccuracies often induced by disguises or facial masks. The integration of a Support Vector Machine (SVM) classifier, when amalgamated with multispectral SWIR imaging, showcased a remarkable reduction in the false acceptance rate, offering a promising avenue for further research and development in this domain.

2.3 Emotional Authenticity and Facial Movements

The subtle and complex world of emotional expression was brought to the forefront in a study that utilized a CNN-based approach (Lee et al., 2020). The methodology, which was designed to discern seven distinct emotions, including authentic and simulated smiles, harnessed the FERC-2013 dataset, providing a comprehensive framework for understanding the nuances of emotional authenticity in the context of deepfake detection. Similarly, (Jafar et al., 2020) implemented a strategy combining a CNN and the DFT-MF technique, which critically assessed mouth movements and successfully detected forged videos and images, thereby achieving a high level of accuracy in analyzing mouth dynamics during speech.

2.4 Eye-Blinking Dynamics and GANs

Navigating through the complexities brought forth by the malicious utilization of Generative Adversarial Networks (GANs), researchers turned their focus toward eye-blinking, a physiological attribute often poorly replicated in forged videos (Ciftci et al., 2020)(Menotti et al., 2015). By employing Long-Term Recurrent Convolutional Networks (LRCN) to scrutinize the dynamics of eye-blinking, researchers have encountered limitations in videos featuring frequent blinking or altered facial features, yet offered a novel perspective on utilizing physiological characteristics for detection. Jung et al.(Jung et al., 2020) introduced Deep Vision, an algorithm that capitalized on predictable eye-blinking patterns as a mechanism to differentiate between genuine and manipulated videos.

2.5 Biological Data and Feature Analysis

In an innovative approach, (Ciftci et al., 2020) introduced a methodology grounded in the intricate analysis of biological data, employing variables such as heart rate to differentiate genuine content from manipulated videos. By training SVM and CNN models on both temporal and spatial facial features, the methodology demonstrated a promising potential in the realm of deepfake detection, albeit with potential susceptibilities when dimensionality reduction techniques were employed, thus necessitating further exploration and refinement in future research endeavors.

2.6 Present Work

In the context of this research, our focus is meticulously centered on the detailed analysis of facial features, with the objective of identifying the most accurate and reliable facial regions for deepfake detection. Our exploration involves a comparative analysis, scrutinizing the efficacy of both internal (such as eyes, nose, and mouth) and external (such as hairline and jawline) facial features, with a hypothesis that internal features may furnish more robust and informative data for the purpose of deepfake detection, offering a novel perspective in the ongoing battle against digital media manipulation.

3 METHOD

3.1 Face-Cut-Out

The Face-Cutout technique serves as a data augmentation method for enhancing the training of Convolutional Neural Networks (CNNs) aimed at improving deepfake detection. This technique generates training images with varying occlusions, relying on facial landmark information without regard to orientation. Facial landmarks include the positions of key facial features such as eyes, ears, nose, mouth, jawline, and forehead. Google's Media Pipe Face Mesh, a facial landmark detection model, can accurately identify 468 unique landmark positions on a human face in real-time.

To perform face cut-out, we grouped certain landmark positions to create polygons that were then occluded in the training images. We used two groups of polygons:

- baseline : images without augmentation
- Cut-out 1: chin, hair, jawline, and mouth (landmark positions 211-150, 103-67, 57-43, and 425-280, respectively)
- Cut-out 2: left eye ,right eye ,both eyes and the nose (landmark positions from 386 to 446 for the left eye, 53 to 340 for both eyes, and 6 to 419 for the nose) By using these positions to calculate polygons for the face cut-out, we generated training images with different occlusions for improved deepfake detection.

We applied Cut-outs 1 and 2 to the selected datasets, as shown in Figure 1.

Figure 1: The figure illustrates examples of the datasets generated in this study, comprising three distinct groups:(1) Baseline Images representing original, unaltered faces;(2) Face Cut-Out 1, which involves cutting out specific regions such as the left eye, right eye, both eyes, and nose; and (3) Face Cut-Out 2, which cuts out the forehead, chin, mouth, and jawline.

3.2 Dataset Selection

In the pursuit of empirically evaluating and training models tailored to deepfake detection, this study rigorously selected and incorporated two datasets, widely acknowledged and utilized in the research community for their robustness and comprehensiveness: the FaceForensics++ (FF++) dataset (Mhou et al., 2017) and the Celeb-DF dataset.

The FF++ dataset has garnered recognition as a benchmark in the sphere of face forgery detection, serving as a vital resource for researchers and practitioners alike. This dataset furnishes a comprehensive compilation, comprising over 1,000 original videos, along with their manipulated counterparts, each of which has been meticulously generated employing a range of deep learning-based face manipulation techniques. The inclusion of these manipulated videos, each varying in complexity and method of generation, offers a rich, diverse, and challenging dataset for training and evaluating deepfake detection models.

Conversely, the Celeb-DF dataset (Jafar et al., 2020) provides a distinctly large-scale resource, encapsulating 590 original videos and a staggering 5,639 deepfake videos, thereby providing a substantial volume of data for model training and evaluation. Moreover, the Celeb-DF dataset offers a richly diversified array of subjects spanning across varied ages,

Table 1.

Dental measurement	FF++			Celeb-DF		
	ACC	AUC	logloss	ACC	AUC	logloss
EfficientNet-B7 + Cut-out 1	0.66	0.78	1.12	0.90	0.88	0.44
EfficientNet-B7+ Cut-out2	0.80	0.84	0.53	0.92	0.93	0.25
EfficientNet-B7+ Baseline	0.77	0.81	0.59	0.91	0.89	0.51
Xception + Cut-out 1	0.75	0.81	0.90	0.90	0.91	0.35
Xception+ Cut-out 2	0.75	0.83	0.76	0.91	0.92	0.29
Xception+ Baseline	0.77	0.77	0.78	0.84	0.79	0.80

ethnic groups, and genders. This demographic diversity furnishes a holistic platform, enabling the exploration and evaluation of deepfake forensics research in a multitude of contexts and scenarios, ensuring the developed models are inclusive and effective across varied subject matter.

However,the choice of Celeb-DF and FF++ datasets for deepfake detection is due to their realistic, high-quality deepfakes, diverse methods, and large data volumes, providing a robust training platform and setting benchmarks in deepfake research.

In the process of utilizing these datasets, meticulous care was undertaken to adhere to conventional practices for dataset division, ensuring that the models were trained, validated, and tested in a rigorous and standardized manner. Specifically, the datasets were allocated with 80% of the data designated for training, facilitating the models to learn and adapt to the complexities and nuances of the deepfake videos. Subsequently, 10% of the data was reserved for validation, assisting in tuning and optimizing the models during the training process. Finally, the remaining 10% of the data was strictly utilized for testing purposes, ensuring an unbiased evaluation of the models' performance and capabilities in identifying deepfakes, providing a rigorous and thorough assessment of their applicability and effectiveness in real-world scenarios.

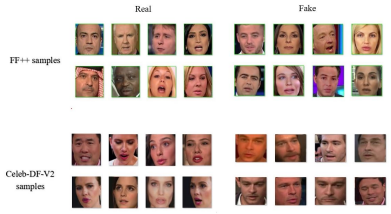

Figure 2: Sample images from each dataset that we used.

3.3 Model Selection

Two deep convolutional models, EfficientNet-B7 and XceptionNet, were chosen as feature extractors for the deepfake detection algorithm. Both models were initialized with pre-trained ImageNet weights, enabling them to leverage rich feature representations. XceptionNet employs depth-wise separable convolutions to optimize computation, and it has shown excellent performance in deepfake detection.

EfficientNet-B7, the largest variant in the EfficientNet architecture family, is known for its high performance and efficiency. It has achieved state-of-the-art results and is pre-trained using the Noisy Student technique, enhancing robustness.

3.4 Pre-Processing and Training Set-up

The dataset was pre-processed to capture every 10th frame from each video, and the OpenCV library was used to crop the images to focus on facial regions. Facial landmarks, as defined in section 3.1, were allocated using the MediaPipe library. Images were then divided based on the facial regions subjected to cut-outs. For training, images were normalized, resized to 224 x 224 resolution, and subjected to augmentations, including Image Compression, Gaussian Noise, and Flipping. The Rectified Adam optimizer was employed, with a learning rate scheduling strategy. Binary Cross-entropy Loss was used for model training, which was limited to 20 epochs with early stopping.A batch size of 64 was utilized for all experiments, and a GPU was employed for training.

3.5 Testing the Models

he model's evaluation was conducted using a pristine, non-augmented test dataset, specifically set aside as 10% of the total data for testing purposes. It is crucial to note that this test dataset was entirely separate from those used in the training and validation stages. This separation was deliberately chosen to ensure the model's performance could be accurately assessed on new, unseen data. For this evaluation, only the facial regions within the images were utilized. The focus of the assessment was on the model's ability to differentiate between genuine and manipulated faces, relying on the subtleties of facial features as key criteria.

3.6 Libraries and Toolkits Used

The research harnessed the capabilities of various Python libraries and toolkits, including but not limited to OpenCV for extracting video frames, ImageDataGenerator for data preparation and augmentation, and MediaPipe for facial detection and landmark allocation. Additionally, libraries such as Matplotlib, NumPy, Pandas, and scikit-learn were utilized for data visualization, manipulation, model evaluation, and metrics calculation, respectively, ensuring a comprehensive and robust methodological approach.

4 RESULTS

The evaluation of our proposed method's performance involves several aspects, including the application of face cut-out augmentations to the FaceForensics++ (FF++) and Celeb-DF datasets, comparative analysis with different training settings, and benchmarking against state-of-the-art deepfake detection techniques.A comparative analysis of results was conducted under three distinct settings. Baseline(Original faces without any augmentation),Cut-out 1(Four cut-outs placed strategically on the chin, mouth, jawline, and forehead regions) and Cut-out 2(Four cut-outs placed strategically on the left eye, right eye, both eyes, and nose regions.)

- Phase One: Cut-out Technique Evaluation with Each Dataset During this phase, three image groups were created from each dataset: Baseline, Cut-out 1, and Cut-out 2. Subsequently, these groups were trained using the selected deep convolutional models, EfficientNet-B7 and XceptionNet. The results demonstrated that models trained with the Cut-out 2 group significantly outperformed those in the Baseline and Cut-out 1 groups (Figure.3). Interestingly, the Cut-out 1 group occasionally underperformed the Baseline group, as observed in training with the EfficientNet-B7 model (Table1).

  The results indicated substantial improvements in the performance of the EfficientNet and Xception models when trained with the Cut-out 2 group, with accuracy gains ranging from 1.23% to 17.7% compared to the Baseline group. These findings highlight the effectiveness of the Cut-Out 2 dataset in training more robust facial recognition models. This can be attributed to the enforced learning within the Cut-Out 2 dataset, emphasizing distinguishing facial regions that are critical in differentiating fake from genuine faces.

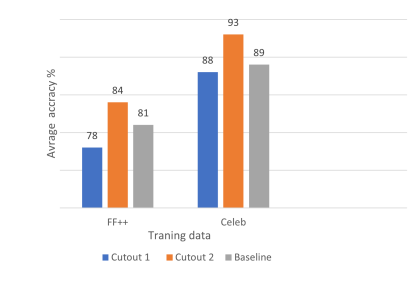

Figure 3: Test results in phase 1.

In the context of the Celeb-DF dataset, the Cut-out 2 group achieved log-loss results 43.18% better than the Cut-out 1 group when using the EfficientNet-B7 model, closely matched by the Xception model. Models trained with face Cut-out 2 augmentations consistently demonstrated superior performance. These findings suggest that training with Cut-out 2 images led models to prioritize the exposed facial regions, such as the forehead, cheeks, and chin. Consequently, it can be inferred that regions of the face outside the central features (eyes and nose) provide more significant information for discerning differences between authentic and synthetic faces.

This aligns with a study by Huang et al.(Huang et al., 2012), which found that even when facial expressions were occluded, models could identify a majority of facial expressions by relying on external facial features for cues. Similarly, our study suggests that features beyond the central region of the face contain crucial information for detecting disparities between similar faces. In cases where faces are very similar, such as with deepfakes, focusing on facial features beyond the central region of the face can lead to more accurate detection of differences.

- Phase Two: Evaluation of the Performance with a Combined Dataset

In this phase, the datasets from Phase One were combined to increase the overall volume of training data. This augmentation aimed to improve the model's generalization and performance with previously unseen data while exposing it to more diverse examples. Table 2 presents the results of the second phase, which evaluated the performance of three different groups (Baseline, Cut-out 1, and Cut-out 2) with a combined dataset of face images. The models were trained for 20 epochs, and their performance was assessed based on AUC, ACC (accuracy), and log-loss metrics.

EfficientNet-B7 with Cut-out 2 achieved the best performance, with an AUC of 0.89, ACC of 0.91, and log-loss of 0.45. The Xception model with Cut-out 2

Table 2: Phase Two results.

Models/Cutout type	Combined data-set		
	ACC	AUC	logloss
EfficientNet-B7 + Cut-out 1	0.89	0.90	0.48
EfficientNet-B7+ Cut-out2	0.89	0.91	0.45
EfficientNet-B7+ Baseline	0.90	0.87	0.69
Xception + Cut-out 1	0.83	0.85	0.84
Xception+ Cut-out 2	0.86	0.88	0.85
Xception+ Baseline	0.77	0.73	0.94

also performed well, with an AUC of 0.86, ACC of 0.88, and log-loss of 0.85. The EfficientNet-B7 and Xception models with baseline datasets showed similar performance. These results suggest that the use of Cut-out 2 effectively improves the performance of facial recognition models compared to Cut-out 1 and the Baseline dataset. The study also highlights that the EfficientNet-B7 model is more effective for face recognition than the Xception model. Cut-out 2 likely helps preserve facial features, allowing the model to better learn and recognize them, leading to improved performance in face recognition tasks. Additionally, it emphasizes that external facial regions, such as the forehead, cheeks, and chin, are vital for deepfake detection. The results revealed unexpected insights into the significance of external facial regions, such as the forehead, cheeks, and chin, in identifying deepfakes. Contrary to expectations, AI models exhibited improved performance when focusing on these external areas. This contrasts with previous findings in prosopagnosia research, where individuals with face recognition difficulties improved by concentrating on core facial features like the eyes and nose. Furthermore, it is surprising that the AI model could detect deepfakes even when the eyes, traditionally considered essential for face recognition, were not visible in cases like the Cut-out 2 group, where the eyes and nose regions were removed from facial images.

5 CONCLUSIONS

In conclusion, our experiments, reinforced by the discoveries of fellow researchers, demonstrate that in cases where facial resemblances are prominent, directing attention to facial features located beyond the central regions of the face results in more precise distinctions between faces.

Our findings that external facial regions may play a crucial role in deepfake detection open several avenues for future research in the field of deepfake detection and computer vision.One promising avenue for future research is to compare the performance of AI models focusing on external regions with human perception. This would involve analyzing situations where humans excel in recognizing deepfakes compared to AI models and vice versa. Such an investigation could provide insights into how to leverage the strengths of both humans and AI to develop more effective deepfake detection systems.

Another important area of future research is to conduct real-world testing of AI models that focus on external facial regions. This would involve testing the models in practical scenarios where internal facial regions may be obscured due to factors like masks, sunglasses, or low lighting conditions. This research would help to assess the feasibility of using these models in real-world deepfake detection applications. Overall, our findings suggest that external facial regions may play a crucial role in deepfake detection. This opens up several promising avenues for future research in the field of deepfake detection and computer vision.

REFERENCES

Ciftci, U. A., Demir, I., and Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

Dagar, D. and Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International journal of multimedia information retrieval*, 11(3):219–289.

Huang, X., Zhao, G., Zheng, W., and Pietikäinen, M. (2012). Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, 33(16):2181–2191.

Jafar, M. T., Ababneh, M., Al-Zoube, M., and Elhassan, A. (2020). Forensics and analysis of deepfake videos. In *2020 11th international conference on information and communication systems (ICICS)*, pages 053–058. IEEE.

Jung, T., Kim, S., and Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154.

Korshunov, P. and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Lee, M., Lee, Y. K., Lim, M.-T., and Kang, T.-K. (2020). Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. *Applied Sciences*, 10(10):3501.

Menotti, D., Chiachia, G., Pinto, A., Schwartz, W. R., Pedrini, H., Falcao, A. X., and Rocha, A. (2015). Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):864–879.

Mhou, K., van der Haar, D., and Leung, W. S. (2017). Face spoof detection using light reflection in moderate to low lighting. In *2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 47–52. IEEE.

Steiner, H., Sporrer, S., Kolb, A., Jung, N., et al. (2016). Design of an active multispectral swir camera system for skin detection and face verification. *Journal of Sensors*, 2016.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

Xu, Y., Zhang, R., Yang, C., Zhang, Y., Yang, Z., and Liu, J. (2021). New advances in remote heart rate estimation and its application to deepfake detection. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 387–392. IEEE.