Detection of Digital Manipulation in Facial Images (Student Abstract)

Aman Mehra[1], **Akshay Agarwal**[1,2], **Mayank Vatsa**[3], **Richa Singh**[3]
[1]IIIT-Delhi, India
[2]Texas A&M University, Kingsville, USA
[3]IIT Jodhpur, India
{aman17017, akshaya}@iiitd.ac.in, mvatsa@iitj.ac.in, richa@iitj.ac.in

Abstract

Advances in deep learning have enabled the creation of photo-realistic DeepFakes by switching the identity or expression of individuals. Such technology in the wrong hands can seed chaos through blackmail, extortion, and forging false statements of influential individuals. This work proposes a novel approach to detect forged videos by magnifying their temporal inconsistencies. A study is also conducted to understand role of ethnicity bias due to skewed datasets on deepfake detection. A new dataset comprising forged videos of Indian ethnicity individuals is presented to facilitate this study.

Introduction

With the rise of machine learning algorithms, it has become possible to switch the face of one individual with another (Agarwal et al. 2017; Singh et al. 2020; Majumdar et al. 2019). Recent advances have made such forgeries, popularly known as Deepfakes, become increasingly photo-realistic. The limited compute and expertise needed for their synthesis make them ubiquitous. Being nearly indiscernible to the human eye, these deepfakes pose a severe threat to society. Forgeries mimicking influential individuals can cause stock manipulation, diplomatic turmoil, and the swaying of voter sentiment. Thus, it is imperative to have detection techniques capable of accurately identifying them.

Most digital media undergoes some form of compression online, which deteriorates the performance of forgery detectors. An effective system to detect forgery must be resilient to compression while maintaining a similar performance across different demographics. The data used to build such systems plays an important role in preventing demographics biases. To build unbiased systems, it is first important to understand the extent of their impact. In this paper, we study the impact of ethnicity bias by varying the proportion of Caucasians (skin tone #1 and #2 in the Fitzpatrick's skin color codes) and Indians (mostly skin tone #3, #4, #5 in the Fitzpatrick's skin color codes) in a controlled dataset. To enable this, a new dataset comprising deepfakes of Indian individuals is presented. Subsequently, a novel approach amplifying temporal inconsistencies in doctored videos through motion magnification is proposed. It employs an optimized 3D ConvNet architecture designed for the task of video forgery detection. We analyze the proposed approach and demonstrate its efficacy on two datasets - FaceForensics++ (Rössler et al. 2019) and Celeb-DF (Li et al. 2020).

Methods	Majority	Merging Fraction α				
		0.0	0.1	0.3	0.5	1.0
MesoNet	Indian	8.77	11.39	7.95	11.21	4.11
	Caucasian	0.64	6.56	4.81	0.35	4.27
XceptionNet	Indian	14.82	15.14	7.27	1.69	2.82
	Caucasian	3.04	6.52	11.94	7.63	0.01
DenseNet	Indian	19.92	7.18	7.39	11.14	4.28
	Caucasian	10.99	2.43	6.14	14.14	2.85

Table 1: Ethnicity bias as the absolute generalization gap % between test sets of the majority and minority ethnicities.

Ethnicity Bias in Deepfake Detection

Indian Forgery Dataset: A new dataset is presented to study ethnicity bias. It contains 200 pristine and 248 forged videos of Indian subjects, created using FSGAN (Nirkin, Keller, and Hassner 2019), a subject agnostic face-swapping technique. These videos are derived from 149 unique identities and have a gender ratio of 1.6 (Male: Female).

Ethnicity Bias Study: We investigate whether skewed ethnicity distributions hinder the ability to identify deepfakes in under-represented ethnicities. Indian and Caucasian are selected with nearly equal pristine and forged videos (roughly 110 each). Caucasian videos are taken from FaceForensics++, while the Indian videos are taken from the newly created Indian Forgery Dataset. We maintain a common deepfake generation technique and gender ratio across ethnicities to prevent extrinsic factors from affecting the experiments.

Multiple architectures are trained on various combinations of the two ethnicities, governed by a fractional factor α. Table 1 shows the generalization gap as a proxy for the bias. The majority of the models possess bias against the minority ethnicity when trained on skewed datasets (low α). On increasing the under-represented ethnicity to over 30% of the majority, this effect diminishes, indicating that the effect of bias can be kept in check by introducing limited samples of the under-represented ethnicity.

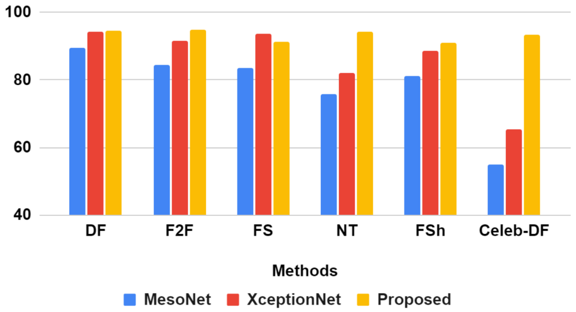

Figure 1: Results on Celeb-DF (AUC in %) and all manipulations of FaceForensics++ (Accuracy in %). NT is neural texture, FS is FaceSwap, F2F is Face2Face, DF is Deepfake, and FSh is FaceShifter.

Proposed DeepFake Detection Algorithm

It is hypothesized that the lack of enforced temporal coherency in the synthesis pipeline of forged videos leaves behind subtle motion inconsistencies. These artifacts can be amplified to become perceptible to the classification network, even in the presence of heavy compression. For that, in this research, face cropped video clips are pre-processed with deep learning-based motion magnification. The magnification follows the Eulerian perspective of motion, observing variations in individual pixels over time rather than tracking movement across frames as in optical flows. Later, the magnified clip is concatenated with the difference between them and the original clip. The intuition behind this is that both inputs complement each other to learn different features. For example, the magnified clips account for the lack of color information in the difference clips.

We propose a modified 3D-SE-ResNet (Hara, Kataoka, and Satoh 2017) by incorporating dense skip connections to detect forged videos. However, unlike the DenseNet (Huang et al. 2017), these connections are placed between residual blocks rather than within them, enabling better optimization without incurring the high memory demands of the DenseNet. The network is trained on fixed-length clips using binary cross-entropy loss.

Results

The results on FaceForensics++ (c40) are reported as the classification accuracy at the video level. The classification threshold is calculated on the validation set of all clip level predictions associated with each video. On Celeb-DF, results are reported as the clip level (16 contiguous frames) area under the curve (AUC) to facilitate comparison with existing techniques. Figure 1 demonstrates the efficacy of the proposed approach by outperforming existing techniques across manipulation types. The results on FaceForensics++, observed at high compression, highlight the proposed method's resilience to compression. Additionally, an ablation study is conducted to ascertain the influence of each component. The results shown in Figure 2 demonstrate an improving trend with the addition of each element (dense connections, magnified input, and the difference between magnified and original), thus validating the inclusion of each of them.

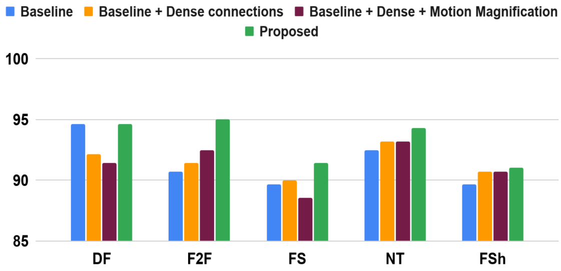

Figure 2: Ablation results (%) on FaceForensics++(c40). The 3D-SE-ResNet is the baseline and accuracies are presented for each additional component.

Conclusion

In this paper, we showcase ethnicity bias on forgery detection for the first time in literature. It is shown that unbalanced ethnicity distribution in the training data can deteriorate the performance for the under-represented testing subgroup. To facilitate this study, a new dataset of Indian individuals is presented. Additionally, a novel compression resilient forgery detection approach is proposed. The proposed algorithm employs motion magnification to amplify temporal inconsistencies in videos and demonstrate state-of-the-art performance on multiple datasets.

Acknowledgements

The research is partially supported by the Visvesvaraya Ph.D. Fellowship, Swarnajayanti Fellowship by the Government of India (GoI), and a research grant from MeitY, GoI.

References

Agarwal, A.; Singh, R.; Vatsa, M.; and Noore, A. 2017. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE IJCB*, 659–665.

Hara, K.; Kataoka, H.; and Satoh, Y. 2017. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. In *IEEE ICCVW*, 3154–3160.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *IEEE CVPR*, 2261–2269.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE/CVF CVPR*, 3207–3216.

Majumdar, P.; Agarwal, A.; Singh, R.; and Vatsa, M. 2019. Evading face recognition via partial tampering of faces. In *IEEE CVPRW*.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE ICCV*, 7184–7193.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*, 1–11.

Singh, R.; Agarwal, A.; Singh, M.; Nagpal, S.; and Vatsa, M. 2020. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, 3583–13589.