Recurrent-Convolution Approach to DeepFake Detection – State-Of-Art Results on FaceForensics++

Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, Prem Natarajan
USC Information Sciences Institute, Marina del Rey, CA, USA
{esabir, chengjia, ajaiswal, wamageed, iacopo, pnataraj}@isi.edu

Abstract

Spread of misinformation has become a significant problem, raising the importance of relevant detection methods. While there are different manifestations of misinformation, in this work we focus on detecting face manipulations in videos. Specifically, we attempt to detect Deepfake, Face2Face and FaceSwap manipulations in videos. We exploit the temporal dynamics of videos with a recurrent approach. Evaluation is done on FaceForensics++ dataset and our method improves upon the previous state-of-the-art up to 4.55%.

1. Introduction

A spate of recent incidents have increased the scrutiny of online misinformation [7, 2]. This has spurred research for both analysis and detection of misinformation [24, 18]. Misinformation can be manifested in different ways - deliberate manipulation of information or presenting unmanipulated content in a misleading context. Digital image manipulation such as copy-move and splicing [26, 25] are examples of deliberate manipulation, while image repurposing [10, 21, 11] is an example of misleading context. Out of the wide range of manipulations on different modalities, face manipulation in videos has recently garnered interest. Increased proliferation of fake videos may be attributed to two reasons:

1. Transposing a person's identity or expression with someone else's, is easier now [5].
2. A video is more likely to be believable since it demonstrates the activity in progress.

In this work, we attempt to solve the problem of face manipulation in videos. We leverage existing work from activity recognition to exploit temporal discrepancies in manipulated videos for improving upon the state-of-the-art. We also experiment with face alignment which shows an improvement.

2. Related Work

Activity Recognition in videos has well developed literature and can be used to gain insight for processing videos. There are two major approaches in this area. The first is a two stream network approach [22], where a video frame and optical flow are processed in two separate branches followed by fusion. The second is a single stream network, such as a recurrent convolutional network where each frame is processed by a convolutional neural network (CNN) and extracted features are fused by a recurrent network [6].

Datasets for face manipulation detection in videos had been lacking until recently with the release of FaceForensics [19] and FaceForensics++ [20] datasets. FaceForensics released Face2Face manipulated videos [23]. FaceForensics++ is an extension with *Deepfake* [5], *Face2Face* [23] and *FaceSwap* [15] manipulations. The dataset comprises 1000 videos with 720 in training and 140 each for validation and test. All videos are collected from youtube. *Deepfakes* are an autoencoder based approach to face identity manipulation, while *FaceSwap* is a graphics based approach for the same. *Face2Face* is a graphics based approach for performing facial reenactment instead of face identity swap. The resulting face acquires a new expression, whilst maintaining identity. There are three versions of this dataset based on compression: high quality (no compression), low quality (mild compression) and very low quality (heavy compression).

Copy move and splicing detection datasets and methods are abundant for both images and videos [26, 17]. However, due to the recent emergence of face manipulation problem the literature is relatively sparse. Rossler *et al.* [20] introduced baselines based on existing methods and trained an XceptionNet [4] architecture to establish state-of-the-art. MesoNet [1] also introduces two CNN based architectures for face manipulation detection. They take a mesoscopic approach to manipulation detection which combines information from both low level (microscopic) and high level (macroscopic) features.

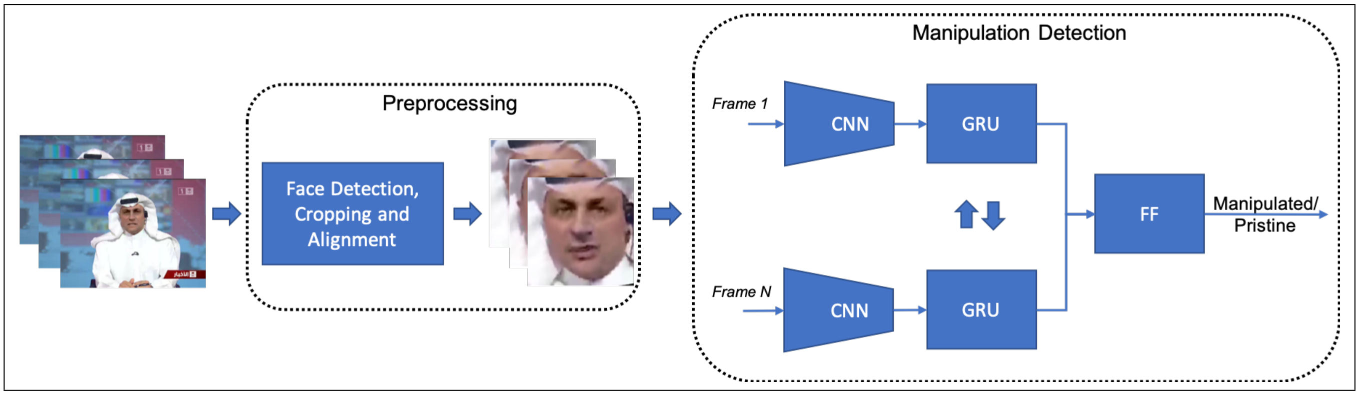

Figure 1: The overall pipeline is a two step process. The first step detects, crops and aligns faces on a sequence of frames. The second step is manipulation detection with our recurrent convolutional model.

3. Method

The overall approach for manipulation detection involves preprocessing - detection, cropping and alignment of faces from video frames, followed by manipulation detection over the preprocessed region.

For cropping the face region, we use the masks provided by [20], generated using computer graphics [23]. Since it has been shown in [16] that face alignment is beneficial for face recognition, we employ face alignment here as well. Face images are aligned using a simple similarity transformation (4 DoF), compensating for isotropic scale, in-plane rotation, and 2D translation. Most of the faces are near-frontal and thus it was sufficient to employ an accurate yet fast landmark detector method [13] implemented through [14]. Though [13] returns dense landmarks on the face, we select only a set of seven sparse points located on the most discriminative features of the faces (corners of the eyes, the tip of the nose, and corners of the mouth). Following the similarity transformation, faces are aligned with a loose crop at a 224×224 resolution.

For manipulation detection, we use a recurrent-convolutional network similar to [6], where the input is a sequence of frames from the query video. The intuition behind this model is to exploit temporal discrepancies across frames. Temporal discrepancies are expected to occur in images, since manipulations are performed on a frame by frame basis. As such, low level artifacts caused by manipulations on faces are expected to further manifest themselves as temporal artifacts with inconsistent features across frames. There are two differences from [6] in our implementation: (1) instead of using CaffeNet [12], we explore more suitable CNN architectures for the problem; (2) Instead of averaging recurrent features across all timesteps, we extract the final output of the recurrent network. Fig. 1 shows the model diagram. In our experiments, we explore ResNet [8] and DenseNet [9]. There are two reasons for exploring these CNNs. FaceForensics++ [20] is a low resource dataset with a 1,000 videos and to avoid overfitting, authors had to use pre-trained XceptionNet [4] with fixed feature extraction layers. For end to end trainability of CNN, we chose ResNet [8] which has been shown to be easily trainable. Additionally, manipulation artifacts exhibit low level features (such as discontinuous jawlines, blurred eyes etc.) which do not require high level face semantic features. DenseNet is a suitable CNN architecture, which extracts features at different levels of hierarchy [9].

4. Experiments

Our evaluation metric is accuracy for a fair comparison with baselines in [20]. Additionally, we report area under roc curve (AUC) scores for all experiments. All numbers are reported on FaceForensics++. For training, we use Adam optimizer with 1e-4 learning rate. We use GRU cells [3] for our recurrent network. Additionally, all results are on the heavily compressed version of dataset from [20]. We do not evaluate on high and low quality videos since the baseline performance for those is already above 98% in [20].

Table 1 shows our results on *Deepfakes*, *Face2Face* and *FaceSwap* manipulation. We consistently find DenseNet to outperform ResNet, face alignment to give improvement and a sequence of images to be better than single frame input. For our experiments, we process five frames in a sequence. Table 2 reports ROC plots for the same results. Since the false alarm region is hard to discern, we also show the semi-log plots.

5. Conclusion

Misinformation in online content is increasing and there is an exigent need for detecting such content. Face manipulation in videos is one aspect of the large problem. In this work we showed that a recurrent-convolutional approach improves upon the state-of-the-art.

Manipulation	Frames	FF++ [20]	ResNet50	DenseNet	ResNet50 + Alignment	DenseNet + Alignment	ResNet50 + Alignment + BiDir	DenseNet + Alignment + BiDir
Deepfake	1	93.46	94.8	94.5	96.1	96.4	-	-
	5	-	94.6	94.7	96.0	96.7	94.9	**96.9**
Face2Face	1	89.8	90.25	90.65	89.31	87.18	-	-
	5	-	90.25	89.8	92.4	93.21	93.05	**94.35**
FaceSwap	1	92.72	91.34	91.04	93.85	96.1	-	-
	5	-	90.95	93.11	95.07	95.8	95.4	**96.3**

Table 1: Accuracy for manipulation detection across all manipulation types. DenseNet with alignment and bidirectional recurrent network is found to perform best. FF++ [20] is the baseline in these experiments.

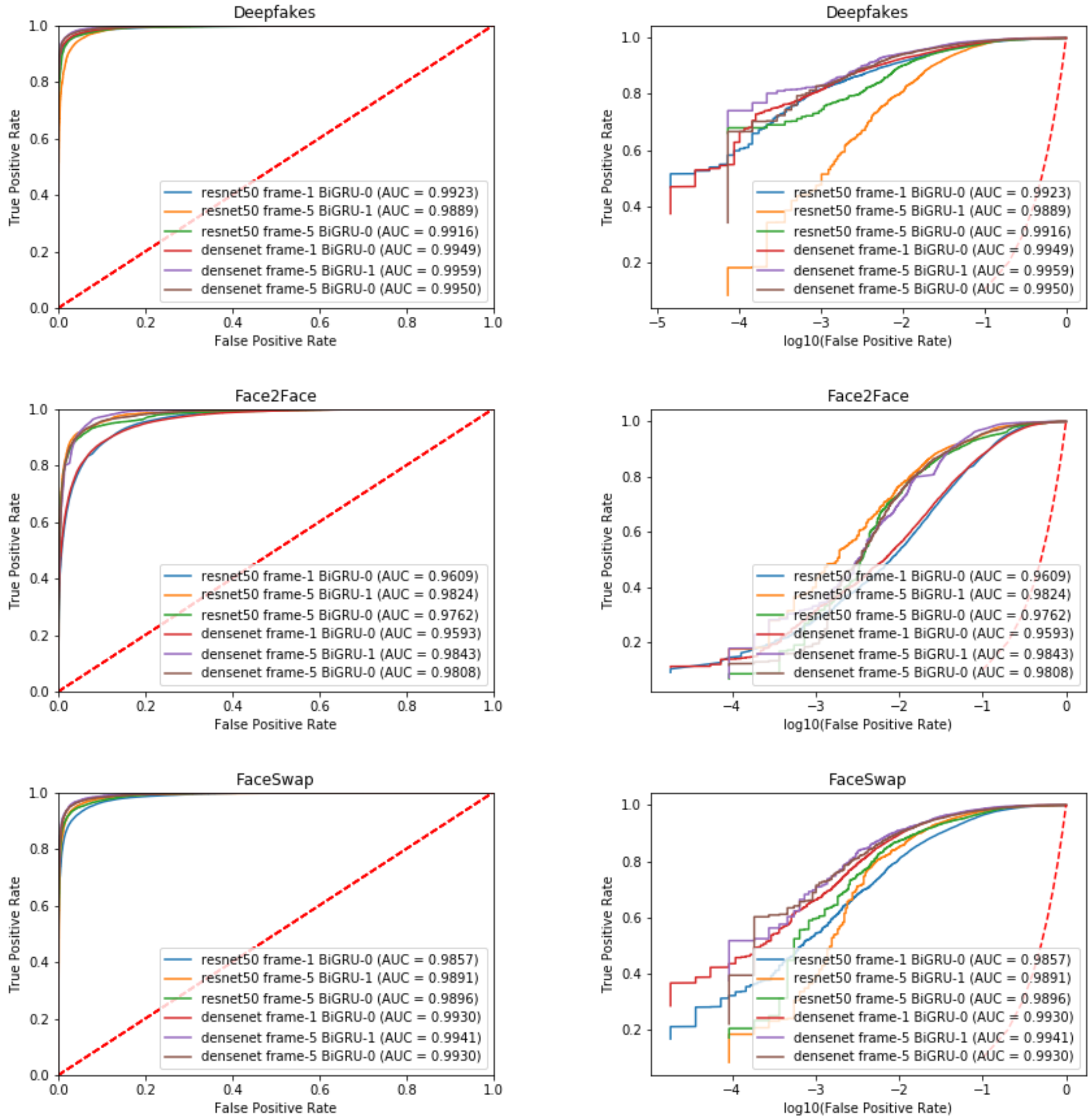

Table 2: ROC plots for all manipulation types. Each row corresponds to a different manipulation type. The left column is a linear plot, while the right column has semi-log plots to better analyze the false alarm region.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec. 2018. 1

[2] H. Allcott and M. Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017. 1

[3] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*, June 2014. arXiv: 1406.1078. 2

[4] F. Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. pages 1251–1258, 2017. 1, 2

[5] deepfakes. Non official project based on original /r/Deepfakes thread. Many thanks to him!: deepfakes/faceswap, Apr. 2019. original-date: 2017-12-19T09:44:13Z. 1

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. pages 2625–2634, 2015. 1, 2

[7] E. Ferrara. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. SSRN Scholarly Paper ID 2995809, Social Science Research Network, Rochester, NY, June 2017. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016. 2

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. pages 4700–4708, 2017. 2

[10] A. Jaiswal, E. Sabir, W. AbdAlmageed, and P. Natarajan. Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 1465–1471, New York, NY, USA, 2017. ACM. event-place: Mountain View, California, USA. 1

[11] A. Jaiswal, Y. Wu, W. AbdAlmageed, I. Masi, and P. Natarajan. AIRD: Adversarial Learning Framework for Image Repurposing Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 2

[13] V. Kazemi and J. Sullivan. One Millisecond Face Alignment with an Ensemble of Regression Trees. pages 1867–1874, 2014. 2

[14] D. E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. 2

[15] Marek. 3d face swapping implemented in Python. Contribute to MarekKowalski/FaceSwap development by creating an account on GitHub, Apr. 2019. original-date: 2016-06-19T00:09:07Z. 1

[16] I. Masi, F.-J. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, and T. Hassner. Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):379–393, 2019. 2

[17] R. C. Pandey, S. K. Singh, and K. K. Shukla. Passive copy-move forgery detection in videos. In *2014 International Conference on Computer and Communication Technology (ICCCT)*, pages 301–306, Sept. 2014. 1

[18] N. Ruchansky, S. Seo, and Y. Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806, New York, NY, USA, 2017. ACM. event-place: Singapore, Singapore. 1

[19] A. Rssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niener. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv:1803.09179 [cs]*, Mar. 2018. arXiv: 1803.09179. 1

[20] A. Rssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niener. FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv:1901.08971 [cs]*, Jan. 2019. arXiv: 1901.08971. 1, 2, 3

[21] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan. Deep Multimodal Image-Repurposing Detection. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 1337–1345, New York, NY, USA, 2018. ACM. event-place: Seoul, Republic of Korea. 1

[22] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014. 1

[23] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Face2face: Real-Time Face Capture and Reenactment of RGB Videos. pages 2387–2395, 2016. 1, 2

[24] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. 1

[25] Y. Wu, W. Abd-Almageed, and P. Natarajan. Deep Matching and Validation Network: An End-to-End Solution to Constrained Image Splicing Localization and Detection. In *Pro-*

ceedings of the 25th ACM International Conference on Multimedia, MM '17, pages 1480–1502, New York, NY, USA, 2017. ACM. event-place: Mountain View, California, USA. 1

[26] Y. Wu, W. Abd-Almageed, and P. Natarajan. BusterNet: Detecting Copy-Move Image Forgery with Source/Target Localization. pages 168–184, 2018. 1