PRESENTATION ON
RECURRENT CONVOLUTIONAL STRATEGIES FOR FACE MANIPULATION DETECTION IN VIDEOS
(ARXIV: 1905.00582)

Pratikkumar.Prajapati@sjsu.edu

Apr/26/2020

OVERVIEW

- *The original paper at [1]*
- *Features: Image + Temporal Information*
- *Model: CNN + RNN*
- *Dataset: FaceForensics++ (FF++) [2]*
  - *1000 videos = 720 training + 140 validation + 140 test*
- *Accuracy:*
  - *AUC of 96.9% DeepFake, LQ*
  - *AUC of 96.3% FaceSwap, LQ*

OVERALL ARCHITECTURE

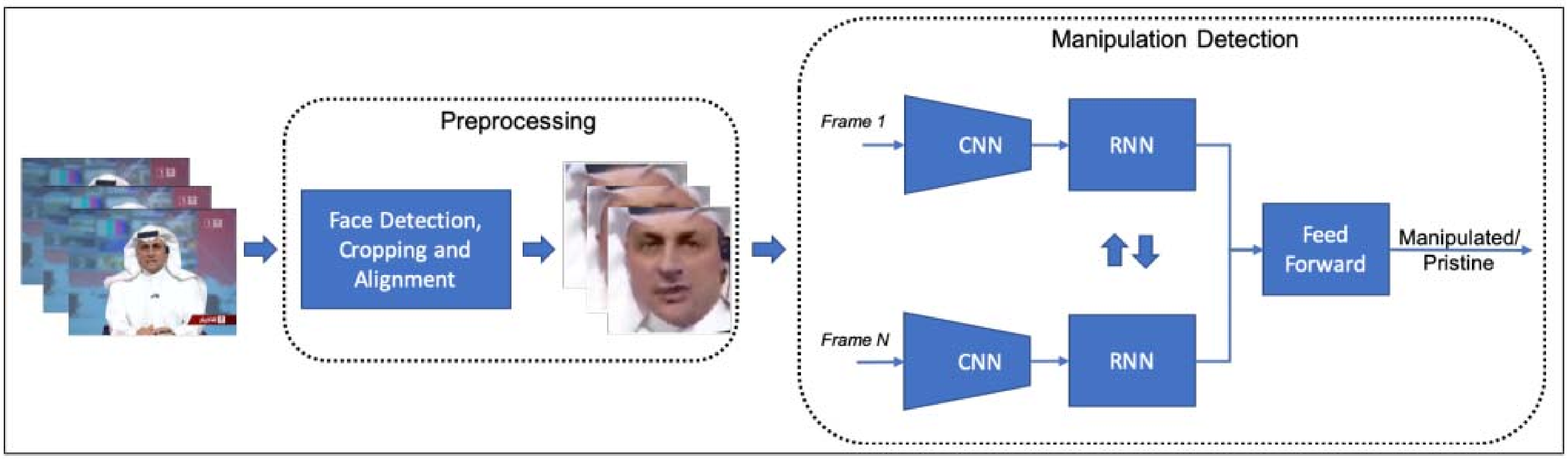

Figure 1: The overall pipeline is a two step process. The first step detects, crops and aligns faces on a sequence of frames. The second step is manipulation detection with our recurrent convolutional model.

PREPROCESSING

- Uses masks provided by [2] to crop the face region
- Face alignment
  1. Landmark-based alignment
     - seven sparse points of the face are used.
     - corners of the eyes, the tip of the nose, and corners of the mouth.
  2. Spatial Transformer Network (STN)
     - performs spatial alignment of data with learnable affine transformation parameters.
     - a differentiable module which can be inserted into other CNN models to learn about features and landmarks.

MANIPULATION DETECTION

- Inputs are sequence of frames from the target video
- CNN
  - To learn about features of the image (frame(s) of video)
  - ResNet and DenseNet as backbone
- RNN
  - Uses GRU to exploit temporal discrepancies across frames.
  - Temporal discrepancies are expected to occur in images, since manipulations are performed on a frame-by-frame basis.
- The backbone n/w is first trained on FF++ training split minimizing cross-entropy loss to train to detect real from synthesized faces.
- The backbone is then extended with RNN and re-trained end-to-end.
- Adam optimizer

RNN TRAINING STRATEGIES

1. A single RNN on top of final features leant from backbone n/w.
2. Multiple RNN at different levels of hierarchy of the backbone net.
   - To utilize micro, meso and macroscopic features

RESULTS OF EXPERIMENTS (1/2)

Manipulation	Frames	FF++ [34]	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++ [34] is the baseline in these experiments.

RESULTS OF EXPERIMENTS (2/2)

Manipulation	Base	Variation	
		Spatial Transformer	Multi Recurrence
Deepfake	**96.9**	91.7	94.4
Face2Face	**94.35**	87.46	89.9
FaceSwap	**96.3**	93.2	94.8

Table 2: Results on using variations to the recurrent convolutional architecture. Both spatial-transformer networks and multi-recurrent networks exhibit a decline in performance.

REFERENCES

[1]. E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in Proc. Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[2]. A. Rossler, 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