CS282 Fakebusters: Robust Speaker Identification System

Akshay Punhani, Daniel Rincón, Shruti Agarwal, Tanya Piplani

May 2019

1 Abstract

The creation of sophisticated fake videos has been largely relegated to Hollywood studios or sophisticated state actors. Recent advances in deep learning, however, have made it significantly easier to create sophisticated and compelling fake videos. With relatively modest amounts of data and computing power, the average person can, for example, create a video of a world leader confessing to colluding with a foreign government to win an election. These so called deep fakes [1] pose a significant threat to our democracy and national security. To contend with this growing threat, we want to develop a forensic technique that models facial movements to typify a individual's speaking pattern. Although not visually apparent, these correlations are often violated by the nature of how deep fake videos are created and can, therefore, be used for authentication.

2 Introduction

Harmful lies are ubiquitous. But the ability to distort reality has taken an exponential leap forward with DeepFakes [**?**] technology. This capability makes it possible to create audio and video of real people saying and doing things they never said or did. Machine learning techniques are escalating the technology's sophistication, making deep fakes ever more realistic and increasingly resistant to detection. While deep-fake technology will bring with it certain benefits, it also will introduce many harms. Individuals and businesses will face different forms of exploitation, intimidation, and personal sabotage.

3 Related Research

3.1 Face/Mouth based video forgery methods

Multiple approaches that create face-based manipulated videos are proposed recently. For example, real-time transfer of facial expression into an existing video [19, 9]; re-synthesis of mouth in an existing video [18, 11]. This technique trains a model to replicate the mouth movements that are associated with specific audio snippets. Therefore when the audio if faked that mouth movements are 'biometricaly' accurate but misaligned; face swapping of the person in a video with another person [2, 3]; automatic generation of talking head generation [14, 20].

3.2 Forgery detection methods

In order to detect the above video manipulations, a number of detection techniques have also been proposed [12, 21, 16, 15]. The method is based on the observations that current DeepFake algorithm can only generate images of limited resolutions, which need to be further warped to match the original faces in the source video. Such transforms leave distinctive artifacts in the resulting DeepFake videos, and can be effectively captured by convolutional neural networks.

3.3 Speaker recognition using audio-video information

There have been multiple research to recognize a person using the audio and video information [6, 13]. These results demonstrate a network can be used to

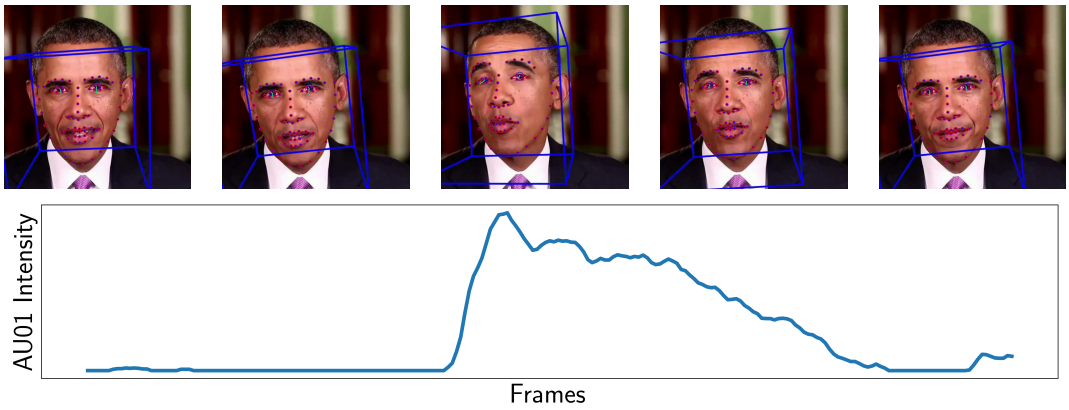

Figure 1: Shown in Figure 1 an example annotated of the annotated landmark positions, head rotation, and the intensity of the inner brow raiser AU01.

identify a speaker based on the visual and audio information during a speech. It is also shown that a CNN can be trained to find synchronization between audio and visual. signals [8].

4 Data

We believe that the facial movements can be used in the task of speaker identification. Given that face swapping techniques can forge facial features and speech but the facial expressions are not the ones of the real person but of the 'actor' that is wearing their 'mask', a classifier that relies on this signal will be resistant to these kind of fakes.

In an ideal scenario, we would like to test our hypothesis on a large dataset with thousands of different people [7]. However, given the computation and time limitations we started with the videos of 12 world and national leaders. Described below is our data collection and processing steps.

4.1 Data Download

For each of the 12 persons, we collected the videos of the person talking in a formal setting like a news interview, a weekly addresses, debate, public speech etc. The number of videos downloaded for each leader and the duration of the downloaded videos (in hours) are given in columns (a) and (b) of Table 1.

4.2 Data Pre-processing

The YouTube videos are then cleaned to extract the continuous video segments that satisfied the following criterions:

- the segment should be atleast 10-seconds long
- only one person i.e. the leader should be present in the entire video segment
- the person should be talking during the entire video segment
- the camera should be stationary during the segment
- the speaker should be mostly frontal face in the segment

The number and duration (in hours) of non-overlapping segments extracted for each leader is given in the columns (c) and (d) of the Table 1. Each of the segment was saved in H264 format at a frame rate of 30 and good compression quality of 20. For each leader, we selected random 0.5 hours of segments, which corresponds to the minimum duration

of videos for Michelle Obama. The segments were finally divided into a 0.8 : 0.1 : 0.1 train:validation:test split for all the experiments.

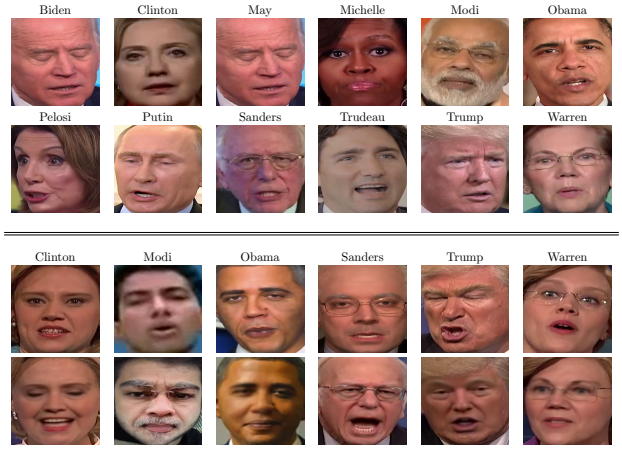

Figure 2: Shown in above are random sample frames from the real videos for each leader (top two rows), frames from the impersonator and face-swap deepfake videos for 6 leaders (third and fourth rows).

	(a)	(b)	(c)	(d)	(e)	(f)
Biden	26	5.39	197	52.03	0	0
Clinton	50	5.55	150	141.00	10.40	10.40
May	61	4.52	261	132.68	7.73	0
Michelle	17	0.54	43	31.86	0	0
Modi	9	5.61	84	82.17	16.44	6.80
Obama	304	18.93	1,064	824.40	10.23	10.23
Pelosi	24	4.39	206	93.41	0	0
Putin	31	13.33	436	165.85	0	0
Sanders	71	8.18	367	225.47	6.62	4.35
Trudeau	48	5.08	361	161.30	4.50	0
Trump	92	11.21	881	364.63	11.28	10.33
Warren	33	4.44	261	132.80	2.41	2.40

Table 1: Shown above are for each of the 12 leaders (a) the count of videos downloaded from YouTube (b) duration of videos in hours (c) count of segments after video pre-processing (d) total duration of segments in minutes (e) duration in minutes of impersonator videos (f) duration in minutes for face-swap videos.

4.3 Feature Extraction

We use OpenFace2.0 [4] open source library to extract the facial landmark, facial action units and head motion of each speaker. Using the OpenFace library, we tracked the following facial features:

- 3-D rotation of the head about the x-axis, y-axis, and z-axis
- 68×2 3-D facial landmarks. Each landmark was rotated to remove above 3-D rotation and points were scaled to have a unit distance between the center of the eyes. We discard the Z-coordinate and used only x and y coordinates thus giving a total of 68×2 features.
- 17 Action Units: The OpenFace2 toolkit provides the intensity and occurrence for 17 AUs: inner brow raiser (AU01), outer brow raiser (AU02), brow lowerer (AU04), upper lid raiser (AU05), cheek raiser (AU06), lid tightener (AU07), nose wrinkler (AU09), upper lip raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), lip stretcher (AU20), lip tightener (AU23), lip part (AU25), jaw drop (AU26), eye blink (AU45)

Shown in Figure 1 an example annotated of the annotated landmark positions, head rotation, and the intensity of the inner brow raiser AU01.

4.4 Deep Fakes

For fake videos, we used two types of datasets. As we exploit the facial expressions of the person when they talk, we used videos where the leaders speaking style is impersonated by someone else, for example the videos of Alec Baldwin impersonating the talking style of US president Donald Trump on Saturday Night Live. In addition to this, we used face-swap [1]

[1] github.com/shaoanlu/faceswap-GAN

videos for five leaders. [2]

The duration (in minutes) of impersonators and face-swap videos for each leader are given in columns (e) and (f) of Table 1. Notice there were no face-swap or imposter videos collected for some leaders. We believe, however, there are enough videos to validate our technique. Also, shown in Figure 2 are random sample frames from the real videos for each leader (top two rows), frames from the impersonator and face-swap deep-fake videos for 6 leaders (third and fourth rows).

5 Method

5.1 Raw pixel based deep fake detection

5.1.1 Data Preparation

The videos for the 12 speakers are varying length sequence. In order to process them we uniformly subsample the videos sequences to have T frames. Hence each sequence $S = \{s_1, ..., s_T\}$ of input data is video tensor of shape $\{T, H, W, C\}$ where H is the height of each frame and W is the width of each frame, and C is the number of channels.

5.1.2 Model

In order to deal with both the structural and temporal properties of these video sequences we use a hybrid architecture. This architecture consists of a convolutional neural network (f) to extract spatial features for each frame, and a LSTM architecture (l) to extract the temporal information from the video sequence. Each frame is first passed through a state of the art convolutional neural network according to the following equation:

$$x_t = f(s_t)$$

where s_t is the frame at timestamp t in the video sequence and $t \in 1, ..., T$.

[2] We used the face-swap videos generated for another unpublished research project by a team in University of Southern California.

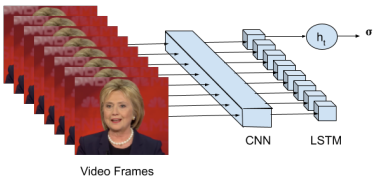

Figure 3: Raw pixel based deep fake detection. The video sequence is first fed into a state of the art CNN network. The extracted features are fed into a LSTM framework. The entire sequence representation is then extract as the last hidden vector of the LSTM at timestamp T. This representation is fed into a softmax layer for maximum likelihood estimation.

One we have obtained all the features extracted from the convolutional neural network, we have effectively converted out input video sequence $S = \{s_1, ..., s_T\}$ into a sequence of features extracted from the convolutional network $X = \{x_1, ..., x_T\}$. Now this input sequence is fed into the LSTM network as follows:

$$[h!]h_t = l(x_t, h_{t-1})$$

Thus, we extract h_T as the latent space representation of the entire video sequence. This representation is then fed into a n class classifier softmax layer that performs maximum likelihood cross entropy based estimate of the most likely speaker. The cross entropy loss in backpropogated through the entire architecture so that both the convolutional neural network and the LSTM learn the required representation for recognizing the speaker as part of the temporal sequence effectively.

	Clinton	Modi	Obama	Sanders	Trump	Warren
[17]	1.0	0.5	1.0	1.0	1.0	1.0

Table 2: Shown are the overall AUC for detecting face-swap videos of 6 leaders using the FaceForensics CNN based approach

5.2 Facial Landmark based deep fake detection

5.2.1 Data Preparation

The facial landmark features were extracted using the OpenFace 2.0 library for the 12 leaders. This data was balanced such that all the leaders are represented equally and then reshaped into Numpy arrays (Tensor) of size (Number of frames/Timesteps, Timesteps, Number of features). The number of frames for each leader were on average 47000, the Timesteps were fixed at 100 while the number of facial features as described above were 156.

5.2.2 Model

In order to deal with the temporal properties of these facial features i.e change in facial expression over a fixed time frame, the data was fed into a state of the art Long Short-Term Memory (LSTM) model. The architecture consists of 3 stacked layers of LSTM having 156, 100 and 32 cells respectively. Similar to the previous model, this representation is then fed into a n class classifier softmax layer that performs maximum likelihood cross entropy based estimate of the most likely speaker. The cross entropy loss in backpropogated through the entire architecture and is optimized using RMSProp so that the LSTM learns the required representation for recognizing the speaker as part of the temporal sequence effectively.

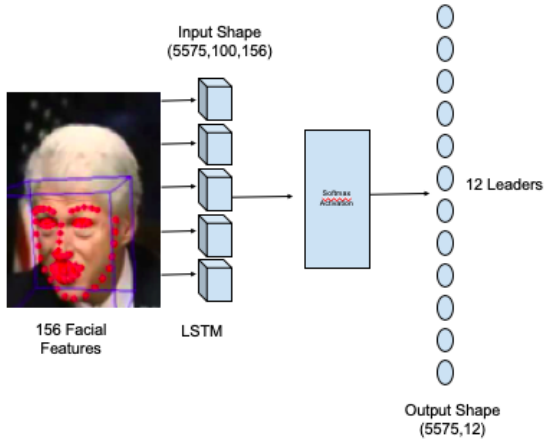

Figure 4: Facial landmark based deep fake detection. The extracted facial landmarks sequence is fed into a state of the art 3 layered LSTM network. This representation is fed into a softmax layer for maximum likelihood estimation.

6 Experiments and Results

6.1 Baseline

6.1.1 CNN based Deep fake detection

We compare our results with a CNN-based approach used in FaceForensics++ [17] in which the authors have trained multiple models to detect 3 types of face manipulations including face-swap deep fakes. We used the higher-performing models trained using XceptionNet [5] architecture with cropped faces as input. The performance of these models was tested on face-swap videos of 6 leaders. We tested the models[3] made available by the authors without any fine-tuning for our dataset. The per-frame CNN output for the real class was used to compute the accuracies in terms of AUC. The overall accuracies for each leader is given in Table 2 first row. The accuracies obtained for each leader using this approach is perfect expect in case of Modi. This is mainly because the face-swap forgery for Modi were downloaded from

[3] niessnerlab.org/projects/roessler2019faceforensicspp.html

YoutTube and were heavily compressed, which might have destroyed all the artifacts on which the CNN relies.

6.1.2 Face image based deep fake detection

As an additional benchmark we ran a speaker identification classifier that randomly sampled one frame from a video sequence and ran it through a classifier that trained the last layer of a VGGFace (ResNet) Network . The classifier obtained 97% accuracy on test data, 36% in imposter data, and 60% in fake data, showing that by simply running classification on static images the classifier is not robust to video forgery, given that impostor videos receive a boost in accuracy when the fake 'mask' is added (from 36% to 60%). Our goal is to build a classifier in which this increase in accuracy is less when fakes are introduced to 'mask' imposter videos.

Figure 5: Training and validation loss curve for raw pixel based method

6.2 Our Results

6.2.1 Raw pixel based deep fake detection

For our experimental setup we chose to represent the video as a sequence of T frames. In our experiments we chose $T = 32$, as that seemed to give the best results and run-time trade-off. Longer sequence marginally performed better, but were much slower. We also represented each frame as a $780x1280x3$ image. For the state of the art neural network we used a ResNet-50 [10] architecture that was pretrained on ImageNet. For effectively employing transfer learning we do backpropogate the cross entropy loss through the entire framework, that includes both the convolutions and the LSTM. However the learning rate for different parts of the feature extraction backbone are different. We used a learning rate of $1e{-}5$ for the convolutions and $1e-3$ for the LSTM. We also clipped the norm of the gradient to 10, which is standard for training LSTM architectures. Adam optimizer with momentum was used. A batch size of 32 videos was fed at once for stochastic mini batch gradient descent.

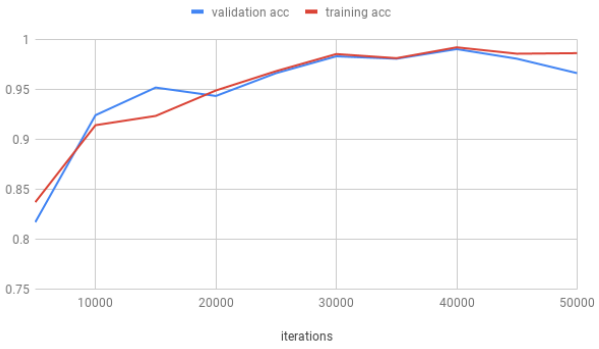

Figure 6: Training and validation accuracy curve for raw pixel based method

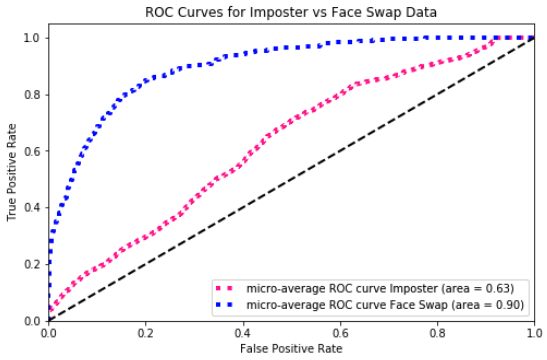

Figure 7: Comparision of Region Operating Characteristic for Imposter and Face Swap Data

6.2.2 Facial Landmark based deep fake detection

The 3 stacked LSTM architecture seemed to give the best results for the temporal sequences of the facial features. On the real test video data for the 12 leaders, the model was able to achieve an accuracy of 96% with an AUC of 1.0. While on the imposter and the face swap test data the accuracy reasonably decreased to 16.5% and 48.9% for 9 and 6 leaders respectively. The average AUC also decreased to 0.63 and 0.90 proving the hypothesis that the probabilities of correctly identifying the speakers decreased considerably. A batch size of 64 was selected for the this experiment and a total of 50 epochs were performed to get the best results.

7 Conclusion

In this project we showed that facial expressions can be used for speaker identification. The facial expressions based identification system is more robust towards face-swap deep fakes than conventional image based identification system. In future, we would like to the test robustness of our identification system on a larger dataset with several hundred people. Also, we would like to combine audio and visual modalities along with facial expressions to improve the robustness of the identification system.

References

[1] What are deepfakes why the future of porn is terrifying. https://www.highsnobiety.com/p/what-are-deepfakes-ai-porn/. (Accessed on 05/29/2018).

[2] https://www.faceapp.com/. (Accessed on 05/29/2018).

[3] Fakeapp . https://www.fakeapp.org/. (Accessed on 05/29/2018).

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pages 1086–1090, 2018.

[7] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[8] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[9] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 4217–4224, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-

nition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[12] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[13] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] Hai X Pham, Yuting Wang, and Vladimir Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *arXiv preprint arXiv:1803.07716*, 2018.

[15] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017.

[16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[17] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019.

[18] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[19] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[20] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018.

[21] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.