Protecting World Leaders Against Deep Fakes

Shruti Agarwal and Hany Farid
University of California, Berkeley
Berkeley CA, USA
{shruti_agarwal, hfarid}@berkeley.edu

Yuming Gu, Mingming He, Koki Nagano, and Hao Li
University of Southern California / USC Institute for Creative Technologies
Los Angeles CA, USA
{ygu, he}@ict.usc.edu, koki.nagano0219@gmail.com, hao@hao-li.com

Abstract

The creation of sophisticated fake videos has been largely relegated to Hollywood studios or 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 illegal activity leading to a constitutional crisis, a military leader saying something racially insensitive leading to civil unrest in an area of military activity, or a corporate titan claiming that their profits are weak leading to global stock manipulation. These so called deep fakes pose a significant threat to our democracy, national security, and society. To contend with this growing threat, we describe a forensic technique that models facial expressions and movements that typify an 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.

1. Introduction

While convincing manipulations of digital images and videos have been demonstrated for several decades through the use of visual effects, recent advances in deep learning have led to a dramatic increase in the realism of fake content and the accessibility in which it can be created [27, 14, 29, 6, 19, 21]. These so called AI-synthesized media (popularly referred to as deep fakes) fall into one of three categories: (1) face-swap, in which the face in a video is automatically replaced with another person's face. This type of technique has been used to insert famous actors into a variety of movie clips in which they never appeared [5], and used to create non-consensual pornography in which one person's likeliness in an original video is replaced with another person's likeliness [13]; (2) lip-sync, in which a source video is modified so that the mouth region is consistent with an arbitrary audio recording. For instance, the actor and director Jordan Peele produced a particularly compelling example of such media where a video of President Obama is altered to say things like "President Trump is a total and complete dip-****."; and (3) puppet-master, in which a target person is animated (head movements, eye movements, facial expressions) by a performer sitting in front of a camera and acting out what they want their puppet to say and do.

While there are certainly entertaining and non-nefarious applications of these methods, concerns have been raised about a possible weaponization of such technologies [7]. For example, the past few years have seen a troubling rise in serious consequences of misinformation from violence against our citizens to election tampering [22, 28, 26]. The addition of sophisticated and compelling fake videos may make misinformation campaigns even more dangerous.

There is a large body of literature on image and video forensics [11]. But, because AI-synthesized content is a relatively new phenomena, there is a paucity of forensic techniques for specifically detecting deep fakes. One such example is based on the clever observation that the individuals depicted in the first generation of face-swap deep fakes either didn't blink or didn't blink at the expected frequency [15]. This artifact was due to the fact that the data used to synthesize faces typically did not depict the person with their eyes closed. Somewhat predictably, shortly after this forensic technique was made public, the next generation of synthesis techniques incorporated blinking into their systems so that this technique is now less effective. This same team also developed a technique [31] for detecting

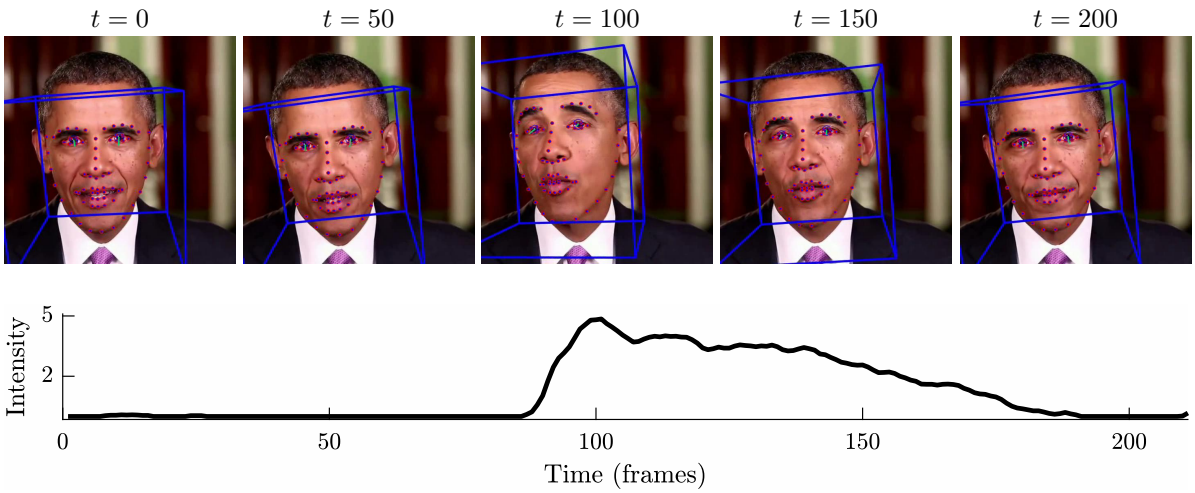

Figure 1. Shown above are five equally spaced frames from a 250-frame clip annotated with the results of OpenFace tracking. Shown below is the intensity of one action unit AU01 (eye brow lift) measured over this video clip.

face-swap deep fakes by leveraging differences in the estimated 3-D head pose as computed from features around the entire face and features in only the central (potentially swapped) facial region. While effective at detecting face-swaps, this approach is not effective at detecting lip-sync or puppet-master deep fakes.

Other forensic techniques exploit low-level pixel artifacts introduced during synthesis [16, 1, 20, 23, 32, 12, 24, 18]. Although these techniques detect a variety of fakes with relatively high accuracy, they suffer, like other pixel-based techniques, from simple laundering counter-measures which can easily destroy the measured artifact (e.g., additive noise, recompression, resizing).

We describe a forensic technique that is designed to detect deep fakes of an individual. We customize our forensic technique for specific individuals and, because of the risk to society and democratic elections, focus on world and national leaders and candidates for high office. Specifically, we first show that when individuals speak, they exhibit relatively distinct patterns of facial and head movements (see for example [9] as well as [30] in which upper-body movements was used for speaker identification). We also show that the creation of all three types of deep fakes tends to disrupt these patterns because the expressions are being controlled by an impersonator (face-swap and puppet-master) or the mouth is decoupled from the rest of the face (lip-sync). We exploit these regularities by building what we term soft bio-metric models of high-profile individuals and use this model to distinguish between real and fake videos. We show the efficacy of this approach on a large number of deep fakes of a range of U.S. politicians ranging from Hillary Clinton, Barack Obama, Bernie Sanders, Donald Trump, and Elizabeth Warren. This approach, unlike previous approaches, is resilient to laundering because it relies on relatively coarse measurements that are not easily destroyed, and is able to detect all three forms of deep fakes.

2. Methods

We hypothesize that as an individual speaks, they have distinct (but probably not unique) facial expressions and movements. Given a single video as input, we begin by tracking facial and head movements and then extracting the presence and strength of specific action units [10]. We then build a novelty detection model (one-class support vector machine (SVM) [25]) that distinguishes an individual from other individuals as well comedic impersonators and deep-fake impersonators.

2.1. Facial Tracking and Measurement

We use the open-source facial behavior analysis toolkit OpenFace2 [3, 2, 4] to extract facial and head movements in a video. This library provides 2-D and 3-D facial landmark positions, head pose, eye gaze, and facial action units for each frame in a given video. An example of the extracted measurements is shown in Figure 1.

The movements of facial muscles can be encoded using facial action units (AU) [10]. 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).

Our model incorporates 16 AUs – the eye blink AU was eliminated because it was found to not be sufficiently distinctive for our purposes. These 16 AUs are augmented

person of interest (POI)	video (hours)	segments (hours)	segment (count)	10-second clips (count)
real				
Hillary Clinton	5.56	2.37	150	22,059
Barack Obama	18.93	12.51	972	207,590
Bernie Sanders	8.18	4.14	405	63,624
Donald Trump	11.21	6.08	881	72,522
Elizabeth Warren	4.44	2.22	260	31,713
comedic impersonator				
Hillary Clinton	0.82	0.17	28	1,529
Barack Obama	0.70	0.17	21	2,308
Bernie Sanders	0.39	0.11	12	1,519
Donald Trump	0.53	0.19	24	2,616
Elizabeth Warren	0.11	0.04	10	264
face-swap deep fake				
Hillary Clinton	0.20	0.16	25	1,576
Barack Obama	0.20	11	12	1,691
Bernie Sanders	0.07	0.06	5	1,084
Donald Trump	0.22	0.19	24	2,460
Elizabeth Warren	0.04	0.04	10	277
lip-sync deep fake				
Barack Obama	0.99	0.99	111	13,176
puppet-master deep fake				
Barack Obama	0.19	0.20	20	2,516

Table 1. Total duration of downloaded videos and segments in which the POI is speaking, and the total number of segments and 10-second clips extracted from the segments.

Figure 2. Shown from top to bottom, are five example frames of a 10-second clip from original, lip-sync deep fake, comedic impersonator, face-swap deep fake, and puppet-master deep fake.

with the following four features: (1) head rotation about the x-axis (pitch); (2) head rotation about the z-axis (roll); (3) the 3-D horizontal distance between the corners of the mouth (mouth_h); and (4) the 3-D vertical distance between the lower and upper lip (mouth_v). The first pair of features capture general head motion (we don't consider the rotation around the y-axis (yaw) because of the differences when speaking directly to an individual as opposed to a large crowd). The second pair of these features captures mouth stretch (AU27) and lip suck (AU28), which are not captured the default 16 AUs.

We use the Pearson correlation to measure the linearity between these features in order to characterize an individual's motion signature. With a total of 20 facial/head features, we compute the Pearson correlation between all 20 of these features, yielding ${}_{20}C_2 = (20 \times 19)/2 = 190$ pairs of features across all 10-second overlapping video clips (see Section 2.2). Each 10-second video clip is therefore reduced to a feature vector of dimension 190 which, as described next, is then used to classify a video as real or fake.

2.2. Data set

We concentrate on the videos of persons of interest (POIs) talking in a formal setting, for example, weekly address, news interview, and public speech. All videos were manually downloaded from YouTube where the POI is primarily facing towards the camera. For each downloaded video, we manually extracted video *segments* that met the following requirements: (1) the segment is at least 10 seconds in length; (2) the POI is talking during the entire segment; (3) only one face – the POI – is visible in the segment; and (4) the camera is relatively stationary during the segment (a slow zoom was allowed). All of the segments were saved at 30 fps using an mp4-format at a relatively high-quality of 20. Each segment was then partitioned into overlapping 10-second clips (the clips were extracted by sliding a window across the segment five frames at a time). Shown in Table 1 are the duration of video and segment durations, and the number of clips extracted for five POIs.

We tested our approach with the following data sets: 1) 5.6 hours of video segments of 1,004 unique people, yielding 30,683 10-second clips, from the FaceForensics data set [23]; 2) comedic impersonators for each POI, (Table 1); 3) face-swap deep fakes, lip-sync deep fakes, and puppet-master deep fakes (Table 1). Shown in Figure 2 are five example frames from a 10-second clip of an original video, a lip-sync deep fake, a comedic impersonator, a face-swap deep fake, and puppet-master deep fake of Barack Obama.

2.2.1 Deep Fakes

Using videos of their comedic impersonators as a base, we generated face-swap deep fakes for each POI. To swap faces between each POI and their impersonator, a generative adversarial network (GAN) was trained based on the Deepfake architecture [1]. Each GAN was trained with approximately 5000 images per POI. The GAN then replaces the

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

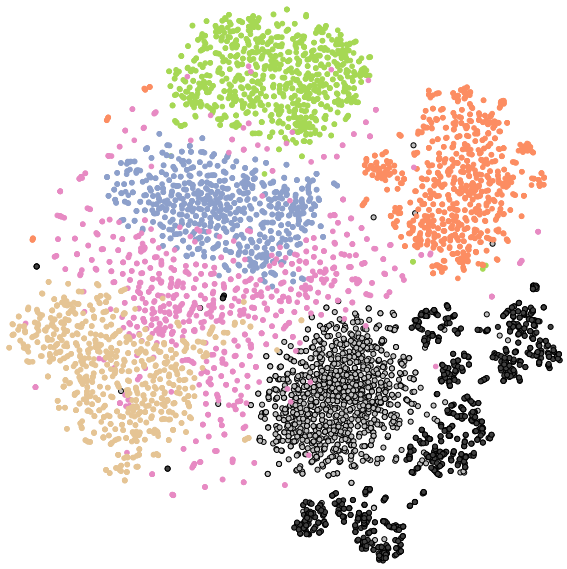

Figure 3. Shown is a 2-D visualization of the 190-D features for Hillary Clinton (brown), Barack Obama (light gray with a black border), Bernie Sanders (green), Donald Trump (orange), Elizabeth Warren (blue), random people [23] (pink), and lip-sync deep fake of Barack Obama (dark gray with a black border).

impersonator's face with the POI's face, matching the impersonator's expression and head pose on each video frame. We first detect the facial landmarks and facial bounding box using `dlib`. A central 82% of the bounding box is used to generate the POI's face. The generated face is then aligned with the original face using facial landmarks. The facial landmark contour is used to generate a mask for post-processing that includes alpha blending and color matching to improve the spatio-temporal consistency of the final face-swap video.

Using comedic impersonators of Barack Obama, we also generated puppet-master deep fakes for Obama. The photo-real avatar GAN (paGAN)[19] synthesizes photo-realistic faces from a single picture. This basic process generates videos of only a floating head on a static black background. In addition to creating these types of fakes, we modified this synthesis process by removing the face masks during training, allowing us to generate videos with intact backgrounds. The temporal consistency of these videos was improved by conditioning the network with multiple frames allowing the network to see in time[14]. This modified model was trained using only images of Barack Obama.

While both of these types of fakes are visually compelling, they do occasionally contain spatio-temporal glitches[2]. These glitches, however, are continually being reduced and it is our expectation that future versions will result in videos with little to no glitches.

[2] All synthesized videos will be made publicly available.

2.3. Modeling

Shown in Figure 3 is a 2-D t-SNE [17] visualization of the 190-dimensional features for Hillary Clinton, Barack Obama, Bernie Sanders, Donald Trump, Elizabeth Warren, random people [23], and lip-sync deep fake of Barack Obama. Notice that in this low-dimensional representation, the POIs are well separated from each other. This shows that the proposed correlations of action units and head movements can be used to discriminate between individuals. We also note that this visualization supports the decision to use a one-class support vector machine (SVM). In particular, were we to train a two-class SVM to distinguish Barack Obama (light gray) from random people (pink), then this classifier would almost entirely misclassify deep fakes (dark gray with black border).

In the ideal world, we would build a large data set of authentic videos of an individual and a large data set of fake videos of that same person. In practice, however, this is not practical because it requires a broad set of fake videos at a time when the techniques for creating fakes is rapidly evolving. As such, we train a novelty detection model (one-class SVM [25]) that requires only authentic videos of a POI. Acquiring this data is relatively easy for world and national leaders and candidates for high office who have a large presence on video-sharing sites like YouTube.

The SVM hyper-parameters γ and ν that control the Gaussian kernel width and outlier percentage are optimized using 10% of the video clips of random people taken from the FaceForensics original video data set [23]. Specifically, we performed a grid search over γ and ν and selected the hyper-parameters that yielded the highest discrimination between the POI and these random people. These hyper-parameters were tuned for each POI. The SVM is trained on the 190 features extracted from overlapping 10-second clips. During testing, the input to the SVM sign decision function is used as a classification score for a new 10-second clip [25] (a negative score corresponds to a fake video, a positive score corresponds to a real video, and the magnitude of the score corresponds to the distance from the decision boundary and can be used as a measure of confidence).

We next report the testing accuracy of our classifiers, where the 10-second video clips of POI are split into a 80:20 training:testing data sets in which we ensured that there was no overlap in content between training and testing.

3. Results

The performance of each POI-specific model is tested using the POI-specific comedic impersonators and deep fakes, Section 2.2. We report the testing accuracy as the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and the true positive rate (TPR) of correctly recognizing an original at fixed false positive rates

	random people	comedic impersonator	face-swap	lip-sync	puppet-master
190-features					
10-second clip					
TPR (1% FPR)	0.62	0.56	0.61	0.30	0.40
TPR (5% FPR)	0.79	0.75	0.81	0.49	0.85
TPR (10% FPR)	0.87	0.84	0.87	0.60	0.96
AUC	0.95	0.94	0.95	0.83	0.97
segment					
TPR (1% FPR)	0.78	0.97	0.96	0.70	0.93
TPR (5% FPR)	0.85	0.98	0.96	0.76	0.93
TPR (10% FPR)	0.99	0.98	0.97	0.88	1.00
AUC	0.98	0.99	0.99	0.93	1.00
29-features					
10-second clip					
AUC	**0.98**	0.94	0.93	**0.95**	**0.98**
segment					
AUC	**1.00**	0.98	0.96	**0.99**	1.00

Table 2. Shown are the overall accuracies for Barack Obama reported as the area under the curve (AUC) and the true-positive rate (TPR) for three different false positive rates (FPR). The top half corresponds to the accuracy for 10-second video clips and full video segments using the complete set of 190 features. The lower-half corresponds to using only 29 features (see Section 3.1.1).

(FPR) of 1%, 5%, and 10%. These accuracies are reported for both the 10-second clips and the full-video segments. A video segment is classified based on the median SVM score of all overlapping 10-second clips. We first present a detailed analysis of the original and fake Barack Obama videos, followed by an analysis of the other POIs.

3.1. Barack Obama

Shown in top half of Table 2 are the accuracies for classifying videos of Barack Obama based on 190 features. The first four rows correspond to the accuracy for 10-second clips and the next four rows correspond to the accuracy for full-video segments. The average AUC for 10-second clips and full segments is 0.93 and 0.98. The lowest clip and segment AUC for lip-sync fakes, at 0.83 and 0.93, is likely because, as compared to the other fakes, these fakes only manipulate the mouth region. As a result, many of the facial expressions and movements are preserved in these fakes. As shown next, however, the accuracy for lip-sync fakes can be improved with a simple feature-pruning technique.

To select the best features for classification, a 190 models were iteratively trained with between 1 and 190 features. Specifically, on the first iteration, 190 models were trained using only a single feature. The feature that gave the best overall training accuracy was selected. On the second iteration, 189 models were trained using two features, the first of which was determined on the first iteration. The second feature that gave the best overall training accuracy was selected. This entire process was repeated 190 times. Shown in Figure 4 is the testing accuracy as a function of the number of features for the first 29 iterations of this process (the

	random people	comedic impersonator	face-swap
Hillary Clinton			
TPR (1% FPR)	0.31	0.22	0.48
TPR (5% FPR)	0.60	0.55	0.77
TPR (10% FPR)	0.75	0.76	0.89
AUC	0.91	0.93	0.95
Bernie Sanders			
TPR (1% FPR)	0.78	0.48	0.58
TPR (5% FPR)	0.92	0.70	0.84
TPR (10% FPR)	0.95	0.84	0.92
AUC	0.98	0.94	0.96
Donald Trump			
TPR (1% FPR)	0.30	0.39	0.31
TPR (5% FPR)	0.65	0.72	0.60
TPR (10% FPR)	0.77	0.83	0.74
AUC	0.92	0.94	0.90
Elizabeth Warren			
TPR (1% FPR)	0.75	0.97	0.86
TPR (5% FPR)	0.91	0.98	0.91
TPR (10% FPR)	0.95	0.99	0.92
AUC	0.98	1.00	0.98

Table 3. Shown are the overall accuracies for 10-second video clips of Hillary Clinton, Bernie Sanders, Donald Trump, and Elizabeth Warren. The accuracies are reported as the area under the curve (AUC) and the true-positive rate (TPR) for three different false positive rates (FPR).

training accuracy reached a maximum at 29 features). This iterative training was performed on a random 10% of the 10-second videos clips of random people, comedic impersonators and all three types of deep fakes.

With only 13 features the AUC nearly plateaus at an average of 0.95. Not shown in this figure is the fact that accuracy starts to slowly reduce after including 30 features. The top five distinguishing features are the correlation between: (1) upper-lip raiser (AU10) and 3-D horizontal distance between the corners of the mouth (mouth_h); (2) lip-corner depressor (AU15) and mouth_h; (3) head rotation about the x-axis (pitch) and mouth_v; (4) dimpler (AU14) and pitch; and (5) lip-corer depressor (AU15) and lips part (AU25). Interestingly, these top-five correlations have at least one component that corresponds to the mouth. We hypothesize that these features are most important because of the nature of lip-sync fakes that only modify the mouth region, and the face-swap, puppet-master, and comedic impersonators are simply not able to capture the subtle mouth movements.

Shown in the bottom half of Table 2 is a comparison of the accuracy for the full 190 features and the 29 features enumerated in Figure 4. The bold-face values in this table denote those accuracies that are improved relative to the full 190 feature set. We next test the robustness of these 29

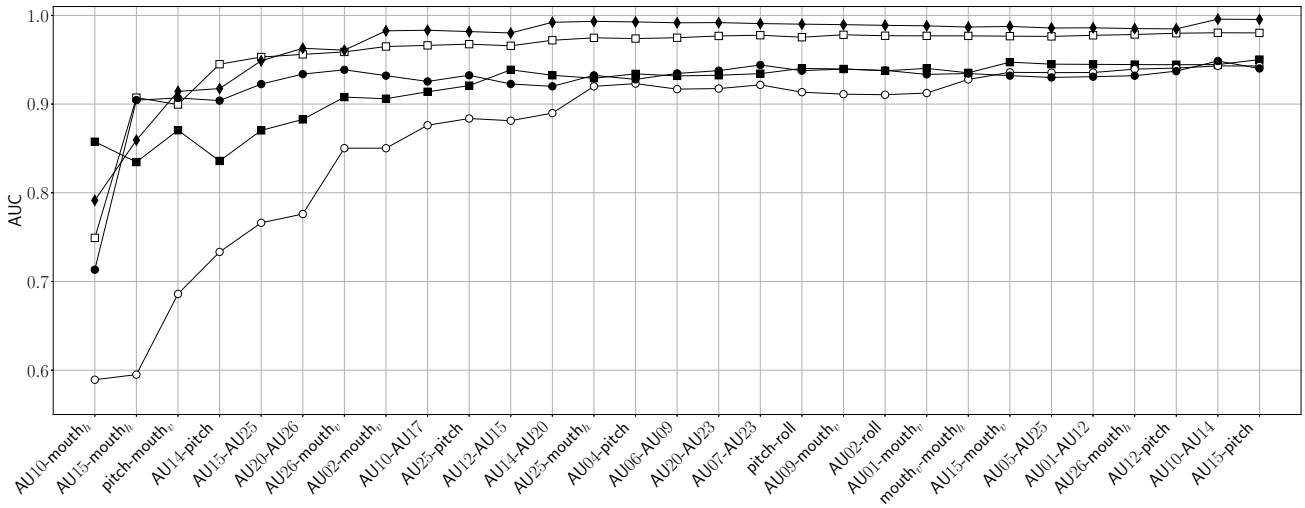

Figure 4. The accuracy (as AUC) for comedic impersonator (black square), random people (white square), lip-sync deep fake (black circle), face-swap deep fake (white circle), and puppet-master (black diamond) for a classifier trained on between one and 29 features as enumerated on the horizontal axis. In particular, the accuracy for AU10-mouth$_h$ corresponds to an SVM trained on only this feature. The accuracy for AU15-mouth$_h$ corresponds to an SVM trained on two features, AU10-mouth$_h$ and AU15-mouth$_h$. Overall accuracy plateaus at approximately 13 features.

features to a simple laundering attack, to the length of the extracted video clip, and to the speaking context.

3.1.1 Robustness

As mentioned earlier, many forensic techniques fail in the face of simple attacks like recompression, and so we tested the robustness of our approach to this type of laundering. Each original and fake video segments were initially saved at an H.264 quantization quality of 20. Each segment was then recompressed at a lower quality of 40. The AUCs for differentiating 10-second clips of Barack Obama from random people, comedic impersonators, face-swap, lip-sync, and puppet-master deep fakes after this laundering are: 0.97, 0.93, 0.93, 0.92, and 0.96, virtually unchanged from the higher-quality videos (see Table 2). As expected, because our analysis does not rely on pixel-level artifacts, our technique is robust to a simple laundering attack.

In order to determine the robustness to clip length, we retrained four new models using clips of length 2, 5, 15, and 20 seconds. The average AUCs across all videos are 0.80, 0.91, 0.97, and 0.98, as compared to an AUC of 0.96 for a clip-length of 10 seconds. As expected, accuracy drops for shorter clips, but is largely unaffected by clip lengths between 10 and 20 seconds.

The talking style and facial behavior of a person can vary with the context in which the person is talking. Facial behavior while delivering a prepared speech, for instance, can differ significantly as compared to answering a stressful question during a live interview. In two followup experiments, we test the robustness of our Obama model against a variety of contexts different than the weekly addresses used for training.

In the first experiment, we collected videos where, like weekly addresses, Barack Obama was talking to a camera. These videos, however, spanned a variety of contexts ranging from an announcement about Osama Bin Laden's death to a presidential debate video, and a promotional video. We collected a total of 1.5 hours of such videos which yielded 91 video segments of 1.3 hours duration and 21,152 overlapping 10-second clips. The average accuracy in terms of AUC to distinguish these videos from comedic impersonators, random people, lip-sync fake, face-swap fake and puppet-master fake is 0.91 for 10-second clips and 0.98 for the full segments, as compared to the previous accuracy of 0.96 and 0.99. Despite the differences in context, our model seems to generalize reasonably well to these new contexts.

In the second experiment, we collected another round of videos of Obama in even more significantly different contexts ranging from an interview in which he was looking at the interviewer and not the camera to a live interview in which he paused significantly more during his answer and tended to look downward contemplatively. We collected a total of 4.1 hours of videos which yielded 140 video segments of 1.5 hours duration and 19,855 overlapping 10-second clips. The average AUC dropped significantly to 0.61 for 10-second clips and 0.66 for segments. In this case, the context of the videos was significantly different so that our original model did not capture the necessary features. However, on re-training the Obama model on the original

data set and these interview-style videos, the AUC increased to 0.82 and 0.87 for the 10-second clips and segments. Despite the improvement, we see that the accuracy is not as high as before suggesting that we may have to train POI and context specific models and/or expand the current features with more stable and POI-specific characteristics.

3.1.2 Comparison to FaceForensics++

We compare our technique with the CNN-based approach used in FaceForensics++ [24] in which multiple models were trained to detect three types of face manipulations including face-swap deep fakes. For our evaluation, we used the higher-performing models trained using XceptionNet [8] architecture with cropped faces as input. The performance of these models was tested on the real, face-swap deep fake, lip-sync deep fake, and puppet-master deep fake Obama videos saved at high and low qualities (the comedic impersonator and random people data sets were not used as they are not synthesized content). 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 (AUC). The overall accuracies for detecting frames of face-swaps, puppet-master and lip-sync deep fakes at quality $20/40$ are $0.84/0.71$, $0.53/0.76$, and $0.50/0.50$, as compared to our average AUC of $0.96/0.94$. Even though FaceForensics++ works reasonably well on face-swap deep fakes, it fails to generalize to lip-sync deep fakes which it has not seen during the training process.

3.2. Other Leaders/Candidates

In this section, we analyse the performance of SVM models trained for Hillary Clinton, Bernie Sanders, Donald Trump, and Elizabeth Warren. Shown in Figure 5 are sample frames from videos collected for these four leaders (see Table 1). For each POI, a model was trained using the full set of 190 features. Shown in Table 3 are the accuracies for classifying 10-second clips of Hillary Clinton, Bernie Sanders, Donald Trump, and Elizabeth Warren. The average AUC for these POIs are 0.93, 0.96, 0.92, and 0.98.

4. Discussion

We described a forensic approach that exploits distinct and consistent facial expressions to detect deep fakes. We showed that the correlations between facial expressions and head movements can be used to distinguish a person from other people as well as deep-fake videos of them. The robustness of this technique was tested against compression, video clip length and the context in which the person is talking. In contrast to existing pixel-based detection methods, our technique is robust against compression. We found, however, that the applicability of our approach is vulnerable to different contexts in which the person is speaking (e.g., formal prepared remarks looking directly into the camera versus a live interview looking off-camera). We propose to contend with this limitation in one of two ways. Simply collect a larger and more diverse set of videos in a wide range of contexts, or build POI- and context-specific models. In addition to this context effect, we find that when the POI is consistently looking away from the camera, the reliability of the action units may be significantly compromised. To address these limitations, we propose to augment our models with a linguistic analysis that captures correlations between what is being said and how it is being said.

Figure 5. Shown are sample frames for (a) real; (b) comedic impersonator; and (c) face-swap for four POIs.

Acknowledgment

This research funded by the U.S. Government, Google, Microsoft, and the Defense Advanced Research Projects Agency (DARPA FA8750-16-C-0166). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Hao Li is affiliated with USC, USC/ICT, and Pinscreen. This project was not funded by nor conducted at Pinscreen. Koki Nagano who is affiliated with Pinscreen has worked on this project as part of his employment at USC/ICT. We thank Ira Kemelmacher-Shlizerman and Supasorn

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

Suwajanakorn for the lip-sync deep fake examples.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7. IEEE, 2018. 2

[2] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, volume 6, pages 1–6. IEEE, 2015. 2

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–10. IEEE, 2016. 2

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66. IEEE, 2018. 2

[5] Jennifer Finney Boylan. Will deep-fake technology destroy democracy?, 2018. 1

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 1

[7] Robert Chesney and Danielle Keats Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. Technical Report Public Law Research Paper No. 692, University of Texas Law, 2018. 1

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 7

[9] Jeffrey F Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *4th IEEE International Conference on Multimodal Interfaces*, page 491. IEEE Computer Society, 2002. 2

[10] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976. 2

[11] H. Farid. *Photo Forensics*. MIT Press, 2016. 1

[12] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2018. 2

[13] Drew Harwell. Scarlett Johansson on fake AI-generated sex videos: nothing can stop someone from cutting and pasting my image, 2018. 1

[14] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics*, 2018. 1, 4

[15] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *IEEE Workshop on Information Forensics and Security*, Hong Kong, 2018. 1

[16] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 2

[17] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 4

[18] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops*, pages 83–92. IEEE, 2019. 2

[19] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. paGAN: real-time avatars using dynamic textures. In *SIGGRAPH Asia Technical Papers*, page 258. ACM, 2018. 1, 4

[20] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. *arXiv preprint arXiv:1810.11215*, 2018. 2

[21] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. *CoRR*, abs/1807.09251, 2018. 1

[22] Kevin Roose and Paul Mozur. Zuckerberg Was Called Out Over Myanmar Violence. Here's His Apology, 2018. 1

[23] 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. 2, 3, 4

[24] 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. 2, 7

[25] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471. 2, 4

[26] Scott Shane and Mark Mazzetti. Inside a 3-Year Russian Campaign to Influence U.S. Voters, 2018. 1

[27] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):95, 2017. 1

[28] Amanda Taub and Max Fisher. Where Countries Are Tinderboxes and Facebook Is a Match, 2018. 1

[29] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics*, 2018. 1

[30] George Williams, Graham Taylor, Kirill Smolskiy, and Chris Bregler. Body motion analysis for multi-modal identity verification. In *20th International Conference on Pattern Recognition*, pages 2198–2201. IEEE, 2010. 2

[31] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Bristol, United Kingdom, 2019. 1

[32] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to GANs: Analyzing fingerprints in generated images. *CoRR*, abs/1811.08180, 2018. 2