Detecting Synthetic Faces by Understanding Real Faces

Shruti Agarwal

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-171
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-171.html

August 4, 2021

Detecting Synthetic Faces by Understanding Real Faces

By

Shruti Agarwal

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Hany Farid, Chair
Professor Emily Cooper
Professor Alexei A. Efros
Professor Ren Ng

Summer 2021

Detecting Synthetic Faces by Understanding Real Faces

Abstract

Detecting Synthetic Faces by Understanding Real Faces

by

Shruti Agarwal

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Hany Farid, Chair

The creation of sophisticated fake videos has been largely relegated to Hollywood studios or state actors. Recent advances in deep learning, however, have democratized the creation of sophisticated and compelling fake images, videos, and audios. This synthetically-generated media – so-called deep fakes – continue to capture the imagination of the computer-graphics and computer-vision communities. At the same time, the easy access to technology that can create deep fakes of anybody saying anything continues to be of concern because of its power to disrupt democratic elections, commit small to large-scale fraud, fuel dis- and mis-information campaigns, and create non-consensual pornography.

To contend with this growing threat, I describe a diverse set of techniques to detect state-of-the-art deep-fake videos. One set of these techniques are identity-specific, exploiting soft- and hard-biometric cues like dynamic facial motion and static facial appearance. Another set of these techniques are identity-independent, exploiting the dynamics of lip and ear motion.

Given the large-scale presence of deep fakes and the poor scalability of forensic techniques on the internet, the reliance on human perception to detect deep fakes is inevitable. Therefore, I also present several perceptual studies to understand the human visual system's ability to detect synthetic faces.

To my loving mother and sisters for their unwavering belief in me.

Contents

List of Figures

List of Tables

Acknowledgments

It is an overwhelming mix of emotions to think back to the six years of my Ph.D. During this long journey, I have spent equal time at two of the most prestigious academic institutions in America, Dartmouth College in Hanover and the University of California in Berkeley. I am thankful to the Computer Science departments of both institutions to have given me this opportunity. I cannot think of a better place to pursue my degree.

I visited Dartmouth College (and America) for the first time in 2015 for an open house. I remember being immediately in love with Hanover's tranquility and Hany's charisma (and, of course, his research interests). I am fortunate that Hany decided to take me as his Ph.D. student, given that I had minimal research experience. I thank Hany for believing in my abilities and research ideas even when I did not believe in them. I thank him for giving me the freedom (and funding) to work on my ideas and for always going the extra mile to ensure my well-being. I deeply value the many life lessons I have directly or indirectly learned from him. Thanks for introducing me to this quote by Albert Einstein: "Insanity is doing the same thing over and over again and expecting different results.". I am glad to have Hany as my mentor, supporter, advocator, and many times as my life coach.

I want to thank Emily Cooper for being very supportive throughout my Ph.D. I thank her for waiting patiently when my meetings with Hany kept getting extended and got her late for dinner. She has always encouraged me, appreciated my work, and guided me to exciting career possibilities. I thank Alyosha Efros and Ren Ng for agreeing to be a part of my committee. Alyosha's research has inspired me to design better forensics algorithms and think about the applicability of my research in the real world. I am thankful to Ren for his valuable insights into my research during the qualification exam.

During my Ph.D., I have been fortunate to meet several fantastic peers. I have thoroughly enjoyed working with the master and postdoc students in our lab. I want to thank Sophie Nightingale for the long conversations we had during coffee breaks. Moving to a new place seemed to pose similar challenges for both of us. It was helpful to have those discussions with her. I want to thank all of my research colleagues;

most of them are also my dear friends. I thank Maksim Bolonkin, Benedikt Lorch, Wei Fan, Tarek El-Gaaly, Ser-Nam Lim, and Ohad Fried. It was an honor to work with them, and I could not have done it without them. The discussions with them have made me a better researcher, inspired innovative ideas, and helped me improve the quality of my work.

I want to extend my deepest gratitude to all my dearest friends. Their love kept me going through this journey. I am glad to have been blessed with many compassionate and fun people who are no less than a family to me. I want to thank Suman Bera for being a mentor to me throughout the Ph.D. He has always supported me and pushed me to reach my highest potential – sometimes even pushing too hard. Thanks, Suman, for motivating me patiently during the many low phases. I would like to thank Dhananjay Beri (aka Beri) for introducing me to my love for nature. I thank him for organizing all the adventurous outings, some of which are my most valuable memories. I want to thank all my girlfriends for being an inspiration and also an emotional anchor. Thanks, Sisira Kanhirathingal, Kalika Srivastava, Vasanta Kommineni, Nithya Ramgopal, and Jenny, for always getting it like no one else. I would not have made it through without all of them. I would like to thank David Hermele, Vivant Sakore, and Ashish Sur for their loving heart and for being there for me during the extreme isolation of the pandemic.

Finally, I would like to dedicate the successful completion of my Ph.D. to my family. My mother, Dr. Asha Agarwal, whose life inspires me to be strong and always strive to be of service to the world. My sisters, Ekta Agarwal and Amita Agarwal, who always supported my dreams and encouraged me to excel. My brother Amit Bansal whose composure, humor, and warmth kept me grounded. My two adorable nephews Aryav and Adwik, whose laughter and positive energy always filled my heart with love. Radhasoami!

Chapter 1

Introduction

A 2020 study showed that more than half (53%) of American adults get their "news" from social media [1]. In just a three month period of this same year, Facebook removed 7 million false coronavirus-related information [2], and reported 1.8 billion fake-news engagements around U.S. elections [3]. At the same time, 56% of Facebook users said they cannot recognize fake news if it aligns with their beliefs [4]. These numbers suggest an alarming presence of misinformation in our lives and its potential harms. Though the adverse effects of misinformation have plagued our society for decades, technology has exacerbated the problem.

Today, the internet has democratized access to publishing and consuming information. Coupled with the speed and widespread usage of social media, any single piece of information – regardless of its credibility – has the potential to instantaneously reach billions of people. To deal with the large volume of information and to maximize user engagement, social media platforms rely on recommendation algorithms to curate the content which is seen by their users. These algorithms, unfortunately, driven mainly to increase profits, promote sensational fake content over a factual news article. For example, researchers at Facebook reported that unchecked recommendation algorithms will promote "more and more divisive content in an effort to gain user attention and increase time on the platform." [5] The promotion of more divisive content results in online echo chambers where people are only served information that is emotionally engaging and aligns closely to their worldview, regardless of whether it is true or not. In this discordant online world where truth is already losing its meaning, recent advances have added a new threat: the ability to create hyper-realistic manipulated images, video, and audio – so-called deep fakes.

Advances in artificial intelligence and machine learning have made it significantly easier to create sophisticated and compelling fake media. 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 or a military leader saying something racially insensitive leading to civil unrest in an area of military activity. While people are susceptible to falling prey to this fake content, on the other hand the existence of such technology gives people probable deniability (liar's dividend) [6]. The ease to create hyper-realistic media puts doubt on the authenticity of every image, video, or audio. Depending on their prior beliefs, people can now chose to dismiss an image or video as fake regardless of its truthfulness, thus, further increasing the divisiveness in our society. In this dissertation, I discuss one such type of AI-synthesized content, popularly called deep fake, that is a dangerous addition to the misinformation phenomenon and poses a significant threat to our democracy, national security, and society.

The term “deep fake" emerged in 2017 when a Reddit user named “deepfakes", along with other Reddit users, began using advances in machine learning to digitally insert celebrity faces into sexually explicit material. Over the intervening few years, the term is now used to describe any AI-synthesized images, audios, and videos (typically of people) generated from a range of different techniques. There are many commercially available apps such as DeepFake, ZAO3, and FaceApp4 to quickly and easily create compelling manipulations. This approach has been popularized by adding the actor Nicholas Cage's face into movies in which he never appeared, including his highly entertaining appearance in The Sound of Music [1]. However, such a technique has also been used to create non-consensual pornography, to instigate small- and large-scale fraud, and to produce dis-information. Therefore, the general consensus today is, while synthetic media can be entertaining, it can also easily be weaponized and may lead to a general lack of trust in what we see and hear online.

To deal with this growing threat, I present a set of techniques to detect state-of-the-art deep-fake videos of people. To recognize manipulated faces, we have developed models to understand real faces and show that these models are disrupted by the nature of how deep-fake videos are created. I also present perceptual studies to understand the ability of humans to detect deep-fake faces. Described next are different types of deep fakes, the existing deep-fake creation and detection techniques, and a brief overview of this dissertation.

1.1 Related Work

We begin by describing the most relevant work in both the creation and detection of deep fakes.

[1] https://youtu.be/MHkZEpfUnAA

1.1.1 Creation

One of the earliest examples of what we now generically call deep-fake videos dates back to 1997 [7]. In this seminal work, videos of a person mouthing words she did not speak are synthesized by reordering the mouth region in a training video to match specific phonemes in a new audio track (today, w e would call this a lip-sync deep fake). The intervening two decades have seen tremendous advances in computer-graphics and -vision based rendering, synthesis, and understanding, thanks to i) the accessibility to large-scale public data and ii) the evolution of deep learning techniques that eliminate many manual editing steps such as Autoencoders and Generative Adversarial Networks (GAN).

In 2014 Goodfellow et al. proposed a GAN architecture that contains two networks, a generator and a discriminator [8]. The generator synthesizes images; the discriminator tries to distinguish synthetic images from real ones. Both networks are trained together competitively, at the end of which the generator can synthesize plausible images achieving impressive results in a variety of image synthesis and editing applications. In image generation, GAN can be applied in the following scenarios: 1) taking noise as input to produce an image [9–11]; 2) taking an image from one semantic category (such as a horse) as input to produce an image of another semantic category (such as zebra) [12–14]; 3) taking a sketch or a pixel-level semantic map as input to produce a realistic image that is constrained by the layout of the sketch [12, 15]. Compared with other generative models such as variational autoencoders [16], images generated by GANs are usually more realistic.

Using this underlying GAN-based image synthesis, several deep-fake video generation techniques have been proposed. These deep-fake videos can be broadly categorized into (1) face-swap, in which the face in a video is automatically replaced with another person's face [17–20]. This type of technique has been used to insert famous actors into a variety of movie clips in which they never appeared [21], and used to create non-consensual pornography in which one person's likeliness in an original video is replaced with another person's likeliness [22]; (2) lip-sync, in which a source video is modified so that the mouth region is consistent with an arbitrary audio recording or textual input [23–26]. 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 [27–30], 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.

1.1.2 Detection

There is a significant literature in the general area of digital forensics [31]. Here we focus only on techniques for detecting the types of deep-fake images and videos described above. In response to deep fakes, several authentication techniques have emerged that can be roughly categorized into one of three categories:

1. **Forensic Analysis**: based on the assumption that manipulation or synthesis will leave behind some statistical, geometric, or physical artifact, this class of approaches analyzes content for explicit traces of manipulation or synthesis [31]. The benefit of this approach is it can be applied to a broad class of content and requires little to no prior assumptions. The drawback is, to date, most forensic techniques cannot operate at a speed or accuracy for deployment at an internet-scale of billions of daily uploads.

2. **Digital Signatures**: this class of approaches tackles the authentication from a different direction, focusing on authenticating content at the point of recording [32]. In either software or hardware, a specialized camera app extracts date/time, geo-location, and pixel data at the point of recording, and hashes and cryptographically signs this data. The resulting digital signature can be used downstream to verify that the content has not been altered from the time of recording, and localize where and when the content was recorded. The benefit of this approach is it can, at an internet scale, verify recorded content quickly and accurately. The drawback of this approach is it requires a specialized camera app and is unable to verify content not recorded through such an app.

3. **Digital Watermarks**: this class of approaches incorporates directly into a synthesis pipeline a digital watermark that can be used downstream to identify deep-fake content [33, 34]. The benefit of this approach is it can quickly and accurately identify deep-fake content. The drawback is it requires a specific infrastructure incorporated into all synthesis pipelines and is vulnerable to attacks designed to remove (or add) watermarks [35].

In this dissertation, I describe detection techniques falling into category 1. Depending upon the type of features used for detection, these forensic techniques can be further divided into:

- **Low-level approaches**: detect pixel-level artifacts introduced by the synthesis process. Some of these techniques detect generic artifacts [36–39], while others detect explicit artifacts that result from, for example, image warping [40], image blending [41] and inconsistencies between the image and metadata [42]. The benefit of low-level approaches is that they can detect artifacts that may not be visibly apparent. The drawback is that they can be sensitive to unintentional laundering (e.g., transcoding or resizing) or intentional adversarial attacks (e.g., [43]) and extrapolation to novel datasets. In contrast, the high-level approaches described next tend to be more resilient to these types of laundering and attacks and more likely to generalize to novel datasets.

- **High-level approaches**: focus on more semantically meaningful features. For example, [44] recognized that the creation of face-swap deep fakes introduces inconsistencies in the head pose as estimated from the central, swapped portion of the face and the surrounding, original head. These inconsistencies leverage 3-D geometry which is currently difficult for synthesis techniques to correct. Because training data sets often do not depict people with their eyes closed, it was observed that early face-swap deep fakes contained an abnormally low number of eye blinks [45]. More recent deep fakes, however, seem to have corrected this problem. A related technique [46] exploits spatial and temporal physiological signals that appear not to be consistent across real videos and disrupted in face-swap deep fakes. We believe that detection techniques based on these types of semantic and temporal dynamics is essential to staying slightly ahead of the cat-and-mouse game of synthesis and detection.

1.2 Overview

The three types of deep-fake videos – face-swap, lip-sync, and puppet-master – have one thing in common: they tend to disrupt the nature of how an individual speaks. In face-swap and puppet-master deep fakes, for instance, the face is of one person but the expressions are being controlled by an impersonator. Similarly, in lip-sync deep fakes, the mouth is decoupled from the rest of the face and synchronized with a new audio. In chapters 2 and 3, we exploit these discrepancies in the facial behavior of individuals to detect their deep fakes. In each approach, we first model facial-behavioral patterns that are distinct (but not necessarily unique) to an individual during a speech. These patterns are then used to distinguish between real and fake videos of that individual.

In [47], it is shown that the facial expressions convey unique information about a person's identity and in [48] the authors used upper-body movements for speaker identification. Using these observations, in **Chapter 2**, we use the facial expressions and head movements of individuals during speech to build their soft-biometric models.

For this, 20 facial movements are measured and simple hand-crafted correlation-based features are computed to build mannerism models for five U.S. politicians ranging from Hillary Clinton to Barack Obama, Bernie Sanders, Donald Trump, and Elizabeth Warren. The deep-fake videos are then shown to be inconsistent with the learned mannerisms of an individual. Even though this technique performs well on all three types of deep-fakes, there are some limitations: 1) building a new model for every individual can be a significant effort; 2) it is unlikely that the hand-picked 20 facial features are optimal to distinguish between a large number of identities; 3) the measurement of facial movements is reliable only if the person is talking with their face towards the camera; 4) the person-specific behavioral patterns changed with the context in which the person is speaking (e.g., formal prepared speech versus an informal interview).

In **Chapter 3**, we aim to resolve some of the above limitations. For this a convolutional neural network is used to learn the behavioral biometric of several thousand individuals. The features from this network are paired with static facial features to determine if a person's facial identity is consistent with their behavioral identity. This technique is specifically designed to detect face-swap deep fakes where the facial appearance of one person (source) is mapped to the facial movements of another person (target). As a result the facial appearance matches the source identity whereas the facial behavior matches the target identity. We use this discrepancy to detect face-swap deep fakes from multiple large-scale datasets. In contrast to the previous approach where the models were built for only five individuals, this technique is shown to detect deep-fake videos of thousands of identities.

While the facial identity in a face-swap deep fake may accurately depict the co-opted identity, the ears belong to the original identity. While the mouth in a lip-sync deep fake may be well synchronized with the audio, the dynamics of the ear motion will be de-coupled from the mouth and jaw motion. Therefore, statically, the shape of the human ear can be used as a biometric cue to detect person specific deep fakes. Dynamically, movement of the mandible (lower jaw) in correlation with the aural movement can be used to detect lip-sync deep fakes. In **Chapter 4**, we exploit this observation to build static-appearance and dynamic-behavioral models of human ears to detect deep fakes.

The above approaches capture soft-biometric cues that current deep-fake synthesis techniques are not (yet) able to synthesize well. These techniques, however, are generally most effective when confronted with face-swap and puppet-master deep fakes in which the facial behavior of a person changes significantly, but are less effective for lip-sync deep fakes where most of the facial behavior remains of the actual person. Also, such techniques require a large number of videos to train person-specific models. To this end, in **Chapter 5** we developed a person-independent forensic technique for detecting lip-sync deep fakes. This approach exploits the fact that in lip-sync deep fakes, the dynamics of the mouth shape – so-called visemes – are occasionally inconsistent with a spoken phoneme. Try, for example, to say a word that begins with `M`, `B`, or `P` – mother, brother, parent – and you will notice that your lips have to completely close. We observe that this phoneme to viseme mapping is occasionally violated in lip-sync deep fakes, even if it is not immediately apparent upon casual inspection. We leveraged these inconsistencies to detect audio-based and text-based lip-sync deep fakes.

Recently, Twitter found a network of more than $3,000$ accounts using GAN-generated images, 900 of which used synthetic female faces [49]. However, the above detection methods, like other forensic techniques, suffer from poor scalability to internet-scale. Therefore, often we still rely on human observers to identify a deep fake that appears online. Therefore, in **Chapter 6**, we shift our attention from detection to perception. We design experiments to analyze human's ability to detect synthetic faces generated with two types of techniques: StyleGAN2 and face-morphing. In each experiment, human participants are asked to detect a synthetic or real face from a high-quality dataset with a diverse set of gender, race, and age. Human participants are found to struggle in all perceptual tasks, supporting the need for effective computational solutions to better protect us from deep fakes.

Chapter 2

Detecting Deep Fakes from Facial Behavior

The current deep fakes, all three kinds, have one thing in common, they change the way a person talks. For example, in face-swap and puppet master deep fakes we see the face of one person, but the expressions and head movements are controlled by another person. On the other hand, in lip-sync deep fakes the mouth is altered to a new speech, whereas the rest of the face is taken from some other context. This will create a mis-match between the movement of mouth and rest of the face. Exploiting this observation, we describe a forensic technique that models facial expressions and movements that typify an individual's speaking pattern. Although not visually apparent, we show these correlations are often violated by the nature of how deep-fake videos are created and can, therefore, be used for authentication [1].

2.1 Introduction

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 [47] as well as [48] in which upper-body movements were 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

[1] This work was first published as *Protecting World Leaders Against Deep Fakes* in CVPRW, 2019 [50]

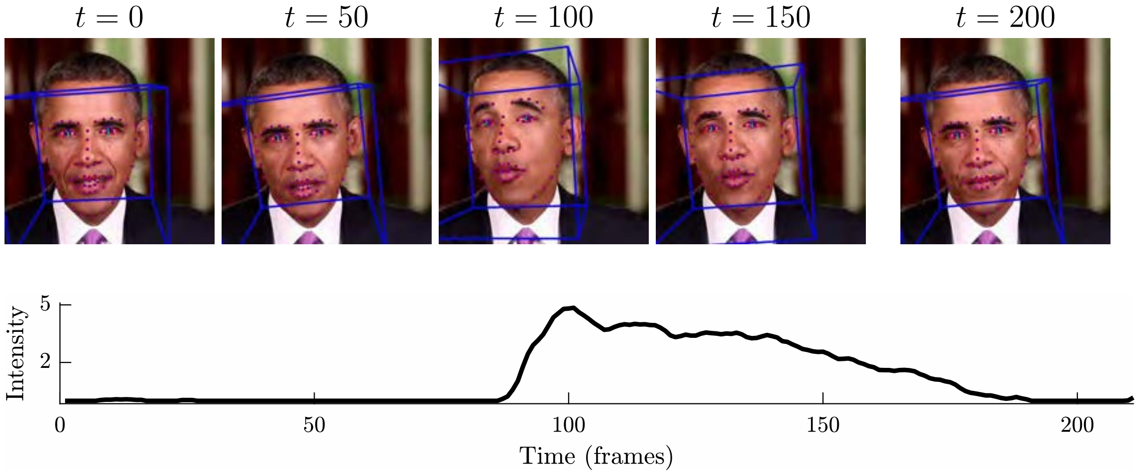


Figure 2.1: Shown above are five equally spaced frames from a 250-frame clip annotated with the results of OpenFace tracking. Shown below is the intensity of one action unit AU01 (eyebrow lift) measured over this video clip.

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 as soft biometric models of high-profile individuals and use these models 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.

2.2 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 [51]. We then build a novelty detection model (one-class support vector machine (SVM) [52]) that distinguishes an individual from other individuals as well as comedic impersonators and deep-fake impersonators.

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	0.11	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 2.1: 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.

2.2.1 Datasets

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 2.1 are the video and segment duration 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 [53]; 2) comedic impersonators for each POI, (Table 2.1); 3) face-swap deep fakes, lip-sync deep fakes, and puppet-master deep fakes (Table 2.1). Shown in Figure 2.2 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.

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 [2]. Each GAN was trained with approximately 5000 images per POI. The GAN then replaces the 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) [27] 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 [30]. 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. 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] https://github.com/shaoanlu/faceswap-GAN

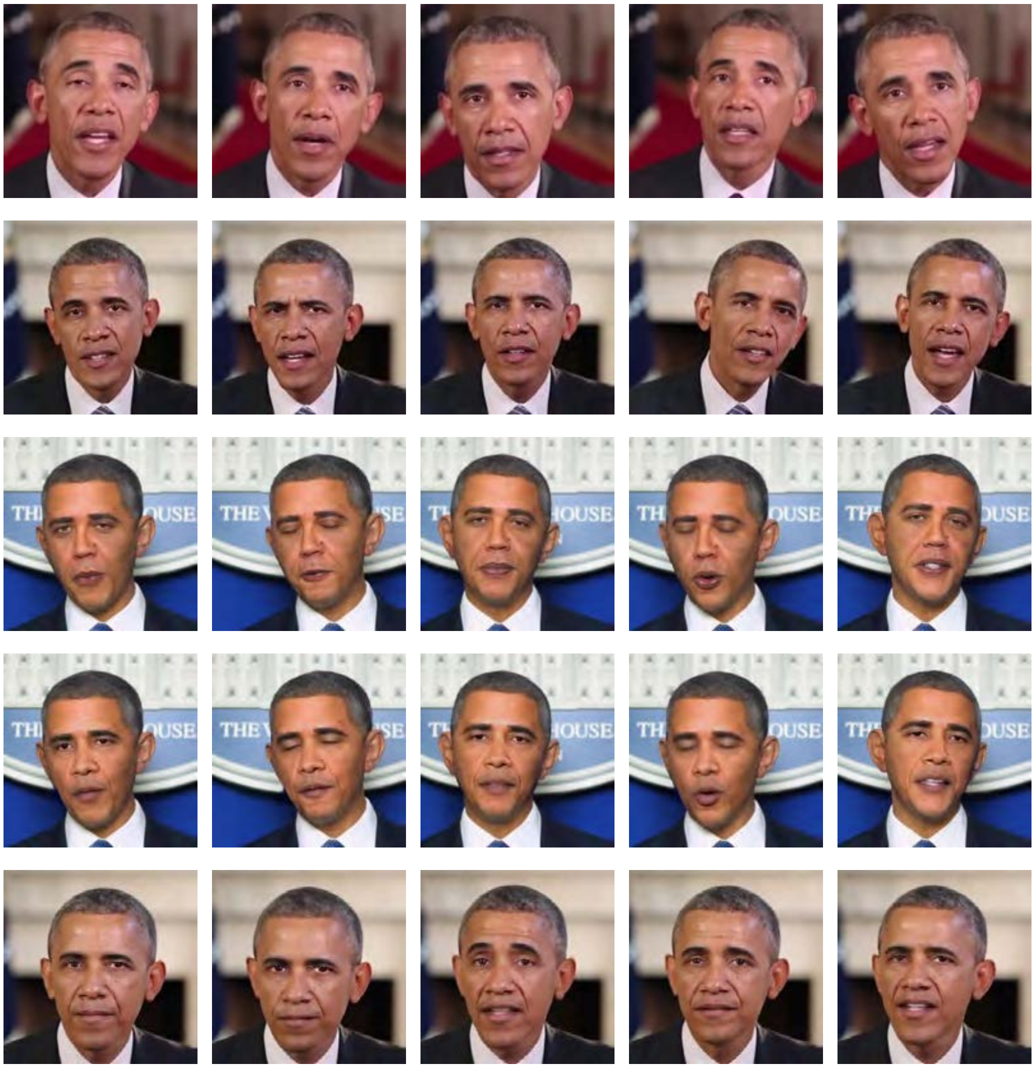


Figure 2.2: Shown from top to bottom, are five example frames of a 10-second clip from original, lip-sync deep fake, comedic impersonator, face-swap deep fake, and puppet-master deep fake.

2.2.2 Facial Tracking and Measurement

We use the open-source facial behavior analysis toolkit OpenFace2 [54–56] 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 2.1.

The movements of facial muscles can be encoded using facial action units (AU) [51]. 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), and 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 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 captures 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 by 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 2.2.1). 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.

2.2.3 Modeling

Shown in Figure 2.3 is a 2-D t-SNE [57] visualization of the 190-dimensional features for Hillary Clinton, Barack Obama, Bernie Sanders, Donald Trump, Elizabeth Warren, random people [53], 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

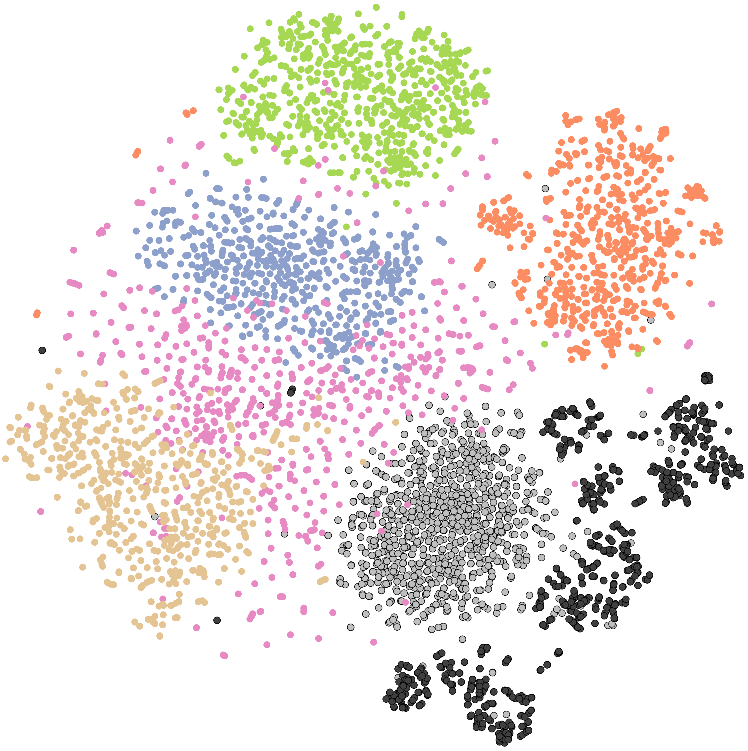


Figure 2.3: Shown is a 2-D visualization of the 190-D features for Hillary Clinton (brown), Barack Obama (light gray with a black border), Bernie Sanders (green), Donald Trump (orange), Elizabeth Warren (blue), random people [53] (pink), and lip-sync deep fake of Barack Obama (dark gray with a black border).

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 are rapidly evolving. As such, we train a novelty detection model (one-class SVM [52]) 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 [53]. 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 [52] (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 all 10-second video clips are split into 80:20 training:testing data sets, in which there was no overlap in the training and testing video segments.

2.3 Results

The performance of each POI-specific model is tested using the POI-specific comedic impersonators and deep fakes, Section 2.2.1. 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 (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.

2.3.1 Barack Obama

Shown in top half of Table 2.2 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.

	random	comedic			puppet-
	people	impersonator	face-swap	lip-sync	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 2.2: 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.

To select the best features for classification, 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 2.4 is the testing accuracy as a function of the number of features for the first 29 iterations of this process (the training accuracy reached a maximum at 29 features). This iterative training was performed on 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

	random	comedic	
	people	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 2.3: 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).

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-corner 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 2.2 is a comparison of the accuracy for the full 190 features and the 29 features enumerated in Figure 2.4. The bold-face values in this table denote those accuracies that are improved relative to the full 190 feature

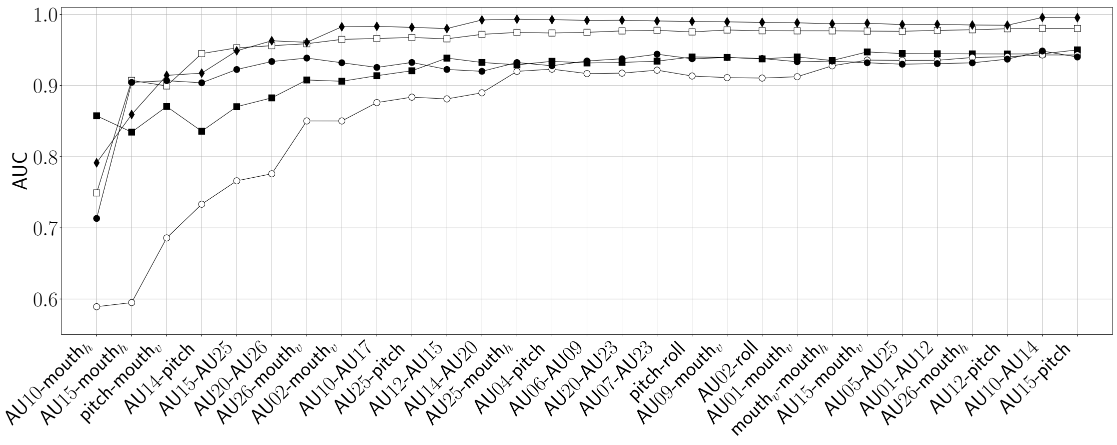


Figure 2.4: The accuracy (as AUC) for comedic impersonator (black square), random people (white square), lip-sync deep fake (black circle), face-swap deep fake (white circle), and puppet-master (black diamond) for a classifier trained on between one and 29 features as enumerated on the horizontal axis. 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.

set. We next test the robustness of these 29 features to a simple laundering attack, to the length of the extracted video clip, and to the speaking context.

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 2.2). 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.

Comparison to FaceForensics++

We compare our technique with the CNN-based approach used in FaceForensics++ [53] in which multiple models were trained to detect three types of face manipulations including face-swap deep fakes. We evaluated the higher-performing models trained using XceptionNet [58] architecture with cropped faces as input. The performance of these models was tested on the real, face-swap deep fake, lip-sync

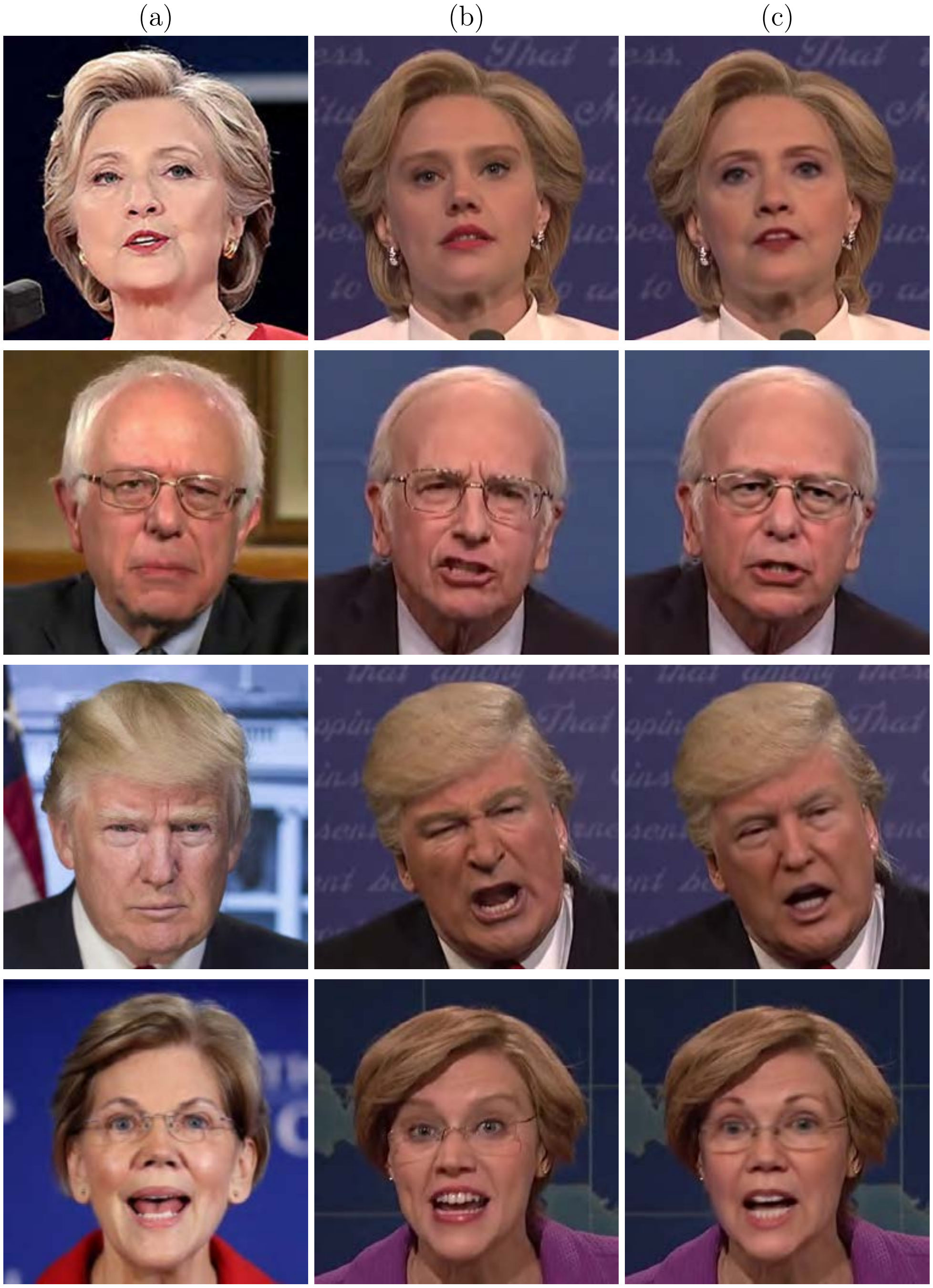


Figure 2.5: Shown are sample frames for (a) real; (b) comedic impersonator; and (c) face-swap for four POIs.

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 data set.

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.

2.3.2 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 2.5 are sample frames from videos collected for these four leaders (see Table 2.1). For each POI, a model was trained using the full set of 190 features. Shown in Table 2.3 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.

2.4 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 can augment our

[3] https://www.niessnerlab.org/projects/roessler2019faceforensicspp.html

models with a linguistic analysis that captures correlations between what is being said and how it is being said.

Chapter 3

Detecting Deep Fakes from Facial Appearance and Behavior

Face-swap deep fakes has been the most widely spread type of deep fakes on the internet. One of the most prevalent use of face-swap deep fakes has been seen against women, where more than 90% of the fakes are used for creating non-consensual pornography. Here we describe a biometric-based forensic technique for detecting face-swap deep fakes. This technique combines a static biometric based on facial recognition with a temporal, behavioral biometric based on facial expressions and head movements, where the behavioral embedding is learned using a CNN with a metric-learning objective function. We show the efficacy of this approach across several large-scale video datasets, as well as in-the-wild deep fakes [1].

3.1 Introduction

The creation of non-consensual pornography was the first use of deep fakes, and continues to pose a threat particularly to women, ranging from celebrities to journalists, and those that simply attract unwanted attention [6]. In response, several U.S. states have recently passed legislation trying to mitigate the harm posed by this content, and similar legislation is being considered at the U.S. federal and international levels. In addition, the democratization of access to sophisticated technology to synthesize highly realistic fake audio, image, and videos promises to add to our struggle to contend with dis- and mis-information campaigns designed to commit small- to large-scale fraud, disrupt democratic elections, and sow civil

[1] This work was first published as *Detecting deep-fake videos from appearance and behavior* in WIFS, 2020 [59]

unrest.

We describe a forensic technique to authenticate face-swap deep-fake videos in which a person's facial identity is replaced with another's. The most common approach to detecting these deep fakes leverages low-level pixel artifacts introduced during the synthesis process. These approaches suffer from vulnerability to simple counter-measures including trans-coding and resizing, and often struggle to generalize to new synthesis techniques (see Section 1.1 for more details).

In contrast, in our approach we leverage a more fundamental flaw in deep fakes: the face-swap deep fake is simply not the person it purports to be. In particular, we combine a static biometric based on facial identity with a temporal, behavioral biometric based on facial expressions and head movements. The former leverages standard techniques from face recognition, while the latter leverages a learned behavioral embedding using a convolutional neural network (CNN) powered by a metric-learning objective function. These two biometric signals are used because we observe that the facial behaviors in a face-swap deep fake remain those of the original individual, while the facial identity is of a different individual. By matching the behavioral and facial identities against a set of authentic reference videos, inconsistencies in the matching identities can reveal face-swap deep fakes. Our experimental results against thousands of unique identities spanning five large datasets support this hypothesis.

Our behavioral model is constructed by stacking together static FAb-Net features [60] over time (four seconds). By combining many FAb-Net features, which themselves capture static head pose, facial landmarks, and facial expression, we are able to capture spatiotemporal behaviors. Unlike previous chapter for modeling spatiotemporal human behavior that required a specific model for each person, we will show that the metric-learning objective used by our CNN to learn this behavioral feature allows us to build a generic model that can be trained on one group of people in one dataset and generalize to previously unseen people in different datasets. We summarize our primary contributions as:

- a novel spatiotemporal behavior model for capturing facial expressions and head movement that generalizes to previously unseen people;
- a novel combination of appearance and behavioral biometrics for detecting face-swap deep-fake videos;
- a large-scale evaluation across five large data sets consisting of thousands of real and deep-fake videos, the results of which show that our approach is highly effective at detecting face-swap deep fakes; and

- an analysis of the underlying methodology and results that provides insight into the specific nature of the learned features, and the robustness of our approach across different datasets, manipulations, and qualities of deep fakes.

In the next section, we describe our methods in detail and show the efficacy of our approach across five large-scale video datasets, as well as in-the-wild deep fakes.

3.2 Methods

We begin by describing the five datasets used for validation and analysis. We next describe two biometric measurements that underlie our forensic detection scheme. These include a biometric based on temporal behavioral (facial expressions and head movements) and a biometric based on static facial features.

3.2.1 Datasets

The world leaders dataset (WLDR) [50] consists of several hours of real videos of five U.S. political figures, their political impersonators, and face-swap deep fakes between each political figure and their corresponding impersonator. We augmented this dataset with five new U.S. political figures.

The FaceForensics++ dataset (FF) [53] consists of 1000 YouTube videos of 1000 different people, mostly news anchors and video bloggers. Each video was used to create four types of deep fakes: DeepFake, FaceSwap, Face2Face, and Neural Textures. We only use the first two categories of fakes as only these are face-swap deep fakes. After removing videos with multiple people or with identities overlapping to other datasets, we were left with 990 real videos and the corresponding 1980 deep-fake videos.

The DeepFake Detection dataset (DFD) [61] by Google/Jigsaw consists of 363 real and 3068 face-swap deep fakes of 28 paid and consenting actors. Each individual was made to perform tasks like walking, hugging, talking, etc. in different expressions ranging from happy, to angry, neutral, or disgust. For our analysis, we selected only those videos where the individual was talking, resulting in 185 real and 1577 deep-fake videos.

The Deep Fake Detection Challenge Preview dataset (DFDC-P) [62] consists of 1131 real and 4113 face-swap deep fakes videos of 66 consenting individuals of various genders, ages and ethnic groups. It is one of the largest deep fake dataset with videos of various quality, viewpoints, lighting conditions and scenes.

The Celeb-DF (Ver. 2) dataset (CDF) [63] is currently the largest publicly available deep-fake dataset. It is reported as containing 5639 face-swap deep fakes generated from 590 YouTube videos of 61 celebrities speaking in different settings ranging from interviews, to TV-shows, and award functions (we, however, only identified 59 unique identities in the downloaded dataset).

For each identity in the WLDR, DFD, and DFDC-P datsets, a random 80% of the real videos are used for the reference set and the remaining 20% are used for testing. In these three datasets there were sufficient videos of each individual in similar contexts. In contrast, the FF and CDF datasets had either only a small number of videos per individual or the context for each individual varied drastically. For these two datasets, therefore, we take a different approach to creating the reference/testing sets. In particular, each real video is divided in half, the first half of which is used for reference, and the second half used for testing. Similarly, we split each fake video in half, discard the first half and subject the second half to testing. The first half is discarded because the real counterpart of this video is used for reference, thus avoiding any overlap in utterances between the reference and testing. We recognize that this split is not ideal as video halves are not independent, but as we will see below, there is little difference in the results between the 80/20 splits and these 50/50 splits.

Each reference and testing video is re-saved at a frame-rate of 25fps (and a ffmpeg quality of 20). This consistent frame-rate allows us to partition each video into overlapping 4-second clips, each of 100 frames, with a 5-frame sliding window.

3.2.2 Behavior

FAb-Net nicely captures the frame-based facial movements and expressions but is, by design, identity-agnostic. We seek to learn a modified embedding that both captures facial movements and expressions, but also distinguishes these features across individuals. That is, starting with the static FAb-Net features, we learn a low-dimensional mapping that encodes identity-specific spatiotemporal behavior.

Given FAb-Net feature matrices for n, t-frame video clips $X_1, \dots, X_n$ with identity labels $y_1, \dots, y_n$, we learn a mapping $f(\cdot) : \mathbb{R}^{256 \times t} \to \mathbb{R}^d$, that projects X_i to an embedding space such that the similarity S_{ij} between $f(X_i)$ and $f(X_j)$ is high if $y_i = y_j$ (positive sample) and S_{ij} is low if $y_i \neq y_j$ (negative sample). Because, the output $f(X_i)$ is normalized to lie on a unit sphere, a cosine similarity, between two vector-based representations, is used to compute S_{ij}.

To learn the mapping $f(\cdot)$, a CNN is trained with a multi-similarity metric-learning objective function [64]. Following the approach in [64], the loss for a mini-batch is computed as follows. First, for every input X_i, hard positive and negative samples are selected. For hard negative samples (where $y_i \neq y_j$), a sample X_j is selected if $S_{ij} > \min\{S_{ik} - \epsilon\}$, for all k such that $y_i = y_k$, and where ϵ is a small margin. This formulation selects the most confusing negative samples whose similarity with the input is larger than the minimum similarity between the input and all positive samples. Similarly, for hard positive samples (where $y_i = y_j$), a sample X_j is selected if $S_{ij} < \max\{S_{ik} + \epsilon\}$, for all k such that $y_i \neq y_k$. Here, the most meaningful positive samples are selected by comparing to the negative samples most similar to the input.

A soft weighting is then applied to rank these selected samples according to their importance for learning the desired embedding space. For a given input X_i, let $\mathcal{N}_i$ and $\mathcal{P}_i$ represents the selected negative and positive samples that are weighted as follows:

$$w_{ij}^{-} = \frac{e^{\beta(S_{ij}-\lambda)}}{1 + \sum_{k \in \mathcal{N}_i} e^{\beta(S_{ik}-\lambda)}} \quad \text{and} \quad w_{ij}^{+} = \frac{e^{-\alpha(S_{ij}-\lambda)}}{1 + \sum_{k \in \mathcal{P}_i} e^{-\alpha(S_{ik}-\lambda)}}, \tag{3.1}$$

where α, β, and λ are hyper-parameters. Finally, the loss $\mathcal{L}$ over a mini-batch of size m is:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{k \in \mathcal{P}_i} e^{-\alpha(S_{ik}-\lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{k \in \mathcal{N}_i} e^{\beta(S_{ik}-\lambda)} \right] \right\}. \tag{3.2}$$

By performing supervised training using the identity labels in the training data, the network is encouraged to learn an embedding space that clusters the biometric signatures by identity.

Our model is trained on the VoxCeleb2 dataset [65], containing over a million utterances from $5,994$ unique identities. The size of the input feature matrix is fixed to $t = 100$, corresponding to a 4-second video clip at 25 frames/second (this clip size was selected as it was the minimum clip size of the VoxCeleb2 utterances). We used the ResNet-101 network architecture [66], where the input layer of the network is modified to the size of our feature matrix (256×100). A fully-connected output layer of size $d = 512$ is added on top of this network, forming our final feature vector, which is normalized to be zero-mean and unit-length before computing the loss. We name this network Behavior-Net.

The CNN training is performed for $10,000$ iterations with a mini-batch of size 256. Following [64], in each mini-batch, 32 identities are randomly selected, for which eight utterance videos (each of variable length) are randomly selected, from which a randomly selected 100-frame sequence is extracted. All other optimization hyper-parameters are the same as in [64].

Even though the Behavior-Net features are trained only on the VoxCeleb2 dataset, as described below, these features will be used to classify different identities across different datasets. This generalizability is both practically useful and suggests that the underlying Behavior-Net captures intrinsic properties of people.

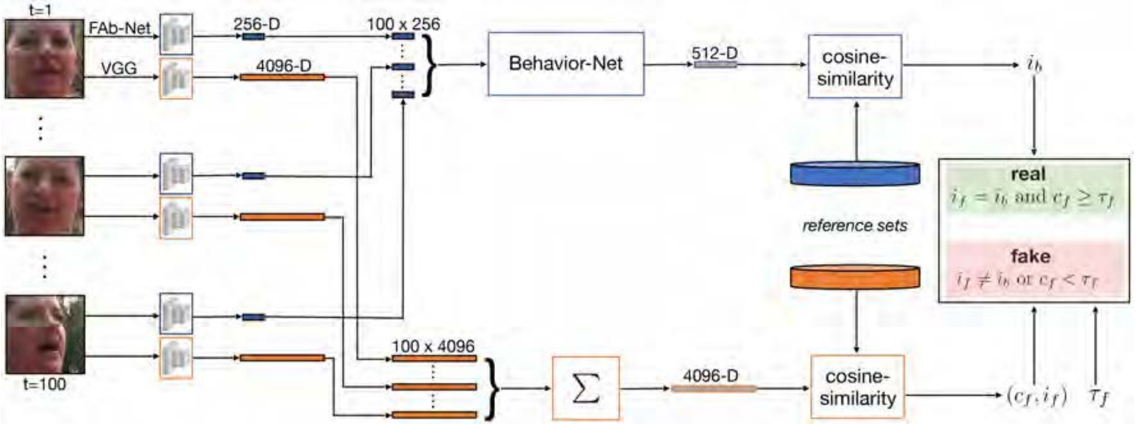


Figure 3.1: An overview of our authentication pipeline (see Section 3.2.4).

3.2.3 Appearance

Rapid advances in deep learning and access to large datasets have led to a revolution in face recognition. We leverage one such fairly straight-forward approach, VGG [67], a 16-layer CNN trained to perform face recognition on a dataset consisting of $2,622$ identities. VGG yields a distinct 4096-D face descriptor per face, per video frame. These descriptors are averaged over the 100 frames of the 4-second video clip to yield a single facial descriptor.

Faces for this facial biometric and the behavioral biometric are extracted using OpenFace [54]. Once localized and extracted from a video frame, each face is aligned and re-scaled to a size of 256×256 pixels.

3.2.4 Authentication

Given a authentic 4-second video clips for all unique identities, two reference sets are created with the VGG facial and Behavior-Net features. Define F_i to be the $4096 \times m_i$ real-valued matrix consisting of the VGG features for m_i video clips of identity i. Similarly, define B_i to be the $512 \times m_i$ real-valued matrix consisting of the Behavior-Net features for the same m_i video clips, also of identity i. Each column of the matrices F_i and B_i contains the VGG and Behavior-Net features for a single video clip.

Given these reference sets, a previously unseen 4-second video clip is authenticated as follows. First, extract the facial and Behavior-Net features, $\vec{f} \in \mathbb{R}^{4096}$ and $\vec{b} \in \mathbb{R}^{512}$. Next, find the identities, i_f and i_b in the reference sets with the most similar features

	Average	WLDR	FF	DFD	DFDC-P	CDF
real	96.5%	99.6%	99.2%	93.1%	93.1%	97.6%
fake	91.8%	95.8%	98.7%	93.2%	71.7%	99.4%
average	94.2%	97.7%	98.9%	93.2%	82.4%	98.5%

Table 3.1: Classification accuracies corresponding to the ROCs in Fig. 3.2 at a fixed threshold of $\tau_f = 0.86$.

using a cosine-similarity metric:

$$i_f \;=\; \arg\max_i\{\max(\vec{f^t}\cdot F_i)\} \qquad \text{and} \qquad i_b \;=\; \arg\max_i\{\max(\vec{b^t}\cdot B_i)\} \tag{3.3}$$

With these matched identities, a video clip is classified as real or fake following two simple rules (see also Fig. 3.1):

1. A video clip is classified as real if the facial and Behavior-Net identities are the same, $i_f = i_b$, and if the facial similarity is above a specified threshold, $c_f >= \tau_f$, where $c_f = \max(\vec{f^t}\cdot F_{i_f})$ (i.e., a close facial match is found).
2. A video clip is classified as fake if either
   (a) the matched identities are different, $i_f \neq i_b$, or
   (b) the facial similarity is below threshold, $c_f < \tau_f$.

The rationale for the asymmetric treatment of the facial and Behavior-Net similarities is that in a face-swap deep fake, the facial identity of a person is modified but typically not the behavior. As a result, it is possible for a person's facial identity to be significantly different in a test video than in their reference videos, in which case, we should not be confident of the facial identity match.

3.3 Results

We describe the overall accuracy of detection followed by an analysis of robustness and relative importance of the appearance and behavioral features;

3.3.1 Identification

Shown in Fig. 3.2 are the receiver operating curves (ROC) for each of the five datasets enumerated in the previous section, along with the average across all datasets. The

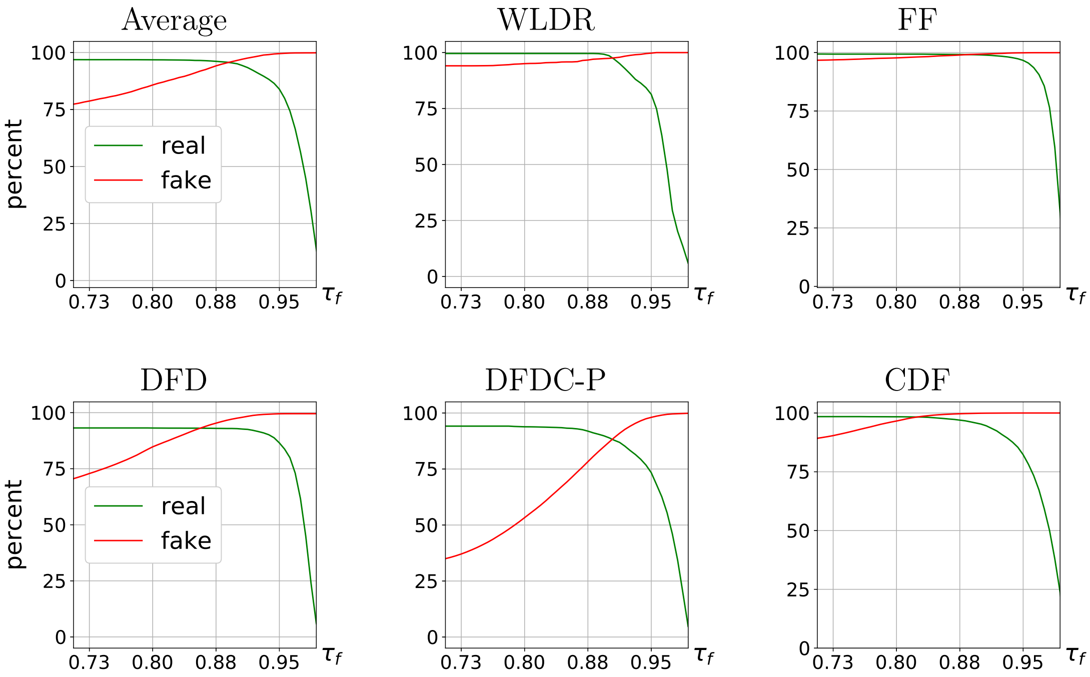


Figure 3.2: Shown are receiver operating curves (ROC) for each of five datasets and the average across all datasets (top-left panel). The green/red curves correspond to the accuracy of classifying real/fake videos. The horizontal axis corresponds to the VGG threshold (τ_f).

green/red curves correspond to the accuracy of classifying real/fake videos. The horizontal axis corresponds to the facial VGG threshold (τ_f) used in determining if a video clip should be classified as real or fake (see Section 3.2.4 and Fig. 3.1).

As expected, as the threshold increases, the detection accuracy for fake (red) increases while the detection accuracy for real (green) decreases, particularly dramatically for threshold τ_f values that approach the maximum value of 1.0. Recall that these accuracies are on a single 4-second clip.

The cross-over accuracies in Fig. 3.2 are 95.5% (Average), 97.3% (WLDR), 99.1% (FF), 93.1% (DFD), 88.4% (DFDC-P), and 98.3% (CDF). These cross-over points, however, come at varying τ_f threshold values. Shown in Table 3.1 is the detection accuracy, ranging from 82.4% for DFDC-P to 98.9% for FF, for a fixed threshold of $\tau_f = 0.86$.

Note that the accuracy for the DFDC-P is unusually low. This is because many of the fake videos in this dataset failed to convincingly map the facial appearance of the desired source identity into the target video. Shown in Fig. 3.3 is a representative example of this problem. Shown is one frame from the source video, one frame from the target video, and the corresponding frame from the face-swap deep-fake video in which the source

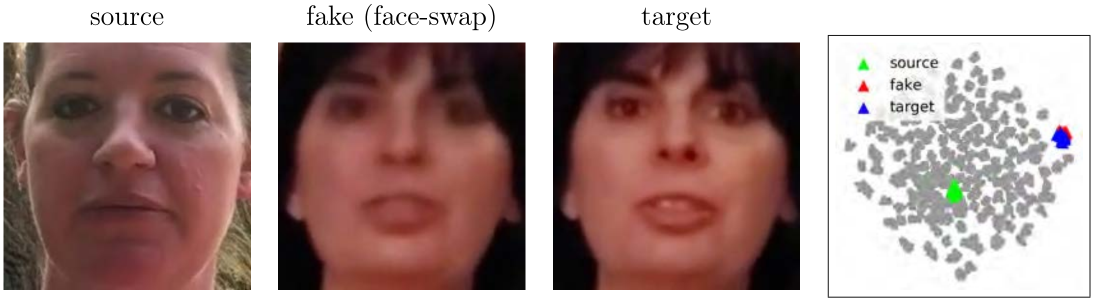


Figure 3.3: Shown is an example frame of a face-swap deep fake (second panel) from the DFDC-P dataset, in which the source identity (first panel) should be mapped onto the target (third panel), which is clearly not the case in this example. Shown in the fourth panel is the dimensionality-reduced visualization of the 4096-D VGG features from all the identities (gray), source identity (green), target identity (blue), and the face-swap identity (red). This visualization shows that the source identity is not successfully mapped onto the deep fake (see also Fig. 3.4).

identity should be mapped into the target video. In this example drawn from the DFDC-P dataset, we can clearly see that the source identity was not mapped into the target video, but rather continues to look like the target. Shown in Fig. 3.4 is confirmation that this problem persists throughout the DFDC-P dataset. In particular, shown in the first and third columns are, for each dataset, the distribution of similarities in facial identities (as measured by the facial VGG cosine similarity) between all faces in the fake videos and their corresponding source identities. Shown in the second and fourth columns are the similarity in facial identities all faces in the fake videos and their corresponding target identities. In a successful face swap, in which the identity in the target is replaced with that in the source, the facial similarity between the source and fake should be higher than the target and fake. Correspondingly, for each dataset, except DFDC-P, the average facial similarity of the fakes is higher relative to the source than the target. For the DFDC-P dataset, however, the fakes are on average closer to the target than the source. This difference accounts for the low accuracy on the DFDC-P dataset as both behavior and appearance of the fakes correspond to the target identity and are thus classified as real by our algorithm. Although this effect is most pronounced in the DFDC-P dataset, the DFD dataset also suffers from a similar problem, failing to convincingly map the source to the target identity. These failures justify our use of a confidence threshold in the facial similarity matching (case 2(b) in Section 3.2.4).

We next evaluate our detection algorithm against three in-the-wild, face-swap deep-fake videos downloaded from YouTube. These three deep fakes were created using the following source and target combinations: 1) Steve Buscemi mapped onto Jennifer Lawrence [2]; 2)

[2] https://www.youtube.com/watch?v=VWrhRBb-1Ig

	WLDR	FF	DFD	DFDC-P	CDF
Protecting World Leaders [50]	0.93	–	–	–	–
2-stream [36]	–	0.70	0.52	0.61	0.53
XceptionNet-c23 [63]	–	**0.99**	0.85	0.72	0.65
Head Pose [44]	–	0.47	0.56	0.55	0.54
MesoNet [68]	–	0.84	0.76	0.75	0.54
Face Warping [40]	–	0.80	0.74	0.72	0.56
Ours: Appearance and Behavior	**0.99**	**0.99**	**0.93**	**0.95**	**0.99**

Table 3.2: Comparison of our approach with previous work over multiple benchmarks [63]. The reported values correspond to the AUC. Although not a perfect comparison due to significantly different underlying methodologies, our approach does perform well. The FF dataset in this comparison consists of the FaceSwap and Deepfake categories.

Tom Cruise mapped onto Bill Hader [3]; and 3) Billie Eilish mapped onto Angela Martin [4]. Because, only Jennifer Lawerence was already in our reference set (CDF), real videos for the other five identities were downloaded from YouTube to augment our reference set. This included three minutes of videos of Angela Martin from The Office and 20 minutes of interview videos for each of Billie Eilish, Steve Buscemi, Bill Hader, and Tom Cruise. The accuracy rate for each of these face-swap deep fakes is 100%.

Lastly, shown in Table 3.2 is a comparison of our detection accuracy, measured using area under the curve (AUC), to six previous deep-fake detection schemes. Our scheme outperforms or is equal to previous approaches across all datasets. Note, however, that this is not a perfect comparison because our approach has access to a reference set of only real videos to compare against, as compared to these other fully-supervised approaches with access to real and fake reference videos.

3.3.2 Analysis

Our Behavior-Net feature was designed to capture spatiotemporal behavior, while the VGG feature captures facial identity. Here we analyze our results in more detail to ensure that these two features are not entangled and that the Behavior-Net does in fact capture temporal properties not captured by the static FAb-Net features.

In the first analysis, we show that Behavior-Net does in fact capture behavior and not just a person's facial identity. Shown in Fig. 3.5(a) are the distributions of Behavior-Net similarities between source (blue)/target (orange) identities relative to their face-swap deep

[3] https://www.youtube.com/watch?v=r1jng79a5xc

[4] https://www.instagram.com/p/B6lXvJlIU92/

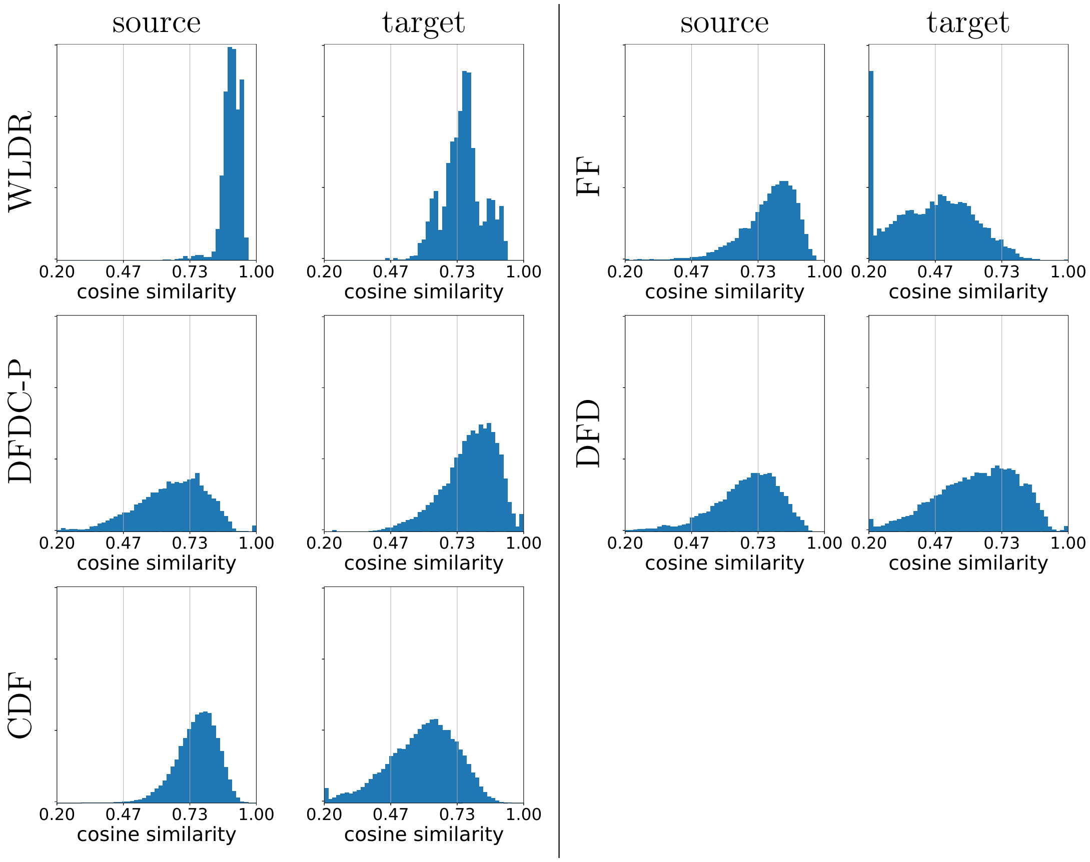


Figure 3.4: The distributions in the first and third column correspond to the facial similarity between the faces in the source and fake videos (as computed by the cosine similarity between corresponding VGG features). The distributions in the second and fourth column correspond to the facial similarity between the faces in the target and fake videos. In a successful face-swap deep fake, the source to fake similarity will be higher than the target to fake similarity, as is the case for the WLDR dataset. For the DFDC-P dataset, however, these distributions are reversed (see also Fig. 3.3).

fakes (recall that a face-swap deep fake is created by mapping an identity in a source video to a target video). The similarity of the target behavior relative to the face-swap deep fakes is much higher than the source, meaning that even though the facial identity in the deep fake matches the source, the behavioral identity still matches the target. This indicates that the Behavior-Net is capturing more information than just facial identity.

In the second analysis, we show that Behavior-Net captures identity-specific behaviors and not just identity-agnostic expressions or behaviors. This analysis is based on the real videos in the DFD dataset, where each of the 28 actors were recorded talking in different contexts ranging from a casual conversation sitting on a couch to a speech at a podium. Each of these contexts captured a specific facial expression ranging from neutral, to angry, happy, and laughing. And, each of these contexts were recorded twice, once with a still camera and once with moving camera. Shown in Fig. 3.5(b) are the distributions of Behavior-Net similarities between the same person in the same context (blue), the same person in different contexts (orange), and different people in the same context (green). When different people are recorded in the same context, we see that their Behavior-Net features are not similar, indicating that Behavior-Net captures identity-specific behaviors and not just specific contexts. At the same time, however, we see, that context can change an individual behavior (the orange vs. blue distributions). For example, a person is likely to have a different behavior when they are speaking casually to their friends as opposed to giving a formal speech to a large crowd. Nevertheless, our Behavior-Net captures identity-specific behaviors, albeit somewhat context dependent. Shown in Fig. 3.5(c) are the same distributions as in panel (b) but for only the static FAb-Net features. The distributions for the same person in the same context (blue), the same person in different contexts (orange), and different people in the same context (green) are all nearly identical, revealing that the static FAb-Net features does not capture identity-specific information.

In the third analysis, we analyse the amount of data required to build a reference set for an individual. For this analysis, the same reference set as before was used for the identities in FF, DFD, DFDC-P, and CDF. For the identities in the WLDR dataset (the only one with hours of video per person), the reference sets consists of between 1 and 2000 randomly selected 4-second clips. With 2, 30, 50, 100, 1000, and 2000 video clips, the average detection accuracy for identities in the WLDR dataset are 65.4%, 92.2%, 93.2%, 94.0%, 97.3%, and 97.7%, respectively. This rapid increase in accuracy and leveling off shows that large reference sets are not needed, assuming, again, that the context in which the individual is depicted is similar.

In this fourth, and final, analysis, we analyse the robustness of classification against a simple compression laundering operation. The video clips in our reference and testing sets, Section 3.3.1, are each encoded at a relatively high ffmpeg quality of qp=20 (the lower this value, the higher the quality). Each testing video clip was recompressed at a lower quality of qp= 40 and classified against the original reference set. For the same threshold ($\tau_f = 0.86$), the average detection accuracy remains high at 94.5% (WLDR), 98.1% (FF), 93.2% (DFD), 80.9% (DFDC-P), and 93.3% (CDF). These results are almost identical to

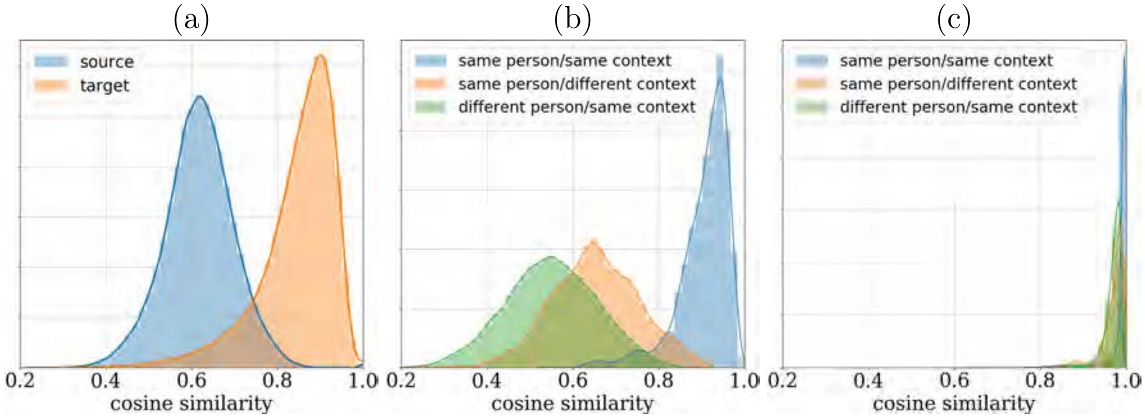


Figure 3.5: Shown in panels (a) and (b) are the distributions of spatiotemporal behavior similarity, measured as the cosine similarity between Behavior-Net feature vectors. Shown in panel (c) is the distribution of spatial FAb-net similarity. See text for a detailed explanation of each panel.

the high-quality videos in Table 3.1.

3.4 Discussion

We have developed a novel technique for detecting face-swap deep fakes. This technique leverages a fundamental flaw in these deep fakes in that the person depicted in the video is simply not the person that it purports to be. We have shown that a combination of a facial and behavioral biometric is highly effective at detecting these face-swap deep fakes. Unlike many other techniques, this approach is less vulnerable to counter attack and generalizes well to previously unseen deep fakes with previously unseen people.

Our forensic technique should generalize to so-called puppet-master deep fakes in which one person's facial expressions and head movements are mapped onto another person. These deep fakes suffer from the same basic problem as face-swap deep fakes in that the underlying behavior of the person is not that who it purports to be. As such, our combined facial and behavioral biometric should be able to detect these deep fakes.

We will, however, likely struggle to classify so-called lip-sync deep fakes in which only the mouth has been modified to be consistent with a new audio track. The facial identity and the vast majority of the behavior in these deep fakes will be consistent with the person depicted. To overcome this limitation, we seek to customize our behavioral model to learn explicit inconsistencies between the mouth and the rest of the face and/or underlying audio signal.

There is little question that the arms-race of synthesis and detection will continue.

While it may not be possible to entirely stop the creation and distribution of deep fakes, our, and related approaches, promise to make the creation of convincing deep fakes more difficult and time consuming. This will eventually take it out of the hands of the average person and relegate it to the hands of a fewer and fewer experts. While the threat of deep fakes will remain, this will surely be a more manageable threat.

Chapter 4

Detecting Deep Fakes from Aural and Oral Dynamics

A face-swap deep fake replaces a person's face – from eyebrows to chin – with another face. A lip-sync deep fake replaces a person's mouth region to be consistent with an impersonated or synthesized audio track. An overlooked aspect in the creation of these deep-fake videos is the human ear. Statically, the shape of the human ear has been shown to provide a biometric signal. Dynamically, movement of the mandible (lower jaw) causes changes in the shape of the ear and ear canal. While the facial identity in a face-swap deep fake may accurately depict the co-opted identity, the ears belong to the original identity. While the mouth in a lip-sync deep fake may be well synchronized with the audio, the dynamics of the ear motion will be de-coupled from the mouth and jaw motion. We describe a forensic technique that exploits these static and dynamic aural properties [1].

4.1 Introduction

Here we describe a high-level forensic technique to detect lip-sync and face-swap deep fakes. Most of the focus on creating deep-fake videos has been on facial expressions, the mouth, and audio-video synchronization. The creation of a lip-sync deep fake, for example, requires a detailed synthesis of the mouth region, teeth, and tongue, all the while making sure the mouth is properly synthesized with the audio and spoken phonemes. An overlooked aspect in the creation of these deep-fake videos is the human ear.

The reason for this is probably two-fold. The structure and movement of the human ear is complex, and it is likely our attention is not drawn to a person's ear when they are talking. Eye tracking studies on face perception have consistently revealed a Y-shaped

[1] This work was first published as *Detecting deep fakes from aural and oral dynamics* in CVPRW, 2021 [69]

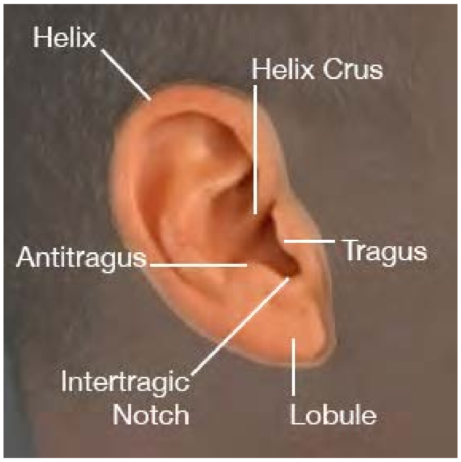


Figure 4.1: The human ear and a few of its parts.

pattern of fixations over the eye, nose and mouth regions [70, 71]. Janik et al. [72] found subjects spend 40% of the time looking at the eyes while free viewing facial photographs.

Both statically – the shape of the human ear provides a biometric signal [73–77] – and dynamically – movement of the mandible (lower jaw) causes changes in the shape of the ear and ear canal [78–80] – the human ear provides a rich source of forensic information. Specifically, while the facial identity in a face-swap deep fake may accurately depict the co-opted identity, the ears belong to the original identity. And, while the mouth in a lip-sync deep fake may be well synchronized with the audio, the dynamics of the ear motion will be de-coupled from the mouth and jaw motion. We describe a forensic technique that exploits these static and dynamic aural properties.

Biometric identification based on aural features has a well-established literature dating back as far as the 1890s [73–77]. The general, albeit not unanimous, conclusions today is certain aural features are distinct and stable over a person's lifetime. It remains unclear, however, if these aural features are distinct enough to work on a large scale, and if extraction of these features is sufficiently robust to work in the wild. For our purposes of deep-fake detection, however, the demands of distinctiveness are significantly less than in a biometric setting, and with our focus on video, feature extraction should be more robust than from only a single image.

In the next section, we place our work in context relative to previous forensic techniques. We then describe our underlying methodology and show the efficacy of our approach across simulated lip-sync deep fakes, production-quality deep fakes, and in-the-wild deep fakes.

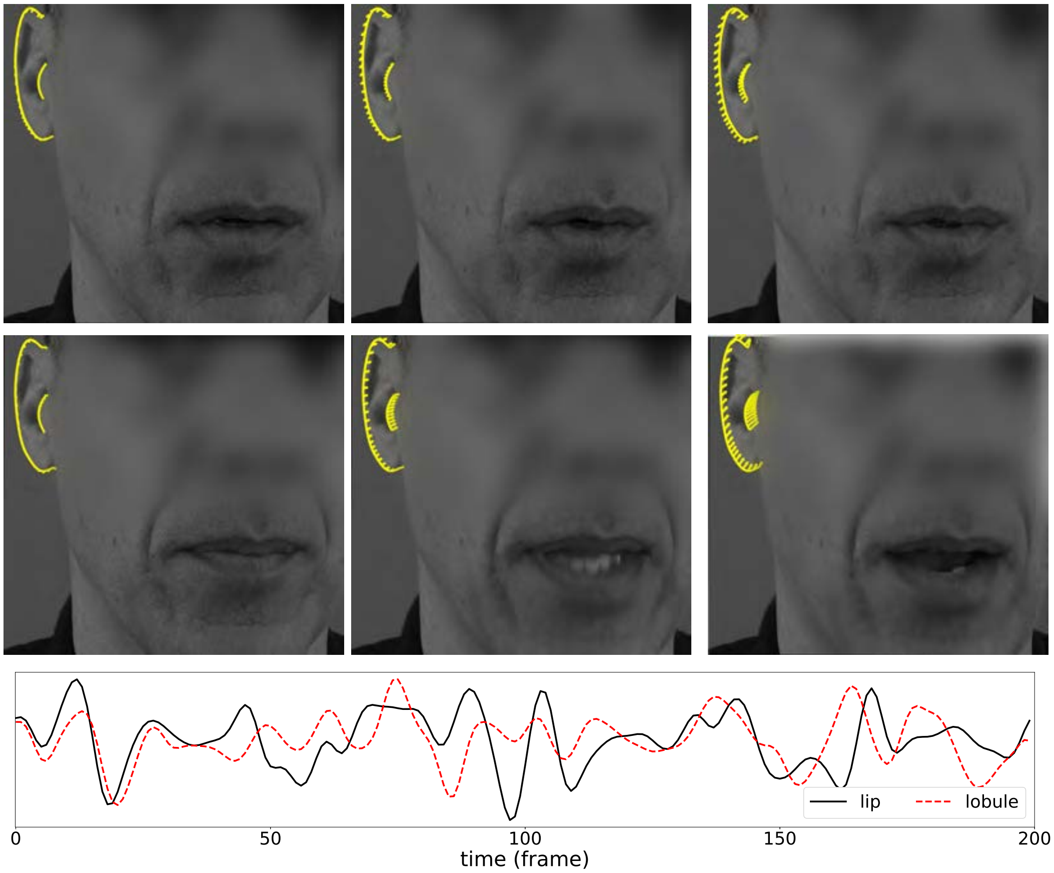


Figure 4.2: Shown in the first two rows are three equally-spaced frames in which the subject is speaking. Shown in each panel is a tracked Bezier curve corresponding to the ear's helix and lobule (larger outer curve) and tragus (smaller inner curve). The small vectors along each curve correspond to the estimated local motion (scaled by 5x), revealing how the ear moves during speech. Facial expressions such as raised eyebrows, smiling, and surprise induce similar aural motion. Shown in the lower panel is the measured horizontal lobule motion (red, dashed) and vertical lip distance (black, solid), revealing a correlation (r = 0.34) between these two signals.

	video (count)	total (seconds)	minimum (seconds)	maximum (seconds)
Joe Biden	21	769	12	59
Angel Merkel	10	228	12	42
Donald Trump	16	547	16	54
Mark Zuckerberg	17	484	20	29

Table 4.1: The number of videos in our data set, along with the total, minimum, and maximum video duration for each of four individuals.

4.2 Methods

We describe the dataset and aural dynamics methodology, followed by the aural biometric methodology.

4.2.1 Datasets

A total of 64 videos were downloaded from YouTube of Joe Biden, Angela Merkel, Donald Trump, and Mark Zuckerberg. These videos spanned in length between 12 and 59 seconds, Table 4.1. We ensured the left or right ear was visible throughout each video. For each frame of each video, we measured the aural motion, the vertical distance between the lips, and the audio RMSE. Due to large head movements, the feature tracking occasionally failed (4-5 times per video) and was corrected by manually re-annotating the necessary features.

In a lip-sync deep fake, the mouth movements of an existing video are modified to match a new audio. We used the following three strategies to generate such lip-sync deep fakes: (1) a lip-sync deep fake is simulated by simply correlating the aural movements from one video segment to the oral signal from a randomly selected segment of the same length; (2) visually compelling lip-sync deep fakes were generated for Biden, Merkel, Trump, and Zuckerberg, in which the mouth region is GAN-synthesized to be consistent with a new audio and optimized for visual quality and temporal coherence (courtesy of Kristof Szabo, Zoltan Kovacs, and Dominik Mate Kovacs). A total of six fakes were created for each of the four identities by swapping the original audio with a randomly selected audio from the same individual; and (3) three in-the-wild lip-sync deep fakes were downloaded from YouTube and Instagram, two for Donald Trump [2] and one for Mark Zuckerberg [3].

[2] https://www.instagram.com/p/ByPhCKuF22h/, https://youtu.be/VWMEDacz3L4

[3] https://youtu.be/cnUdOTpuoXI

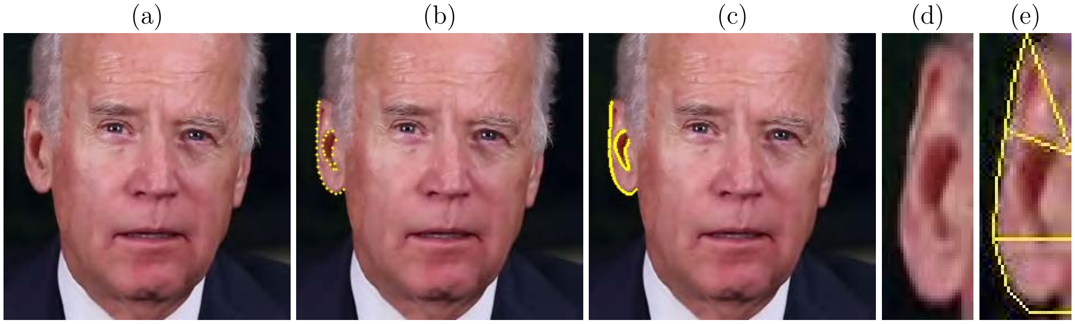


Figure 4.3: Shown are (a) a single video frame where the face has been tracked, aligned, and cropped; (b) 35 manually annotated aural landmarks; (c) 100 points on each of two Bezier fitted curves; (d) rotated and cropped ear; and (e) three regions from which local aural motion are averaged.

4.2.2 Aural Dynamics

The human ear has three primary sections: the inner- and middle-ear, and the outer-ear consisting of visible features like the lobule, tragus, and helix, Figure 4.1. Movement in the ear canal – connecting the outer-ear and middle-ear – has been studied in relationship with the movement of the mandible (lower jaw) [78–80]. Additional studies reveal the middle ear muscles to be responsive to face and head movements, onset of vocalization, yawning, swallowing, coughing, and laughing [81].

We observe such physiological movements in the middle ear can also be observed in movements of the outer-ear's lobule, tragus, and helix, Figure 4.2. We hypothesize that because deep fakes focus on the synthesis of the face, these aural movements will be absent or disrupted in deep fakes. We next describe techniques for measuring aural motion and correlating this motion to oral signals consisting of facial movements and auditory signals.

We describe the estimation of aural motion in a video in which it is assumed a single person is talking with their left or right ear visible throughout the video segment. This estimation is composed of four parts, as enumerated below.

Face Alignment: For each video frame, 68, 2D facial landmarks are extracted using Dlib [82]. Using these landmarks, the face in each frame is aligned such that the endpoints of the jaw (landmarks 0 and 16) lie on a horizontal line, are scaled to have a fixed distance of 164 pixels, and translated to a fixed location (pixel locations $(46, 90)$ and $(210, 90)$). After this alignment, a 256×256 pixel region is cropped around the face and ears, Figure 4.3(a).

Feature Tracking: In order to localize the ears in each frame, we begin by manually annotating 35 aural landmarks on the first aligned video-frame. The first set of 20 landmarks are on the outer portion of the ear, from the helix to the lobule, and the remaining set of 15 landmarks are around the tragus, Figure 4.3(b). We fit to each of these sets of landmarks, a Bezier curve of order 8 and 10, respectively. A total of 100 points are uniformly sampled from each of these curves, Figure 4.3(c). Lastly, the 200 Bezier points are tracked across all frames using the Kanade-Lucas-Tomasi (KLT) tracker [83].

Aural Alignment: In order to measure the local aural motion due to facial expressions and speech, we first eliminate the global motion due to head movements. In each frame, the tracked aural landmarks are affine-aligned to the landmarks in the previous frame. Each frame is then rotated such that the axes of the bounding box containing all of the aural landmarks are parallel to the image axes, Figure 4.3(d).

Motion Estimation: The local aural motion due to facial expressions and speech is estimated using dense optical flow between each consecutive aligned frames. The average 2D motion in the horizontal and vertical directions is computed in three aural regions around the helix, tragus, and lobule, Figure 4.3(e), yielding a total of six estimated aural motions. We next describe how these motions are correlated to oral signals consisting of facial expressions and speech.

Aural/Oral Correlations: We observe the per-frame aural movements are correlated to the per-frame vertical distance between the lips (i.e., the openness of the mouth) and audio root mean square energy (RMSE) (i.e., the loudness of the speech). While there are other facial and auditory correlations, we focus here on just these two. For each video frame, 68, 3D facial landmarks are estimated using the OpenFace toolkit [54]. The landmarks corresponding to the center of the top and bottom lip (landmarks 51 and 57) are used to compute vertical distance between the lips.

The audio RMSE is measured using the open-source python package LibROSA [84] over a sliding 0.032-second window and a hop length of 0.033 seconds. Given an audio with 16kHz sampling rate and a corresponding video of 30 fps, this yields a single audio RMSE value for each video frame.

The Pearson correlation between the horizontal and vertical aural motion in each of three ear regions and the above two oral signals is computed over a sliding 10-second segment with a 0.033-second shift. This yields a total of 12 correlations per each video segment. Shown in Figure 4.2 (bottom panel) is a representative example of the measured tragus horizontal movement (red) and lip vertical distance (black), from which the correlation is computed.

4.2.3 Aural Biometrics

The aural dynamics described above are designed to detect lip-sync deep fakes in which the aural and oral signals are desynchronized. In a face-swap deep fake, however, these signals are likely to be consistent with the original speaker. But, in a face-swap deep fake the ears in the video belong to the original identity and not to the person it purports to depict. As a result, we can leverage aural biometrics to verify the true identity in the video. There is a significant literature on aural biometrics including 2D, image-based features [85],

	Biden	Merkel	Trump	Zuckerberg
training (all)	0.97	0.90	0.93	0.87
simulated	0.97	0.82	0.93	0.87
GAN-generated	0.78	0.90	0.98	0.78
in-the-wild	–	–	0.70	0.71
training (ind)	0.99	0.96	0.99	0.97
simulated	0.96	0.80	0.98	0.86
GAN-generated	0.97	0.85	0.97	0.82
in-the-wild	–	–	0.76	0.77

Table 4.2: The performance (reported as area under the curve, AUC) for a single model trained on all four individuals (top), and separate models trained on each individual (bottom). All video segments are 10 seconds in length. Results are reported for the training dataset, and three different types of fakes: simulated, GAN-generated, and in-the-wild.

3D model-based features [86], and learned features [87].

Here we adopt a simple approach based on 2D, image-based features which capture the general shape of the ear. Any of a number of other techniques would be equally viable. In our approach, we first manually annotate 20 landmarks equally spaced from the helix to the lobule, and another 15 landmarks equally spaced around the tragus, Figure 4.7(c). The overall shape of the helix and tragus are characterized using two Bezier curves of order 8 and 10.

The shape of an ear is compared for similarity to a reference ear by first aligning the 35 aural landmarks, as described in the previous section. Because the ear may be imaged from any camera angle, the resulting perspective projection can significantly alter its appearance in the image. We assume, therefore, a reference ear in which the ear is parallel to the imaging plane, thus minimizing any perspective distortion. The comparison ear is then aligned to this reference ear using a planar homography [88] applied to the 35 aural landmarks. Although the ear is not perfectly planar, this homography is reasonable given the relatively small depth change along the ear as compared to a typical distance to the camera.

Once aligned, two ears are compared for similarity by measuring the average Euclidean distance between 100 equally sampled points on each of two Bezier curves and their closest point in the reference ear. This average distance is used as our measure of biometric similarity between two ears.

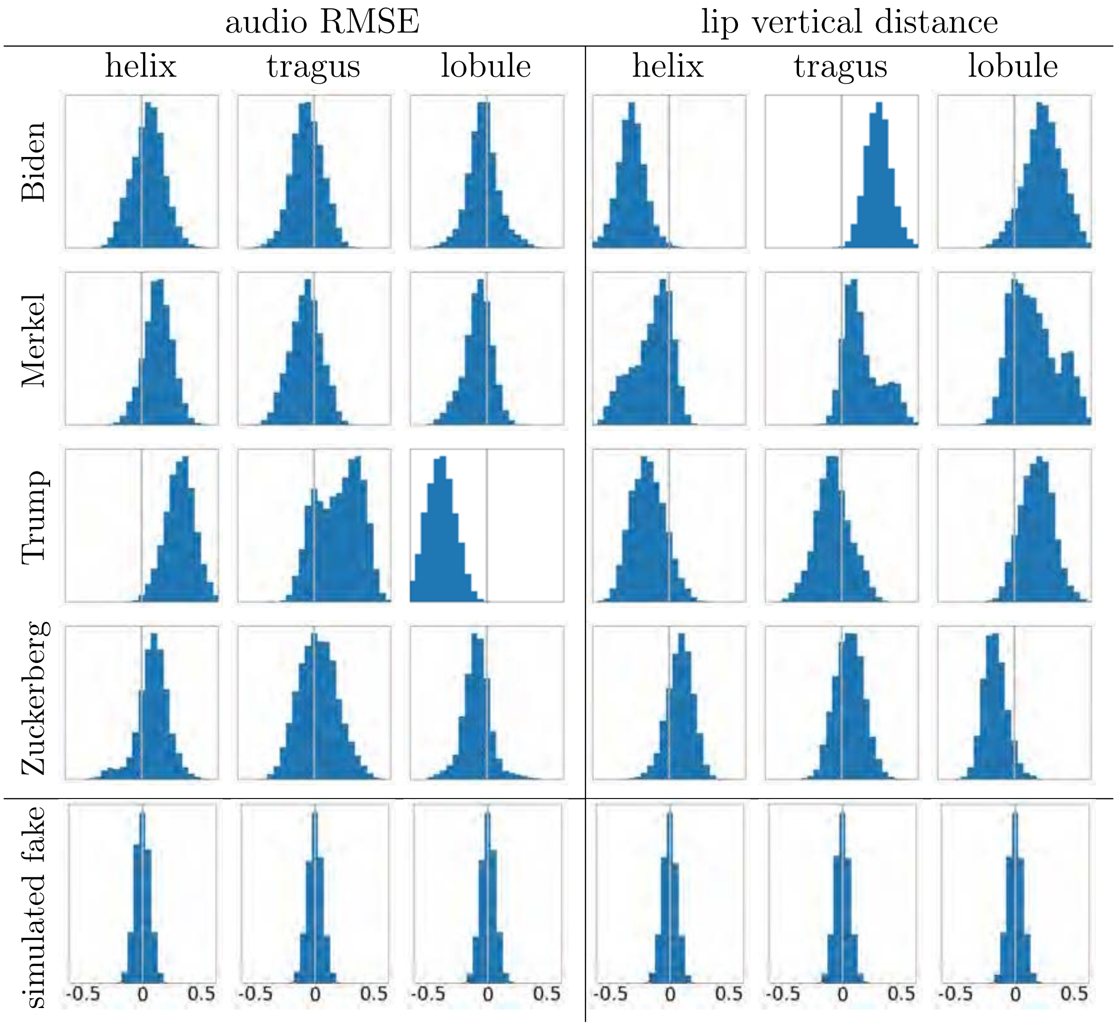


Figure 4.4: Shown are the distribution of correlations between audio (left) and lip vertical distance (right) and the *horizontal* motion of three aural areas. From top to bottom are the results for four individuals, and simulated fakes. While the fakes have no correlation, we see strong, but not necessarily consistent, correlations across individuals.

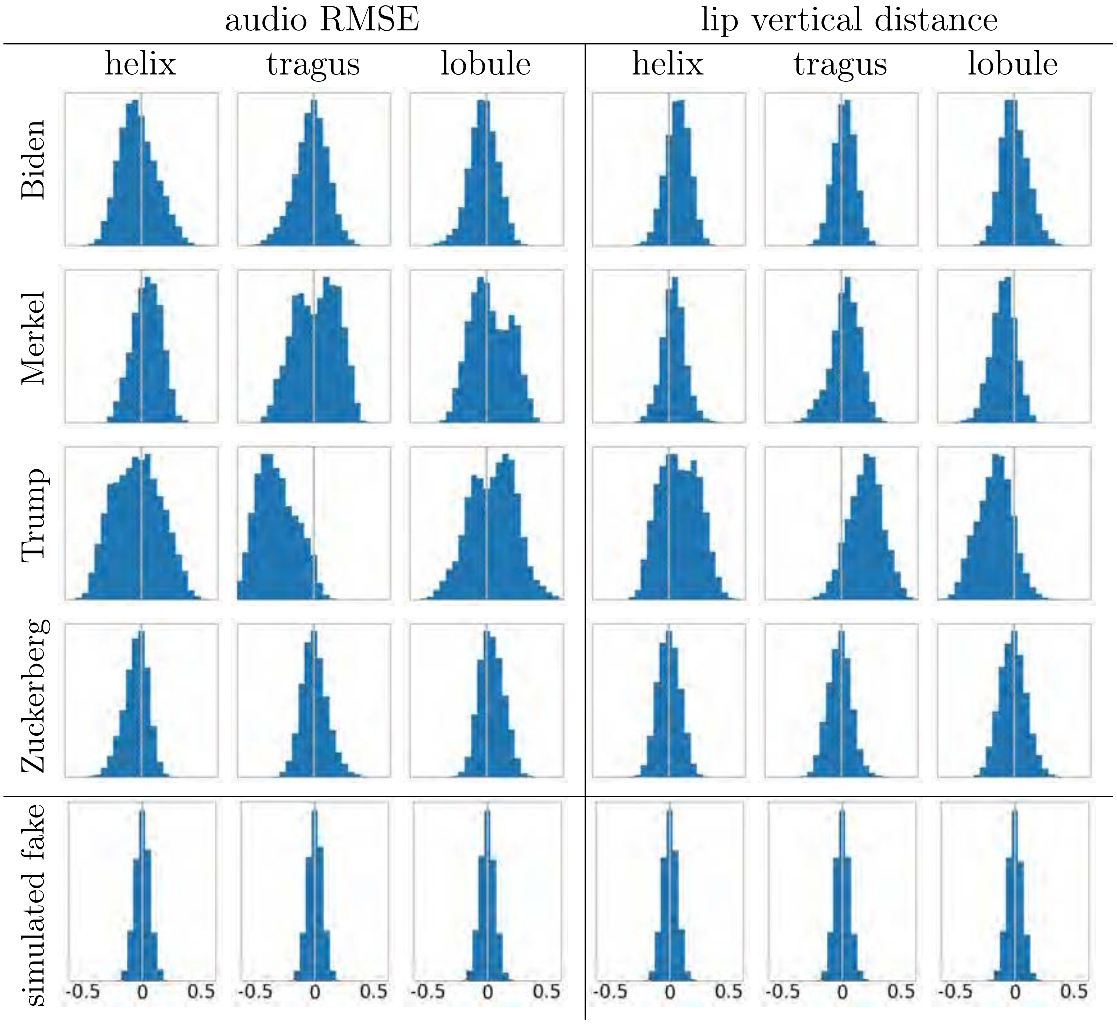


Figure 4.5: Shown are the distribution of correlations between audio (left) and lip vertical distance (right) and the *vertical* motion of three aural areas. From top to bottom are the results for four individuals, and simulated fakes. While the fakes have no correlation, we see strong, but not necessarily consistent, correlations across individuals.

4.3 Results

4.3.1 Aural Dynamics

For all Biden, Merkel, Trump, and Zuckerberg videos, the distribution of audio, facial, and aural correlations are shown in Figures 4.4 and 4.5. These correlations are computed between the horizontal (Figure 4.4) and vertical (Figure 4.5) motion in the helix, tragus, or lobule with the facial (lip vertical) or audio (RMSE) signal. Shown in the last row of these figures are the correlations for simulated fakes in which the aural movements from one video segment are paired to the oral signal from a randomly selected video segment.

The nature of the correlations are somewhat person-specific. For horizontal aural motion, for example, the tragus motion is strongly positively correlated with audio for Trump, but weakly negatively correlated for Biden, Merkel, and Zuckerberg. Similarly, the horizontal lobule motion is strongly negatively correlated for Trump, but not for the others. Additionally, the horizontal tragus motion is positively correlated to the lip vertical distance for Biden, Merkel, and Zuckerberg, but not Trump. For vertical aural motion, this basic pattern continues. The tragus motion is strongly negatively correlated with audio for Trump, but not for the others.

By comparison, in all cases, the simulated fakes (last row of Figures 4.4 and 4.5), we see a complete lack of correlation between these aural and oral signals.

In order to evaluate the efficacy of these dynamic aural features to detect lip-sync deep fakes, a linear classifier is trained as follows. For each individual, the available videos are split into non-overlapping, 80%/20% training and testing sets. A logistic regression model is trained on the 12 aural/oral correlations for the original videos and simulated-fake videos. This model was then evaluated on the testing original videos, and all three types of fake videos: simulated, GAN-generated, and in-the-wild.

Shown in Table 4.2 (top), is the average accuracy reported as the area under the curve (AUC) for 20 random training/testing splits. The average training AUC is 0.91, and the average testing AUC is 0.84, ranging from a low of 0.70 for the Trump in-the-wild fakes, to a high of 0.98 for the GAN-generated Trump fakes. This predictor was trained on all four identities. As we saw in Figures 4.4 and 4.5, however, the nature of the correlations is somewhat person-specific.

Shown in Table 4.2 (bottom) are the results of training four separate logistic regression models, trained on original and fake videos from one individual with, again, 20 random training/testing splits. With this person-specific training, the average training AUC increases from 0.91 to 0.98, and the average testing accuracy increases from 0.84 to 0.87.

Despite only analyzing short 10-second segments, overall accuracy is fairly high. This accuracy can be improved by integrating over an entire video with a simple majority rule.

4.3.2 Aural Biometrics

We demonstrate the use of our aural shape features on the TikTok viral, deep-fake videos of Tom Cruise created by @deeptomcruise [89]. We collected 13 images of Cruise from various internet sources where the left or right ear was visible. Although it has been shown that the left and right ears exhibit some symmetry, there also exist some asymmetries [90]. Despite these asymmetries, we compare all ears to a single right-ear reference, Figure 4.7(c).

Shown in Figure 4.7 (top), is the comparison of all 12 ears to the reference ear. The

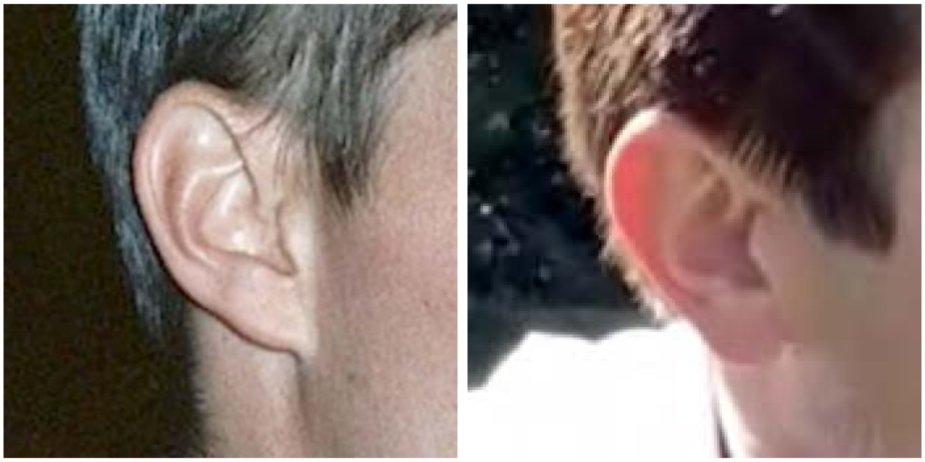


Figure 4.6: The right ear of the real Tom Cruise (left) and @deeptomcruise of TikTok fame [89] (right), from which we see significant differences to the overall shape and earlobe connectivity to the upper jaw.

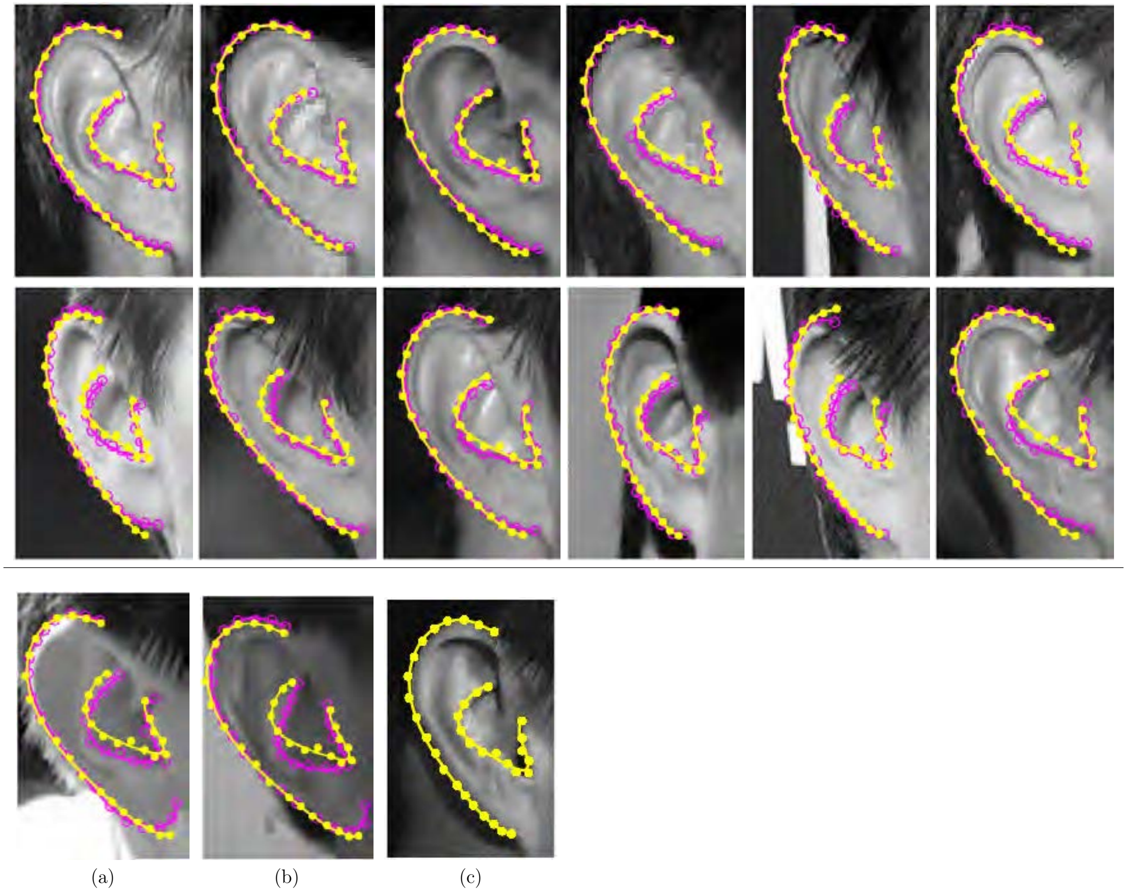
average difference, as described in Section 4.2.3, across all ears is 0.28 with a minimum and maximum difference of 0.19 and 0.37 (these differences are unitless because the aural landmarks are normalized into a range of $[-1, 1]$).

By comparison, shown in Figure 4.7(a-b) is a comparison between the @deeptomcruise ears and the reference ear. Here, the shape difference is 0.51 and 0.58, more than 35% larger than the largest difference across authentic ears.

4.4 Discussion

TikTok's @deeptomcruise recently produced what is arguably some of the most compelling and sophisticated deep-fake videos to date [89]. Beyond the excellent face-swap synthesis, these videos also benefit from a talented performer who resembles the real Tom Cruise and is capable of imitating his mannerisms and voice. The impersonator, however, left behind biometric clues to his identity: the shape and structure of the ears and – not discussed here, but worthy of further investigation – distinct characteristics of the hands. Because deep-fake synthesis has understandably focused on the face, these additional biometric signals should prove a useful addition to the forensic analyst's toolkit.

Our methodology of exploiting aural biometrics and aural and oral correlations are, however, not without limitations. Long hair, for example, will impede any measurements of the shape or dynamics of the ear; large head movements make tracking and aural motion estimation challenging; large head movements may bring the ear into and out of view; the static biometric analysis requires a reference ear of the individual in question; and the dynamic aural motion analysis is most effective with video of the individual in question.



(a) (b) (c)

Figure 4.7: Twelve images of Tom Cruise's ear, two images of @deeptomcruise's ear (a-b), and a reference ear for the real Cruise (c). The yellow annotation (filled circle) corresponds to the shape of the reference ear, and the magenta (open circle) corresponds to the comparison ear's shape. A total of 20 landmarks are annotated from the helix to lobule, and 15 along the tragus, to which two Bezier curves are fit.

Lastly, accurate tracking of the ears has proven to be challenging, requiring some human assistance to correct for tracking slippage. Our approach would benefit from more robust tracking.

A benefit of our dynamic aural and oral analysis is the measured signal unfolds over hundreds of frames, whereas current synthesis techniques typically operate on one or only a few video frames. In addition to the two oral correlations explored here (mouth movement and audio), other facial and audio signals can be exploited including raised eyebrows, smiling, frowning, and audio pitch.

More generally, focusing on high-level, soft- and hard-biometric signals such as the ear, hand, mannerisms, and iris provide a rich forensic signal, striking at the heart of all forms of deep fakes that simply don't depict the person they purport to be.

Chapter 5

Detecting Deep Fakes from Phoneme-Viseme Mismatches

Detection of deep fakes with only a small spatial and temporal manipulation is particularly challenging. We describe a technique to detect such manipulated videos by exploiting the fact that the dynamics of the mouth shape – visemes – are occasionally inconsistent with a spoken phoneme. We focus on the visemes associated with words having the sound `M` (mama), `B` (baba), or `P` (papa) in which the mouth must completely close in order to pronounce these phonemes. We observe that this is not the case in many deep-fake videos. Such phoneme-viseme mismatches can, therefore, be used to detect even spatially small and temporally localized manipulations. We demonstrate the efficacy and robustness of this approach to detect different types of deep-fake videos, including in-the-wild deep fakes [1].

5.1 Introduction

The biometric-based techniques discussed in the previous chapters struggle to detect the lip-sync deep fakes. Part of the reason is that lip-sync deep fakes perform only partial face manipulation, preserving the facial identity of the person. Additionally, some of the lip-sync techniques only modify a single word, or short sentences instead of a longer sequence required for building behavioral biometrics. As a result, most of the biometric signals are preserved in these types of fakes, making it difficult for biometric-based detection techniques to detect them. As these fakes manipulate only a small region of the face (only lower half of the face), leading to fewer creation artifacts making the fakes difficult to detect even by pixel-based detectors. Many examples of lip-sync deep fakes of leaders, including the lip-sync deep fake of Obama with Jordan Peele's voice, have garnered attention of the people and researchers towards the realism of this type of deep fakes.

[1] This work was first published as *Detecting deep-fake videos from phoneme-viseme mismatches* in CVPRW, 2020 [91]

	video (count)	video (seconds)	MBP (count)
original	79	$2,226$	$1,582$
A2V [24]	111	$3,552$	$2,323$
T2V-L [26]	59	308	166
T2V-S [26]	24	156	57
in-the-wild	4	87	66

Table 5.1: The number of videos, duration of videos, and total number of visemes MBP for each dataset.

We describe a forensic technique for detecting lip-sync deep fakes, focusing on high-level techniques in order to be robust to a range of different synthesis techniques and to be more robust to intentional or unintentional laundering. Our technique exploits the fact that, although lip-sync deep fakes are often highly compelling, the dynamics of the mouth shape – so-called visemes – are occasionally inconsistent with a spoken phoneme. Try, for example, to say a word that begins with M, B, or P – mother, brother, parent – and you will notice that your lips have to completely close. If you are not a ventriloquist, you will have trouble properly enunciating “mother" without closing your lips. We observe that this type of phoneme to viseme mapping is occasionally violated, even if it is not immediately apparent upon casual inspection. We describe how these inconsistencies can be leveraged to detect audio-based and text-based lip-sync deep fakes and evaluate this technique on videos of our creation as well as in-the-wild deep fakes.

We start by introducing the concept of various phoneme and visemes pairing, followed by the description of how they are used in our technique. We then present the deep-fake detection results and conclude the chapter with future discussions.

5.2 Methods

5.2.1 Datasets

We analyse lip-sync deep fakes created using three synthesis techniques, Audio-to-Video [24] (A2V) and Text-to-Video [26] in which only short utterances are manipulated (T2V-S), and Text-to-Video in which longer utterances are manipulated (T2V-L). The A2V synthesis technique takes as input a video of a person speaking and a new audio recording, and synthesizes a new video in which the person's mouth is synchronized with the new audio. The T2V synthesis techniques take as input a video of a person speaking and the desired text to be spoken, and synthesize a new video in which the person's mouth is synchronized with the new words. The videos in the T2V-S dataset are taken directly from the original publication [26]. The videos in the T2V-L dataset are generated using

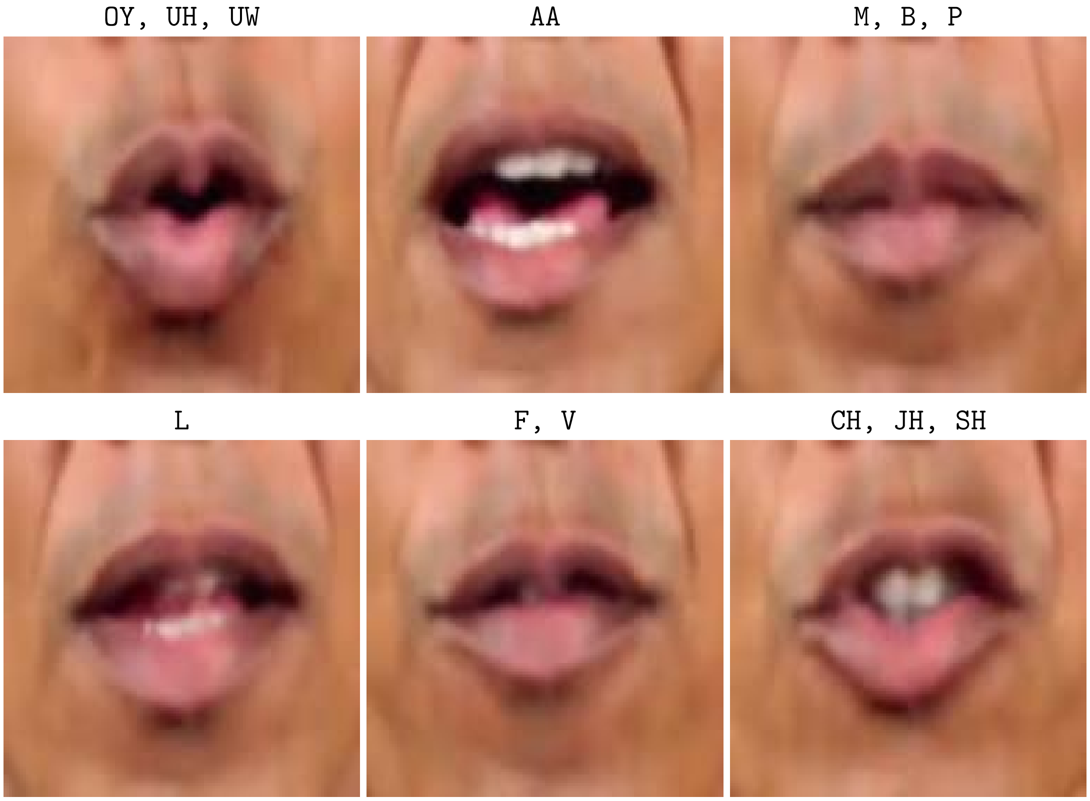


Figure 5.1: Six example visemes and their corresponding phonemes. The phonemes in the top-right (`M, B, P`), for example, correspond to the sound you make when you say “mother", “brother", or “parent". To make this sound, you must tightly press your lips together, leading to the shown viseme.

the implementation of [26] generalized from short to longer utterances. We also apply our analysis to four in-the-wild lip-sync deep fakes downloaded from Instagram and YouTube[2].

For each lip-sync video, we also collected, when available, the original video that was used to create the fake. For each video, the face in each frame was localized, aligned, and cropped (to 256 × 256 pixels) using OpenFace [54], and resaved at a frame-rate of 30 fps. Shown in Table 5.1 are the count and duration (in seconds) of the lip-sync and original videos in our testing dataset.

[2] https://www.instagram.com/bill_posters_uk and https://youtu.be/VWMEDacz3L4

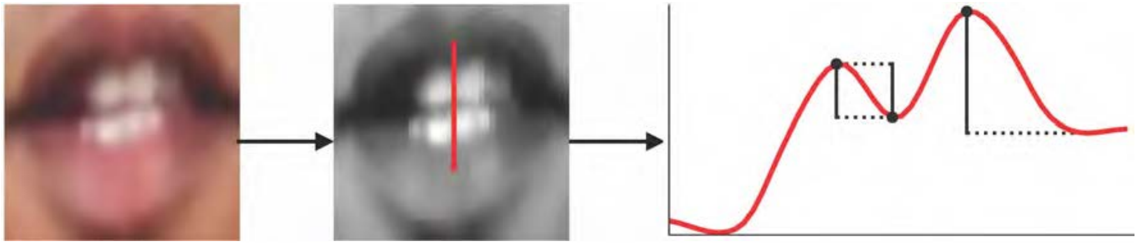


Figure 5.2: Overview of the profile feature extraction used to measure the mouth-closed viseme. The input image is first converted to grayscale and a vertical intensity profile is extracted from the center of the mouth. Shown on the right is the intensity profile with the location of local minima and maxima (black dots) and their corresponding prominences measured as the height, denoted by the dashed horizontal lines, relative to a neighboring minima/maxima.

5.2.2 Phonemes and Visemes

In spoken language, phonemes are perceptually distinct units of sound. A viseme, the visual counterpart of a phoneme, corresponds to the mouth shape needed to enunciate a phoneme. Shown in Figure 5.1 are a subset of six visemes with their corresponding phonemes (a single viseme may correspond to more than one phoneme) [92].

In order to pronounce chair (`CH`), jar (`JH`), or shelf (`SH`), for example, you need to bring your teeth close together and move your lips forward and round them, causing the teeth to be visible through the open mouth. Whereas, in order to pronounce toy (`OY`), open (`UH`), or row (`UW`), the lips again need to be rounded but the teeth are not brought together and therefore not visible through the open mouth. The phoneme group of `M` (mother), `B` (brother), and `P` (parent), on the other hand, requires the mouth to be completely closed for the pronunciation.

The specific shape of various visemes may depend on other speech characteristics like emphasis or volume. The `M, B, P` phoneme group (`MBP`), however, always requires the mouth to be completely closed regardless of other speech characteristics (with the exception of ventriloquists). We focus, therefore, our analysis on this consistent phoneme/viseme mapping.

5.2.3 Extracting Phonemes

In order to analyse a viseme during a spoken `MBP` phoneme, we first extract the location of all phonemes as follows. Google's Speech-to-Text API [93] is used to automatically transcribe the audio track associated with a video. The transcription is manually checked to remove any errors and then aligned to the audio using P2FA [94]. This alignment generates a sequence of phonemes along with their start and end time in the input audio/video. Here, only the `MBP` phonemes will be considered. Shown in the last column of Table 5.1 are the number of MBP phoneme occurrences extracted for each dataset.

5.2.4 Measuring Visemes (manual)

For a given MBP occurrence, the associated viseme is searched in six video frames around the start of the occurrence. We consider multiple frames to adjust for small phoneme to audio alignment errors. Only the frames around the start of the occurrence are analysed because the mouth should be closed before the MBP phoneme sound is made.

Given six frames for an MBP occurrence, we take three approaches to determine if the expected mouth-close viseme is present in any of the frames. The first approach is purely manual where an analyst is presented with six video frames and a reference frame from the same video where the mouth is clearly closed. The analyst is then asked to label each presented sequence as “open" or “closed." A closed sequence is one in which the mouth is completely closed for at least one video frame. This approach provides the ground-truth for an automatic computational approach to determining if the mouth shape associated with a MBP phoneme is open or closed. This type of manual analysis might also be applicable in one-off, high-stakes analyses.

5.2.5 Measuring Visemes (profile)

In the second approach, a mouth-close viseme is automatically detected in any of the six frames centered around an MBP occurrence. For each frame, the lip region is extracted from 68 facial landmarks [82]. The extracted lip region is rescaled to 50×50 pixels and converted from RGB to grayscale. A vertical intensity profile is then extracted from the middle of the mouth (Figure 5.2). We expect this intensity profile to be qualitatively different when the mouth is open or closed. Shown in the top middle panel of Figure 5.1, for example, is a mouth open in which the vertical intensity profile will change from skin tone to bright (teeth), to dark (the back of the mouth), to bright (teeth), and then back to skin tone. In contrast shown in the top right panel of Figure 5.1, is a mouth closed in which the vertical intensity will be largely uniform skin tone.

The overall profile shape is quantified by computing the sum of the prominences of the local minima, l, and maxima, h, in the intensity profile (as determined using MATLAB's `findpeaks` function, with the default parameters), Figure 5.2. The measurements l and h capture how much the intensity along the profile decreases (e.g., when the back of the mouth is visible) and increases (e.g., when the teeth are visible). These measuremtns are made for each of the six frames, l_i and h_i, $i \in [1, 6]$, and compared to the reference measurements l_r and h_r in which the mouth is closed, Figure 5.3. The measure of similarity to a reference frame in the six-frame sequence is the minimum of $(|l_i - l_r| + |h_i - h_r|)$, $i \in [1, 6]$.

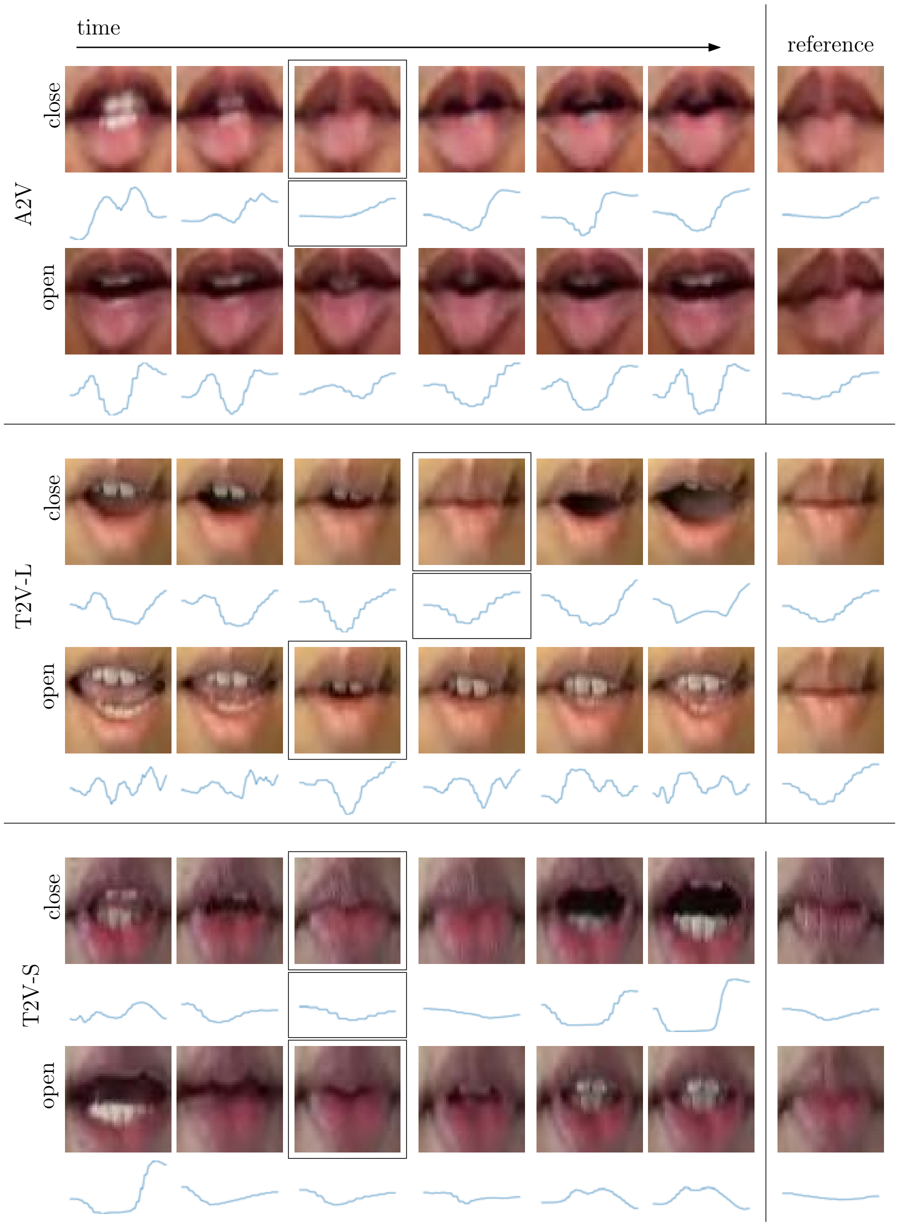


Figure 5.3: Six sequential frames extracted from a single MBP occurrence in different deep-fake videos. Shown on the right is a reference frame where the mouth is clearly closed. Shown below each frame is a 1-D intensity profile used to automatically classify the mouth as open or close. The bounding box corresponds to a frame that matched the reference frame shown to the right (only the closed-mouth sequences match).

dataset	correct	incorrect	total
original	0.709	0.001	0.710
A2V	0.49	0.15	0.64
T2V-L	0.09	0.43	0.52
T2V-S	0.26	0.11	0.37
in-the-wild	0.64	0.05	0.69

Table 5.2: The average number of correct, incorrect, and total viseme occurrences/second of video.

5.2.6 Measuring Visemes (CNN)

In a third approach, we explored if a more modern learning-based approach can outperform the hand-crafted profile feature. Specifically, we trained a convolutional neural network (CNN) to classify if a mouth is open or closed in a single video frame. The input to the network is a color image cropped around the mouth and rescaled to a 128×128 pixels (Figure 5.1). The output, c, of the network is real-valued number in $[0, 1]$ corresponding to an “open" (0) or “closed" (1) mouth. The open/closed classification in a six-frame sequence is the maximum of $c_i,\ i \in [1, 6]$.

The network is trained using videos of Barack Obama for whom the lip-sync deep fakes were created in the A2V dataset. This training dataset consists of original videos disjoint from the testing videos reported in Table 5.1. In total, we manually labelled $15,600$ video frames where the mouth is open (8258 instances) or closed (7342 instances). In each frame, OpenFace [54] is used to automatically detect, scale, and rotate (in-plane) the face to a normalized pose and resolution.

The Xception architecture [58] is used to train a classifier using 90%/10% images for training/validation. The network is trained for 50,000 iterations with a mini-batch of size 64. In each mini-batch, equal number of images were randomly sampled from each label. The initial learning rate of 0.01 was reduced twice at iterations 20,000 and 40,000. The weights were optimized using Adam optimizer and a cross-entropy loss function.

5.2.7 Global Audio-to-Video Alignment

We previously used P2FA to ensure that the phonemes were correctly synchronized with the underlying audio. Here we also ensure that the audio is correctly synchronized with the underlying video. This audio-to-video alignment is done through a brute-force search of the global shift in the audio (in the range $[-1, 1]$ seconds, in steps of $1/10$ seconds) that creates the best agreement between all `MBP` phonemes and the correct mouth-closed viseme. This alignment contends with slight audio to video desynchronization that might occur from transcoding or innocuous video editing.

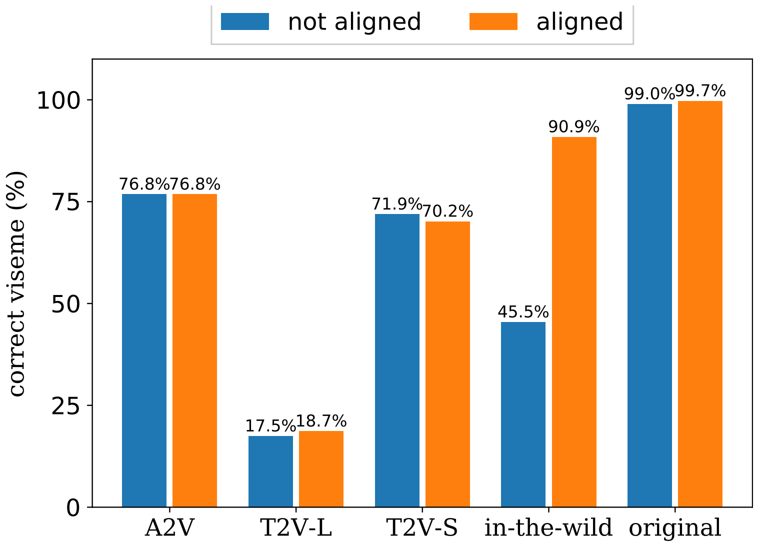


Figure 5.4: The number of correct MBP phoneme to viseme pairings before (blue) and after (orange) audio to video alignment. The T2V-L lip-sync deep fakes are the least well matched, while the (aligned) in-the-wild deep fakes are correctly matched more than 90% of the time.

5.3 Results

Detecting Deep Fakes (manually): We evaluate the efficacy of detecting deep fakes first by using the manual annotation for determining if the phoneme and viseme pairing is correct. Shown in Figure 5.4 are the percent of MBP phoneme occurrences where the correct viseme is observed. For each dataset, the percent is reported before (blue) and after (orange) the global audio to video alignment. The problem of misalignment is most salient for in-the-wild videos where before alignment only 45.5% of the visemes were correct, as compared to 90.9% after alignment. For each of the other datasets, misalignment was not an issue.

For the four deep-fake data sets (A2V, T2V-S, T2V-L, in-the-wild), the percentage of correct phoneme to viseme pairing (after alignment) ranges from a high of 90.9% of 66 occurrences (in-the-wild), to 76.8% of 2,323 occurrences (A2V), and 70.2% of 57 occurrences (T2V-S), and 18.7% of 166 occurrences (T2V-L). The phoneme to viseme pairing in original videos is correct for 99.7% of 1,582 occurrences (the small number of errors are due either to manual annotation or transcription error).

Shown in Table 5.2 is the rate (per second) at which MBP phonemes occur (*total* column) and the rate at which phoneme-viseme mismatches occur (*incorrect* column). The rate of spoken MBP phonemes varies from 0.71 (original) to 0.37 (T2V-S), and so it is important to compare to the appropriate base rate when considering overall accuracy.

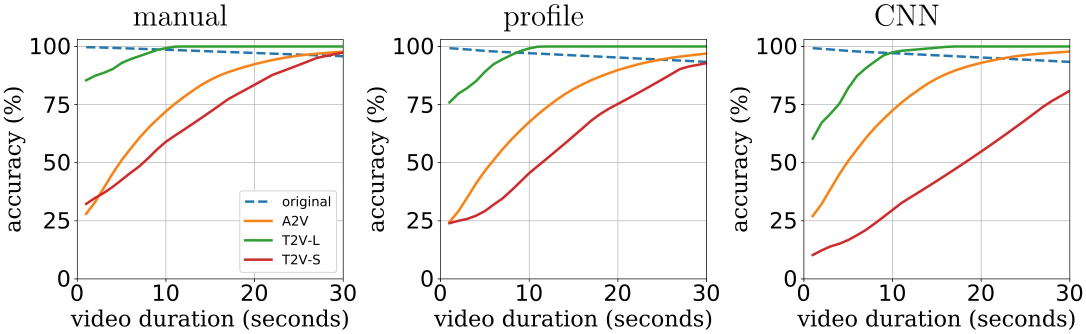


Figure 5.5: Shown in each panel is the accuracy with which lip-sync deep fakes are detected using mismatched `MBP` phoneme to viseme pairings. Each solid curve (orange, green, red, and purple) corresponds to a different deep-fake dataset and the dashed curve (blue) corresponds to the original dataset. Each panel corresponds to a different technique for determining if a mouth is open or closed. Detection accuracy improves steadily as the length of the video increases from 1 to 30 seconds.

Even a relatively low number of say 10% incorrect phoneme to viseme pairings can, over time, lead to an effective detection strategy. In particular, shown in the left-most panel of Figure 5.5 is the percent of videos that are correctly identified as fake as a function of video duration, from 1 to 30 seconds. A video is detected as fake if the number of incorrect phoneme to viseme mismatches exceeds the expected mismatch of 0.3% found in original video (Figure 5.4. As expected, the detection accuracy increases as the video length increases. At a length of 30 seconds, for example, nearly all of the A2V, T2V-L, and T2V-S videos are classified correctly, while only 4% of original videos are misclassified.

Detecting Deep Fakes (automatically): We next evaluate the accuracy of automatically determining if a mouth is open or closed and how these automatic classifications impact the accuracy of detecting a video as real or fake. Throughout, the manual annotation described above are used as ground truth.

Shown in Table 5.3 is the accuracy of the two automatic techniques (profile and CNN) to detect if a mouth is open or closed. Each classifier was configured to have an average false alarm rate of 0.5% (i.e., misclassifying a closed mouth as open). The performance of both the profile and CNN techniques are high on the A2V dataset with an average accuracy above 96%. On the T2V-L and T2V-S datasets, however, the profile technique performs better than the CNN which was only trained on videos of Barack Obama (somewhat surprisingly, however, the CNN generalizes to the in-the-wild videos).

Shown in the central and right-most panel of Figure 5.5 is the video detection accuracy when the manual annotation of mouth open or closed is replaced with the automatic detection based on intensity profiles (center) and CNN classification (right). Using the

dataset	profile	CNN
original	99.4%	99.6%
A2V	96.6%	96.9%
T2V-L	83.7%	71.1%
T2V-S	89.5%	80.7%
in-the-wild	93.9%	97.0%

Table 5.3: The accuracy of the two automatic techniques (profile and CNN) to detect if a mouth is open or closed. The accuracies are computed at a fixed threshold corresponding to average false alarm rate of 0.5% (i.e., misclassifying a closed mouth as open).

profile technique, the video detection accuracy is only slightly degraded as compared to the manual annotation (left-most panel): at 30 seconds, for example, the manual annotation has an accuracy on the original, A2V, and T2V-S datasets of 96.0%, 97.8%, and 97.4%, as compared to the automatic profile technique with an accuracy of 93.4%, 97.0%, and 92.8%.

For the CNN technique, the video detection accuracy for the original and A2V datasets remains comparable to the manual and profile annotations: at 30 seconds, the accuracy on the original and A2V datasets is 93.4% and 97.8%. For the T2V-S dataset, however, the accuracy drops from 97.4% to 81.0%. This is because the CNN was trained only on videos of Barack Obama exclusively in the A2V dataset, and thus does not generalize well to different people in the T2V-S dataset. We hypothesize that this accuracy can be improved by training a CNN with different people.

Failures: Shown in Figure 5.7 are two six-frame sequences where the profile technique misclassified a closed mouth as open (top) and an open mouth as closed (bottom). The first failure is because the shape of the lips is different from the reference frame. The second failure is because the mouth is asymmetrically open. While these failure cases are somewhat inevitable when using automatic techniques, they are easily flagged by a manual annotator.

Robustness: We next examine the robustness of the two automatic detection techniques against two simple laundering operations, recompression and resizing. Each video was laundered using `ffmpeg` by: (1) reencoding at a lower quality of qp=40 (typical videos are encoded at higher quality of qp $\in [10, 20]$); or (2) resizing to half-resolution and scaling back to the original resolution (effectively, blurring each video frame). The average accuracy of the profile and CNN technique in detecting open or closed mouth after recompression is 90.46% and 88.32%. The average accuracy of the profile and CNN technique after resizing is 83.80% and 89.92%.

Resizing has a significant impact on accuracy for the profile technique. This is because resizing reduces the prominence of the local minima and maxima. As a result, the open

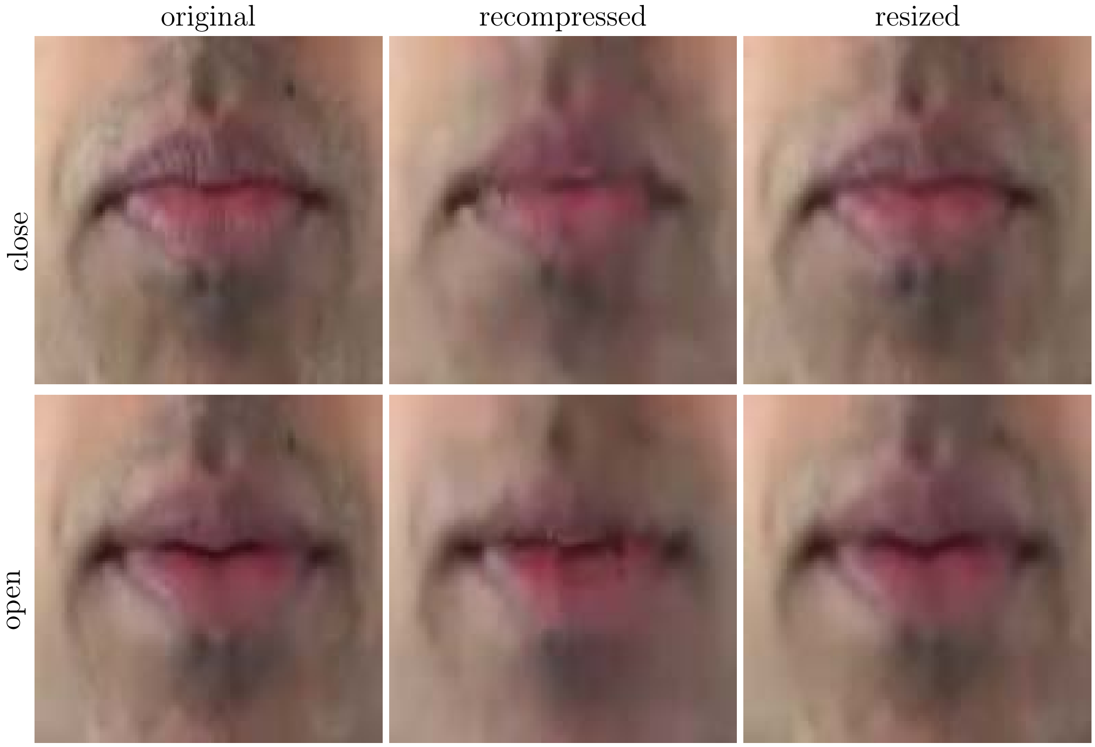


Figure 5.6: Shown is a closed (top) and open (bottom) mouth before (first column) and after recompression (second column) and after resizing (third column). Although our automatic techniques correctly classified the closed-mouth, they misclassified as closed the recompressed and resized open mouth. A human analyst can, however, still identify the small opening between the lips even after recompression or resizing.

mouth are more likely to be mis-classified as closed. For such low quality videos, therefore, manual annotation can be more robust than the automatic detection (Figure 5.6).

5.4 Discussion

We described a forensic technique that uses phoneme-viseme mismatches to detect deep-fake videos. Our main insight is that while many visemes can vary, the sounds associated with the `M`, `B`, and `P` phonemes require complete mouth closure, which is often not synthesized correctly in deep-fake videos. For high-stakes cases, we show that an analyst can manually verify video authenticity. For large-scale applications, we show the efficacy of two automatic approaches: one using hand-crafted features that requires no large training data, and one using a CNN.

While we had good reason to look only at `MBP` phonemes, we believe that including all visemes in the analysis will improve results even further. This extension, however, is not trivial and will require modeling the possible variance of each viseme and co-articulation. It will, however, allow us to use a larger portion of a video for analysis, ultimately leading to better detection.

Our CNN results, trained only on videos of Barack Obama, are person specific and perform much better on videos of Obama. We expect better results using a network that is trained on a large corpus of people. Obtaining such a large labelled dataset is challenging — especially since we care mostly about the hard cases in which a mouth is almost closed or open, with just a few pixel difference. Such labels currently cannot be accurately extracted from face landmark detectors. Thus, it would be beneficial to develop unsupervised methods to automatically differentiate between complete and almost complete mouth closure.

Even with these limitations, our method can already detect state-of-the-art, lip-sync deep fakes. We expect future synthesis techniques to continue the cat-and-mouse game, taking into more careful account the phoneme to viseme matching. We view deep-fake detection using phoneme-viseme mismatches as one more tool in the forensic expert toolkit, to be developed and used together with other complementary techniques.

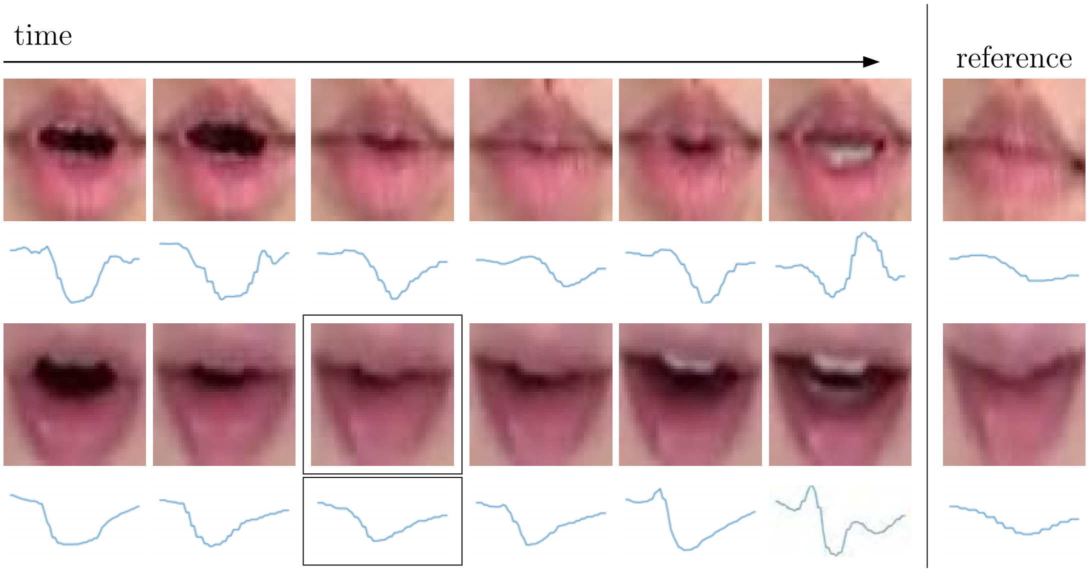


Figure 5.7: Shown are two six-frame sequences where the automatic profile technique failed to correctly classify the mouth as open or closed. The mouth in the upper sequence was incorrectly classified as open, whereas in the lower sequence, the mouth was incorrectly classified as closed. Shown on the far right is the reference frame and shown below each frame is the intensity profile used for classification.

Chapter 6

Perceptual Detection of Faces

In today's digital world, facial images are used as a proof of identity at many venues – from social media accounts to national identification documents. With recent advances in techniques for face synthesis and manipulations, it is becoming easier to create fake identities using synthetic faces. In this chapter, we perform perceptual studies to determine how well humans can detect original faces from synthetic and manipulated faces. We use StyleGAN2 and traditional face morphing to create two high-quality synthetic-face datasets, with a diverse set of people across gender, race, and age. For both techniques, we examine people's ability to detect synthetic faces with and without the presence of feedback and training. We show that human participants struggle in all perceptual tasks, supporting the need for effective computational solutions to better protect us from fraudulent identities [1].

6.1 Introduction

With the advances in artificial intelligence tools, specifically generative adversarial networks (GANs), synthetic faces have become more and more indistinguishable from real faces. The existence and ease-of-use of such technology has consequences across almost all domains, from law enforcement and national security through to politics, media, and entertainment. There has been multiple instances of fraudulent social media accounts created using GAN faces, including the creation of a fictional candidate for U.S. Congress [96]. Such instances have generated a huge concerns among policy makers, research communities, and general public. Many computational techniques have been proposed in order to detect GAN images [38], which has shown poor generalizability and scalability in the real-world scenario [97]. As a result we rely, more often than not, on people's ability to distinguish between original and GAN faces.

[1] The work on morphed faces was first published as *Perceptual and computational detection of face morphing* in JOV, 2021 [95]

Given our reliance on people's judgments about face authenticity, it is important to better understand human ability to detect GAN generated faces. Even though the faces generated with state-of-the-art StyleGAN [9] and StyleGAN2 [10] architectures look remarkably real, there is a lack of a formal perceptual study to prove the same. Here, we try to bridge this gap by analyzing people's ability to recognize high-quality GAN faces generated using StyleGAN2. We perform a large-scale perceptual study to answer the following: 1) how well humans can discriminate between real faces and GAN faces; 2) can we train humans to recognize common artifacts in GAN faces in order to achieve better discriminative performance.

Another relatively new type of identity theft uses morphed facial images in identification documents in which images of two individuals are digitally blended to create an image that maintains a likeness to each of the original identities. We frequently rely on photo-based identity documents to verify identity in critical settings such as border control. Much research, however, has shown that matching pairs of unfamiliar faces is a difficult task [98–100], including for trained identification-checkers [101]. The difficulty of this task leaves identity verification processes vulnerable to fraudulent attacks. The use of morphed passport photos is a recent type of fraud that border control agencies are facing. Early research exploring human detection of morphed images indicates that people frequently accept both low-quality and high-quality morphed images as genuine [102–104]. The face databases used in these previous studies, however, have a number of limitations, most notably low-quality morphs and/or limited diversity in terms of the race, gender, and age of the faces used to create the stimuli. Here, we extend this previous literature by examining human and computer-based detection of face morphing using a diverse set of facial images to generate high-quality morphs.

In the next section, we review the related perceptual studies in this domain. Then the rest of the chapter is divided into two broad sections to discuss the datasets, methods, and experimental results for perceptual detection of GAN faces and morphed faces.

6.2 Related Work

Despite the popular belief that StyleGAN2-generated faces are hyper-realistic [105, 106], it has not been formally shown that these images can consistently fool the human visual system. There are, however, two related studies to understand humans' perception of GAN-synthesized facial videos. In [53], the authors asked human subjects to classify 30 real- and 30 synthesized-video frames and demonstrated that humans achieved almost 80% accuracy. The fake videos in this study, however, were created using earlier GAN-based video-synthesis techniques that leave behind noticeable artifacts in the facial region. A more comprehensive study was performed in [107], where the authors hand-picked 60 real and 60 synthesized videos from state-of-the-art DFDC dataset [62]. The synthetic videos spanned various visual artifacts, from easily noticeable to imperceptible ones. The human accuracy was shown to reduce with reduction in visual artifacts, with performance falling below chance for videos with least artifacts.

Even though the previous studies suggest some degree of human ability to detect synthetic videos, it is not sure if this will be the case for StyleGAN2-generated static faces. This is firstly because, the visual artifacts in the state-of-the-art synthetic faces are different and maybe more subtle than the previously analyzed synthetic videos. Secondly, the per-frame manipulation of facial identity or expressions of a person in a video can lead to additional artifacts due to facial motion. These artifacts can provide cues to humans when analyzing a video which are not there when analyzing a static face. Therefore, in this work we aim to obtain better understanding about how difficult it is for humans to detect these high quality synthetic faces.

On the other hand, there has been a number of perceptual studies in the past to analyze the threats of morphed faces. Early research indicates that people frequently accept both low-quality and high-quality morphed images as genuine [102–104]. In [102,103], the authors performed similar perceptual experiments as presented in this paper. There are, however, a number of important limitations with these two studies. First, although the morphs were created using advanced morphing software, there was no manual editing stage to remove obvious artifacts that are known to result from the morphing process, such as the outline of another person's hair. In fact, such artifacts were precisely what the authors guided participants to look for to help them to detect morphs. This limitation might artificially inflate the effect of the guidance and training manipulation. Second, the stimuli were created using facial images from the Glasgow Face Matching Test [100] which includes mostly White individuals. This lack of racial diversity in the stimuli further limits the extent that the results of these studies are likely to be representative of the detection of morphs in the real world.

In 2019, another research group sought to address the first limitation noted above by replicating the study by [103] using higher-quality face morphs [104]. Using these higher-quality morphs, they also found that initial detection was poor, but unlike [103], training had no effect on accuracy. In one experiment, the authors mimicked the real world scenario where the participants completed a live face matching task rather than a computer-based one. For this task, 48 models (44 White) were photographed and paired with a visually similar model (foil). The models approached participants on a university campus and presented either a photo of the 1) model (match), 2) model morphed with the foil individual (50/50 morph), or 3) foil (mismatch). The participants were asked to indicate if they thought the photo was of the model or not. For the match and mismatch conditions, average accuracy was 83% and 84%, however, the morph photos were accepted nearly half of the time (49%). The pairing of models to create the morph photos revealed an interesting finding—for the majority of the pairs, the morph was accepted as a valid identification for one model more frequently than for the other model. [104] conclude that even when generating 50/50 morphs, the morph does not represent each of the original faces equally. Because in previous work [102] the morphs were only presented alongside one of the original identities the results might not accurately represent true morph detection

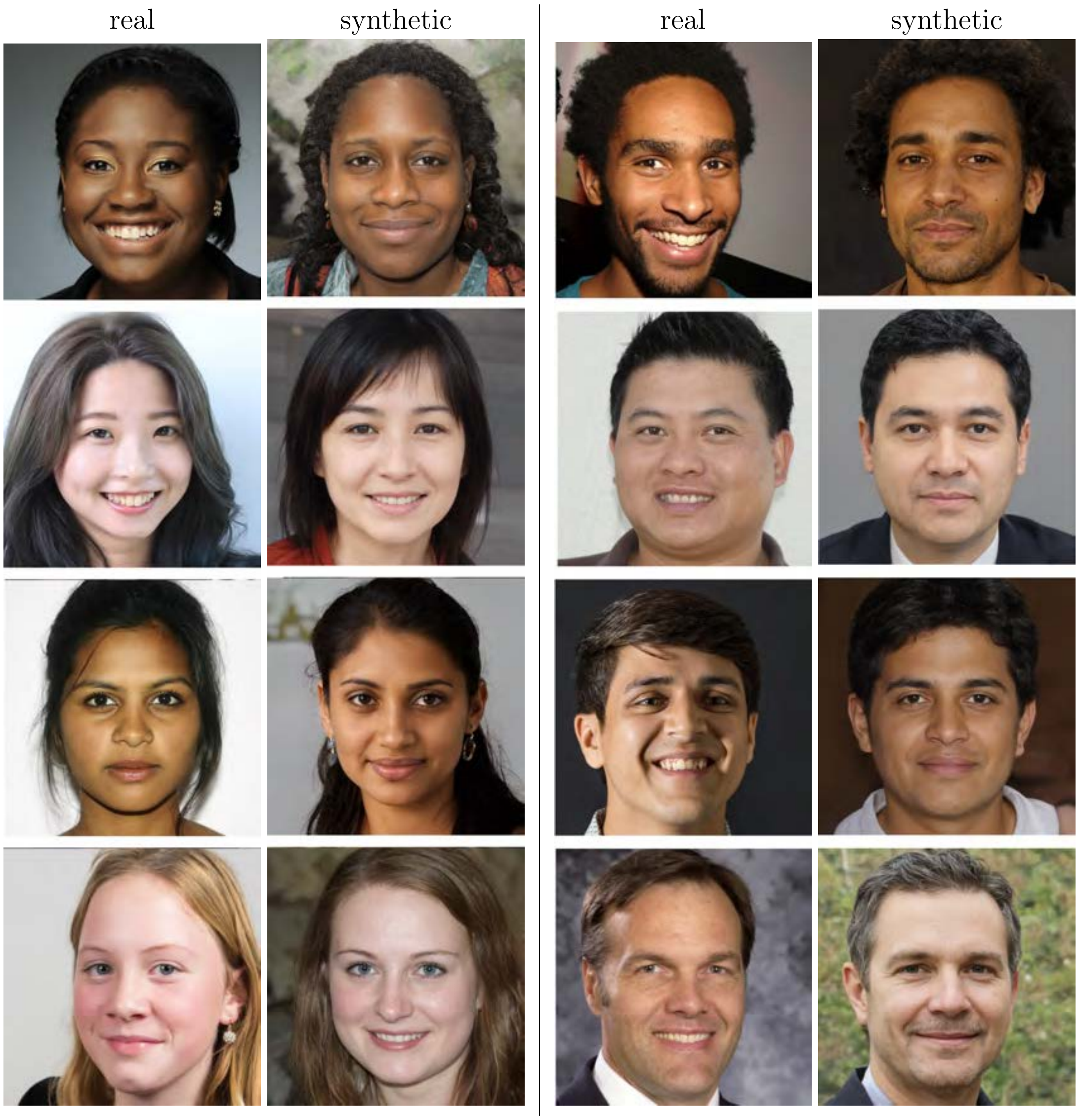


Figure 6.1: Shown above are 16 example faces from real and GAN generated dataset. The images span across different races and genders used in our dataset.

rates.

In [104] the authors also tested whether a simple computational model could outperform human ability to detect morphs. Principal components analysis was used to extract a low-dimensional representation of the faces. These representations were then used to train a linear discriminant analysis model with two classes, morph and original. Using this model to classify the remaining images that were not used in the training set resulted in an average accuracy of 68% corresponding to a sensitivity of 1.01. This result suggests that a simple computational model can outperform humans at morph detection but remains a far from perfect classifier.

Even though the previous studies perform similar experiments as ours, there is one common limitation in all of these previous studies. They used low-quality morphs and/or limited diversity in terms of the race, gender, and age of the faces used to create the stimuli. Here, we extend this previous literature by examining human and computer-based detection of face morphing using a diverse set of facial images to generate high-quality morphs.

6.3 Synthetic Faces

6.3.1 Datasets

We employed state-of-the-art StyleGAN2 architecture to generate 400 high-quality faces of resolution 1024 × 1024. It was ensured that the generated faces didn't have noticeable background artifacts and included a diverse range of gender, age, and race. There were 100 African American, 100 East Asian, 100 South Asian, and 100 Caucasian. Of these 200 were women and 200 were men, spanning a range of apparent ages. In order to minimize facial differences between real and synthetic faces, for every synthetic face we selected the most similar looking real face from Flickr-Faces-HQ (FFHQ) dataset [9]. A standard convolutional neural network descriptor (termed VGG) [67] was used to extract a low-dimensional, perceptually meaningful, representation of each face in the full dataset. The extracted representation – a 4096-D real-valued vector – for each of the 400 synthetic faces was compared with all other representations in the FFHQ dataset to find the most similar face as defined by the face whose representation is most similar – in terms of Euclidean distance – with the synthetic face [108, 109]. Shown in Figure 6.1 are example images from the real and GAN generated face datasets. There was no further post-processing done to the images.

6.3.2 Experiment 1: classification (original or synthetic)

Methods

This experiment was done with 315 participants on Amazon's Mechanical Turk (AMT). The participants self-reported as: 176 men, 137 women, 2 prefer not to say; between 24-73

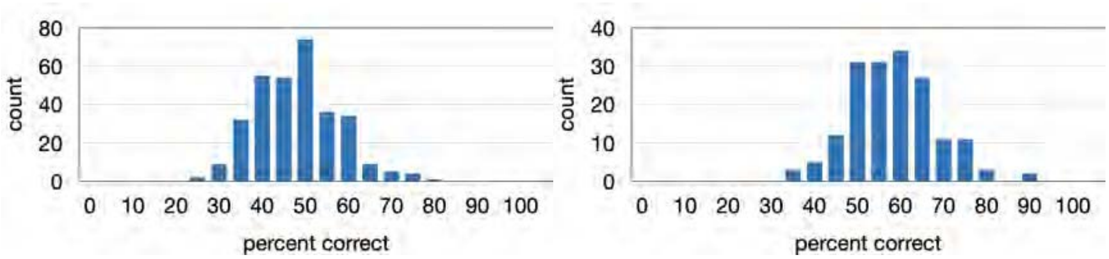


Figure 6.2: Shown on left/right are the distribution of percent accuracies for 315/170 participants in Experiment 1/2 of Section 6.3. The average percent accuracy of participants in Experiment 1 and 2 are 48.0% and 52.0%.

years of age (μ = 40.9; σ = 10.5); 229 White, 32 South Asian, 18 East Asian, and 12 African American. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile.

A within-subject design was employed in which each trial consisted of a single face image (either original or GAN generated). Each participants viewed 128 trials, across which, the faces were equally balanced in terms of gender and race. Although unknown to the participant, half of the faces were real and half were synthesized. On each trial, participants were instructed to specify if the image is a real face or a GAN generated face. In this classification task, chance performance is 50%. We also created 50 attention-check trials, examples of which are shown in Figure 6.3. These trials were intentionally easy comprising of very noticeable synthetic images.

Participants first received task instructions including a brief description about GAN generated faces and then shown a few examples of synthesized and real faces. Participants had an unlimited amount of time to indicate whether the face is real or synthesized.

Following the practice trial, participants completed the 128 trials plus 10 attention check trial. At the end of the session participants were asked a few basic demographic questions.

Results

The average accuracy of identifying whether the shown image is real or not is 48.0% (chance is 50%), corresponding to a sensitivity of $d' = -0.09$ and bias of $\beta = 0.99$, where the bias corresponds to a tendency to label faces as “real", Figure 6.2 (left). The accuracy for real/synthetic faces was 52.0%/45.0% – i.e. participants were not biased towards saying either synthetic or real face.

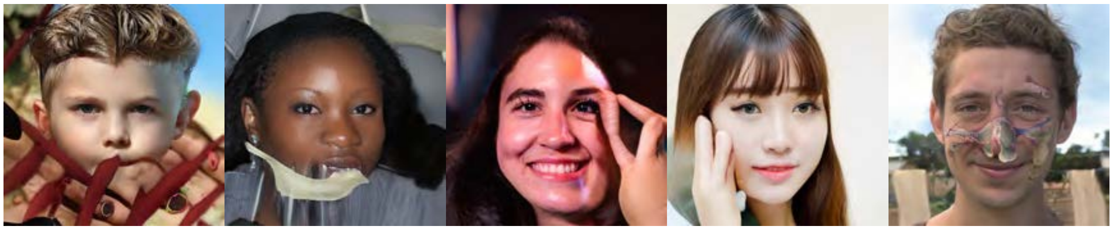


Figure 6.3: Five examples catch trial used during Experiment 1 and 2 to ensure participants are paying attention, see Section 6.3. These images have noticeable unnatural distortions on the face.

Discussion

The results of this experiment suggest images synthesized by StyleGAN2 are realistic enough to fool naive observers. However, another possible explanation for this poor performance can be: the participants are not aware of the common facial artifacts that are left behind by GANs. Therefore, the next experiment is designed to assist participants in spotting these facial clues. We want to understand if such training and feedback can improve the performance.

6.3.3 Experiment 2: classification (original or synthetic) with training and feedback

This experiment was done with 170 participants on AMT. The participants self-reported as: 83 men, 85 women, 2 prefer not to say; between 18-77 years of age ($\mu = 40.7$; $\sigma = 11.6$); 123 White, 21 South Asian, and 11 African American, 7 East Asian, and 2 other/prefer not to say. Participants received \$5 for completing this experiment. As an incentive to encourage effort on the task, a \$5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile. There was no overlap between the participants in this experiment and previous one.

Other than the inclusion of a training session as described below and accuracy feedback after each trial, the design, underlying stimuli, and procedure were identical to that used in the previous experiment.

Before starting the trials, participants were shown some training images that will help them spot mistakes in synthetic faces. The following are the facial artifacts that the participants were asked to notice while performing the trials:

1. **Look at the ears:** Ears might have an unusual shape and they might even be different sizes, shapes, and/or misaligned, Figure 6.4 (a) and (b).

2. **Look at the glasses:** Glasses tend to have thin frames, with end pieces that might not match, Figure 6.4 (c).

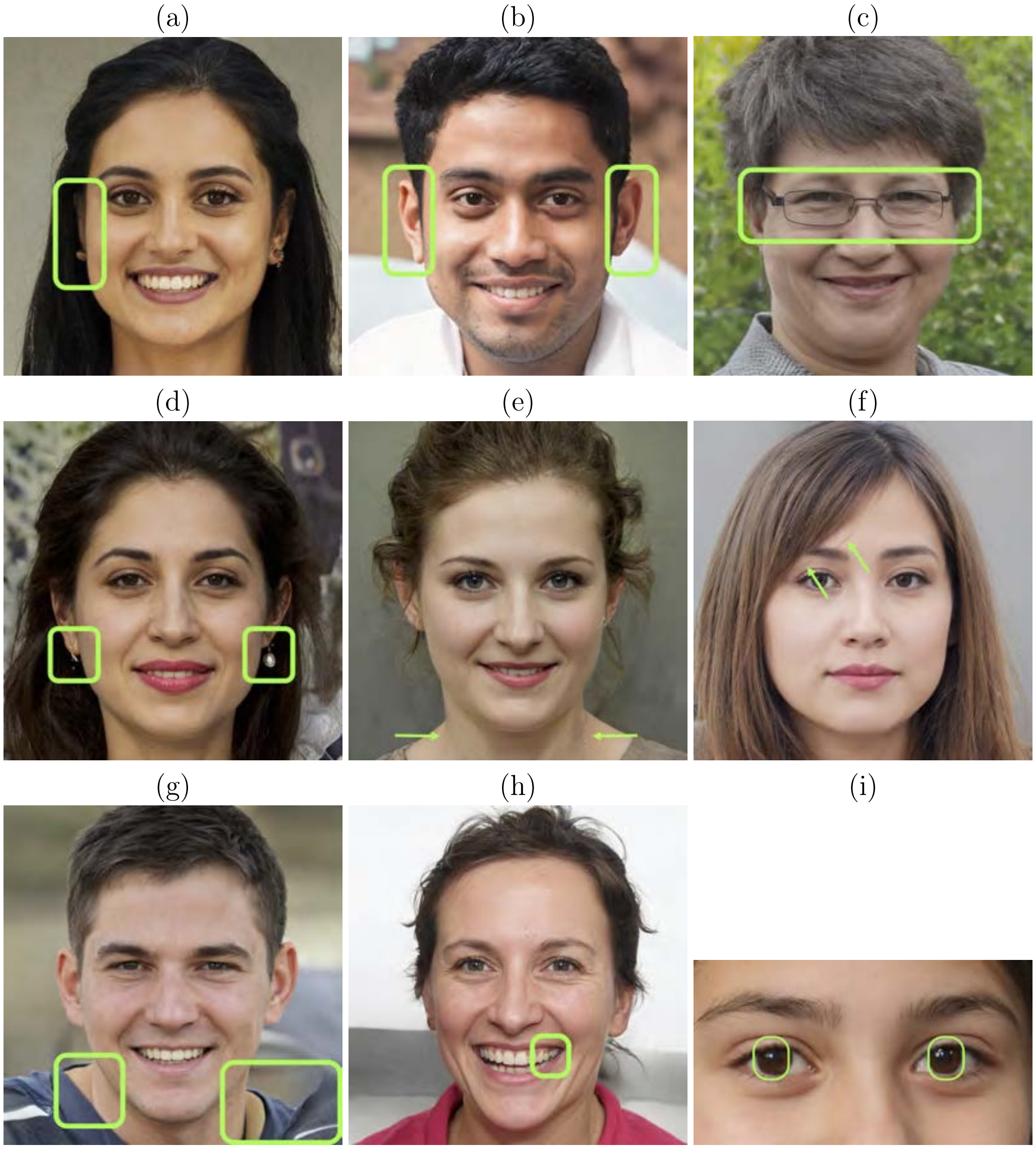


Figure 6.4: Shown are the images used for training participants in Experiment 2, see Section 6.3.3. Each image highlight a specific artifact in StyleGAN2 synthesized faces.

3. **Look at the accessories:** Earrings might not match and necklaces might appear on only part of the neck, Figure 6.4 (d) and (e).
4. **Look at the hair and hairline:** Hair might stick out beyond the outline of the head or appear unnaturally on the face/neck, Figure 6.4 (f).
5. **Look at the neck and shoulders:** Clothing might dissolve into the background. Shoulders and/or neck might be misshapen, Figure 6.4 (g).
6. **Look at the teeth:** Teeth might be an odd shape or color or might blur into each other, Figure 6.4 (h).
7. **Look at the eyes:** Reflections in the eyes might not match sometimes, Figure 6.4 (i).

To check that participants paid attention to the training, they were then asked to select the three artifacts from a list of eleven possible options, where the three incorrect answers were easily identifiable as they were not mentioned in the training. Participants were given the option to view the training session a second time if they were unsure of the correct options. The other change from the previous experiment was that after responding on each trial, participants were provided with feedback indicating whether their response was correct or not.

Results

After training, the average accuracy of identifying a face as synthetic or not was 58.0%, corresponding to a sensitivity of $d' = 0.41$ and bias of $\beta = 0.98$ (compared to previously 48.0%, $d' = 0.09$, $\beta = 0.99$, Figure 6.2 (right). The accuracy for real/synthetic faces were 57.0%/60.0%. The overall accuracy of the participants increase by 17.2% after training.

Discussion

The training and feedback in this experiment helped participants to recognize synthetic faces better. This gives a hope that raising awareness about the common artifacts in GAN synthesized faces can increase human performance to detect fraudulent attacks using GAN faces. However, the accuracy improved only slightly and it is not yet known if the same accuracy will hold for low-resolution and low-quality faces. The poor accuracies the two experiments suggests that people are unable to reliably detect a GAN synthesized face from the real face.

6.4 Morphed Faces

6.4.1 Datasets

We note four main limitations of other datasets that have been used in similar research [102–104]: 1) obvious visual morphing artifacts; 2) a small number of faces from which to select matching faces; 3) a manual process to match similar faces; and 4) a lack of racial/gender diversity. We address these limitations to create a high quality and diverse dataset [2].

We collected 3,500 passport-format facial images from 13 face databases [110–124]. These 3,500 images included a diverse range of individuals across gender, age, and race. To ensure diversity in our final stimulus set, we manually selected 54 individuals constituting 6 African American, 16 East Asian, 16 South Asian, and 16 Caucasian. Of these 26 were women and 28 were men, spanning a range of apparent ages. Some of the face databases specified the race/gender of each face. Where this information was not available, we relied on the subjective judgement of the three authors.

We matched each of these 54 individuals with their most similar looking counterpart in the remaining faces in our dataset. A standard convolutional neural network descriptor (termed VGG) [67] was used to extract a low-dimensional, perceptually meaningful, representation of each face in the full dataset. The extracted representation – a 4096-D real-valued vector – for each of the 54 manually selected target faces was compared with all other representations in the dataset to find the most similar face as defined by the face whose representation is most similar – in terms of Euclidean distance – with the target face [108, 109]. A mid-way morph was then generated for each pair of matched faces as follows.

A total of 68 corresponding points on the two faces were extracted using a standard facial landmark detector [82], Figure 6.5(a)-(b). These points were augmented with an average of 116 manually selected points along the hairline and top of the head, ears, and neck Figure 6.5(d)-(e). These manually selected points improved the overall visual quality of the generated morphs, Figure 6.5(c) versus (f). After extracting corresponding facial landmarks, and prior to generating the facial morphs, the two faces were aligned by an affine transform, consisting of anisotropic scaling, shearing, rotation, and translation. This alignment ensured that facial features did not significantly move during the morphing process. In particular, denote the 68 corresponding feature points on each face as (x_i, y_i) and (u_i, v_i), $i \in [1, 68]$. The six-parameter affine is given by:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} a_5 \\ a_6 \end{pmatrix}, \tag{6.1}$$

where the affine parameters a_1, a_2, a_3, and a_4 embody the anisotropic scaling, shearing, and rotation, and the parameters a_5 and a_6 embody the horizontal and vertical translation.

[2] All images and morphs will be made available upon request.

The transforms that best, in the least-squares sense, aligns the features on each face is estimated by first defining the following quadratic error function:

$$
\begin{aligned}
E(\vec{a}) &= \left\| \begin{pmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{68} & y_{68} & 0 & 0 & 1 & 0 \\ 0 & 0 & x_{68} & y_{68} & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} - \begin{pmatrix} u_1 \\ v_1 \\ \vdots \\ u_{68} \\ v_{68} \end{pmatrix} \right\|^2 && (6.2) \\
&= \|M\vec{a} - \vec{b}\|^2. && (6.3)
\end{aligned}
$$

This quadratic error function is minimized using standard least-squares estimation: differentiate with respect to $\vec{a}$, set the result equal to 0, and solve for $\vec{a}$ to yield the least-squares estimate of the aligning affine transform:

$$\vec{a} \ = \ (M^t M)^{-1} M^t \vec{b}. \qquad (6.4)$$

Given two aligned faces f and g (with VGG-representations $\vec{v}_f$ and $\vec{v}_g$), a morphed face m_{fg} is generated using a standard image-warping technique [125]. Briefly, a triangular mesh is created on each face using the facial landmarks as vertices, Figure 6.5(g)-(h). A mid-way morph is created by geometrically warping each triangular patch according to a morphing parameter $\alpha \in [0, 1]$, where a value of 0 corresponds to the source image f, a value of 1 corresponds to the source image g, and an intermediate value corresponds to a mid-way morph. The underlying pixel values are similarly computed as a weighted combination, $(1 - \alpha)f + \alpha g$, of the original pixel values (applied separately to each image color channel).

In past studies, morphs have often been generated with $\alpha = 0.5$ which means that the original faces f and g are weighted equally to generate a 50/50 morphed face m_{fg}. Previous research, however, has indicated that, when creating a morphed face, if one individual in the pair is more distinct than the other, then the 50/50 morph typically resembles the more distinct individual [104,126]. To compensate for this effect, we generated a range of morphs m_{fg}^{α} with α ranging from 0.1 to 0.9, in steps of 0.1. The value of α was selected that led to a morphed face m_{fg}^{α} that was, in the Euclidean sense on the underlying VGG-representation, mid-way between the source images f and g.

To improve overall contrast, each morph m_{fg}^{α} was gamma-corrected with $\gamma = 1.5$. The morphs were then tightly cropped around the face and manually edited to remove obvious morphing artifacts, Figure 6.5(f) versus (i). Lastly, to ensure that the morphing process did not create any obvious artifacts, the images were matched in terms of luminance, color, and sharpness. In particular, the mean luminance of each source image f and g was matched to the mean luminance of the morph image m_{fg}, and the source image mean and variance of the chrominance channels (Cb/Cr) were matched to the morphed image. Because image morphing tends to lead to blurring, each RGB color channel of each source and morph

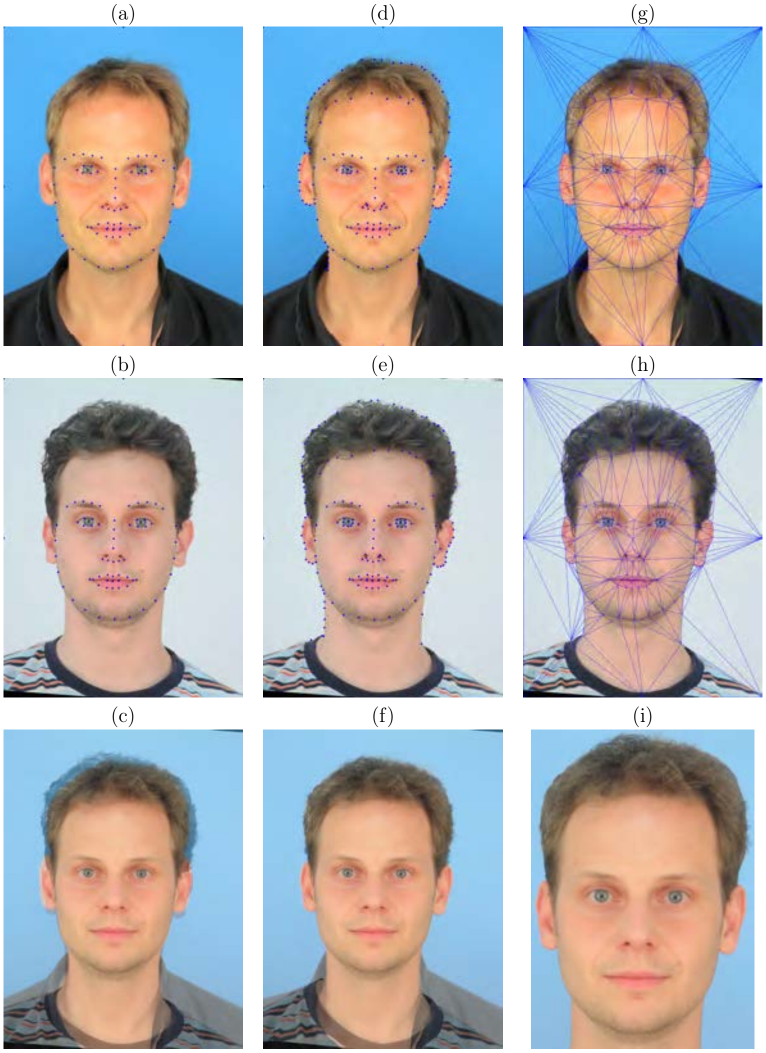


Figure 6.5: Shown are original faces of two different people f and g (panels (a) and (b)) with the automatically extracted facial landmark points overlaid (blue dots). Their mid-way morph m_{fg} generated using only these automatically extracted landmarks is shown in panel (c). Shown in panels (d) and (e) are the same original faces f and g now with both the automatically and manually selected facial landmark points overlaid (blue dots). Shown in panels (g) and (h) are a tessellation of the faces used for the image morphing. The mid-way morph generated using both the automatically extracted and manually selected landmarks is shown in panel(f). Shown in panel (i) is the tightly cropped and manually touched-up and color and sharpness adjusted final image. (Original image sources: Utrecht ECVP [117] and PUT face database [116].)

image was high-pass filtered until the average gradient of each image channel matched the maximum gradient across all three images. The resulting 54 pairs of different individuals and their mid-way morph comprise our *different-individual* dataset

An analogous *same-individual* dataset was created by selecting a new set of 54 facial images from the original dataset of $3,500$ for which there were two or more distinct images of the same person. We manually selected individuals to match the gender, age, and race distribution of our different-individuals dataset. A mid-way morph was created for each pair of images using the same technique described above.

In summary, our dataset consists of 108 face pairs, 54 of two different individuals and 54 of the same individual taken at different times, each with a mid-way morph. Representative sets of these different and same faces and morphs are shown in Figure 6.6 and Figure 6.7 [3].

6.4.2 Experiment 1a: identification (original and morph)

In this first experiment we examined people's ability to determine whether two facial images, one original and one morphed, are of the same person or not [4].

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as: 65 men, 34 women, 1 prefer not to say; between 22-72 years of age ($\mu = 36.8$; $\sigma = 9.6$); 74 White, 14 South Asian, 7 East Asian, and 5 African American. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile.

A within-subject design was employed in which each trial consisted of two images (one original, one morph) displayed side-by-side in one of eight configurations. Denote the original images of different people as f and g and their mid-way morph as m_{fg}, and denote the original images of the same people as h and $\tilde{h}$ and their mid-way morph as $m_{h\tilde{h}}$. There are four configurations for each dataset consisting of 54 pairs from the "different-individual" dataset with one image on the left and one on the right: $f + m_{fg}$; $m_{fg} + f$; $g + m_{fg}$; $m_{fg} + g$, and 54 pairs from the "same-individual" dataset: $h + m_{h\tilde{h}}$; $m_{h\tilde{h}} + h$; $\tilde{h} + m_{h\tilde{h}}$; $m_{h\tilde{h}} + \tilde{h}$, for a total of 432 possible displays. Each participants viewed only 108 image pairs using the following fully counterbalanced block design. Four blocks were created each containing 27 trials for a total of 108 trials. The first and second block each consisted of 14 different

[3] (Original image sources: MR2 [CC BY-NC-SA 4.0], Chicago Face Database [permission to publish images granted], CUHK student database [118], NIST color FERET [image publication permitted under fair use policy], and CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy]).

[4] All experiments reported in this paper were approved by University of California Berkeley's Office for Protection of Human Subjects (OPHS), Protocol ID: 2019-07-12422. Participants gave fully informed consent prior to taking part.

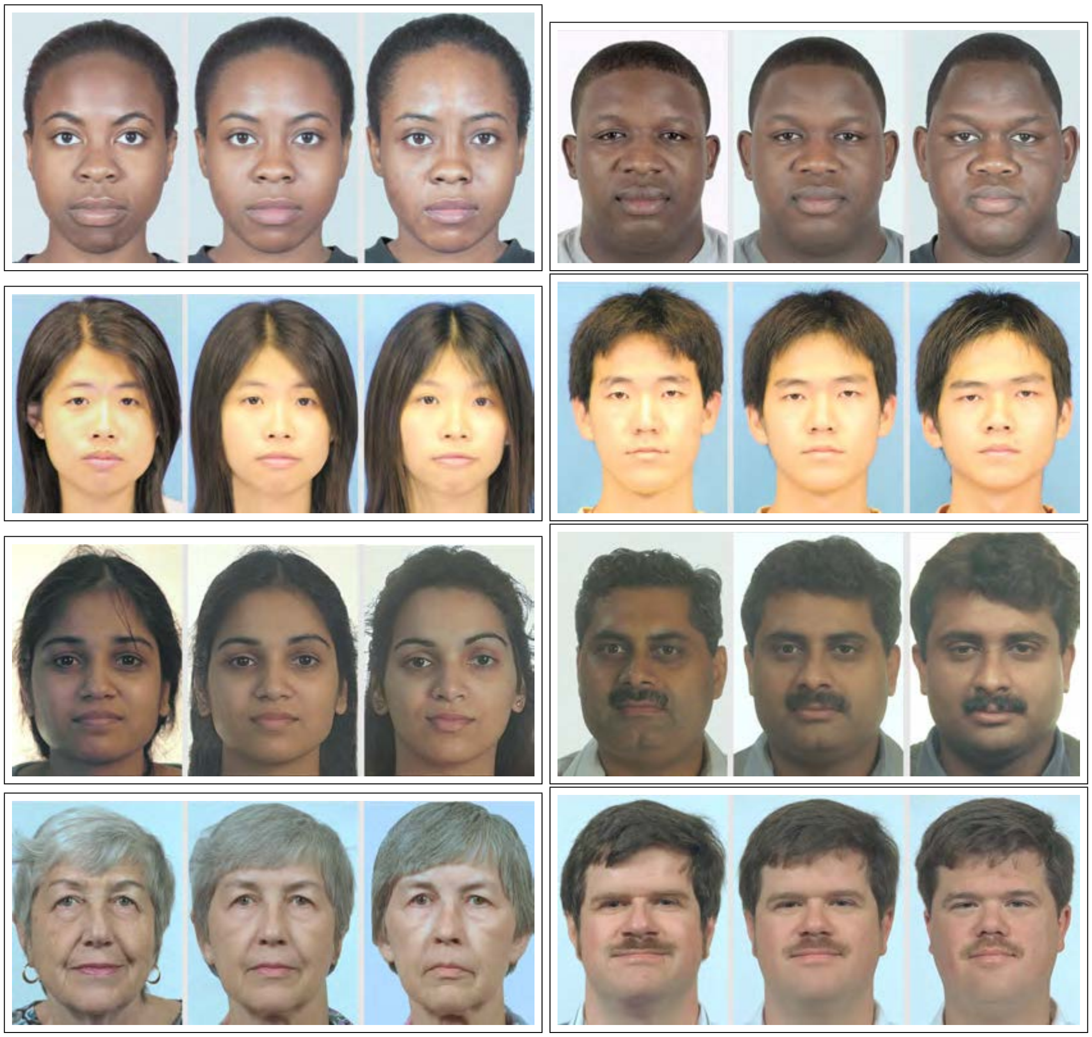


Figure 6.6: Shown are the example images from different-individual dataset. Shown for each set of three images are two original images (left/right) and their mid-way morph (center).

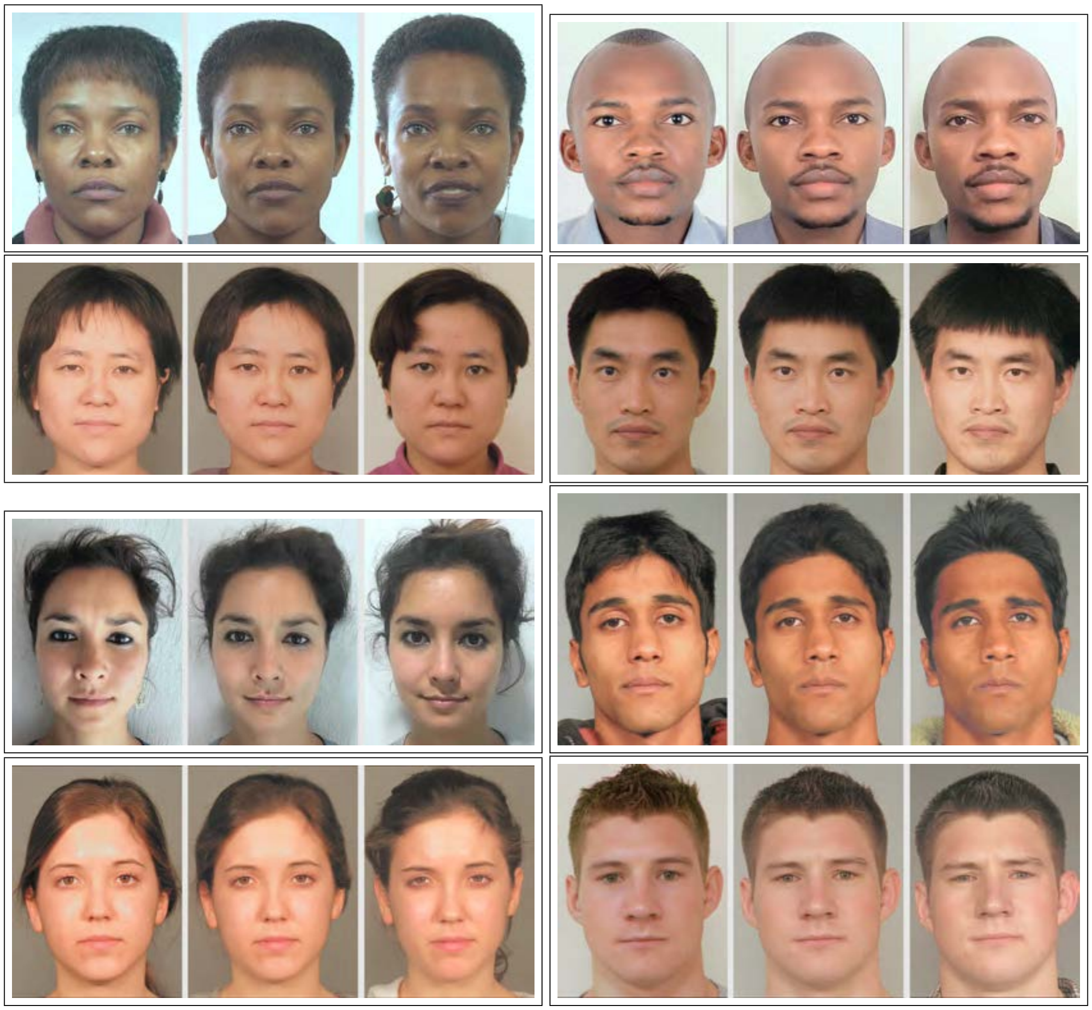


Figure 6.7: Shown are the example images from same-individual dataset. Shown for each set of three images are two original images (left/right) and their mid-way morph (center).

and 13 same image pairs; the third and fourth blocks each consisted of 13 different and 14 same image pairs. Each block consisted of the same number of men, women, and racial groups.

On each trial, participants were instructed to specify if the images were of the same person or not and asked to rate the confidence in their response. In this discrimination task, chance performance is 50%. Four attention-check trials were created, one for each block. These trials were intentionally easy comprising of two images of distinctly different looking people, one male and one female, Figure 6.8.

Participants first received task instructions including a brief description of what face morphing is and how it can be used to commit identity fraud. Participants then completed a practice trial; they viewed two images on the screen, an original image and a morph. Participants had an unlimited amount of time to indicate whether or not they thought that the images were of the same individual. After responding to the same/different individual question, participants rated their confidence in their decision using a 6-point Likert-type scale, from 1 (*50% - Guessing*) to 6 (*100% - Absolutely certain*).

Following the practice trial, participants completed the 108 trials in blocks of 27 plus one attention check trial per block, shown in a randomized order within each block. Blocks were shown in one of four possible counterbalanced orders. At the end of the session participants were asked a few basic demographic questions.

A precision-for-planning analysis revealed that at least 99 participants would provide a margin of error that is 0.2 of the population standard deviation with 95% assurance [127,128]. This analysis applies to all reported experiments.

Results

The average accuracy of identifying a facial image as the same person or not was 59.2% (chance is 50%), corresponding to a sensitivity of $d' = 0.68$ and bias of $\beta = 1.81$, where the bias corresponds to a tendency to label faces as “same", Table 6.1. The accuracy for faces of different/same individuals was 29.5%/88.8% – participants were heavily biased to saying that faces were of the same individual.

Shown in Figure 6.9(1a) is, for each level of participant-reported confidence (on a scale of 1 to 6 (certain)), the participant accuracy. With similar average accuracy across all levels of confidence, we see that participants are not well calibrated in their response and confidence.

Discussion

The results of this experiment suggest that human participants have a limited ability to reliably determine the identity in a mid-way facial morph. There are two possible interpretations of this result: (1) the morphs are of high enough quality and similarity as to mask identity; or (2) participants are simply unable to accurately distinguish two unfamiliar faces. In the next experiment (1b), we seek to differentiate between these two

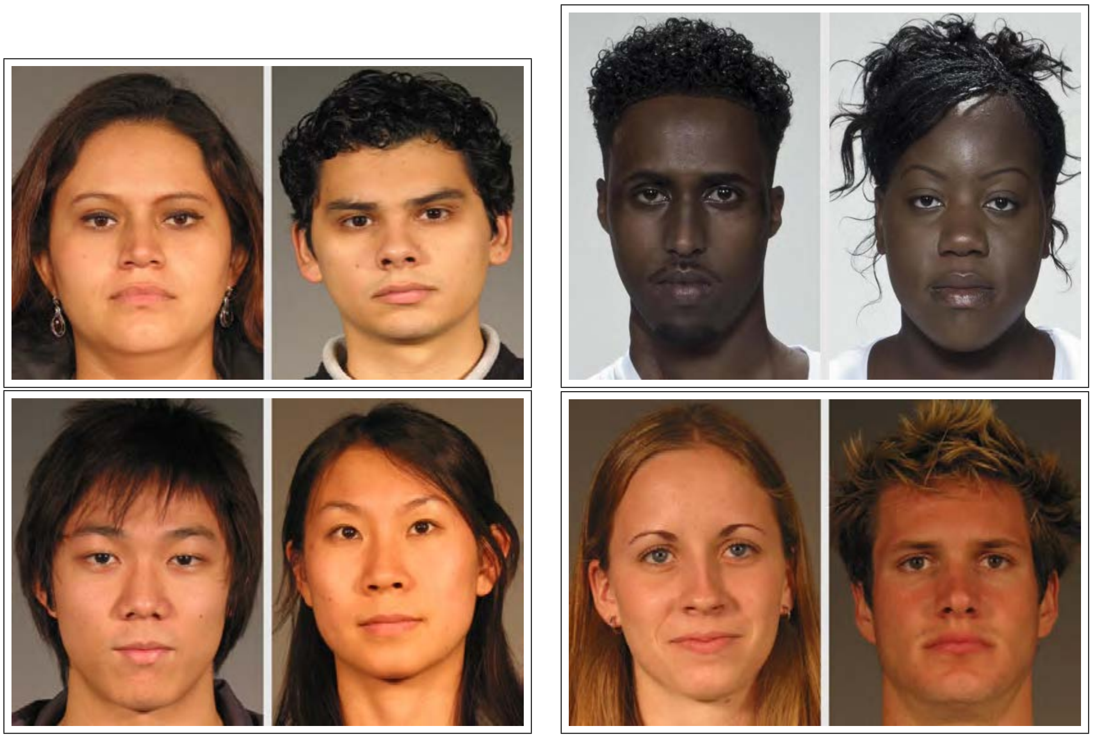


Figure 6.8: Catch trials used in Experiments 1a, 1b, and 1c to ensure that participants were paying attention to the task. (Original image sources: Face Research Lab London Set [CC BY 4.0] and CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy].)

possibilities by asking participants to distinguish between two original non-morphed facial images consisting of the same or different person.

6.4.3 Experiment 1b: identification (original and original)

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as: 53 men, 47 women; between 23-69 years of age ($\mu = 39.4$; $\sigma = 10.9$); 74 White, 14 South Asian, and 6 African American, 2 East Asian, and 4 other/prefer not to say. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile. A further two participants were excluded because they responded incorrectly on at least one of the attention check questions. There was no overlap between the participants in this experiment and Experiment 1a.

Experiment	d'	β	% correct [95% CIs]
1a	0.68	1.81	59.2 [57.6, 60.7]
1b	1.74	1.03	80.8 [78.8, 82.8]
1c	0.57	1.44	59.2 [57.9, 60.6]
2a	0.21	0.98	54.1 [52.5, 55.5]
2b	0.53	0.92	60.4 [58.9, 61.9]

Table 6.1: Participant accuracy in five experiments, reported as sensitivity (d'), bias (β), and accuracy (% correct with [95% confidence intervals]). Experiment 1a: determine if two images, one original and one morphed, are of the same person; Experiment 1b: determine if two images, both original, are of the same person; Experiment 1c: replication of Experiment 1a with training; Experiment 2a: determine if a single facial image is a morph or not; and Experiment 2b: replication of Experiment 2a with training and feedback.

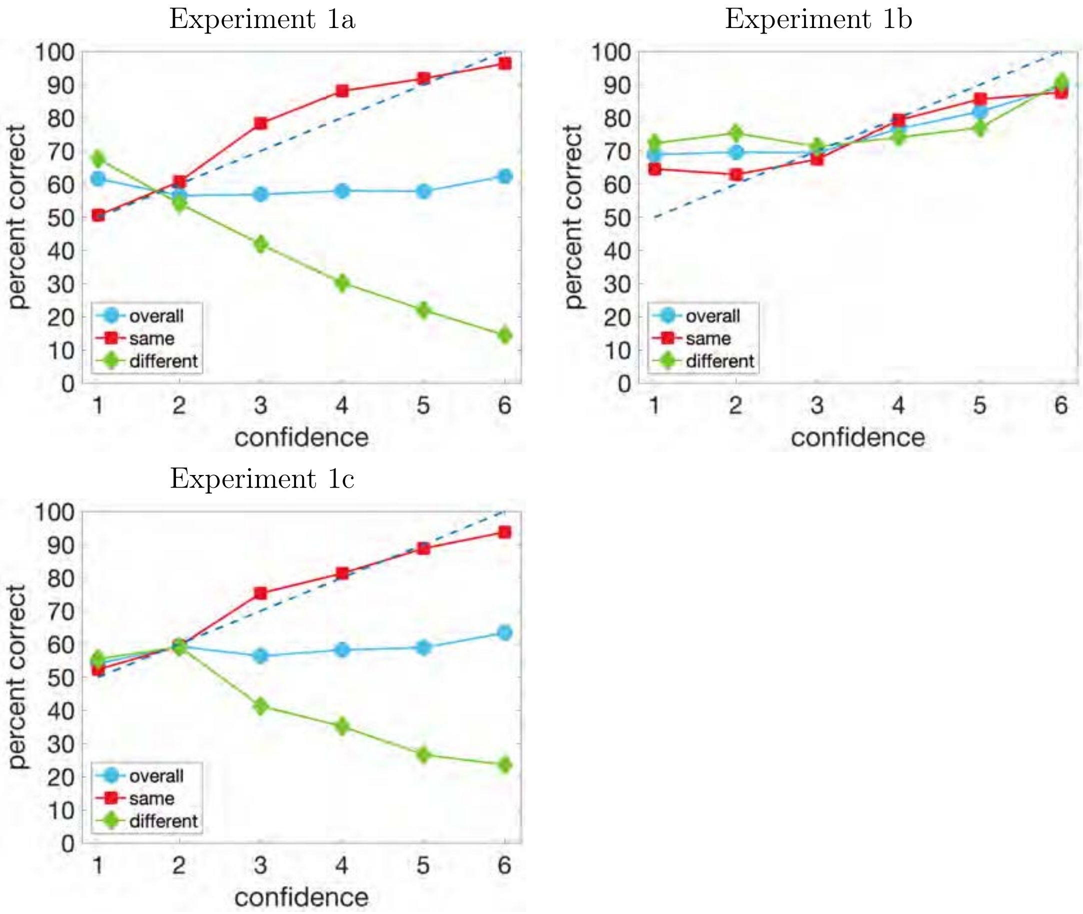
A within-subject design was employed in which each trial consisted of two original images – from the same dataset as used in Experiment 1a – displayed side-by-side in one of four configurations. Each participant saw 54 different individuals (f, g) with two possible configurations: $f + g$ or $g + f$, and 54 same individuals (h, $\tilde{h}$) with two possible configurations: $h + \tilde{h}$ or $\tilde{h} + h$, for a total of 216 possible displays. Each participant viewed 108 image pairs using the counterbalanced block design per Experiment 1a. Given that this experiment did not include any morphed facial images, we removed the brief explanation of face morphing from the task instructions and amended the practice trial to show two distinct original images of the same person. The procedure was otherwise identical to that of Experiment 1a.

Results

The average accuracy of identifying two facial images as depicting the same person or not was 80.8%, corresponding to a sensitivity of $d' = 1.74$ and bias of $\beta = 1.03$. The accuracy for faces of different/same individuals was 80.4%/81.3% – unlike the previous experiment, participants were not biased in their responses. Shown in Figure 6.9(1b) is, for each level of participant-reported confidence, the participant accuracy. With slightly higher accuracy at the higher levels of confidence, it appears that participants are fairly well calibrated in their response and confidence.

Discussion

The results of this experiment suggest that participants can reliably determine whether two original facial images depict the same person or two different people. Therefore,


Experiment 1a

Experiment 1b

Experiment 1c

Figure 6.9: Confidence-accuracy curves for Experiment 1a, 1b, and 1c. The dashed line represents perfect accuracy-confidence calibration.

participants' limited ability in the task in Experiment 1a may be interpreted as a result of the morphs being of high enough quality and similarity to mask identity.

Given that participants can distinguish two unfamiliar faces with reasonable accuracy, we next examine whether participant accuracy in distinguishing identity in morphed faces can be improved with training. To develop our training initiative, we draw on the finding that attending to certain facial features when comparing two faces can help to improve the accuracy of face matching decisions [129, 130]. In the next experiment (1c), we replicate Experiment 1a but this time using masked facial images that allow participants to only compare the eyes, nose, and mouth.

6.4.4 Experiment 1c: identification (original and morph) with masking

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as: 51 women, 48 men, 1 prefer not to say; between 22-65 years of age ($\mu = 40.3$; $\sigma = 9.6$); 81 White, 7 East Asian, 6 African American, 5 South Asian, and 1 other/prefer not to say. As in the previous two experiments, participants received \$5 for completing this experiment. As an incentive to encourage effort on the task, a \$5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile. There was no overlap between the participants in this experiment and Experiments 1a and 1b.

Other than the two exceptions enumerated below, the design, underlying stimuli, and procedure were identical to that used in Experiment 1a.

1. In each trial, participants saw a pair of images (one original and one morph) displayed side-by side. One image pair revealed only the eyes, and the other image pair revealed only the nose/mouth region, Figure 6.10. The creation of these masks was automated as follows. The pixel locations of the facial features (eyes, nose, mouth) were extracted using OpenFace [54]. For the eyes, a bounding box was extracted that contained all of the features on both eyes. For the nose/mouth, a bounding polygon was extracted that contained all of the features on the nose and mouth. To ensure that the mask did not occlude the features, these bounding boxes were enlarged by 5% of their original size. The final images, Figure 6.10, were generated by reducing the contrast of all pixels outside of the mask to 15% of full contrast, making it difficult for participants to use the entire face for recognition, while leaving enough context for the visible features.

2. The order in which participants viewed the two feature regions was counterbalanced resulting in twice the number of display configurations as in Experiment 1a. Participants saw 27 of each of the different people configurations ($f + m_{fg}$; $m_{fg} + f$; $g + m_{fg}$; $m_{fg} + g$) with the eye region shown first and the mouth and nose region shown second and 27 with the mouth and nose region shown first and the eye region shown second. Participants saw 27 of each of the same people configurations ($h + m_{h\tilde{h}}$; $m_{h\tilde{h}} + h$; $\tilde{h} + m_{h\tilde{h}}$; $m_{h\tilde{h}} + \tilde{h}$) with the eye region shown first and the mouth and nose region shown second and 27 with the mouth and nose region shown first and the eye region shown second. Each participant viewed a total of 108 image pairs.

Results

The average accuracy of identifying a facial image as the same person or not was 59.2%, corresponding to a sensitivity of $d' = 0.57$ and bias of $\beta = 1.44$ (cf Experiment 1a: 59.2%,

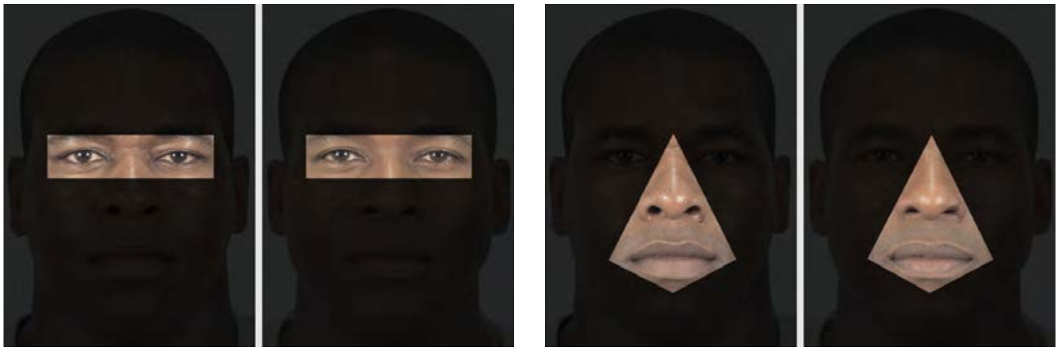


Figure 6.10: Example masked stimuli from Experiment 1c. Shown in each image pair is an original image of an individual f (left) and the mid-way morph m_{fg} (right) to a different person g, with only the eye region or mouth and nose region visible. (Original image source: Chicago Face Database [permission to publish images granted].)

$d' = 0.68$, $\beta = 1.81$, Table 6.1). The accuracy for faces of different/same individuals was 36.1%/82.3% – although overall accuracy was still not high, the masking reduced the bias to report that faces were of the same individual. As in Experiment 1a, average participant accuracy is similar across all levels of confidence, suggesting that participants are still not well calibrated in their response and confidence, Figure 6.9(1c).

Discussion

Had the strategy of focusing participants' attention to facial features been successful in increasing accuracy or decreasing bias, this would have been a simple strategy for a passport issuance office to adopt. The results of this experiment, however, reveal that facial-feature comparison did not significantly improve participants' accuracy in determining identity of morphed faces. Compared to Experiment 1a, however, participants showed a smaller bias to respond “same". Participants clearly struggle to distinguish identity when presented with two images, one of which is a mid-way morph.

Distinguishing identity, however, is only one way in which a fraudulent identity might be determined. The other way is to simply identify a face as having been morphed relative to some unknown face. In the next set of experiment, we examine participant's ability to perform this task.

6.4.5 Experiment 2a: classification (original or morph)

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as: 57 men, 43 women; between 24-72 years of age ($\mu = 40.4$; $\sigma = 10.2$); 78 White, 10 South Asian, 5 African American, 4 East Asian, and 3 other/prefer not to say. Participants received \$5 for completing this experiment. As an incentive to encourage effort on the task, a \$5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile. A further three participants were excluded because they responded incorrectly on at least one of the attention check questions. There was no overlap between the participants in this experiment and Experiments 1a, 1b, or 1c.

A within-subject design was employed in which each trial consisted of a single original or morphed face. For the morphed face trials, each participant saw 27 different people mid-way morphs (m_{fg}) and 27 same person mid-way morphs ($m_{h\tilde{h}}$). For the original image trials, participants saw 27 images, either f or g, from the different individual image pairs and 27 images, either h or $\tilde{h}$, from the same individual image pairs.

Each participant viewed 108 images using the following fully counterbalanced block design. Four blocks were created each containing 27 trials for a total of 108 trials. The first and second block each consisted of 14 original face trials and 13 morphed face trials; the third and fourth blocks each consisted of 14 morphed face trials and 13 original face trials. The selection of the original or mid-way morph from each of the 108 image sets was also counterbalanced resulting in two versions of each block.

On each trial, participants were instructed to specify if the image was a morph or not and asked to rate the confidence in their response. In this task, chance performance is 50%. Four attention-check trials were created, one for each block. These trials were intentionally easy comprising a morphed face of a person with an image of a cartoon character, Figure 6.11.

Participants first received task instructions including a brief description of what face morphing is and how it can be used to commit identity fraud. Participants then viewed four videos demonstrating how two faces can be digitally combined to create a morph of those two faces. To create these videos, we selected four additional face pairs from the original dataset of $3,500$ faces. For each face pair, we generated morphs using the same method as described previously (see Dataset section). To demonstrate the gradual morphing of two faces we generated five morphs with a different blending value α ranging from 0.1 to 0.5 in steps of 0.1. The $0.5 - \alpha$ morph was then manually edited to remove obvious morphing artifacts. In the video, the two original images (f and g) were displayed on either side of their morph (m_{fg}). The six versions of the morph appeared sequentially, starting with the $0.1 - \alpha$ morph. Each version of the morph remained on the screen for one second. Participants viewed all four videos and were asked to indicate if they were able to see the videos clearly. Participants then completed a practice trial consisting of a single image on the screen, an original image or a mid-way morph.

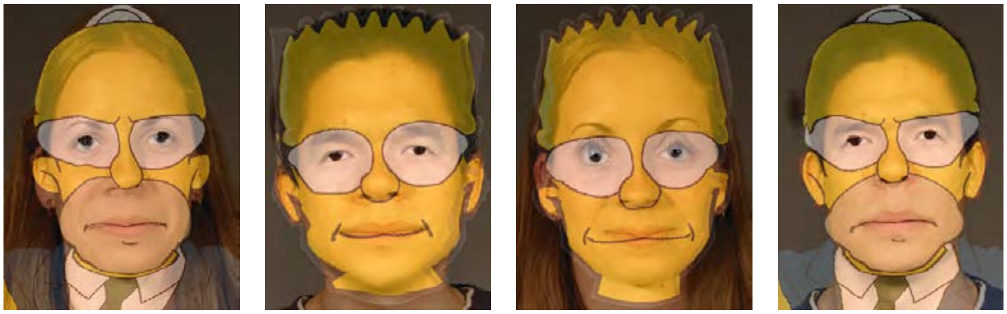


Figure 6.11: Catch trials used in Experiments 2a and 2b to ensure that participants were paying attention to the task. (Original image sources: CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy] and Pixy.org [CC BY-NC-ND 4.0].)

Following the practice trial, participants completed the 108 trials in blocks of 27 plus one attention check trial per block, shown in a randomized order within each block. Blocks were shown in one of eight possible counterbalanced orders. At the end of the session participants were asked a few basic demographic questions.

Participants had an unlimited amount of time to indicate whether they thought that the image was a morph or not. After responding to the morph or not question, participants rated their confidence in their decision using a 6-point Likert-type scale, from 1 (*Guessing*) to 6 (*Absolutely certain*).

Results

The average accuracy of identifying a face as a morph or not was 54.1%, corresponding to a sensitivity of $d' = 0.21$ and bias of $\beta = 0.98$. The accuracy for original/morphed faces was 50.0%/58.1%. As in Experiments 1a and 1c, average participant accuracy was similar across all levels of confidence, again suggesting that participants are not well calibrated in their response and confidence, Figure 6.12(2a).

Discussion

The results of this experiment suggest that human participants cannot reliably determine when a facial image has been morphed and when it is an original image. Two possible explanations for this result are that (1) the morphed faces are of high enough quality that there are no artifacts that can be reliably perceived by human participants; or (2) participants are unaware of the artifacts to look for in morphed faces. To try to determine which of these two possibilities best accounts for our results, in the next experiment we replicate Experiment 2a but before attempting the task, participants completed a short training session that highlights some common morphing artifacts to look for in the images.

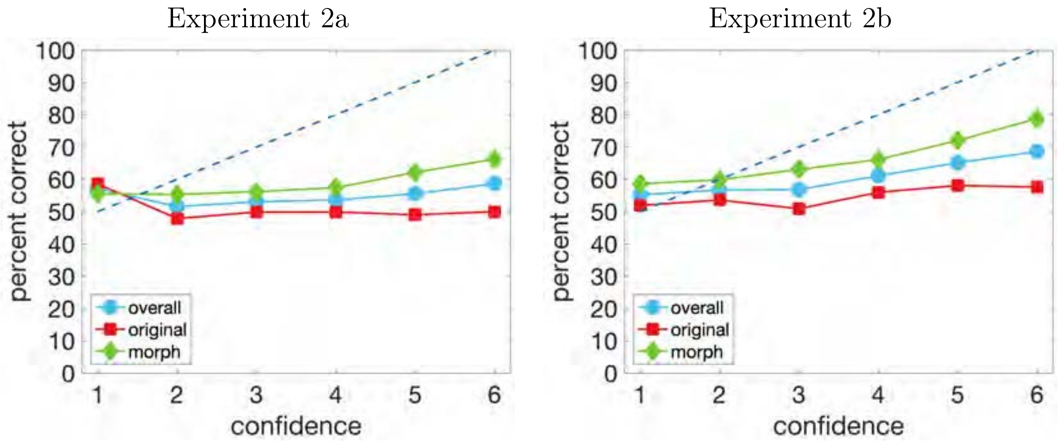


Figure 6.12: Confidence-accuracy curves for Experiment 2a and 2b. The dashed line represents perfect accuracy-confidence calibration.

6.4.6 Experiment 2b: classification (original or morph) with training/feedback

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as: 58 men, 41 women, 1 prefer not to say; between 24—68 years of age ($\mu = 39.7$; $\sigma = 9.9$); 81 White, 9 South Asian, 4 African American, 2 East Asian, and 4 other/prefer not to say. As in the previous experiments, participants received \$5 for completing this experiment. As an incentive to encourage effort on the task, a \$5 bonus was offered and paid for those achieving an accuracy in the top 20 percentile. A further five participants were excluded because they responded incorrectly on at least one of the attention check questions. There was no overlap between the participants in this experiment and Experiments 1a, 1b, 1c, or 2a.

Other than the inclusion of a training session as described below and accuracy feedback after each trial, the design, underlying stimuli, and procedure were identical to that used in Experiment 2a.

After viewing the four videos demonstrating how two faces can be digitally combined to create a morph of those two faces participants completed a short training. The training provided information about common morphing artifacts that may be helpful to look for when deciding if the facial images were morphs or not. Participants were told that:

1. Morphed faces tend to look less sharp: the complexion of a morphed face is usually smoother with a more uniform appearance.

2. Morphing hair can be difficult, and morphs often have fewer wayward strands of hair

and ghosting (a ghost-like outline of another person's hair).

3. When morphing images of two people with different postures or different clothing/hair coverage, the editing process might make the neck line appear unnaturally straight and flat.

To check that participants paid attention to the training, they were then asked to select the three artifacts from a list of six possible options, where the three incorrect answers were easily identifiable as they were not mentioned in the training. Participants were given the option to view the training session a second time if they were unsure of the correct options. The other change from Experiment 2a was that after responding on each trial, participants were provided with feedback indicating whether their response was correct or not.

Results

The average accuracy of identifying a face as a morph or not was 60.4%, corresponding to a sensitivity of $d' = 0.53$ and bias of $\beta = 0.92$ (cf Experiment 2a: 54.1%, $d' = 0.21$, $\beta = 0.98$, Table 6.1). The accuracy for original/morphed faces was 54.6%/66.2%. Average accuracy was only slightly higher at the higher levels of confidence (4-6) than the lower levels (1-3) of confidence suggesting participants had a limited ability to calibrate their response and confidence, Figure 6.12(2b).

Discussion

The results of this experiment indicate that raising awareness of morphing artifacts and providing feedback led to only a small improvement in participants' accuracy in determining whether a facial image has been morphed or not. Even with this training, participants struggled to reliably identify a face as having been morphed relative to an unknown face.

Taken together, the results of our five experiments suggest that people are unable to reliably detect face morphing, neither by distinguishing identity nor by classifying a face as having been morphed. Next we examine whether computational approaches can be used to detect face morphing.

6.4.7 Crowd Wisdom

To determine whether groups are more accurate in the detection of face morphing than individual decision makers we next examined whether there is "wisdom in the crowd" [131]. For each experiment, we aggregated the 100 participant responses for each of the 108 trials using a majority rules criterion.

Using this crowd-based approach in Experiment 1a resulted in an average accuracy of 53.8% for identifying a facial image as the same person or not, which was not reliably different to the average of individual responses (59.2%). In Experiment 1b, however, average accuracy using the crowd-based approach was 16.4% higher than the averaged individual responses (97.2% vs. 80.8%, 95% CI [12.7%, 20.1%]). In addition, in Experiment 1c, where participants received training, the crowd-based approach resulted in an average accuracy similar to the individual approach (58.3% vs. 59.2%). The improved accuracy in Experiment 1b suggests that participants make different mistakes and so pooling across multiple responses improves overall accuracy.

In Experiment 2a, accuracy in classifying images as original or morphs was similar when averaging individual (54.1%) and crowd (58.4%) responses. When participants received training and feedback (Experiment 2b), the crowd-based approach resulted in an average accuracy 12.2% higher than the individual approach (60.4% vs. 72.6%, 95% CI [6.1%, 18.4%]). Without any training (Experiment 2a) participants typically made the same mistakes but having received training (Experiment 2b) there was greater variation in which of the trials participants responded correctly on. This result suggests that with training and feedback the crowd becomes wiser.

It is possible that aggregating across identity verification decisions of multiple passport officers might lead to greater accuracy in real-world passport issuance. Of course, this additional effort might not be feasible and we also note that when people know a decision is group-based it can lead to social loafing [131].

6.4.8 Computational Identification

The results of Experiments 1a and 1c show that participant's ability to determine whether two facial images, one original and one morphed, are of the same person or not is limited. We next examine whether computational techniques can perform this task. A standard convolutional neural network [67] was used to extract a low-dimensional, perceptually meaningful [108, 109], representation of each face in our dataset of 108 face pairs and their corresponding mid-way morph. This is the same VGG representation used earlier to determine the similarity between two faces and to compute the mid-way morph.

For each of the 54 pairs of faces of the same individual taken at different times (h and $\tilde{h}$), the similarity between these faces was measured as the Euclidean distance between the VGG representation of the two original faces. Similarly, the distance was computed between the VGG representation for each of the 54 pairs of different individuals (f and g) and their mid-way morph (m_{fg}).

Shown in Figure 6.13(a) is the receiver operating curve (ROC) plotted as the true positive rate (correctly identifying the same individual) as a function of the false positive rate (incorrectly identifying an individual and their mid-way morph as the same). The area under the curve (AUC) is 0.38, where a chance classifier would have an AUC of 0.5, showing that even a state-of-the-art, machine-learning, face recognition algorithm is not able to perform this identification task. We note that, in a flipped classifier, this result effectively corresponds to an AUC of 0.62, still illustrating a fairly limited performance. Interestingly, however, the below chance AUC result indicates that a morphed face is highly similar to the source faces; a finding that we draw on in the subsequent Computational

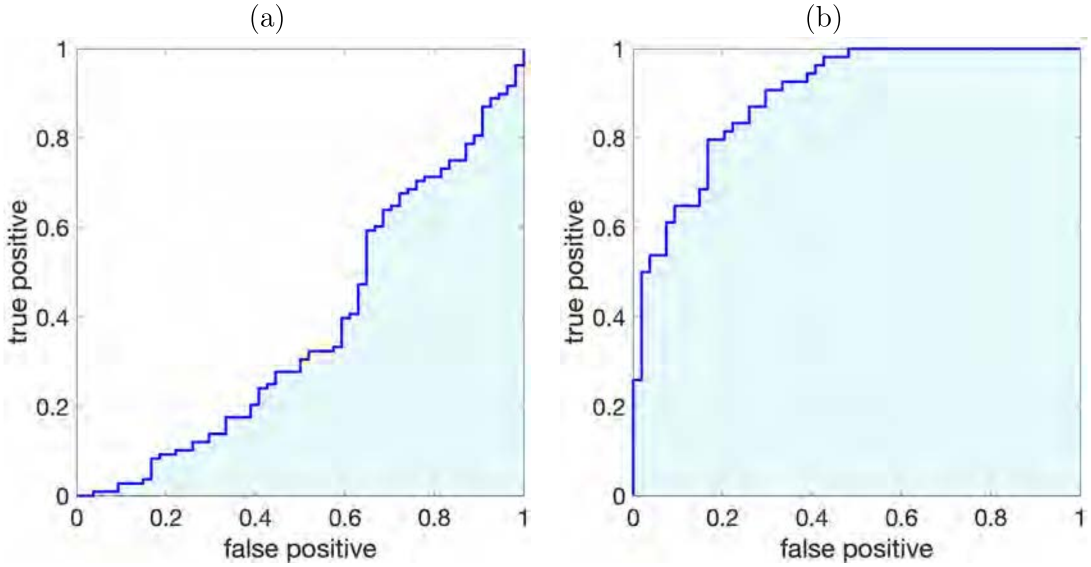


Figure 6.13: Receiver-operating-characteristic curves for (a) same individuals (true positive) and different individuals and their mid-way morph (false positive). (b) same individuals (true positive) and different individuals (false positive).

Classification section.

We next evaluate if this computational approach can perform the task of face recognition outside of the issue of morphing (as in Experiment 1b). The distance was computed between the VGG representation for each of the 54 pairs of faces of two different individuals (f and g). Shown in Figure 6.13(b) is the ROC for this task, now with an AUC of 0.90. Although VGG-based facial recognition is generally effective in distinguishing between different individuals, it struggles to distinguish between morphed faces, reinforcing just how difficult this task is.

Facial recognition systems have been shown to perform worse at recognizing faces of women and Black individuals [132, 133]. We next examined whether these biases were present when using the VGG-based face recognition algorithm to perform our identification task. When identifying pairs of faces of the same individual taken at different times (h and $\tilde{h}$) and the 54 pairs of different individuals (f and g) and their mid-way morph (m_{fg}) the face recognition algorithm performed worse on Black faces (AUC = 0.00) than East Asian (AUC = 0.29), South Asian (AUC = 0.61), or White (AUC = 0.29) faces. The algorithm performed slightly better for women (AUC = 0.43) than for men (AUC = 0.33). In addition, in the absence of morph faces, the VGG-based facial recognition performed worse for Black faces (AUC = 0.75) than for the other races (all AUCs > 0.90). There was no difference in face recognition performance for women and men.

6.4.9 Computational Classification

In the previous sections, we saw that a mid-way morph between two different people looks similar enough to each person so as to cause consistent misidentification by human participants and state-of-the-art facial recognition. In this section, we attempt to leverage the unusual similarity between two photos of a person as a possible indication that one of the photos is a mid-way morph.

Consider, for example, the use of a mid-way morph in identity theft in which person f is attempting to steal person g's identity, and submits a request for a new passport with a mid-way morph photo m_{fg}. The passport office will compare the original photo g with the new photo m_{fg} to make sure that it is the same person. Per our earlier results, and assuming a high-quality morph, the faces will look similar enough to match. The two photos g and m_{fg}, however, will share significant geometric and photometric properties because the morphed image m_{fg} is composed of one half of the original image g, as compared to a completely new photo of person g which will almost certainly differ somewhat in terms of head pose, facial expression, lighting, etc.

We hypothesize, therefore, that a pair of images of an individual, one of which is a morph, will be more geometrically and photometrically similar than two separately photographed images of an individual. Each of the original images was registered to the morphed image using a standard local and nonrigid registration (using Matlab's `imregdemons`), parameterized as a local 2-D motion field (v_x, v_y). The magnitude of the geometric distortion between two images is quantified as the average magnitude of the gradient of the underlying motion field $(\sqrt{v_x^2 + v_y^2})$. Once aligned, the photometric similarity between two images is quantified as the mutual information [134] on the luminance channel.

Shown in Figure 6.14 are the geometric and photometric measurements for morphed (filled red circles) and original (open blue circles) images. Each original data point corresponds to the average difference h aligned to $\tilde{h}$, and $\tilde{h}$ aligned to h, where h and $\tilde{h}$ correspond to distinct images of the same person. Each morphed data point corresponds to the average difference between f aligned to m_{fg}, and g aligned to m_{fg}, where f and g correspond to images of different people. The morphed images are distinctly more photometrically similar (having a higher luminance mutual information) and more geometrically similar (having a smaller warp-field gradient). This, again, is as it should be given how the morphed image is created.

Because all of the images in our dataset are passport-style photos, this unusually high similarity among the morphed images is not simply an artifact of the style of the photographs. This unusual similarity can, therefore, be used as a cue to flag potentially suspicious similar images.

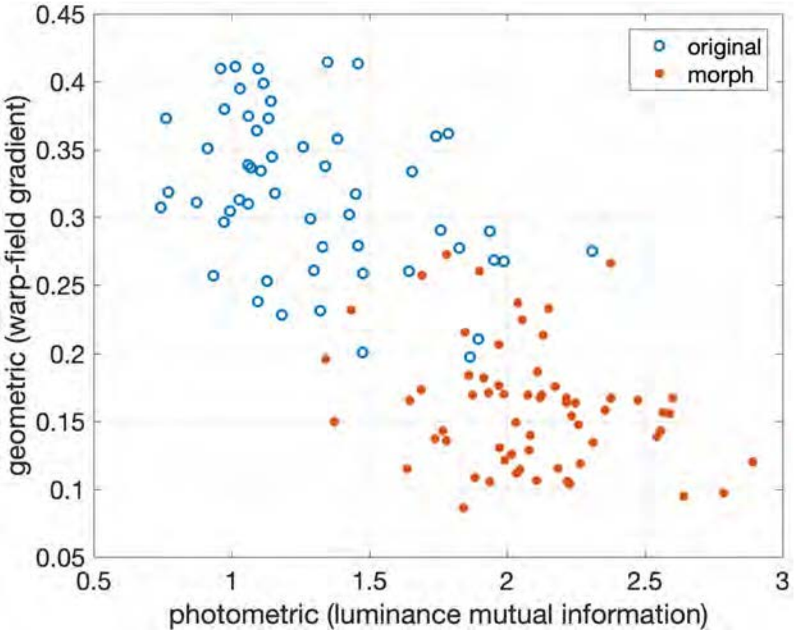


Figure 6.14: Photometric and geometric measurements between two distinct images of the same individual (original) and an image of an individual and a mid-way morph to another individual (morph). A lower geometric measure corresponds to a higher degree of similarity, and a higher photometric measure corresponds to a higher degree of similarity.

6.5 Discussion

Images synthesized by StyleGAN2 are realistic enough to fool naive observers. Even when told about specific synthesis artifacts, observers are unable to reliably discriminate the real from the synthetic. Similarly, in case of morphed faces, human participants showed a limited ability to detect face morphing, both by distinguishing identity (Experiment 1a) and by classifying a morphed face (Experiment 2a). Training did not significantly improve performance in the identification task (Experiment 1c) and training and feedback resulted in only a small improvement in performance in the classification task (Experiment 2b). Additionally, we found that even a state-of-the-art, machine-learning, face recognition algorithm could not reliably distinguish one person from a mid-way morph. We did, however, identify a computational technique to leverage the unusual similarity between a pair of images when one is a mid-way morph. This technique could be implemented at passport issuance to help in flagging suspicious applications for further processing.

As synthetic media continues to improve in realism and sophistication, it will become increasingly more difficult to visually discriminate between the real and the fake. It seems, therefore, that the possible advantage of training people to look for certain artifacts may be of limited value. Going forward it is important to develop robust computational techniques and stricter policies to protect people from these types of attacks.

Chapter 7

Conclusions and Discussion

Each of the forensic techniques presented in this dissertation exploit an intrinsic flaw in deep-fake videos. The first two person-specific techniques described in Chapters 2 and 3 exploit the fact that deep-fake videos are often driven by an impersonator who is not the person depicted in the video. As a result, the facial mannerisms distinct to an individual are disrupted. This type of soft-biometric model, built over long temporal windows, is difficult to circumvent in the current per-frame deep-fake synthesis techniques.

While these two techniques exploit facial mannerisms, other soft-biometric features are violated in the creation of deep fakes. Because deep-fake synthesis focuses exclusively on the facial region, other parts of the head are left unattended. In Chapter 4, we leverage human ear biometrics and show that the ears in face-swap deep fakes can be identified as the impersonator's, and not the person it purports to be.

These biometric cues are most effective when an impersonator is used in the synthesis of deep fakes, primarily face-swap and puppet-master deep fakes. Lip-sync deep fakes, on the other hand, preserves most of the biometric cues, but other physiological properties may be violated. In Chapter 4 we exploited inconsistencies between the movement of the mouth and ear. In Chapter 5 we exploited inconsistencies between the shape of the mouth and the underlying audio signal. These inconsistencies arise because the mouth is synthesized independent of the rest of the face and head.

However, there are limitations to these detection techniques. Similar to many other forensic techniques, most of the above methods also rely on some manual intervention. This limits their scalability for use in automatic detection on large-scale platforms. As a result, online users are often tasked with deciding whether an image or video is real or not. In Chapter 6, therefore, we turned our attention to understanding the limits of the human visual system to recognize manipulated faces. It was shown that humans are only slightly better than chance at detecting synthetic faces, both those generated by state-of-the-art deep-fake synthesis and those generated by more traditional manipulation techniques.

Current deep-fake videos, even though highly realistic, often ignore many aspects of a person's identity. In the future, the person-specific techniques presented in this dissertation can be augmented with other biometric cues. To create convincing deep-fake video, for example, it is necessary to accompany it with convincing audio. Currently, deep-fake videos are associated with either 1) an impersonator's voice or 2) a synthesized voice. As a result the audio signal can be used as a biometric cue to detect person-specific deep fakes. Additionally, the face-swap deep fakes using an impersonator's body can leave other tell-tales like impersonator's hand or body motion or gait which, in the past, have been shown to provide biometric information.

Additionally, in most deep-fake videos people are made to say things they have never said. This can disrupt the statistical properties of the underlying text of the speech. For instance, a public speech given by a world leader is generally pre-meditated and in many cases written by a professional speech writer. Therefore, similar to author attribution, an analysis of spoken text can also provide important biometric signals. The knowledge about the person's identity behind the words can be valuable in high-stake cases like political speeches. Additionally, we may be able to use textual and visual cues together to extract high-level semantic relationships between the two modalities. The current audio-based synthesis techniques, however, may not understand such semantics of the underlying text and the commonly associated gestures with it.

Lastly, it is also important to simultaneously understand the human perceptual ability to detect deep-fake videos. In Chapter 6, experiments were performed on static faces and many interesting research questions are still unanswered for deep-fake video perception. For example, there are multiple types of deep-fake videos, each of which pose a different set of forensic challenges. However, it is not known which of the three types of deep fakes are the most difficult to detect perceptually. This analysis can help researchers to calibrate their focus on specific types of deep fakes. One of the most popular type of deep fakes on the internet are face-swaps, on which the biometric-based forensic techniques are shown to work the best. However, it is not known if humans can also be trained to recognize person-specific cues to detect face-swaps. More generally, do we become less susceptible to deep fakes when we know the person in the fake? If yes, then what are the person-specific cues that help us recognize the fakes? Such perceptual studies can guide the development of biometric-based detection algorithms.

Often I'm asked if detection will eventually win the ongoing cat-and-mouse game of synthesis and detection. The simple answer here is "probably not"; historically synthesis techniques have always been able to evolve and avoid detection. What then makes the detection techniques interesting? The democratization of deep-fake synthesis tools is one of major concerns around these sophisticated manipulated media. The ease of availability of these tools enable an average user to create, for example, a deep-fake video of a celebrity or a national leader saying or doing things they never said or did. I believe the detection techniques, like the ones presented in this dissertation, mitigate this problem by discovering non-trivial inconsistencies in the deep fakes. This makes it more difficult and time-consuming for an average person to create a convincing fake that can fool an expert's eye. The forensic techniques, which are computed over long temporal windows, provide an edge over the per-frame synthesis methods that fail to simulate temporal consistency. Additionally, each of the high-level features used here are shown to be robust against the common laundering operations like compression or resizing. This robustness is valuable for the analysis of online videos which are often poor quality and low resolution.

Given the limited scalability of forensic techniques and poor performance of the human visual system, let us now briefly discuss other potential solutions. The problem of deep fakes is multifaceted and is only a small part of the larger misinformation ecosystem. Even though forensic techniques provide valuable tools for expert analysis, there are other problems: 1) the widespread availability of sophisticated synthesis techniques, 2) the promotion of fake content on social media platforms reaching billions, and 3) consequently, the growing mistrust in all forms of media threatening the existence of our society and democracy. The main stakeholders in this ecosystem are the creators, the publishers, and the consumers, each of which needs to take cohesive actions to fight this plague of misinformation. Described below is one potential measure that can bring all these pieces together.

The Content Authenticity Initiative (CAI) [32], in collaboration with Adobe, Twitter, Qualcomm, and Truepic, is developing a standard for media attribution and provenance. The new standard will enable trustworthy media with verifiable information including when, where, and how the media was created and edited. This glass-to-glass attribution, starting from the camera lens and ending at the user's screen, will be associated with all pieces of recorded media. Because this approach has shifted the burden from the consumer to the producer, it holds larger potential to operate at internet scale. These efforts, however, are only effective if they are adopted by all stakeholders in the information ecosystem. The social media platforms, who currently favor user engagement over content authenticity, need to adjust their approach and promote truth. Their recommendation algorithms can use the above attribution to favor authoritative and trustworthy content. Content creators need to become more accustomed to creating verifiable content through secure hardwares and softwares. Lastly, as consumers of online content, we all need to be responsible towards seeking authentic content and validating before sharing. Going a step further, even the researchers and the developers of AI tools can proactively adopt solutions to mitigate the misuse of their technology. For instance, artificially fingerprinting the generative models can help in identifying and tracing the source of GAN images and deep-fake videos [33, 34]. I believe a shared understanding of truth in imagery is essential for everyone to regain trust in the media we see every day.

Bibliography

[1] Pew Research Center, “More than eight-in-ten Americans get news from digital devices..” https://pewrsr.ch/2MZqns7.

[2] ABC News, “Facebook removes seven million posts for sharing false information on coronavirus.” https://www.nbcnews.com/tech/tech-news/facebook-removes-seven-million-posts-sharing-false-information-coronav-rcna77.

[3] The German Marshall Fund of the United States, “New Study by Digital New Deal Finds Engagement with Deceptive Outlets Higher on Facebook Today Than Run-up to 2016 Election..” https://shar.es/aoPPto.

[4] P. Moravec, R. Minas, and A. R. Dennis, “Fake news on social media: People believe what they want to believe when it makes no sense at all,” *Kelley School of Business Research Paper*, 2018.

[5] The Wall Street Journal, “Facebook Executives Shut Down Efforts to Make the Site Less Divisive.” https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?reflink=desktopwebshare-permalink.

[6] B. Chesney and D. Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security,” *Calif. L. Rev.*, 2019.

[7] C. Bregler, M. Covell, and M. Slaney, “Video rewrite: Driving visual speech with audio,” in *Conference on Computer Graphics and Interactive Techniques*, 1997.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in *Conference on Neural Information Processing Systems*, 2014.

[9] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2018.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *IEEE International Conference on Computer Vision*, 2017.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[17] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject agnostic face swapping and reenactment,” in *IEEE International Conference on Computer Vision*, 2019.

[18] “Faceswap.” https://github.com/MarekKowalski/FaceSwap.

[19] “Faceswap-GAN.” https://github.com/shaoanlu/faceswap-GAN.

[20] “Deepfakes faceswap.” https://github.com/deepfakes/faceswap.

[21] J. F. Boylan, “Will deep-fake technology destroy democracy?.” https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html.

[22] D. Harwell, “Scarlett Johansson on fake AI-generated sex videos: ‘nothing can stop someone from cutting and pasting my image’.” https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image.

[23] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *ACM Transactions on Graphics*, 2019.

[24] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning lip sync from audio,” *ACM Transactions on Graphics*, 2017.

[25] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM International Conference on Multimedia*, 2020.

[26] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics*, 2019.

[27] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, and H. Li, "PaGAN: Real-time avatars using dynamic textures," *ACM Transactions on Graphics*, 2018.

[28] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[29] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European Conference on Computer Vision*, 2020.

[30] 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.

[31] H. Farid, *Photo Forensics.* MIT press, 2016.

[32] "Content Authenticity Initiative." https://contentauthenticity.org/.

[33] N. Yu, V. Skripniuk, D. Chen, L. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," *arXiv preprint arXiv: 2012.08726*, 2021.

[34] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial GAN fingerprints: Rooting deepfake attribution in training data," *arXiv preprint arXiv: 2007.08457*, 2020.

[35] S. Voloshynovskiy, S. Pereira, T. Pun, J. J. Eggers, and J. K. Su, "Attacks on digital watermarks: classification, estimation based attacks, and benchmarks," *IEEE Communications Magazine*, 2001.

[36] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[37] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *IEEE International Conference on Computer Vision*, 2018.

[38] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and simulating artifacts in gan fake images,” in *IEEE International Workshop on Information Forensics and Security*, 2019.

[39] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to spot...for now,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[40] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[41] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for more general face forgery detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[42] M. Huh, A. Liu, A. Owens, and A. A. Efros, “Fighting fake news: Image splice detection via learned self-consistency,” in *European Conference on Computer Vision*, 2018.

[43] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *IEEE Symposium on Security and Privacy*, 2017.

[44] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[45] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing AI created fake videos by detecting eye blinking,” in *IEEE International Workshop on Information Forensics and Security*, 2018.

[46] U. A. Ciftci, I. Demir, and L. Yin, “FakeCatcher: Detection of synthetic portrait videos using biological signals,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[47] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman, “Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification,” in *IEEE International Conference on Multimodal Interfaces*, 2002.

[48] G. Williams, G. Taylor, K. Smolskiy, and C. Bregler, “Body motion analysis for multi-modal identity verification,” in *IEEE International Conference on Pattern Recognition*, 2010.

[49] Vice, “Twitter Accounts With AI-Generated Cat Avatars at Center of Turkish Porn Bot Ring.” https://www.vice.com/en/article/z3v579/twitter-accounts-with-ai-generated-cat-avatars-at-center-of-turkish-porn-bot-ring.

[50] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[51] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, 1976.

[52] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, 2001.

[53] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *IEEE International Conference on Computer Vision*, 2019.

[54] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[55] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," 2015.

[56] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," 2018.

[57] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[58] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[59] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deep-fake videos from appearance and behavior," in *IEEE International Workshop on Information Forensics and Security*, 2020.

[60] O. Wiles, A. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," in *British Machine Vision Conference*, 2018.

[61] N. Dufour and A. Gully, "Contributing Data to Deepfake Detection Research." https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[62] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[63] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[64] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[65] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, 2018.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[67] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[68] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security*, 2018.

[69] S. Agarwal and H. Farid, "Detecting deep fakes from aural and oral dynamics," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[70] G. J. Walker-Smith, A. G. Gale, and J. M. Findlay, "Eye movement strategies involved in face perception," *Perception*, 1977.

[71] J. M. Henderson, C. C. Williams, and R. J. Falk, "Eye movements are functional during face learning," *Memory and Cognition*, 2005.

[72] S. W. Janik, A. R. Wellens, M. L. Goldberg, and L. F. Dell'Osso, "Eyes as the center of focus in the visual examination of human faces," *Perceptual and Motor Skills*, 1978.

[73] M. Burge and W. Burger, "Ear biometrics," in *Biometrics*, 1996.

[74] M. Burge and W. Burger, "Ear biometrics in computer vision," in *IEEE International Conference on Pattern Recognition*, 2000.

[75] B. Victor, K. Bowyer, and S. Sarkar, "An evaluation of face and ear biometrics," in *Object Recognition Supported by User Interaction for Service Robots*, 2002.

[76] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Computing Surveys*, 2013.

[77] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, 2017.

[78] R. J. Oliveira, B. Hammer, A. Stillman, J. Holm, C. Jons, and R. H. Margolis, "A look at ear canal changes with jaw motion," *Ear and Hearing*, 1992.

[79] M. J. Grenness, J. Osborn, and W. L. Weller, "Mapping ear canal movement using area-based surface matching," *The Journal of the Acoustical Society of America*, 2002.

[80] S. Darkner, R. Larsen, and R. R. Paulsen, "Analysis of deformation of the human ear and canal caused by mandibular movement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2007.

[81] G. Salomon and A. Starr, "Electromyography of middle ear muscles in man during motor activities," *Acta Neurologica Scandinavica*, 1963.

[82] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, 2009.

[83] C. Tomasi and T. Kanade, "Detection and tracking of point," tech. rep., CMU-CS-91-132, Carnegie, Mellon University, 1991.

[84] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Python in Science Conference*, 2015.

[85] Y. Zhou and S. Zaferiou, "Deformable models of ears in-the-wild for alignment and recognition," 2017.

[86] H. Dai, N. Pears, and W. Smith, "A data-augmented 3D morphable model of the ear," 2018.

[87] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, "A deep learning approach for person identification using ear biometrics," *Applied Intelligence*, 2020.

[88] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[89] deeptomcruise. https://www.tiktok.com/@deeptomcruise.

[90] A. Abaza and A. Ross, "Towards understanding the symmetry of human ears: A biometric perspective," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2010.

[91] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[92] "Annosoft lipsync tool." http://www.annosoft.com/docs/Visemes17.html.

[93] "Google speech-to-text." https://cloud.google.com/speech-to-text/docs.

[94] S. Rubin, F. Berthouzoz, G. J. Mysore, W. Li, and M. Agrawala, "Content-based tools for editing audio stories," in *ACM Symposium on User Interface Software and Technology*, 2013.

[95] S. J. Nightingale, S. Agarwal, and H. Farid, "Perceptual and computational detection of face morphing," *Journal of Vision*, 2021.

[96] CNN Business, "A high school student created a fake 2020 candidate. Twitter verified it." https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html.

[97] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *European Conference on Computer Vision*, 2020.

[98] V. Bruce, Z. Henderson, K. Greenwood, P. J. Hancock, A. M. Burton, and P. Miller, "Verification of face identities from images captured on video.," *Journal of Experimental Psychology: Applied*, 1999.

[99] A. M. Megreya and A. M. Burton, "Hits and false positives in face matching: A familiarity-based dissociation," *Perception and Psychophysics*, 2007.

[100] A. M. Burton, D. White, and A. McNeill, "The glasgow face matching test," *Behavior Research Methods*, 2010.

[101] D. White, R. I. Kemp, R. Jenkins, M. Matheson, and A. M. Burton, "Passport officers' errors in face matching," *Public Library of Science One*, 2014.

[102] D. J. Robertson, R. S. Kramer, and A. M. Burton, "Fraudulent ID using face morphs: Experiments on human and automatic recognition," *Public Library of Science One*, 2017.

[103] D. J. Robertson, A. Mungall, D. G. Watson, K. A. Wade, S. J. Nightingale, and S. Butler, "Detecting morphed passport photos: a training and individual differences approach," *Cognitive Research: Principles and Implications*, 2018.

[104] R. S. Kramer, M. O. Mireku, T. R. Flack, and K. L. Ritchie, "Face morphing attacks: Investigating detection with humans and computers," *Cognitive Research: Principles and Implications*, 2019.

[105] Medium.com, "NVIDIA Open-Sources Hyper-Realistic Face Generator StyleGAN." https://medium.com/syncedreview/nvidia-open-sources-hyper-realistic-face-generator-stylegan-f346e1a73826.

[106] The Verge, "Can you tell the difference between a real face and an AI-generated fake?." https://www.theverge.com/2019/3/3/18244984/ai-generated-fake-which-face-is-real-test-stylegan.

[107] P. Korshunov and S. Marcel, "Deepfake detection: humans vs. machines," *arXiv preprint arXiv:2009.03155*, 2020.

[108] T. Tariq, O. T. Tursun, M. Kim, and P. Didyk, "Why are deep representations good perceptual quality features?," in *European Conference on Computer Vision*, 2020.

[109] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[110] J. P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, 1998.

[111] J. P. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[112] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *International Conference on Audio-and Video-Based Biometric Person Authentication*, Springer, 2003.

[113] B. Weyrauch, B. Heisele, J. Huang, and V. Blanz, "Component-based face recognition with 3d morphable models," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

[114] J. P. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[115] B. Azam, A. N. Melika, and M. D. Mohammad, "Iranian face database with age, pose and expression," in *International Conference on Machine Vision*, 2007.

[116] A. Kasiński, A. Florek, and A. Schmidt, "The put face database," *Image Processing and Communications*, 2008.

[117] "Utrecht ECVP face database." `http://pics.stir.ac.uk/2D_face_sets.htm`.

[118] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[119] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, 2010.

[120] C. I. Watson, "Multiple Encounter Dataset I (MEDS-I)," *NIST Interagency/Internal Report*, 2010.

[121] T. F. Vieira, A. Bottino, A. Laurentini, and M. De Simone, "Detecting siblings in image pairs," *The Visual Computer*, 2014.

[122] D. S. Ma, J. Correll, and B. Wittenbrink, "The Chicago face database: A free stimulus set of faces and norming data," *Behavior Research Methods*, 2015.

[123] N. Strohminger, K. Gray, V. Chituc, J. Heffner, C. Schein, and T. B. Heagins, "The MR2: A multi-racial, mega-resolution database of facial stimuli," *Behavior Research Methods*, 2016.

[124] L. DeBruine and B. Jones, "Face research lab london set," 2017.

[125] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Science and Business Media, 2010.

[126] J. Tanaka, M. Giles, S. Kremen, and V. Simon, "Mapping attractor fields in face space: the atypicality bias in face recognition," *Cognition*, 1998.

[127] G. Cumming, F. Fidler, P. Kalinowski, and J. Lai, "The statistical recommendations of the american psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis," *Australian Journal of Psychology*, 2012.

[128] G. Cumming, *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, 2013.

[129] R. I. Kemp, A. Caon, M. Howard, and K. R. Brooks, "Improving unfamiliar face matching by masking the external facial features," *Applied Cognitive Psychology, year=2016, publisher=Wiley Online Library*.

[130] A. Towler, D. White, and R. I. Kemp, "Evaluating the feature comparison strategy for forensic face identification.," *Journal of Experimental Psychology: Applied*, 2017.

[131] R. Hastie and T. Kameda, "The robust beauty of majority rules in group decisions.," *Psychological Review*, 2005.

[132] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, 2012.

[133] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018.

[134] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.