usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos

Yi Xu, True Price, Jan-Michael Frahm, and Fabian Monrose,
The University of North Carolina at Chapel Hill

https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/xu

This paper is included in the Proceedings of the
25th USENIX Security Symposium

August 10–12, 2016 • Austin, TX

ISBN 978-1-931971-32-4

Virtual U: Defeating Face Liveness Detection by Building Virtual Models From Your Public Photos

Yi Xu, True Price, Jan-Michael Frahm, Fabian Monrose

Department of Computer Science, University of North Carolina at Chapel Hill

{yix, jtprice, jmf, fabian}@cs.unc.edu

Abstract

In this paper, we introduce a novel approach to bypass modern face authentication systems. More specifically, by leveraging a handful of pictures of the target user taken from social media, we show how to create realistic, textured, 3D facial models that undermine the security of widely used face authentication solutions. Our framework makes use of virtual reality (VR) systems, incorporating along the way the ability to perform animations (*e.g.,* raising an eyebrow or smiling) of the facial model, in order to trick liveness detectors into believing that the 3D model is a real human face. The synthetic face of the user is displayed on the screen of the VR device, and as the device rotates and translates in the real world, the 3D face moves accordingly. To an observing face authentication system, the depth and motion cues of the display match what would be expected for a human face.

We argue that such VR-based spoofing attacks constitute a fundamentally new class of attacks that point to a serious weaknesses in camera-based authentication systems: Unless they incorporate other sources of verifiable data, systems relying on color image data and camera motion are prone to attacks via virtual realism. To demonstrate the practical nature of this threat, we conduct thorough experiments using an end-to-end implementation of our approach and show how it undermines the security of several face authentication solutions that include both motion-based and liveness detectors.

1 Introduction

Over the past few years, face authentication systems have become increasingly popular as an enhanced security feature in both mobile devices and desktop computers. As the underlying computer vision algorithms have matured, many application designers and nascent specialist vendors have jumped in and started to offer solutions for mobile devices with varying degrees of security and usability. Other more well-known players, like Apple and Google, are posed to enter the market with their own solutions, having already acquired several facial recognition software companies[1]. While the market is segmented based on the type of technology offered (*e.g.,* 2D facial recognition, 3D recognition, and facial analytics/face biometric authentication), Gartner research estimates that the overall market will grow to over $6.5 billion in 2018 (compared to roughly $2 billion today) [13].

With this push to market, improving the accuracy of face recognition technologies remains an active area of research in academia and industry. Google's FaceNet system, which achieved near-perfect accuracy on the Labeled Faces in the Wild dataset [47], exemplifies one such effort. Additionally, recent advances with deep learning algorithms [38, 53] show much promise in strengthening the robustness of the face identification and authentication techniques used today. Indeed, state-of-the-art face identification systems can now outperform their human counterparts [36], and this high accuracy is one of the driving factors behind the increased use of face recognition systems.

However, even given the high accuracy of modern face recognition technologies, their application in face authentication systems has left much to be desired. For instance, at the Black Hat security conference in 2009, Duc and Minh [10] demonstrated the weaknesses of popular face authentication systems from commodity vendors like Lenovo, Asus, and Toshiba. Amusingly, Duc and Minh [10] were able to reliably bypass face-locked computers simply by presenting the software with photographs and fake pictures of faces. Essentially, the security of these systems rested solely on the problem of face detection, rather than face authentication. This widely publicized event led to subsequent integration of more robust face authentication protocols. One prominent example is Android OS, which augmented its face authen-

[1] See, for example, "Apple Acquires Face Recognition, Expression Analysis firm, Emotient", TechTimes, Jan, 2016; "Google Acquires Facial Recognition Software Company PittPar," WSJ, 2011.

tication approach in 2012 to require users to blink while authenticating (*i.e.,* as a countermeasure to still-image spoofing attacks). Unfortunately, this approach was also shown to provide little protection, and can be easily bypassed by presenting the system with two alternating images — one with the user's eyes open, and one with her eyes closed.[2] These attacks underscore the fact that face authentication systems require robust security features beyond mere recognition in order to foil spoofing attacks.

Loosely speaking, three types of such spoofing attacks have been used in the past, to varying degrees of success: (*i*) still-image-based spoofing, (*ii*) video-based spoofing, and (*iii*) 3D-mask-based spoofing. As the name suggests, still-image-based spoofing attacks present one or more still images of the user to the authentication camera; each image is either printed on paper or shown with a digitized display. Video-based spoofing, on the other hand, presents a pre-recorded video of the victim's moving face in an attempt to trick the system into falsely recognizing motion as an indication of liveness. The 3D-mask-based approach, wherein 3D-printed facial masks are used, was recently explored by Erdogmus and Marcel [11].

As is the typical case in the field of computer security, the cleverness of skilled, motivated adversaries drove system designers to incorporate defensive techniques in the biometric solutions they develop. This cat-and-mouse game continues to play out in the realm of face authentication systems, and the current recommendation calls for the use of well-designed face liveness detection schemes (that attempt to distinguish a real user from a spoofed one). Indeed, most modern systems now require more active participation compared to simple blink detection, often asking the user to rotate her head or raise an eyebrow during login. Motion-based techniques that check, for example, that the input captured during login exhibits sufficient 3D behavior, are also an active area of research in face authentication.

One such example is the recent work of Li et al. [34] that appeared in CCS'2015. In that work, the use of liveness detection was proposed as a solution to thwarting video-based attacks by checking the consistency of the recorded data with inertial sensors. Such a detection scheme relies on the fact that as a camera moves relative to a user's stationary head, the facial features it detects will also move in a predictable way. Thus, a 2D video of the victim would have to be captured under the exact same camera motion in order to fool the system.

As mentioned in [34], 3D-printed facial reconstructions offer one option for defeating motion-based liveness detection schemes. In our view, a more realizable approach is to present the system with a 3D facial mesh in a virtual reality (VR) environment. Here, the motion of the authenticating camera is tracked, and the VR system internally rotates and translates the mesh to match. In this fashion, the camera observes exactly the same movement of facial features as it would for a real face, fulfilling the requirements for liveness detection. Such an attack defeats color-image- and motion-based face authentication on a fundamental level because, with sufficient effort, a VR system can display an environment that is essentially indistinguishable from real-world input.

In this paper, we show that it is possible to undermine modern face authentication systems using one such attack. Moreover, we show that an accurate facial model can be built using *only* a handful of publicly accessible photos — collected, for example, from social network websites — of the victim. From a pragmatic point of view, we are confronted with two main challenges: *i*) the number of photos of the target may be limited, and *ii*) for each available photo, the illumination setting is unknown and the user's pose and expression are not constrained. To overcome these challenges, we leverage robust, publicly available 3D face reconstruction methods from the field of computer vision, and adapt these techniques to fit our needs. Once a credible synthetic model of a user is obtained, we then employ entry-level virtual reality displays to defeat the state of the art in liveness detection.

The rest of the paper is laid out as follows: §2 provides background and related work related to face authentication, exploitation of users' online photos, and 3D facial reconstruction. §3 outlines the steps we take to perform our VR-based attack. In §4, we evaluate the performance of our method on 5 commercial face authentication systems and, additionally, on a proposed state-of-the-art system for liveness detection. We suggest steps that could be taken to mitigate our attack in §5, and we address the implications of our successful attack strategy in §6.

2 Background and Related Work

Before delving into the details of our approach, we first present pertinent background information needed to understanding the remainder of this paper.

First, we note that given the three prominent classes of spoofing attacks mentioned earlier, it should be clear that while still-image-based attacks are the easiest to perform, they can be easily countered by detecting the 3D structure of the face. Video-based spoofing is more difficult to accomplish because facial videos of the target user may be harder to come by; moreover, such attacks can also be successfully defeated, for example, using the recently suggested techniques of Li et al. [34] (which we discuss in more detail later). 3D-mask-based approaches, on the other hand, are harder to counter. That said, building a 3D mask is arguably more time-consuming and also requires specialized equipment. Nevertheless, because

[2] https://www.youtube.com/watch?v=zYxphDK6s3I

of the threat this attack vector poses, much research has gone into detecting the textures of 3D masks [11].

2.1 Modern Defenses Against Spoofing

Just as new types of spoofing attacks have been introduced to fool face authentication systems, so too have more advanced methods for countering these attacks been developed. Nowadays, the most popular liveness detection techniques can be categorized as either texture-based approaches, motion-based approaches, or liveness assessment approaches. We discuss each in turn.

Texture-based approaches [11, 25, 37, 40, 54, 60] attempt to identify spoofing attacks based on the assumption that a spoofed face will have a distinctly different texture from a real face. Specifically, they assume that due to properties of its generation, a spoofed face (irrespective of whether it is printed on paper, shown on a display, or made as a 3D mask) will be different from a real face in terms of shape, detail, micro-textures, resolution, blurring, gamma correction, and shading. That is, these techniques rely on perceived limitations of image displays and printing techniques. However, with the advent of high-resolution displays (*e.g.,* 5K), the difference in visual quality between a spoofed image and a living face is hard to notice. Another limitation is that these techniques often require training on every possible spoofing material, which is not practical for real systems.

Motion-based approaches [3, 27, 29, 32, 57] detect spoofing attacks by using motion of the user's head to infer 3D shape. Techniques such as optical flow and focal-length analysis are typically used. The basic assumption is that structures recovered from genuine faces usually contain sufficient 3D information, whereas structures from fake faces (photos) are usually planar in depth. For instance, the approach of Li et al. [34] checks the consistency of movement between the mobile device's internal motion sensors and the observed change in head pose computed from the recorded video taken while the claimant attempts to authenticate herself to the device. Such 3D reasoning provides a formidable defense against both still-image and video-based attacks.

Lastly, liveness assessment techniques [19, 30, 31, 49] require the user to perform certain tasks during the authentication stage. For the systems we evaluated, the user is typically asked to follow certain guidelines during registration, and to perform a random series of actions (*e.g.,* eye movement, lip movement, and blinking) at login. The requested gestures help to defeat contemporary spoofing attacks.

Take-away: For real-world systems, liveness detection schemes are often combined with motion-based approaches to provide better security protection than either can provide on their own. With these ensemble techniques, traditional spoofing attacks can be reliably detected. For that reason, the combination of motion-based systems and liveness detectors has gained traction and is now widely adopted in many commercial systems, including popular face authentication systems offered by companies like KeyLemon, Rohos, and Biomids. For the remainder of this paper, we consider this combination as the state of the art in defenses against spoofing attacks for face authentication systems.

2.2 Online Photos and Face Authentication

It should come as no surprise that personal photos from online social networks can compromise privacy. Major social network sites advise users to set privacy settings for the images they upload, but the vast majority of these photos are often accessible to the public or set to 'friend-only' viewing' [14, 26, 35]. Users also do not have direct control over the accessibility of photos of themselves posted by other users, although they can remove ('untag') the association of such photos with their account.

A notable use of social network photos for online security is Facebook's social authentication (SA) system [15], an extension of CAPTCHAs that seeks to bolster identity verification by requiring the user to identify photos of their friends. While this method does require more specific knowledge than general CAPTCHAs, Polakis et al. [42] demonstrated that facial recognition could be applied to a user's public photos to discover their social relationships and solve 22% of SA tests automatically.

Given that one's online photo presence is not entirely controlled by the user alone — but by their collective social circles — many avenues exist for an attacker to uncover the facial appearance of a user, even when the user makes private their own personal photos. In an effort to curb such easy access, work by Ilia et al. [17] has explored the automatic privatization of user data across a social network. This method uses face detection and photo tags to selectively blur the face of a user when the viewing party does not have permission to see the photo. In the future, such an approach may help decrease the public accessibility of users' personal photos, but it is unlikely that an individual's appearance can ever be completely obfuscated from attackers across all social media sites and image stores on the Internet.

Clearly, the availability of online user photos is a boon for an adversary tasked with the challenge of undermining face authentication systems. The most germane on this front is the work of Li et al. [33]. There, the authors proposed an attack that defeated commonly used face authentication systems by using photos of the target user gathered from online social networks. Li et al. [33] reported that 77% of the users in their test set were vulnerable to their proposed attack. However, their work is targeted at face recognition systems that *do not incorporate face liveness detection*. As noted in §2, in modern face authentication software, sophisticated liveness detection approaches are already in use, and these techniques thwart still-image spoofing attacks of the kind performed by Li et al. [33].

2.3 3D Facial Reconstruction

Constructing a 3D facial model from a small number of personal photos involves the application of powerful techniques from the field of computer vision. Fortunately, there exists a variety of reconstruction approaches that make this task less daunting than it may seem on first blush, and many techniques have been introduced for facial reconstruction from single images [4, 23, 24, 43], videos [20, 48, 51], and combinations of both [52]. For pedagogical reasons, we briefly review concepts that help the reader better understand our approach.

The most popular facial model reconstruction approaches can be categorized into three classes: shape from shading (SFS), structure from motion (SFM) combined with dense stereoscopic depth estimation, and statistical facial models. The SFS approach [24] uses a model of scene illumination and reflectance to recover face structure. Using this technique, a 3D facial model can be reconstructed from only a single input photo. SFS relies on the assumption that the brightness level and gradient of the face image reveals the 3D structure of the face. However, the constraints of the illumination model used in SFS require a relatively simple illumination setting and, therefore, cannot typically be applied to real-world photo samples, where the configuration of the light sources is unknown and often complicated.

As an alternative, the structure from motion approach [12] makes use of multiple photos to triangulate spatial positions of 3D points. It then leverages stereoscopic techniques across the different viewpoints to recover the complete 3D surface of the face. With this method, the reconstruction of a dense and accurate model often requires many consistent views of the surface from different angles; moreover, non-rigid variations (*e.g.,* facial expressions) in the images can easily cause SFM methods to fail. In our scenario, these requirements make such an approach less usable: for many individuals, only a limited number of images might be publicly available online, and the dynamic nature of the face makes it difficult to find multiple images having a consistent appearance (*i.e.,* the exact same facial expression).

Unlike SFS and SFM, statistical facial models [4, 43] seek to perform facial reconstruction on an image using a training set of existing facial models. The basis for this type of facial reconstruction is the 3D morphable model (3DMM) of Blanz and Vetter [6, 7], which learns the principal variations of face shape and appearance that occur within a population, then fits these properties to images of a specific face. Training the morphable models can be performed either on a controlled set of images [8, 39] or from internet photo-collections [23]. The underlying variations fall on a continuum and capture both expression (*e.g.,* a frowning-to-smiling spectrum) and identity (*e.g.,* a skinny-to-heavy or a male-to-female spectrum). In 3DMM and its derivatives, both 3D shape and texture information are cast into a high-dimensional linear space, which can be analyzed with principal component analysis (PCA) [22]. By optimizing over the weights of different eigenvectors in PCA, any particular human face model can be approximated. Statistical facial models have shown to be very robust and only require a few photos for high-precision reconstruction. For instance, the approach of Baumberger et al. [4] achieves good reconstruction quality using only two images.

To make the process fully automatic, recent 3D facial reconstruction approaches have relied on a few facial landmark points instead of operating on the whole model. These landmarks can be accurately detected using the supervised descent method (SDM) [59] or deep convolutional networks [50]. By first identifying these 2D features in an image and then mapping them to points in 3D space, the entire 3D facial surface can be efficiently reconstructed with high accuracy. In this process, the main challenge is the localization of facial landmarks within the images, especially contour landmarks (along the cheekbones), which are half-occluded in non-frontal views; we introduce a new method for solving this problem when multiple input images are available.

The end result of 3D reconstruction is a untextured (*i.e.,* lacking skin color, eye color, etc.) facial surface. Texturing is then applied using source image(s), creating a realistic final face model. We next detail our process for building such a facial model from a user's publicly available internet photos, and we outline how this model can be leveraged for a VR-based face authentication attack.

3 Our Approach

A high-level overview of our approach for creating a synthetic face model is shown in Figure 1. Given one or more photos of the target user, we first automatically extract the landmarks of the user's face (stage ❶). These landmarks capture the pose, shape, and expression of the user. Next, we estimate a 3D facial model for the user, optimizing the geometry to match the observed 2D landmarks (stage ❷). Once we have recovered the shape of the user's face, we use a single image to transfer texture information to the 3D mesh. Transferring the texture is non-trivial since parts of the face might be self-occluded

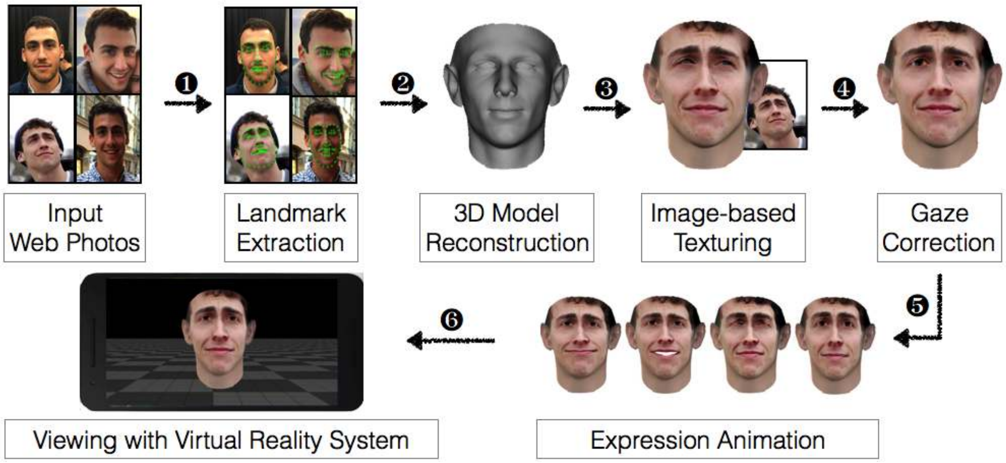

Figure 1: Overview of our proposed approach.

(*e.g.,* when the photo is taken from the side). The texture of these occluded parts must be estimated in a manner that does not introduce too many artifacts (stage ❸). Once the texture is filled, we have a realistic 3D model of the user's face based on a single image.

However, despite its realism, the output of stage ❸ is still not able to fool modern face authentication systems. The primary reason for this is that modern face authentication systems use the subject's gaze direction as a strong feature, requiring the user to look at the camera in order to pass the system. Therefore, we must also automatically correct the direction of the user's gaze on the textured mesh (stage ❹). The adjusted model can then be deformed to produce animation for different facial expressions, such as smiling, blinking, and raising the eyebrows (stage ❺). These expressions are often used as liveness clues in face authentication systems, and as such, we need to be able to automatically reproduce them on our 3D model. Finally, we output the textured 3D model into a virtual reality system (stage ❻).

Using this framework, an adversary can bypass both the face recognition and liveness detection components of modern face authentication systems. In what follows, we discuss the approach we take to solve each of the various challenges that arise in our six-staged process.

3.1 Facial Landmark Extraction

Starting from multiple input photos of the user, our first task is to perform facial landmark extraction. Following the approach of Zhu et al. [63], we extract 68 2D facial landmarks in each image using the supervised descent method (SDM) [59]. SDM successfully identifies facial landmarks under relatively large pose differences (± 45deg yaw, ± 90deg roll, ± 30deg pitch). We chose the technique of Zhu et al. [63] because it achieves a median alignment error of 2.7 pixels on well-known datasets [1] and outperforms other commonly used techniques (*e.g.,* [5]) for landmark extraction.

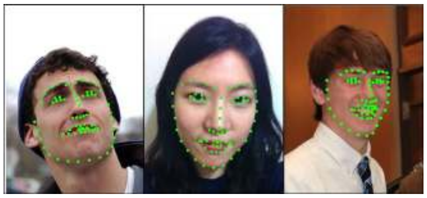

Figure 2: Examples of facial landmark extraction

For our needs, SDM works well on most online images, even those where the face is captured at a low resolution (*e.g.,* 40×50 pixels). It does, however, fail on a handful of the online photos we collected (less than 5%) where the pose is beyond the tolerance level of the algorithm. If this occurs, we simply discard the image. In our experiments, the landmark extraction results are manually checked for correctness, although an automatic scoring system could potentially be devised for this task. Example landmark extractions are shown in Figure 2.

3.2 3D Model Reconstruction

The 68 extracted 3D point landmarks from each of the N input images provide us with a set of coordinates $s_{i,j} \in \mathbb{R}^2$, with $1 \leq i \leq 68, 1 \leq j \leq N$. The projection of the 3D points $S_{i,j} \in \mathbb{R}^3$ on the face onto the image coordinates $s_{i,j}$ follows what is called the "weak perspective projection" (WPP) model [16], computed as follows:

$$s_{i,j} = f_j P R_j \left(S_{i,j} + t_j\right), \qquad (1)$$

where f_j is a uniform scaling factor; P is the projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$; R_j is a 3×3 rotation matrix defined by the pitch, yaw, and roll, respectively, of the face relative to the camera; and $t_j \in \mathbb{R}^3$ is the translation of the face with respect to the camera. Among these parameters, only $s_{i,j}$ and P are known, and so we must estimate the others.

Fortunately, a large body of work exists on the shape statistics of human faces. Following Zhu et al. [63], we capture face characteristics using the 3D Morphable Model (3DMM) [39] with an expression extension proposed by Chu et al. [9]. This method characterizes variations in face shape for a population using principal component analysis (PCA), with each individual's 68 3D point landmarks being concatenated into a single feature vector for the analysis. These variations can be split into two categories: constant factors related to an individual's distinct appearance (identity), and non-constant factors related to expression. The identity axes capture characteristics such as face width, brow placement, or lip size, while the expression axes capture variations like smiling versus frowning. Example axes for variations in expression are shown in Figure 6.

More formally, for any given individual, the 3D coordinates $S_{i,j}$ on the face can be modeled as

$$S_{i,j} = \bar{S}_i + A_i^{id}\alpha^{id} + A_i^{exp}\alpha_j^{exp}, \qquad (2)$$

where $\bar{S}_i$ is the statistical average of $S_{i,j}$ among the individuals in the population, A_i^{id} is the set of principal axes of variation related to identity, and A_i^{exp} is the set of principal axes related to expression. α^{id} and α_j^{exp} are the identity and expression weight vectors, respectively, that determine *person-specific* facial characteristics and expression-specific facial appearance. We obtain $\bar{S}_i$ and A_i^{id} using the 3D Morphable Model [39] and A_i^{exp} from Face Warehouse [8].

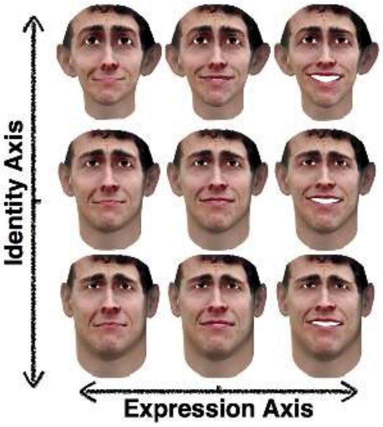

Figure 3: Illustration of identity axes (heavy-set to thin) and expression axes (pursed lips to open smile).

When combining Eqs. (1) and (2), we inevitably run into the so-called "correspondence problem." That is, given each identified facial landmark $s_{i,j}$ in the input image, we need to find the corresponding 3D point $S_{i',j}$ on the underlying face model. For landmarks such as the corners of the eyes and mouth, this correspondence is self-evident and consistent across images. However, for contour landmarks marking the edge of the face in an image, the associated 3D point on the user's facial model is pose-dependent: When the user is directly facing the camera, their jawline and cheekbones are fully in view, and the observed 2D landmarks lie on the fiducial boundary on the user's 3D facial model. When the user rotates their face left (or right), however, the previously observed 2D contour landmarks on the left (resp. right) side of the face shift out of view. As a result, the observed 2D landmarks on the edge of the face correspond to 3D points closer to the center of the face. This 3D point displacement must be taken into account when recovering the underlying facial model.

Qu et al. [44] deal with contour landmarks using constraints on surface normal direction, based on the observation that points on the edge of the face in the image will have surface normals perpendicular to the viewing direction. However, this approach is less robust because the normal direction cannot always be accurately estimated and, as such, requires careful parameter tuning. Zhu et al. [63] proposed a "landmark marching" scheme that iteratively estimates 3D head pose and 2D contour landmark position. While their approach is efficient and robust against different face angles and surface shapes, it can only handle a single image and cannot refine the reconstruction result using additional images.

Our solution to the correspondence problem is to model 3D point variance for each facial landmark using a pre-trained Gaussian distribution (see Appendix A). Unlike the approach of Zhu et al. [63] which is based on single image input, we solve for pose, perspective, expression, and neutral-expression parameters over *all* images jointly. From this, we obtain a neutral-expression model S_i of the user's face. A typical reconstruction, S_i, is presented in Figure 4.

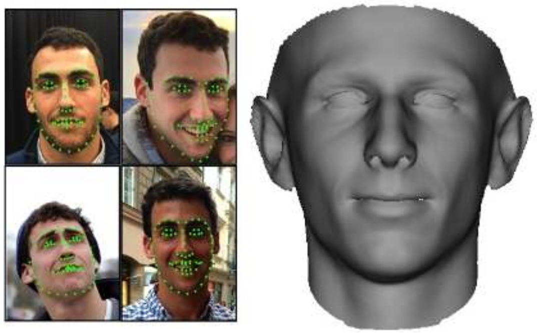

Figure 4: 3D facial model (right) built from facial landmarks extracted from 4 images (left).

3.3 Facial Texture Patching

Given the 3D facial model, the next step is to patch the model with realistic textures that can be recognized by the face authentication systems. Due to the appearance variation across social media photos, we have to achieve this by mapping the pixels in a *single* captured photo onto the 3D facial model, which avoids the challenges of mixing different illuminations of the face. However, this still leaves many of the regions without texture, and those untextured spots will be noticeable to modern face authentication systems. To fill these missing regions, the naïve approach is to utilize the vertical symmetry of the face and fill the missing texture regions with their symmetrical complements. However, doing so would lead to strong artifacts at the boundary of missing regions. A realistic textured model should be free of these artifacts.

To lessen the presence of these artifacts, one approach is to iteratively average the color of neighboring vertices as a color trend and then mix this trend with texture details [45]. However, such an approach over-simplifies the problem and fails to realistically model the illumination of facial surfaces. Instead, we follow the suggestion of Zhu et al. [63] and estimate facial illumination using spherical harmonics [61], then fill in texture details with Poisson editing [41]. In this way, the output model will appears to have a more natural illumination. Sadly, we cannot use their approach directly as it reconstructs a planar normalized face, instead of a 3D facial model, and so we must extend their technique to the 3D surface mesh.

The idea we implemented for improving our initial textured 3D model was as follows: Starting from the single photo chosen as the main texture source, we first estimate and subsequently remove the illumination conditions present in the photo. Next, we map the textured facial model onto a plane via a conformal mapping, then impute the unknown texture using 2D Poisson editing. We further extend their approach to three dimensions and perform Poisson editing directly on the surface of the facial model. Intuitively, the idea behind Poisson editing is to keep the detailed texture in the editing region while enforcing the texture's smoothness across the boundary. This process is defined mathematically as

$$\Delta f = \Delta g, s.t f|_{\partial\Omega} = f^0|_{\partial\Omega}, \tag{3}$$

where Ω is the editing region, f is the editing result, f^0 is the known original texture value, and g is the texture value in the editing region that is unknown and needs to be patched with its reflection complement. On a 3D surface mesh, every vertex is connected with 2 to 8 neighbors. Transforming Eq. 3 into discrete form, we have

$$|N_p|f_p - \sum_{q\in N_p\cap\Omega} f_q = \sum_{q\in N_p\cap\overline{\Omega}} f_q^0 + (\Delta g)_p, \tag{4}$$

where N_p is the neighborhood of point p on the mesh. Our enhancement is a natural extension of the Poisson editing method suggested in the seminal work of Pérez et al. [41], although no formulation was given for 3D. By solving Eq. 4 instead of projecting the texture onto a plane and solving Eq. 3, we obtain more realistic texture on the facial model, as shown in Figure 5.

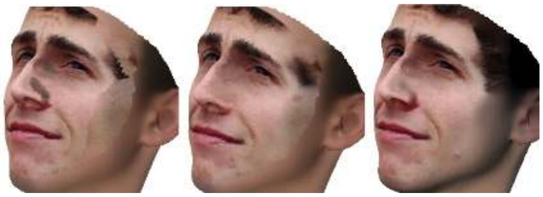

Figure 5: Naïve symmetrical patching (left); Planar Poisson editing (middle); 3D Poisson editing (right).

3.4 Gaze Correction

We now have a realistic 3D facial model of the user. Yet, we found that models at stage ❸ were unable to bypass most well-known face recognition systems. Digging deeper into the reasons why, we observed that most recognition systems rely heavily on gaze direction during authentication, *i.e.,* they fail-close if the user is not looking at the device. To address this, we introduce a simple, but effective, approach to correct the gaze direction of our synthetic model (Figure 1, Stage ❹).

The idea is as follows. Since we have already reconstructed the texture of the facial model, we can synthesize the texture data in the eye region. These data contain the color information from the sclera, cornea, and pupil and form a three-dimensional distribution in the RGB color space. We estimate this color distribution with a 3D Gaussian function whose three principle components can be computed as (b_1, b_2, b_3) with weight $(\sigma_1, \sigma_2, \sigma_3), \sigma_1 \geq \sigma_2 \geq \sigma_3 > 0$. We perform the same analysis for the eye region of the average face model obtained from 3DMM [39], whose eye is looking straight towards the camera, and we similarly obtain principle color components $(b_1^{std}, b_2^{std}, b_3^{std})$ with weight $(\sigma_1^{std}, \sigma_2^{std}, \sigma_3^{std}), \sigma_1^{std} \geq \sigma_2^{std} \geq \sigma_3^{std} > 0$. Then, we convert the eye texture from the average model into the eye texture of the user. For a texture pixel c in the eye region of average texture, we convert it to

$$c_{convert} = \sum_{i=1}^{3} \frac{\sigma_i}{\sigma_i^{std}} (c' b_i^{std}) b_i. \tag{5}$$

In effect, we align the color distribution of the average eye texture with the color distribution of the user's eye texture. By patching the eye region of the facial model with this converted average texture, we realistically capture the user's eye appearance with forward gaze.

3.5 Adding Facial Animations

Some of the liveness detection methods that we test require that the user performs specific actions in order to unlock the system. To mimic these actions, we can simply animate our facial model using a pre-defined set of facial expressions (*e.g.,* from FaceWarehouse [8]). Recall that in deriving in Eq. 2, we have already computed the weight for the identity axis α^{id}, which captures the user-specific face structure in a neutral expression. We can adjust the expression of the model by substituting a specific, known expression weight vector α^{exp}_{std} into Eq. 2. By interpolating the model's expression weight from 0 to α^{exp}_{std}, we are able to animate the 3D facial model to smile, laugh, blink, and raise the eyebrows (see Figure 6).

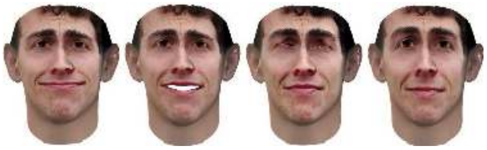

Figure 6: Animated expressions. From left to right: smiling, laughing, closing the eyes, and raising the eyebrows.

3.6 Leveraging Virtual Reality

While the previous steps were necessary to recover a realistic, animated model of a targeted user's face, our driving insight is that virtual reality systems can be leveraged to display this model as if it were a real, three-dimensional face. This VR-based spoofing constitutes a fundamentally new class of attacks that exploit weaknesses in camera-based authentication systems.

In the VR system, the synthetic 3D face of the user is displayed on the screen of the VR device, and as the device rotates and translates in the real world, the 3D face moves accordingly. To an observing face authentication system, the depth and motion cues of the display exactly match what would be expected for a human face. Our experimental VR setup consists of custom 3D-rendering software displayed on a Nexus 5X smart phone. Given the ubiquity of smart phones in modern society, our implementation is practical and comes at no additional hardware cost to an attacker. In practice, any device with similar rendering capabilities and inertial sensors could be used.

On smart phones, accelerometers and gyroscopes work in tandem to provide the device with a sense of self-motion. An example use case is detecting when the device is rotated from a portrait view to a landscape view, and rotating the display, in response. However, these sensors are not able to recover absolute *translation* — that is, the device is unable to determine how its position has changed in 3D space. This presents a challenge because without knowledge of how the device has moved in 3D space, we cannot move our 3D facial model in a realistic fashion. As a result, the observed 3D facial motion will not agree with the device's inertial sensors, causing our method to fail on methods like that of Li et al. [34] that use such data for liveness detection.

Fortunately, it is possible to track the 3D position of a moving smart phone using its outward-facing camera with structure from motion (see §2.3). Using the camera's video stream as input, the method works by tracking points in the surrounding environment (*e.g.,* the corners of tables) and then estimating their position in 3D space. At the same time, the 3D position of the camera is recovered relative to the tracked points, thus inferring the camera's change in 3D position. Several computer vision approaches have been recently introduced to solve this problem accurately and in real time on mobile devices [28, 46, 55, 56]. In our experiments, we make use of a printed marker[3] placed on a wall in front of the camera, rather than tracking arbitrary objects in the surrounding scene; however, the end result is the same. By incorporating this module into our proof of concept, the perspective of the viewed model due to camera translation can be simulated with high consistency and low latency.[4]

An example setup for our attack is shown in Figure 7. The VR system consists of a Nexus 5X unit using its outward-facing camera to track a printed marker in the environment. On the Nexus 5X screen, the system displays a 3D facial model whose perspective is always consistent with the spatial position and orientation of the authentication device. The authenticating camera views the facial model on the VR display, and it is successfully duped into believing it is viewing the real face of the user.

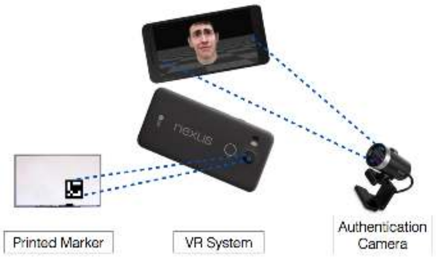

Figure 7: Example setup using virtual reality to mimic 3D structure from motion. The authentication system observes a virtual display of a user's 3D facial model that rotates and translates and the device moves. To recover the 3D translation of the VR device, an outward-facing camera is used to track a marker in the surrounding environment.

[3] See Goggle Paper at http://gogglepaper.com/

[4] Specialized VR systems such as the Oculus Rift could be used to further improve the precision and latency of camera tracking. Such advanced, yet easily obtainable, hardware has the potential to deliver even more sophisticated VR attacks compared to what is presented here.

4 Evaluation

We now demonstrate that our proposed spoofing method constitutes a significant security threat to modern face authentication systems. Using real social media photos from consenting users, we successfully broke five commercial authentication systems with a practical, end-to-end implementation of our approach. To better understand the threat, we further systematically run lab experiments to test the capabilities and limitations of our proposed method. Moreover, we successfully test our proposed approach with the latest motion-based liveness detection approach by Li et al. [34], which is not yet available in commercial systems.

Participants

We recruited 20 volunteers for our tests of commercial face authentication systems. The volunteers were recruited by word of mouth and span graduate students and faculty in two separate research labs. Consultation with our IRB departmental liaison revealed that no application was needed. There was no compensation for participating in the lab study. The ages of the participants range between 24 and 44 years, and the sample consists of 6 females and 14 males. The participants come from a variety of ethnic backgrounds (as stated by the volunteers): 6 are of Asian descent, 4 are Indian, 1 is African-American, 1 is Hispanic, and 8 are Caucasian. With their consent, we collected public photos from the users' Facebook and Google+ social media pages; we also collected any photos we could find of the users on personal or community web pages, as well as via image search on the web. The smallest number of photos we collected for an individual was 3, and the largest number was 27. The average number of photos was 15, with a standard deviation of approximately 6 photos. No private information about the subjects was recorded beside storage of the photographs they consented too. Any images of subjects displayed in this paper was done with the consent of that particular volunteer.

For our experiments, we manually extracted the region around user's face in each image. An adversary could also perform this action automatically using tag information on social media sites, when available. One interesting aspect of social media photos is they may capture significant physical changes of users over time. For instance, one of our participants lost 20 pounds in the last 6 months, and our reconstruction had to utilize images from before and after this change. Two other users had frequent changes in facial hair styles – beards, moustaches, and clean-shaven – all of which we used for our reconstruction. Another user had only uploaded 2 photos to social media in the past 3 years. These varieties all present challenges for our framework, both for initially reconstructing the user's face and for creating a likeness that matches their current appearance.

Industry-leading Solutions

We tested our approach on five advanced commercial face authentication systems: KeyLemon[5], Mobius[6], True Key [18], BioID [21], and 1U App[7]. Table 1 summarizes the training data required by each system when learning a user's facial appearance, as well as the approximate number of users for each system, when available. All systems incorporate some degree of liveness detection into their authentication protocol. KeyLemon and the 1U App require users to perform an action such as blinking, smiling, rotating the head, and raising the eyebrows. In addition, the 1U App requests these actions in a random fashion, making it more resilient to video-based attacks. BioID, Mobius and True Key are motion-based systems and detect 3D facial structure as the user turns their head. It is also possible that these five systems employ other advanced liveness detection approaches, such as texture-based detection schemes, but such information has not been made available to the public.

Methodology

System	Training Method	# Installs
KeyLemon[3]	Single video	∼100,000
Mobius[2]	10 still images	18 reviews
True Key[1]	Single video	50,000-100,000
BioID[2]	4 videos	unknown
1U App[1]	1 still image	50-100

Table 1: Summary of the face authentication systems evaluated. The second column lists how each system acquires training data for learning a user's face, and the third column shows the number approximate number of installations or reviews each system has received according to (1) the Google Play Store, (2) the iTunes store, or (3) softpedia.com. BioID is a relatively new app and does not yet have customer reviews on iTunes.

All participants were registered with the 5 face authentication systems under indoor illumination. The average length of time spent by each of the volunteers to register across all systems was 20 minutes. As a control, we first verified that all systems were able to correctly identify the users in the same environment. Next, before testing our method using textures obtained via social media, we evaluated whether our system could spoof the recognition systems using photos taken in this environment. We

[5] http://www.keylemon.com
[6] http://www.biomids.com
[7] http://www.1uapps.com

thus captured one front-view photo for each user under the same indoor illumination and then created their 3D facial model with our proposed approach. We found that these 3D facial models were able to spoof each of the 5 candidate systems with a 100% sucess rate, which is shown in the second column of Table 2

Following this, we reconstructed each user's 3D facial model using the images collected from public online sources. As a reminder, any source image can be used as the main image when texturing the model. Since not all textures will successfully spoof the recognition systems, we created textured reconstructions from all source images and iteratively presented them to the system (in order of what we believed to be the best reconstruction, followed by the second best, and so on) until either authentication succeeded or all reconstructions had been tested.

Findings

We summarize the spoofing success rate for each system in Table 2. Except for the 1U system, all facial recognition systems were successfully spoofed for the majority of participants when using social media photos, and all systems were spoofed using indoor, frontal view photos. Out of our 20 participants, there were only 2 individuals for whom none of the systems was spoofed via the social-media-based attack.

Looking into the social media photos we collected of our participants, we observe a few trends among our results. First, we note that moderate- to high-resolution photos lend substantial realism to the textured models. In particular, photos taken by professional photographers (*e.g.,* wedding photos or family portaits) lead to high-quality facial texturing. Such photos are prime targets for facial reconstruction because they are often posted by other users and made publicly available. Second, we note that group photos provide consistent frontal views of individuals, albeit with lower resolution. In cases where high-resolution photos are not available, such frontal views can be used to accurately recover a user's 3D facial structure. These photos are easily accessible via friends of users, as well. Third, we note that the least spoof-able users were not those who necessarily had a low number of personal photos, but rather users who had few forward-facing photos and/or no photos with sufficiently high resolution. From this observation, it seems that creating a realistic texture for user recognition is the primary factor in determining whether a face authentication method will be fooled by our approach. Only a small number of photos are necessary in order to defeat facial recognition systems.

	Indoor	Social Media	
	Spoof %	Spoof %	Avg. # Tries
KeyLemon	100%	85%	1.6
Mobius	100%	80%	1.5
True Key	100%	70%	1.3
BioID	100%	55%	1.7
1U App	100%	0%	—

Table 2: Success rate for 5 face authentication systems using a model built from (second column) an image of the user taken in an indoor environment and (third and fourth columns) images obtained on users' social media accounts. The fourth column shows the average number of attempts needed before successfully spoofing the target user.

We found that our failure to spoof the 1U App, as well as our lower performance on BioID, using social media photos was directly related to the poor usability of those systems. Specifically, we found the systems have a very high false rejection rate when live users attempt to authenticate themselves in different illumination conditions. To test this, we had 5 participants register their faces indoors on the 4 mobile systems.[8] We then had each user attempt to log in to each system 10 times indoors and 10 times outdoors on a sunny day, and we counted the number of accepted logins in each environment for each system. True Key and Mobius, which we found were easier to defeat, correctly authenticated the users 98% and 100% of the time for indoor logins, respectively, and 96% and 100% of the time for outdoor logins. Meanwhile, the indoor/outdoor login rates of BioID and the 1U App were 50%/14% and 96%/48%, respectively. The high false rejection rates under outdoor illumination show that the two systems have substantial difficulty with their authentication when the user's environment changes. Our impression is that 1U's single-image user registration simply lacks the training data necessary to accommodate to different illumination settings. BioID is very sensitive to a variety of factors including head rotation and illumination, which leads to many false rejections. (Possibly realizing this, the makers of BioID therefore grant the user 3 trials per login attempt.) Even so, as evidenced by the second column in Table 2, our method still handily defeats the liveness detection modules of these systems given images of the user in the original illumination conditions, which suggests that all the systems we tested are vunerable to our VR-based attack.

Our findings also suggest that our approach is able to successfully handle significant changes in facial expression, illumination, and for the most part, physical characteristics such as weight and facial hair. Moreover, the approach appears to generalize to users regardless of gender or ethnicity. Given that it has shown to work on a varied collection of real-world data, we believe that the at-

[8] As it is a desktop application, KeyLemon was excluded.

tack presented herein represents a realistic security threat model that could be exploited in the present day.

Next, to gain a deeper understanding of the realism of this threat, we take a closer look at what conditions are necessary for our method to bypass the various face authentication systems we tested. We also consider what main factors contribute to the failure cases of our method.

4.1 Evaluating System Robustness

To further understand the limitations of the proposed spoofing system, we test its robustness against resolution and viewing angle, which are two important factors for the social media photos users upload. Specifically, we answer the question: what is the minimum resolution and maximum head rotation allowed in an uploaded photo before it becomes unusable for spoofing attacks like ours? We further explore how low-resolution frontal images can be used to improve our success rates when high-resolution side-view images are not available.

4.1.1 Blurry, Grainy Pictures Still Say A Lot

To assess our ability to spoof face authentication systems when provided only low-resolution images of a user's face, we texture the 3D facial models of our sample users using an indoor, frontal view photo. This photo is then downsampled at various resolutions such that the distance between the user's chin and forehead ranges between 20 and 50 pixels. Then, we attempt to spoof the True Key, BioId, and KeyLemon systems with facial models textured using the down-sampled photos.[9] If we are successful at a certain resolution, that implies that that resolution leaks the user's identity information to our spoofing system. The spoofing success rate for various image resolutions is shown in Figure 8.

The result indicates that our approach robustly spoofs face authentication systems when the height of the face in the image is at least 50 pixels. If the resolution of an uploaded photo is less than 30 pixels, the photo is likely of too low-resolution to reliably encode useful features for identifying the user. In our sample set, 88% of users had more than 6 online photos with a chin-to-forehead distance greater than 100 pixels, which easily satisfies the resolution requirement of our proposed spoofing system.

4.1.2 A Little to the Left, a Little to the Right

To identify the robustness of the proposed system against head rotation, we first evaluate the maximum yaw angle allowed for our system to spoof baseline systems using a single image. For all 20 sample users, we collect multiple indoor photos with yaw angle varying from 5 degrees (approximately frontal view) to 40 degrees (significantly rotated view). We then perform 3D reconstruction for each image, for each user, on the same three face authentication systems. The spoofing success rate for a single input image as a function of head rotation is illustrated in Figure 9 (left). It can be seen that the proposed method successfully spoofs all the baseline systems when the input image has a largely frontal view. As yaw angle increases, it becomes more difficult to infer the user's frontal view from the image, leading to a decreased spoofing success rate.

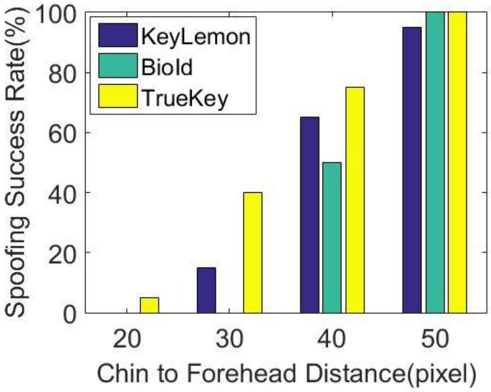

Figure 8: Spoofing success rate with texture taken from photos of different resolution.

4.1.3 For Want of a Selfie

The results of Figure 9 (left) indicate that our success rate falls dramatically if given only a single image with a yaw angle larger than 20 degrees. However, we argue that these high-resolution side-angle views can serve as base images for facial texturing if additional low-resolution frontal views of the user are available. We test this hypothesis by taking, for each user, the rotated images from the previous section along with 1 or 2 low-resolution frontal view photos (chin-to-forehead distance of 30 pixels). We then reconstruct each user's facial model and use it to spoof our baseline systems. Alone, the provided low-resolution images provide insufficient texture for spoofing, and the higher-resolution side view does not provide adequate facial structure. As shown in Figure 9 (right), by *using the low-resolution front views to guide 3D reconstruction and then using the side view for texturing*, the spoofing success rate for large-angle head rotation increases substantially. From a practical standpoint, low-resolution frontal views are relatively easy to obtain, since they can often be found in publicly posted group photos.

[9] We skip analysis of Mobius because its detection method is similar to True Key, and our method did not perform as well on True Key. We also do not investigate the robustness of our method in the 1U system because of our inability to spoof this system using online photos.

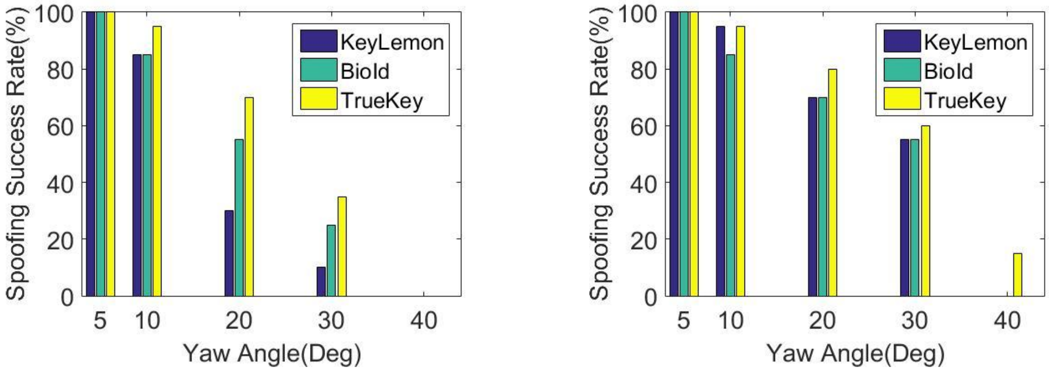

Figure 9: Spoofing success rate with different yaw angles. Left: Using only a single image at the specified angle. Right: Supplementing the single image with low-resolution frontal views, which aid in 3D reconstruction.

4.2 Seeing Your Face *Is* Enough

Our approach not only defeats existing commercial systems having liveness detection — it fundamentally undermines the process of liveness detection based on color images, entirely. To illustrate this, we use our method to attack the recently proposed authentication approach of Li et al. [34], which obtains a high rate of success in guarding against video-based spoofing attacks. This system adds another layer to motion-based liveness detection by requiring that the movement of the face in the captured video be consistent with the data obtained from the motion sensor of the device. Fortunately, as discussed in §3, the data consistency requirement is automatically satisfied with our virtual reality spoofing system because the 3D model rotates in tandem with the camera motion.

Central to Li et al. [34]'s approach is to build a classifier that evaluates the consistency of captured video and motion sensor data. In turn, the learned classifier is used to distinguish real faces from spoofed ones. Since their code and training samples have not been made public, we implemented our own version of Li et al. [34]'s liveness detection system and trained a classifier with our own training data. We refer the reader to [34] for a full overview of the method.

Following the methodology of [34], we capture video samples (and inertial sensor data) of ∼4 seconds from the front-facing camera of a mobile phone. In each sample, the phone is held at a distance of 40cm from the subject and moved back-and-forth 20cm to the left and right. We capture 40 samples of real subjects moving the phone in front of their face, 40 samples where a prerecorded video of a user is presented to the camera, and 30 samples where the camera is presented with a 3D reconstruction of a user in our VR environment. For training, we use a binary logistic regression classifier trained on 20 samples from each class, with the other samples used for testing. Due to the relatively small size of our training sets, we repeat our classification experiments 4 times, with random train/test splits in each trial, and we report the average performance over all four trials.

Training Data	Real	Video	VR
Real+Video	19.50 / 20	0.25 / 20	9.75 / 10
Real+Video+VR	14.00 / 20	0.00 / 20	5.00 / 10
Real+VR	14.75 / 20	—	5.00 / 10

Table 3: Number of testing samples classified as real users. Values in the first column represent true positive rates, and the second and third columns represent false positives. Each row shows the classification results after training on the classes in the first column. The results were averaged over four trials.

The results of our experiments are shown in Table 3. For each class (real user data, video spoof data, and VR data), we report the average number (over 4 trials) of test samples classified as real user data. We experiment with three different training configurations, which are listed in the first column of the table. The first row shows the results when using real user data as positive samples and video spoof data as negative samples. In this case, it can easily be seen that the real-versus-video identification is almost perfect, matching the results of [34]. However, our VR-based attack is able to spoof this training configuration nearly 100% of the time. The second and third rows of Table 3 show the classification performance when VR spoof data is included in the training data. In both cases, our approach defeats the liveness detector in 50% of trials, and the real user data is correctly identified as such less than 75% of the time.

All three training configurations clearly point to the fact that our VR system presents motion features that are close to real user data. Even if the liveness detector of [34] is specifically trained to look for our VR-based attack, 1 out of every 2 attacks will still succeed, with the false rejection rate also increasing. Any system using this detector will need to require multiple log-in attempts to account for the decreased recall rate; allowing multiple log-in attempts, however, allows our method more opportunties to succeed. Overall, the results indicate that the proposed VR-based attack successfully spoofs Li et al. [34]'s approach, which is to our knowledge the state of the art in motion-based liveness detection.

5 Defense in Depth

While current facial authentication systems succumb to our VR-based attack, several features could be added to these systems to confound our approach. Here, we detail three such features, namely random projection of light patterns, detection of minor skin tone fluctuations related to pulse, and the use of illuminated infrared (IR) sensors. Of these, the first two could still be bypassed with additional adversary effort, while the third presents a significantly different hardware configuration that would require non-trivial alterations to our method.

Light Projection The principle of using light projection for liveness detection is simple: Using an outward-facing light source (*e.g.,* the flashlight commonly included on camera-equipped mobile phones), flash light on the user's face at random intervals. If the observed change in illumination does not match the random pattern, then face authentication fails. The simplicity of this approach makes it appealing and easily implementable; however, an adversary could modify our proposed approach to detect the random flashes of light and, with low latency, subsequently add rendered light to the VR scene. Random projections of structured light [62], *i.e.,* checkerboard patterns and lines, would increase the difficulty of such an attack, as the 3D-rendering system must be able to quickly and accurately render the projected illumination patterns on a model. However, structured light projection requires specialized hardware that typically is not found on smart phones and similar devices, which decreases the feasibility of this mitigation.

Pulse Detection Recent computer vision research [2, 58] has explored the prospect of video magnification, which transforms micro-scale fluctuations over time into strong visual changes. One such application is the detection of human pulse from a standard video of a human face. The method detects small, periodic color changes related to pulse in the region of the face and then amplifies this effect such that the face appears to undergo strong changes in brightness and hue. This amplification could be used as an additional method for liveness detection by requiring that the observed face have a detectable pulse. Similar ideas have been applied to fingerprint systems that check for blood flow using light emitted from beneath a prism. Of course, an attacker using our proposed approach could simply add subtle color variation to the 3D model to approximate this effect. Nevertheless, such a method would provide another layer of defense against spoofed facial models.

Infrared Illumination Microsoft released Windows Hello as a more personal way to sign into Windows 10 devices with just a look or a touch. The new interface supports biometric authentication that includes face, iris, or fingerprint authentication. The platform includes Intel's RealSense IR-based, rather than a color-based, facial authentication method. In principle, their approach works in the same way as contemporary face authentication methods, but instead uses an IR camera to capture a video of the user's face. The attack presented in this paper would fail to bypass this approach because typical VR displays are not built to project IR light; however, specialized IR display hardware could potentially be used to overcome this limitation.

One limiting factor that may make IR-based techniques less common (especially on mobile devices) is the requirement for additional hardware to support this enhanced form of face authentication. Indeed, as of this writing, only a handful of personal computers support Windows Hello.[10] Nevertheless, the use of infrared illumination offers intriguing possibilities for the future.

Takeaway In our opinion, it is highly unlikely that robust facial authentication systems will be able to operate using solely web/mobile camera input. Given the widespread nature of high-resolution personal online photos, today's adversaries have a goldmine of information at their disposal for synthetically creating fake face data. Moreover, even if a system is able to robustly detect a certain type of attack – be it using a paper printout, a 3D-printed mask, or our proposed method – generalizing to all possible attacks will increase the possibility of false rejections and therefore limit the overall usability of the system. The strongest facial authentication systems will need to incorporate non-public imagery of the user that cannot be easily printed or reconstructed (*e.g.,* a skin heat map from special IR sensors).

6 Discussion

Our work outlines several important lessons for both the present state and the future state of security, particularly as it relates to face authentication systems. First, our exploitation of social media photos to perform facial reconstruction underscores the notion that online privacy of one's appearance is tantamount to online privacy of other personal information, such as age and location.

[10] See "PC platforms that support Windows Hello" for more info.

The ability of an adversary to recover an individual's facial characteristics through online photos is an immediate and very serious threat, albeit one that clearly cannot be completely neutralized in the age of social media. Therefore, it is prudent that face recognition tools become increasingly robust against such threats in order to remain a viable security option in the future.

At a minimum, it is imperative that face authentication systems be able to reject synthetic faces with low-resolution textures, as we show in our evaluations. Of more concern, however, is the increasing threat of virtual reality, as well as computer vision, as an adversarial tool. It appears to us that the designers of face authentication systems have assumed a rather weak adversarial model wherein attackers may have limited technical skills and be limited to inexpensive materials. This practice is risky, at best. Unfortunately, VR itself is quickly becoming commonplace, cheap, and easy-to-use. Moreover, VR visualizations are increasingly *convincing*, making it easier and easier to create realistic 3D environments that can be used to fool visual security systems. As such, it is our belief that authentication mechanisms of the future must aggressively anticipate and adapt to the rapid developments in the virtual and online realms.

Appendix

A Multi-Image Facial Model Estimation

In §3.2, we outline how to associate 2D facial landmarks with corresponding 3D points on an underlying facial model. Contour landmarks pose a substantial difficulty for this 2D-to-3D correspondence problem because the associated set of 3D points for these features is pose-dependent. Zhu et al. [63] compensate for this phenomenon by modeling contour landmarks with parallel curved line segments and iteratively optimizing head orientation and 2D-to-3D correspondence. For a specific head orientation R_j, the corresponding landmark points on the 3D model are found using an explicit function based on rotation angle:

$$\begin{aligned} s_{i,j} &= f_j P R_j (S_{i',j} + t_j) \\ S_{i',j} &= \bar{S}_{i'} + A_{i'}^{id}\alpha^{id} + A_{i'}^{exp}\alpha_j^{exp} \\ i' &= land(i, R_j), \end{aligned} \qquad (6)$$

where $land(i,R_j)$ is the pre-calculated mapping function that computes the position of landmarks i on the 3D model when the orientation is R_j. Ideally, the first equation in Eq. (6) should hold for all the landmark points in all the images. However, this is not the case due to the alignment error introduced by landmark extraction. Generally, contour landmarks introduce more error than corner landmarks, and this approach actually leads to inferior results when multiple input images are used.

Therefore, different from Zhu et al. [63], we compute the 3D facial model with Maximum a Posteriori (MAP) estimation. We assume the alignment error of each 3D landmark independently follows a Gaussian distribution. Then, the most probable parameters $\theta := (\{f_j\}, \{R_j\}, \{t_j\}, \{\alpha_j^{exp}\}, \alpha^{id})$ can be estimated by minimizing the cost function

$$\theta = \underset{\theta}{\operatorname{argmax}}\{\sum_{i=1}^{68}\sum_{j=1}^{N}\frac{1}{(\sigma_i^s)^2}||s_{i,j} - f_j P R_j (S_{i',j} + t_j)||^2 + \sum_{j=1}^{N}(\alpha_j^{exp})'\Sigma_{exp}^{-1}\alpha_j^{exp} + (\alpha^{id})'\Sigma_{id}^{-1}\alpha^{id}\}. \qquad (7)$$

Here, $S_{i',j}$ is computed using Eq. (6). Σ_{id} and Σ_{exp} are covariance matrices of α^{id} and α_j^{exp}, which can be obtained from the pre-existing face model. $(\sigma_i^s)^2$ is the variance of alignment error of the i-th landmark and is obtained from a separate training set consisting 20 images with hand-labeled landmarks. Eq. (7) can be computed efficiently, leading to the estimated identity weight α^{id}, with which we can compute the neutral-expression model $S_i (= \bar{S}_{i'} + A_{i'}^{id}\alpha^{id})$.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision (IJCV)*, 56 (3):221–255, 2004.

[2] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, 2013.

[3] W. Bao, H. Li, N. Li, and W. Jiang. A liveness detection method for face recognition based on optical flow field. In *Image Analysis and Signal Processing, International Conference on*, pages 233–236, 2009.

[4] C. Baumberger, M. Reyes, M. Constantinescu, R. Olariu, E. De Aguiar, and T. Oliveira Santos. 3d face reconstruction from video using 3d morphable model and silhouette. In *Graphics, Patterns and Images (SIBGRAPI), Conference on*, pages 1–8, 2014.

[5] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2930–2940, 2013.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[7] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003.

[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *Visualization and Computer Graphics, IEEE Transactions on*, 20 (3):413–425, 2014.

[9] B. Chu, S. Romdhani, and L. Chen. 3d-aided face recognition robust to expression and pose variations. In *Computer Vision and Pattern Recognition (CVPR), Conference on*, pages 1907–1914, 2014.

[10] N. Duc and B. Minh. Your face is not your password. In *Black Hat Conference*, volume 1, 2009.

[11] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *Information Forensics and Security, IEEE Transactions on*, 9(7):1084–1097, 2014.

[12] D. Fidaleo and G. Medioni. Model-assisted 3d face reconstruction from video. In *Analysis and modeling of faces and gestures*, pages 124–138. Springer, 2007.

[13] Gartner. Gartner backs biometrics for enterprise mobile authentication. *Biometric Technology Today*, Feb. 2014.

[14] S. Golder. Measuring social networks with digital photograph collections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 43–48, 2008.

[15] M. Hicks. A continued commitment to security, 2011. URL https://www.facebook.com/notes/facebook/a-continued-commitment-to-security/486790652130/.

[16] R. Horaud, F. Dornaika, and B. Lamiroy. Object pose: The link between weak perspective, paraperspective, and full perspective. *International Journal of Computer Vision*, 22(2):173–189, 1997.

[17] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security*, pages 781–792, 2015.

[18] Intel Security. True Key^{TM} by Intel Security: Security white paper 1.0, 2015. URL https://b.tkassets.com/shared/TrueKey-SecurityWhitePaper-v1.0-EN.pdf.

[19] H.-K. Jee, S.-U. Jung, and J.-H. Yoo. Liveness detection for embedded face recognition system. *International Journal of Biological and Medical Sciences*, 1(4):235–238, 2006.

[20] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[21] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[22] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[23] I. Kemelmacher-Shlizerman. Internet based morphable model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3256–3263, 2013.

[24] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2): 394–405, 2011.

[25] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park, and J. Kim. Face liveness detection based on texture and frequency analyses. In *Biometrics (ICB), 5th IAPR International Conference on*, pages 67–72, 2012.

[26] H.-N. Kim, A. El Saddik, and J.-G. Jung. Leveraging personal photos to inferring friendships in social network services. *Expert Systems with Applications*, 39(8):6955–6966, 2012.

[27] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee. Face liveness detection using variable focusing. In *Biometrics (ICB), 2013 International Conference on*, pages 1–6, 2013.

[28] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys. Turning mobile phones into 3d scanners. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 3946–3953, 2014.

[29] K. Kollreider, H. Fronthaler, and J. Bigun. Evaluating liveness by face images and the structure tensor. In *Automatic Identification Advanced Technologies, Fourth IEEE Workshop on*, pages 75–80. IEEE, 2005.

[30] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in liveness assessment. *Information Forensics and Security, IEEE Transactions on*, 2(3):548–558, 2007.

[31] K. Kollreider, H. Fronthaler, and J. Bigun. Verifying liveness by multiple experts in face biometrics. In *Computer Vision and Pattern Recognition Workshops, IEEE Computer Society Conference on*, pages 1–6, 2008.

[32] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes, and S. Sridharan. Liveness detection based on 3d face shape analysis. In *Biometrics and Forensics (IWBF), International Workshop on*, pages 1–4, 2013.

[33] Y. Li, K. Xu, Q. Yan, Y. Li, and R. H. Deng. Understanding osn-based facial disclosure against face authentication systems. In *Proceedings of the ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, pages 413–424. ACM, 2014.

[34] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng. Seeing your face is not enough: An inertial sensor-based liveness detection for face authentication. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security*, pages 1558–1569, 2015.

[35] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70. ACM, 2011.

[36] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. *arXiv preprint arXiv:1404.3840*, 2014.

[37] J. Määttä, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using micro-texture analysis. In *Biometrics (IJCB), International Joint Conference on*, pages 1–7, 2011.

[38] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[39] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009.

[40] B. Peixoto, C. Michelassi, and A. Rocha. Face liveness detection under bad illumination conditions. In *Image Processing (ICIP), 18th IEEE International Conference on*, pages 3557–3560, 2011.

[41] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics (TOG)*, 22(3):313–318, 2003.

[42] I. Polakis, M. Lancini, G. Kontaxis, F. Maggi, S. Ioannidis, A. D. Keromytis, and S. Zanero. All your face are belong to us: Breaking facebook's social authentication. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 399–408, 2012.

[43] C. Qu, E. Monari, T. Schuchert, and J. Beyerer. Fast, robust and automatic 3d face model reconstruction from videos. In *Advanced Video and Signal Based Surveillance (AVSS), 11th IEEE International Conference on*, pages 113–118, 2014.

[44] C. Qu, E. Monari, T. Schuchert, and J. Beyerer. Adaptive contour fitting for pose-invariant 3d face shape reconstruction. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12, 2015.

[45] C. Qu, E. Monari, T. Schuchert, and J. Beyerer. Realistic texture extraction for 3d face models robust to self-occlusion. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015.

[46] T. Schops, T. Sattler, C. Hane, and M. Pollefeys. 3d modeling on the go: Interactive 3d reconstruction of large-scale scenes on mobile devices. In *3D Vision (3DV), International Conference on*, pages 291–299, 2015.

[47] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015.

[48] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):222, 2014.

[49] L. Sun, G. Pan, Z. Wu, and S. Lao. Blinking-based live face detection using conditional random fields. In *Advances in Biometrics*, pages 252–260. Springer, 2007.

[50] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 3476–3483, 2013.

[51] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *Computer Vision–ECCV 2014*, pages 796–812. Springer, 2014.

[52] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3952–3960, 2015.

[53] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1701–1708, 2014.

[54] X. Tan, Y. Li, J. Liu, and L. Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision (ECCV)*, pages 504–517. 2010.

[55] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 65–72, 2013.

[56] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. Global localization from monocular slam on a mobile phone. *Visualization and Computer Graphics, IEEE Transactions on*, 20(4):531–539, 2014.

[57] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection using 3d structure recovered from a single camera. In *Biometrics (ICB), International Conference on*, pages 1–6, 2013.

[58] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31 (4), 2012.

[59] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 532–539, 2013.

[60] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *Biometrics (ICB), International Conference on*, pages 1–6, 2013.

[61] L. Zhang and D. Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):351–363, 2006.

[62] L. Zhang, B. Curless, and S. M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *3D Data Processing Visualization and Transmission, First International Symposium on*, pages 24–36, 2002.

[63] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.