HHS Public Access

Author manuscript

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2021 July 01.

Published in final edited form as:

IEEE Trans Pattern Anal Mach Intell. 2020 July ; 42(7): 1582–1593. doi:10.1109/TPAMI.2020.2986951.

Differential 3D Facial Recognition: Adding 3D to Your State-of-the-Art 2D Method

J. Matías Di Martino,
Department of Physics, School of Engineering, Universidad de la República, Montevideo 11200, Uruguay, and also with the Department of Electrical Engineering, Duke University, Durham, NC 27708.

Fernando Suzacq,
Department of Electrical Engineering, Universidad de la República, Montevideo 11200, Uruguay.

Mauricio Delbracio,
Department of Electrical Engineering, Universidad de la República, Montevideo 11200, Uruguay.

Qiang Qiu,
Department of Electrical Engineering, Duke University, Durham, NC 27708.

Guillermo Sapiro [Fellow, IEEE]
Department of Electrical Engineering, Duke University, Durham, NC 27708.

Abstract

Active illumination is a prominent complement to enhance 2D face recognition and make it more robust, e.g., to spoofing attacks and low-light conditions. In the present work we show that it is possible to adopt active illumination to enhance state-of-the-art 2D face recognition approaches with 3D features, while bypassing the complicated task of 3D reconstruction. The key idea is to project over the test face a high spatial frequency pattern, which allows us to simultaneously recover real 3D information plus a standard 2D facial image. Therefore, state-of-the-art 2D face recognition solution can be transparently applied, while from the high frequency component of the input image, complementary 3D facial features are extracted. Experimental results on ND-2006 dataset show that the proposed ideas can significantly boost face recognition performance and dramatically improve the robustness to spoofing attacks.

Keywords

Differential 3D; active stereo; face recognition; spoofing detection; 3D facial analysis

1 Introduction

TWO-DIMENSIONAL face recognition has become extremely popular as it can be ubiquitously deployed and large datasets are available. In the past several years, tremendous progress has been achieved in making 2D approaches more robust and useful in real-world

Corresponding authors: J. Matias Di Martino, matias.di.martino.uy@gmail.com.

applications. Though 2D face recognition has surpassed human performance in certain conditions, challenges remain to make it robust to facial poses, uncontrolled ambient illumination, aging, low-light conditions, and spoofing attacks [1], [2], [3], [4]. In the present work we address some of these issues by enhancing the captured RGB facial image with 3D information as illustrated in Fig. 1.

High resolution cameras became ubiquitous, although for 2D face recognition, we only need a facial image of moderate or low resolution. For example latest phones frontal camera have a very high resolution (e.g., 3088 × 2320 pixels) while the resolution of the input to most face recognition systems is limited to 224 × 224 pixels [4], [5], [6], [7], [8]. This means that, in the context of face recognition, we are drastically underutilizing most of the resolution of captured images. We propose an alternative to use the discarded portion of the spectra and extract real 3D information by projecting a high frequency light pattern. Hence, a low resolution version of the RGB image remains approximately invariant allowing the use of standard 2D approaches, while 3D information is extracted efficiently from the local deformation of the projected patterns.

The proposed solution to extract 3D facial features has key differences with the two common approaches presented in existing literature: 3D hallucination [9], [10], [11], [12] and 3D reconstruction [13], [14]. We will discuss these differences in detail in the following section. We illustrate the main limitation of 3D hallucination in the context of face recognition in Fig. 2, which emphasizes the lack of real 3D information on a standard RGB input image. We demonstrate that it is possible to extract actual 3D facial features bypassing the ill-posed problem of explicit depth estimation. Our contributions are summarized as follows:

- Analyzing the spectral content of thousands of facial images, we design a high frequency light pattern that simultaneously allow us to retrieve a standard 2D low resolution facial image plus a 3D gradient facial representation.
- We propose an effective and modular solution that achieves 2D and 3D information decomposition and facial feature extraction in a data-driven fashion (bypassing a 3D facial reconstruction).
- We show that by defining an adequate distance function in the space of the feature embedding, we can leverage the advantages of both 2D and 3D features. We can transparently exploit existing state-of-the-art 2D methods and improve their robustness, e.g., to spoofing attacks.

2 Related Work

To recognize or validate the identity of a subject from a 2D color photograph is a longstanding problem of computer vision and has been largely studied for over forty years [15], [16]. Recent advances in machine learning, and in particular, the success of deep neural networks, reshaped the field and yielded more efficient, accurate, and reliable 2D methods such as: ArcFace [5], VGG-Face [6], DeepFace [4], and FaceNet [7].

In spite of this, spoofing attacks and variations in pose, expression and illumination are still active challenges and significant efforts are being made to address them [14], [17], [18],

[19], [20], [21], [22], [23], [24], [25]. For example, Deng *et al.* [26] attempt to handle large pose discrepancy between samples. To that end, they propose an adversarial facial UV map completion GAN. Complementing previous approaches that seek for robust feature representations, several works propose more robust loss and metric functions [27], [28].

3D hallucination From Single RGB.

To enhance 2D approaches a common trend is to hallucinate a 3D representation from an input RGB image which is used to extract 3D features [9], [10], [11], [12], [29], [30]. For example, Cui *et al.* [31] introduce a cascade of networks that simultaneously recover depth from an RGB input while seeking for separability of individual subjects. The estimated depth information is then used as a complementary modality to RGB.*3D Face Recognition.*The approaches described previously share an important practical advantage that at the same time is their weakness, they extract all the information from a standard (RGB) 2D photograph of the face. As depicted in Fig. 2a single image does not contain actual 3D information. To overcome this intrinsic limitation different ideas have been proposed and datasets with 3D facial information are becoming more popular [8]. For example, Zafeiriou *et al.* [13] propose a four-light source photometric stereo (PS). A similar idea is elaborated by Zou *et al.* [14] who propose to use active near-infrared illumination and combine a pair of input images to extract an illumination invariant face representation.

Despite the previous mentioned techniques, performing a 3D facial reconstruction is still a challenging and complicated task. Many strategies have been proposed to tackle this problem, including time delay based [33], image cue based [9], [34], [35], [36], [37], and triangulation based methods [38], [39], [40], [41], [42]. Although there has been great recent development, available technology for 3D scanning is still too complicated to be ubiquitously deployed [32], [43], [44], [45].

The proposed solution has two key features that make it, to the best of our knowledge, different from existing alternatives. (a) Because the projected pattern is of a high spatial frequency, we can recover a standard (low resolution) RGB facial image that can be fed into state-of-the-art 2D face recognition methods. (b) We avoid the complicated task of 3D facial reconstruction and instead, extract local 3D features from the local deformation of the projected pattern. In that sense our ideas can be implemented exploiting existing and future 2D solutions. In addition, our approach is different from those that hallucinate 3D information. As discussed before and illustrated in Fig. 2 this task requires a strong prior of the scene which is ineffective, for example, if a spoofing attack is presented (see the example provided in Fig. 15 in the supplementary material), available online.

3 Proposed Approach

Notation.—Let $\mathcal{I} \subset \mathbb{R}^{H \times W \times C}$ denote the space of images with $H \times W$ pixels and C color channels, and $\mathcal{X}_n \subset \mathbb{R}^n$ a space of n-dimensional column vectors (in the context of this work associated to a facial feature embedding). $\mathcal{I}_{rgb}$ denotes the set of RGB images ($C = 3$), while $\mathcal{I}_{\nabla z}$ is used to denote the space of two channel images ($C = 2$) associated to the gradient of a single-channel image $z \in \mathbb{R}^{H \times W \times 1}$. (The first/second channel represents the

partial derivative with respect to the first/second coordinate.)*Combining Depth and RGB Information.* The proposed approach consists of three main modules as illustrated in Fig. 3: $g: \mathcal{I}_{rgb} \rightarrow \mathcal{I}_{rgb} \times \mathcal{I}_{\nabla z}$ performs a decomposition of the input image into texture and depth information, $f_{rgb}: \mathcal{I}_{rgb} \rightarrow \mathcal{X}_{n/2}$, and $f_{\nabla z}: \mathcal{I}_{\nabla z} \rightarrow \mathcal{X}_{n/2}$ extract facial features associated to the facial texture and depth respectively. These three components are illustrated in Fig. 3 in blue, yellow, and green, respectively. (We decided to have three modules instead of a single *end-to-end* design for several reasons that will be discussed below.)

Algorithm 1.

Compute 2D Facial Features Enhanced With 3D Information

procedure FacialEmbedding(I)

Decompose the input image into texture and depth gradient information.

$\{I_{rgb}, I_{\nabla z}\} = g(I)$

Extract facial information from each component.

$x_{rgb} = f_{rgb}(I_{rgb})$

$x_{\nabla z} = f_{\nabla z}(I_{\nabla z})$

Combine texture and depth information.

x = Concatenate $(x_{rgb}, x_{\nabla z})$

return x ▷ Facial embedding

end procedure

We denote the facial feature extraction from the input image as $f_\theta: \mathcal{I}_{rgb} \rightarrow \mathcal{X}_n$, where $f_\theta(I) = \left(f_{rgb}(I_{rgb}), f_{\nabla z}(I_{\nabla z})\right)^T$ with $\{I_{rgb}, I_{\nabla z}\} = g(I)$. The subscript θ represent the parameters of the mapping f, which can be decomposed in three groups $\theta = (\theta_g, \theta_{rgb}, \theta_{\nabla z})$, associated to the image decomposition, RGB feature extraction, and depth feature extraction respectively. In the following we discuss how these parameters are optimized for each specific task, which is one of the advantages of formulating the problem in a modular fashion.

Once texture and depth facial information is extracted into a suitable vector representation $x = f_\theta(I)$ (as illustrated in Algorithm 1), we can select a distance measure $d: \mathcal{X}_n \times \mathcal{X}_n \rightarrow \mathbb{R}^+$ to compare facial samples and estimate whether they have a high likelihood of belonging to the same subject or not. It is worth noticing that faces are embedded into a space in which the first half of the dimensions are associated to information extracted from the RGB representation while the other half codes depth information. These two sources of information may have associated different confidence levels (depending on the conditions at deployment). We address this in detail in Section 3.3 and propose an anisotropic distance adapted to our solution, and capable of leveraging the good performance of 2D solutions in certain conditions, while improving robustness and handling spoofing attacks in a continuous and unified fashion.

3.1 Pattern Design

When a pattern of light $p(x, y)$ is projected over a surface with a height map $z(x, y)$, it is perceived by a camera located along the x-axis with a deformation given by $p(x + \phi(x, y), y)$ ($\phi(x, y) \propto z(x, y)$). A detailed description of active stereo geometry is provided in the supplementary material Section B, available online. Let us denote $I_0(x, y)$ the image we would acquire under homogeneous illumination, and $p(x, y)$ the intensity profile of the projected light. Without loss of generality we assume the system baseline is parallel to the x axis. The image acquired by the camera when the projected light is modulated with a profile $p(x, y)$ is

$$I(x, y) = I_0(x, y)p(x + \phi(x, y), y)\,. \tag{1}$$

We will restrict to periodic modulation patterns and let T denote the pattern spatial period, we also define $f_0 \overset{def}{=} \frac{1}{T}$. To simplify the system design and analysis, lets also restrict to periodic patterns that are invariant to the y coordinate. In these conditions we can express $p(x, y) = \sum_{n=-\infty}^{+\infty} a_n e^{i2\pi n f_0 x}$ where a_n represent the coefficients of the Fourier series of p. (Note that because of the invariance with respect to the y coordinate, the coefficients a_n are constant instead of a function of y.) Equation (1) can be expressed as

$$I(x, y) = \sum_{n=-\infty}^{+\infty} I_0(x, y)a_n e^{i2\pi n f_0(x + \phi(x, y))}\,. \tag{2}$$

Defining $q_n(x, y) \overset{def}{=} I_0(x, y)a_n e^{i2\pi n f_0 \phi(x, y)}$, Equation (2) can be expressed as [46]

$$I(x, y) = \sum_{n=-\infty}^{+\infty} q_n(x, y)e^{i2\pi n f_0 x}\,. \tag{3}$$

Applying the 2D Fourier Transform (FT) in both sides of Equation (3) and using standard properties of the FT [47] we obtain

$$\tilde{I}(f_x, f_y) = \sum_{n=-\infty}^{+\infty} \tilde{q}_n(f_x - nf_0, f_y)\,. \tag{4}$$

We denote as $\tilde{I}$ the FT of I and use (f_x, f_y) to represent the 2D frequency domain associated to x and y axis respectively.

Equation (4) shows that the FT of the acquired image can be decomposed into the components $\tilde{q}_n$ centered at $(nf_0, 0)$. In the context of this section, we refer to a function $h(x, y)$ being smooth if

$$\frac{\| \tilde{h}(f_x, f_y) \|}{\| \tilde{h}(0,0) \|} < 10^{-3} \quad \forall \quad |f_x| > \frac{f_0}{2}. \tag{5}$$

Assuming $I_0(x, y)$ and $\phi(x, y)$ are smooth (we empirically validate this hypothesis below), the components $\tilde{q}_n$ can be isolated as illustrated in Fig. 4. The central component is of particular interest, $q_0(x, y) = a_0\, I_0(x, y)$ captures the facial texture information and can be recovered from $I(x, y)$ if f_0 is large enough (we provide a more precise quantitative analysis in what follows). On the other hand, relative (gradient) 3D information can be retrieved from the components $\{q_0, q_1\}$ as we show in Proposition 1.

Proposition 1: *Gradient depth information is encoded in the components* $\{q_0(x, y), q_1(x, y)\}$.

Proof: We define the wrapping function $\mathscr{W}(u) = \text{atan}(\tan(u))$. This function wraps the real set into the interval $(-\pi/2, \pi/2]$ [48]. This definition can be extended to vector inputs wrapping the modulus of the vector field while keeping its direction unchanged, i.e., $\mathscr{W}(\vec{u}) = \frac{\mathscr{W}(\| \vec{u} \|)}{\| \vec{u} \|}\vec{u}$ if $\| \vec{u} \| \neq 0$ and $\mathscr{W}(\vec{u}) = \vec{0}$ if $\| \vec{u} \| = 0$. From $q_1(x, y)$ and $q_0(x, y)$ we can compute[1]

$$\phi_{\mathscr{W}}(x, y) = \frac{1}{2\pi f_0}\text{atan}\left(\frac{\text{Im}\left\{\frac{q_1(x,y)}{q_0(x,y)}\right\}}{\text{Re}\left\{\frac{q_1(x,y)}{q_0(x,y)}\right\}}\right), \tag{6}$$

where $\phi_{\mathscr{W}}$ denotes the wrapped version of ϕ. Moreover, $\phi_{\mathscr{W}}(x, y) = \phi(x, y) + \pi k(x, y)$ with $k(x, y) \in \mathbb{N}$ (wrapping introduces shifts of magnitude multiple of π). Computing the gradient both sides leads to $\nabla\phi_{\mathscr{W}}(x, y) = \nabla\phi(x, y) + \pi\nabla k(x, y)$ where $\| \nabla k(x, y) \| \in \mathbb{N}$. Assuming the magnitude of the gradient of $\phi(x, y)$ is bounded by $\pi/2$ and considering that $\| \nabla k(x, y) \| \in \mathbb{N}$, we can apply the wrapping function both sides of the previous equality to obtain $\mathscr{W}(\nabla\phi_{\mathscr{W}})(x, y) = \nabla\phi(x, y)$ which proves (recall Equation (6)) that the gradient of ϕ can be extracted from the components q_0 and q_1. To conclude the proof, we use the property of linearity of the gradient operation and the fact that $\phi(x, y)$ is proportional to the depth map of the scene (see Equation (12) and Section B in the supplementary material), available online. □

Analytic versus Data-Driven Texture and Gradient Depth Extraction.—The previous analysis shows that closed forms can be obtained to extract texture and depth gradient information. However, to compute these expressions is necessary to isolate different spectral components $\tilde{q}_n$. To that end, filters need to be carefully designed. The design of these filters is challenging, e.g., one need to control over-smoothing versus introducing

[1]We assume images are extended in an even fashion outside the image domain, to guaranteed that $a_1 \in \mathbb{R}$ and avoid an additional offset term.

ringing artifact which are drastically amplified by a posterior gradient computation [39], [42]. To overcome these challenges, we chose to perform a depth (gradient) and texture decomposition in a data-driven fashion, which as we showin Section 4, provides an efficient and effective solution.

Bounds on f_0 and Optimal Spectral Orientation.—As discussed above, the projected pattern $p(x, y)$ should have a large fundamental frequency f_0. In addition, the orientation of the fringes and the system baseline can be optimized if faces present a narrower spectral content in a particular direction. We study the texture and depth spectrum of the facial images of ND-2006 dataset (this dataset provides ground truth facial texture and depth information). We observed (see Fig. 5) that for facial images sampled at a 480 × 480 spatial resolution, most of the energy is concentrated in a third of the discrete spectral domain (observe the extracted one dimensional profiles of the spectrum shown at the left side of Fig. 5). In addition, we observe that the spectral content of facial images is approximately isotropic. See, for example, Fig. 5 and observe how for 1-dimensional sections across different orientations the 2D spectra envelope is almost constant. We conclude that the orientation of the fringes does not play a significant role in the context of facial analysis. In addition, we conclude that the fringes width should be smaller than 7mm (distance measure over the face).[2]

3.2 Network Training and the Advantages of Modularity

As described previously, the parameters of the proposed solution can be split in three groups $\theta = (\theta_g, \theta_{rgb}, \theta_{\nabla z})$. This is an important practical property and we designed the proposed solution to meet this condition (in contrast to an end-to-end approach).

Let us define $\mathscr{B}_1, \mathscr{B}_2,$ and $\mathscr{B}_3$ three datasets containing ground truth depth information, ground truth identity for rgb facial images, and ground truth identity for depth facial images, respectively. More precisely,

$\mathscr{B}_1 = \{(I_i(x, y), I_{0i}(x, y), z_i(x, y)), i = 1, \ldots, n_1\}$, $\mathscr{B}_2 = \{(I_{0i}(x, y), y_i), i = 1, \ldots, n_2\}$, and $\mathscr{B}_3$, $= \{(z_i(x, y), y_i), i = 1, \ldots, n_3\}$

where $I_i(x, y)$ denotes a (facial or generic) RGB image acquired under the projection of the designed pattern, $I_{0i}(x, y)$ represents (facial or generic) standard RGB images, $z_i(x, y)$ denotes a gray image representing the depth of the scene, and y_i a scalar integer representing the subject id.

We denote as $\{g_1(I), g_2(I)\} = g(I)$ the RGB and gradient depth components estimated by the decomposition operation g. We partitioned the parameters of g into two sets of dedicated kernels $\theta_g = \{\theta_{g_1}, \theta_{g_2}\}$, the first group focuses on retrieving the texture component while the second group retrieves the depth gradient. These parameters can be optimized as

[2]This numerical results is obtained by approximating the bounding box of the face as a $20cm \times 20cm$ region, sampled with 480 × 480 pixels which corresponds to a pixel length of $2.4mm$, a third of the spectral band correspond to signal of a period of 6 pixels which leads to a binary fringe of at least $7.2mm$ wide.

$$\theta_{g1} = argmin \sum_{(I_{0i},\, I_i) \in \mathscr{B}_1} \|g_1(I_i) - I_{0i}\|_2^2 \tag{7}$$

$$\theta_{g2} = argmin \sum_{(z_i,\, I_i) \in \mathscr{B}_1} \|g_2(I_i) - \nabla z_i\|_2^2 . \tag{8}$$

(We also evaluated training a shared set of kernels trained with an unified loss, this alternative is harder to train in practice, due to the natural difference between the dynamic range and sparsity of gradient images compared with texture images.)

For texture and depth facial feature extraction, we tested models inspired in the Xception architecture [49]). Additional details are provided in the supplementary material Section D, available online. To train these models we add an auxiliary fully connected layer on top of the facial embedding (with as many neurons as identities in the train set) and minimize the cross-entropy between the ground truth and the predicted labels. More precisely, let us denote $\hat{f}_{rgb}(I_{rgb}) = [p_1, \ldots, p_c]$ the output of the fully connected layer associated to the embedding $f_{rgb}(I_{rgb})$ where p_i denotes the probability associated to the id i,

$$\theta_{rgb} = argmin \sum_{(I_{0i},\, y_i) \in \mathscr{B}_2} \sum_{c} -\mathbf{1}_{y_i = c} \log\left(\hat{f}_{rgb}(I_{0i})[c]\right) \tag{9}$$

$$\theta_{\nabla z} = argmin \sum_{(z_i,\, y_i) \in \mathscr{B}_3} \sum_{c} -\mathbf{1}_{y_i = c} \log\left(\hat{f}_{\nabla z}(\nabla z_i)[c]\right), \tag{10}$$

where $\mathbf{1}_{y_i = c}$ denotes the indicator function. (Of course one can choose other alternative losses to train these modules, see e.g., [5], [27], [28], [50].)

As described above, the proposed design allows to leverage information from three types of datasets ($\mathscr{B}_1, \mathscr{B}_2, \mathscr{B}_3$). This has an important practical advantage as 2D facial and 3D generic datasets are more abundant, and the pattern dependant set $\mathscr{B}_1$ can be of modest size as $\#(\theta_g) \ll \#(\theta_{rgb})$.

3.3 Distance Design

Once different modules are set we can compute the facial embedding of test images following the procedure described in Algorithm 1. Let us define $x^a \in \mathscr{X}_n$ and $x^b \in \mathscr{X}_n$ the feature embedding of two facial images I_a and I_b respectively. Recall that the first $n/2$ elements of x are associated to features extracted from (a recovered) RGB facial image while the remaining elements are associated to depth information, i.e.,
$x = \left(x_{rgb}[1], \ldots, x_{rgb}[n/2], x_{\nabla z}[1], \ldots, x_{\nabla z}[n/2]\right)^T$.

We define the distance between two feature representations
$x^a = \left(x^a_{rgb}, x^a_{\nabla z}\right)$, and $x^b = \left(x^b_{rgb}, x^b_{\nabla z}\right)$ as

$$d_{\alpha,\beta,\gamma}\left(x^a, x^b\right) \stackrel{def}{=} (1-\gamma) d_c\left(x^a_{rgb}, x^b_{rgb}\right) + \gamma d_c\left(x^a_{\nabla z}, x^b_{\nabla z}\right)\left(1 + \left(\frac{d_c\left(x^a_{\nabla z},\ x^b_{\nabla z}\right)}{\beta}\right)^{\alpha}\right). \tag{11}$$

$d_c: \mathcal{X}_{n/2} \times \mathcal{X}_{n/2} \rightarrow [0,1]$ denotes the cosine distance, $\gamma \in [0,1]$ sets the relative weight of RGB and depth features, and $\alpha, \beta \in \mathbb{R}$ define a non-linear response for the distance between depth features. As we will describe in the following, this provides robustness against common cases of spoofing attacks.

Intuitively, γ allows us to set the relative confidence associated to RGB and depth features, for example, $\gamma = 1/2$ gives the same weight to RGB and depth features, while $\gamma = 0$ ($\gamma = 1$) ignores the distance between samples in the depth (RGB) embedding space. This is important in practice, as is common to obtain substantially more data to train RGB models than depth ones ($|\mathscr{B}_2| \gg |\mathscr{B}_3|$). This suggests that in good test conditions (e.g., good lighting) one may trust more RGB features over depth features ($\gamma < 1/2$). As we will empirically show in the following section, when two facial candidates are compared, $d_{\alpha,\infty,\gamma}\left(x^a, x^b\right) = (1-\gamma)\delta\left(x^a_{rgb}, x^b_{rgb}\right) + \gamma\delta\left(x^a_{\nabla z}, x^b_{\nabla z}\right)$ is an effective distance choice. However, it does not handle robustly common cases of spoofing attacks. The most common deployments of spoofing attacks imitate the facial texture more accurately than the facial depth [51], [52], [53], therefore, the global distance between two samples should be large when the distance of the depth features is large (i.e., above a certain threshold). To that end, we introduce an additional non-linear term controlled by parameters β and α, for $\delta\left(x^a_{\nabla z}, x^b_{\nabla z}\right) < \beta$ the standard cosine distance dominates while for large values the distance will be amplified in a non-linear fashion.

4 Experiments and Discussion

Data.

Three public dataset are used for experimental validation: FaceScrub [54], CASIA Anti-spoofing [55], and ND-2006 [56]. FaceScrub contains $100k$ RGB (2D) facial images of 530 different subjects, and is used to train the texture-based facial embedding. CASIA dataset contains 150 genuine videos (recording a person) and 450 videos of different types of spoofing attacks, the data was collected for 50 subjects. We use this dataset to simulate and imitate the texture properties of images of spoofing attacks. ND-2006 is one of the larges publicly available datasets with 2D and 3D facial information, it contains $13k$ images of 888 subjects. We used this set to demonstrate that differential 3D features can be extracted from a single RGB input, to compare RGB features with 3D features extracted from the differential 3D input, and to show that when 2D and 3D information is properly combined, the best properties of each can be obtained.

Texture and Differential 3D Decomposition.

In Section 3.1 we discussed how real 3D information and texture information can be coded and later extracted using a single RGB image. In addition, we argue that this decomposition

can be learned efficiently and effectively in a data-driven fashion. To that end, we tested simple network architectures composed of standard convolutional layers (a full description of these architectures and the training protocols are provided as supplementary material), available online. Using ground truth texture and depth facial information, we simulated the projection of the designed pattern over the 888 subjects provided in ND-2006 dataset. Illustrative results are presented in Fig. 6 and in the supplementary material, available online. The 3D geometrical model and a detailed description of the simulation process is provided in Section D.1. Though the simulation of the deformation of a projected pattern can be computed in a relatively simple manner (if the depth information is known), the inverse problem is analytically hard [39], [41], [45].

Despite the previous, we observed that a stack of convolutional layers can efficiently learn how to infer from the image with the projected pattern, both depth gradient information, and the standard (2D) facial image. Fig. 7 illustrates some results for subjects in the test set. The first column corresponds to the input to the network, the second column the ground truth texture information, and the third column the retrieved texture information. The architecture of the network and the training protocol is described in detail in the supplementary material Section D, available online. As we can see in the examples illustrated in Fig. 7, an accurate low resolution texture representation of the face can be achieved in general, and visible artifact are observed only in the regions where the depth is discontinuous (see for example, the regions illustrated at the bottom of Fig. 7).

Fig. 8 illustrates the ground truth and the retrieved depth gradient (again, for random samples from the test set). To estimate the 3D information, we feed to a different branch of convolutional layers the gray version of the input image. These layers are fully described in the supplementary material Table 5, available online. A gray input image is considered instead of a color one because the projected pattern is achromatic, and therefore, no 3D information is encoded in the colors of the image. In addition, we crop the input image to exclude the edges of the face. (Facial registration and cropping is performed automatically using dlib [57] facial landmarks.) As discussed in Section 3, and in particular, in the proof of Proposition 1, the deformation of the projected fringes only provide local gradient information if the norm of the gradient of the depth is bounded. In other words, where the scene present depth discontinuities, no local depth information can be extracted by our proposed approach. This is one of the main reasons why differential 3D information can be exploited for face recognition, while bypassing the more complicated task of a 3D facial reconstruction.

One of the advantages of the proposed approach is that it extracts local depth information, and therefore, the existence of depth discontinuities does not affect the estimation on the smooth portion of the face. This is illustrated in Figs. 9a and 9b, where a larger facial patch is fed into the network. The decomposition module is composed exclusively of convolutional layers, and therefore, images of arbitrary size can be evaluated. Fig. 9a shows the input to the network, and Fig. 9b the first channel of the output (for compactness we display only the x-partial derivative). As we can see, the existence of depth discontinuities does not affect the prediction in the interior of the face (we consider the prediction outside this region as noise and we replace it by 0 for visualization).

Several algorithms have been proposed to hallucinate 3D information from a 2D facial image [9], [10], [11], [12]. In order to verify that our decomposition network is extracting real depth information (in lieu of hallucinating it from texture cues), we simulated an image where the pattern is projected over a surface with identical texture but with a planar 3D shape (as in the example illustrated in Fig. 2). Fig. 9a shows the image acquired when the fringes are projected over the ground truth facial depth, and (c) when instead the depth is set to 0 (without modifying the texture information). The first component of the output (x-partial derivative) is shown in (b) and (d), as we can see, the network is actually extracting true depth information (from the deformation of the fringes) and not hallucinating 3D information from texture cues. (As we will see next, this property is particularly useful for joint face recognition and spoofing prevention.)

2D and 3D Face Recognition.

Once the input image is decomposed into a (standard) texture image and depth gradient information, we can proceed to extract 2D and 3D facial features from each component. To this end, state-of-the-art network architectures are evaluated. Our method is agnostic to the RGB and depth feature extractors, moreover, as the retrieved texture image is close to a standard RGB facial images (in sense of the L2-norm), any pre-train 2D feature extractor can be used (e.g., [4], [5], [6], [7], [8]). In the experiments presented in this section we tested a network based on the Xception architecture [49] (details are provided as supplementary material), available online. For the extraction of texture features, the network is trained using FaceScrub [54] dataset (as we previously described, this is a public dataset of 2D facial images). The module that extracts 3D facial features is trained using 2/3 of the subjects of ND-2006 dataset, leaving the remaining subjects exclusively for testing. The output of each module is a 512-dimensional feature vector (see, e.g., Fig. 3), hence the concatenation of 2D +3D features leads to a 1,024-dimensional feature vector. Fig. 10 illustrates a 2D embedding of the texture features, the depth features, and the combination of both. The 2D mapping is learned by optimizing the t-SNE [58] over the train partition, then a random subset of test subjects are mapped for visualization. As we can see, 3D features favor the compactness and increase the distance between clusters associated to different subjects.

To test the recognition performance, the images of the test subjects are partitioned into two sets: gallery and probe. For all the images in both sets, the 2D and 3D feature embedding is computed (using the pre-trained networks described before). Then, for each image in the probe set, the n nearest neighbors in the gallery set are selected. The distance between each sample (in the embedding space) is measured using the distance defined in Section 3, Equation (11). For each sample in the probe set, we consider the classification as accurate, if at least one of the n nearest neighbors is a sample from the same subject. The Rank-n accuracy is the percentage of samples in the probe set accurately classified.

Fig. 11 and Table 1 show the Rank-n accuracy when: only 2D features ($\gamma = 0$), only 3D features ($\gamma = 1$), or a combination of both ($0 < \gamma < 1$) is considered. As explained in Section 3.3, the value of γ can be used to balance the weight of texture and depth features. As we can see, in all the cases a combination of texture and depth information outperforms each of them individually. This is an expected result as classification tends to improve when

independent sources of information are combined [59]. γ is an hyper-parameter that should be set depending on the conditions at deployment. In our particular experiments the best results are obtained for $\gamma = 0.3$, which suggests that RGB features are slightly more reliable than depth features. This is an expected result as the module that extract RGB features is typically trained in a much larger datasets (2D facial images became ubiquitous). We believe this may change if, for example, testing is performed under low light conditions [21]. Testing this hypothesis is one of the potential path for future research. In the experiment discussed so far, we ignored the role of β and α (i.e., we set $\beta = \infty$ and $\alpha = 1$). As we will discuss in the following, these parameters become relevant to achieve jointly face recognition and spoofing prevention.

Robustness to Spoofing Attacks.

Spoofing attack are simulated to test face recognition models, in particular, how robust these frameworks are under (unseen) spoofing attacks. As in the present work we focus on the combination of texture and depth based features, the simulation of spoofing attacks must account for realistic texture and depth models. The models for the synthesis of spoofing attacks are described in detail in the supplementary material Section D.3, available online.

Fig. 12 illustrates spoofing samples (first four rows) and genuine samples (bottom five rows). The first two columns correspond to the ground truth texture and depth information, the third column illustrates the input to our system, and the last three columns correspond to the outputs of the decomposition network. These three last images are fed into the feature extraction modules for the extraction of texture and depth based features respectively, as illustrated in Fig. 3. It is extremely important to highlight, that spoofing samples are included exclusively at testing time. In other worlds, during all the training process the entire framework is agnostic to the existence of spoofing examples. If the proposed framework is capable of extracting real 3D facial features, it should be inherently robust to most common types of spoofing attacks.

As discussed before, the combination of texture and depth based features improves recognition accuracy. On the other hand, when spoofing attacks are included, we observe that texture based features are more vulnerable to spoofing attacks (see for example Figs. 12 and 14). To simultaneously exploit the best of each feature component, we design a non-linear distance as described in Equation (11). Fig. 13 illustrates the properties of the defined distance for different values of α and β. As it can be observed, for those genuine samples (relative distances lower than β) the non linear component can be ignored and the distance behave as the euclidean distance with a relative modulation set by γ. On the other hand, if the distance between the depth components is above the threshold β, it will dominate the overall distance achieving a more robust response to spoofing attacks.

To quantitatively evaluate the robustness against spoofing attacks, spoofing samples are generated for all the subjects in the test set. As before, the test set is separated into a gallery and a probe set and the generated spoofing samples are aggregated into the probe set. For each image in the probe set, the distance to a sample of the same subject in the gallery set is evaluated. If this distance is below a certain threshold λ, the image is labeled as genuine, otherwise, the image is labeled as spoofing. Comparing the classification label with the

ground truth label we obtain the number of true positive (genuine classified as genuine), false positive (spoofing classified as genuine), true negative (spoofing classified as spoofing), and false negative (genuine classified as spoofing). Changing the value of the threshold λ we can control the number of false positive versus the number of false negatives as illustrated in Fig. 14.

Fig. 14 shows the ratio of false positive and false negative for $\lambda \in [0, 2]$. As before the distance between the samples is computed using the definition provided in (11), in blue/red the RGB/depth baseline is illustrated, the other set of curves (displayed in green tones) correspond to a combination of texture and depth features with $\gamma = 0.3$ and different values of α and β. In Table 2 the ratio of true positive is reported for a fixed ratio of false positives. The ACER measure (last column) corresponds to the average between the ratio of spoofing and genuine samples misclassified.

Testing Variations on the Ambient Illumination.

To test the impact of variations on lighting conditions we simulated test samples under different ambient illumination, implementation details are described in the supplementary material Section D.4, available online. Table 3 compares the rank-5 accuracy of 2D features and 2D+3D features as the power of the ambient illumination increases. As described in the supplementary material, available online, the ambient illumination is modeled with random orientation, and therefore, the more powerful the illumination is the more diversity between the test and the gallery samples is introduced.

In the present experiments, we assumed that both the projected pattern and the ambient illumination have similar spectral content. In practice, one can project the pattern, e.g., on the infrared band. This would make the system invisible to the user, and reduce the sensitivity of 3D features to variations on the ambient illuminations. We provide a hardware implementation feasibility study and illustrate how the proposed ideas can be deployed in practice in the supplementary material Section E, available online.

Improving State of the Art 2D Face Recognition.

To test how the proposed ideas can impact the performance of state-of-the-art 2D face recognition systems, we evaluated our features in combination with texture based features obtained with ArcFace [5]. ArcFace is a powerful method pre-trained on very large datasets, on ND-2006 examples it achieves perfect recognition accuracy (100 percent rank-1 accuracy). When ArcFace is combined with the proposed 3D features, the accuracy remains excellent (100 percent rank-1 accuracy), i.e., adding the proposed 3D features does not negatively affects robust 2D solutions. On the other hand, 3D features improve ArcFace on challenging conditions as we discuss in the following. Interesting results are observed when ArcFace is tested under spoofing attacks, as we show in Table 4, ArcFace fails to detect spoofing attacks. ArcFace becomes more robust when it is combined with 3D features, improving from nearly 0 TPR@FPR(10^{-3}) to 84 percent. In summary, as 2D methods improve and become more accurate, our 3D features do not affect them negatively when they work well, while improve their robustness in challenging situations.

5 Conclusion

We proposed an effective and modular alternative to enhance 2D face recognition methods with actual 3D information. A high frequency pattern is designed to exploit the high resolution cameras ubiquitous in modern smartphones and personal devices. Depth gradient information is coded in the high frequency spectrum of the captured image while a standard texture facial image can be recovered to exploit state-of-the-art 2D face recognition methods. We show that the proposed method can be used to simultaneously leverage 3D information and texture information. This allows us to enhance state-of-the-art 2D methods improving their accuracy and making them robust, e.g., to spoofing attack.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by ARO, ONR, NSF, and NGA.

Biography

J. Matias Di Martino received the BSc and PhD degrees in electrical engineering from Universidad de la Republica, Uruguay, in 2011 and 2015 respectively. During 2016–2017, he was a research associate at Ecole Normale Superieure de Cachan, Paris. Since 2017, he has been working as assistant professor at the Physics Department of the School of Engineer, Universidad de la Republica, and since 2019, he has also been a research assistant professor with the Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina. His main areas of interest are applied optics, machine learning, facial and behavioral analysis, and image processing.

Fernando Suzacq received the BSc degree in computer engineering from Universidad de la República, Uruguay, in 2013. Currently, he is working toward the MSc degree in electrical engineering at Universidad de la República, Uruguay. He has been working actively in the industry more than the past ten years building software and machine learning systems. His main areas of interest are computer vision, particularly face recognition, and machine learning.

Mauricio Delbracio received the BSc degree in electrical engineering from UdelaR, Montevideo, in 2006, and the MSc and PhD degrees in applied mathematics from cole Normale Suprieure de Cachan (ENS-Cachan), France, in 2009 and 2013 respectively. He is currently a research scientist at Google Research. Before joining Google, in 2019, he was an assistant professor with the Department of Electrical Engineering, Universidad de la República (UdelaR), Uruguay. From 2013 to 2016 he was a postdoctoral researcher with the ECE Department, Duke University, Durham, North Carolina. His research interests include image and signal processing, computer graphics, computational imaging, and machine learning. His current research focuses on algorithms, data analysis and applications of machine learning to image and signal processing. In 2016 he was awarded the Early Career Prize from the Society for Industrial and Applied Mathematics (SIAM) Activity Group on Imaging Science, in 2016 for his important contributions to image processing.

Qian Qiu received the bachelor's degree (with first class honors) in computer science, in 2001, the master's degree in computer science from National University of Singapore, Singapore, in 2002, and the PhD degree in computer science from University of Maryland, College Park, in 2013. During 2002–2007, he was a senior research engineer at Institute for Infocomm Research, Singapore. He is currently an assistant research professor with the Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina. His research interests include computer vision and machine learning, specifically on face recognition, human activity recognition, image classification, and representation learning.

Guillermo Sapiro (Fellow, IEEE) received the BSc (summa cum laude), MSc, and PhD degrees from the Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel, in 1989, 1991, and 1993 respectively. After postdoctoral research at MIT, he became Member of Technical Staff at the research facilities of HP Labs in Palo Alto, California. He was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota, where he held the position of Distinguished McKnight University professor and Vincentine Hermes-Luh chair in

Electrical and Computer Engineering. Currently he is the Edmund T. Pratt, Jr. school professor with Duke University, Durham, North Carolina. He works on theory and applications in computer vision, computer graphics, medical imaging, image analysis, and machine learning. He has authored and coauthored more than 400 papers in these areas and has written a book published by Cambridge University Press, January 2001. He was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies, in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work, in 1992, the Rothschild Fellowship for Post-Doctoral Studies, in 1993, the Office of Naval Research Young Investigator Award, in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE), in 1998, the National Science Foundation Career Award, in 1999, and the National Security Science and Engineering Faculty Fellowship, in 2010. He received the test of time award at ICCV 2011 and ICML 2019. He is a fellow of the American Academy of Arts and Sciences, and SIAM. He was the founding editorin-chief of the *SIAM Journal on Imaging sciences*.

References

[1]. Ding C and Tao D, “A comprehensive survey on pose-invariant face recognition,” ACM Trans. Intell. Syst. Technol, vol. 7, no. 3, 2016, Art. no. 37.

[2]. Kemelmacher-Shlizerman I, Seitz SM, Miller D, and Brossard E, “The MegaFace benchmark: 1 million faces for recognition at scale,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4873–4882.

[3]. Nech A and Kemelmacher-shlizerman I, “Level playing field for million scale face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7044–7053.

[4]. Taigman Y, Yang M, Ranzato M, and Wolf L, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.

[5]. Deng J, Guo J, Xue N, and Zafeiriou S, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4690–4699.

[6]. Parkhi OM. Deep face recognition; in Proc. Brit. Mach. Vis. Conf.; 2015. 41

[7]. Schroff F, Kalenichenko D, and Philbin J, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 815–823.

[8]. Zulqarnain S, Ajmal G, Science C, and Engineering S, “Learning from millions of 3D scans for large-scale 3D face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1895–1905.

[9]. Eigen D, Puhrsch C, and Fergus R, “Depth map prediction from a single image using a multi-scale deep network,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2014, pp. 2366–2374.

[10]. Huber P, Hu G, Tena R, Mortazavian P, and Koppen WP, “A multiresolution 3D morphable face model and fitting framework,” in Proc. 11th Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory Appl., 2015, pp. 79–86.

[11]. Liu F, Shen C, and Lin G, “Deep convolutional neural fields for depth estimation from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 5162–5170.

[12]. Pini S, Grazioli F, Borghi G, Vezzani R, Emilia R, and Cucchiara R, “Learning to generate facial depth maps,” in Proc. Int. Conf. 3D Vis., 2018, pp. 634–642.

[13]. Zafeiriou S et al., “Face recognition and verification using photometric stereo: The photoface database and a comprehensive evaluation,” IEEE Trans. Inf. Forensics Security, vol. 8, no. 1, pp. 121–135, 1 2013.

[14]. Zou X, Kittler J, Messer K, and Kingdom U, “Face Recognition Using Active Near-IR Illumination,” in Proc. Brit. Mach. Vis. Conf., 2005, pp. 24.1–24.11.

[15]. Kaya Y and Kobayashi K, “A basic study on human face recognition,” Front. Pattern Recogn, vol. 1, pp. 265–289, 1972.

[16]. Zhao W, Chellappa R, Phillips PJ, and Rosenfeld A, “Face recognition: A literature survey,” ACM Comput. Surv, vol. 35, no. 4, pp. 399–458, 2003.

[17]. Cao K, Rong Y, Li C, Tang X, and Loy CC, “Pose-robust face recognition via deep residual equivariant mapping,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5187–5196.

[18]. Hayat M, Khan SH, Werghi N, Goecke R, Dhabi A, and Emirates UA, “Joint registration and representation learning for unconstrained face identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2767–2776.

[19]. He L, Li H, Zhang Q, Sun Z, and Technology I, “Dynamic feature learning for partial face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7054–7063.

[20]. Kumar A and Chellappa R, “Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 430–439.

[21]. Lezama J, Qiu Q, and Sapiro G, “Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6807–6816.

[22]. Liu Y. Exploring disentangled feature representation beyond face identification; in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.; 2018. 2080–2089.

[23]. Tran L, Yin X, and Liu X, “Disentangled representation learning GAN for pose-invariant face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1283–1292.

[24]. Yu X and Porikli F, “Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5367–5375.

[25]. Zhao J. Towards pose invariant face recognition in the wild; in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.; 2018. 2207–2216.

[26]. Deng J, Cheng S, Xue N, Zhou Y, and Zafeiriou S, “UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7093–7102.

[27]. Liu W, Wen Y, Yu Z, Li M, Raj B, and Song L, “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 212–220.

[28]. Wang H. CosFace: Large margin cosine loss for deep face recognition; in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.; 2018. 5265–5274.

[29]. Blanz V and Vetter T, “Face recognition based on fitting a 3D morphable model,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 25, no. 9, pp. 1063–1074, 9 2003.

[30]. Dou P, Shah SK, and Kakadiaris IA, “End-to-end 3D face reconstruction with deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5908–5917.

[31]. Cui J, Zhang H, Han H, Shan S, Chen X, and Technology I, “Improving 2D face recognition via discriminative face depth estimation,” in Proc. Int. Conf. Biometrics, 2018, pp. 1–8.

[32]. Hartley R and Zisserman A, Multiple View Geometry in Computer Vision. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[33]. Marks M, “System and devices for time delay 3D,” U.S Patent 5,151,821, Sep. 29, 1992.

[34]. Eigen D and Fergus R, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 2650–2658.

[35]. Laina I, Rupprecht C, Belagiannis V, Tombari F, and Navab N, “Deeper depth prediction with fully convolutional residual networks,” in Proc. 4th Int. Conf. 3D Vis., 2016, pp. 239–248.

[36]. Prados E and Faugeras O, “Shape from shading,” in Handbook of Mathematical Models in Computer Vision. Berlin, Germany: Springer, 2006, pp. 375–388.

[37]. Saxena A, Sun M, and Ng AY, “Make3D: Learning 3D scene structure from a single still image,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 31, no. 5, pp. 824–840, 5 2009. [PubMed: 19299858]

[38]. Ayubi GA, Ayubi JA, Di Martino JM, and Ferrari JA, “Pulse-width modulation in defocused three-dimensional fringe projection,” Opt. Lett, vol. 35, no. 21, pp. 3682–3684, 2010. [PubMed: 21042390]

[39]. Di Martino JM, Fernandez A, and Ferrari JA, “One-shot 3D gradient field scanning,” Opt. Lasers Eng, vol. 72, pp. 26–38, 2015.

[40]. Li B, Wang Y, Dai J, Lohry W, and Zhang S, “Some recent advances on superfast 3D shape measurement with digital binary defocusing techniques,” Opt. Lasers Eng, vol. 54, pp. 236–246, 2014.

[41]. Rosman G and Rus D, “Information-driven adaptive structured-light scanners,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 874–883.

[42]. Zhang S and Yau S-T, “High-resolution, real-time 3D absolute coordinate measurement based on a phase-shifting method,” Opt. Express, vol. 14, no. 7, pp. 2644–2649, 2006. [PubMed: 19516395]

[43]. Di Martino M, Flores J, and Ferrari JA, “One-shot 3D scanning by combining sparse landmarks with dense gradient information,” Opt. Lasers Eng, vol. 105, pp. 188–197, 2018.

[44]. Zhang S, “Recent progresses on real-time 3D shape measurement using digital fringe projection techniques,” Opt. Lasers Eng, vol. 48, no. 2, pp. 149–158, 2010.

[45]. Zhang S, Handbook of 3D Machine Vision: Optical Metrology and Imaging. Boca Raton, FL, USA: CRC Press, 2013.

[46]. Takeda M and Mutoh K, “Fourier transform profilometry for the automatic measurement of 3-D object shapes,” Appl. Opt, vol. 22, no. 24, pp. 3977–3982, 1983. [PubMed: 18200299]

[47]. Schwartz L, “Théorie des distributions à valeurs vectorielles,” in Annales de l’institut Fourier, vol. 7, pp. 1–141, 1957.

[48]. Pritt MD and Ghiglia DC, Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software. Hoboken, NJ, USA: Wiley, 1998.

[49]. Chollet F, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.

[50]. Zheng Y, Pal DK, and Savvides M, “Ring loss: Convex feature normalization for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5089–5097.

[51]. Boulkenafet Z, Komulainen J, Li L, Feng X, and Hadid A, “OULU-NPU: A mobile face presentation attack database with real-world variations,” in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., 2017, pp. 612–618.

[52]. Liu Y, Jourabloo A, and Liu X, “Learning deep models for face anti-spoofing: Binary or auxiliary supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 389–398.

[53]. Zhang X, Hu X, Chen C, and Peng S, “Face spoofing detection based on 3D lighting environment analysis of image pair,” in Proc. IEEE Int. Conf. Pattern Recognit., 2016, pp. 2995–3000.

[54]. Ng H-W and Winkler S, “A data-driven approach to cleaning large face datasets,” in Proc. IEEE Int. Conf. Image Process., 2014, pp. 343–347.

[55]. Zhang Z, Yan J, Liu S, Lei Z, Yi D, and Li SZ, “A face anti-spoofing database with diverse attacks,” in Proc. Int. Conf. Biometrics, 2012, pp. 26–31.

[56]. Faltemier KBTC and Flynn P, “Using a multi-instance enrollment representation to improve 3D face recognition,” in Proc. 1st IEEE Int. Conf. Biometrics: Theory Appl. Syst., 2007, pp. 1–6.

[57]. King DE, “Dlib-ml: A machine learning toolkit,” J. Mach. Learn. Res, vol. 10, no. Jul, pp. 1755–1758, 2009.

[58]. Maaten LVD and Hinton G, “Visualizing data using t-SNE,” J. Mach. Learn. Res, vol. 9, no. Nov, pp. 2579–2605, 2008.

[59]. Kuncheva LI, Combining Pattern Classifiers: Methods and Algorithms. New York, NY, USA: Wiley-Interscience, 2004.

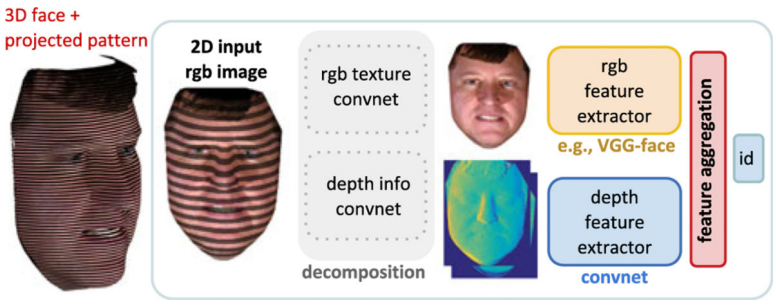

Fig. 1.

Real 3D face recognition is possible by capturing one single RGB image if a high frequency pattern is projected. The low frequency components of the captured image can be fed into a state-of-the-art 2D face recognition method, while the high frequency components encode local depth information that can be used to extract 3D facial features. It is important to highlight that, in contrast with most existing 3D alternatives, the proposed approach provides real 3D information, not 3D hallucination from the RGB input. As a result, state-of-the-art 2D face recognition methods can be enhanced with real 3D information.

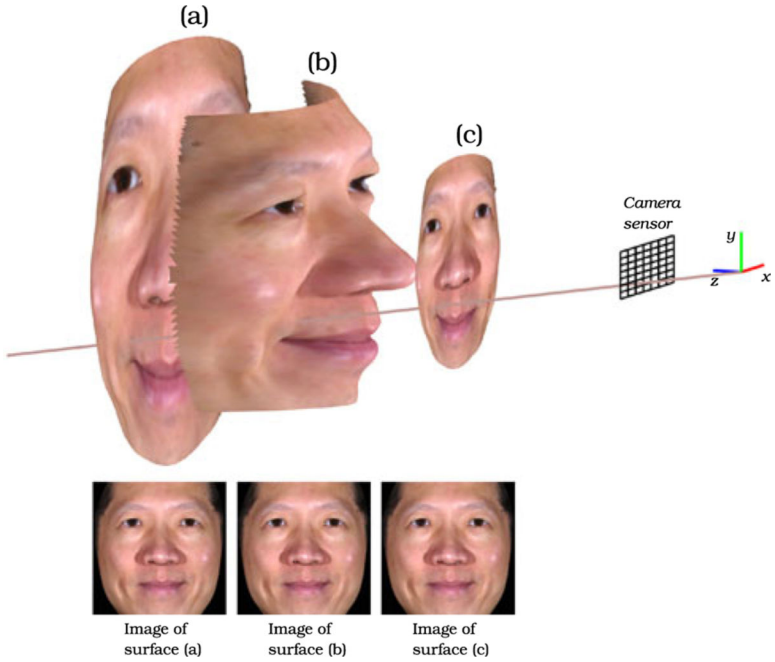

Fig. 2.

Illustration of three different 3D surfaces that look equivalent from a monocular view (single RGB image). On top, three surfaces (a), (b) and (c) are simulated, being (a) and (c) flat and (b) the 3D shape of a test subject. We use classic projective geometry [32] and simulate the image we obtain when photographing (a), (b) and (c) respectively. The resulting images are shown at the bottom. As we illustrate with this simple example, the relation between images and 3D scenes is not bijective and the problem of 3D hallucination is ill-posed. To overcome this, 3D hallucination solutions enforce important priors about the geometry of the scene. This is why we argue, that these methods do not really add to the face recognition task, actual 3D information. (A complementary example is presented in Fig. 15 in the supplementary material), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.2986951.

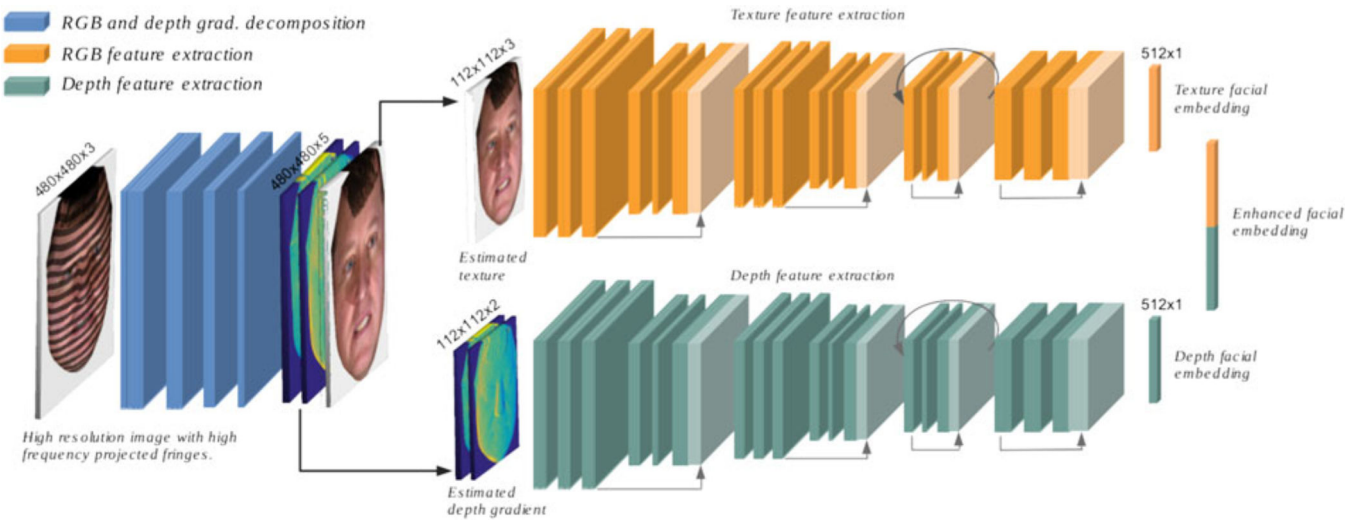

Fig. 3.

Architecture overview. First a network (illustrated in blue) is used to decompose the input image that contains overlapped high frequency fringes into a lower resolution (standard) texture facial image and depth gradient information. The former is used as the input of a state-of-the-art 2D face recognition DNN (yellow blocks). The depth information is fed to another network (green blocks) trained to extract discriminative (depth-based) facial features. Different network architectures are tested, we provide implementation details in Section D in the supplementary material, available online.

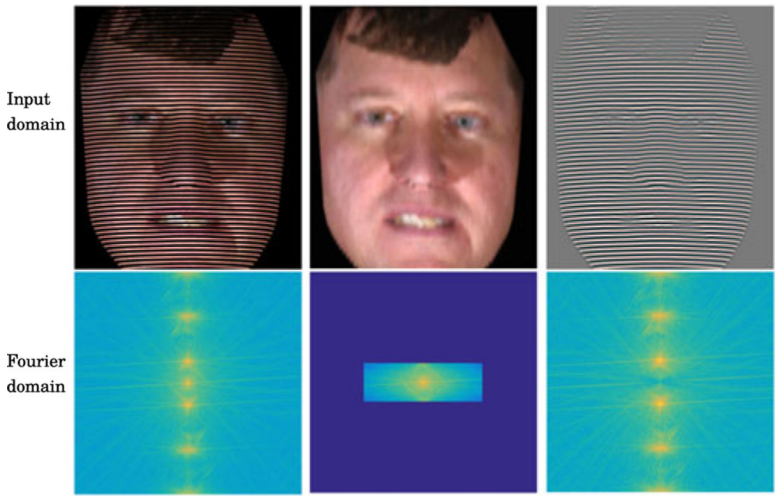

Fig. 4.

2D plus real 3D in a single rgb image. The first column illustrates the RGB image acquired by a (standard) camera when horizontal stripes are projected over the face. The second column isolates the low frequency components of the input image, and the third column corresponds to the residual high frequency components. (In all the cases the absolute value of the Fourier Transform is represented in logarithmic scale). As can be seen, high frequency patterns can be used to extract 3D information of the face (third column) while preserving a lower resolution version of the facial texture (middle column).

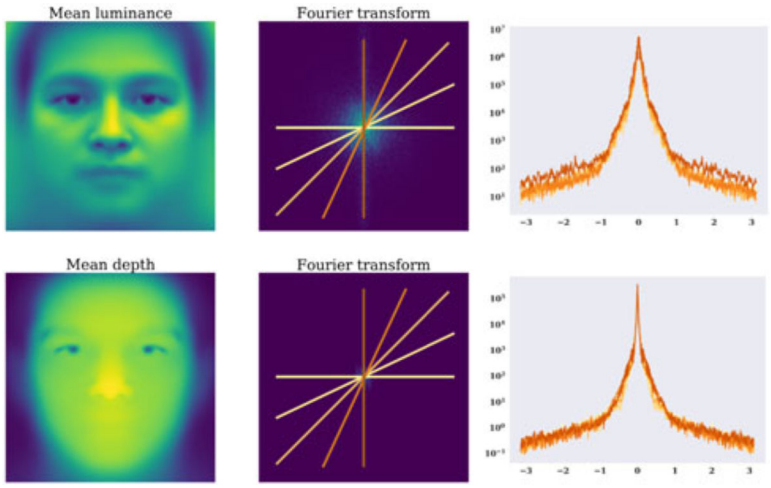

Fig. 5.

Faces average spectral content. The first column illustrates the mean luminance and depth map for the faces in the dataset ND-2006. The second column shows the mean Fourier Transform of the faces luminance and depth respectively. The third column shows the profile across different sections of the 2D Fourier domain. Columns two and three represent the absolute value of the Fourier transform in logarithmic scale. Faces are registered using the eyes landmarks and the size normalized to 480 × 480 pixels.

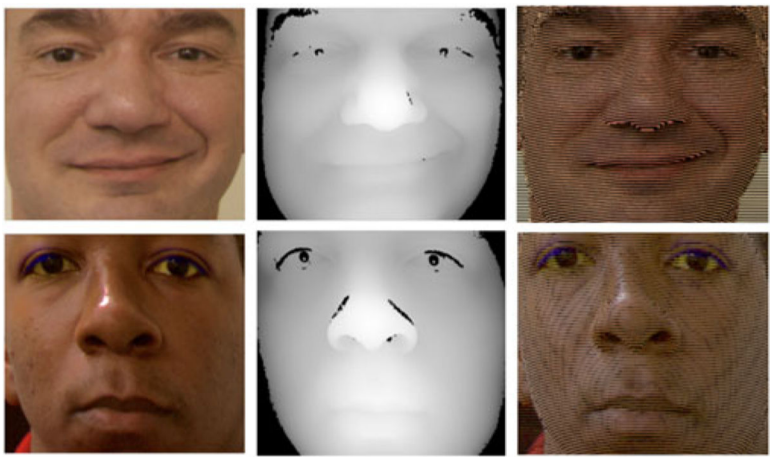

Fig. 6.

Active light projection. From left to right: ground truth RGB facial image, 3D facial scanner, and finally the image we would acquire if the designed high frequency pattern is projected over the face. Two random samples from ND-2006 are illustrated.

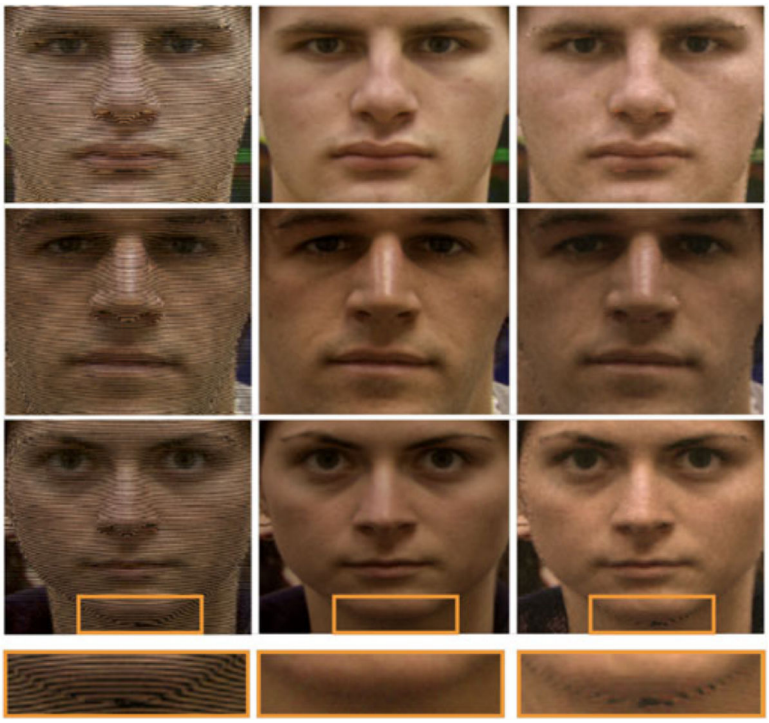

Fig. 7.

Examples of the facial texture recovered from the image with the projected pattern. The first column, shows the input image (denoted as I in Algorithm 1). The second column shows the ground truth, and the third column the texture recovered by the network I_{rgb}. This examples are from the test set and the images associated to these subjects were never seen during the training phase.

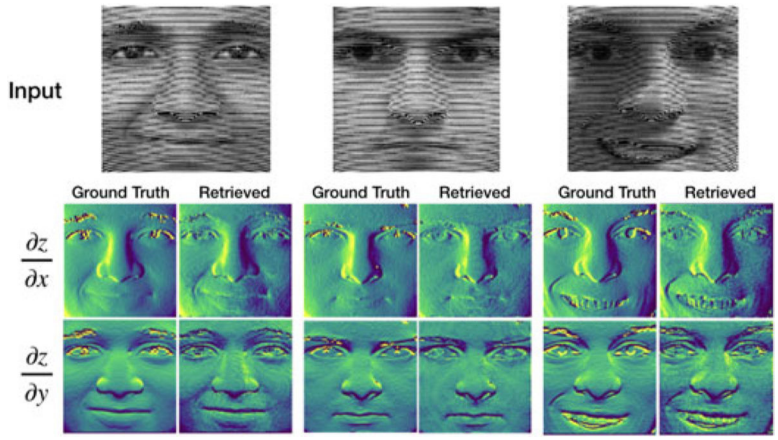

Fig. 8.

Differential depth information extracted from the image with the projected pattern. The first row illustrates the input image (depth information can be extracted from a gray version of the input as the designed patter is achromatic). The second and third row show the ground truth and the retrieved x and y partial derivatives of the depth respectively.

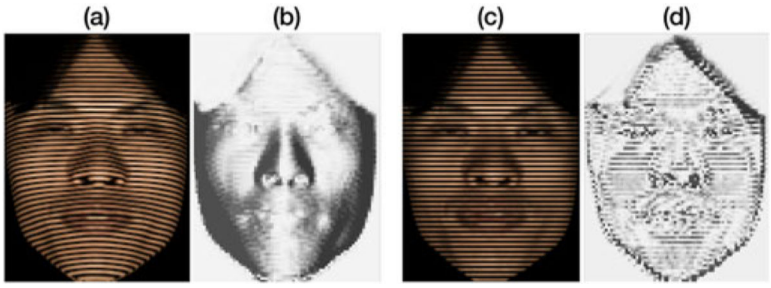

Fig. 9.

Is the network really extracting depth information? In this figure we show the output of the network for two inputs generated using identical facial texture but different depth ground truth data. (a) Image obtained when the projected pattern is projected over the face with the real texture and the real 3D profile. (b) Output of the network when we input (a) (only the x-partial derivative is displayed for compactness). (c) Image obtained when the projected pattern is projected over a flat surface with the texture of the real face. (d) Output of the network when the input is (c). None of these images were seen during training.

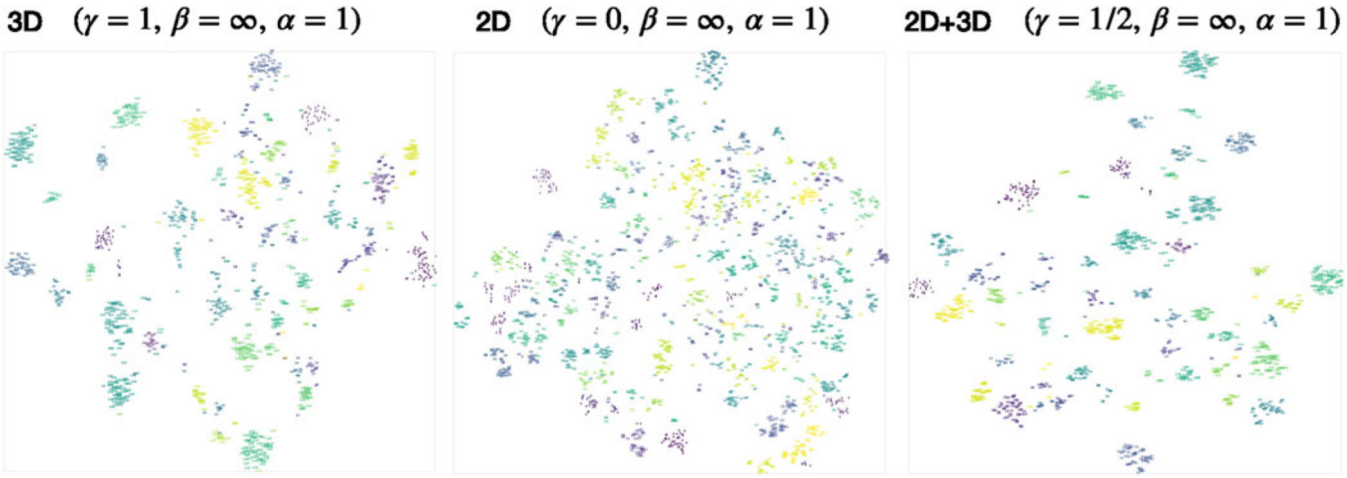

Fig. 10.

Facial features low dimensional embedding (for visualization purposes only). We illustrate texture-based and depth-based features in a low dimensional embedding space. A random set of subject of the test set is shown. From left to right: the embedding of depth-features, texture-based features, and finally, the combination of texture and depth features. t-SNE [58] algorithm is used for the low-dimensional embedding.

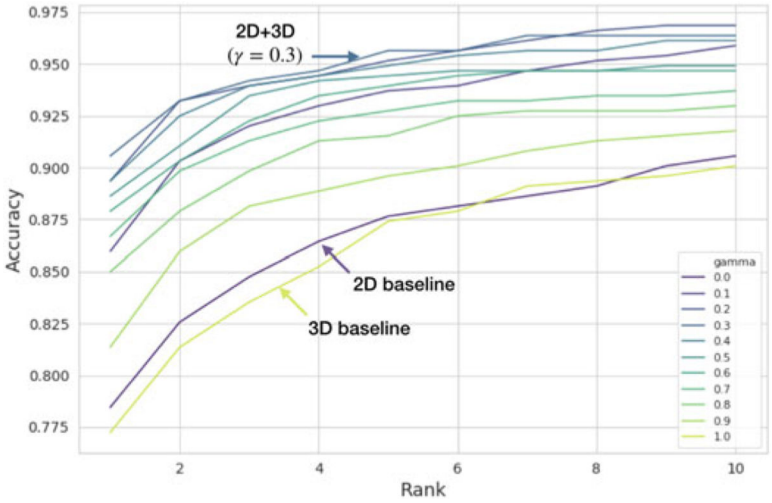

Fig. 11.

Rank-n accuracy for 2D, 3D, and 2D+3D face recognition. As discussed in Section 3 the value of γ can be set to weight texture and depth information in the classification decision. The extreme cases are $\gamma = 0$ (only texture is considered) and $\gamma = 1$ (only depth is considered). These extreme cases are illustrated in yellow and blue respectively, while intermediate solutions ($0 < \gamma < 1$) are presented in tones of green.

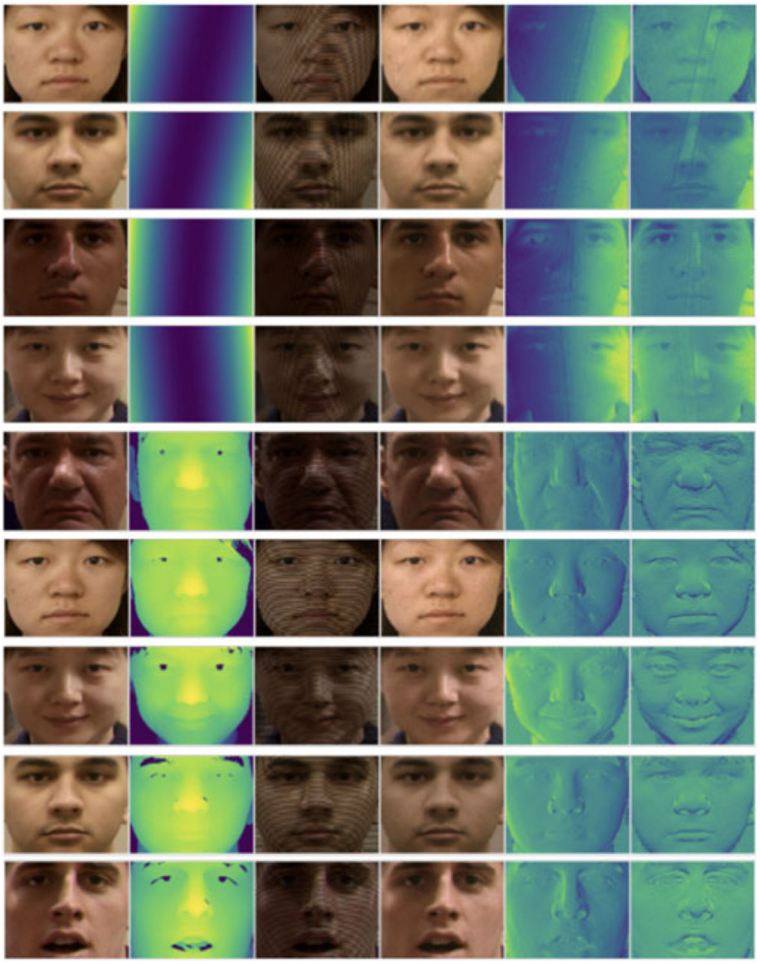

Fig. 12.

Examples of samples from live subjects and spoofing attacks. From left to right: (1) the ground truth texture, (2) the ground truth depth, (3) the input to our system (image with the projected pattern), (4) the recovered texture component (one of the outputs of the decomposition network), (5)/(6) recovered x/y depth partial derivative. The first four rows correspond to spoofing samples (as explained in Section D.3), and the bottom five rows to genuine samples from live subjects.

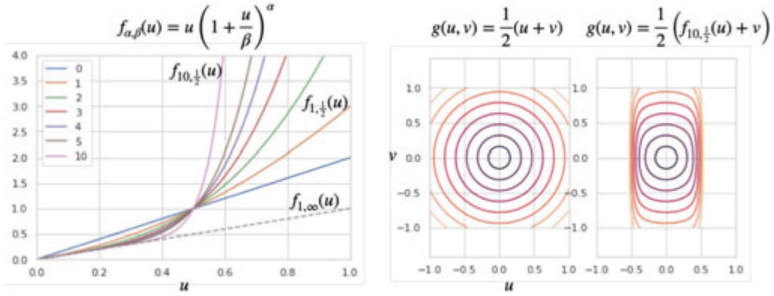

Fig. 13.

Illustration of the properties of the distance function defined in (11). On the left side we illustrate the role of the parameter α, and on the right, we compare the proposed distance and the standard euclidean distance. As can be observed, both measures are numerically equivalent in the region $[-\beta/2, \beta/2] \times [-\beta/2, \beta/2]$, but the proposed measure gives a higher penalty to vectors whose u coordinate exceeds the value β.

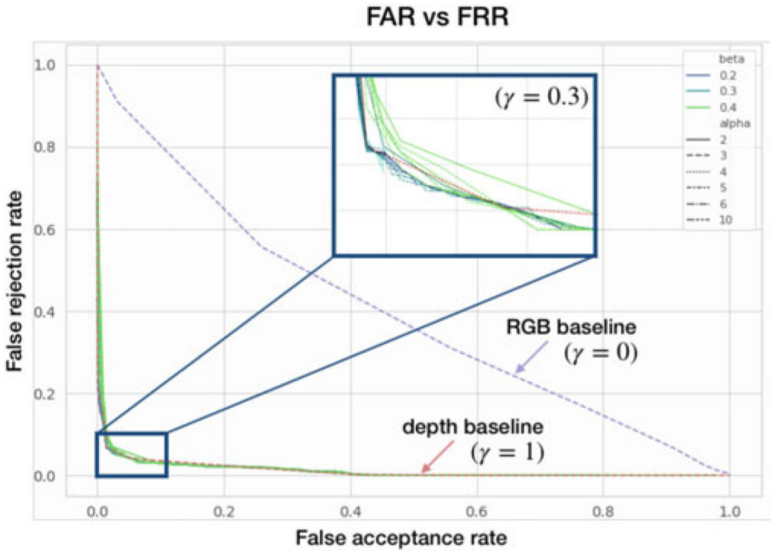

Fig. 14.

False acceptance rate and false rejection rate under the presence of spoofing attacks. On color blue we illustrate the RGB baseline ($\gamma = 0$), on the other extreme, the red curve illustrates the performance when only depth features are considered. The combination of RGB and depth features is illustrated in tones of green for different values of α and β (in this experiment we set $\gamma = 0.3$).

TABLE 1

Rank-n Accuracy for 2D, 3D, and 2D+3D Face Recognition

Rank-n Accuracy	**1**	**2**	**5**	**10**
RGB baseline ($\gamma = 0$)	78.5	82.6	87.7	90.6
(Depth baseline ($\gamma = 1$)	77.2	81.4	87.4	90.1
(our) $\gamma = 0.3$	**90.6**	**93.2**	**95.6**	**96.4**
(our) $\gamma = 0.5$	88.6	91.0	94.4	94.9
(our) $\gamma = 0.8$	85.0	87.9	91.5	93.0

As discussed in Section 3 the value of γ can be set to weight the impact of texture and depth information. The extreme cases are $\gamma = 0$ (only texture is considered) and $\gamma = 1$ (only depth is considered).

TABLE 2

Spoofing detection results

	TPR% @FPR = 10^{-3}	**TPR% @FPR = 10^{-2}**	**ACER %**
RGB baseline ($\gamma = 0$)	21.8	24.0	38.9
Depth baseline ($\gamma = 1$)	**88.4**	**97.1**	4.0
(our) $\gamma = 0.3$, $\beta = 0.35$ $\alpha = 2$	85.5	96.9	4.5
(our) $\gamma = 0.3$, $\beta = 0.35$ $\alpha = 5$	83.8	**97.1**	4.0
(our) $\gamma = 0.3$, $\beta = 0.35$ $\alpha = 10$	85.0	95.6	**3.9**
(our) $\gamma = 0.3$, $\beta = 0.4$ $\alpha = 2$	82.6	96.9	4.7
(our) $\gamma = 0.3$, $\beta = 0.4$ $\alpha = 5$	86.4	**97.1**	4.4
(our) $\gamma = 0.3$, $\beta = 0.4$ $\alpha = 10$	81.8	**97.1**	4.1
(our) $\gamma = 0.3$, $\beta = 0.5$ $\alpha = 2$	86.4	96.4	5.3
(our) $\gamma = 0.3$, $\beta = 0.5$ $\alpha = 5$	82.8	95.6	5.7
(our) $\gamma = 0.3$, $\beta = 0.5$ $\alpha = 10$	85.0	94.4	5.9

The ratio of true positive for a fixed ratio of false positive and the ACER measure are reported. Texture and depth facial features are combined using the distance defined in (11). As we can see, the parameters γ, α, and β can be set to obtain better facial recognition performance and robustness against spoofing detection.

TABLE 3

Recognition Accuracy Under Different Ambient Illumination Conditions

Rank-5 Accuracy	**power = 100%**	**power = 150%**	**power = 200%**
RGB baseline ($\gamma = 0$)	89.2	81.2	53.9
(our) $\gamma = 0.5$	93.6	90.7	80.7

The power of the additional ambient light is provided relative to the power of the projected light, i.e., power = 200% means that the added ambient illumination is twice as bright as the projected pattern.

TABLE 4

Spoofing Detection Results for ArcFace and ArcFace Enhanced With 3D Features

	TPR% @FPR = 10^{-3}	**TPR% @FPR = 10^{-2}**	**ACER %**
ArcFace ($\gamma = 0$)	0	0	46.2
(ArcFace + 3D) ($\gamma = 0.5$)	84.7	94.7	7.9

Like in Table 2, the ratio of true positive for a fixed ratio of false positive and the ACER measure are reported.