Detailed 3D Face Reconstruction from a Single RGB Image

Gemma Rotger[1]
grotger@cvc.uab.es

Francesc Moreno-Noguer[2]
fmoreno@iri.upc.edu

Felipe Lumbreras[1]
felipe@cvc.uab.es

Antonio Agudo[2]
aagudo@iri.upc.edu

[1] Computer Vision Center and Departament Ciències de la Computació, UAB
UAB Campus O Building, (08193) Cerdanyola del Vallès, Spain

[2] Institut de Robòtica i Informàtica Industrial, CSIC-UPC
4-6 Llorens i Artigas, (08028) Barcelona, Spain

ABSTRACT

This paper introduces a method to obtain a detailed 3D reconstruction of facial skin from a single RGB image. To this end, we propose the exclusive use of an input image without requiring any information about the observed material nor training data to model the wrinkle properties. They are detected and characterized directly from the image via a simple and effective parametric model, determining several features such as location, orientation, width, and height. With these ingredients, we propose to minimize a photometric error to retrieve the final detailed 3D map, which is initialized by current techniques based on deep learning. In contrast with other approaches, we only require estimating a depth parameter, making our approach fast and intuitive. Extensive experimental evaluation is presented in a wide variety of synthetic and real images, including different skin properties and facial expressions. In all cases, our method outperforms the current approaches regarding 3D reconstruction accuracy, providing striking results for both large and fine wrinkles.

Keywords

3D Wrinkle Reconstruction, Face Analysis, Optimization.

1 INTRODUCTION

Reconstructing the 3D human geometry from RGB images has been extensively studied in computer vision and computer graphics in the past two decades [AMCMN16, AMAC17, GVWT13, GZC+16, LZZL16, SWTC14, ZTG+18, CCC+18]. These results can be applied to many everyday applications from the movie industry to medical purposes, robotics, gaming, or human-computer interaction, to name a few. A significant area of research is focused on the acquisition of human faces. Unfortunately, it is well known that human faces are highly variable, differing widely between gender, age, ethnicity, and gesture expression, making the task harder. For this reason, the study of high-detailed face geometry has transcended amongst other methods.

Face reconstruction from monocular information is known to be a highly under-constrained problem, requiring the use of additional priors to constrain the solution. Early approaches proposed to address the problem through the use of facial 3D morphable methods (3DMM) [BV99, CWZ+14], allowing to recover the large-scale geometry. Thanks to the parametric fitting problem nature, they yield in a considerable dimensionality reduction of the solution space. Nevertheless, due to the generality of these approaches, they are not suitable to retrieve subject-specific details, such as wrinkles and scars, being necessary the use of refining algorithms.

Recently, the use of Deep Learning (DL) strategies has been proposed for face reconstruction from RGB information. In this context, many efforts have been done by proposing 3DMM-based models [RSOEK17], volumetric regression [JBAT17], and learning from synthetic data [RSK16]. While these techniques have proved a good performance, the solutions have a remarkable lack of individual detail even after training specific refining networks. A direct solution would be increasing the amount of training data, exploring the solution space of the wrinkles, which often is not trivial in practice. Notably, most of the before mentioned methods need a refinement step to retrieve individual details.

In this paper, we present a novel optimization framework that combines the use of wrinkle properties, directly extracted from the input image with a photomet-

ric energy term, in order to recover detailed 3D faces from a single image with the highest level of detail possible. On the one hand, we take advantage of current solutions based on deep learning [JBAT17] to obtain a coarse initialization. While this initializing solution cannot recover fine details, it is accurate enough to estimate an initial 3D mesh shape. Recall that we only use deep learning for initialization, i.e., no training data is employed to model wrinkles. On the other hand, we estimate the illumination parameters required later to define a photometric energy term. Finally, we encode wrinkles by using smooth functions. For this purpose, we find wrinkles that can be accurately described in the image up to a depth parameter, since relief caused by wrinkles produce visible shading in the image. We then group the connected pixels after filtering noisy measurements (removing unconnected pixels and pixels with low response values), since they have similar displacement behavior. After that, we operate locally fitting a second-order polynomial that better adapts to the pixel coordinates of each wrinkle detected in the image. With the distance from the pixel to the polynomial curve, the overall wrinkle location, orientation, height, and width, we can model the wrinkle displacement and transfer it to the node which is related to the pixel. We have to take into account that, depending on the mesh and image resolutions, not all the pixels must have a correspondent vertex. As much the pixel-to-vertex correspondences lead to one-to-one, the better the performance.

Our approach offers a realistic solution for high-resolution single image 3D face reconstruction, which differs with current methods since it is both fast and effective, as well as can be easily adapted as a refinement strategy for most of the initializing formulations. We present experimental validation in a wide variety of synthetic and real images, including different skin properties and facial expressions, showing the suitability of our approach even in extreme cases like special effects facial makeup.

2 RELATED WORK

The 3D reconstruction of faces from RGB images has remained an appealing topic. Several methods addressed the problem from different perspectives. First, parametric methods represent a face solution as a linear combination of shape bases, which can be inferred directly from data [SKSS14, AMN19, AMAC17], or predefined in advance [LXC+15]. From the perspective of RGB-D data, other approaches [ZNI+14] required a template to reconstruct a fully rigged face. However, the shape and size variance of the face becomes a limitation to this model. In particular, when the skin produces subject-specific wrinkles and folds on aging and expression, the capture of these medium and fine-scale details compounds the problem.

In the recent years, certain multi-camera [GSSM15], binocular [GFT+11, VWB+12], and monocular approaches [AMN17, AMN18, GVWT13, GZC+16, LZZL16, SWTC14] have obtained remarkable detailed results. Garrido *et al.* [GVWT13] appended detail to a personalized blend shape model, and Shi *et al.* [SWTC14] proposed an iterative process between large-scale reconstruction and fine-scale per-pixel shading cues. More recently, in [GZC+16] has combined shape from shading and learning of a generative detail deformation model, whilst in [LZZL16] proposed to regress a cascaded 3D face shapes model in 2D and 3D spaces. All of them presented accurate solutions over monocular videos but they need a complex setup, as a manual initialization [GVWT13]. Eventually, in [CBZB15] was presented as the first real-time approach by training local regressors to predict wrinkles in monocular sequences. While these approaches produced a great advance in the topic, most of them relied on spatio-temporal priors which become not suitable for single image reconstruction.

Estimating the 3D reconstruction from a single image is a severely under-constrained problem. Early approaches [GMMB00, WWY06, WBS01] were hardly able to recover simple objects like geometric figures and wall-like scenes. The 3D reconstruction of human faces arose with the introduction of 3DMM [BV99, CWZ+14], where the problem was simplified to a low-rank fitting problem, thus not the full geometry was estimated but a set of parameters controlling identity and expression. We can find two different groups in the literature: model-based [HHT+16, JZD+18, RV05] and data-based [RSK16, RSOEK17] approaches. Regarding model-based approaches, they frequently use a non-linear least squares multi-feature fitting, which retrieves a 3D facial structure and the corresponding texture map [RV05], performs under different resolution levels [HHT+16], and fits a photometric minimization energy term to find the parameters that better adjusts the input image [JZD+18]. Even these methods achieve a considerable level of detail, they require plenty of time to offer detailed results. Furthermore, the 3DMM may not yield in a correct result if the given face it is not well represented in the low-rank model.

As in most of the fields, DL approaches attempt to provide an accurate solution to detailed face reconstruction from a single image [JBAT17, RSK16, RSOEK17, TTHM+18, CCL18]. However, we find in DL solutions the lack of a detailed reconstruction –the most obvious of the alleged detailed methods–, since the current DL approaches are not able to directly predict a face with detailed facial features such as wrinkles or scars. In the specific case of [RSK16] they refine their solution by

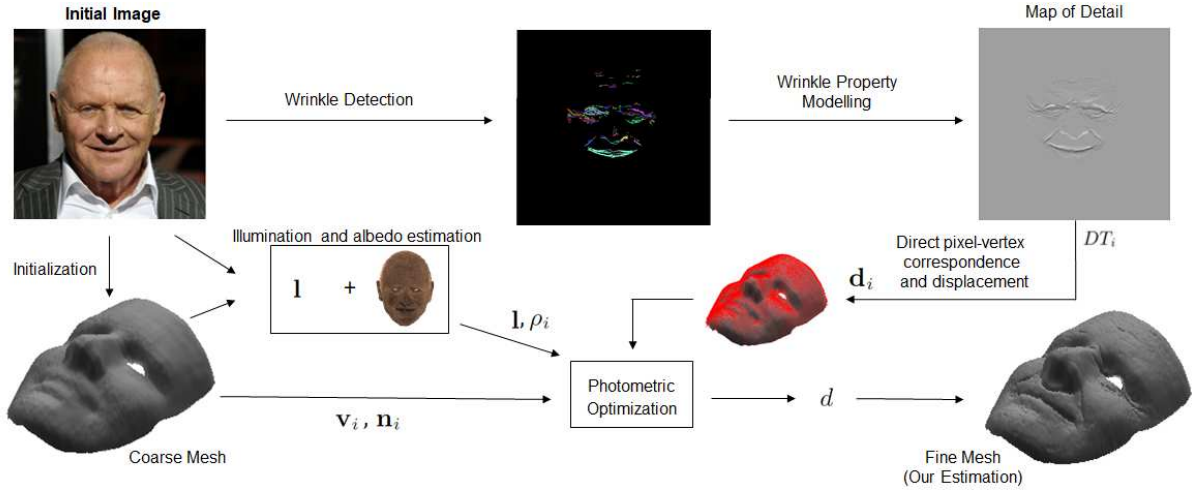

Figure 1: **An overview of our fine-detail reconstruction approach from an RGB image.** Our approach consists of two flows: 1) we use a deep-learning approach to obtain a 3D coarse mesh from the image. From this geometry can be extracted the 3D locations $\mathbf{v}_i$ and normals $\mathbf{n}_i$ for the i-th point. 2) Just considering the input image, we run a wrinkle detection algorithm based on image partial derivatives, and then group and model each of them. With this information, we can obtain a map of details DT_i to establish a direct correspondence between pixels in the image and vertices in the mesh, as well as the corresponding displacement vector $\mathbf{d}_i$. Finally, along with the illumination properties and albedo $(\mathbf{l}, \rho_i)$, we formulate the problem by means of a photometric optimization problem to recover the final detailed 3D mesh parametrized by the parameter d. As can be seen, our refining approach can recover detailed areas in 3D, in contrast to initializing approaches.

shape-from-shading techniques. In [RSOEK17], they need to train a different network to recover detail, which is extremely time and data consuming. Both of them fail to recover fine-scale detail. In [JBAT17] was presented a volumetric method as an alternative to direct fitted morphable models. However, the detail level obtained is still unsatisfactory.

Our main contribution is to develop a unified and unsupervised method that retrieves a detailed 3D mesh from a single image. It is worth noting that we do not need training data to code the wrinkle space, providing striking results even for complex shapes.

3 OUR APPROACH

In this paper, we propose a novel formulation to recover detailed 3D human faces from a single RGB image. It is worth mentioning that our approach is also capable of refining solutions given by coarse deformable face models, allowing us retrieving fine details such as wrinkles, furrows, and folds. To accomplish this goal, we define the 3D human face as an irregular and non-rigid surface that may vary due to age or facial expressions. Based on those variations, we define a parametric method that allows coding the skin wrinkles over a smooth and undetailed initialization.

To this end, we find inspiration on previous works [BKN02, LXC+15], where wrinkles were automatically modeled by means of given parameters. Nevertheless, our formulation is more general being capable of extracting from the input image all the parameters required to define the parametric model, except the depth of the furrows, which can be estimated by solving an energy minimization problem. This means no further information about the 3D face (or its corresponding projection in the image) is required. Even though it seems a simple model, due to the adaptability over different scenarios, it works properly over almost all the wrinkle types we can find in real images. An overview of our approach is displayed in Fig. 1. As we will introduce later, a map of detail is computed from the image partial derivatives, and then it is applied over the mesh. It is fast to compute even working locally on all the wrinkles, and as it can be seen in the experimental section, it provides satisfactory results outperforming current state-of-the-art techniques.

3.1 Initial 3D Face from Single Image

To obtain an initial 3D face estimation from a single image, we propose to use the convolutional volumetric regression model depicted in [JBAT17]. Later, we adjust the output to our model which works on a triangulated mesh by retrieving the surface of the volume. Since in the CNN resultant cropped image the detail is almost imperceptible, we align the mesh with the original image using the facial features of both images extracted with [ABS16]. We also estimate the scene lighting field in terms of spherical harmonics [BJ03].

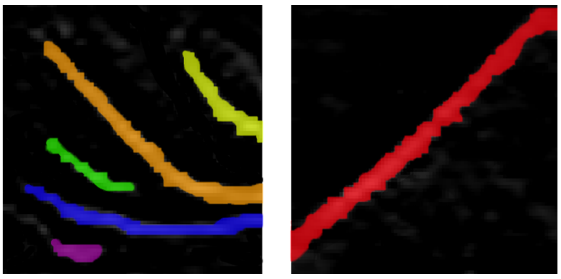

Figure 2: **Wrinkle detection.** For a specific image, we detect wrinkles across the image by clustering pixels in images partial derivatives. We obtain connected regions with an area greater than a threshold. Each wrinkle is accurately labeled and reconstructed separately since not all the wrinkles yield the same parameterization. **Left:** We observe a detection and clustering of multiple eye lines. Each color corresponds to a different wrinkle. **Right:** An example of a single nasolabial wrinkle is displayed, where small areas are ignored.

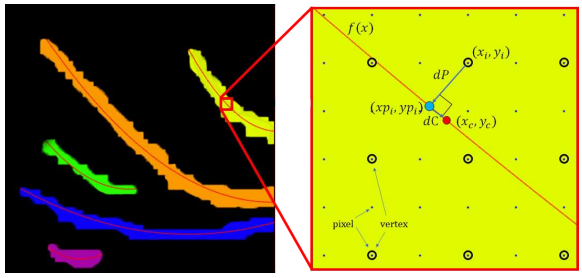

Figure 3: **Curve fitting and distance metrics. Left:** We can appreciate how the curves fit the given data. **Right:** Zooming view where we can observe the image pixel locations represented by blue dots, and these which have a corresponding mesh point by a black circle. The blue circle represents the perpendicular distance dP between a point (x_i^k, y_i^k) and $f(x)^k$, the red circle indicates the center of the wrinkle, and dC is the distance between the projected blue point (xp_i^k, yp_i^k) and the central point (x_c^k, y_c^k).

Assuming Lambertian reflectance, we can consider an initial albedo to be the mean skin color over the face. The reflection model to the i-th point can be written as $\mathbf{I}_i = \rho_i \cdot (\mathbf{l} \cdot \mathbf{H}(\mathbf{n}_i))$, where $\mathbf{I}_i$ represents the image value, ρ_i is the albedo, $\mathbf{l}$ is a 1×9 vector to represent the spherical harmonic coefficients, $\mathbf{H}$ is a 9×1 vector to indicate the spherical harmonic basis, and $\mathbf{n}_i$ is the surface normal. The process to estimate $\mathbf{l}$ and ρ_i may be iterated until the variance is sufficiently low. This model is simple and fast, but it has a limitation, does not accomplish well among cast shadows.

3.2 Wrinkle Properties

Modeling the wrinkles of a human face through a parametric formulation is a relevant area in 3D facial animation [BKN02, WWY06], with applications on face acquisition in real time [CBZB15, LWYZ17]. However, in order to produce very realistic results, these approaches normally require a large number of parameters to be known, such as the wrinkle properties as well as the material (skin) behavior. As a common practice, the location of the wrinkle furrow is represented by parametric models, such as the cubic Bezier curve [BKN02]. Unfortunately, previous formulations require these parameters to be known, or they have to be defined by the user in advance, limiting its applicability in real scenarios. To solve this limitation, we present a simple and intuitive method to retrieve the detailed furrows that better adapt over the image plane. Our method is able to extract from the image all those parameters, without the need for any training data at all. In addition, our approach can work on different mesh resolutions, and recover from large- to fine-scale details. It is suitable to describe fine-scale furrows since they can be adjusted with only three points. As a limitation, a furrow with less than three associated points cannot be retrieved.

Localization and clustering of wrinkle pixels. First of all, we set the images partial derivatives in both x- and y-directions. Both directions are used to determine regions of the image where changes in texture or geometry occur, i.e., allowing us to detect the wrinkles too. To avoid miss-detection from noisy measurements, we select the absolute value of magnitudes which are higher than a specific threshold (0.25 in our experiments). Once the initial candidates are detected, we analyze the connectivity between selected pixels, determining so different regions where all the pixels are connected (some examples are displayed in Fig. 2). We consider K sets of 8-connected pixels where each has been previously detected a wrinkle at pixel-level. We group all the pixels in a determined wrinkle k as $\mathbf{I}(x_i^k, y_i^k)$.

Curve Fitting and Distance Metrics. We now describe how these clustered points can affect to the geometric relief. Following classical curve fitting methods for data analysis, we introduce a second-order polynomial that better fits every curve. Note that for partial derivative with a major component in x, we swap axis since a well-defined function associates one, and only one, output to any particular input. So for each region k, we define a function $f(x)^k$ that better fits the data, such that:

$$f(x)^k = a^k x^2 + b^k x + c^k \,, \tag{1}$$

where the tuple (a^k, b^k, c^k) have to be estimated for every curve k. To define our parametric model over the defined furrow in a region, we propose to use the location of pixels (x_i^k, y_i^k) in the image, the curve defined by $f(x)^k$, and its corresponding centroid (x_c^k, y_c^k) such as:

$$(x_c^k, y_c^k) = \left(\frac{1}{R} \sum_i x_i^k, \frac{1}{R} \sum_i y_i^k \right) \quad (2)$$

where R represents the number of pixels in the k-th region.

Wrinkle Modeling. We next exploit the information available from the image and the previously extracted data to model a wrinkle as a furrow or relief. For every group k, let us define dP as the perpendicular distance between the point (x_i^k, y_i^k) and $f(x)^k$. In the same manner, we also introduce dC as the distance between a projected point (xp_i^k, yp_i^k) and the corresponding central point (x_c^k, y_c^k) projection on $f(x)^k$ (both distances are represented in Fig. 3).

Considering previous definitions, we now assume that the maximum influence of the depth parameter is where dP becomes null, with the global maximum at the central point (x_c^k, y_c^k) projection, then it loses influence while distances dP and dC increase.

To obtain the point (xp_i^k, yp_i^k), we first find the tangent to $f(x)^k$ at point x_i^k as $T_s = 2a^k x + b^k$. After that, we find the negative reciprocal slope, with is the normal slope $N_s = -1/T_s$, and the normal line to $f(x)^k$. Finally, we compute the two possible solutions of the second-order polynomial and choose the one with smaller distance to the point (x_i^k, y_i^k). On summary, both locations can be computed as:

$$xp_i^k = \frac{-b^k + N_s \pm \sqrt{(b^k - N_s)^2 - 4a^k(c^k + N_s x_i^k - y_i^k)}}{2a^k},$$
$$yp_i^k = a^k xp_i^2 + b^k xp_i^k + c^k .$$

We also introduce the height h of the wrinkle, and it can be defined by the maximum distance between any point (x_i^k, y_i^k) and the centroid (x_c^k, y_c^k) projection on $f(x)^k$. Additionally, it is worth noting that the width of the wrinkle we denote as w is the maximum distance that a point can have with $f(x)^k$ in the perpendicular direction.

According to previous definitions, we can model the indentation/relief of every point belonging to the region according to the parameters dP, dC, w, and h. So the effect of the furrow on the center is maximum when both dP and dC are zero and soften as the distances get larger. We now define a detail map DT which is a map with the same size as the image at full resolution and contain information about how the skin should wrinkle at each position. Its value for every i-th pixel in the region is defined as:

$$DT_i = \left(1 - \frac{dC_i}{h}\right) \cdot \left(-exp\left(\frac{-dP_i}{w}\right)\right) . \quad (3)$$

Note that this map gives information on how the skin should wrinkle along with the image depending on the distances dP and dC, that provides a smooth map of simulated wrinkles. Thus, the smoothness property of the resultant detailed shape is guaranteed since the pixels that are close also have a similar value. However, meshes with a very poor resolution may need a smoothing step to adjust the result to the mesh geometry, since the resultant detail may be too sharp. Finally, we normalize DT between -1 and 1. Then, while the wrinkle sunk in the lowest values of DT, the remaining pixels on the outer part (with larger values in DT), slightly lift to produce a more realistic effect. The values not referring to any wrinkle are set as zero.

The transference from the map of detail to the mesh is done by direct pixel-vertex correspondence. Each vertex of the mesh $\mathbf{v}_i$ is displaced along its normal direction $\mathbf{n}_i$ according to the value of DT_i in order to determine the corresponding displacement $\mathbf{d}_i$ defined as:

$$\mathbf{d}_i = \mathbf{n}_i \cdot d \cdot DT_i , \quad (4)$$

where $d \cdot DT_i$ is a scalar representing the displacement magnitude in the direction of the normal vector.

3.3 Depth Estimation via Energy Minimization

In order to retrieve the 3D reconstruction from a single image, we propose to minimize a photometric loss function. As DT is normalized between -1 and 1, d determines the exact scale of the shift. To automatically find the d parameter, we minimize the following photometric energy:

$$\mathscr{E}(d) = \arg\min_d \sum_{i=1}^{I} \|\mathbf{I}_i - \rho_i \cdot (\mathbf{l} \cdot \mathbf{H}(\mathbf{n}_i(\mathbf{v}_i + \mathbf{d}_i)))\|_2^2$$
$$\text{subject to } \mathbf{d}_i = \mathbf{n}_i \cdot d \cdot DT_i \quad (5)$$

where the displacement vector $\mathbf{d}_i$, is a function of the parameter d we have to estimate. $\mathbf{n}_i(\mathbf{v}_i + \mathbf{d}_i)$ refers to the normal of the surface with the estimated displacement. This optimization is resolved using the Matlab least-squares solver.

4 EXPERIMENTAL EVALUATION

We next present quantitative and qualitative results of our method on a wide range of subjects with different gender, ethnics, age, and facial gestures. Unfortunately, we cannot directly evaluate our approach due to a one-to-one correspondence between 3D ground truth, and the corresponding estimation is not available in practice. To solve this, and to provide a quantitative evaluation, we propose to align both 3D meshes and then computing an error between closest points on the meshes by following the normal direction. The error ε_{3D} can then be defined as $\varepsilon_{3D} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{v}(i)' - \mathbf{v}_{gt}(i)\|$, where

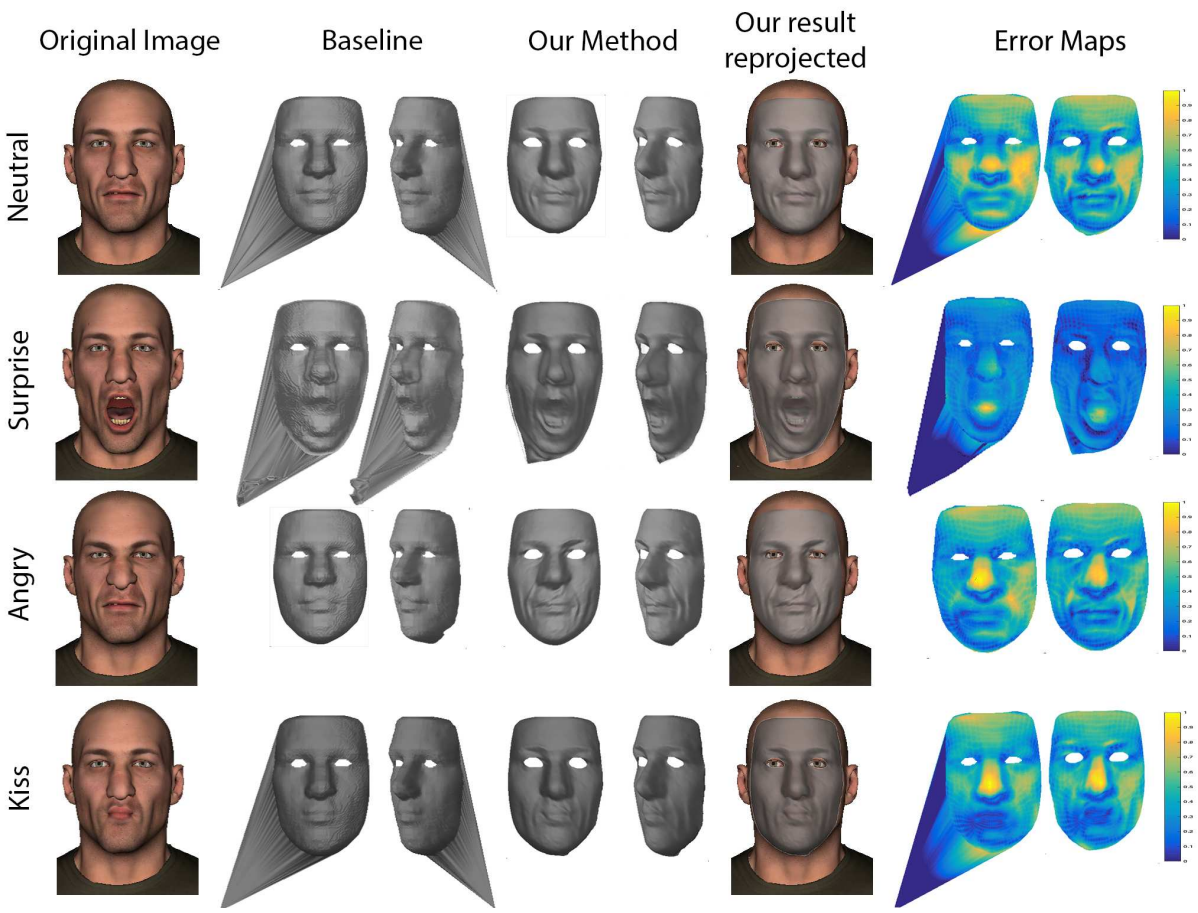

Figure 4: **Detailed 3D face reconstruction from a single image on synthetic data.** We represent a qualitative evaluation and comparison of four different facial expressions, one per row. **First column:** Rendered synthetic image we use as input. **Second and Third columns:** Frontal and side views of the 3D face reconstruction we obtain by using [JBAT17]. Particularly, we use this estimation as an initialization. **Fourth and Fifth columns:** We display the same estimations after applying our formulation, which is also reprojected over the original image in the **Sixth column**. **Seventh and Eighth columns:** A vertex error map is represented between the baseline [JBAT17] and our estimation with respect to the 3D ground truth, respectively. As can be seen, our approach can recover a larger amount of fine details in 3D. Best viewed in color.

$\mathbf{v}(i)'$ represents our 3D estimation of the i-th point and $\mathbf{v}_{gt}(i)$ is the corresponding ground truth, respectively. To compute the error, we follow the process depicted in the literature [RLMNA18] to find correspondences between aligned meshes with different elements. For further details, we refer the reader to this paper. Finally, we also establish a comparison with respect to state-of-the-art approaches.

4.1 Synthetic Images

Firstly, we evaluate our approach on synthetic data by using the Victor dataset [GVWT13], which includes several 3D meshes on a wide variety of facial gestures: pain, angry, sing, smile, kiss, sad, or surprise to name a few. This means this dataset provides plenty of medium-level wrinkles and strong cheek and nasolabial lines. It also includes tiny details around the eyes, however, it is very subtle and cannot be easily captured with the given image resolution. We propose to use this dataset to present a quantitative evaluation since the 3D ground truth is available for every mesh. As we previously introduced, to initialize our approach we obtain an initial 3D face by applying the CNN-based baseline [JBAT17]. As it was also claimed, this solution cannot recover properly the high-frequency geometry, producing smooth solutions where most of the detailed shapes are missed. However, this solution is accurate enough for initialization. Thanks to our formulation we can model every wrinkle in the shape, obtaining more detailed and realistic solutions.

We numerically evaluate our approach on eight facial expressions and establish a comparison with respect to the baseline [JBAT17]. These results are summarized in Table 1. As it is shown, our approach outperforms current solutions in terms of 3D accuracy. Unfortunately, the numerical improvement in the geometric er-

Expre.		Neutral	Sad	Angry	Pain	Surprise	Kiss	Happy	Sing	Average
[JBAT17]	$\varepsilon_{3Dtotal}$	0.022	0.040	0.039	0.037	0.064	0.050	0.055	0.059	0.046
	$\varepsilon_{3Dwrink.}$	0.808	0.811	0.814	0.814	0.613	0.820	0.811	0.813	0.788
	$\mathscr{E}(d)_{total}$	0.393	0.518	0.261	0.494	0.259	0.422	0.240	0.133	0.340
	$\mathscr{E}(d)_{wrink.}$	0.157	0.155	0.160	0.154	0.171	0.154	0.163	0.149	0.158
Ours	$\varepsilon_{3Dtotal}$	**0.021**	**0.034**	**0.036**	**0.036**	**0.042**	**0.044**	**0.051**	**0.054**	**0.039**
	$\varepsilon_{3Dwrink.}$	**0.791**	**0.797**	**0.801**	**0.801**	**0.602**	**0.801**	**0.797**	**0.799**	**0.774**
	$\mathscr{E}(d)_{total}$	0.389	0.508	0.253	0.486	0.254	0.414	0.233	0.126	0.332
	$\mathscr{E}(d)_{wrink.}$	0.125	0.123	0.127	0.128	0.151	0.122	0.134	0.136	0.131
Ours	nW	21	28	27	23	27	22	22	19	23.6
	$t(s)$	4.996	2.573	2.570	2.436	7.282	7.306	2.593	2.491	4.031

Table 1: **Quantitative evaluation and comparison on synthetic images.** The table reports the 3D error ε_{3D} together with the energy value $\mathscr{E}(d)$ in Eq. (5) for the baseline [JBAT17] and for our approach. Both metrics are computed toward the full set of points and only on the adjusted vertices, denoted by *wrink*. We also add the number of detected wrinkles nW, and the computation time $t(s)$ in seconds for our approach.

Input		Img1	Img2	Img3	Img4	Img5	Img6
[JBAT17]	$\mathscr{E}(d)_{total}$	0.2093	0.2543	2.2219	0.2969	0.3298	0.3067
	$\mathscr{E}(d)_{wrink.}$	0.1615	0.1810	0.2344	0.2109	0.3089	0.2130
Ours	$\mathscr{E}(d)_{total}$	**0.2070**	**0.2505**	**2.2103**	**0.2929**	**0.3290**	**0.2968**
	$\mathscr{E}(d)_{wrink.}$	**0.1570**	**0.1729**	**0.2285**	**0.2002**	**0.3037**	**0.2101**
Ours	nW	307	16	21	73	124	18
	$t(s)$	7.844	5.3001	5.1339	6.8080	7.6930	4.6443

Table 2: **Quantitative evaluation on real images.** The table reports the photometric energy error $\mathscr{E}(d)$ (see Eq. (5)) for the baseline [JBAT17] and for our approach toward the full set of points and uniquely over the affected vertices to properly visualize the influence of wrinkles. As in the previous analysis, we also add to our approach the number of detected wrinkles nW and the computation time $t(s)$ in seconds, respectively. Images are denoted as Img1, Img2, and Img3, and they correspond to the first column in Fig. 5. In the same manner, Img4, Img5 and Img6 correspond to the second column of the cited figure.

ror is not much striking, since the distance the vertices are shifted is short, we can observe a greater difference if we only interpret the same metric on the displaced vertices. However, the differences between both estimations can be observed in a qualitative manner. Figure 4 represents a qualitative evaluation and comparison of some evaluated facial expressions between our estimation and the baseline [JBAT17]. As can be seen, our approach is able to recover the cheek lines successfully and some other details around the facial geometry, which is not recovered properly by current methods. Particularly, it can be seen how the CNN-based solution fails to retrieve medium and fine details in some expressions (see first and third rows in Fig. 4), even producing large artifacts by placing points at one of the image corners (such as neutral, surprise and kiss expressions in the same figure). All these artifacts can be solved by our approach in an efficient and effective manner.

We also provide in Table 1 the number of wrinkles detected by our algorithm. We consider this information can also be an indirect metric to evaluate the performance of our approach since a priori a higher number of detected wrinkles should provide more detailed surfaces. However, this number should not be too high in order to avoid false wrinkles due to noisy measurements. Eventually, we include the time budget on a standard desktop computer Intel(R) Xenon(R) CPU ES-1620 v3 at 3.506GHz to solve wrinkle detection, modeling, and estimation. When the CPU permits a parallel computation, the wrinkles are estimated so, and its number does not affect drastically to the total computation budget. On balance, it is worth mentioning that all of this took just a few seconds on a commodity laptop (see the last row in Tables 1 and 2).

4.2 Real Images

We next evaluate our approach on several real images. To prove the generality of our approach we propose to process a set of images with a wide range of geometries, including subjects with different gender, age, ethnic group, and facial gesture. Since no ground truth is available for these images, we provide qualitative evaluation and comparison with the baseline [JBAT17]. For completeness, we also provide the energy value obtained by using both formulations on Table 2. We can notice the same error toward the overall vertices and exclusively on the displaced vertices. As we can appreciate, the relation between the two values establishes

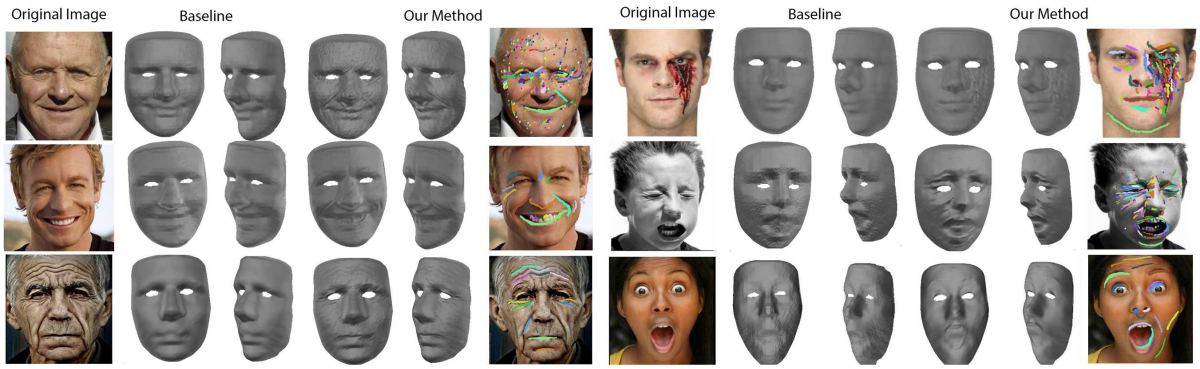

Figure 5: **Detailed 3D face reconstruction from a single real image**. Six different scenarios –varying age, gender, ethnic group, and facial expression– are displayed in rows. We represent two blocks, with the same information. **First column:** Input image. **Second and third column:** Camera and side views of the 3D reconstruction obtained by [JBAT17]. **Fourth and fifth column:** Camera and side views of our 3D reconstruction. **Sixth column:** Detected wrinkles. Best viewed in color.

Figure 6: **Detailed 3D reconstruction on real images**. Left and right display the same. **First row:** Four input images at different expressions. **From second to fourth row:** Reprojected mesh of the 3D reconstruction using [GVWT13], [JBAT17] and ours, respectively. Note that the solution provided by [GVWT13] includes a 200k-point, instead of using 20k points like us.

an indirect metric to measure the number of detections. For instance, in Img3 not all the wrinkles were properly captured due to noise (caused by filters added to the image for aesthetic purposes). The difference between values is high in comparison with other images. We can see how our method outperforms [JBAT17] toward all the metrics. Some examples from different viewpoints are displayed in Fig. 5. Again, it is worth pointing out that our method can obtain more accurate solutions than state of the art in terms of geometric details, as it can be observed in the before mentioned figure. Different types of wrinkles, as well as scars or blood marks, are satisfactorily recovered under different statuses (uncontrolled illumination, different resolution, and noise).

Next, we use four indoor images for two subjects taken from [GVWT13]. As before, we can only present a

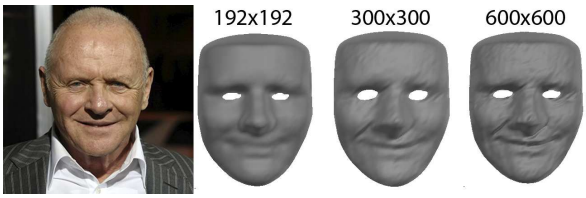

Figure 7: **3D reconstruction as a function of the image resolution.** To show the impact of the image resolution over the final result, we run our approach on different down-sampled images. While most of the details are not detected in low-resolution images, they become more accurately detected as the resolution in pictures increases.

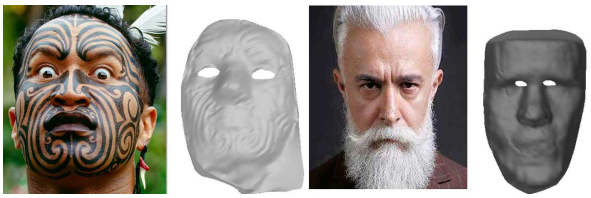

Figure 8: **Failure cases**. Two ambiguous examples our approach cannot solve properly. **Left:** A dark tattooed Maori where our algorithm fails in differentiating texture and shadow areas. **Right:** A big beard produces a self-occlusion in the face. Although the estimation is visually correct, it does not consider nor recover the human beard.

qualitative evaluation and comparison, including this time the results provided by [GVWT13]. We display the corresponding 3D reconstructions on Fig 6. Despite other techniques exploiting temporal priors and a higher resolution [GVWT13], our approach still seems to provide more detailed solutions than the remainder of the evaluated methods, even when reconstructing significantly fewer points.

In general terms, real images include more sophisticated details compared to synthetic data. Note that the design of realistic wrinkles is still a challenging problem in 3D modeling and animation. Therefore, the acquisition of realistic models from vision is essential to geometrically analyze these local details. In this context, the input image resolution is a key factor to obtain good results. As it is represented in Fig. 7, we observe a more fine acquisition when the image resolution increases.

Failure Cases. Since our formulation relies on image partial derivatives to detect wrinkles, some conditions such as shadows or texture-varying areas could produce ambiguous situations. For instance, obscure tattoos, strong cast shadows, and significant occlusions are some examples in which recovering detail becomes a hard task, and our method fails (observe some examples in Fig. 8). In the case of facial tattoos, or strong cast shadows, the algorithm proceeds to reconstruct the wrinkles where the geometry is not varying.

5 CONCLUSION

In this paper, we have proposed an intuitive and effective approach to retrieve detailed 3D reconstruction of faces from a single RGB image. Considering only the input image, our approach can obtain several features to parametrize the wrinkles without having been observed previously. This scheme allows us to model even person-specific attributes, such as scars or several shapes as a consequence of aging. Additionally, our approach is efficient since only need few seconds to solve the problem, by sorting out a photometric optimization problem. We have extensively evaluated our approach on both synthetic and real images, considering a wide range of variability in which we have outperformed existing state-of-the-art solutions. An interesting avenue for future research is to extend our formulation to handle more severe occlusions as well as a validation in real time at frame-rate.

Acknowledgments: This work has been partially supported by the Spanish Ministry of Science and Innovation under projects FireDMMI TIN2014-56919-C3-2-R, BOSSS TIN2017-89723-P, and HuMoUR TIN2017-90086-R; by the CSIC project R3OBJ 201850I099, and by the Spanish State Research Agency through the María de Maeztu Seal of Excellence to IRI MDM-2016-0656.

6 REFERENCES

[ABS16] B. Amos, L. Bartosz, and M. Satyanarayanan. OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, 2016.

[AMAC17] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video. *JMIV*, 57(1):75–98, 2017.

[AMCMN16] A. Agudo, J. M. M. Montiel, B. Calvo, and F. Moreno-Noguer. Mode-shape interpretation: Re-thinking modal space for recovering deformable shapes. In *WACV*, 2016.

[AMN17] A. Agudo and F. Moreno-Noguer. Combining local-physical and global-statistical models for sequential deformable shape from motion. *IJCV*, 122(2):371–387, 2017.

[AMN18] A. Agudo and F. Moreno-Noguer. Force-based representation for non-rigid shape and elastic model estimation. *TPAMI*, 40(9):2137–2150, 2018.

[AMN19] A. Agudo and F. Moreno-Noguer. Shape basis interpretation for monocular deformable 3D reconstruction. *TMM*, 21(4):821–834, 2019.

[BJ03] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *TPAMI*, 25(2):218–233, 2003.

[BKN02] Y. Bando, T. Kuratate, and T. Nishita. A simple method for modeling wrinkles on human skin. In *CG&A*, 2002.

[BV99] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999.

[CBZB15] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *TOG*, 34(4):46, 2015.

[CCC+18] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, and Jingyi Yu. Sparse photometric 3D face reconstruction guided by morphable models. In *CVPR*, 2018.

[CCL18] N. Chinaev, A. Chigorin, and I. Laptev. Mobileface: 3D face reconstruction with efficient cnn regression. In *ECCV*, 2018.

[CWZ+14] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014.

[GFT+11] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *TOG*, 30(6), 2011.

[GMMB00] E. Guillou, D. Meneveaux, E. Maisel, and K. Bouatouch. Using vanishing points for camera calibration and coarse 3D reconstruction from a single image. *VC*, 16(7):396–410, 2000.

[GSSM15] P. F. U. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3D reconstruction. In *ICCV*, 2015.

[GVWT13] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *TOG*, 32(6):158–1, 2013.

[GZC+16] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *TOG*, 35(3):28, 2016.

[HHT+16] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3D morphable face model and fitting framework. In *VISIGRAPP*, 2016.

[JBAT17] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *ICCV*, 2017.

[JZD+18] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3D face reconstruction with geometry details from a single image. *TIP*, 27(10):4756–4770, 2018.

[LWYZ17] S. Liu, Z. Wang, X. Yang, and J. Zhang. Realtime dynamic 3D facial reconstruction for monocular video in-the-wild. In *CVPR*, 2017.

[LXC+15] J. Li, W. Xu, Z. Cheng, K. Xu, and R. Klein. Lightweight wrinkle synthesis for 3D facial modeling and animation. *Computer-Aided Design*, 58:117–122, 2015.

[LZZL16] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, 2016.

[RLMNA18] G. Rotger, F. Lumbreras, F. Moreno-Noguer, and A. Agudo. 2D-to-3D facial expression transfer. In *ICPR*, 2018.

[RSK16] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016.

[RSOEK17] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.

[RV05] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, 2005.

[SKSS14] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, 2014.

[SWTC14] F. Shi, H.T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *TOG*, 33(6):222, 2014.

[TTHM+18] A. Tuan Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *CVPR*, 2018.

[VWB+12] L. Valgaerts, C. Wu, A. Bruhn, H.P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *TOG*, 31(6):187–1, 2012.

[WBS01] M. Wilczkowiak, E. Boyer, and P. Sturm. Camera calibration and 3D reconstruction from single images using parallelepipeds. In *ICCV*, 2001.

[WWY06] Y. Wang, C.C.L. Wang, and M.M.F. Yuen. Fast energy-based surface wrinkle modeling. *Computers & Graphics*, 30(1):111–125, 2006.

[ZNI+14] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an RGB-D camera. *TOG*, 33(4):156, 2014.

[ZTG+18] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum*, 37(2), 2018.