Combining Dense Nonrigid Structure from Motion and 3D Morphable Models for Monocular 4D Face Reconstruction

Mohammad Rami Koujan
University of Exeter
Exeter, United Kingdom
mk538@exeter.ac.uk

Anastasios Roussos
University of Exeter[1], Imperial College London[2]
Exeter, United Kingdom[1]
London, United Kingdon[2]
a.roussos@exeter.ac.uk

ABSTRACT

Monocular 4D face reconstruction is a challenging problem, especially in the case that the input video is captured under unconstrained conditions, i.e. "in the wild". The majority of the state-of-the-art approaches build upon 3D Morphable Modelling (3DMM), which has been proven to be more robust than model-free approaches such as Shape from Shading (SfS) or Structure from Motion (SfM). While offering visually plausible shape reconstruction results that resemble real faces, 3DMMs adhere to the model space learned from exemplar faces during the training phase, often yielding facial reconstructions that are excessively smooth and look too similar even across captured faces with completely different facial characteristics. This is due to the fact that 3DMMs are typically used as hard constraints on the reconstructed 3D shape. To overcome these limitations, in this paper we propose to combine 3DMMs with Dense Nonrigid Structure from Motion (DNSM), which is much less robust but has the potential of reconstructing fine details and capturing the subject-specific facial characteristics of every input. We effectively combine the best of both worlds by introducing a novel dense variational framework, which we solve efficiently by designing a convex optimisation strategy. In contrast to previous methods, we incorporate 3DMM as a soft constraint, penalizing both departure of reconstructed faces from the 3DMM subspace and variation of the identity component of the 3DMM over different frames of the input video. As demonstrated in qualitative and quantitative experiments, our method is robust, accurately estimates the 3D facial shape over time and outperforms other state-of-the-art methods of 4D face reconstruction.

KEYWORDS

3D morphable models, Structure from motion, Face reconstruction, Monocular videos, 3D faces, 4D reconstruction.

1 INTRODUCTION

Monocular 4D face reconstruction is the problem of recovering the 3D facial geometry in every frame of an input face video. It has attracted increased attention by the scientific community, especially during the last years, see e.g. [6, 20, 24, 29, 30, 40–43]. It has a plethora of applications, ranging from marker-less performance capture and augmented reality, to facial expression recognition for human-computer interaction.

Solving this problem under unconstrained conditions (commonly referred to as "in the wild") is particularly challenging and can be considered as an open problem of Computer Vision. This can be attributed to the fact that estimating the varying 3D facial shape over time is a highly ill-posed problem that is impossible to solve without incorporating priors. There are two broad categories of approaches in using priors for monocular 4D face reconstruction:

Face-specific, model-based priors, as e.g. used in [6, 24, 42]. In this case, 3DMMs are usually used to constrain the 3D facial shape in a low-dimensional subspace. In this way, the facial shape is represented by a relatively small number of parameters. These priors are very strong, making the reconstructions robust to challenging conditions, such as occlusions, large pose variations and low-resolution input. On the other hand, the reconstructions are typically overly smooth and do not capture high-frequency details of the 3D shape, due to the low dimensionality of the considered 3DMM subspaces. In addition, the reconstructed faces often resemble a generic face rather than the real input face, especially in cases of in-the-wild videos. This is because this type of reconstruction methods have to heavily rely on the 3DMM prior as the only way to compensate for the challenges of the input.

Generic, model-free priors, as e.g. used in [15, 18, 31, 46]. In this case, generic priors on the shape and dynamics of the captured object are used. These are applicable on any object and not only on faces. However, it is worth mentioning that most of the results that are typically shown in the relevant papers are on face videos. These methods do not use on a model of shape variation and do not require any training data. The reconstructions in this case are data-driven and solely rely on the observed input, using very generic constraints, such as temporal consistency and piece-wise smoothness of the recovered shape. In this way, they can recover a dynamic 3D geometry that is very characteristic to the specific input. On the other hand, these methods require very specific acquisition conditions to yield accurate reconstructions: Since these methods heavily rely on the observed input, the videos should be captured under controlled conditions, avoiding e.g. excessive occlusions or low resolution input. In addition, it is required that there is substantial temporal variation on the relative 3D pose between the camera and the captured object, so that it is possible to resolve in a fully data-driven way the geometric ambiguities that are related to the camera projection. When this kind of acquisition conditions are not met, these methods fail, yielding highly inaccurate reconstructions.

In this paper, we combine the best of both worlds (approaches based on model-based and model-free priors) by introducing a novel dense variational framework. Our framework combines model-free multi-frame optical flow, dense non-rigid structure from motion and 3D Morphable Model fitting. In more detail, we extend the dense variational formulation of [18] by adding face-specific priors. In contrast to previous 3DMM-based methods, the priors are incorporated as soft constraints, allowing deviations from the 3DMM subspace, so that the solution can capture facial shapes that cannot be represented

by the face model. With the proposed framework, we achieve dense 4D reconstructions that not only are robust to in-the-wild conditions but also include fine details and facial shape and dynamics that are specific to the captured face.

2 RELATED WORK

3D reconstruction of objects commonly found in images has played a significant role in a wide range of computer vision applications, such as object detection and recognition, scene interpretation and understanding, human-machine interaction, etc. Human faces are an archetype of those objects with ever-increasing interest for their potential impact and crucial applications.

By nature, the task of reconstructing the 3D geometry of human faces appearing in videos or images is rather problematic due to its ill-posed characteristics, with several associated ambiguities. Recently, many solutions have been presented for tackling this problem, incorporating myriad of priors and imposing different constraints.

Shape from Shading techniques, such as [2, 28, 36–39, 45], rely on simplified lighting and illumination models, with some other face-specific priors, to aid the reconstruction process. Those methods are prone to the in-the-wild conditions encountered in most real-world videos, being attributed to the oversimplified assumptions about light propagation models that fail to simulate real world scenarios.

After its first introduction by Blanz & Vetter [4], 3D Morphable Models (3DMM) have been used extensively in the literature with several additions [1, 6, 17, 24, 30, 33, 34, 34, 42, 44]. With the very recent framework in [6], it is even feasible to fit the 3DMM to in-the-wild images and videos, reconstructing both facial geometry and texture.

Dense Structure from Motion techniques approach this problem distinctly [18, 19]. Mainly, they incorporate geometric constraints to perform the reconstruction task. However, they are commonly criticised for the complicated and time-consuming frameworks they propose, mostly due to the infamous high-dimensionality curse. Additionally, the optical flow estimations required as an input usually for such methods ought to be accurately tracked among frames for producing satisfactory results.

Deep neural networks are another way of approaching this problem with ever-increasing interest for their promising results [9–11, 16, 26, 27].

There have been also some attempts to combine more than one of the aforementioned schemes [21, 22, 25], gaining the advantages of each. As opposed to our framework, most of those techniques, when combining 3DMM with other methods, impose 3DMM as a hard constraint, limiting their capacity to capture fine-scale details. Others only deal with rigid face deformations, unlike our method which assumes non-rigid deformations in the input videos.

3 PROPOSED FRAMEWORK

The method we put forward in this paper (henceforth referred to as DSfM-3DMM) benefits from two combined schemes, namely: 1) multi-frame subspace flow, where motion flow field is estimated from the input frames starting from a reference image, and 2) a 3D Morphable Model (3DMM) that plays a key role in the initialisation and final energy formulation of the entire framework.

Fig.1 demonstrates the different stages adopted in our DSfM-3DMM approach for doing the dense 3D reconstruction and tracking task. After parsing the input video into a sequence of frames, two steps are carried out concurrently: 1) following the paper of R. Garg et al. [19], dense optical flow is computed from a reference frame, not necessarily the first frame, to each of the other frames in the input sequence (section 3.3), 2) a Large Scale Facial Model (LSFM) [5, 7], the largest-scale 3D morphable model of facial identity learned from around 10,000 scans of different individuals, is used to provide a rigid estimation of the human face captured in the input video, with the aid of 68 facial landmarks (section 3.1), as well as an estimation of the camera pose parameters. Next, a correspondence is established in our approach between the rigid estimation, represented in the 3DMM space, and the dense 2D tracks extracted with [19] (section 3.5). In the final step, we aim to minimise an energy function we formulate in section 3.6, so that we can densely reconstruct in 3D and track the subject's face appearing in the input monocular video.

3.1 3D Morphable Models (3DMM)

3D Morphable Models (3DMM) were first introduced in the seminal work of Blanz and Vetter [4] as a linear point distribution parametric model for 3D representation of human faces accompanied by a fitting framework to surfaces and 2D images. Under such a model, an instance 3D face shape, say $\mathbf{x} \in \mathbb{R}^{3N}$, in the vectorized form ($\mathbf{x} = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^T$) can be represented as:

$$\mathbf{x}(\mathbf{p}, \mathbf{q}) = \bar{\mathbf{x}} + \mathbf{U}_{id}\mathbf{p} + \mathbf{U}_{exp}\mathbf{q} \tag{1}$$

where $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is the mean shape vector for both identity and expression, $\mathbf{U}_{id} \in \mathbb{R}^{3N \times np}$ is the orthonormal basis with the np most representative/significant principal components out of $M-1$, M being number of training faces used while building the identity part of the 3DMM, $\mathbf{U}_{exp} \in \mathbb{R}^{3N \times nq}$ is the orthonormal basis with the nq most representative/significant principal components out of $T-1$, T being number of training faces used while building the expression part of the 3DMM, and $\mathbf{p} \in \mathbb{R}^{np}$, $\mathbf{q} \in \mathbb{R}^{nq}$ are the identity and expression parameters. Therefore, a 3DMM ($\mathbf{x}$) in this case is a function of both identity and expression coefficients ($\mathbf{x}(\mathbf{p}, \mathbf{q})$). For the identity part of the 3DMM, Large Scale Morphable Model (LSFM) [5, 7], which was built from approximately 10,000 scans of different people with varied demographic information, was adopted, while the blendshapes model of Facewarehouse [12] was used for the expression.

3.2 UV Mapping and Model Space Sub-Sampling

The objective (energy) functional formulated in this paper (as detailed later in section 3.6) has a term E_{reg} that works as an edge-preserving spatial regularizer. Such a term is defined based on an unwrapped version of the 3D shape S_f we need to estimated for each frame. The aim of the unwrapping is, therefore, to establish a one-to-one mapping $f(v) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ that maintains a 3D to 2D correspondence between each 3D face vertex (v_i) and a 2D point on an image grid. Towards that goal and following [8], we choose to perform an optimal cylindrical unwrapping of the 3DMM mean face ($\bar{\mathbf{x}}$) in the UV space, a space in which the manifold of the face is flattened into adjoining 2D atlas. This results in a 3-channel image/UV-map ($\mathcal{U}$) such that $\mathcal{U}(c_i) = v_i$, where c_i is a point on this

unevenly distributed map. After placing the center of mass of the mean 3DMM face shape ($\bar{\mathbf{x}}$) at the origin, $c_i = [\theta_i, z_i']^T = f(v_i)$ is computed as follows:

$$\theta_i = arctan(\frac{x}{z}), \quad z_i' = y \tag{2}$$

where $v_i = [x, y, z]^T$ is the corresponding vertex on the 3DMM mean face $\bar{\mathbf{x}}$.

A key advantage of defining this bijective mapping is that neighbouring vertices on the mesh will be neighbours in UV space, resulting in a shared topology T, where $T = [\mathbf{t}_1^T, \mathbf{t}_2^T, ..., \mathbf{t}_m^T]$, $\mathbf{t}_i = [t_1^i, t_2^i, t_3^i]$, $t_j^i \in \{\mathbb{Z}^+ | t_j^i \leq N\}$, given that $\mathbf{t}_i$ is the ith triangle index and T is provided with the utilised 3DMM.

For the sake of making this step non video-specific, we define a fixed regular grid $\mathbf{G}$ overlaid on top of the UV map ($\mathcal{U}$) computed from $\bar{\mathbf{x}}$ and find its corresponding 3D face shape, call it $\bar{\mathbf{x}}_s$, by subsampling the 3DMM space. By adopting such an approach, the dependency of the UV map is decoupled from the input video, leading to a predetermined correspondence between $\mathbf{G}$ and any reconstructed 3D face $\mathbf{x}_s$ in the subsampled 3DMM.

With a step size: $D_z = rD_\theta$, the 2D grid $\mathbf{G}$ is pre-computed, given that r is the radius of the optimal unwrapping cylinder, computed as in [8] based on annotations, and D_θ is chosen in our paper so that 2D grid points density is comparable to the original 3DMM resolution N.

Since the computed UV map ($\mathcal{U}$) shares the same topology as the 3DMM mean face ($\bar{\mathbf{x}}$), 3D vertices corresponding to the 2D grid ($\mathbf{G}$) points are computed as the barycentric coordinates of the triangles defined in T and overlaid on top of grid in the UV space. This gives rise to a sub-sampled 3DMM mean face shape ($\bar{\mathbf{x}}_s \in \mathbb{R}^{3\times Q}$, Q being the sub-sampled face resolution) derived as follow:

$$\bar{\mathbf{x}}_s = [v_1, ..., v_Q] = \bar{\mathbf{x}}\mathbf{B} \tag{3}$$

$\mathbf{B} = [B_1, ..., B_Q] \in \mathbb{R}^{N\times Q}$ is the matrix storing the barycentric coordinates $B_i \in \mathbb{R}^N$ of a vertex v_i that corresponds to 2D grid point lying inside a triangle $\mathbf{t}_j = [t_1^j, t_2^j, t_3^j]$ in $\mathbf{G}$. B_i is actually a sparse vector (having at least $N-3$ zeros) with $B_i(t_1^j), B_i(t_2^j), B_i(t_3^j)$ embodying the barycentric coordinates of v_i. To adjust the incorporated 3DMM model in section 3.1 accordingly, we sample $\mathbf{U}_{id}$ and $\mathbf{U}_{exp}$ in the same manner based on $\mathbf{B}$, producing the following subsampled 3DMM:

$$\mathbf{x}_s = \bar{\mathbf{x}}_s + \mathbf{U^s}_{id}\mathbf{p} + \mathbf{U^s}_{exp}\mathbf{q} \tag{4}$$

where $\bar{\mathbf{x}}_s$, $\mathbf{U^s}_{id}$, $\mathbf{U^s}_{exp}$ denote the sub-sampled mean face, identity and expression bases, respectively. Henceforth, this sub-sampled model will be used and referred to in this paper, even though we omit the superscript $\mathbf{s}$ in equation 4 and consider $Q \approx N$. It is worth noting that after subsampling both $\mathbf{U}_{id}$ and $\mathbf{U}_{exp}$ we need to orthonormalise them again to be used in our E_{dmm} and E_{id} of the energy functional we formulate in equation 11. We choose to orthonormalise both bases as follows:

$$SVD(\mathbf{U^s}_{exp}\mathbf{U^s}^T_{exp}) = \tilde{\mathbf{U}}^{\mathbf{s}}_{exp}\Lambda_{exp}\tilde{\mathbf{U}}^{\mathbf{sT}}_{exp} \tag{5}$$

$$\tilde{\mathbf{U}}^{\mathbf{s}}_{id}\Lambda_{id}\tilde{\mathbf{U}}^{\mathbf{sT}}_{id} = SVD((\mathbf{I}_{3Q\times 3Q} - \tilde{\mathbf{U}}^{s}_{exp}\tilde{\mathbf{U}}^{\mathbf{sT}}_{exp})\mathbf{U^s}_{id}) \tag{6}$$

noting that $SVD(.)$ is the singular value decomposition operator, $\tilde{\mathbf{U}}^{\mathbf{s}}_{id}$, and $\tilde{\mathbf{U}}^{\mathbf{s}}_{exp}$ are the orthonormalised versions of the sub-sampled identity and expression bases, respectively, and Λ_{id}, Λ_{exp} are their corresponding eigenvalues .

3.3 Multi-Frame Subspace Flow

Starting from the observation that 2D trajectories of various points on the same non-rigid surface exhibits high degree correlation over time, R. Garg et al. put forward in [19] a procedure for the computation of optical flow from a reference frame to all other frames in a sequence. Their key remark is that the amount of correlation existing between moving points (pixels) on a non-rigid surface over time can be expressed in a compact form as a linear combination of a low-rank motion basis. This results in a subspace constraint acting as a spatial regularisation term, along with a brightness consistency term, in their formulated energy functional. The final outcome is a reduction in the notoriously high-dimensionality associated with this kind of problems, and a temporally smooth estimations.

While generating the motion field estimation results in our framework, we incorporate the gradient of pixel intensities, rather than their absolute values, in the brightness consistency term suggested in [19] for more robust performance, given our test videos are challenging (in-the-wild), of varying resolutions and affected by noise.

3.4 3D Rigid Initialisation Using 3DMM

In this paper, we opt for computing the rigid 3D face shape initialisation of our framework using 3DMMs. This is mainly due to the fact that relying on Rigid Structure from Motion (RSfM) techniques, as it is done usually in similar frameworks, for the rigid initialisation computation has significant shortcomings in some challenging scenarios: 1) human faces captured in the input video should exhibit enough rotation for the SfM to work robustly, 2) No significant occlusion can be present, 3) 2D facial features should be tracked quite accurately among input frames sequence. All the aforementioned issues can be encountered quite often in in-the-wild videos. Hence, a powerful alternative approach is to use 3DMM fitting on sparse facial landmarks, leading to plausible face reconstruction suitable as an initialisation and overcoming the previously stated barriers.

A two-step procedure was followed to produce the rigid 3D shape estimation. While in the first camera parameters are estimated, the second step tackles the calculation of identity and expression parameters ($\mathbf{p}, \mathbf{q}$), assumed in our rigid case to be fixed over all input frames sequence.

Camera Matrix Estimation For the sake of computations ease, we postulate in our experiments an orthographic camera model of the form:

$$\hat{v} = \rho\hat{\Pi} \times v \tag{7}$$

$v = [x, y, z]$ being an object-centered face vertex in the camera coordinate system and ρ is the scaling factor that accounts for global changes in depth d. Let $\hat{\mathbf{L}}$ be $2n_f \times L$ matrix storing the 2D landmarks of all the frames, where each column has the (x, y) coordinates of the same landmark in all the frames, $\hat{\Pi} = [\Pi_1^T ... \Pi_{nf}^T]^T$ a $2nf \times 3$ matrix that stacks the scaled orthographic projection ($\mathcal{P} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, mapping 3D points to image pixels) matrices $\Pi_f \in \mathbb{R}^{2\times 3}$ from all the frames f. The following least squares problem (LSM) was minimised

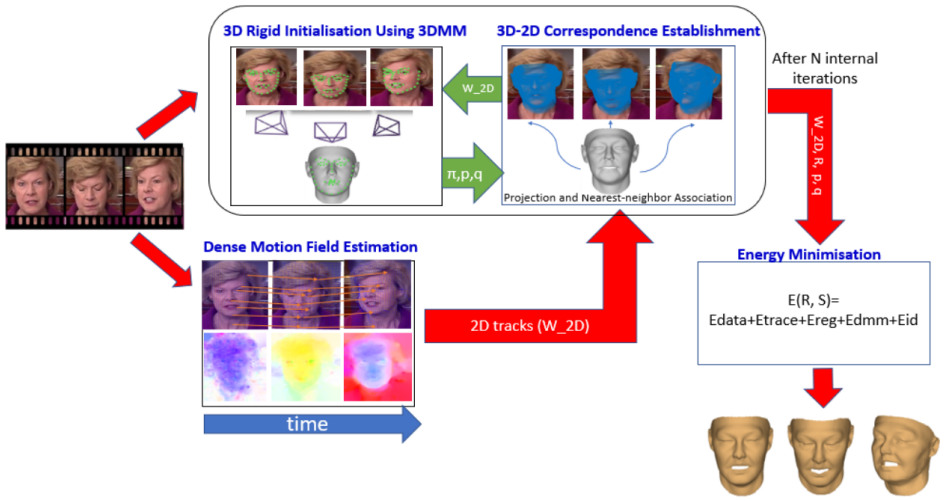

Figure 1: Our proposed method marked by the different adopted stages

to find the scaled orthographic camera projection matrix:

$$\min_{\hat{\Pi}} ||\hat{\Pi} \mathcal{S}_{rig} - \hat{\mathbf{L}}||_{\mathcal{F}}^2 \tag{8}$$

given that $\mathcal{S}_{rig}$ is a $3 \times L$ matrix with the sparse 3D landmarks on the mean 3DMM shape (denoted as $\bar{\mathbf{x}}$ in section 3.1), with landmarks stored column-wise, and $\mathcal{F}$ is the Frobenius norm.

Shape Parameters Estimation having obtained a per-frame estimation of the projection (camera) matrix, the following LSM was put together to compute the 3DMM parameters (identity and expression) of the input face:

$$\begin{aligned} \min_{\mathbf{p},\mathbf{q}} \sum_{f=1}^{nf} ||(I_L \otimes \Pi_f)(\bar{\mathbf{x}} + \mathbf{U}_{id}\mathbf{p} + \mathbf{U}_{exp}\mathbf{q}) - l_f||^2, \\ \textit{subject to} \quad -w_{id} \leq \mathbf{p} \leq w_{id} \quad \textit{and} \\ -w_{exp} \leq \mathbf{q} \leq w_{exp} \end{aligned} \tag{9}$$

where $\otimes$ denotes Kronecker product, such that the multiplication with the $2L \times 3L$ matrix $I_L \otimes \Pi_f$ implements the application of the camera projection Π_f on each one of the L landmarks, $l_f \in \mathbb{R}^{2L}$ is the 2D facial landmarks extracted from frame f, w_{id} and w_{exp} are the box constraints imposed on the identity and expression and have the same size as $\mathbf{p}$ and $\mathbf{q}$, respectively. The introduction of the box constraints (w_{id} and w_{exp}) in equation 9 is to compensate for the inaccuracies associated with the input landmarks relied on for fitting and the encountered occlusions. Note that the 3DMM parameters ($\mathbf{p}$ and $\mathbf{q}$) are constant over all the input frames f, complying with our postulated assumption about the rigidity of estimated face at this stage. The facial shape rigidity assumption throughout the whole video is rather tight. However, as verified experimentally in [47], once provided with a significant number of frames, it provides a very robust initialisation of the camera parameters even in cases of large facial deformation.

3.5 2D-3D Correspondence Establishment

Having obtained a sequence of 2D tracks ($W_{2D} \in \mathbb{R}^{2nf \times K}$, K being the number of dense $2D$ image points tracked between frames) lying on the subject's face (section 3.3) and a rigid initialisation ($\mathbf{x} = \bar{\mathbf{x}} + \mathbf{U}_{id}\mathbf{p} + \mathbf{U}_{exp}\mathbf{q}$) of its 3D shape (section 3.4), a correspondence between these 2D tracks and the rigid 3D shape vertices should be constructed.

Using the estimated camera matrix of the reference frame, say $\Pi_{f^\star}$, the rigid shape $\mathbf{x} \in \mathbb{R}^{3N}$ is projected first onto the reference frame $f^\star$, after being rearranged as a matrix $\mathbf{x}'_{f^\star} \in 3 \times N$ with (x, y, z) coordintes of each vertex placed in a column-wise order:

$$\mathcal{P}(\hat{\Pi}_{f^\star}, \mathbf{x}'_{f^\star}) = \hat{\mathbf{x}}_{f^\star} = \hat{\Pi}_{f^\star} \mathbf{x}'_{f^\star} \tag{10}$$

where $\mathcal{P}$: $\mathbb{R}^{3\times N} \rightarrow \mathbb{R}^{2\times N}$ is a linear view transformation mapping 3D to 2D points, $\Pi_{f^\star}$ is a 2×3 camera matrix of the reference frame $f^\star$, estimated in section 3.4, and $\hat{\mathbf{x}}_{f^\star} \in \mathbb{R}^{2\times N}$ is the projected vertices of the rigid 3D face on the reference frame $f^\star$. Let

$$\mathbf{w}_{f^\star} = \begin{bmatrix} x^1_{f^\star} & \dots & x^K_{f^\star} \\ y^1_{f^\star} & \dots & y^K_{f^\star} \end{bmatrix} \in \mathbb{R}^{2\times K}$$

represent the traced face dense 2D points in the reference frame $f^\star$, with $W_{2D} = [\mathbf{w}_1^T, ..., \mathbf{w}_{nf}^T]^T \in \mathbb{R}^{2nf\times K}$. A correspondence between $\mathbf{w}_{f^\star}$ and $\hat{\mathbf{x}}_{f^\star}$ is created by choosing the nearest neighbour (column), based on the euclidean distance, in $\mathbf{w}_{f^\star}$ for each vertex (column) in $\hat{\mathbf{x}}_{f^\star}$. This results in a matrix, say $\mathbf{w}^{nn}_{f^\star}$, of size $2 \times N$, where N is the resolution (number of vertices) of the utilised and sub-sampled 3DMM (see section 3.2). Since the dense point tracks are known with respect to the reference frame, the matrix W_{2D} is updated by adding the track of each point in $\mathbf{w}_{f^\star}$, resulting in a new matrix, call it W^{nn}_{2D}. Following an iterative approach, W^{nn}_{2D} is used again for refining the rigid estimation ($\hat{\Pi}$ and $\mathbf{x}$) obtained in section 3.4, but this time using all dense 2D tracks N, which have been put in correspondence with the rigid 3D estimation, rather than the 68

facial landmarks employed in the first iteration. Experimentally, we found that 2-3 iterations is enough for most of the tested videos.

3.6 Energy Formulation

Let $I_1, ..., I_F$ be the input video frame sequence to be densely reconstructed in 3D and tracked, F the number of frames, I_{ref} the reference frame, N the number of pixels tracked starting from the reference frame after establishing the correspondence with the utilised sub-sampled 3DMM as described in section 3.5. With the aim of 3D dense reconstruction and tracking from monocular videos in mind, we compose an objective function of the form:

$$\begin{aligned} E(\mathbf{R}, \mathbf{S}) = \lambda E_{data}(\mathbf{R}, \mathbf{S}) + E_{reg}(S) + \tau E_{trace}(\mathbf{S}) \\ + c_{dmm} E_{dmm}(\mathbf{S}) + c_{id} E_{id}(\mathbf{S}) \end{aligned} \tag{11}$$

Data term (E_{data}) a geometric data term that aims at minimising the reprojection of reconstructed shapes into input frames. This term takes the following quadratic form, with $||.||_{\mathcal{F}}$ denoting Frobenius norm:

$$E_{data} = \frac{1}{2}||\mathbf{W} - \mathbf{RS}||^2_{\mathcal{F}} \tag{12}$$

where $\mathbf{W}$ is the 2D tracks matrix of size $2F \times N$, storing N tracked points on I_{ref} throughout the input sequence (see W_{2D} section 3.5), $\mathbf{R}$ is $2F \times 3F$ reprojection matrix with diagonal 2×3 elements implementing the per-frame reprojection, $\mathbf{S}$ is $3F \times N$ matrix stacking vertically per-frame 3D shapes $\mathbf{S}_f$

$$\mathbf{S}_f = \begin{bmatrix} x_f^1 & ... & x_f^N \\ y_f^1 & ... & y_f^N \\ z_f^1 & ... & z_f^N \end{bmatrix}$$

with columns having the x, y, z coordinates of each 3D shape vertex.

Regularisation term (E_{reg}) an edge-preserving spatial regularization on the dense 3D trajectories that comprise the columns of $\mathbf{S}$. Consider $\mathbf{S}_f^i$ as the ith ($i = 1, 2, 3$) row of a frame f 3D shape $\mathbf{S}_f$. E_{reg} is defined as a total variation term:

$$E_{reg} = \sum_{f=1}^{F} \sum_{i=1}^{3} TV\{S_f^i\} = \sum_{f=1}^{F} \sum_{i=1}^{3} \sum_{p=1}^{N} ||\nabla \mathbf{S}_f^i(p)|| \tag{13}$$

with $\mathbf{S}_f^i$ representing a discrete 2D image of the same size as the 2D grid (mask) $\mathbf{G}$ defined in section 3.2 and $\nabla \mathbf{S}_f^i(p)$ denoting the gradient of $\mathbf{S}_f^i$ at pixel p. Since each vertex in $\mathbf{S}_f^i$ gives rise to a pixel on I_f, reshaping $\mathbf{S}_f^i$, based on $\mathbf{G}$, as a 2D discrete image holds valid and allows the computation of the gradient as forward differences in both horizontal and vertical directions (interested readers are referred to [13] for more details).

Trace term (E_{trace}) as the name implies, this term favours a smaller rank of the time-evolving shape matrix, minimising the number of principal components needed to represent such a shape over time.

$$E_{trace} = ||P(\mathbf{S})||_{\star} = \sum_{j=1}^{\min(F, 3N)} \Lambda_j \tag{14}$$

given that $||.||_{\star}$ is the nuclear norm and $P(\mathbf{S})$ is an $F \times 3N$ matrix, with row-wise per-frame shapes.

Additionally, we propose to add two new terms (E_{dmm} and E_{id}) that are face-specific and act as a soft constraint on: 1) departure of $\mathbf{S}_f$ from our 3DMM space, and 2) deviation from the mean identity in the input frame sequence, respectively.

Distance from 3DMM space (E_{dmm}) this term penalises the deviation of reconstructed shapes from the 3D Morphable Model space. It is formulated as a quadratic cost between the per-frame reconstructed shapes and their projection onto the subspace spanned by the 3DMM.

$$E_{dmm} = \frac{1}{2} \sum_{f=1}^{F} c_{dmm} ||(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T)(\mathbf{S}_f - \bar{\mathbf{x}})||^2 \tag{15}$$

given that $\bar{\mathbf{x}}$ is the mean 3DMM shape of size $3N$, $\mathbf{I}$ is a $3N \times 3N$ identity matrix, $\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_{id} \tilde{\mathbf{U}}_{exp}]$ is a $3N \times (np + nq)$ combined basis comprised of the orthonormalised and subsampled version of Large Scale Morphable Model (LSFM) [5, 7] and Facewarehouse [12], respectively, as explained in section 3.2. $np + nq$ symbolises the principal components kept for explaining the identity and expression of reconstructed faces in the 3DMM subsapce.

Identity unification (E_{id}) with the objective of consolidating identity of reconstructed 3D shapes ($\mathbf{S}_f$) throughout the input sequence, this term focuses on keeping the projection of each obtained 3D shape on the identity basis of our 3DMM subspace as close as possible to the mean projections onto the same basis over all frames. Mathematically, this term is put together as below:

$$E_{id} = \frac{1}{2} c_{id} \sum_{f=1}^{F} ||\tilde{\mathbf{U}}_{id}^T (\mathbf{S}_f - \bar{\mathbf{x}}) - \bar{d}_{id}||^2 \tag{16}$$

$$\bar{d}_{id} = \frac{1}{F} \sum_{f=1}^{F} \tilde{\mathbf{U}}_{id}^T (\mathbf{S}_f - \bar{\mathbf{x}}) \tag{17}$$

3.6.1 *Optimisation of the Formulated Energy.*

To minimise the proposed energy functional in equation (11), we adopt a similar minimisation procedure to the one suggested in [18]. We alternate between the estimation of the motion matrix $\mathbf{R}$ and the shape matrix $\mathbf{S}$, maintaining the other unalterable.

Minimising equation (11) w.r.t $\mathbf{R}$ while fixing $\mathbf{S}$ is fairly straightforward, boiling down to minimising the only dependent term (E_{data}) using Levenberg-Marquardt. On the other hand, in a second step, estimating $\mathbf{S}$ while keeping $\mathbf{R}$ constant is a non-trivial task. Basically, the problem can be divided into two sub-problems, with the aim of decoupling the nuclear norm and TV regularisation terms of the energy, as demonstrated in equations (18) and (19).

$$\begin{aligned} &\min_{\mathbf{S}} \frac{1}{2\theta} ||\mathbf{S} - \bar{\mathbf{S}}||^2_{\mathcal{F}} + \frac{\lambda}{2} ||\mathbf{W} - \mathbf{RS}||^2_{\mathcal{F}} + \sum_{f,i,p} ||\nabla \mathbf{S}_f^i(p)|| + \\ &\frac{c_{dmm}}{2} \sum_{f=1}^{F} ||(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T)(\mathbf{S}_f - \bar{\mathbf{x}})||^2 + \\ &\frac{c_{id}}{2} \sum_{f=1}^{F} ||\tilde{\mathbf{U}}_{id}^T (\mathbf{S}_f - \bar{\mathbf{x}}) - \bar{d}_{id}||^2 \end{aligned} \tag{18}$$

$$\min_{\mathbf{S}} \frac{1}{2\theta} ||\mathbf{S} - \bar{\mathbf{S}}||^2_{\mathcal{F}} + \tau ||P(\bar{\mathbf{S}})||_{\star} \tag{19}$$

where θ has the role of a quadratic relaxation parameter that is relatively small so that the optimal $\mathbf{S}$ and $\bar{\mathbf{S}}$ are similar. Although equation (18) is convex, it is non-differentiable due to the presence of the

edge-preserving spatial regularisation term (E_{reg}). Circumventing such a problem can be achieved by dualising the regularization term in (18) and rewriting the corresponding minimisation in its primal-dual form. Algorithm 1 summarises the approach for minimising (18).

Algorithmus 1 : Primal dual algorithm for Eq. (18)

Input : Measurement matrix $\mathbf{W}$, current motion matrix estimates $\mathbf{R}$ and low rank shapes $\bar{\mathbf{S}}$

Output : Spatial smooth shapes $\mathbf{S}$

Parameters : λ, θ, and step size σ of dual update

Initialise : the dual variable q using the estimates from the previous iteration of this algorithm (0 in the first)

while *not converge* **do**

$$\mathbf{D}_q = \begin{bmatrix} \nabla^\star q_1^1(1) & .. & q_1^1(N) \\ . & .. & . \\ . & .. & . \\ \nabla^\star q_F^3(1) & .. & \nabla^\star q_F^3(N) \end{bmatrix}$$

for $f = 1$ **to** F **do**

$$\mathbf{S}_{f(3N\times 1)} = (\lambda\hat{\mathbf{R}}^T\hat{\mathbf{R}} + \tfrac{1}{\theta}\mathbf{I}_{3N\times 3N} + c_{id}\mathbf{U}_{id}\mathbf{U}_{id}^T + c_{dmm}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\mathbf{T}})^{-1}(\lambda\hat{\mathbf{R}}^T\mathbf{W}_{f(2N\times 1)} + \frac{\bar{\mathbf{S}}_f}{\theta} + c_{dmm}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\mathbf{T}}\bar{\mathbf{x}} + c_{id}\mathbf{U}_{id}\mathbf{U}_{id}^T\bar{\mathbf{x}} + c_{id}\mathbf{U}_{id}\bar{\mathbf{d}}_{id} - \mathbf{D}^{\mathbf{f}}_{\mathbf{q}(3\times\mathbf{N})});$$

for $f = 1$ **to** F, $i = 1$ **to** 3, $p = 1$, **to** N **do**

$$\mathbf{q}_f^i(p) = \frac{\mathbf{q}_f^i(p)+\sigma\nabla S_f^i(p)}{\max(1, ||\mathbf{q}_f^i(p)+\sigma\nabla S_f^i(p)||)};$$

In Algorithm (1), we chose to decouple overall 3D shapes estimation into per-frame independent problem, so that it is feasible to solve it in parallel using GPU (Graphics Processing Unit). Note that, in Algorithm 1, $\hat{\mathbf{R}} = (\mathbf{R}_{f(2\times 3)} \otimes I_{N\times N})$ is a $2N$ by $3N$ matrix which implements the orthographic projection of $\mathbf{S}_{f(3N\times 1)}$ onto the corresponding frame resulting in $\mathbf{W}_{f(2N\times 1)}$, given that N is the resolution of the subsampled 3DMM, which is also equivalent to the number of tracked points in each input frame after establishing the correspondence as explained in section 3.5.

To minimise equation 19, we use the soft impute algorithm, see [32] for more details.

4 EXPERIMENTAL RESULTS

In this section, we present the results obtained while conducting both quantitative and qualitative experiments to evaluate our method against other state-of-the-art methods, namely: 4Dface [23], 3DMMedges [3], DV-NRSfM [18]. We use the code provided by the authors for the methods we compare against without any modifications. As stated before, our 3DMM model of choice is a combination of the Large Scale Facial Model (LSFM), which is made up of 10,000 faces of both sexes and varying ages [5, 7] (for the identity part), and Facewarehouse [12], which is composed of 150 individuals aged 7-80 from various ethnic backgrounds, for the expression. The LSFM and Facewarehouse models were registered using Nonrigid ICP [14] algorithm. Motion field estimation was generated in our framework using the code provided by the authors of [19], but with incorporating gradient information computed from input frames sequence in addition to the direct intensity values.

4.1 Quantitative Results

To quantitatively evaluate our presented method, we generated two synthetic videos each consisting of 440 frames and exhibiting various natural expressions and head pose variations. Those two videos were acquired from high-resolution face scans generated by a DI4DTM face scanner, with the (virtual) camera undergoing a periodic rotation. Such videos facilitate the quantitative evaluation of the 4D face reconstructions for every tested frame. Fig. 3 shows 3 selected frames with different poses and expressions from each of the synthetic videos we produce. The original size of each frame including the black background is 512×512 pixels.

To evaluate our per-frame reconstructed 3D faces against the ones produced by 4Dface [23], 3DMMedges [3], and DV-NRSfM [18], we calculate a per-frame error representing the average per-vertex discrepancy between the recovered mesh and the corresponding ground truth. While generating the results, the same 68 facial landmarks were made available as an input for all the tested methods, including ours. In addition, all generated faces were aligned with the ground truth meshes before calculating any quantitative comparative measures. Fig. 4 demonstrates the cumulative error across all frames with four different methods. Our method (termed as DSfM-3DMM) outperforms the other three methods in both videos, followed by 3DMMedges, 4Dface, and DV-NSfM, respectively. The performance of DV-NSfM is the worst, since it struggles to reconstruct faces in the lack of proper camera rotation around the synthesised faces in the videos. On the other hand, our initialisation proves its robustness in such a challenging scenario.

4.2 Effect of Edmm and Eid Terms

To appraise the decisive role of incorporating E_{dmm} and E_{id} terms in the final energy equation 11, we generated around 1000 synthetic frames with 2 different subjects, a male and a female, exhibiting various facial expressions while the synthetic camera is rotating around them incrementally, left to right. We tested our method, along with some variants of which, on those frames and computed the per-vertex error for each reconstructed frame. The final result is demonstrated in Fig.6 as a cumulative error across all frames reconstructed using:

(1) $DSFM - 3DMM$: our proposed method in this paper, with all the energy terms in equation 11.
(2) $DSFM - E_{dmm}$ our proposed method without the E_{id} term in equation 11.
(3) $DSFM$: equation 11 without E_{dmm} and E_{id}.
(4) $3DMM$: classical 3DMM fitting using sparse landmarks only, see section 3.4

Analysing Fig. 6 reveals that, as argued earlier, using 3D Morphable Models (3DMM) alone limits the reconstruction results in terms of capturing the fine scale details, vindicated in the figure with the smallest area under the corresponding curve. Using a dense nonrigid structure from motion (DSFM) produces more accurate reconstructions compared to using 3DMMs alone, as can be seen in Fig. 6. Combining DSfM with 3DMM fitting approaches leads to better reconstructions as we claim in this paper, which can be justified in the results visualised in Fig. 6. The combination of E_{dmm} and E_{id} proves fruitful when compared against relying on only the distance from the learned 3DMM manifold (E_{dmm} term).

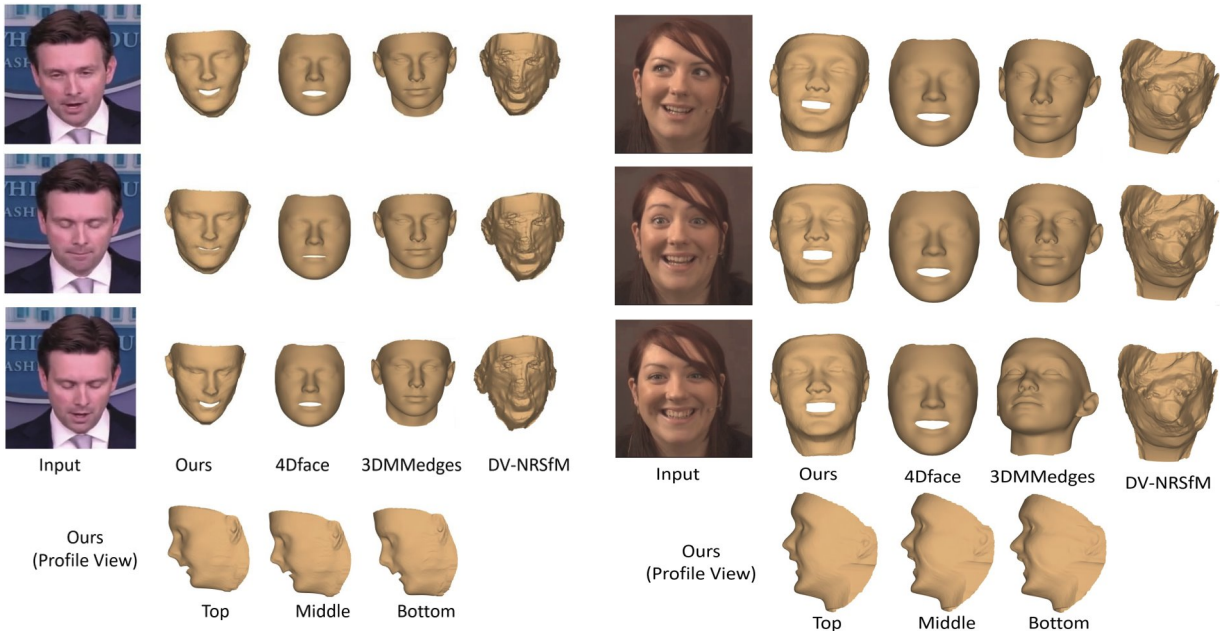

Figure 2: Reconstruction results generated from our method, 3DMMedges [3], 4Dface [23], and DV-NRSfM [18] on a male and female in-the-wild videos with 80 and 100 interocular resolution, respectively. Only 3 selected frames with their results are shown from each video.

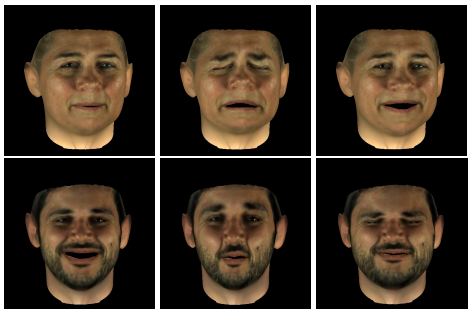

Figure 3: Three frames, showing dissimilar facial expressions, from the two synthetic videos we generate in this paper. Top row: video 1, bottom row: video 2. These are used for quantitative evaluation.

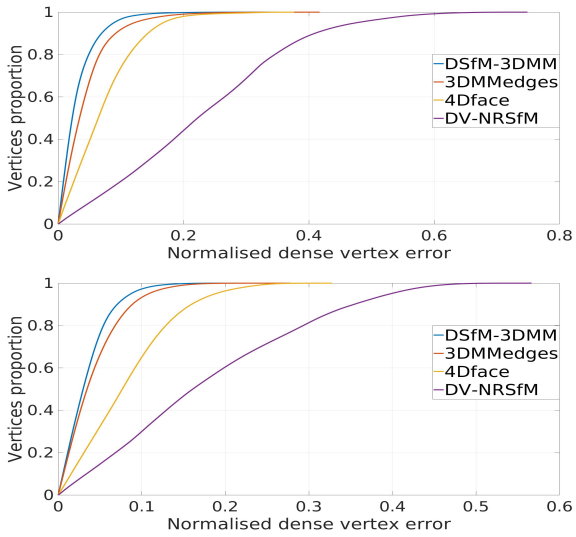

Figure 4: Quantitative evaluation of the compared methods on synthetic video 1 (top) and 2 (bottom).

4.3 Qualitative Results

For the qualitative evaluation, in-the-wild videos from the 300VW [35] dataset were selected. The aforementioned dataset is characterised by challenging videos accompanied with noise, occlusions, and low resolution, rendering the process of reconstruction very demanding. Fig. 2 reveals a qualitative comparison between the reconstructed faces from two test videos by our scheme (DSfM-3DMM), 3DMMedges [3], 4Dface [23], and DV-NRSfM [18]. Looking at Fig. 2 in more details, some apparent trends can be noticed. First, DV-NRSfM produces the worst results with several noticeable deformations, which can be attributed to the fact that both videos have low-resolution and lack enough camera rotation for this method. Shapes obtained by our method look more similar to the actual subjects and are marked by person-specific characteristics, e.g. the nose shape and eye closure (middle and bottom frames) in the male video, and the rise of the eyebrows in the female video (middle frame). 3DMMedges and 4Dface methods generate faces of less similarity to the captured subjects compared to ours. 3DMMedges method fails

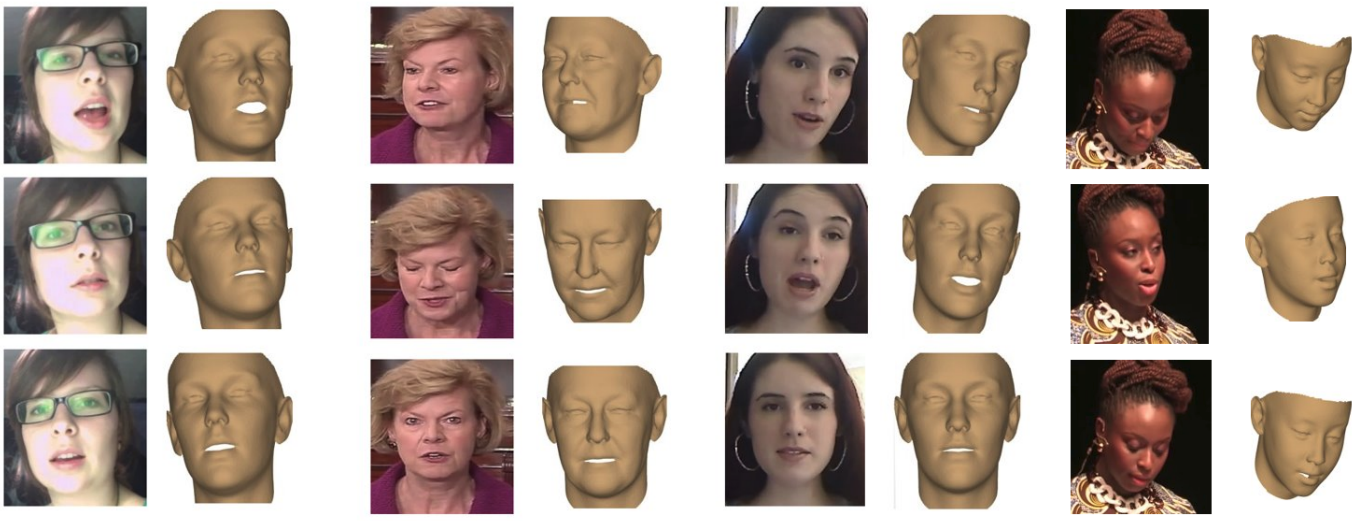

Figure 5: Reconstruction results produced by our method on four dissimilar videos concatenated column-wise along with the generated 3D faces. Each row depicts a frame from one video and to its right is the corresponding reconstructed 3D face

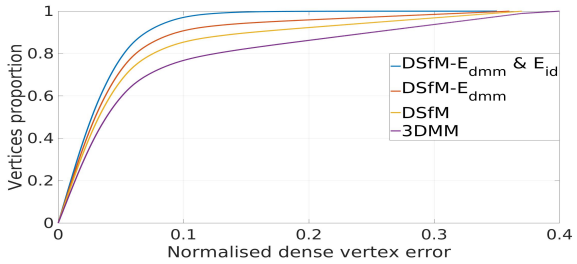

Figure 6: Quantitative evaluation of our method against some of its variants, produced by deleting some terms of Eq. 11, on 1000 synthetic frames.

sometimes completely in estimating the correct subject's pose, e.g bottom frame of the female-video. Fig. 5 presents the reconstruction results obtained by our proposed method when applied on some videos from in-the-wild 300VW [35] dataset.

5 CONCLUSION

In this paper, we propose a solution for the problem of 4D face reconstruction and tracking from monocular videos. Our suggested framework capitalises on both Dense Nonrigid Structure from Motion (DNSfM) and 3D Morphable Models (3DMM). The result is a more robust and accurate methodology when dealing with challenging (in-the-wild) videos that have low-resolution and lack proper camera rotation around the subject's face, which affects considerably the DSfM when used alone. At the same time, this combination produces 3D shapes that have somewhat the freedom to depart from the 3DMM space and capture details that cannot be expressed by the incorporated 3DMM. We have validated the potential of our proposed approach both quantitatively, using a set of synthetic videos we generated, and qualitatively, on the 300VW [35] dataset for in-the-wild videos, and outperformed other state-of-the-art methods tested on the same videos. The effect of adding the two energy terms (E_{dmm} and E_{id}) acting as soft constraints on the 3DMM manifold were evaluated separately on a synthetic set of videos and shown to offer a rewarding combination.

REFERENCES

[1] Brian Amberg. 2011. *Editing faces in videos*. Ph.D. Dissertation. University_of_Basel.

[2] Jonathan T Barron and Jitendra Malik. 2015. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2015), 1670–1687.

[3] Anil Bas, William AP Smith, Timo Bolkart, and Stefanie Wuhrer. 2016. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*. Springer, 377–391.

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

[5] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. 2018. Large scale 3d morphable models. *International Journal of Computer Vision* 126, 2-4 (2018), 233–254.

[6] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Poumpis, Yannis Panagakis, and Stefanos P Zafeiriou. 2018. 3D Reconstruction of" In-the-Wild" Faces in Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. 2016. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5543–5552.

[8] James Booth and Stefanos Zafeiriou. 2014. Optimal uv spaces for facial morphable model construction. In *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 4672–4676.

[9] Adrian Bulat and Georgios Tzimiropoulos. 2017. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 1. 4.

[10] Adrian Bulat and Georgios Tzimiropoulos. 2017. Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. *arXiv preprint arXiv:1712.02765* (2017).

[11] Adrian Bulat and Georgios Tzimiropoulos. 2018. Hierarchical binary CNNs for landmark localization with limited resources. *arXiv preprint arXiv:1808.04803* (2018).

[12] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions*

on Visualization and Computer Graphics 20, 3 (2014), 413–425.

[13] Antonin Chambolle and Thomas Pock. 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision* 40, 1 (2011), 120–145.

[14] Shiyang Cheng, Ioannis Marras, Stefanos Zafeiriou, and Maja Pantic. 2017. Statistical non-rigid ICP algorithm and its application to 3D face alignment. *Image and Vision Computing* 58 (2017), 3–12.

[15] Yuchao Dai, Hongdong Li, and Mingyi He. 2014. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision* 107, 2 (2014), 101–122.

[16] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. 2018. Cascade Multi-View Hourglass Model for Robust 3D Face Alignment. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 399–403.

[17] Nathan Faggian, Andrew Paplinski, and Jamie Sherrah. 2008. 3D morphable model fitting from multiple views. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 1–6.

[18] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. 2013. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 1272–1279.

[19] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. 2013. A variational approach to video registration with subspace constraints. *International journal of computer vision* 104, 3 (2013), 286–314.

[20] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 28.

[21] Tomoya Hara, Hiroyuki Kubo, Akinobu Maejima, and Shigeo Morishima. 2012. Fast-accurate 3d face model generation using a single video camera. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 1269–1272.

[22] Matthias Hernandez, Tal Hassner, Jongmoo Choi, and Gerard Medioni. 2017. Accurate 3D face reconstruction via prior constrained structure from motion. *Computers & Graphics* 66 (2017), 14–22.

[23] Patrik Huber, William Christmas, Adrian Hilton, Josef Kittler, and Matthias Rätsch. 2016. Real-time 3D face super-resolution from monocular in-the-wild videos. In *ACM SIGGRAPH 2016 Posters*. ACM, 67.

[24] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. 2016. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[25] Patrik Huber, Philipp Kopp, Matthias Rätsch, William Christmas, and Josef Kittler. 2016. 3D face tracking and texture fusion in the wild. *arXiv preprint arXiv:1605.06764* (2016).

[26] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1031–1039.

[27] Amin Jourabloo and Xiaoming Liu. 2016. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4188–4196.

[28] Ira Kemelmacher-Shlizerman. 2013. Internet based morphable model. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 3256–3263.

[29] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. 2017. Inversefacenet: Deep single-shot inverse face rendering from a single image. *arXiv preprint arXiv:1703.10956* (2017).

[30] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 194.

[31] Qi Liu-Yin, Rui Yu, Lourdes Agapito, Andrew Fitzgibbon, and Chris Russell. 2017. Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. *arXiv preprint arXiv:1708.01654* (2017).

[32] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *journal of machine learning research* 11, Aug (2010), 2287–2322.

[33] Sami Romdhani and Thomas Vetter. 2003. Efficient, Robust and Accurate Fitting of a 3D Morphable Model.. In *ICCV*, Vol. 3. 59–66.

[34] Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 2. IEEE, 986–993.

[35] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 50–58.

[36] William AP Smith and Edwin R Hancock. 2006. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 12 (2006), 1914–1930.

[37] William AP Smith and Edwin R Hancock. 2008. Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics. *International Journal of Computer Vision* 76, 1 (2008), 71–91.

[38] Patrick Snape, Yannis Panagakis, Stefanos Zafeiriou, et al. 2015. Automatic construction Of robust spherical harmonic subspaces.. In *CVPR*. 91–100.

[39] Patrick Snape and Stefanos Zafeiriou. 2014. Kernel-pca analysis of surface normals for shape-from-shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1059–1066.

[40] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2017. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *arXiv preprint arXiv:1712.02859* (2017).

[41] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2. 5.

[42] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2387–2395.

[43] Anh Tuân Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. 2018. Extreme 3D face reconstruction: Seeing through occlusions. In *Proc. CVPR*.

[44] Luan Tran and Xiaoming Liu. 2018. Nonlinear 3D Face Morphable Model. *arXiv preprint arXiv:1804.03786* (2018).

[45] Philip L Worthington and Edwin R Hancock. 1999. New constraints on data-closeness and needle map consistency for shape-from-shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 12 (1999), 1250–1267.

[46] Rui Yu, Chris Russell, Neill Campbell, and Lourdes Agapito. 2015. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *IEEE International Conference on Computer Vision (ICCV 2015)*. University of Bath.

[47] Stefanos Zafeiriou, G Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. 2018. The 3d menpo facial landmark tracking challenge. (2018).