The 3D Menpo Facial Landmark Tracking Challenge

Stefanos Zafeiriou[*,1,2] Grigorios G. Chrysos[*,1] Anastasios Roussos[*,1,3]
Evangelos Ververas[1] Jiankang Deng[1]
George Trigeorgis[1]
[1] Department of Computing, Imperial College London, UK
[2] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[3] Department of Computer Science, University of Exeter, UK
{s.zafeiriou, g.chrysos}@imperial.ac.uk

Abstract

Recently, deformable face alignment is synonymous to the task of locating a set of 2D sparse landmarks in intensity images. Currently, discriminatively trained Deep Convolutional Neural Networks (DCNNs) are the state-of-the-art in the task of face alignment. DCNNs exploit large amount of high quality annotations that emerged the last few years. Nevertheless, the provided 2D annotations rarely capture the 3D structure of the face (this is especially evident in the facial boundary). That is, the annotations neither provide an estimate of the depth nor correspond to the 2D projections of the 3D facial structure. This paper summarises our efforts to develop (a) a very large database suitable to be used to train 3D face alignment algorithms in images captured "in-the-wild" and (b) to train and evaluate new methods for 3D face landmark tracking. Finally, we report the results of the first challenge in 3D face tracking "in-the-wild".

1. Introduction

Face alignment and tracking on images/videos captured under unconstrained recording conditions has recently received considerable attention due to the numerous applications such as entertainment, security, human computer interaction, graphics etc.

The current state-of-the-art in face alignment revolves around Deep Convolutional Neural Networks (DCNN) equipped with a multiscale structure, alleged Hourglass architecture [15] [1] or structures that combine a convolutional network for feature extractions and Recurrent Neural Networks (RNNs) for solving non-linear least square problems [19]. The landscape is not different in deformable face tracking, where DCNNs currently hold the state-of-the-art [19, 12] [2].

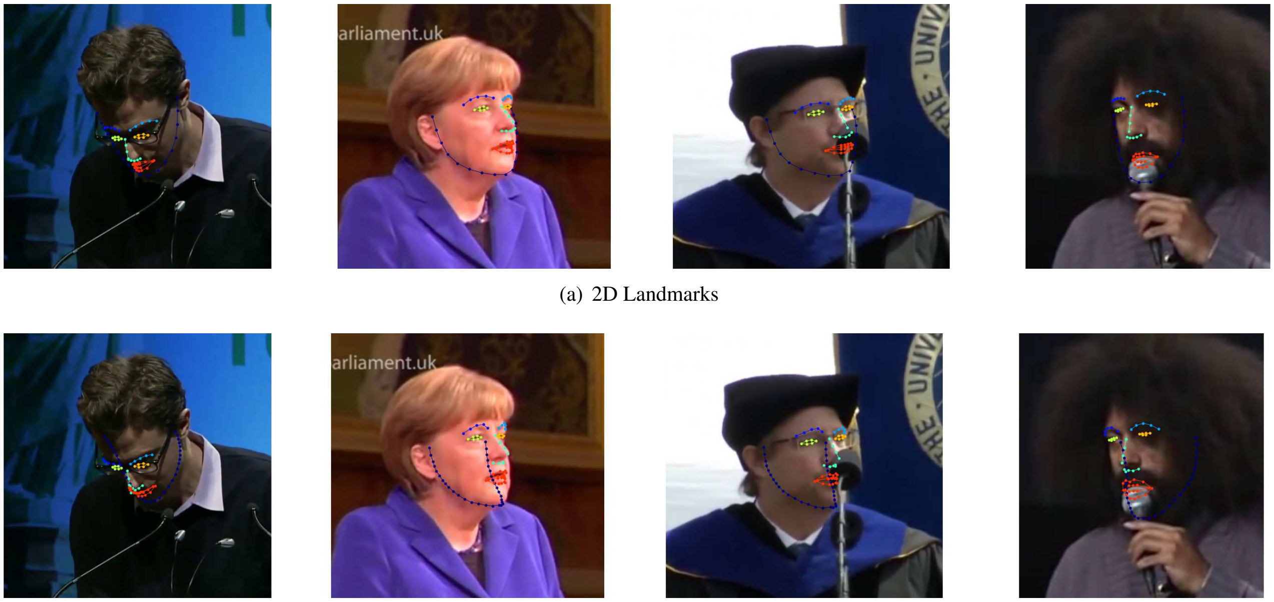
Currently, it is feasible to robustly train DCNNs for face alignment, since our group has provided large scale landmark annotations [17, 16, 18, 24]. In the first challenge, i.e. 300W [17], our group provided annotations for over 4,350 "in-the-wild" images (approximately 5,000 faces). In the 300VW [18] our group provided annotations for 114 videos, aiming at evaluating efforts for deformable face tracking. The 300W and 300VW benchmarks provided annotations with regards to a frontal face shape of 68 landmarks. A step forward was made in CVPR 2017 by our group in the so-called Menpo Challenge [24]. In Menpo challenge we provided annotations for over 12,000 faces including, for the first time, annotations for over 4,000 profile faces (with regards to to 39 landmarks). All the above benchmarks constitute a very valuable asset for 2D deformable face alignment and tracking and have used to drive the research in the field.

Even though all the above databases provide annotations that correspond to semantically meaningful parts of the face many of the landmarks hardly correspond to the 3D structure of the human face. That is, they do not accurately correspond to the projections, in the image plane, of any landmarks of the 3D facial structure. Furthermore, the 2D annotations of the above benchmarks do not bare any information regarding the depth of the 3D face. In this paper, we call the 2D projections of the 3D landmarks in the image plane as 3DA-2D landmarks to distinguish them from the 3D coordinates of the facial landmarks in the 3D scene, which we call 3D landmarks in this paper. An example of 2D landmark annotations provided by 300VW and the cor-

[*] S. Zafeiriou, G. Chrysos and A. Roussos contributed equally and have joint first authorship.

[1] Hourglass networks won the recent Menpo Challenge on multi-view face alignment [22] and the recent 3D face alignment competition [3].

[2] For state-of-the-art techniques the readers may refer to the recent comprehensive survey [7].



(a) 2D Landmarks

(b) 3DA-2D Landmarks

Figure 1. First row (a): The annotated 2D landmarks provided by the 300VW competition. Second row (b): The estimated 3DA-2D landmarks provided by Menpo 3D challenge.

responding 3DA-2D landmarks, estimated by the proposed procedure are shown in Fig. 1.

The major problem regarding extracting 3D and 3DA-2D landmark annotations in images captured "in-the-wild" is that: (a) the faithful reconstruction of the 3D facial surface remains very challenging in unconstrained recording conditions; (b) photo-realistic synthesis of face in arbitrary poses and illumination conditions is not possible without the facial albedo, which requires special setups in order to be precisely captured (e.g., a light stage [11]). This is why the first 3D landmark localisation challenge, which was organised in conjunction with ECCV 2016, used only data captured in controlled conditions (i.e., Multi-PIE [13]) or synthetically generated data using simple techniques (i.e., rendering a 3d face captured in controlled conditions using arbitrary backgrounds [14]).

In this paper, we make a significant step further and provide large scale 3DA-2D facial landmarks, as well as 3D facial landmarks in a normalised facial model space. These annotations can be used for training algorithms for estimating 3DA-2D, as well as 3D landmarks in "in-the-wild" images. We use these landmarks to train a DCNN based on the Hourglass architecture [12] for estimating 3DA-2D landmarks. The trained DCNN was used to provide a first estimate of the 3DA-2D locations of landmarks in facial videos. Then, an elaborate procedure combining Structure from Motion (SfM) and 3DMM fitting is used to convert these estimates to ground annotations which can be used for training and evaluating algorithms 3DA-2D and 3D facial landmark tracking algorithms. Finally, we used these data to evaluate efforts in 3D face tracking and present the results. All in all, our contributions in this paper are the following:

- We provide a large scale database of facial images with 3DA-2D and 3D facial landmarks by applying the state-of-the-art 3DMM fitting algorithm of [1] driven by the ground-truth 2D landmarks.
- We propose an elaborate procedure for estimating 3DA-2D and 3D landmarks in arbitrary "in-the-wild" videos. The procedure is highly accurate and was used to provide more than 280,000 annotated frames.
- We present the results of the first challenge on 3DA-2D and 3D landmark tracking.

2. Creating a Large Scale Database with 3DA-2D and 3D landmarks

Recently in [25] a facial 3DMM has been fitted on the 2D landmarks and used in order to train a DCNN for the estimation of the 3D facial surface. In order to produce a large scale dataset of 3DA-2D and 3D landmarks we utilised the recently introduced 3DMM fitting strategy which is applicable to "in-the-wild" images. The difference between the method used in [25] and the one used in this work is that our 3DMM fitting strategy not only uses the facial landmarks but the facial texture as well. Furthermore, in order to improve accuracy we annotated all the images with regards to (a) gender, (b) ethnicity and (d) apparent age and used the bespoke 3DMMs from the LSFM model [2]. We provide 3DA-2D and 3D landmarks for all the databases that we have annotated with 2D landmarks, i.e 300W, Menpo etc. It is worth noting that the 3D landmarks are provided in the normalised space of the model.

Fitting the 3DMM in hundreds of thousands of video frames is computationally expensive, we opted to train a DCNN, based on the hourglass architecture, that regresses to 3DA-2D landmarks. In particular, after the coarse step of the architecture of [12], it regresses to 3DA-2D landmarks (using as auxiliary input the 2D landmark locations) [3].

3. Creation of Ground Truth 3D Facial Landmarks on Videos

To extract accurate 3D landmarks from facial videos, a semi-automated procedure is followed as described below (the core steps are depicted in Figure 2). Initially, we employ the aforementioned DCNN network to estimate the per frame 3DA-2D landmarks. The automatic personalisation method of [8] was utilised for refining certain facial parts (i.e. the eyes). Sequentially, an energy minimisation method was used to fit our combined identity and expression models on the landmarks of all frames of the video simultaneously. We apply this fitting twice, first by using the global LSFM model for the identity variation and second by using the corresponding bespoke LSFM model, based on manual annotation of the demographics of the input face. Finally, we sample the dense facial mesh that is generated by the fitting result at every frame on the sparse landmark locations. Via visual inspection of both the dense 3D and the reprojected sparse 2D landmarks results in all frames, we choose the best of the two results (global versus bespoke identity models) and we retain it as ground truth only if the result is plausible in all frames.

3.1. Dense 3D Face Shape Modelling

Let us denote the 3D mesh (shape) of a face with N vertexes as a $3N \times 1$ vector

$$\mathbf{s} = \left[\mathbf{x}_1^{\mathsf{T}}, \ldots, \mathbf{x}_N^{\mathsf{T}}\right]^{\mathsf{T}} = \left[x_1, y_1, z_1, \ldots, x_N, y_N, z_N\right]^{\mathsf{T}} \quad (1)$$

where $\mathbf{x}_i = \left[x_i, y_i, z_i\right]^{\mathsf{T}}$ are the object-centered Cartesian coordinates of the i-th vertex.

In this work we unbundle the identity from the expression variation and then combine them to articulate the 3D facial shape of any identity. An identity shape model is considered first, i.e. a model of shape variation across different individuals, assuming that all shapes are under neutral expression. For this, we adopt our LSFM models [2], which consist the largest models of 3D Morphable Modelling (3DMM) of facial identity built from approximately 10,000 scans of different individuals[4]. The dataset that LSFM models are trained on includes rich demographic information about each subject, allowing the construction of not only a global 3DMM model but also *bespoke models* tailored for specific age, gender or ethnicity groups. In this work, we utilise both the global and the bespoke LSFM models.

Each LSFM model (global or bespoke) forms a shape subspace that allows the expression of any new mesh. To construct such an LSFM model initially a set of 3D training meshes are brought into dense correspondence so that each mesh is described with the same number of vertices and all samples have a shared semantic ordering. The rigid transformations are removed from these semantically similar meshes, $\{\mathbf{s}_i\}$, by applying Generalised Procrustes Analysis. Sequentially, Principal Component Analysis (PCA) is performed which results in $\{\bar{\mathbf{s}}_{id}, \mathbf{U}_{id}, \mathbf{\Sigma}_{id}\}$, where $\bar{\mathbf{s}}_{id} \in \mathbb{R}^{3N}$ is the mean shape vector, $\mathbf{U}_{id} \in \mathbb{R}^{3N \times n_p}$ is the orthonormal basis after keeping the first n_p principal components and $\mathbf{\Sigma}_{id} \in \mathbb{R}^{n_p \times n_p}$ is a diagonal matrix with the standard deviations of the corresponding principal components. Let $\widetilde{\mathbf{U}}_{id} = \mathbf{U}_{id}\mathbf{\Sigma}_{id}$ be the identity basis with basis vectors that have absorbed the standard deviation of the corresponding mode of variation so that the shape parameters $\mathbf{p} = \left[p_1, \ldots, p_{n_p}\right]^{\mathsf{T}}$ are normalised to have unit variance. Therefore, assuming normal prior distributions, we have $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n_p})$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

A 3D shape instance of a novel identity can be generated using this PCA model as a function of the parameters $\mathbf{p}$:

$$\mathcal{S}_{id}(\mathbf{p}) = \bar{\mathbf{s}}_{id} + \widetilde{\mathbf{U}}_{id}\mathbf{p} \quad (2)$$

Furthermore, we also consider a 3D shape model of expression variations, as offsets from a given identity shape $\mathcal{S}_{id}$. The blendshapes model of Facewarehouse [5] are utilised for this module. We adopt Nonrigid ICP [6] to accurately register this model with the LSFM identity model. Then the expression model can be represented with the triplet $\{\bar{\mathbf{s}}_{exp}, \mathbf{U}_{exp}, \mathbf{\Sigma}_{exp}\}$, where $\bar{\mathbf{s}}_{exp} \in \mathbb{R}^{3N}$ is the mean expression offset, $\mathbf{U}_{exp} \in \mathbb{R}^{3N \times n_q}$ is the orthonormal expression basis having n_q principal components and $\mathbf{\Sigma}_{exp} \in \mathbb{R}^{n_q \times n_q}$ is the diagonal matrix with the corresponding standard deviations. Similarly with the identity model, we consider the basis $\widetilde{\mathbf{U}}_{exp} = \mathbf{U}_{exp}\mathbf{\Sigma}_{exp}$ and the associated normalised parameters $\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n_q})$.

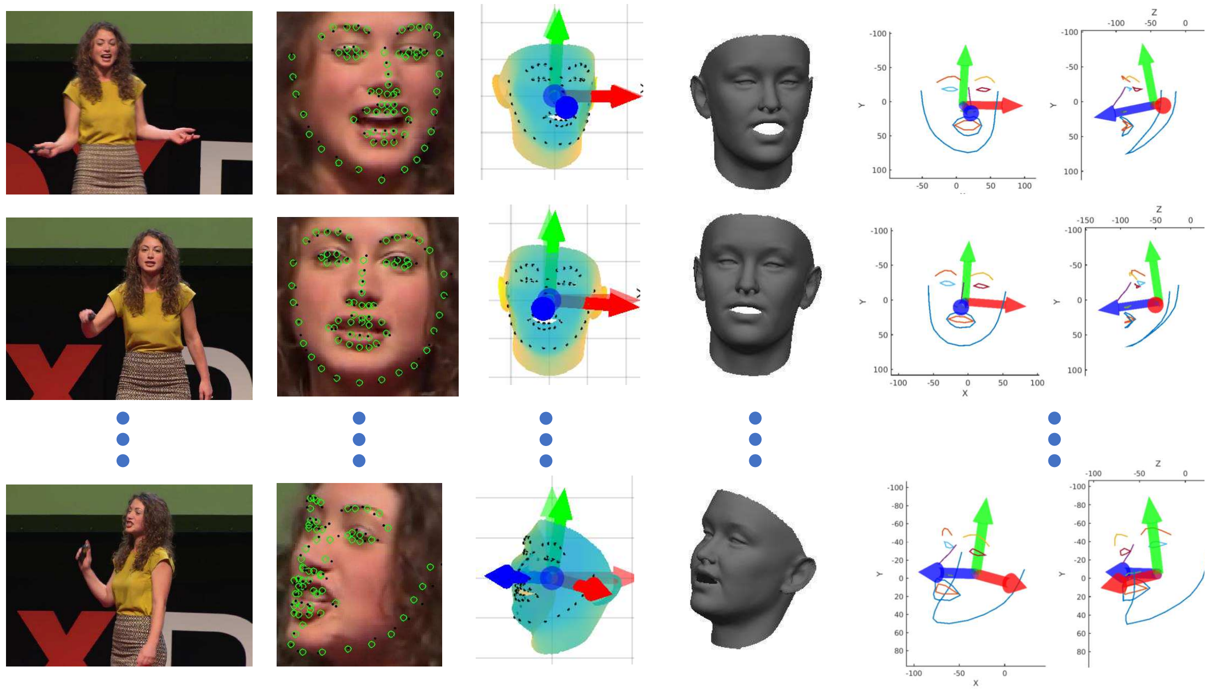
Combining the two aforementioned models, we end up with the following combined model that represents the 3D facial shape of any identity under any expression:

$$\mathcal{S}(\mathbf{p}, \mathbf{q}) = \bar{\mathbf{s}} + \widetilde{\mathbf{U}}_{id}\mathbf{p} + \widetilde{\mathbf{U}}_{exp}\mathbf{q} \quad (3)$$

where $\bar{\mathbf{s}} = \bar{\mathbf{s}}_{id} + \bar{\mathbf{s}}_{exp}$ is the overall mean shape, $\mathbf{p}$ is the vector with the identity parameters and $\mathbf{q}$ is the vector with the

[3] Simultaneously to this work we found that a similar method has been proposed in [4] for transferring 2D to 3DA-2D landmarks.

[4] The LSFM models have recently become available upon application: http://www.ibug.doc.ic.ac.uk/resources/lsfm.



(a) Input video

(b) Landmark localisation

(c) Camera estimation (rigid SfM)

(d) Dense 3D shape estimation

(e) Sampling of 3D shape on face landmarks

Figure 2. Main steps of the adopted pipeline to create ground truth 3D facial landmarks on videos. We are based on a state-of-the-art landmarker and an energy minimisation approach to fitting powerful dense 3D face models on the sequence of landmarks.

expression parameters. We construct one combined identity and expression model for each LSFM model (global or bespoke). For example, Figure 3 visualises the first few components of identity and expression for the case of global LSFM model.

3.2. Dense 3D Model Fitting

First of all, on every frame of the input video the 2D coordinates of a sparse set of facial landmarks are estimated by using the state-of-the-art facial landmarker of [3, 12], which can work under unconstrained conditions; see Figure 2(b). Crucially, this method provides a reliable estimation of the 2D projection of the real 3D positions of self-occluded landmarks even in cases of head poses close to profile views. Afterwards we fit our LSFM models on the extracted 2D landmarks locations. The rich dynamic information available in sequential frames enables us to provide very precise estimations of the ground truth shape, see Figure 2(d). More precisely, thanks to our combined identity and expression shape model, we can fix the identity parameters throughout the whole video. This is an important constraint that greatly helps our estimations. In addition, we impose temporal smoothness on the expression parameters, which improves the estimation of the 3D facial deformations of the individual observed in the input video.

3.2.1 Camera Model

The purpose of a camera model is to map (project) the object-centered Cartesian coordinates of a 3D mesh instance $\mathbf{s}$ into 2D Cartesian coordinates on an image plane.

The projection of a 3D point $\mathbf{x} = [x, y, z]^{\mathsf{T}}$ into its 2D location in the image plane $\mathbf{x}' = [x', y']^{\mathsf{T}}$ involves two steps. First, the 3D point is rotated and translated using a linear *view transformation* to bring it in the camera reference frame:

$$\mathbf{v} = [v_x, v_y, v_z]^{\mathsf{T}} = \mathbf{R}_v \mathbf{x} + \mathbf{t}_v \qquad (4)$$

where $\mathbf{R}_v \in \mathbb{R}^{3\times3}$ and $\mathbf{t}_v = [t_x, t_y, t_z]^{\mathsf{T}}$ are the camera's 3D rotation and translation components, respectively. This is based on the fact that, without loss of generality, we can assume that the observed facial shape is still and that the relative change in 3D pose between camera and object is only due to camera motion.

Then, the camera projection is applied. For the sake of computational efficiency and stability of the estimations, we consider a scaled orthographic camera, which simplifies the involved optimisation problems by making the camera pro-

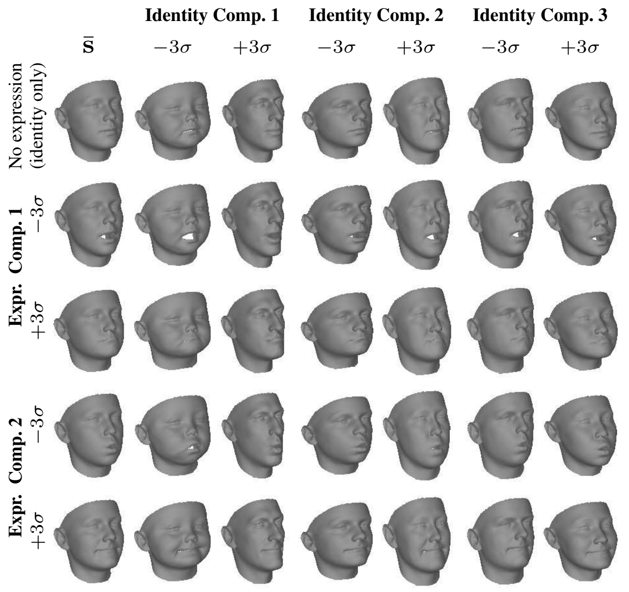


Figure 3. Principal components of identity (case of LSFM global model), expression variation and their combination, using the first 3 principal components for identity and the first 2 components for expression. Note that the first row corresponds to the identity model only.

jection function to be linear. In more detail, the 2D location of the 3D point $\mathbf{x}$ is given by:

$$\mathbf{x}' = \sigma\left[v_x, v_y\right]^{\mathsf{T}} \tag{5}$$

where σ is the scale parameter of the camera. Note that since in the scaled orthographic case the translation component t_z is ambiguous, we will consider it fixed and omit it from the subsequent formulations.

In addition, we represent the 3D rotation $\mathbf{R}_v$ using the three parameters of the axis-angle parametrisation $\mathbf{q} = \left[q_1, q_2, q_3\right]^{\mathsf{T}}$.

Camera function. The projection operation performed by the camera model of the 3DMM can be expressed with the function $\mathcal{P}(\mathbf{s}, \mathbf{c}) : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{2N}$, which applies the transformations of Eqs. (4) and (5) on the points of provided 3D mesh $\mathbf{s}$ with

$$\mathbf{c} = \left[\sigma, q_1, q_2, q_3, t_x, t_y\right]^{\mathsf{T}} \in \mathbb{R}^6 \tag{6}$$

being the vector of *camera parameters* with length $n_c = 6$. For abbreviation purposes, we represent the camera model of the 3DMM with the function $\mathcal{W} : \mathbb{R}^{n_p, n_c} \rightarrow \mathbb{R}^{2N}$ as

$$\mathcal{W}(\mathbf{p}, \mathbf{q}, \mathbf{c}) \equiv \mathcal{P}\left(\mathcal{S}(\mathbf{p}, \mathbf{q}), \mathbf{c}\right) \tag{7}$$

where $\mathcal{S}(\mathbf{p}, \mathbf{q})$ is a 3D mesh instance using Eq. (2). Finally, we denote by $\mathcal{W}_l(\mathbf{p}, \mathbf{q}_f, \mathbf{c}_f) : \mathbb{R}^{n_p, n_c} \rightarrow \mathbb{R}^{2L}$, where L is the number of the considered sparse landmarks, the selection of the elements of $\mathcal{W}(\mathbf{p}, \mathbf{q}, \mathbf{c})$ that correspond to the x, y and z coordinates of the 3D shape vertices associated with the facial landmarks.

3.2.2 Energy Formulation

To achieve highly-accurate fitting results, even in especially challenging cases, we design an energy minimisation strategy that is tailored for video input and exploits the rich dynamic information usually contained in facial videos. Since these estimations are intended for the creation of ground truth and we are not constrained by the need of real-time performance, we follow a batch approach, where we assume that all frames of the video are available from the beginning.

Let $\boldsymbol{\ell}_f = \left[x_{1f}, y_{1f}, \ldots, x_{Lf}, y_{Lf}\right]^{\mathsf{T}}$ be the 2D facial landmarks for the f-th frame estimated by the method of [3]. Even though we consider the identity parameters $\mathbf{p}$ as fixed over the frames of the video, we expect that every frame has its own expression, camera, and texture parameters vectors, which we denote by $\mathbf{q}_f$, $\mathbf{c}_f$ and $\boldsymbol{\lambda}_f$ respectively. We also denote by $\hat{\mathbf{q}}$, $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\lambda}}$ the concatenation of the corresponding parameter vectors over all frames (with n_f being the number of frames of the video): $\hat{\mathbf{q}}^{\mathsf{T}} = \left[\mathbf{q}_1^{\mathsf{T}}, \ldots, \mathbf{q}_{n_f}^{\mathsf{T}}\right]$, $\hat{\mathbf{c}}^{\mathsf{T}} = \left[\mathbf{c}_1^{\mathsf{T}}, \ldots, \mathbf{c}_{n_f}^{\mathsf{T}}\right]$ and $\hat{\boldsymbol{\lambda}}^{\mathsf{T}} = \left[\boldsymbol{\lambda}_1^{\mathsf{T}}, \ldots, \boldsymbol{\lambda}_{n_f}^{\mathsf{T}}\right]$

To fit a 3D face model on the facial landmarks, we propose to minimise the following energy:

$$\begin{aligned}\hat{E}(\mathbf{p}, \hat{\mathbf{q}}, \hat{\mathbf{c}}) &= \hat{E}_{\text{land}}(\mathbf{p}, \hat{\mathbf{q}}, \hat{\mathbf{c}}) \\ &\quad + \hat{E}_{\text{priors}}(\mathbf{p}, \hat{\mathbf{q}}) + c_{sm}\hat{E}_{\text{smooth}}(\hat{\mathbf{q}})\end{aligned} \tag{8}$$

where $\hat{E}_{\text{land}}$, $\hat{E}_{\text{priors}}$ and $\hat{E}_{\text{smooth}}$ are a multi-frame 2D landmarks term, a prior regularisation term and a temporal smoothness term respectively. Also c_{sm} is a balancing weights for the temporal smoothness term.

The multi-frame **2D landmarks term** ($\hat{E}_{\text{land}}$) is a summation of the reprojection error of the sparse 2D landmarks for all frames:

$$\hat{E}_{\text{land}}(\mathbf{p}, \hat{\mathbf{q}}, \hat{\mathbf{c}}) = \sum_{f=1}^{n_f} \left\|\mathcal{W}_l(\mathbf{p}, \mathbf{q}_f, \mathbf{c}_f) - \boldsymbol{\ell}_f\right\|^2 \tag{9}$$

The **shape priors term** ($\hat{E}_{\text{priors}}$) imposes priors on the reconstructed 3D facial shape of every frame. Since the facial shape at every frame is derived from the (zero-mean and unit-variance) identity parameter vector $\mathbf{p}$ and the frame-specific expression parameter vector $\mathbf{q}_f$ (also zero-mean and unit-variance), we define this term as:

$$\begin{aligned}\hat{E}_{\text{priors}}(\mathbf{p}, \hat{\mathbf{q}}) &= \hat{c}_{id}\left\|\mathbf{p}\right\|^2 + c_{exp}\sum_{f=1}^{n_f}\left\|\mathbf{q}_f\right\|^2 \\ &= \hat{c}_{id}\left\|\mathbf{p}\right\|^2 + c_{exp}\left\|\hat{\mathbf{q}}\right\|^2\end{aligned} \tag{10}$$

where $\hat{c}_{id}$ and c_{exp} are the balancing weights for the prior terms of identity and expression respectively.

The **temporal smoothness term** ($\hat{E}_{\text{smooth}}$) enforces smoothness on the expression parameters vector $\mathbf{q}_f$ by penalising the squared norm of the discrimination of its 2nd temporal derivative. This corresponds to the regularisation imposed in smoothing splines and leads to naturally smooth trajectories over time. More specifically, this term is defined as:

$$\hat{E}_{\text{smooth}}(\hat{\mathbf{q}}) = \sum_{f=2}^{n_f-1} \|\mathbf{q}_{f-1} - 2\mathbf{q}_f + \mathbf{q}_{f+1}\|^2 = \left\|\mathbf{D}^2\hat{\mathbf{q}}\right\|^2 \tag{11}$$

where the summation is done over all frames for which the discretised 2nd derivative can be expressed without having to assume any form of padding outside the temporal window of the video. Also $\mathbf{D}^2 : \mathbb{R}^{n_q n_f} \to \mathbb{R}^{n_q(n_f-2)}$ is the linear operator that instantiates the discretised 2nd derivative of the n_q-dimensional vector $\mathbf{q}_f$. This means that $\mathbf{D}^2\hat{\mathbf{q}}$ is a vector that stacks the vectors $(\mathbf{q}_{f-1} - 2\mathbf{q}_f + \mathbf{q}_{f+1})$, for f=$2, \ldots, n_f - 1$. It is worth mentioning that we could have imposed temporal smoothness on the parameters $\mathbf{c}_f$, $\boldsymbol{\lambda}_f$ too. However, we have empirically observed that the temporal smoothness on $\mathbf{q}_f$, in conjunction with fixing the identity parameters $\mathbf{p}$ over time, is adequate for accurate and temporally smooth estimations. Following the Occam's razor principle, our design choice is to avoid expanding the energy with additional unnecessary terms (it also keeps the number of hyper-parameters as low as possible).

3.2.3 Optimisation of the Proposed Energy

As described next, we first estimate the camera parameters $\hat{\mathbf{c}}$ (see Figure 2(c)) and afterwards the shape parameters $(\mathbf{p}, \hat{\mathbf{q}})$ (see Figure 2(d)).

Camera Parameters Estimation. In this initial step, we solely consider the 2D landmarks term $\hat{E}_{\text{land}}$, which is the only term of the energy $\hat{E}(\mathbf{p}, \hat{\mathbf{q}}, \hat{\mathbf{c}})$ that depends on $\hat{\mathbf{c}}$. We minimise $\hat{E}_{\text{land}}$ by assuming that the unknown facial shape is fixed over all frames, but does not necessarily lie on the subspace defined by the combined shape model of Eq. (2). In other words, the facial shape $\mathcal{S}$ is considered to have $3N$ free parameters, corresponding to the 3D coordinates of the N vertices of the 3D shape. However, since in this step the energy that is minimized involves only the sparse landmarks, only the 3D coordinates of the vertices that correspond to the sparse landmarks can actually be estimated. (i.e., $3L$ parameters in total for the 3D shape).

Note that the estimation of the rigid shape is only done to facilitate the camera parameters' estimation, which is the main goal of this step. The assumption of facial shape rigidity during the whole video is over-simplistic. However, as verified experimentally, it provides a very robust initialisation of the camera parameters even in cases of large facial deformation, provided that it is fed with significant amount of frames. This is due to the nature of physical deformations observed in human faces, which can be modelled as relatively localised deviations from a rigid shape.

Under the aforementioned assumptions, the 2D landmarks term can be written as:

$$\hat{E}_{\text{land}}(\mathcal{S}_{\text{rig}}, \hat{\boldsymbol{\Pi}}) = \left\|\widehat{\mathbf{L}} - \hat{\boldsymbol{\Pi}}\,\mathcal{S}_{\text{rig}}\right\|_F^2 \tag{12}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm and $\mathcal{S}_{\text{rig}}$ is a $3 \times L$ matrix with the unknown sparse rigid shape, where every column of $\mathcal{S}_{\text{rig}}$ contains the 3D coordinates of a specific landmark point. Also, $\widehat{\mathbf{L}}$ is a $2n_f \times L$ matrix that stacks the matrices $\widetilde{\mathbf{L}}_f$ (f=1, .., n_f), which are the re-arrangements of the landmarks vectors $\boldsymbol{\ell}_f$ into $2 \times L$ matrices:

$$\widehat{\mathbf{L}} = \begin{bmatrix} \widetilde{\mathbf{L}}_1 \\ \vdots \\ \widetilde{\mathbf{L}}_{n_f} \end{bmatrix}, \widetilde{\mathbf{L}}_f = \begin{bmatrix} \tilde{x}_{1f} & \cdots & \tilde{x}_{Lf} \\ \tilde{y}_{1f} & \cdots & \tilde{y}_{Lf} \end{bmatrix} \tag{13}$$

Note that, without loss of generality, the landmarks $\widetilde{\mathbf{L}}_f$ are considered to have their centroid at the origin $(0, 0)$. This means that the landmark coordinates $(\tilde{x}_{if}, \tilde{y}_{if})$ are derived from the original coordinates (x_{if}, y_{if}) after subtracting their per-frame centroid.

In addition, $\hat{\boldsymbol{\Pi}} = \left[\boldsymbol{\Pi}_1^{\mathsf{T}} \cdots \boldsymbol{\Pi}_{n_f}^{\mathsf{T}}\right]^{\mathsf{T}}$ is a $2n_f \times L$ matrix that stacks the scaled orthographic projection matrices $\boldsymbol{\Pi}_f \in \mathbb{R}^{2\times 3}$ from all the frames f. The matrix $\boldsymbol{\Pi}_f$ is derived by the first two rows of the 3D rotation matrix $\mathbf{R}_v$ of the camera (see Eq. (4)), after multiplying them with the scale parameter σ_f of the camera for the frame f. Therefore, an orthogonality constraint should be imposed on each $\boldsymbol{\Pi}_f$:

$$\boldsymbol{\Pi}_f \boldsymbol{\Pi}_f^{\mathsf{T}} = \sigma_f^2 \mathbf{I}_2, \text{ for some } \sigma_f > 0,\ f = 1, \ldots, n_f \tag{14}$$

To summarise, our goal is to minimise $\hat{E}_{\text{land}}$ as described in Eq. (12) with respect to $\mathcal{S}_{\text{rig}}$ and $\hat{\boldsymbol{\Pi}}$, under the constraints of Eq. (14). For this, we employ a simple yet effective rigid Structure from Motion (SfM) approach [20]: We solve the problem based on a rank-3 factorisation of the matrix $\widehat{\mathbf{L}}$.

Regarding the translation part of the camera motion, its x and y components at frame f are derived by the centroid of the original landmarks $\boldsymbol{\ell}_f$ that has been subtracted in the computation of the landmarks $\widetilde{\mathbf{L}}_f$ in Eq. (13). This can be easily verified that is the optimal choice . Regarding the z component of the translation, this is inherently ambiguous due to the orthographic projection, therefore we fix it to a constant value over all frames.

Finally, to yield the camera parameters that will be used in conjunction with the shape model of Eq. (2), we perform a rigid registration between the model's mean shape $\bar{\mathbf{s}}_{id}$ (sampled at the vertices that correspond to the landmarks) and the rigid shape $\mathcal{S}_{\text{rig}}$ estimated by SfM. The similarity transform that registers the two sparse shapes is recovered using Procrustes Analysis and then combined with each frame's similarity transform that is estimated by SfM. This yields a sequence of estimated camera parameters $\mathbf{c}_1, \ldots, \mathbf{c}_{n_f}$. As the final processing for this initialisation step, this sequence is temporally smoothed by using cubic smoothing splines.

Shape Parameters Estimation. Using the estimation of camera parameters $\hat{\mathbf{c}}$, we minimise the energy $\hat{E}$ of Eq. (8) with respect to the shape parameters $\mathbf{p}$ and $\hat{\mathbf{q}}$. This is a linear least squares problem that we can solve very efficiently. In more detail, we can write $\hat{E}$ as follows:

$$
\begin{aligned}
&\hat{E}(\mathbf{p}, \hat{\mathbf{q}}) = \\
&c_\ell \sum_{f=1}^{n_f} \left\| (\mathbf{I}_L \otimes \mathbf{\Pi}_f) \left(\bar{\mathbf{s}}^{(\ell)} + \widetilde{\mathbf{U}}_{id}^{(\ell)} \mathbf{p} + \widetilde{\mathbf{U}}_{exp}^{(\ell)} \mathbf{q}_f \right) - \boldsymbol{\ell}_f \right\|^2 \\
&+ \hat{c}_{id} \left\| \mathbf{p} \right\|^2 + c_{exp} \left\| \hat{\mathbf{q}} \right\|^2 + c_{sm} \left\| \mathbf{D}^2 \hat{\mathbf{q}} \right\|^2
\end{aligned}
\tag{15}
$$

where $\bar{\mathbf{s}}^{(\ell)}$, $\widetilde{\mathbf{U}}_{id}^{(\ell)}$, $\widetilde{\mathbf{U}}_{exp}^{(\ell)}$ are matrices with the rows of $\bar{\mathbf{s}}$, $\widetilde{\mathbf{U}}_{id}$, $\widetilde{\mathbf{U}}_{exp}$ respectively that correspond to the x, y and z coordinates of 3D shape vertices associated with facial landmarks. Also, "$\otimes$" denotes Kronecker product, such that the multiplication with the $2L \times 3L$ matrix $\mathbf{I}_L \otimes \mathbf{\Pi}_f$ implements the application of the camera projection $\mathbf{\Pi}_f$ on each one of the L landmarks.

Note that the sparse landmarks, in conjunction with the adopted high-quality shape models, are able to yield surprisingly plausible estimations of the dynamic facial shape, in most of the cases. However, in some very challenging case (e.g. frames with very strong occlusions or gross errors in the landmarks), this sparse information might not be adequate for satisfactory results. One way to compensate for that would be to increase the regularisation weights $\hat{c}_{id}$ and c_{exp}. Nevertheless, this would strongly affect also the non-pathological cases, where the results are plausible either way, leading to reconstructed shapes and expressions that would be too similar with the mean shape $\bar{\mathbf{s}}$. To avoid that, we follow a different approach by keeping the regularisation weights as low as in the main optimisation and imposing the following *box constraints*:

$$
\begin{aligned}
&|(\mathbf{p})_i| \leq M_p \,, \; i = 1, \ldots, n_p \\
&|(\mathbf{q}_f)_i| \leq M_q \,, \; i = 1, \ldots, n_q \text{ and } f = 1, \ldots, n_f
\end{aligned}
\tag{16}
$$

where $(\cdot)_i$ denotes the selection of the i-th component from a vector. Also, M_p and M_q are positive constants corresponding to the maximum values allowed for the components of identity and expression parameter vectors respectively. These are set so that the corresponding components does not attain a value higher than a certain number of standard deviations (e.g. 4). These constraints are activated only in pathological cases and do not play any role in all the rest cases, which actually are the vast majority. Note also that they are only used in this initialisation step, since when the dense texture information is used as input, they are not required.

To summarise, our goal here is to minimise the energy $\hat{E}$ of Eq. (15) with respect to the shape parameters $\mathbf{p}$ and $\hat{\mathbf{q}}$ under the constraints of Eq. (16). This corresponds to a large-scale linear least squares problem of the form $\arg\min_{\mathbf{x}} \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|^2$, under bound constraints on $\mathbf{x}$, where the matrix $\mathbf{A}$ is sparse. We solve this problem efficiently by adopting the reflective Newton method of [9].

3.3. Sampling on Face Landmarks and Reprojection

After having estimated the shape parameters $(\mathbf{p}, \mathbf{q}_f)$ for every frame of a video, the estimated dense facial mesh in the model space can be synthesised by the model as $\mathcal{S}_f(\mathbf{p}, \mathbf{q}_f) = \bar{\mathbf{s}} + \widetilde{\mathbf{U}}_{id}\mathbf{p} + \widetilde{\mathbf{U}}_{exp}\mathbf{q}_f$. The ground truth 3D landmarks $\mathcal{S}_f^\ell$ are then extracted by keeping the elements of $\mathcal{S}_f$ that contain the x,y and z coordinates of vertices that correspond to the facial landmarks. Note that for the extraction of the 3D landmarks we do not apply the camera parameters, meaning that these landmarks lie on the normalised model space. The reprojected ground truth 2D landmarks (i.e., the 3DA-2D landmarks) are expressed in the image space, therefore to extract them we utilise the estimated camera parameters $\mathbf{c}_f$ and apply the camera function $\mathcal{P}(\cdot)$ to $\mathcal{S}_f^\ell$. This corresponds to the quantity $\mathcal{W}_l(\mathbf{p}, \mathbf{q}_f, \mathbf{c}_f)$, see Sec. 3.2.1.

4. Experiments

During the challenge we provided approximately 14,000 static images with 3DA-2D and 3D landmarks, as well as approximately 90 training videos annotated with the proposed procedure. We believe that the followed procedure, even though semi-automatic, is suitable for providing a high quality ground-truth, since we have tested it in simulated videos and it provided extremely high accuracy (submilimeter accuracy for some landmarks). Additionally, in both the trainset and the testset, the parameter estimation and fitting was performed in the whole video, however we have exported the 3DA-2D and 3D only in the first couple of thousand frames, hence there was information only available to us (latent for participants) to ensure the high quality of our estimations.

The training data have been provided to over 25 groups from all over the world. A tight schedule (a week) was pro-

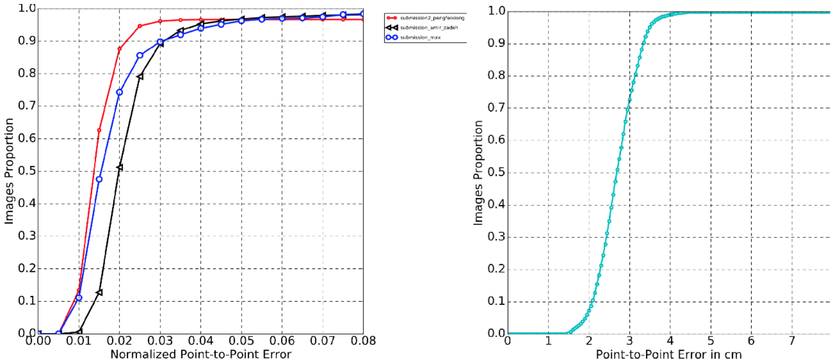


Figure 4. CED curvers for (left graph) the 3DA-2D landmark tracking and (right graph) pure 3D facial landmark tracking (the only group that has sent results for this category was [10]).

vided to return the results on the testset. The testset comprises of 110 videos with 1,000 frames each. The evaluation was performed in the 30 most challenging videos[5]. Results for 3DA-2D landmarks localisation have been returned by three groups, while results for 3D landmarks have been returned by one group only.

For assessing the performance of the submissions we used standard evaluation metrics. That is, for localisation of 3DA-2D landmarks we used normalised root-mean square error (for more details please refer to previous competitions such as [24]). In this challenge, we used the face diagonal as the normalisation factor which is more robust to changes of the face pose. For localisation of 3D landmarks we used the root-mean square error but appropriately normalised first so that it is in cm scale. All three contestants submitted results for 3DA-2D landmark localisation, while we had only one submission that returned results for 3D landmark localisation. In the following, we will briefly describe each participating method:

The method in [10] (abbreviated as submission_max) proposes to jointly estimate facial landmarks and dense facial geometry using a Deep Convolutional Neura Network (DCNN). The geometry is refined by fitting a linear 3D Morphable Model (3DMM) on the estimates from the DCNN.

The method in [23] (submission_amir_zadeh) proposes to apply an extension of the popular Constrained Local Models (CLMs), the so-called Convolutional Experts (CE)-CLMs for the problem of 3DA-2D facial landmark detection. The important module of CE-CLMs is a novel convolutional local detector that brings together the advantages of neural architectures and mixtures. In order to improve further the performance on 3D face tracking the authors use two complementary networks alongside CE-CLM: a network that maps the output of CE-CLM to 84 landmarks called Adjustment Network, and a Deep Residual Network called Correction Networks that learns dataset specific corrections for CE-CLM.

The method in [21] (submission2_pengfeixiong) proposes a two stage shape regression method by combining the powerful local heatmap regression and global shape regression. The base of the method is the now popular stacked hourglass network which is used to generate a set of heatmaps for each 3d shape point by first. While these heatmaps are independent on each other, a hierarchical attention mechanism is applied from global to local heatmaps into the network, in order to model the correlations among neighboring regions. Then, all these heatmaps, alongside the input aligned image are processed by a deep residual network to further leanr the global features and produce the final smooth 3D shape. The CED curves are summarised in Figure 4. The best performing method on 3DA-2D facial landmark tracking was the method [21]. For pure 3D face tracking the only method that competed in this category was [10] .

5. Conclusion

In this paper we presented the 3D Menpo database and the results of the first challenge on 3DA-2D and 3D facial landmark tracking. The 3D Menpo database comprises of (a) around 14,000 static images which are suitable for training or guiding 3D facial landmark localisation algorithms and (b) around 280,000 annotated frames (the combined model space and reprojected space). We introduced an elaborate semi-automatic methodology for providing high quality annotations for training and assessing the performance of 3D facial landmark tracking algorithms. The challenge demonstrates that very good results can be attained for 3DA-2D facial landmark tracking.

6. Acknowledgements

The work of S. Zafeiriou and A. Roussos has been partially funded by the EPSRC Project EP/N007743/1.

[5] There are 4 additional videos with multiple people, however the participants opted for single person evaluation.

References

[1] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017. 2

[2] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 3

[3] A. Bulat and G. Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016. 1, 4, 5

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv:1703.07332*, 2017. 3

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *T-VCG*, 20(3):413–425, 2014. 3

[6] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic. Statistical non-rigid icp algorithm and its application to 3d face alignment. *Image and Vision Computing*, 2016. 3

[7] G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *IJCV*, 2017. 1

[8] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015. 3

[9] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM*, 6(4):1040–1058, 1996. 7

[10] D. Crispell and M. Bazik. Pix2face: Direct 3d face model estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 7, 8

[11] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[12] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *arxiv*. 1, 2, 3, 4

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[14] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision, Workshops*, pages 511–520. Springer, 2016. 2

[15] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 1

[16] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 1

[17] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV-W*, Sydney, Australia, December 2013. 1

[18] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV-W*, December 2015. 1

[19] G. Trigeorgis, P. Snape, M. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment. In *CVPR*, Las Vegas, NV, USA, June 2016. IEEE. 1

[20] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19(1):107–130, 1982. 6

[21] P. Xiong, G. Li, and Y. Sun. 3d face tracking via two-stage hierarchically attentive shape regression network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 7, 8

[22] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017. 1

[23] A. Zadeh, Y. C. Lim, T. Baltrusaitis, and L.-P. Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 7

[24] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR-W*, July 2017. 1, 7

[25] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 2, 3