International Journal of Computer Vision (2019) 127:599–624
https://doi.org/10.1007/s11263-018-1134-y

The Menpo Benchmark for Multi-pose 2D and 3D Facial Landmark Localisation and Tracking

Jiankang Deng[1] · Anastasios Roussos[2] · Grigorios Chrysos[1] · Evangelos Ververas[1] · Irene Kotsia[3] · Jie Shen[1] · Stefanos Zafeiriou[1,4]

Received: 22 February 2018 / Accepted: 14 November 2018 / Published online: 29 November 2018

Abstract

In this article, we present the Menpo 2D and Menpo 3D benchmarks, two new datasets for multi-pose 2D and 3D facial landmark localisation and tracking. In contrast to the previous benchmarks such as 300W and 300VW, the proposed benchmarks contain facial images in both semi-frontal and profile pose. We introduce an elaborate semi-automatic methodology for providing high-quality annotations for both the Menpo 2D and Menpo 3D benchmarks. In Menpo 2D benchmark, different visible landmark configurations are designed for semi-frontal and profile faces, thus making the 2D face alignment full-pose. In Menpo 3D benchmark, a united landmark configuration is designed for both semi-frontal and profile faces based on the correspondence with a 3D face model, thus making face alignment not only full-pose but also corresponding to the real-world 3D space. Based on the considerable number of annotated images, we organised Menpo 2D Challenge and Menpo 3D Challenge for face alignment under large pose variations in conjunction with CVPR 2017 and ICCV 2017, respectively. The results of these challenges demonstrate that recent deep learning architectures, when trained with the abundant data, lead to excellent results. We also provide a very simple, yet effective solution, named Cascade Multi-view Hourglass Model, to 2D and 3D face alignment. In our method, we take advantage of all 2D and 3D facial landmark annotations in a joint way. We not only capitalise on the correspondences between the semi-frontal and profile 2D facial landmarks but also employ joint supervision from both 2D and 3D facial landmarks. Finally, we discuss future directions on the topic of face alignment.

Keywords 2D face alignment · 3D face alignment · Menpo challenge

1 Introduction

Facial landmark localisation and tracking on images and videos captured in unconstrained recording conditions is a problem that has received a lot of attention during the past few years. This is attributed to the fact that it is a necessary pre-processing step for many applications such as face recognition (Taigman et al. 2014), facial behaviour analysis (Eleftheriadis et al. 2016b, a), lip reading (Chung and Zisserman 2016; Chung et al. 2017), 3D face reconstruction (Booth et al. 2016, 2017, 2018) and face editing (Shu et al. 2017), just to name a few.

Currently, methodologies (Xiong and De la Torre 2013; Ren et al. 2014; Zhu et al. 2016a; Trigeorgis et al. 2016; Güler et al. 2017; Bulat and Tzimiropoulos 2017a, b; Honari et al. 2018) that achieve good performance in facial landmark localisation have been presented in recent top-tier computer vision conferences (e.g., CVPR, ICCV, ECCV). This progress would not be feasible without the efforts made by the scientific community to design and develop both benchmarks with high-quality landmark annotations (Sagonas et al. 2013, 2016; Belhumeur et al. 2013; Le et al. 2012; Zhu and Ramanan 2012; Köstinger et al. 2011), as well as rigorous protocols for performance assessment.

Communicated by Rama Chellappa, Xiaoming Liu, Tae-Kyun Kim, Fernando De la Torre, Chen Change Loy.

✉ Jiankang Deng
j.deng16@imperial.ac.uk

[1] Department of Computing, Imperial College London, London, UK

[2] Department of Computer Science, University of Exeter, Exeter, UK

[3] Department of Computer Science, Middlesex University London, London, UK

[4] Centre for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

The recent benchmarks 300W (Sagonas et al. 2013) and 300VW (Shen et al. 2015) are amongst the most popular datasets for facial landmark localisation and tracking and are widely used by both the scientific community and the industry (Trigeorgis et al. 2016; Zhu et al. 2016a; Ren et al. 2014; Xiong and De la Torre 2013; https://megvii.com; https://www.sensetime.com).

The 300W benchmark was developed for the corresponding 300W competition held in conjunction with ICCV 2013. It provides publicly available annotations for images, which originate from the following "in-the-wild" datasets:

- Labeled Face Parts in the Wild (LFPW) dataset (Belhumeur et al. 2013). Since LFPW provides only the source links to download the images and not the actual images, only 1035 images were available (out of 1287).
- Helen dataset (Le et al. 2012) which consists of 2330 high resolution images downloaded from the `flickr.com` web service.
- The Annotated Faces in-the-wild (AFW) (Zhu and Ramanan 2012) dataset which consists of 250 images with 468 faces.
- Two new datasets, namely the IBUG dataset and the 300W test set. IBUG consists of 135 images. In addition, 300W test set consists of 300 images captured indoors and 300 images captured outdoors. The 300W test set was publicly released with the second version of the competition (Sagonas et al. 2016).

In total, the 300W benchmark provides 4350 "in-the-wild" images that contain around 5000 faces. The faces have been annotated using the 68 landmark frontal face markup scheme shown in Fig. 3a.[1]

The next competition on the topic was held in conjunction with ICCV 2015 and revolved around facial landmark tracking "in-the-wild". The challenge introduced the 300VW benchmark (Shen et al. 2015). The 300VW benchmark consists of 114 videos and 218,595 frames. For a recent comparison of the state-of-the-art in 300VW, the interested reader may refer to Chrysos et al. (2018). The 68 frontal face markup scheme was also used for annotating the faces of the 300VW benchmark. Even though the data of 300W and 300VW had a large impact in the computer vision community, there are still two obvious limitations. More specifically, the pose variations and the data size were limited.

In 300W and 300VW, the data were annotated using only the semi-frontal shape (68 landmarks), and there are few faces in extreme poses (e.g. full profile face images). Large-pose face alignment is a very challenging task, until now there are not enough annotated facial images in arbitrary poses, especially with a large number of landmarks. Annotated Facial Landmarks in the Wild (AFLW) (Koestinger et al. 2011) provides a large-scale collection of annotated face images gathered from Flickr, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. In total, about 25k faces are annotated with up to 21 landmarks, but excluding coordinates for invisible landmarks, causing difficulties for model training. Although all invisible landmarks are manually annotated in Zhu et al. (2016a), there is no landmark annotation along the face contour. In Zhu et al. (2016b), a large-scale training dataset in profile views are synthesised from the 300W dataset with the assistance of a 3D Morphable Model (3DMM). However, the generated profile face images have artifacts that affect the alignment accuracy. In Jeni et al. (2016), a 3D landmark localisation challenge was organised in conjunction with ECCV 2016. However, it mainly revolved around images that have been either captured in highly controlled conditions or generated artificially (i.e., rendering a 3D face captured in controlled conditions using arbitrary backgrounds).

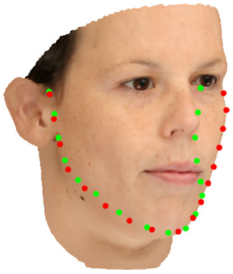


Fig. 1 Landmark annotation on face contour between 2D and 3D views. Red annotation is from 2D view, and green annotation is from 3D view (Color figure online)

In this paper, we introduce multi-view 2D and 3D facial landmark annotations to facilitate large-pose face alignment. As shown in Fig. 1, we consider two kinds of annotation schemes. The first one is the 2D scheme (Red). That is, facial landmarks are always located on the visible face boundary. Faces which exhibit large facial poses are extremely challenging to annotate under this configuration because the landmarks on the invisible face side stack together. To this end, we used different 2D landmark annotation schemes for semi-frontal and profile faces. The second one is the 3D scheme (Green). Since the invisible face contour needs to be consistent with 3D face models, labelling the self-occluded 3D landmarks is quite difficult for human annotators. To this end, we present an elaborate semi-automatic methodology for providing high-quality 3D annotations with the assistance of a state-of-the-art 3DMM model fitting method (Booth et al. 2016).

In order to train recent deep learning architectures, such as ResNets (He et al. 2016) and Stacked Hourglass (Newell et al. 2016), large-scale training data are required. However, the training data provided in 300W and 300VW are quite lim-

[1] Please note that this markup scheme had been also used in Multi-PIE dataset (Gross et al. 2010).

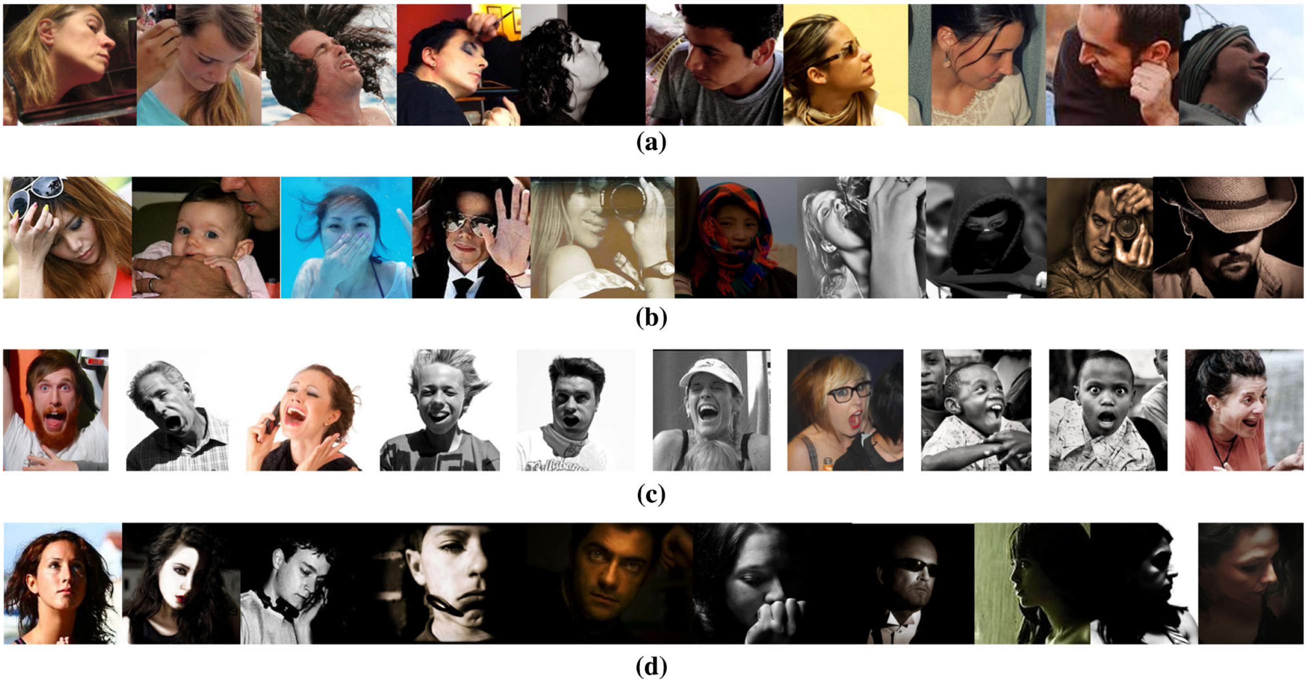


Fig. 2 Some extremely challenging examples (e.g. **a** pose, **b** occlusion, **c** expression, **d** illumination) from the proposed Menpo dataset

ited. To facilitate the training of face alignment models, we provide a large number of facial images with 2D and 3D landmark annotations. In this paper, we make a significant step to face alignment and propose a new comprehensive large pose and large-scale benchmark, which contains both semi-frontal and profile faces, annotated with their corresponding 2D and 3D facial shapes. We introduce an elaborate semi-automatic methodology for providing high-quality 2D and 3D facial landmark annotations. The annotations have been used to set two new benchmarks and two challenges, i.e. Menpo 2D and Menpo 3D. In the Menpo 2D challenge, the participants had access to over 10k annotated semi-frontal faces images (300W and Menpo challenge training data) and 1906 annotated profile images. In the Menpo 3D challenge, the participants had access to around 12k annotated face images and around 280k annotated face frames with regards to the 3D annotation scheme. We outline the results achieved by the participants of the Menpo 2D and 3D challenges. The results demonstrate that recent deep learning architectures, when trained with the abundant data of the Menpo challenges, lead to excellent results. Finally, we also propose a new, simple and very strong baseline, named Cascade Multiview Hourglass Model, for 2D and 3D face alignment. In our method, we not only capitalise on the correspondences between the semi-frontal and profile 2D facial landmarks but also employ joint supervision from both 2D and 3D facial landmarks. Finally, we discuss future directions on the topic of face alignment.

2 Menpo 2D and Menpo 3D Benchmarks

In this section, we present the Menpo 2D and Menpo 3D benchmarks, in terms of the datasets used, the adopted landmark configurations, as well as the creation of ground-truth landmark annotations.

2.1 Datasets

The datasets of Menpo 2D and Menpo 3D benchmarks include face images and videos under completely unconstrained conditions, which exhibit large variations in pose, occlusion, expression and illumination. In Fig. 2, we illustrate some extremely challenging examples from the proposed Menpo dataset. In fact, these four factors have a significant influence on the local facial appearance and thus affect the local feature for a particular face alignment model.

- *Pose* Large pose variations can cause heavy self-occlusion, and some facial components such as half of the facial contour can even be completely missing in a profile face.
- *Occlusion* Occlusion frequently happens on facial contour and some facial organs (e.g. sunglasses on eyes and food on mouths) under uncontrolled conditions. Heavy occlusions can bring great challenges to the in-the-wild face alignment as the facial appearance can be locally changed or even completely missing.

- *Expression* Some inner facial components (e.g. mouth and eyes) have their own variation patterns. Especially, mouth shape is largely affected by some expressions (e.g. supervise and happy), thus it is very challenging for face alignment under exaggerated expressions.
- *Illumination* Illumination changes (e.g. intensity and direction variations) can significantly alter the facial appearance, and even make some detailed textures missing.

In more detail, the dataset of **Menpo 2D Benchmark** consists of the following:

- *Image Training Set* 5658 semi-frontal and 1906 profile facial images from AFLW (Koestinger et al. 2011) and FDDB (Jain and Learned-Miller 2010).
- *Image Test Set* 5335 semi-frontal and 1946 profile facial images from AFLW (Koestinger et al. 2011) and FDDB (Jain and Learned-Miller 2010).

Furthermore, the dataset of **Menpo 3D Benchmark** consists of the following:

- *Image Training Set* 337 images from the Annotated Faces in-the-wild (AFW) (Zhu and Ramanan 2012), 1035 images from the Labeled Face Parts in the Wild (LFPW) (Belhumeur et al. 2013), 2300 images from Helen (Le et al. 2012), 135 images from IBUG (Sagonas et al. 2013), 600 images from 300W (Sagonas et al. 2013, 2016) and 7564 images from Menpo 2D training dataset (Zafeiriou et al. 2017b).
- *Video Training set* 55 videos from 300VW (Shen et al. 2015).
- *Video Test Set* 111 in-the-wild videos manually collected from YouTube.

It is worth noting that for Menpo 3D, the test set contains only videos because this is a tracking challenge. However, an image training set is also provided, which give the ability to participants to train their methods on not only videos but also images.

In Table 1, we list the details (e.g. year, size and annotations) of several previously famous face alignment datasets, ranging from the controlled 2D datasets to the recent in-the-wild challenging 2D datasets and 3D datasets.

Early 2D face alignment datasets [e.g. XM2VTS (Messer et al. 1999), BioID (Jesorsky et al. 2001), FRGC (Phillips et al. 2005), PUT (Kasinski et al. 2008), BUHMAP-DB (Aran et al. 2007), MUCT (Milborrow et al. 2010) and Multi-PIE (Gross et al. 2010)] are collected under controlled conditions with neutral expression, frontal pose and normal lighting. On these controlled datasets, the classic approaches [e.g. ASM (Cootes et al. 1995), AAM (Cootes et al. 2001) and CLM (Cristinacce and Cootes 2006)] have set up state-of-the-art performance, and face alignment under controlled conditions has been well-solved by now.

Recently, some uncontrolled datasets [e.g. LFW (Huang et al. 2008), AFLW (Köstinger et al. 2011), LFPW (Belhumeur et al. 2013), AFW (Zhu and Ramanan 2012), HELEN (Le et al. 2012), COFW (Burgos-Artizzu et al. 2013; Ghiasi and Fowlkes 2015), 300-W (Sagonas et al. 2013, 2016), 300VW (Shen et al. 2015), MTFL (Zhang et al. 2014), MAFL (Zhang et al. 2016b)], which exhibit large appearance variations due to pose, expression, occlusion and illumination, are introduced to investigate the problem of 2D face alignment in-the-wild. From the competition results of 300W (Sagonas et al. 2013, 2016), we can find that cascade regression based methods (Xiong and De la Torre 2013; Yan et al. 2013; Deng et al. 2016) and deep Convolutional Neural Networks (CNN) (Zhou et al. 2013; Fan and Zhou 2016) set up the state-of-the-art performance for the in-the-wild 2D face alignment. On the Menpo 2D challenge (Zafeiriou et al. 2017c), we provide a large-scale and multi-pose dataset, and more advanced deep convolutional structures (Yang et al. 2017; He et al. 2017) are designed to improve the robustness of 2D face alignment in the wild.

Due to the inconsistency of the 2D facial landmark configuration under large pose variations, 3D facial landmarks are introduced into face alignment (Zhu et al. 2016c). Zhu et al. (2016c) proposed a method to synthesise large-scale training samples in profile views [300W-LP (Zhu et al. 2016c)] and employed CNN to fit the dense 3D face model to facial images. In Zafeiriou et al. (2017a), an elaborate semi-automatic methodology was proposed to provide high-quality 3D landmark annotations for face images and videos. From the results of the Menpo 3D challenge (Zafeiriou et al. 2017a), we find stacked hourglass network (Xiong et al. 2017) once again set up state-of-the-art performance.

2.2 Adopted Landmark Configurations

We adopt four different types of landmark configurations:

- *Semi-frontal2D landmarks*, which we use in the Menpo 2D benchmark.
- *Profile 2D landmarks*, which we also use in the Menpo 2D benchmark.
- *3DA-2D (3D Aware 2D) landmarks*, which we use in the Menpo 3D benchmark.
- *3D landmarks*, which we also use in the Menpo 3D benchmark.

In more detail, *Semi-frontal 2D landmarks* correspond to the traditional facial landmarks as typically used in the liter-

Table 1 Datasets for face alignment

Datasets	Year	Faces	Points
XM2VTS (Messer et al. 1999)	1999	2360	68
BioID (Jesorsky et al. 2001)	2001	1521	20
FRGC (Phillips et al. 2005)	2005	4950	68
PUT (Kasinski et al. 2008)	2007	9971	30
BUHMAP-DB (Aran et al. 2007)	2007	2880	52
MUCT (Milborrow et al. 2010)	2010	3755	76
Multi-PIE (Gross et al. 2010) (Semi-frontal)	2010	6665	68
Multi-PIE (Gross et al. 2010) (Profile)	2010	1400	39
LFW (Huang et al. 2008)	2007	13,233	10
AFLW (Köstinger et al. 2011)	2011	25,993	21
LFPW (Belhumeur et al. 2013)	2011	1432	29
AFW (Zhu and Ramanan 2012)	2012	205	6
HELEN (Le et al. 2012)	2012	2330	194
COFW (Burgos-Artizzu et al. 2013)	2013	1007	29
COFW (Ghiasi and Fowlkes 2015)	2015	507	68
300-W (Sagonas et al. 2013, 2016)	2013	3837	68
300VW (Shen et al. 2015)	2015	218k	68
MTFL (Zhang et al. 2014)	2014	12,995	5
MAFL (Zhang et al. 2016b)	2016	20,000	5
Menpo 2D (Zafeiriou et al. 2017c) (Semi-frontal)	2017	10,993	68
Menpo 2D (Zafeiriou et al. 2017c) (Profile)	2017	3852	39
AFLW2000-3D (Zhu et al. 2016c)	2016	2000	68
300W-LP (Zhu et al. 2016c)	2016	61,225	68
Menpo 3D (Zafeiriou et al. 2017a)	2017	11,971 + 280k	84

ature [e.g. in 300W (Sagonas et al. 2013, 2016) and 300VW (Shen et al. 2015) challenges]. They are suitable for poses that are frontal or relatively close to frontal (semi-frontal). This configuration consists of the 68 landmarks (Gross et al. 2010) depicted in Fig. 3a. In case of self-occlusions, these landmarks are placed on the face contour and are annotated along the visible face edges, as shown in Fig. 5a.

In case of views that are nearly profile, the traditional 2D landmarks are not suitable, because a large number of landmarks is self-occluded. Therefore, we use *Profile 2D landmarks*, which are especially designed for profile faces. This configuration consists of the 39 landmarks (Gross et al. 2010) depicted in Fig. 3b, c.

Even though the above landmark configurations correspond to semantically meaningful parts of the face, many of the landmarks are hardly associated with the real 3D geometry of the human face. On contrary, *3D landmarks* and *3DA-2D landmarks* correspond directly to the 3D structure of the human face. *3D landmarks* are defined as the 3D coordinates of the facial landmarks, therefore they bare information regarding the depth of the 3D face. In addition, *3DA-2D landmarks* are defined as the 2D projections of the 3D landmarks on the image plane, see Fig. 5b. The configuration of *3D landmarks* and *3DA-2D landmarks* consists of the 84 landmarks shown in Fig. 4, which is fixed independently from the facial pose. Compared to the 68 landmark configuration in 2D landmarks (semi-frontal), this configuration includes 16 additional landmarks on the facial contour, which correspond to a linear interpolation (in 3D space) of the original 17 landmarks on the facial contour (Fig. 5).

2.3 Creation of Ground-Truth Semi-frontal and Profile 2D Facial Landmarks

We created ground-truth Semi-frontal 2D and Profile 2D landmarks on the images of training and test sets of the Menpo 2D benchmark with the following procedures.

For semi-frontal images, the Semi-frontal 2D landmarks were extracted using a semi-automatic process similar to Sagonas et al. (2013). But instead of an Active Appearance Model (AAM) (Cootes et al. 2001), the method we used was the Mnemonic Descent Method (MDM) (Trigeorgis et al. 2016). In more detail, we trained the model of MDM on the 300W dataset (Sagonas et al. 2013) and then applied it on the images to localise the landmarks. Finally, the output landmarks were inspected and corrected manually using the

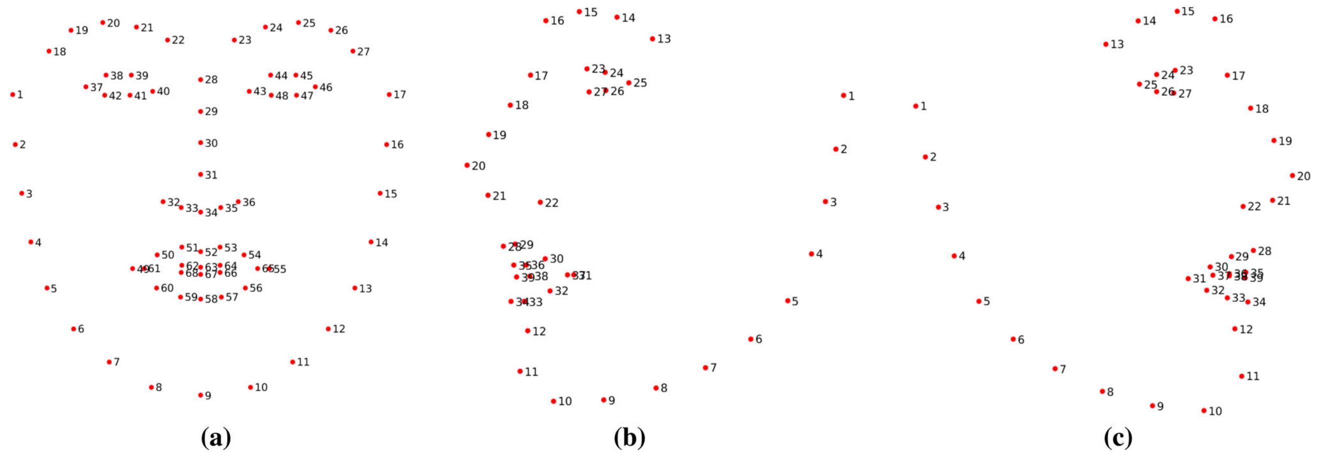


Fig. 3 Landmark configuration for the Menpo 2D benchmark. Semi-frontal face images are annotated by 68 landmarks, and profile face images are annotated by 39 landmarks. **a** Semi-frontal 2D landmarks (68), **b** left profile 2D landmarks (39), **c** right profile 2D landmarks (39)

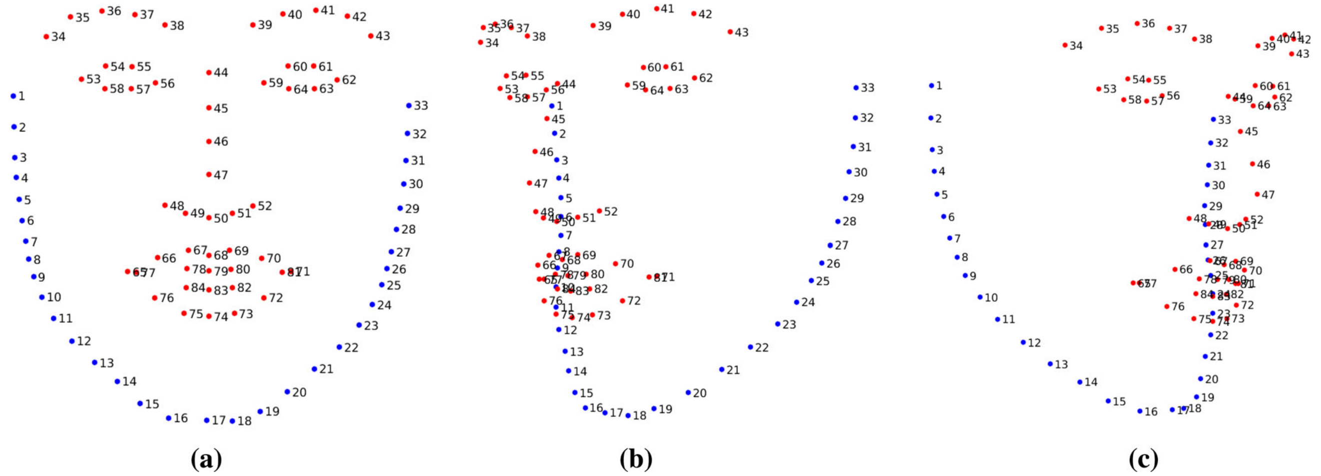


Fig. 4 Configuration for 3D and 3DA-2D landmarks, used in the Menpo 3D benchmark. The configuration includes 84 landmarks. It is independent from the facial pose and corresponds directly to the 3D structure of the human face. **a** Frontal pose, **b** left yaw pose, **c** right yaw pose

`landmarker.io`[2] annotation tool that was developed by our group, as shown in Fig. 6a.

Regarding profile face images, we manually annotated many images from scratch (around 1200), as there was no publicly available in-the-wild image dataset with profile annotations. Same as semi-frontal faces, we used these images to conduct a semi-automatic procedure to annotate the remaining profile images. In more detail, we trained a model of MDM on the annotated profile images. We applied it on the remaining images and then made manual corrections using `landmarker.io`, as illustrated in Fig. 6b, c.

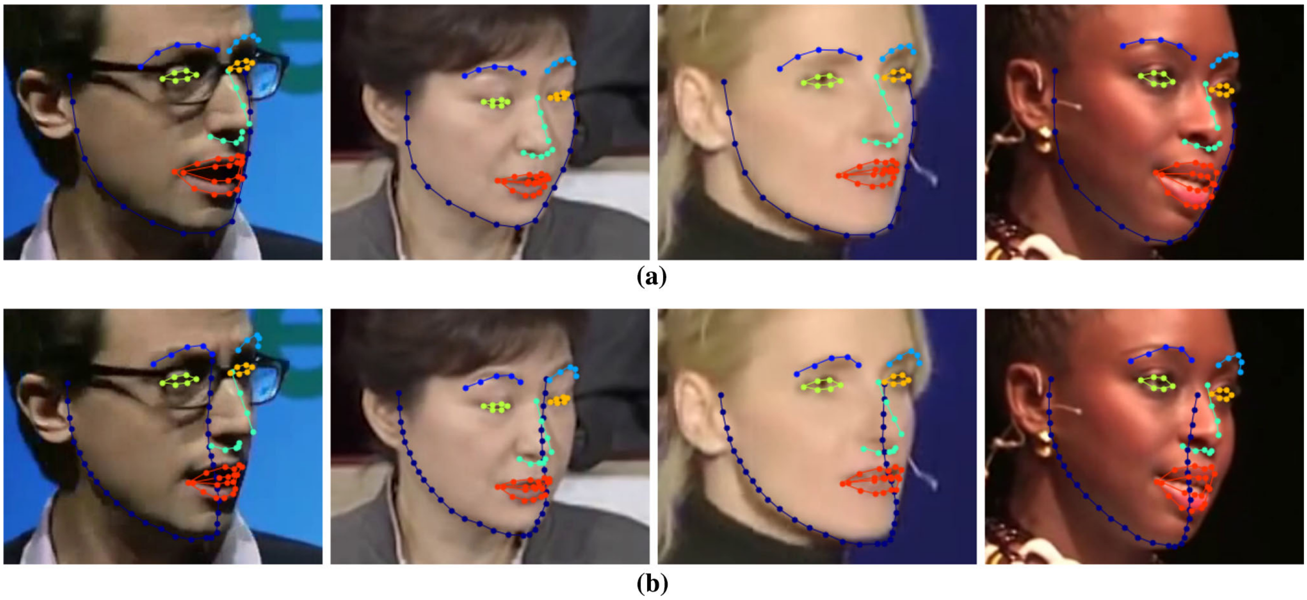
Finally, using the extracted ground-truth landmarks, the images were cropped in a region around the face and the cropped facial images were provided for training and testing.

2.4 Creation of Ground-Truth 3DA-2D and 3D Facial Landmarks

As already presented, the Menpo 3D benchmark consists of both images (used in the training set) and videos (used in training and test sets). In this section, we present the semi-automated procedure that we adopted to create ground-truth 3DA-2D and 3D landmarks on the images and videos of the Menpo 3D benchmark. These are based on fitting our LSFM models (Booth et al. 2016), the largest-scale 3D Morphable Model (3DMM) of faces, on videos and images. We focus our presentation on the case of videos, since this is the most challenging and interesting case for the Menpo 3D benchmark. Please note however that in Sect. 2.4.5, we also provide details for the case of images of the Menpo 3D benchmark.

The core steps of extracting accurate 3D landmarks from facial videos are depicted in Fig. 7. Initially, we employ a DCNN network (Deng et al. 2017) to estimate the per frame

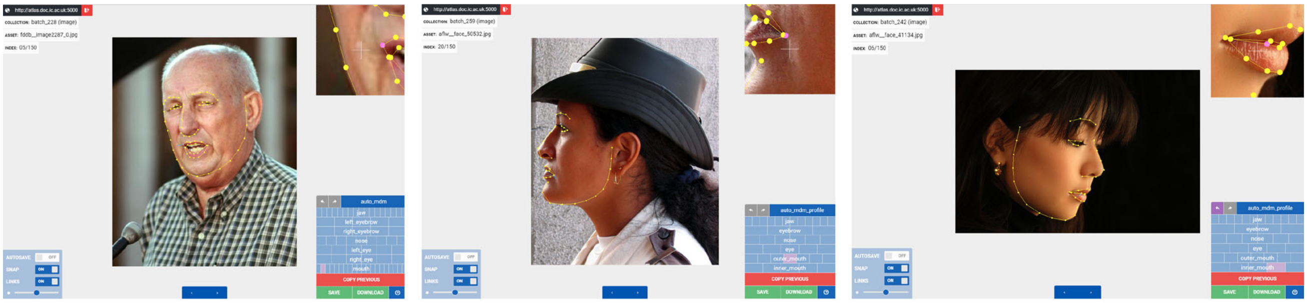
[2] https://www.landmarker.io/.



(a)

(b)

Fig. 5 Visual comparison of the **a** semi-frontal 2D landmarks (68) and **b** 3DA-2D landmarks (84), using examples from frames of videos of the 300VW dataset



http://atlas.doc.ic.ac.uk:5000
COLLECTION: batch_228 (image)
ASSET: fddb__image2287_0.jpg
INDEX: 05/150
AUTOSAVE OFF
SNAP ON
LINKS ON

(a)

http://atlas.doc.ic.ac.uk:5000
COLLECTION: batch_259 (image)
ASSET: aflw__face_50532.jpg
INDEX: 20/150
AUTOSAVE OFF
SNAP ON
LINKS ON

(b)

http://atlas.doc.ic.ac.uk:5000
COLLECTION: batch_242 (image)
ASSET: aflw__face_41134.jpg
INDEX: 06/150
AUTOSAVE OFF
SNAP ON
LINKS ON

(c)

Fig. 6 Examples of manual refinement of semi-frontal 2D (**a**) and profile 2D (**b**, c) landmarks using the `landmarker.io` annotation tool. **a** Semi-frontal 2D landmarks, **b** left profile 2D landmarks, **c** right profile 2D landmarks

3DA-2D landmarks. The automatic personalisation method of Chrysos et al. (2015) is utilised for refining certain facial parts (e.g. the eyes). Afterwards, 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. We then sample the dense facial mesh that is generated by the fitting result at every frame on the sparse landmark locations. Finally, we provide manual feedback by visually inspecting the results and keeping only those that are plausible across all frames of a video. More details follow in the subsequent sections.

2.4.1 Dense 3D Face Shape Modelling

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

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

where $\mathbf{x}_i = [x_i, y_i, z_i]^{\mathsf{T}}$ are the object-centred Cartesian coordinates of the ith 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 (Booth et al. 2016), which consist of the largest models of 3D Morphable Modelling (3DMM) of facial identity built from approximately 10,000

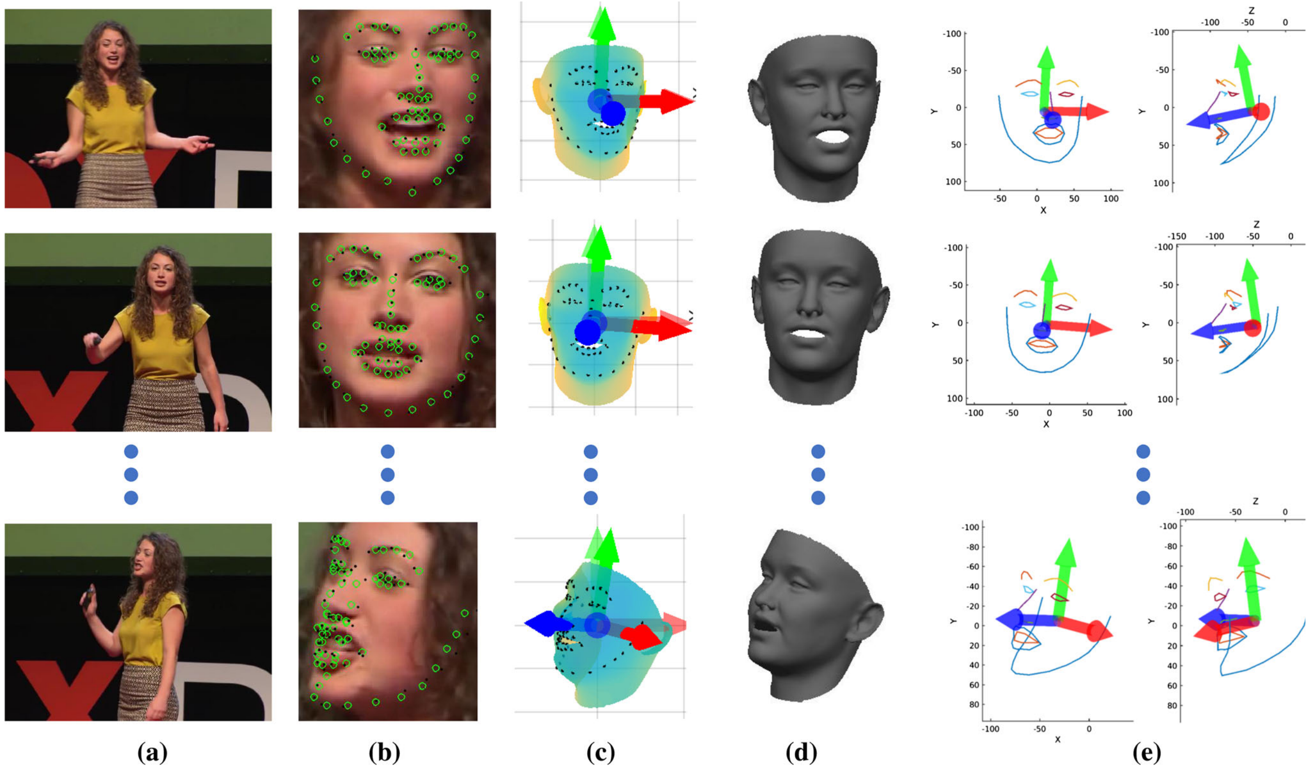


Fig. 7 Main steps of the adopted pipeline to create ground-truth 3D facial landmarks on videos. We are based on a state-of-the-art landmark localisation method and an energy minimisation approach to fitting powerful dense 3D face models on the sequence of landmarks. **a** Input video, **b** landmark localisation, **c** camera estimation (rigid SfM), **d** dense 3D shape estimation, **e** sampling of 3D shape on face landmarks

scans of different individuals. 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 a LSFM model, 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 $\{\bar{\mathbf{s}}_{id}, \mathbf{U}_{id}, \mathbf{6}_{id}\}$, 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 $\mathbf{6}_{id} \in \mathbb{R}^{n_p \times n_p}$ is a diagonal matrix with the standard deviations of the corresponding principal components. Let $\widetilde{\mathbf{U}}_{id} = \mathbf{U}_{id}\mathbf{6}_{id}$ 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 $\mathbf{p}$:

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

Furthermore, we also consider a 3D shape model of expression variations, as offsets from a given identity shape $\mathcal{S}_{id}$. The blend shapes model of Facewarehouse (Cao et al. 2014a) are utilised for this module. We adopt Nonrigid ICP (Cheng et al. 2017) to accurately register this model with the LSFM identity model. Then, the expression model can be represented with the triplet $\left\{\bar{\mathbf{s}}_{exp}, \mathbf{U}_{exp}, \mathbf{6}_{exp}\right\}$, 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 $\mathbf{6}_{exp} \in \mathbb{R}^{n_q \times n_q}$ is the diagonal matrix with the corresponding standard deviations. Similarly with the identity model, we consider the basis $\widetilde{\mathbf{U}}_{exp} = \mathbf{U}_{exp}\mathbf{6}_{exp}$ 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

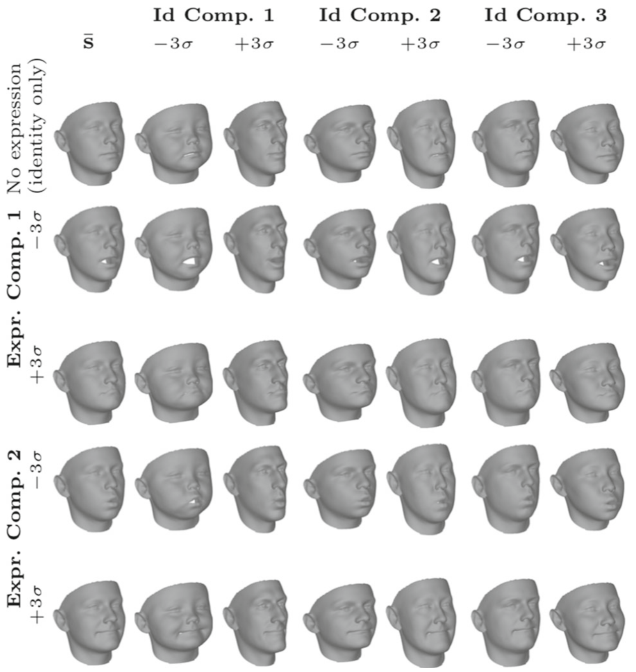


Fig. 8 Principal components of the LSFM global model under identity and expression variations, using the first 3 principal components for identity and the first 2 components for expression. Note that the first row corresponds to the identity model only

facial shape of any identity under any expression:

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

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 expression parameters. We construct one combined identity and expression model for each LSFM model (global or bespoke). For example, Fig. 8 visualises the first few components of identity and expression for the case of global LSFM model.

2.4.2 Camera Model

The purpose of a camera model is to project the object-centred 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:

$$\mathbf{v} = \left[v_x, v_y, v_z\right]^{\mathsf{T}} = \mathbf{R}_v\mathbf{x} + \mathbf{t}_v, \tag{4}$$

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 projection 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 $\mathbf{q} = [q_1, q_2, q_3]^{\mathsf{T}}$. 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. 3. Finally, we denote by $\mathcal{W}(\mathbf{p}, \mathbf{q}_f, \mathbf{c}_f) : \mathbb{R}^{n_p, n_c} \rightarrow \mathbb{R}^{2L}$, 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.

2.4.3 LSFM Fitting on Videos: 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 Bulat and Tzimiropoulos (2016). 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:

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

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 weight for the temporal smoothness term.

The **2D landmarks term** ($\hat{E}_{\text{land}}$) is a summation of the re-projection 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:

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

where the summation is done over all frames for which the 2nd derivative can be expressed without having to assume any form of padding outside the temporal window of the video. Also $\mathbf{D}^2 : \mathbb{R}^{n_q n_f} \rightarrow \mathbb{R}^{n_q(n_f-2)}$ 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, and it keeps the number of hyper-parameters as low as possible.

2.4.4 Optimisation of Energy Function

As described next, we first estimate the camera parameters $\hat{\mathbf{c}}$ (see Fig. 7c) and afterwards the shape parameters $(\mathbf{p}, \hat{\mathbf{q}})$ (see Fig. 7d).

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 minimised 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 $\widetilde{\mathbf{L}}_f$ $(f = 1, \ldots, n_f)$, which are the re-arrangements of the landmarks vectors $\tilde{\boldsymbol{\ell}}_f$ into $2 \times L$ matrices:

$$\widehat{\mathbf{L}} = \begin{bmatrix} \widetilde{\mathbf{L}}_1 \\ \vdots \\ \widetilde{\mathbf{L}}_{n_f} \end{bmatrix}, \ \widetilde{\mathbf{L}}_f = \begin{bmatrix} \tilde{x}_{1f} & \dots & \tilde{x}_{Lf} \\ \tilde{y}_{1f} & \dots & \tilde{y}_{Lf} \end{bmatrix}. \qquad (13)$$

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, $\hat{\mathbf{\Pi}} = \left[\mathbf{\Pi}_1^\mathsf{T} \dots \mathbf{\Pi}_{n_f}^\mathsf{T}\right]^\mathsf{T}$ is a $2n_f \times 3$ matrix that stacks the scaled orthographic projection matrices $\mathbf{\Pi}_f \in \mathbb{R}^{2\times 3}$ from all the frames f. The matrix $\mathbf{\Pi}_f$ 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 $\mathbf{\Pi}_f$:

$$\mathbf{\Pi}_f \mathbf{\Pi}_f^\mathsf{T} = \sigma_f^2 \mathbf{I}_2, \ \text{for some } \sigma_f > 0, \ f = 1, \dots, n_f. \qquad (14)$$

To summarise, our goal is to minimise $\hat{E}_{\text{land}}$ as described in Eq. (12) with respect to $\mathcal{S}_{\text{rig}}$ and $\hat{\mathbf{\Pi}}$, under the constraints of Eq. (14). For this, we employ a simple yet effective rigid Structure from Motion (SfM) approach (Webb and Aggarwal 1982): 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 as 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 $\mathbf{c}_1, \dots, \mathbf{c}_{n_f}$. 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:

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

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*:

$$\begin{aligned} &|(\mathbf{p})_i| \le M_p, \quad i = 1, \dots, n_p, \\ &\left|(\mathbf{q}_f)_i\right| \le M_q, \quad i = 1, \dots, n_q \text{ and } f = 1, \dots, n_f, \end{aligned} \qquad (16)$$

where $(\cdot)_i$ denotes the selection of the ith 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 do 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 $argmin_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, 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 Coleman and Li (1996).

2.4.5 LSFM Fitting on Images

For the images of the Menpo 3D benchmark, we are based on the "in-the-wild" 3DMM fitting strategy that we recently introduced (Booth et al. 2017). Our procedure is similar to the work of Zhu et al. (2016c), where a facial 3DMM is fitted on the 2D landmarks and used in order to train a DCNN for the estimation of the 3D facial surface. However, in contrast to Zhu et al. (2016c), our 3DMM fitting strategy not only uses the facial landmarks but also the facial texture. Furthermore, in order to improve accuracy we annotated every image with regards to (a) gender, (b) ethnicity and (c) apparent age and used the corresponding bespoke LSFM model (Booth et al. 2016) for fitting. In conclusion, this procedure yielded estimated identity **p** and expression **q** parameters of the bespoke LSFM model for every image.

2.4.6 Facial Landmark Sampling and Re-projection for Images and Videos

After having estimated the shape parameters $(\mathbf{p}, \mathbf{q})$ for any image or any video frame from the Menpo 3D dataset, the estimated dense facial mesh in the model space can be synthesised by the model as $\mathcal{S}(\mathbf{p}, \mathbf{q}) = \bar{\mathbf{s}} + \widetilde{\mathbf{U}}_{id}\mathbf{p} + \widetilde{\mathbf{U}}_{exp}\mathbf{q}$. The ground-truth 3D landmarks $\mathcal{S}^{\ell}$ are then extracted by keeping the elements of $\mathcal{S}$ 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 **c** and apply the camera function $\mathcal{P}(\cdot)$ to $\mathcal{S}^{\ell}$. This corresponds to the quantity $\mathcal{W}(\mathbf{p}, \mathbf{q}, \mathbf{c})$.

Via visual inspection of both the dense 3D and the reprojected sparse 2D landmarks results, we choose the best of the two results (global versus bespoke identity models). Finally, in order to retain this result as ground-truth, we require that it is perfect in terms of visual plausibility, otherwise, we omit it. In the case of videos, we make sure that this requirement is met over all video frames.

In conclusion, following our proposed approach, we are able to extract high-quality ground-truth 3DA-2D and 3D facial landmarks for images and videos. We have tested our approach in simulated videos and it provided high accuracy (sub-millimetre accuracy for some landmarks). Furthermore, in the videos of both training and test set, 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.

3 Menpo 2D Challenge

Having introduced the Menpo 2D Benchmark, this section provides a detailed presentation of the Menpo 2D Challenge, in terms of the evaluation metrics used, the participants and the results of the challenge. Finally, we conclude this section by presenting our proposed method for 2D landmark localisation, which was not included in the challenge since we were the organisers. Please note that we organised the Menpo 2D Challenge in conjunction with CVPR 2017 conference.

The Menpo 2D Challenge consisted of two categories:

- localisation of Semi-frontal 2D landmarks in semi-frontal facial images.
- localisation of Profile 2D landmarks in profile facial images.

3.1 Evaluation Metrics

The standard evaluation metric for landmark-wise face alignment is the normalised point-to-point root mean square (RMS) error:

$$\epsilon(\hat{\mathbf{s}}, \mathbf{s}^*) = \frac{\frac{1}{\sqrt{N_L}}\|\hat{\mathbf{s}} - \mathbf{s}^*\|_2}{d_{\text{scale}}}, \tag{17}$$

where $\hat{\mathbf{s}}$ and $\mathbf{s}^*$ are the estimated and ground-truth shape vectors containing all N_L facial landmarks, $\|.\|_2$ is the ℓ_2 norm, and d_{scale} is a normalisation factor to make the error scale-invariant. For the last three face alignment competitions (Sagonas et al. 2013, 2016; Shen et al. 2015), the inter-ocular distance was used as the normalisation factor. However, the inter-ocular distance fails to give a meaningful evaluation metric in the case of profile views as it becomes a very small value in the 2D image. Therefore, we employed the face diagonal as the normalisation factor, which not only achieves scale-invariance but also is more robust to changes of the face pose. This is defined as the length of the diagonal of the bounding box that that tightly surrounds the ground-truth landmarks $\mathbf{s}^*$. It can be mathematically written as:

$$d_{\text{scale}} = \left\| \left(\max_i(x_i^*) - \min_i(x_i^*)\,,\ \max_i(y_i^*) - \min_i(y_i^*) \right) \right\|_2, \tag{18}$$

where x_i^* and y_i^* are the x and y coordinates respectively of the ith ground-truth landmark and the maxima and minima are computed over all landmarks i.

Many works on the topic (Ren et al. 2014) report just the average of the error in (Eq. 17). We believe that mean errors, particularly without accompanying standard deviations, are not a very informative error metric as they can be highly biased by a low number of very poor fits. Therefore, we provide our evaluation in the form of Cumulative Error Distribution (CED) curves. Apart from the mean value, we are also reporting further statistics of the errors of each method, such as the Standard Deviation (Std), the Median, the Median Absolute Deviation (MAD), the Maximum Error, the area-under-the-curve (AUC) (up to error of 0.05) measure of the CED curve and the Failure Rate (We consider any fitting with a point-to-point error greater than 0.05 as a failure). We believe that for the problem of face alignment, these errors metrics are much more representative and provide a much more clear image of the performance of each method.

3.2 Participants

In the Menpo 2D challenge, we allowed entries from the participants in either semi-frontal or profile challenge (i.e., they do not need to submit in both challenges to be considered eligible). We provided the training data accompanied by the corresponding landmark annotations around 30th of January 2017. The test data were released around 22nd of March 2017 and included only the facial images without the corresponding annotations. Furthermore, we provided information regarding whether every image was considered semi-frontal or profile. The participants were allowed to submit results (i.e., the facial landmarks) in a tight schedule after the release of the test data, up until the 31st of March 2017. After this date, the challenge was considered finished. In total, we had 9 participants to the challenge of semi-frontal faces and 8 participants to that of profile faces. In the following, we will briefly describe each participating method. We use an abbreviation based on the name of the first author of the paper.

- *X. Chen* The method in Chen et al. (2017) proposed a four-stage coarse-to-fine framework to tackle the facial landmark localisation problem in-the-wild. In the first stage, they predict the facial landmarks on a coarse level, which sets a good initialisation for the whole framework. Then, the key points are grouped into several components and each component is refined within the local patch. After that, each key point is further refined with multi-scale local patches cropped according to its nearest 3-, 5-, and 7-neighbours, respectively. The results are further fused by an attention gate network. Since the facial landmark configuration is different for semi-frontal and profile faces in the menpo 2D challenge, a linear transformation is finally learned with the least square approximation to adapt the predictions to the competition's subsets.
- *X. Shao* The method in Shao et al. (2017) used a deep architecture to directly detect facial landmarks without using face detection as an initialisation. The architecture consists of two stages, a basic landmark prediction stage and a whole landmark regression stage. At the former stage, given an input image, the basic landmarks of all faces are detected by a sub-network of landmark heatmap and affinity field prediction. At the latter stage, the coarse canonical face and the pose are generated by a Pose Splitting Layer based on the visible basic landmarks. According to its pose, each canonical state is distributed to the corresponding branch of the shape regression sub-networks for the whole landmark detection.
- *Z. He* The method in He et al. (2017) proposed an effective facial landmark detection system, recorded as Robust Fully End-to-end Cascaded Convolutional Neural Network (RFEC-CNN), to characterise the complex non-linearity from face appearance to shape. Moreover, a face bounding box invariant technique is adopted to reduce the landmark localisation sensitivity to the face detector while a model ensemble strategy is adopted to further enhance the landmark localisation performance.
- *Z. Feng* The method in Feng et al. (2017) presented a four-stage framework (face detection, bounding box aggregation, pose estimation and landmark localisation) for robust face detection and landmark localisation in the wild. To achieve a high detection rate, two publicly available CNN-based face detectors and two proprietary detectors are employed. Then, the detected face bounding boxes of each input image are aggregated to reduce false positives and improve face detection accuracy. After that, a cascaded shape regressor, trained using faces with a variety of pose variations, is then employed for pose estimation and image pre-processing. Finally, another cascaded shape regressor is trained for fine-grained landmark localisation, using a large number of training samples with limited pose variations.
- *J. Yang* The method in Yang et al. (2017) explored a two-stage CNN model for robust facial landmark localisation. First, a supervised face transformation network is adopted to remove the translation, scale and rotation variation of each face, in order to reduce the variance of the regression target. Then, a deep convolutional neural network named Stacked Hourglass Network (Newell et al. 2016) is explored to increase the capacity of the regression model.
- *M. Kowalski* The method in Kowalski et al. (2017) used a VGG-based Deep Alignment Network (DAN) for robust face alignment. This method uses entire face images at all stages, contrary to the recently proposed face alignment methods that rely on local patches. The use of entire face images rather than patches allows DAN to handle face images with large variation in head pose and difficult initialisation. DAN consists of multiple stages, where each stage improves the locations of the facial landmarks estimated by the previous stage.

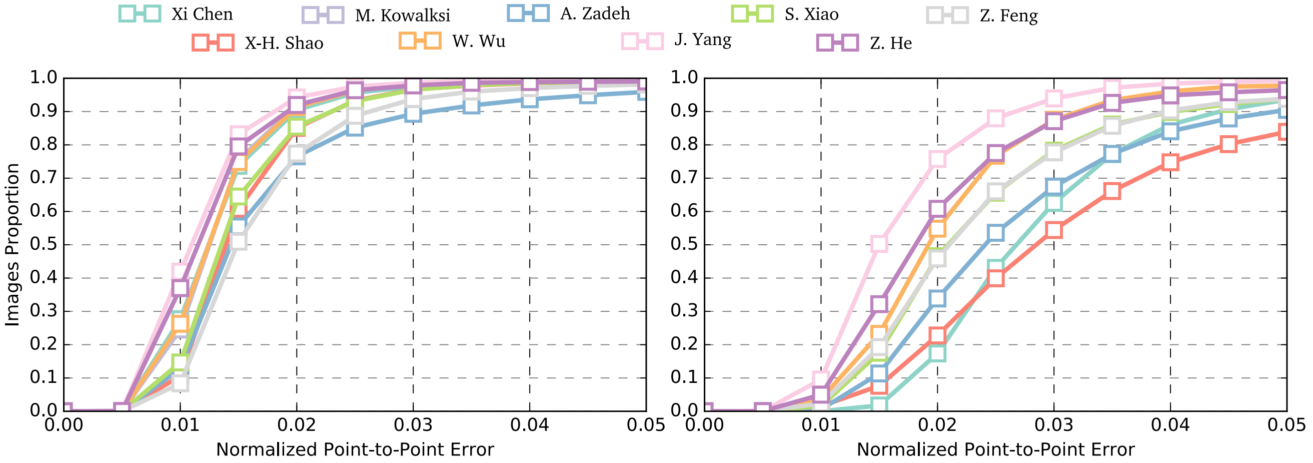


Fig. 9 Quantitative results (CED curves) on the test set of the Menpo 2D Benchmark for both semi-frontal (left) and profile (right) results

Table 2 Key statistics of the performance of the participants on semi-frontal faces (68-points markup)

Methods	Mean	Std	Median	MAD	Max error	AUC$_{0.05}$	Failure rate
Yang et al. (2017)	0.0120	0.0060	0.0107	0.0022	0.1453	**0.7624**	0.0024
He et al. (2017)	0.0139	0.0260	0.0111	0.0023	0.9624	0.7478	0.0096
Wu and Yang (2017)	0.0135	0.0095	0.0120	0.0024	0.5098	0.7337	0.0036
Kowalski et al. (2017)	0.0138	0.0157	0.0120	0.0023	0.6312	0.7337	0.0049
Chen et al. (2017)	0.0200	0.0756	0.0120	0.0026	1.2799	0.7290	0.0111
Xiao et al. (2017)	0.0159	0.0201	0.0133	0.0027	0.6717	0.6986	0.0081
Shao et al. (2017)	0.0165	0.0235	0.0138	0.0027	0.9612	0.6913	0.0101
Feng et al. (2017)	0.0182	0.0179	0.0149	0.0033	0.4661	0.6586	0.0186
Zadeh et al. (2017a)	0.0205	0.0340	0.0143	0.0035	0.9467	0.6479	0.0409

Bold value indicates best result

- *A. Zadeh* The method in Zadeh et al. (2017a) used a novel local detector, Convolutional Experts Network (CEN), in the framework of Constrained Local Model (CLM) for face alignment in the wild. This method brings together the advantages of deep neural architectures and mixtures of experts in an end-to-end framework.
- *S. Xiao* The method in Xiao et al. (2017) proposed a novel 3D-assisted coarse-to-fine extreme-pose facial landmark detection system. For a given face image, the face bounding box is first refined with landmark locations inferred from a 3D face model generated by a Recurrent 3D Regressor (R3R) at a coarse level. Then, another R3R is employed to fit a 3D face model onto the 2D face image cropped with the refined bounding box at fine-scale. 2D landmark locations inferred from the fitted 3D face are further adjusted with the popular 2D regression method, i.e. LBF (Ren et al. 2014). The 3D-assisted coarse-to-fine strategy and the 2D adjustment process explicitly ensure both the robustness to extreme face poses and bounding box disturbance and the accuracy towards pixel-level landmark displacement.
- *W. Wu* The method in Wu and Yang (2017) explored intra-dataset variation and inter-dataset variation to improve face alignment in-the-wild. Intra-dataset variation refers to bias in expression and head pose inside one certain dataset, while inter-dataset variation refers to different bias across different datasets. Model robustness can be significantly improved by leveraging rich variations within and between different datasets. More specifically, Wu and Yang (2017) proposed a novel Deep Variation Leveraging Network (DVLN), which consists of two strong coupling sub-networks, e.g., Dataset-Across Network (DA-Net) and Candidate-Decision Network (CD-Net). In particular, DA-Net takes advantage of different characteristics and distributions across different datasets, while CD-Net makes a final decision on candidate hypotheses given by DA-Net to leverage variations within one certain dataset.

3.3 Competition Results

The CED curves of all 68 landmarks for semi-frontal images and all 39 landmarks for profile images are shown in Fig. 9.

Table 3 Key statistics of the performance of the participants on profile faces (39-points markup)

Methods	Mean	Std	Median	MAD	Max error	$AUC_{0.05}$	Failure rate
Yang et al. (2017)	0.0172	0.0105	0.0150	0.0035	0.2490	**0.6613**	0.0077
He et al. (2017)	0.0247	0.0422	0.0179	0.0048	0.6280	0.5932	0.0355
Wu and Yang (2017)	0.0217	0.0131	0.0193	0.0044	0.2623	0.5802	0.0221
Feng et al. (2017)	0.0285	0.0367	0.0208	0.0057	0.4725	0.5268	0.0617
Xiao et al. (2017)	0.0290	0.0417	0.0209	0.0055	0.6327	0.5237	0.0612
Zadeh et al. (2017a)	0.0375	0.0630	0.0241	0.0071	0.7594	0.4604	0.0951
Chen et al. (2017)	0.0448	0.1162	0.0265	0.0058	1.3698	0.4259	0.0642
Shao et al. (2017)	0.0451	0.0636	0.0282	0.0088	0.7534	0.3891	0.1608

Bold value indicates best result

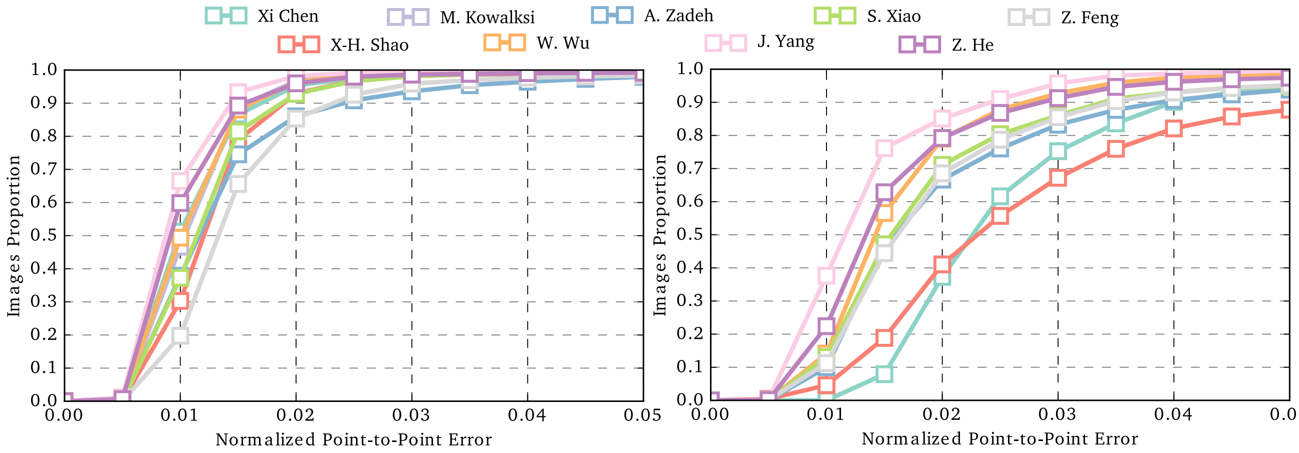


Fig. 10 Quantitative results (CED curves) on the interior landmarks of the test set of the Menpo 2D Benchmark for both semi-frontal (49-points) (left) and profile (28-points) (right) facial images

The key error statistics of the CED curves for the semi-frontal faces are summarised in Table 2, while the key statistics for the profile faces are summarised in Table 3.

From the statistics and the curves, it is evident that in the category of semi-frontal faces the first three entries (Yang et al. 2017; He et al. 2017; Wu and Yang 2017) were quite close. Nevertheless, in the category of the profile faces, the method of Yang et al. (2017) was the clear winner. Overall, the method of Yang et al. (2017) was the best performing method in both semi-frontal and profile categories and is the winner of the competition.

As it is customary in landmark evaluation papers (Sagonas et al. 2013, 2016), we also provide performance graphs excluding the boundary landmarks for both semi-frontal and profile faces. This corresponds to markups of 49 interior Semi-frontal 2D landmarks and 28 interior Profile 2D landmarks. The CED curves of these landmarks in semi-frontal and profile categories are shown in Fig. 10. The corresponding key error statistics are summarised in Tables 4 and 5. We observe that the method of Yang et al. (2017) is still the best performing method.

3.4 A New Strong Baseline for 2D Face Alignment

Since we organised the competition, we could not submit an entry. Nevertheless, we have applied our recent method for face alignment in the wild (Deng et al. 2017) on the Menpo 2D Benchmark. The method is based on hourglass architectures and multi-view training. Interestingly, despite our method is relatively simple, it achieves an improved accuracy on the Menpo 2D Benchmark. Therefore, we consider it as a new strong baseline. In this section, we briefly explain the method and present its experimental results on this benchmark.

As an overview, our work in Deng et al. (2017) proposes a coarse-to-fine joint multi-view deformable face fitting method. The proposed method includes two steps: face region normalisation and multi-view face alignment. Based on face proposals (Zhang et al. 2016a), we train a small network to estimate five landmarks which are then used to remove the similarity transformation. Then, a multi-view Hourglass model is trained to predict the response maps for all landmarks (both 68 landmarks of semi-frontal markup, as well as 39 landmarks of the profile markup). Finally, we train a

Table 4 Key statistics of the performance of the participants on the interior landmarks of semi-frontal faces (49-points markup)

Methods	Mean	Std	Median	MAD	Max error	AUC$_{0.05}$	Failure rate
Yang et al. (2017)	0.0097	0.0053	0.0087	0.0017	0.1719	**0.8084**	0.0022
He et al. (2017)	0.0117	0.0253	0.0093	0.0019	0.9520	0.7886	0.0079
Wu and Yang (2017)	0.0113	0.0085	0.0101	0.0019	0.4752	0.7778	0.0024
Kowalski et al. (2017)	0.0116	0.0147	0.0102	0.0018	0.6720	0.7765	0.0036
Chen et al. (2017)	0.0174	0.0724	0.0099	0.0021	1.2699	0.7746	0.0096
Xiao et al. (2017)	0.0132	0.0188	0.0110	0.0022	0.6411	0.7513	0.0066
Shao et al. (2017)	0.0139	0.0220	0.0115	0.0022	0.9590	0.7420	0.0084
Zadeh et al. (2017a)	0.0162	0.0319	0.0111	0.0026	0.9377	0.7200	0.0204
Feng et al. (2017)	0.0159	0.0164	0.0129	0.0029	0.3686	0.7007	0.0161

Bold value indicates best result

Table 5 Key statistics of the performance of the participants on the interior landmarks of profile faces (28-points markup)

Methods	Mean	Std	Median	MAD	Max error	AUC$_{0.05}$	Failure rate
Yang et al. (2017)	0.0136	0.0093	0.0110	0.0026	0.2162	**0.7319**	0.0036
He et al. (2017)	0.0201	0.0414	0.0132	0.0035	0.6380	0.6778	0.0257
Wu and Yang (2017)	0.0168	0.0109	0.0142	0.0034	0.2252	0.6709	0.0128
Xiao et al. (2017)	0.0233	0.0416	0.0154	0.0042	0.7073	0.6231	0.0509
Feng et al. (2017)	0.0236	0.0361	0.0161	0.0046	0.5141	0.6124	0.0483
Zadeh et al. (2017a)	0.0293	0.0632	0.0157	0.0046	0.8780	0.5990	0.0617
Chen et al. (2017)	0.0409	0.1181	0.0223	0.0051	1.3809	0.4954	0.0493
Shao et al. (2017)	0.0388	0.0636	0.0228	0.0079	0.7769	0.4756	0.1223

Bold value indicates best result

failure checker network based on the shape-indexed feature which is extracted around each landmark predicted by the multi-view Hourglass heatmap. Further details follow.

3.4.1 Face Region Normalisation

Taking the first and second networks from Zhang et al. (2016a), we can get face proposals with high recall. Based on these face proposals, we re-train the third network on the ALFW (Köstinger et al. 2011) and CelebA (Liu et al. 2015) datasets with multi-task loss (Zhang et al. 2016a). For each face box i, the loss function is defined as:

$$L = L_1(p_i, p_i^*) + \lambda_1 p_i^* L_2(t_i, t_i^*) + \lambda_2 p_i^* L_3(l_i, l_i^*), \qquad (19)$$

where p_i is the probability of box i being a face; p_i^* is a binary indicator (1 for positive and 0 for negative examples); the classification loss L_1 is the softmax loss of two classes (face / non-face); $t_i = \{t_x, t_y, t_w, t_h\}_i$ and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}_i$ represent the coordinates of the predicted box and ground-truth box correspondingly. $l_i = \{l_{x_1}, l_{y_1}, \ldots, l_{x_5}, l_{y_5}\}_i$ and $l_i^* = \{l_{x_1}^*, l_{y_1}^*, \ldots, l_{x_5}^*, l_{y_5}^*\}_i$ represent the predicted and ground-truth five facial landmarks. The box and the landmark regression targets are normalised by the face size of the ground-truth. We use $L_2(t_i, t_i^*) = R(t_i - t_i^*)$ and $L_3(l_i, l_i^*) = Rv_i^*(l_i - l_i^*)$ for the box and landmark regression loss, respectively, where R is the robust loss function (smooth-L$_1$) defined in Girshick (2015). In Fig. 11, we give the network structure of the third cascade network with multi-task loss.

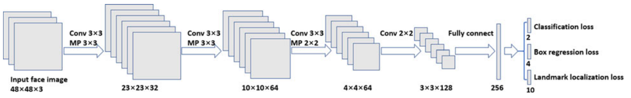


Fig. 11 The architecture of third cascade network. “Conv” means convolution, “MP” means max pooling, and N is the number of landmarks. The step size in convolution and pooling is 1 and 2, respectively

One core idea of our method is to incorporate a spatial transformation (Jaderberg et al. 2015) which is responsible for warping the original image into a canonical representation such that the later alignment task is simplified. Recent work [e.g., (Tadmor et al. 2016)] has explored this idea on face recognition and witnessed an improvement on the performance. In Fig. 12, the five facial landmark localisation network (Fig. 11) as the spatial transform layer is trained to map the original image to the parameters of a warping function (e.g., a similarity transform), such that the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations. Since different training data are used in face region normalisation [AFLW (Köstinger et al. 2011) and CelebA (Liu et al. 2015)] and multi-view align-

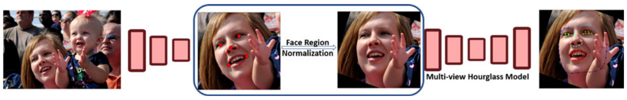


Fig. 12 Face Region Normalisation. The five facial landmark localisation network acts as the spatial transform layer and the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations

ment [300W (Sagonas et al. 2013) and Menpo 2D Benchmark (Zafeiriou et al. 2017c)], end-to-end training of these two networks with intermediate supervision on the face region normalisation step is equal to step-wise training. Therefore, we employ step-wise cascade structure, and the face region normalisation step benefits from larger training data as annotation of the five facial landmarks is much easier than dense annotation.

3.4.2 Multi-view Hourglass Model

Hourglass (Newell et al. 2016) is designed based on Residual blocks (He et al. 2016), which can be represented as follows:

$$x_{n+1} = H(x_n) + F(x_n, W_n), \tag{20}$$

where x_n and x_{n+1} are the input and output of the n-th unit, and F is the stacked convolution, batch normalisation, and ReLU non-linearity. Hourglass is a symmetric top-down and bottom-up full convolutional network. The original signals are branched out before each down-sampling step and combined together before each up-sampling step to keep the resolution information. n scale Hourglass is able to extract features from the original scale to $1/2^n$ scale and there is no resolution loss in the whole network. The increasing depth of network design helps to increase contextual region, which incorporates global shape inference and increases robustness when local observation is blurred.

Based on the Hourglass model (Newell et al. 2016), we formulate the Multi-view Hourglass Model (MHM) which tries to jointly estimate both semi-frontal (68 landmarks) and profile (39 landmarks) face shapes. Unlike other methods which employ distinct models, we try to capitalise on the correspondences between the profile and frontal facial shapes. As shown in Fig. 13, for each landmark on the profile face, the nearest landmark on the frontal face is regarded as its corresponding landmark in the union set, thus we can form the union landmark set with 68 landmarks. During the training, we use the view status to select the corresponding response maps for the loss computation.

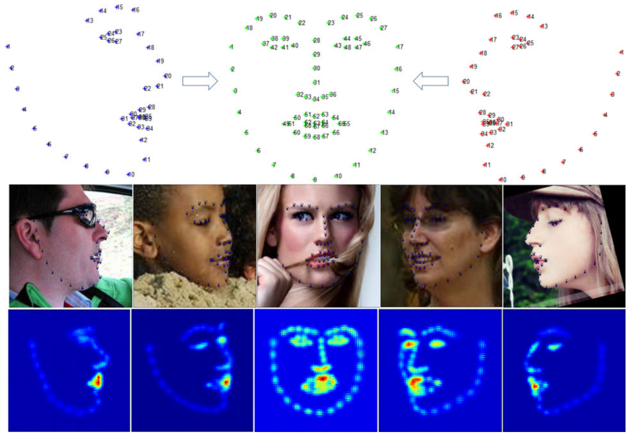


Fig. 13 Multi-view Hourglass Model. First row: facial landmark configuration for frontal (68 landmarks) and profile (39 landmarks) faces (Zafeiriou et al. 2017c). We define a union landmark set with 68 landmarks for frontal and profile shape. For each landmark on the profile face, the nearest landmark on the frontal face is selected as the same definition in the union set. Third row: landmark response maps for all view faces. The response maps for semi-frontal faces (2nd and 4th) benefit from the joint multi-view training

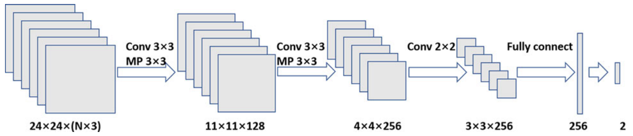


Fig. 14 The architecture of face classifier on the shape-indexed local patches. "Conv" means convolution, "MP" means max pooling, and N is the landmark number. The step size in convolution and pooling is 1 and 2, respectively

$$L = \frac{1}{N} \sum_{n=1}^{N} \left(v_n^* \sum_{ij} \left\| m_n(i,j) - m_n^*(i,j) \right\|_2^2 \right), \tag{21}$$

where $m_n(i, j)$ and $m_n^*(i, j)$ represent the estimated and the ground-truth response maps at pixel location (i, j) for the n-th landmark correspondingly, and $v_n \in \{0, 1\}$ is the indicator to select the corresponding response map to calculate the final loss. We can see from Fig. 13 that the semi-frontal response maps (second and forth examples in third row) benefit from the joint multi-view training, and the proposed method is robust and stable under large pose variations.

Based on the multi-view response maps, we extract shape-indexed patch (24 × 24) around each predicted landmark from the down-sampled face image (128 × 128). As shown in Fig. 14, a small classification network is trained to classify face/non-face. This classifier is employed as a failure checker for deformable face tracking.

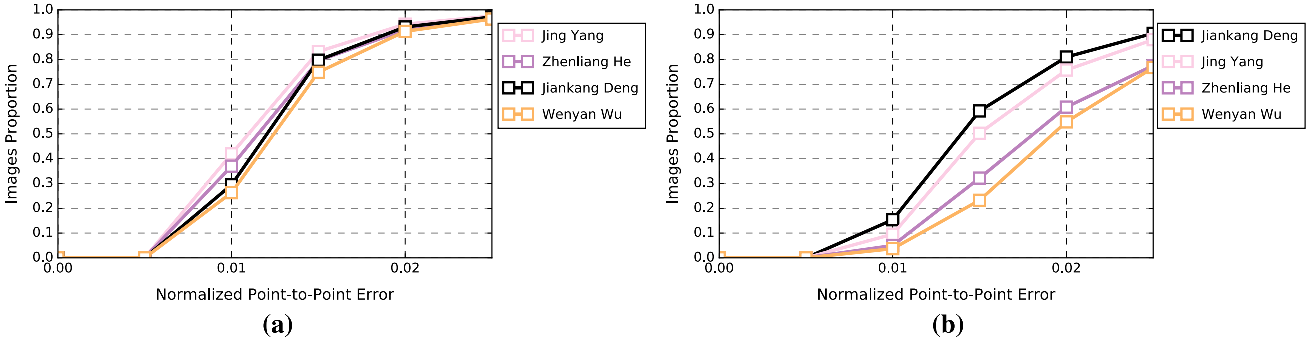


Fig. 15 Landmark localisation results on the Menpo 2D benchmark: comparison of our method (Deng et al. 2017) with the best three entries of Menpo 2D Challenge. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalised with the diagonal of the ground-truth bounding box

3.4.3 Experimental Results

We train the proposed multi-view Hourglass model on the 300W dataset (Sagonas et al. 2013) and the Menpo 2D dataset (Zafeiriou et al. 2017c), where faces are manually annotated with either Semi-frontal 2D (68) or Profile 2D (39) landmarks. The training set of the 300W dataset consists of the LFPW trainset (Belhumeur et al. 2013), the Helen trainset (Le et al. 2012) and the AFW dataset (Zhu and Ramanan 2012), hence, a total of 3148 images are available. The Menpo 2D dataset (Zafeiriou et al. 2017c) consists of 5658 semi-frontal face images and 1906 profile face images.

The training of the proposed multi-view Hourglass model follows a similar design as in the Hourglass Model (Newell et al. 2016). Before the training, several pre-processing steps are undertaken. We firstly remove scale, rotation and translation differences by five facial landmarks among the training face images (referred as the spatial transformer step), then crop and resize the face regions to 256 × 256. We augment the data with rotation (+/− 30°), scaling (0.75–1.25), and translation (+/− 20 pixels) that would help simulate the variations from face detector and five landmark localisation. The full network starts with a 7 × 7 convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64, as it could save GPU memory while preserving alignment accuracy. The network is trained using Tensorflow (Abadi et al. 2016) with an initial learning rate of 10^{-4}, a batch size of 12, and learning steps of 100k. The Mean Squared Error (MSE) loss is applied to compare the predicted heatmaps to the ground-truth heatmaps. Each training step takes 1.2 s on one NVIDIA GTX Titan X (Pascal) GPU card. During testing, face regions are cropped and resized to 256 × 256, and it takes 12.21 ms to generate the response maps.

On the test set of **Menpo 2D Benchmark** (Zafeiriou et al. 2017c), we compare our method with the best three entries (Yang et al. 2017; He et al. 2017; Wu and Yang 2017) of the Menpo 2D Challenge. In Fig. 15, we draw the curve of cumulative error distribution on semi-frontal and profile test sets, separately. The proposed method has similar performance to the best-performing methods in semi-frontal faces. Nevertheless, it outperforms the best-performing method in profile faces. Despite that result on profile data is worse than that on semi-frontal data, both of the fitting errors of our method are remarkably small, approaching 1.48% and 1.27% for profile and semi-frontal faces respectively. In Fig. 16, we give some fitting examples on the Menpo 2D test set. As we can see from the alignment results, the proposed multi-view hourglass model is robust under varying poses, exaggerated expressions, heavy occlusions and sharp illuminations on both semi-frontal and profile subset.

In Fig. 17, we also provide some alignment examples with largest errors predicted by the proposed method on the Menpo 2D semi-frontal and profile dataset. As we can see from these extremely challenging examples, most of the landmark localisation failures occur when local facial appearance is occluded thus the local feature is heavily interfered. When two or more challenging factors (e.g. large pose, exaggerated expression, heavy occlusion and sharp illumination) occur together, the localisation results turn out to be not robust and accurate, and there is still space to improve the performance of the in-the-wild 2D face alignment.

3.5 Development of 2D Face Alignment

It is worth mentioning that, in the first (Sagonas et al. 2013) and second (Sagonas et al. 2016) runs of 300W competition, there were very few competing methods (Zhou et al. 2013; Fan and Zhou 2016) that applied deep learning methods to the problem of face alignment. The state of the art at that time was revolving around feature-based Active Shape Model (ASM) (Milborrow and Nicolls 2014), Active Appearance Model (AAM) (Antonakos et al. 2015; Alabort-i Medina and Zafeiriou 2016) and Constrained Local Model (CLM) (Cristinacce and Cootes 2006; Saragih et al. 2011), as well

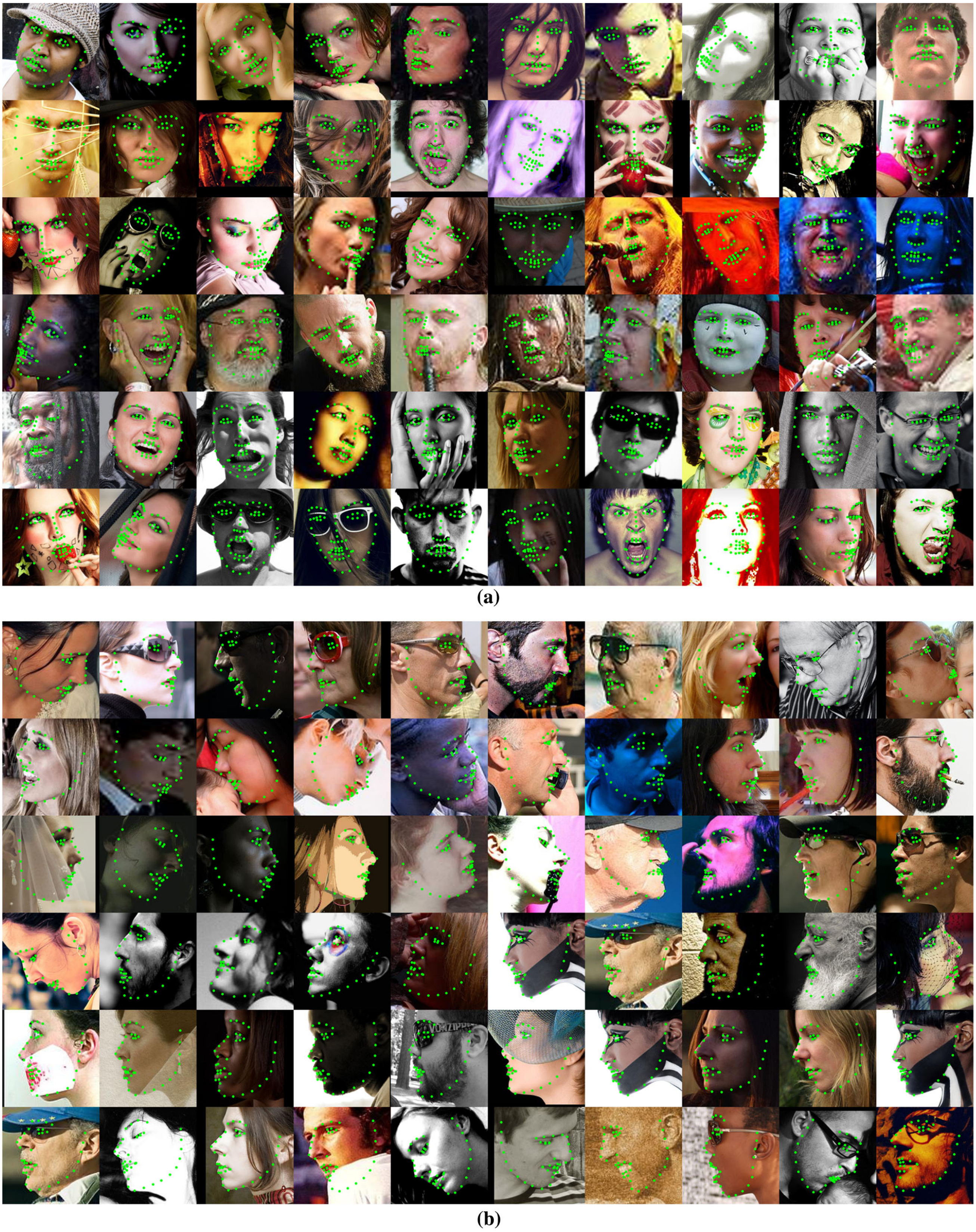


Fig. 16 Example landmark localisation results on the test set of the Menpo 2D benchmark. **a** Menpo semi-frontal, **b** Menpo profile

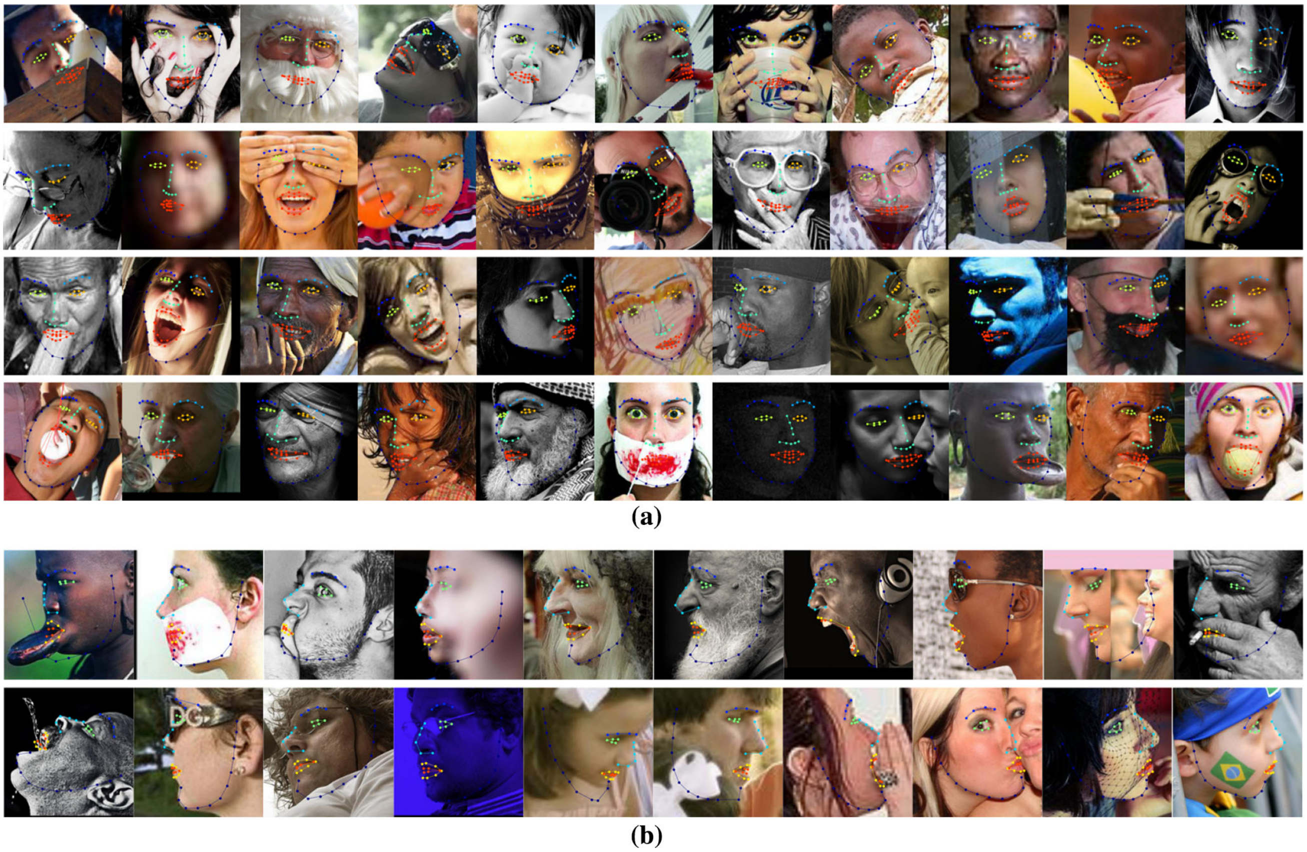


Fig. 17 Alignment examples with largest errors predicted by the proposed joint multi-view alignment method on the Menpo 2D semi-frontal and profile dataset. **a** Menpo 2D semi-frontal, **b** Menpo 2D profile

as cascade regression architectures (Xiong and De la Torre 2013; Cao et al. 2014b; Asthana et al. 2014; Yan et al. 2013; Deng et al. 2016). In the 300VW competition, there was no deep learning entry. Furthermore, the competing methods of 300VW revolved around cascade regression (Yang et al. 2015), CLM (Saragih et al. 2011) and Deformable Part-based Model (DPM) (Zhu and Ramanan 2012).

On contrary, recent state-of-the-art methods on 2D face alignment (Trigeorgis et al. 2016; Güler et al. 2017) extensively employ deep learning methodologies to improve the robustness of the in-the-wild face alignment model. The significant change of the landscape is also reflected in the Menpo 2D Challenge, as all of the participating methods are applying deep learning methodologies to the problem. This is attributed to the success of the recent deep architectures such as ResNets (He et al. 2016) and stacked Hourglasses models (Newell et al. 2016), as well as to the availability of a large amount of training data.

On the semi-frontal test set of **Menpo 2D Benchmark** (Zafeiriou et al. 2017c), we compare some representative classic approaches [e.g. CLM (Cristinacce and Cootes 2006), AAM (Tzimiropoulos and Pantic 2013; Cootes et al. 2001), CEM (Belhumeur et al. 2013) and SDM (Xiong and De la Torre 2013; Deng et al. 2016)] with the best three entries (Yang et al. 2017; He et al. 2017; Wu and Yang 2017) of the Menpo 2D Challenge as well as the proposed joint multi-view alignment method. As we can see from the CED curves in Fig. 18, it is obvious that recent deep convolutional feature based methods outperform the classic approaches by a large margin. M^3CSR (Deng et al. 2016), which is an improved version of SDM (Xiong and De la Torre 2013) and the champion method of the 300W challenge (Sagonas et al. 2016), only obtains the Normalised Mean Error (NME) of 2.17%. By contrast, the best performed method (Yang et al. 2017) achieves the NME of 1.2%.

4 Menpo 3D Challenge

Similarly to the Menpo 2D Challenge that was presented in the previous section, this section provides a detailed presentation of the Menpo 3D Challenge, in terms of the evaluation metrics used, the participants and the results of the challenge. Finally, we conclude this section by presenting our proposed method for 3D landmark localisation, which was not included in the challenge since we were the organisers.

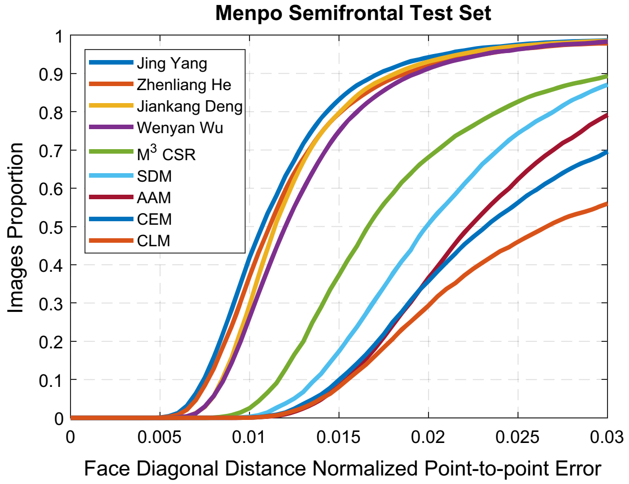


Fig. 18 Recent state-of-the-art methods versus classic approaches on the Menpo 2D semi-frontal test set. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalised with the diagonal of the ground-truth bounding box

Please note that we organised the Menpo 3D Challenge in conjunction with ICCV 2017 conference.

4.1 Evaluation Metrics

For assessing the performance of the submissions, we were once again based the standard evaluation metric, which is the normalised point-to-point error (Eq. 17). For the 3DA-2D landmarks, we used exactly the same evaluation metric and same diagonal-based normalisation as for the Semi-frontal 2D landmarks of the Menpo 2D challenge. For 3D landmarks, we used the point-to-point error in the 3D space of the model, where the scale is in cm of the normalised mean face, which corresponds to the scale of an average adult.

4.2 Participants

During the challenge, we provided approximately 12k static images with 3DA-2D and 3D landmarks, as well as approximately 90 training videos annotated with the proposed procedure. The training data have been provided to over 25 groups from all over the world. A tight schedule (a week) was provided to return the results on the test set. The test set comprises of 110 videos with 1000 frames each. The evaluation was performed in the 30 most challenging videos. Results for 3DA-2D landmarks localisation have been returned by three groups, while results for 3D landmarks have been returned by one group only. In the following, we will briefly describe each participating method.

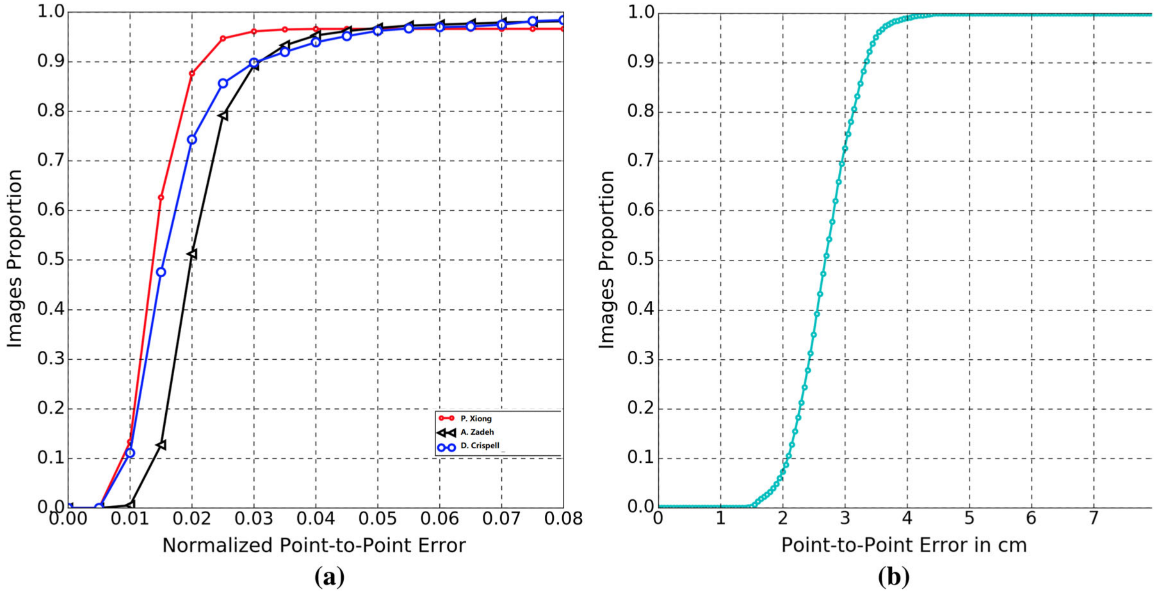
- *D. Crispell* The method in Crispell and Bazik (2017) proposed an efficient and fully automatic method for 3D face shape and pose estimation in the unconstrained 2D images. More specifically, the proposed method jointly estimates a dense set of 3D landmarks and facial geometry using a single pass of a modified version of the popular “U-Net” neural network architecture. In addition, the 3D Morphable Model (3DMM) parameters are directly predicted by using the estimated 3D landmarks and geometry as constraints in a linear system.
- *A. Zadeh* The method in Zadeh et al. (2017b) proposed to apply an extension of the popular Constrained Local Model (CLM), the so-called Convolutional Experts (CE)-CLM for the problem of 3DA-2D facial landmark detection. The important module of CE-CLM is a novel convolutional local detector that brings together the advantages of neural architectures and mixtures. In order to further improve 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.
- *P. Xiong* The method in Xiong et al. (2017) proposed a two-stage shape regression method by combining the powerful local heatmap regression and global shape regression. This method is based on the popular stacked Hourglass network which is used to generate a set of heatmaps for each 3D shape point. Since these heatmaps are independent to each other, a hierarchical attention mechanism is applied from global to local heatmaps into the network, in order to model the correlations among neighbouring regions. Then, all these heatmaps alongside the input aligned image are processed by a deep residual network to further learn the global features and produce the final smooth 3D shape.

4.3 Competition Results

As already mentioned, all three participants submitted results for 3DA-2D landmark localisation, whereas only Zadeh et al. (Crispell and Bazik 2017) submitted additional results for 3D landmark localisation. The CED curves for 3DA-2D and 3D landmarks are summarised in Fig. 19. We observe that, in the case of 3DA-2D landmarks, the best performing method was the method of Xiong et al. (2017). For pure 3D face tracking, the only method that competed in this category was the method of Zadeh (Crispell and Bazik 2017).

4.4 A New Strong Baseline for 3D Face Alignment

Since we organised the Menpo 3D competition, we could not submit an entry. However, as in the case of Menpo 2D competition, we have applied another our recent method (Deng et al. 2018) for localising the 3DA-2D landmarks of the Menpo 3D Benchmark. This method extends our previous



(a) **(b)**

Fig. 19 **a** CED curves for the 3DA-2D landmark localisation. **b** CED curves for the pure 3D landmark localisation (the only group that has sent results for this category was the method of Zadeh Crispell and Bazik (2017)). **a** Menpo 3DA-2D, **b** Menpo 3D

method (Deng et al. 2017) to the case of 3D landmarks and uses joint 2D and 3DA-2D landmark supervision. This extension retains the simplicity and very good performance of the method of Deng et al. (2017), therefore we consider it as a new strong baseline for 3D landmark localisation. In the following, we briefly present our approach and its results on the Menpo 3D benchmark.

4.4.1 Cascade Multi-view Hourglass Model

As shown in Fig. 20, we propose the Cascade Multi-view Hourglass Model (CMHM) for 3DA-2D face alignment, in which two Hourglass models are cascaded with intermediate supervision from 2D and 3DA-2D facial landmarks. For the 2D face alignment, we capitalise on the correspondences between the frontal and profile facial shapes and utilise the Multi-view Hourglass Model (Sect. 3.4) which jointly estimates both semi-frontal and profile 2D facial landmarks. Based on the 2D alignment results, a similarity transformation step is employed (in Fig. 21), and another Hourglass model is performed on the normalised face image to estimate the 3DA-2D facial landmarks. To improve the model capacity and compress the computational complexity of the Hourglass model, we replace the bottleneck block with the parallel and multi-scale Inception block and construct the Inception-ResNet block (Szegedy et al. 2017) as shown in Fig. 22.

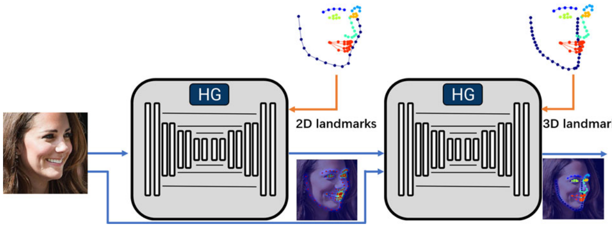


Fig. 20 Cascade Multi-view Hourglass Model for 2D and 3DA-2D face alignment

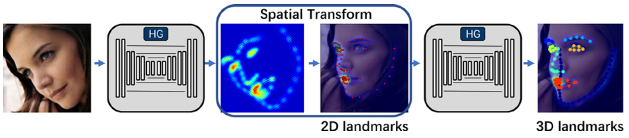
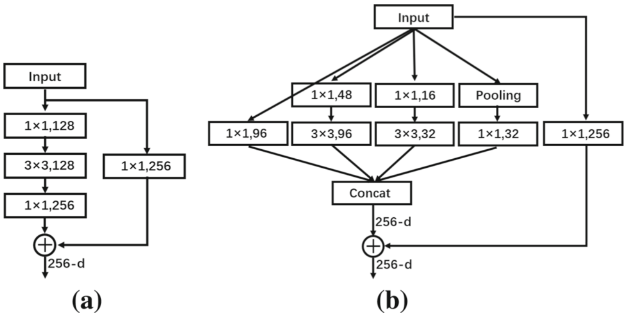


Fig. 21 2D alignment acts as a further spatial transform network for 3DA-2D alignment



(a) **(b)**

Fig. 22 **a** ResNet and **b** Inception-ResNet blocks to construct Hourglass model

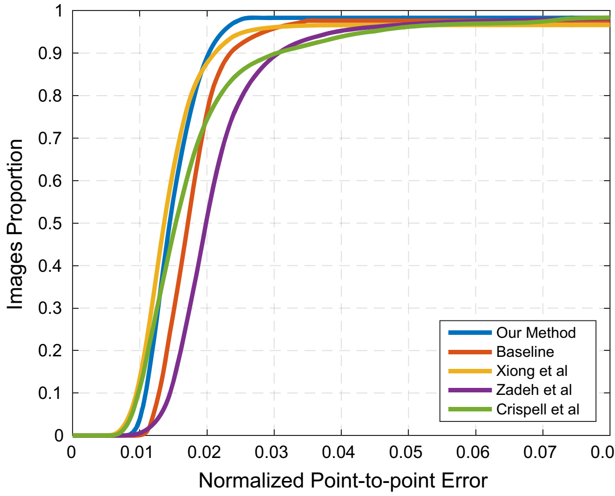


Fig. 23 CED curves on the Menpo 3D tracking testset

4.4.2 Experimental Results

We train the proposed Cascade Multi-view Hourglass Model on the image training sets of Menpo 2D and Menpo 3D benchmarks. The training of the proposed method follows a similar design as the Multi-view Hourglass Model for 2D landmark localisation (Deng et al. 2017). According to the centre and size of the bounding box provided by the face detector (Zhang et al. 2016a), each face region is cropped and scaled to 256 × 256. To improve the robustness of our method, we increase the number of training examples by randomly perturbing the ground-truth image with a different combination of rotation (+/— 45°), scaling (0.75–1.25), and translation (+/— 20 pixels). The full network starts with a 7×7 convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64, as it could reduce GPU memory usage while preserving alignment accuracy. The network is trained using Tensorflow with an initial learning rate of 10^{-4}, a batch size of 8, and 100k learning steps. We drop the learning rate to 10^{-5} after 20 epochs. The Mean Squared Error (MSE) loss is applied to compare the predicted heatmaps to the ground-truth heatmaps. Each training step takes 1.02 s on one NVIDIA GTX Titan X (Pascal). During testing, face regions are cropped and resized to 256 × 256, and it takes 20.76 ms to generate the response maps. By contrast, the baseline method, two-stack Hourglass model (Newell et al. 2016), takes 24.42 ms to generate the response maps.

To track the 3DA-2D landmarks in the videos of Menpo 3D benchmark, we perform a frame-by-frame tracking on the video. Specifically, we always initialise the next frame by the previous facial bounding box unless there is a fitting failure, in which case, a face detector (Zhang et al. 2016a) would be called to re-initialise. The fitting failure is judged by the failure checker as proposed in Sect. 3.4.

Table 6 3DA-2D alignment results on the Menpo 3D tracking testset

Method	AUC	FR (%)
Our method	**0.7977**	1.68
Baseline	0.7605	2.35
Xiong et al. (2017)	0.7935	3.38
Zadeh et al. (2017b)	0.7187	1.83
Crispell and Bazik (2017)	0.7617	1.61

Bold value indicates best result

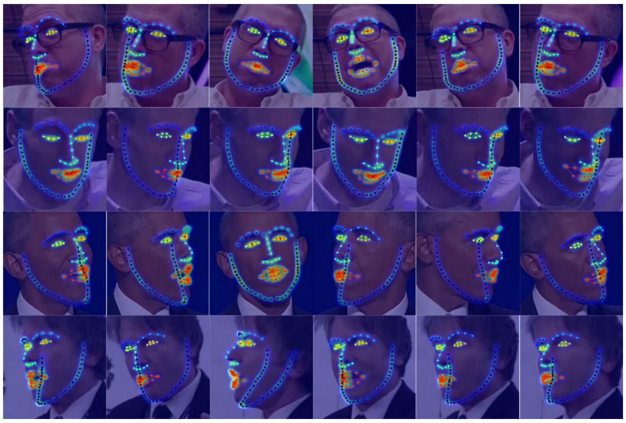


Fig. 24 Example results of our method on the Menpo-3D tracking testset

We expand the Menpo 3D Challenge results on 3DA-2D landmark tracking by including our method (CMHM) as well as the two-stack Hourglass model of Newell et al. (2016), which we consider as the "Baseline" method. Figure 23 reports the Cumulative Error Distribution (CED) curves, and Table 6 reports the Area Under the Curve (AUC) and Failure Rate (FR). We observe from the Table 6 that CMHM obtains a clear improvement (3.72% in AUC) over the baseline two-stack Hourglass model (Newell et al. 2016), and it also achieves the best performance (AUC = 0.7977, FR = 1.68%), which is slightly better than the challenge winner (Xiong et al. 2017), considering that they combined the local heatmap regression and global shape regression. We believe such good performance comes from the robustness of our response maps under large pose variations. This can be visually observed in Fig. 24, where we select some frames from the Menpo 3D tracking testset and plot their corresponding response maps as well as 3DA-2D alignment results. It is evident that the responses remain clear and evident across different poses.

5 Discussion and Conclusions

We have presented two new benchmarks for training and assessing the performance of landmark localisation algorithms in a wide range of poses. More specifically, the Menpo 2D dataset provides different landmark configurations for semi-frontal and profile faces based on the visible landmarks, thus making the 2D face alignment full-pose. By contrast, the Menpo 3D dataset provides a combined landmark configuration for both semi-frontal and profile faces based on the correspondence with a 3D face model, thus making face alignment not only full-pose but also corresponding to the real-world 3D space. We introduced an elaborate semi-automatic methodology for providing high-quality annotations for both the Menpo 2D and 3D datasets. The new benchmarks offer a large number of annotated training and test images for both semi-frontal and profile faces, which helps to boost the performance of 2D and 3DA-2D face alignment under large pose variations.

The state-of-the-art in face landmark localisation 6–7 years ago revolved around variations of Active Shape Models (ASMs), Active Appearance Models (AAMs) and Constrained Local Models (CLMs). Such methods exhibited good generalisation capabilities with few training data. Thanks to the availability of large amount of data and descriptive features such as HoG and SIFT, the state of the art moved towards discriminative methods such as cascade regression. Cascade regression methodologies dominated the field for around 3 years. The main bulk of recent work on cascade regression revolved around how to partition the search space so that to find good updates for various initialisation (Xiong and De la Torre 2015; Zhu et al. 2015). This competition shows that the landscape of facial landmark localisation has changed drastically in the last 2 years. That is, the current trend in landmark localisation, as in many computer vision tasks, involves the application of elaborate deep learning architectures to the problem. This was made feasible due to the large availability of training data, as well as due to recent breakthroughs in deep learning. The Menpo 2D and 3D competitions showed that elaborate deep learning approaches, such as Hourglass networks, achieve striking performance in facial landmark localisation. Furthermore, such fully convolutional architectures are very robust to initialisation/cropping of the face. Based on the Hourglass networks, we provide a unified solution, named Cascade Multi-view Hourglass Model (CMHM), to the 2D and 3D landmark localisation. The proposed method obtains state-of-the-art performance on the Menpo 2D and Menpo 3D datasets.

A crucial question that remains to be answered is “How far are we from solving the problem of face alignment?”. From the competition results, it is evident that large improvement has been achieved during the past few years. Nevertheless, for 10% to 15% of the images, the performance is still unsatisfactory. Especially when two or more challenging factors (e.g. large pose, exaggerated expression, heavy occlusion and sharp illumination) occur together, the alignment results turn out to be not robust and accurate enough because local facial appearance is occluded thus the local observation is consequently inaccurate.

Arguably, the most interesting question that should be answered is the following “Is the current performance good enough?”. Since face alignment is a means to an end of the question, this question could have various answers depending on the application. That is, the current performance could be satisfactory to conduct image normalisation for face recognition, but not for the recognition of complex emotional states or high-quality facial motion capture. In order to answer these questions, the community need to develop benchmarks that contain images/videos with dense annotations that can also be used for other facial analysis tasks.

Since many efforts have been devoted to sparse facial landmark localisation and great advances have been achieved during the last two decades, we are also interested in the question “What is the probable future research direction in face alignment?”. The most promising moving direction of this field might be defining and evaluating methods for predicting ultra-dense face correspondence, a dense version of the face alignment task (Güler et al. 2017; Liu et al. 2017; Feng et al. 2018). However, the main challenge of defining the dense face alignment in the wild is that the ground truth cannot be obtained from special devices (e.g. 3DMD) like the controlled environment or manually annotated from the 2D images but can only be automatically generated by high accurate 3D face model fitting. We will try to define and evaluate dense face alignment in the future.

Regarding licensing. The Menpo challenge further extends already existed databases and benchmarks. In particular, (a) Menpo builds upon 300W and 300VW (which includes parts of LFW, Helen, etc.), hence the user should adhere to the licensing terms of 300W and 300VW (Sagonas et al. 2013; Shen et al. 2015) and (b) uses images from FDDB (Jain and Learned-Miller 2010) and AFLW (Koestinger et al. 2011), hence the user should adhere to the licensing terms of FDDB and AFLW, as well.

Acknowledgements Jiankang Deng was supported by the President’s Scholarship of Imperial College London. This work was partially funded by the EPSRC Project EP/N007743/1 (FACER2VM) and a Google Faculty Fellowship to Dr. Zafeiriou. We thank the NVIDIA Corporation for donating several Titan Xp GPUs used in this work.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467.

Alabort-i Medina, J., & Zafeiriou, S. (2016). A unified framework for compositional fitting of active appearance models. In IJCV (pp. 1–39).

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., & Zafeiriou, S. (2015). Feature-based lucas-kanade and active appearance models. *TIP*, *24*(9), 2617–2632.

Aran, O., Ari, I., Guvensan, A., Haberdar, H., Kurt, Z., Turkmen, I., Uyar, A., & Akarun, L. (2007). A database of non-manual signs in Turkish sign language. In Signal Processing and Communications Applications (pp. 1–4). IEEE.

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In CVPR (pp. 1859–1866).

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *TPAMI*, *35*(12), 2930–2940.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S., et al. (2017). 3D face morphable models in-the-wild. In CVPR.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018). Large scale 3D morphable models. *IJCV*, *126*(2–4), 233–254.

Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., & Dunaway, D. (2016). A 3D morphable model learnt from 10,000 faces. In CVPR.

Bulat, A., & Tzimiropoulos, G. (2016). Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3dfaw) challenge. In ECCV workshops (pp. 616–624). Springer

Bulat, A., & Tzimiropoulos, G. (2017a). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In ICCV.

Bulat, A., & Tzimiropoulos, G. (2017b). How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In ICCV.

Burgos-Artizzu, X. P., Perona, P., & Dollár, P. (2013). Robust face landmark estimation under occlusion. In ICCV (pp. 1513–1520).

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2014a). Facewarehouse: A 3D facial expression database for visual computing. *TVCG*, *20*(3), 413–425.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014b). Face alignment by explicit shape regression. *IJCV*, *107*(2), 177–190.

Chen, X., Zhou, E., Mo, Y., Liu, J., & Cao, Z. (2017). Delving deep into coarse-to-fine framework for facial landmark localization. In CVPR Workshops.

Cheng, S., Marras, I., Zafeiriou, S., & Pantic, M. (2017). Statistical non-rigid ICP algorithm and its application to 3D face alignment. *Image and Vision Computing*, 58, 3–12.

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., & Zafeiriou, S. (2018). A comprehensive performance evaluation of deformable face tracking in-the-wild. *IJCV*, *126*(2–4), 198–232.

Chrysos, G. G., Antonakos, E., Zafeiriou, S., & Snape, P. (2015). Offline deformable face tracking in arbitrary videos. In Proceedings of the IEEE international conference on computer vision workshops (pp. 1–9).

Chung, J. S., Senior, A. W., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In CVPR (pp. 3444–3453).

Chung, J. S., & Zisserman, A. (2016). Lip reading in the wild. In ACCV (pp. 87–103). Springer.

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

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *TPAMI*, *23*(6), 681–685.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *CVIU*, *61*(1), 38–59.

Crispell, D., & Bazik, M. (2017). Pix2face: Direct 3D face model estimation. In: ICCV workshops (pp. 2512–2518). IEEE.

Cristinacce, D., & Cootes, T. F. (2006). Feature detection and tracking with constrained local models. In BMVC.

Deng, J., Liu, Q., Yang, J., & Tao, D. (2016). M3csr: Multi-view, multiscale and multi-component cascade shape regression. *IVC*, *47*, 19–26.

Deng, J., Trigeorgis, G., Zhou, Y., & Zafeiriou, S. (2017). Joint multiview face alignment in the wild. arXiv:1708.06023.

Deng, J., Zhou, Y., Chen, S., & Zafeiriou, S. (2018). Cascade multi-view hourglass model for robust 3D face alignment. In FG.

Eleftheriadis, S., Rudovic, O., Deisenroth, M. P., & Pantic, M. (2016a). Gaussian process domain experts for model adaptation in facial behavior analysis. In CVPR workshops.

Eleftheriadis, S., Rudovic, O., & Pantic, M. (2016b). Joint facial action unit detection and feature fusion: A multi-conditional learning approach. *TIP*, *25*(12), 5727–5742.

Fan, H., & Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *IVC*, *47*, 27–35.

Feng, Y., Wu, F., Shao, X., Wang, Y., & Zhou, X. (2018). Joint 3D face reconstruction and dense alignment with position map regression network. In ECCV.

Feng, Z. H., Kittler, J., Awais, M., Huber, P., & Wu, X. (2017). Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. In CVPR workshops.

Ghiasi, G., & Fowlkes, C. C. (2015). Occlusion coherence: Detecting and localizing occluded faces. arXiv:1506.08347.

Girshick, R. (2015). Fast r-cnn. In ICCV (pp. 1440–1448).

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multipie. *IVC*, *28*(5), 807–813.

Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., & Kokkinos, I. (2017). Densereg: Fully convolutional dense shape regression in-the-wild. In CVPR.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In CVPR (pp. 770–778).

He, Z., Zhang, J., Kan, M., Shan, S., & Chen, X. (2017). Robust feccnn: A high accuracy facial landmark detection system. In CVPR workshops.

Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., & Kautz, J. (2018). Improving landmark localization with semi-supervised learning. In *The IEEE conference on computer vision and pattern recognition* (CVPR)

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Workshop on faces in real-life images: Detection, alignment, and recognition.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In: NIPS (pp. 2017–2025).

Jain, V., & Learned-Miller, E. G. (2010). Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst technical report.

Jeni, L. A., Tulyakov, S., Yin, L., Sebe, N., & Cohn, J. F. (2016). The first 3d face alignment in the wild (3dfaw) challenge. In ECCV (pp. 511–520). Springer.

Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001). Robust face detection using the Hausdorff distance. In Audio and video based biometric person authentication (pp. 90–95). Springer.

Kasinski, A., Florek, A., & Schmidt, A. (2008). The put face database. *Image Processing and Communications*, *13*(3–4), 59–64.

Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In ICCV workshop (pp. 2144–2151). IEEE.

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In ICCV workshops (pp. 2144–2151). IEEE.

Kowalski, M., Naruniec, J., & Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. In CVPR workshops.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. (2012). Interactive facial feature localization. In ECCV (pp. 679–692). Springer.

Liu, Y., Jourabloo, A., Ren, W., & Liu, X. (2017). Dense face alignment. In ICCV workshops (pp. 1619–1628).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In ICCV.

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999). Xm2vtsdb: The extended m2vts database. *Audio and Video based Biometric Person Authentication*, *964*, 965–966.

Milborrow, S., Morkel, J., & Nicolls, F. (2010). The MUCT landmarked face database. *Pattern Recognition Association of South Africa, 201*(0)

Milborrow, S., & Nicolls, F. (2014). Active shape models with sift descriptors and mars. In Computer vision theory and applications (Vol. 2, pp. 380–387). IEEE.

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In ECCV (pp. 483–499). Springer.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., et al. (2005). Overview of the face recognition grand challenge. *CVPR*, *1*, 947–954.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In CVPR (pp. 1685–1692).

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *IVC*, *47*, 3–18.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In ICCV workshops (pp. 397–403).

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *IJCV*, *91*(2), 200–215.

Shao, X., Xing, J., Lv, J. J., Xiao, C., Liu, P., Feng, Y., Cheng, C. (2017). Unconstrained face alignment without face detection. In CVPR workshops.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In ICCV workshops (pp. 1003–1011). IEEE.

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., & Samaras, D. (2017). Neural face editing with intrinsic image disentangling. In CVPR (pp. 5444–5453). IEEE.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI (Vol. 4, p. 12).

Tadmor, O., Rosenwein, T., Shalev-Shwartz, S., Wexler, Y., & Shashua, A. (2016). Learning a metric embedding for face recognition using the multibatch method. In NIPS (pp. 1388–1389).

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In CVPR (pp. 1701–1708).

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., & Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In CVPR (pp. 4177–4187).

Tzimiropoulos, G., & Pantic, M. (2013). Optimization problems for fast aam fitting in-the-wild. In ICCV (pp. 593–600).

Webb, J. A., & Aggarwal, J. K. (1982). Structure from motion of rigid and jointed objects. *Artificial Intelligence*, *19*(1), 107–13.

Wu, W., & Yang, S. (2017). Leveraging intra and inter-dataset variations for robust face alignment. In CVPR workshops.

Xiao, S., Li, J., Chen, Y., Wang, Z., Feng, J., Yan, S., & Kassim, A. A. (2017). 3D-assisted coarse-to-fine extreme-pose facial landmark detection. In CVPR workshops.

Xiong, P., Li, G., & Sun, Y. (2017). Combining local and global features for 3D face tracking. In ICCV workshops. IEEE.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In CVPR (pp. 532–539).

Xiong, X., & De la Torre, F. (2015). Global supervised descent method. In CVPR (pp. 2664–2673).

Yan, J., Lei, Z., Yi, D., & Li, S. (2013). Learn to combine multiple hypotheses for accurate face alignment. In ICCV workshops (pp. 392–396).

Yang, J., Deng, J., Zhang, K., & Liu, Q. (2015). Facial shape tracking via spatio-temporal cascade shape regression. In ICCV workshops (pp. 41–49).

Yang, J., Liu, Q., & Zhang, K. (2017). Stacked hourglass network for robust facial landmark localisation. In CVPR workshops.

Zadeh, A., Baltrusaitis, T., & Morency, L. P. (2017a). Convolutional experts network for facial landmark detection. In CVPR workshops.

Zadeh, A., Lim, Y. C., Baltrusaitis, T., & Morency, L. P. (2017b). Convolutional experts constrained local model for 3D facial landmark detection. In ICCV workshops (Vol. 7).

Zafeiriou, S., Chrysos, G., Roussos, A., Ververas, E., Deng, J., & Trigeorgis, G. (2017a). The 3D menpo facial landmark tracking challenge. In ICCV workshops.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., & Shen, J. (2017b). The menpo facial landmark localisation challenge: A step towards the solution. In CVPR workshops (pp. 2116–2125).

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., & Shen, J. (2017c). The menpo facial landmark localisation challenge: A step towards the solution. In CVPR workshop.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016a). Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, *23*(10), 1499–1503.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. In ECCV (pp. 94–108). Springer.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2016b). Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, *38*(5), 918–930.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., & Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In ICCV workshops (pp. 386–391). IEEE.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In CVPR (pp. 4998–5006).

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2016a). Unconstrained face alignment via cascaded compositional learning. In CVPR (pp. 3409–3417).

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016b). Face alignment across large poses: A 3D solution. In CVPR.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016c). Face alignment across large poses: A 3D solution. In CVPR (pp. 146–155).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In CVPR (pp. 2879–2886). IEEE.