The Menpo Facial Landmark Localisation Challenge: A step towards the solution

Stefanos Zafeiriou* George Trigeorgis Grigorios Chrysos Jiankang Deng
Jie Shen
Department of Computing
Imperial College London

{s.zafeiriou, g.trigeorgis, g.chrysos, j.deng16, jie.shen07}@imperial.ac.uk

Abstract

In this paper, we present a new benchmark (Menpo benchmark) for facial landmark localisation and summarise the results of the recent competition, so-called Menpo Challenge, run in conjunction to CVPR 2017. The Menpo benchmark, contrary to the previous benchmarks such as 300-W and 300-VW, contains facial images both in (nearly) frontal, as well as in profile pose (annotated with a different markup of facial landmarks). Furthermore, we increase considerably the number of annotated images so that deep learning algorithms can be robustly applied to the problem. The results of the Menpo challenge demonstrate that recent deep learning architectures when trained with the abundance of data lead to excellent results. Finally, we discuss directions for future benchmarks in the topic.

1. Introduction

Facial landmark localisation and tracking on images/videos captured in unconstrained recording conditions is a problem that has received a lot of attention the past few years. This is attributed to its numerous applications in face recognition [41], facial behaviour analysis [19, 18], lip reading [14, 13], 3D face reconstruction [49, 7, 8] and face editing [38], just to name a few.

Currently, methodologies that achieve good performance have been presented in recent top-tier computer vision conferences (e.g., ICCV, CVPR, ECCV, BMVC, ACCV etc.). 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 [34, 33, 6, 28, 52, 26], as well as rigorous protocols for performance assessment. The current benchmarks for facial landmark localisation and tracking were presented in satellite workshop-challenges of ICCV 2013 [34] and ICCV 2015 [37] (so-called 300-W and 300-VW benchmarks). The annotated data of 300-W and 300-VW benchmarks are now used by the majority of scientific and industrial community [42, 7, 4, 40, 1] for training and testing facial landmark localisation/tracking algorithms.

Even though the data we provided in 300-W and 300-VW had large impact in the computer vision community, there are still several limitations including

- the data have been annotated using only frontal sparse facial shape,
- annotated test set of 300-W competition is comprised of 600 facial images only.

Motivated by the above we made a significant step further and propose a new comprehensive large-scale benchmark, which contains both semi-frontal and profile faces, annotated with their corresponding facial shape model. Furthermore, we outline the results achieved by the participants of the challenge. The results demonstrate that for frontal faces the performance of the methodologies are starting to converge to an excellent performance. On the other hand for profile faces there is a considerable space for improvement. Finally, we provide some suggestions regarding future challenges on the topic.

2. Menpo Challenge Benchmark

Before presenting the Menpo challenge data we outline the data provided by the previous challenges, i.e. 300-W [34, 33] and 300-VW [37]. Then, we discuss about the aims of the new benchmark and its added value.

The 300-W challenge provides publicly available annotations for over 16,000 images. The "in-the-wild" datasets that have been annotated were

[*] affiliates also with the Department of Computer Science and Engineering, University of Oulu, Finland.

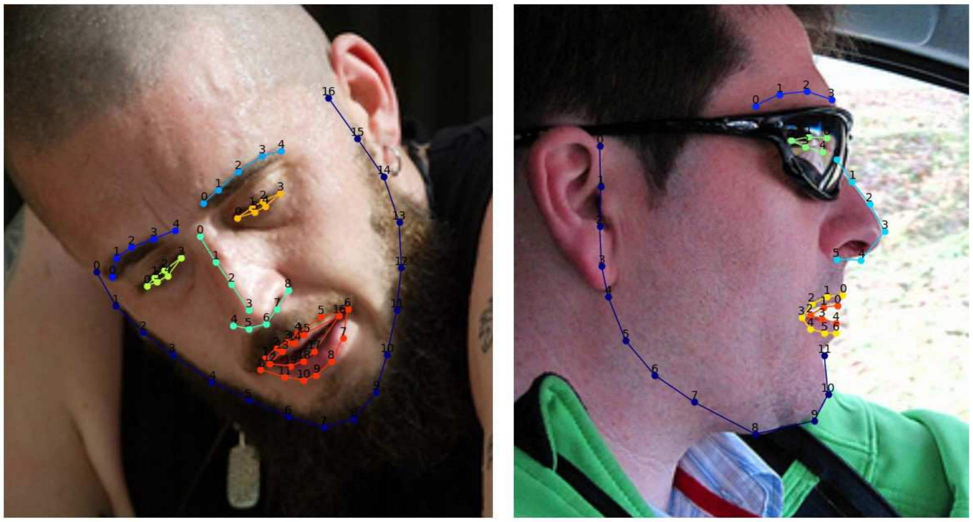

Figure 1: Example facial images (left) annotated with a 68 landmark semi-frontal-face markup and (right) annotated with a 39 landmark profile-face markup.

- Labeled Face Parts in the Wild (LFPW) database [6]. Because LFPW provides only the source links to download the images and not the actual images, only 1035 image were available (out of 1287).
- Helen database [28] which consists of 2330 images downloaded from the `flickr.com` web service.
- The Annotated Faces in-the-wild (AFW) [52] database which consists of 250 images with 468 faces.
- Two new datasets. That is, iBug, which consists of 135 images and the test set of 300-W, which consists of 300 images captured indoors and 300 images captured outdoors. The 300-W test set was publicly released with the second version of the competition [33].

In total the 300-W competition provided 4350 "in-the-wild" images of around 5,000 faces. The faces have been annotated using a 68 landmark frontal face markup scheme that was also used in Multi-PIE (please see Fig. 1 for an example of the 68 landmark mark-up used).

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 300-VW benchmark [37]. The 300-VW benchmark consists of 114 videos and 218,595 frames. For a recent comparison of the state-of-the-art in 300-VW the interested reader may refer to [12]. The 68 frontal face markup scheme was also used for annotating the faces of this 300-VW benchmark.

The two limitations of the previous challenges [1] were that they

- contained few faces in extreme poses (e.g. full profile images); the few images in extreme pose have been annotated with the mark-up of frontal faces;
- the test was consisting of very few images (around 600).

To alleviate the above limitations we decided to introduce the Menpo benchmark and, using this benchmark, to introduce a new challenge in conjunction to CVPR 2017.

The Menpo challenge consists of

- Training set: 5658 semi-frontal and 1906 profile facial images.
- Test set: 5335 frontal and 1946 profile facial images.

The profile facial images have been annotated with a 39 profile landmark scheme (an example is shown in Fig [**?**]). All images have been taken from LFW and FDDB databases and the annotation process was as follows. For semi-frontal images a semi-automatic process was applied similar to [34] but now instead of an Active Appearance Model (AAM) the method we used was the Mnemonic Descent Method (MDM) [42]. That is, an MDM trained on the 300-W data was first applied on the data. The output facial landmarks were inspected and corrected manually and another MDM was trained with the new annotated data etc. Since, at the time the data have been annotated no publicly available datasets of profile faces were available, we had to manually annotate many images from scratch (around 1,200). Using

[1] Last year another competition was held in conjunction with ECCV 2016 revolving around sparse 3D landmark localisation [25]. Nevertheless, it mainly revolved around images that have been either captured in highly controlled conditions or generated artificially.

these images a semi-automatic procedure, as above, was applied for annotating the remaining 2,500 profile images. Finally, using the landmarks, the faces from each image were cropped and the cropped facial images were provided for training and testing.

Another aim of ours was to provide an adequate large number of facial images so that recent deep learning architectures such as ResNets [23], VGG series of networks [39] and Stacked Hourglasses [31] can be robustly trained. Hence, the participants have access to over 11,000 annotated training semi-frontal faces (300-W and Menpo challenge training data) and 1,906 annotated profile faces.

3. Summary of Participants

We decided to allow entries in either semi-frontal or profile challenge. 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 provide an abbreviation based on the name of the first author of the paper):

- **X. Chen**: The method in [11] proposed a 4-stage coarse-to-fine framework to tackle the facial landmark localisation problem in-the-wild. In the first state a Convolutional Neural Network (CNN) first transformed the faces into a canonical orientation and then the first estimate of the landmarks was predicted. Then, fine-scale refinement was performed using linear regressors from patches around the landmarks.

- **X.-H. Shao**: The method in [36] used a CNN to predict a small set of initial landmarks, but at a latter stage, the coarse canonical face and the pose were generated by a Pose Splitting Layer based on the visible basic landmarks. According to its pose, each canonical state was distributed to the corresponding branch of the shape regression sub-networks for the detection of all 68 facial landmarks.

- **Z. He**: The method in [24] used an ensemble of networks where each network is a cascade of sub-networks similar to an MDM [42] to predict the final shape.

- **Z. Feng**: The method in [21] used an ensemble of ridge regressors to predict the final result, separate for the semi-frontal and profile views.

- **J. Yang**: The method in [47] used a CNN to remove similarity transformations from the detected face and then used a Stacked Hourglass Network [] to regress directly to the final result.

- **M. Kowalski**: The method in [27] used a VGG-based alignment network to correct similarity transforms and then a fully-convolutional network that regressed to the final shape.

- **A. Zadeh**: The method in [48] used a fully-convolutional model with a CLM-based loss function.

- **S. Xiao**: The method in [44] used an MDM [42] like recurrent model with a final refinement linear layer for each the semi-frontal and profile views.

- **W. Wu**: The method in [43] used a VGG-16 based network to regress to a parametric form of the shape of multiple datasets.

In the first [34] and second [33] runs of 300-W competition there were very few competing methods that applied deep learning methods to the problem [20]. The state-of-the-art at that time was revolving around Constrained Local Models (CLMs) [15, 35], feature-based Active Shape Models (ASMs) [29] and AAMs [3, 2], as well as cascade regression architectures (e.g., Supervised Descent Methods, Explicit Shape Regression etc. [45, 5, 16, 10]). In the 300-VW competition there was no deep learning entry. The competing methods of 300-VW revolved around CLMs, cascade regression and Deformable Part-based Models (DPMs) [52]. Furthermore, the recent state-of-the-art in 300=W and 300-VW was a neural network architecture that combined a shallow CNN and a recurrent neural network (RNN) (so called Mnemonic Descent Method [42]) and most recently a combination of a dense shape regression method plus MDM [22].

The Menpo challenge demonstrates that the landscape has significantly changed. That is, as can be verified from the above brief analysis of the techniques, all competing methods are applying deep learning methodologies to the problem. This is attributed to the success of the recent deep architectures such as ResNets [23] and stacked Hourglasses models [31], as well as to the availability of a large amount of training data.

4. Competition Results

We allowed participants to compete in either semi-frontal or profile challenges (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 and not the corresponding annotations. Furthermore, we provided information regarding which images were considered semi-frontal and profile. The participants were allowed to submit results (i.e., the facial landmarks) up until the 31st of March after which the challenge was considered finished.

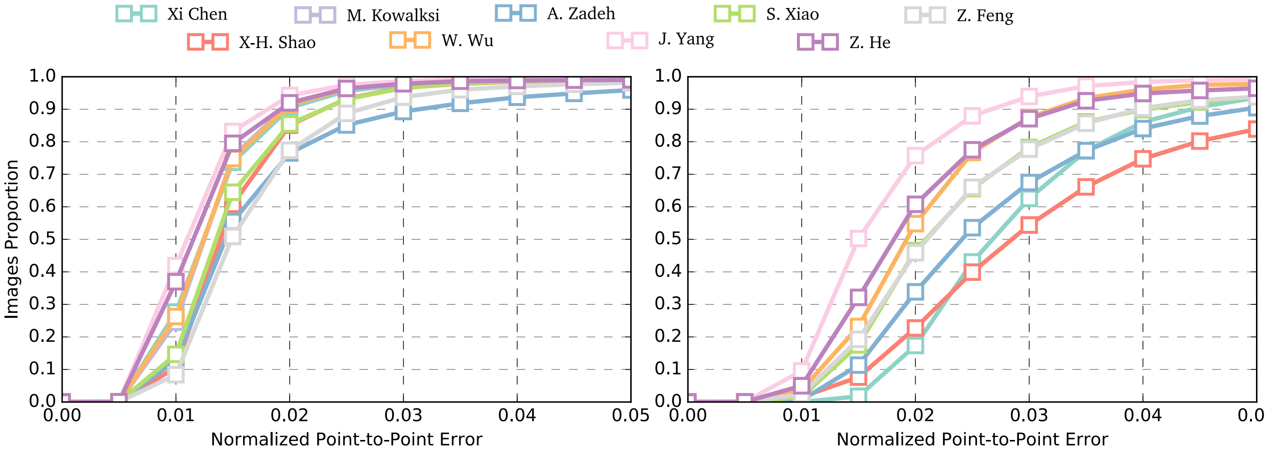

Figure 2: Quantitative results (CED curves) on the test set of the Menpo Benchmark competition for both semi-frontal (left) and profile (right) results.

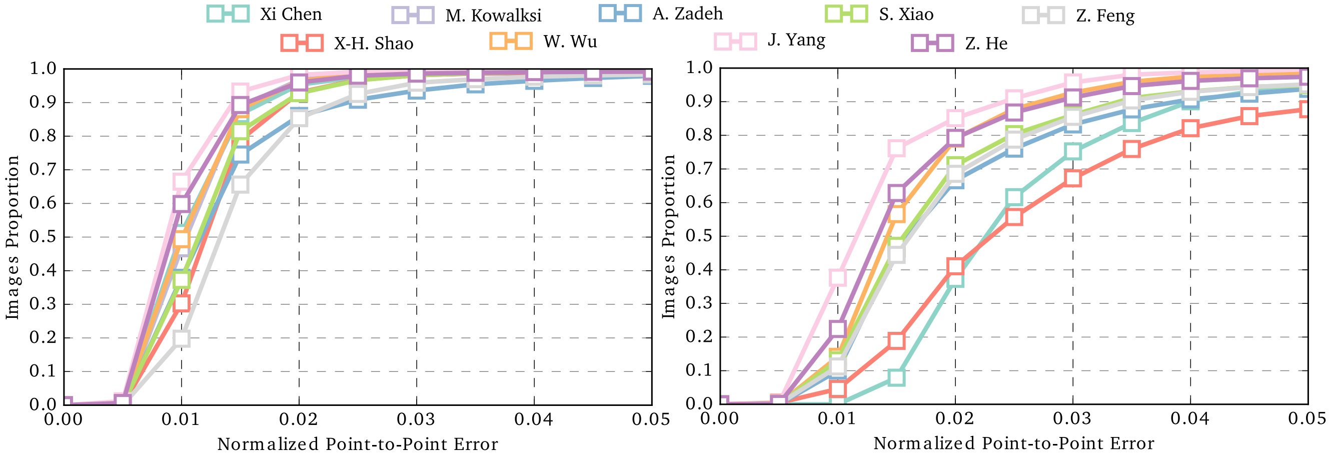

Figure 3: Quantitative results (CED curves) on the interior landmarks of the test set of the Menpo Benchmark competition for both semi-frontal (49-points) (left) and profile (28-points) (right) results.

Evaluation metrics. The standard evaluation metric for landmark-wise face alignment is the normalised root-mean square error

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

where $\hat{\mathbf{s}}$ and $\mathbf{s}^*$ are the estimated and ground truth shape respectively, $\|.\|_2$ is the ℓ_2 norm and d_{scale} is a normalisation factor to make the error scale invariant. For the last two face alignment competitions [34, 33] and most recent works [42] the inter-ocular distance was used as the normalisation factor. Unfortunately the inter-ocular distance fails to give a meaningful localisation metric in the case of profile views as it becomes a very small value. Instead, we used the face diagonal as the normalisation factor which is more robust to changes of the face pose.

Many works on the topic [32] report just the average of the error in (1). 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 CED curves. We have calculated some further statistics from the CED curves such as the area-under-the-curve (AUC) (up to error of 0.05), the failure rate of each method (we consider any fitting with a point-to-point error greater than 0.05 as a failure), the maximum error, the median, Median Absolute Deviation (MAD) and standard deviation (Std). We believe that these are more representative error metrics for the problem of face alignment.

The CED curves for all 68 landmarks for semi-frontal faces and for all 39 landmarks for profile faces is shown in Fig. 2. The key statistics of the CED curve for the semi-

	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.05}$	Failure Rate
J. Yang *et al.* [47]	0.0120	0.0060	0.0107	0.0022	0.1453	0.7624	0.0024
Z. He *et al.* [24]	0.0139	0.0260	0.0111	0.0023	0.9624	0.7478	0.0096
M. Kowalski *et al.* [27]	0.0138	0.0157	0.0120	0.0023	0.6312	0.7337	0.0049
Wenyan Wu [43]	0.0135	0.0095	0.0120	0.0024	0.5098	0.7337	0.0036
Xi chen *et al.* [11]	0.0200	0.0756	0.0120	0.0026	1.2799	0.7290	0.0111
S. Xiao [44] *et al.*	0.0159	0.0201	0.0133	0.0027	0.6717	0.6986	0.0081
X.-H. Shao *et al.* [36]	0.0165	0.0235	0.0138	0.0027	0.9612	0.6913	0.0101
Z. Feng *et al.* [21]	0.0182	0.0179	0.0149	0.0033	0.4661	0.6586	0.0186
A. Zadeh *et al.* [48]	0.0205	0.0340	0.0143	0.0035	0.9467	0.6479	0.0409

Table 1: Key statistics of the performance of the participants in semi-frontal faces (68-points markup).

	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.05}$	Failure Rate
J. Yang *et al.* [47]	0.0172	0.0105	0.0150	0.0035	0.2490	0.6613	0.0077
Z. He *et al.* [24]	0.0247	0.0422	0.0179	0.0048	0.6280	0.5932	0.0355
Wenyan Wu [43]	0.0217	0.0131	0.0193	0.0044	0.2623	0.5802	0.0221
Z. Feng *et al.* [21]	0.0285	0.0367	0.0208	0.0057	0.4725	0.5268	0.0617
S. Xiao [44] *et al.*	0.0290	0.0417	0.0209	0.0055	0.6327	0.5237	0.0612
A. Zadeh *et al.* [48]	0.0375	0.0630	0.0241	0.0071	0.7594	0.4604	0.0951
Xi chen *et al.* [11]	0.0448	0.1162	0.0265	0.0058	1.3698	0.4259	0.0642
X.-H. Shao *et al.* [36]	0.0451	0.0636	0.0282	0.0088	0.7534	0.3891	0.1608

Table 2: Key statistics of the performance of the participants in profile faces (39-points markup).

frontal faces are summarised in Table 1, while the key statistics for the profile faces are summarised in Table 2. From the statistics and the curves it is evident that in the category of semi-frontal faces the first three entries were quite close. Nevertheless, in the category of the profile faces there was a clear winner. In both cases the best performing method was that of [47], which is also the winner of the competition.

As it is customary in landmark evaluation papers we also provide performance graphs excluding the boundary landmarks for both the semi-frontal, as well as the profile faces. The CED curves for the 49 landmarks for semi-frontal faces and for 28 landmarks for profile faces is shown in Fig. 3. The key statistics are summarised in Table 3 and Table 4.

5. Comparison with previous Competitions and State-of-the-art

In this section we will discuss the improvement that can be potentially achieved by using architectures similar to that of the winning entry. Since we organised the competition we could not submit an entry. Nevertheless, we have been experimenting with hourglass architectures [17]. In particular, our work in [17] proposes a CNN that is trained for both the tasks of face detection and landmark localisation. The first network is trained to produce face proposals, as well as to estimate a small set of landmarks which are then used in order to remove the similarity transformation. Finally, a multi-view Hourglass Model is trained to predict the response map for all landmarks (both 68 of semi-frontal mark-up, as well as 39 of the profile mark-up).

A method bearing similarities to ours, independently proposed in [47], won the challenge. Fig. 4 plots the performance of the best three entries of the competition and the performance of our methodology in [17] (abbreviated as Deng et. al. in the Figs. below) in both semi-frontal and profile faces. The method has similar performance to the best performing methods in semi-frontal faces. Nevertheless, it outperforms the best performing method in profile faces.

In order to demonstrate the improvement over the previous state-of-the-art (as submitted in the previous competitions) we run the method in the test sets of 300-W and 300-VW [2]. Fig. 5 plots the competing methods of the second conduct of 300-W competition and the current state-of-the-art method published in CVPR 2017 [22]. Our hourglass network [17] offers large improvement over the state-of-the-art.

Fig. 6 plots the performance of the first two best performing methods of the 300-VW challenge [37], as well as

[2] It worth noting that in order to be directly comparable with the results of 300-W the error has been normalised with the interocular distance and not with the main face diagonal.

	Mean	Std	Median	MAD	Max Error	$\mathrm{AUC}_{0.05}$	Failure Rate
J. Yang *et al.* [47]	0.0097	0.0053	0.0087	0.0017	0.1719	0.8084	0.0022
Z. He *et al.* [24]	0.0117	0.0253	0.0093	0.0019	0.9520	0.7886	0.0079
Wenyan Wu [43]	0.0113	0.0085	0.0101	0.0019	0.4752	0.7778	0.0024
M. Kowalski *et al.* [27]	0.0116	0.0147	0.0102	0.0018	0.6720	0.7765	0.0036
Xi chen *et al.* [11]	0.0174	0.0724	0.0099	0.0021	1.2699	0.7746	0.0096
S. Xiao [44] *et al.*	0.0132	0.0188	0.0110	0.0022	0.6411	0.7513	0.0066
X.-H. Shao *et al.* [36]	0.0139	0.0220	0.0115	0.0022	0.9590	0.7420	0.0084
A. Zadeh *et al.* [48]	0.0162	0.0319	0.0111	0.0026	0.9377	0.7200	0.0204
Z. Feng *et al.* [21]	0.0159	0.0164	0.0129	0.0029	0.3686	0.7007	0.0161

Table 3: Key statistics of the performance of the participants in semi-frontal faces (49-points markup).

	mean	std	median	mad	max	auc	fr
J. Yang *et al.* [47]	0.0136	0.0093	0.0110	0.0026	0.2162	0.7319	0.0036
Z. He *et al.* [24]	0.0201	0.0414	0.0132	0.0035	0.6380	0.6778	0.0257
Wenyan Wu [43]	0.0168	0.0109	0.0142	0.0034	0.2252	0.6709	0.0128
S. Xiao [44] *et al.*	0.0233	0.0416	0.0154	0.0042	0.7073	0.6231	0.0509
Z. Feng *et al.* [21]	0.0236	0.0361	0.0161	0.0046	0.5141	0.6124	0.0483
A. Zadeh *et al.* [48]	0.0293	0.0632	0.0157	0.0046	0.8780	0.5990	0.0617
Xi chen *et al.* [11]	0.0409	0.1181	0.0223	0.0051	1.3809	0.4954	0.0493
X.-H. Shao *et al.* [36]	0.0388	0.0636	0.0228	0.0079	0.7769	0.4756	0.1223

Table 4: Key statistics of the performance of the participants in semi-frontal faces (28-points markup).

the best performing methods of the recent comparison [12] [3]. Again it can be observed that deep learning architectures based on hourglass greatly improve the state-of-the-art [4].

6. Discussion and conclusions

We have presented a new benchmark for training and assessing the performance of landmark localisation algorithms in a wide range of poses. The new benchmark offers a large number of annotated training and test images of both semi-frontal and profile faces (using different mark-ups).

The state-of-the-art in face landmark localisation five-six years ago revolved around variations of CLMs, ASMs , DPMs, and AAMs. Then, with the availability of large amount of data and descriptive features, such as HoGs, 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 initialisations [46, 51]. This competition shows that the landscape of landmark localisation and face alignment has changed drastically the couple of years. That is, the current trend in landmark localisation, as in many computer vision tasks currently, involves the application of elaborate deep learning architectures to the problem. This was made feasible due to large availability of training data, as well as due to recent breakthroughs in deep learning (e.g., residual learning).

This competition 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 (actually they can be used to perform face detection, as well).

A crucial question that remains to be answered is how far away are we from solving the problem. From the 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 not satisfactory. An interesting further research on the topic is to perform analysis of the errors (e.g., are the errors due to occlusion? due to blurring or poor image quality?).

Arguably, the most interesting question that should be answered is the following “is the current achieved performance good enough?”. Since, face alignment is a means

[3] The best performing method in [12] was a pipeline combining of a deep neural network for bounding box tracking (MDNET [30]) or a deep learning method for face detection (MTCNN [50]), a generic method for facial landmark localisation [51] and a Kalman filter for smoothing the output.

[4] A deep learning architecture that uses a hourglass network also won the 3D face alignment competition organised in ECCV 2016 [9].

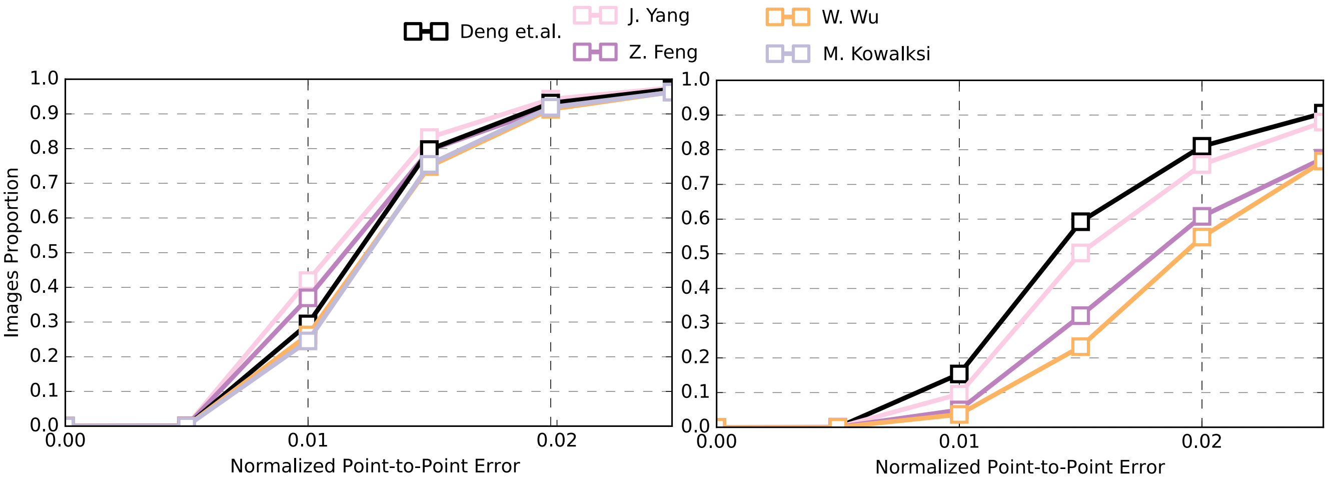

Figure 4: Quantitative results for the top-3 performs on the test set of the Menpo Benchmark competition for both semi-frontal (left) and profile (right) results.

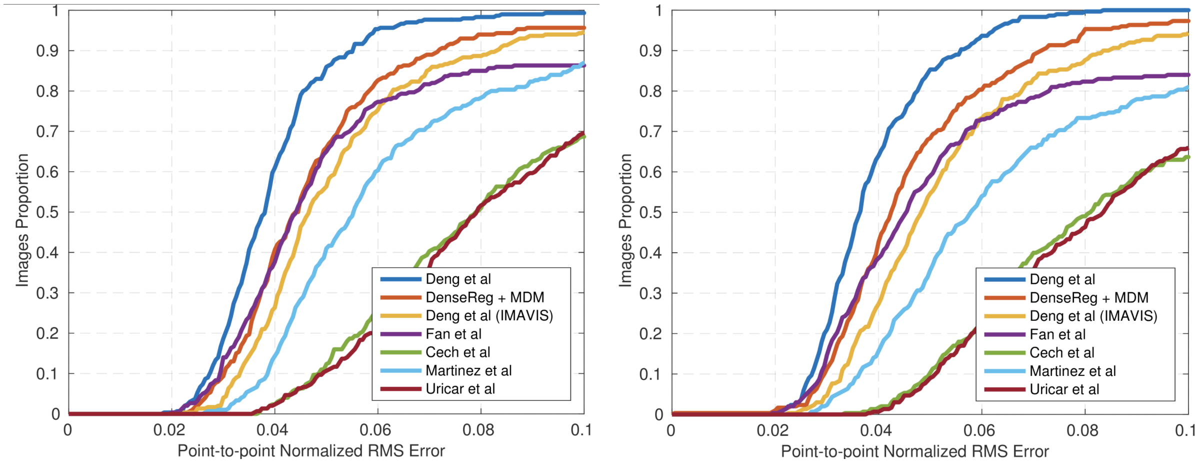

Figure 5: CED curves of the recent state-of-the-art on the test sets, indoor (left) and outdoor (right) of the 300-W benchmark.

to an end the 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 that can also be used for other facial analysis tasks.

7. Acknowledgements

The work of Stefanos Zafeiriou was partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015), as well as by the EPSRC project EP/N007743/1 (FACER2VM). G. Trigeorgis and G. Chrysos were supported by EPSRC DTA award at Imperial College London. J. Deng was supported by the President's Scholarship of Imperial College London.

References

[1] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 679–682. ACM, 2014. 1

[2] J. Alabort-i Medina and S. Zafeiriou. A unified framework for compositional fitting of active appearance models. *International Journal of Computer Vision*, pages 1–39, 2016. 3

[3] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE Transactions on Image Processing*, 24(9):2617–2632, September 2015. 3

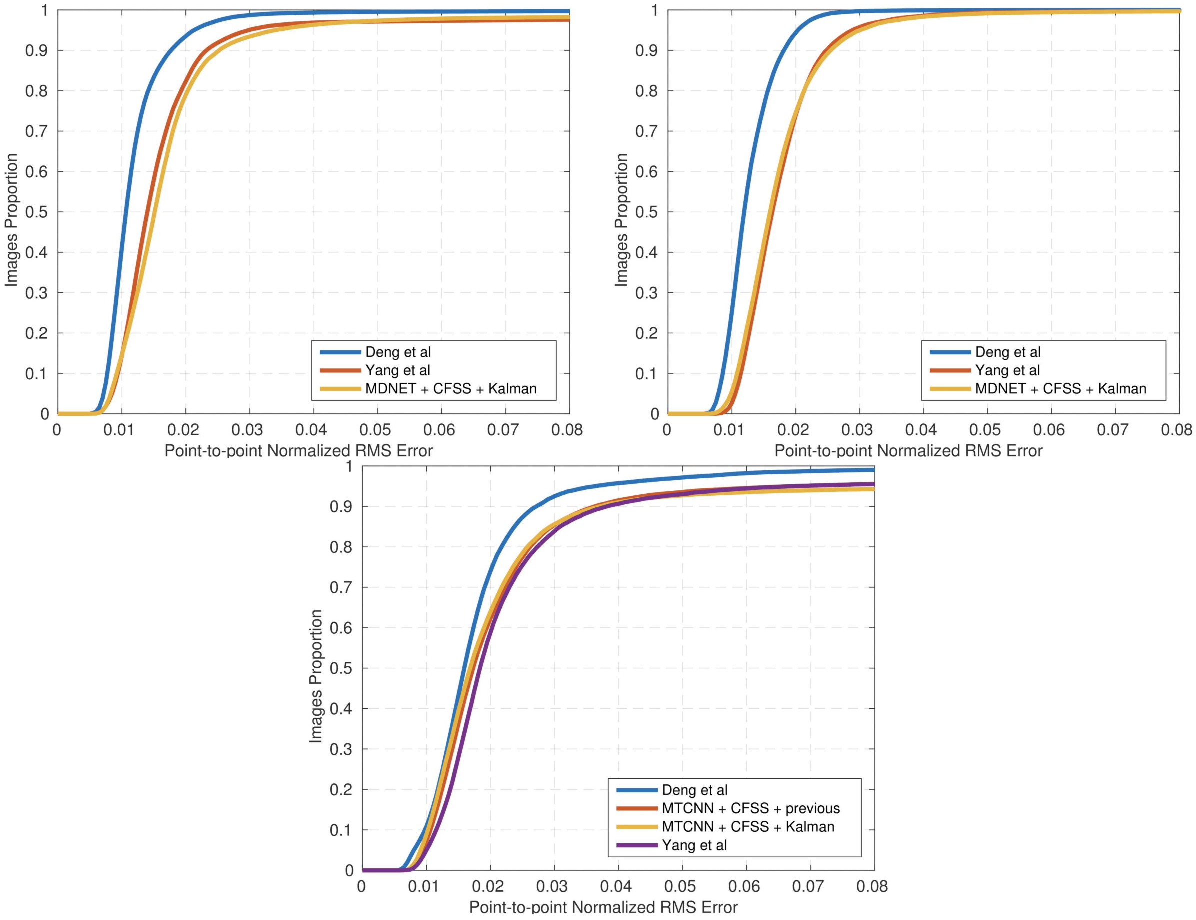

Figure 6: CED curves of the recent state-of-the-art on the test sets, category one (upper-left), category two (upper-right) and category three (down) of the 300-VW benchmark.

[4] E. Antonakos, P. Snape, G. Trigeorgis, and S. Zafeiriou. Adaptive cascaded regression. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016. 1

[5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014. 3

[6] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 35, pages 2930–2940, December 2013. 1, 2

[7] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3D Face Morphable Models ”In-the-Wild”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 1

[8] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 1

[9] A. Bulat and G. Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *Proceedings of the European Conference on Computer Vision Workshop*, pages 616–624. Springer, 2016. 6

[10] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 3

[11] X. Chen, E. Zhou, J. Liu, and Y. Mo. Delving Deep into Coarse-to-fine Framework for Facial Landmark Localization. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[12] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking ”in-the-wild”. *CoRR*, abs/1603.06015, 2016. 2, 6

[13] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[14] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision*, 2016. 1

[15] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference*. 3

[16] J. Deng, Q. Liu, J. Yang, and D. Tao. M 3 csr: multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016. 3

[17] J. Deng, G. Trigeorgis, and S. Zafeiriou. Coarse-to-fine deformable face modelling: A 2.5d approach. *arXiv preprint*, 2017. 5

[18] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Gaussian process domain experts for model adaptation in facial behavior analysis. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshops*, 2016. 1

[19] S. Eleftheriadis, O. Rudovic, and M. Pantic. Joint facial action unit detection and feature fusion: A multi-conditional learning approach. *IEEE Transactions on Image Processing*, 25(12):5727–5742, 2016. 1

[20] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016. 3

[21] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu. Face Detection, Bounding Box Aggregation and Pose Estimation for Robust Facial Landmark Localisation in the Wild. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[22] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *Proceedings of the International Conference on Pattern Recognition and Computer Vision*, 2016. 3, 5

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[24] Z. He, J. Zhang, M. Kan, S. Shan, and X. Chen. Robust FEC-CNN: A High Accuracy Facial Landmark Detection System. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

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

[26] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the International Conference on Computer Vision*, pages 2144–2151. IEEE, 2011. 1

[27] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep Alignment Network: A convolutional neural network for robust face alignment. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[28] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang. Interactive facial feature localization. *Proceedings of the European Conference on Computer Vision*, pages 679–692, 2012. 1, 2

[29] S. Milborrow and F. Nicolls. Active shape models with sift descriptors and mars. In *Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 380–387. IEEE, 2014. 3

[30] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016. 6

[31] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016. 3

[32] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 4

[33] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 1, 2, 3, 4

[34] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 1, 2, 3, 4

[35] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 3

[36] X.-H. Shao, J. Xing, J. Lv, C. Xiao, P. Liu, Y. Feng, C. Cheng, and F. Si. Unconstrained Face Alignment without Face Detection. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[37] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011. IEEE, 2015. 1, 2, 5

[38] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages –. IEEE, 2017. 1

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[40] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1748, 2014. 1

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verifica-

tion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[42] G. Trigeorgis, P. Snape, E. Antonakos, M. A. Nicolaou, and S. Zafeiriou. Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3, 4

[43] W. Wu and S. Yang. Leveraging Intra and Inter-Dataset Variations for Robust Face Alignment. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[44] S. Xiao, J. Li, Y. Chen, Z. Wang, J. Feng, Y. Shuicheng, and A. Kassim. 3D-assisted Coarse-to-fine Extreme-pose Facial Landmark Detection. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[45] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 3

[46] X. Xiong and F. De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015. 6

[47] J. Yang, Q. Liu, and K. Zhang. Stacked Hourglass Network for Robust Facial Landmark Localisation. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[48] A. Zadeh, T. Baltrusaitis, and L.-P. Morency. Convolutional Experts Network for Facial Landmark Detection. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017. 3, 5, 6

[49] S. Zafeiriou, A. Roussos, A. Ponniah, D. Dunaway, and J. Booth. Large scale 3d morphable models. *International Journal of Computer Vision*, 2017. 1

[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 6

[51] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 6

[52] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886. IEEE, 2012. 1, 2, 3