Performance evaluation of face alignment algorithms on "in-the-wild" selfies

Ivan Babanin
Moscow Institute of Physics and Technology, Adorable Inc.
Department of Innovations and High Technology
Institutskiy Pereulok, 9
Russian Federation, 141701, Moscow region, Dolgoprudny
ivan.babanin@phystech.edu

Aleksandr Mashrabov
Moscow Institute of Physics and Technology, Adorable Inc.
Department of Innovations and High Technology
Institutskiy Pereulok, 9
Russian Federation, 141701, Moscow region, Dolgoprudny
mashrabov@phystech.edu

ABSTRACT

Recently mobile apps, which beautify human face or apply cute masks to a human face, become very popular and gain lots of attention in media. These tasks require very precise landmarks localization to avoid "uncanny valley" effect. We introduce the new dataset of selfies, that were taken on mobile devices, and robustly evaluate and compare different state-of-the-art approaches to the task of face alignment. Evidently, our dataset allows to reliably rank face alignment algorithms that is superior to the most popular dataset in that area of research.

Keywords
Benchmark testing, Face, Shape, Machine learning, Robust measurement, Mobile devices, Face alignment

1 INTRODUCTION

The problem of face detection and face alignment has been the focus in computer vision for more than two decades. Recently many research teams have focused on the collection and the annotation of real-world datasets of facial images captured in-the-wild. Such datasets evolve into challenges and encourage many scientists to develop face alignment algorithms that are robust to different pose variations. Although, latest challenges focus on 3D alignment and robust face alignment in a video, although the diversity of datasets with precise annotations for semi-frontal faces is low. However, this case is trendy since people use phones more often than desktop for social media and search on the Internet. This entails the rise of social platforms focused on images messages like Instagram and Snapchat and tools that beautify photos like Snapchat lenses, FaceTune. Another common case that requires exact face alignment is virtual makeup tools. Such applications like Youcam Makeup with more than 100M downloads help to find how you would look if your lips are colored by pink lipstick, if shadows under eyes are green, etc.

We present the first selfies dataset carefully annotated them with 68 fiducial points in a manual manner according to current labeling standards. Our goal is the creation of small dataset to robustly compare state-of-the-art academic and commercial approaches. Also, we check the correlation between overall face alignment quality and quality of tracking key points in specific face areas (mouth, eyes, contour). Example of face annotation is depicted on Figure 1.

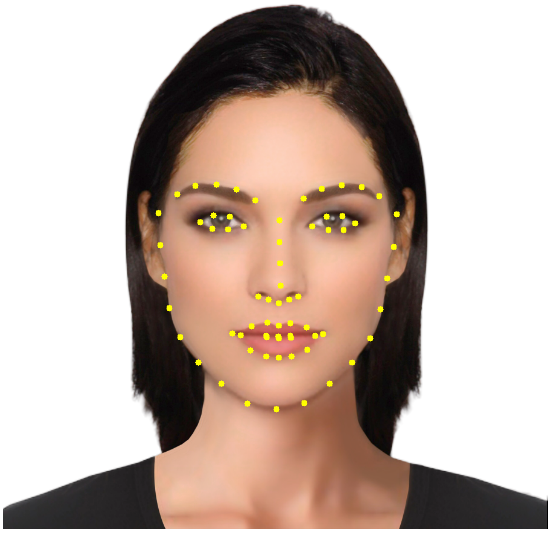

Figure 1: Annotated image with 68 key points [1]

https://doi.org/10.24132/CSRN.2018.2802.10

ISBN 978-80-86943-41-1

2 REVIEW OF EXISTING DATASETS

2.1 Datasets of annotated images

Labeled Face Parts in the Wild (LFPW) database [1] contains 1287 images downloaded from google.com, yahoo.com, and flickr.com. The dataset covers a broad range of appearance variation, including pose, lighting, expression, occlusion, and individual differences. The provided ground truth consists of 35 landmark points.

Helen database [2] contains 2330 images in good resolution downloaded from the flickr.com website. Annotation with 194 landmarks is very precise, but most images are taken not on a mobile phone.

The Annotated Faces in-the-wild (AFW) [3] database which consists of 250 images with 468 faces. Six facial landmark points for each face are provided.

Menpo Challenge database [4] consists of 300W [5] train and test data, iBug dataset. Overall it has 5658 annotated semi-frontal and 1906 annotated profile facial images. Semi-frontal images are provided with 68 landmarks and profile with 39 landmarks. Recently, competition on face alignment was organized on that dataset in July 2017 at top-tier computer vision conference CVPR 2017.

These are the most widely used publicly available databases of images with fiducial points annotation. Although Menpo and Helen databases have enough key points in annotation, original photos in those datasets mostly aren't selfies.

3 RECENT SOLUTIONS

3.1 State-of-the-art academic approaches

W. Wu: Method in [6] used a deep network (VGG-16 and Resnet-18) to regress to a parametric form of the shape of multiple datasets and another network to make the final decision. It showed incredible results in Menpo Challenge 2017 [4] with the 2-nd place and almost real-time performance. The code is not available online; we privately asked authors to evaluate their algorithm on our dataset.

M. Kowalski: Method in [7] used a VGG-based alignment network to correct similarity transforms, extracting features from the entire face images rather than patches around facial key points, and then a fully-convolutional network that finally localizes 68 key points. The code with the pretrained model is available online.

Z. He (Zhenliang): Method in [8] used already known FEC-CNN architecture as a basic method for facial landmark detection with a bounding box invariant algorithm that reduces the prediction sensitivity to face detector and model ensemble technique that is adapted for further performance improvement. The code is not available online; we privately asked authors to evaluate their algorithm on our dataset.

X.-H. Shao: Method in [9] used a sub-network of VGG-19 for landmark heatmap and affinity field prediction at the former stage, and Pose Splitting Layer that regresses basic landmarks at a latter stage. According to its pose, each canonical state is distributed to the corresponding branch of the shape regression sub-networks for the whole landmark detection. The code is not available online; we privately asked authors to evaluate their algorithm on our dataset.

A. Bulat: Method in [10] used a stack of 4 "Hourglass Networks" for landmark localization with a residual block, trained on a very large yet synthetically expanded 2D facial landmark dataset. That leads to remarkable robustness to initialization of parameters and yaw angle of images. The code is open-sourced with pre-trained models.

G. Tzimiropoulos: Method in [13] was implemented in [12] and used parametric linear models of both shape and appearance of an object, typically modeled PCA. The AAM objective function involves the Gauss-Newton minimization of the appearance reconstruction error concerning the shape parameters.

G. Trigeorgis: Method in [17] used a combined and jointly trained convolutional recurrent neural network architecture of cascaded regressors that allows the training of an end-to-end to alleviate problems of existing approaches such as not coherent training process of regressors, the prevalence of handcrafted features. The recurrent module facilitates the joint optimization of the regressors by assuming the cascades are forming a nonlinear dynamical system, in effect, fully utilizing the information between all cascade levels by introducing a memory unit that shares information across all levels. The code is open-sourced.

3.2 Proprietory production systems

Dlib: A very popular fast face alignment library that is widely used as a baseline. It used an ensemble of regression trees under the hood and came with a pre-trained model for 68 facial key points localization. It is open-sourced library and is available at http://dlib.net/.

iOS face alignment: Apple Vision framework that came live with iOS11 in September 2017 provides 65 landmarks. Due to the inconsistency of localization of key points, we compared the accuracy of key points localization only related to mouth region.

4 PROPOSED SOLUTIONS

There are many existing benchmarks for face alignment algorithms, but our goal was to collect a relatively small

[1] http://www.bbc.co.uk/newsbeat/article/32115303/mr-and-mrs-perfect-in-the-real-world

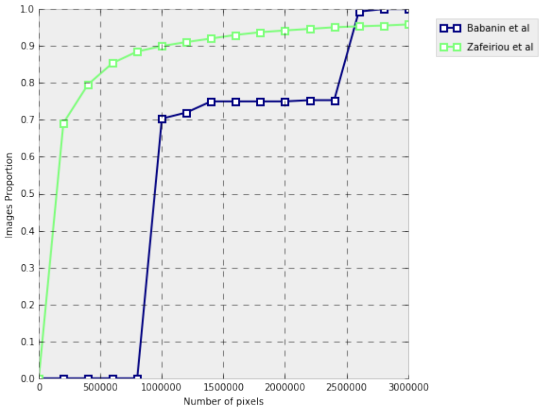

Figure 2: Cumulative distribution of pixels in images

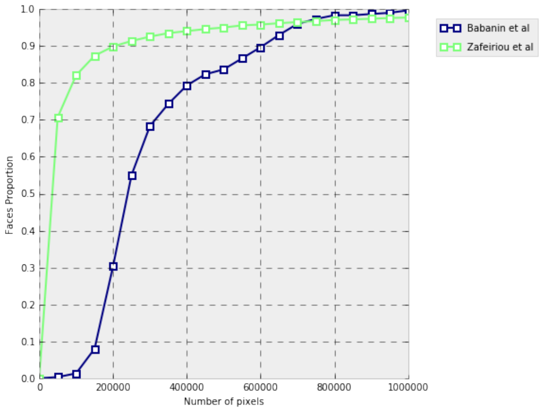

Figure 3: Cumulative distribution of pixels in face rectangles

set of photos that adequately characterize the diversity of selfies. Such images are relatively "easy" compared to almost profile face images [4], so the quality of labeling becomes crucial to make reasonable conclusions. Hence we filtered all photos with an occluded face (by arm, scarf, etc.) and filtered very dark selfies since many popular tasks like face beautification don't make sense in such case.

The number of selfies, that passed initial half-automated filtration, exceeds 5000 images. At last stage, our goal was to ensure the diversity of identities (no more than four photos from each person) that uniformly cover the full range of emotions. At this point, we used 3D Face Morphable Models [15] to fit each image to estimate albedo and shape coefficients. Albedo coefficients describe the identity of a person, helping to limit the number of photos from each person very precisely. Set of shape coefficients describe the full range of emotions [19]; therefore, we applied Principal Components Analysis algorithm [16] to this set to select the photos that demonstrate the diversity of emotions in real-life. Whereas, we used open-source library 4dface [18] to fit each image to 3D face model. 4dface framework operates with local features rather than rough pixel values that results in much more robust fitting against variations in images conditions. The final dataset contains only 300 photos, that allows to compute final metrics very quickly.

We collected dataset of selfies taken on mobile phones by users of the mobile application on behalf of Adorable Inc. All photos were taken on frontal camera and had a resolution at least 720*1080 that is bigger than the majority of images in Menpo dataset [4]. More specifically, 69 percents of photos in Menpo have a resolution less than 200,000 pixels. Thus the majority of pictures in current popular benchmarks is four times smaller than images in our dataset (see Figure 2).

Furthermore, we compared the area of face rectangles in our dataset and Menpo dataset (see Figure 3). 70 percents of face rectangles in Menpo dataset [4] has the area less than 50,000 pixels, although 70 percents of face rectangles in our dataset have the area more than 200,000 pixels.

5 EVALUATION METRICS

In the biggest competitions on face alignment main metric for evaluation is the point-to-point Euclidean distance normalized by the interocular distance [5]. However, as noted in [3], this error metric doesn't provide robust results for profile faces with small interocular distance. Hence, we propose two types of normalizations. In particular, we used the Normalized Mean Error (Normalized Point-to-Point error) defined as:

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{|gt_i - pr_i|_2}{d}, \tag{1}$$

where gt denotes the ground truth landmarks for a given face, pr - the corresponding prediction. And d is:

The diagonal ground truth bounding box [11], computed as $d = \sqrt{w^2_{facebbox} + h^2_{facebbox}}$. This normalization is standard.

The square-root (geometric) of the ground truth bounding box of corresponding face region, computed as $d = \sqrt{w_{bbox} * h_{bbox}}$. This new type of normalization depends on characteristic values of that particular region (e.g. size of mouth is much smaller than size of face).

Moreover, our goal is to compare alignment of different face regions: mouth (N=20 points), eyes and brows (N=22 points), contour (N=17 points), so that we have six error values to compare aforementioned face alignment approaches.

However, as noted in [4] mean errors without corresponding standard deviations are not reliable metrics to compare approaches and to make reasonable conclusions. Therefore, we provide our evaluation in the form of cumulative error distribution (CED) curves. After that we find the area-under-the-curve (AUC) taking

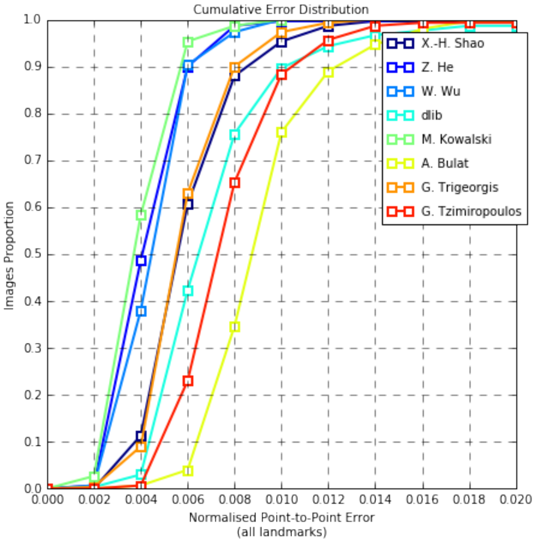

Figure 4: CED curve for entire face, diagonal normalization

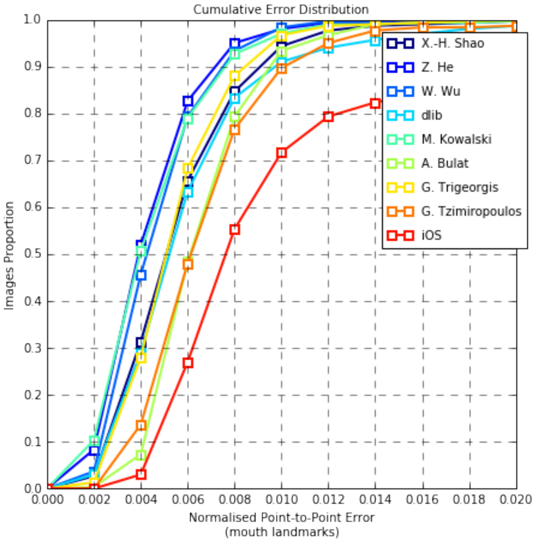

Figure 6: CED curve for mouth region, diagonal normalization

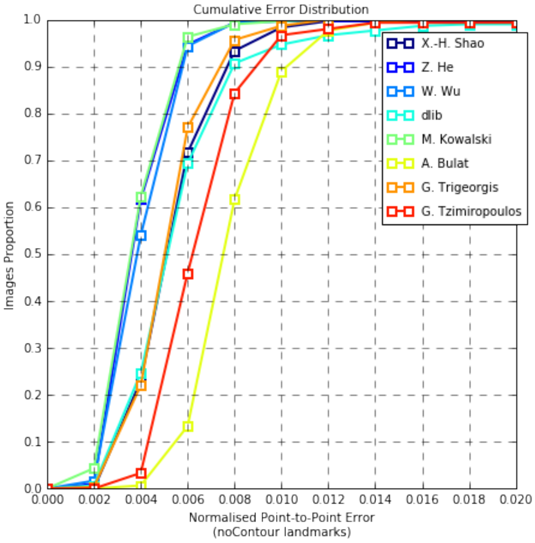

Figure 5: CED curve for entire face without contour, diagonal normalization

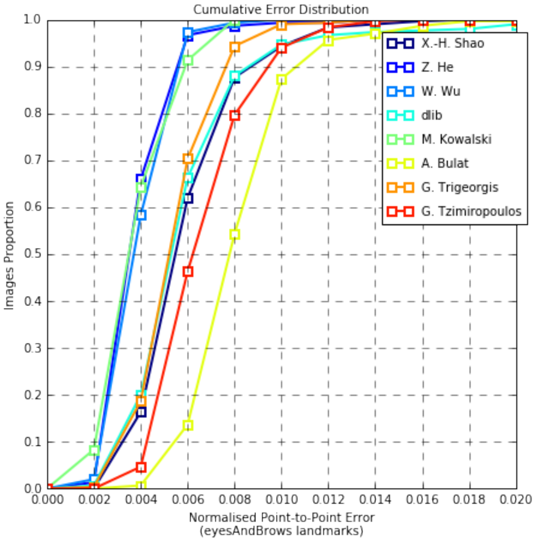

Figure 7: CED curve for eyes and brows region, diagonal normalization

only those images that have the error less than 0.03 [20]. Besides, that error rate for geometric normalization and mouth, eyes and brows, no controur region is 0.30. Another important metric is the failure rate of each method that is a proportion of images with error more than 0.03, which describes very poor face alignment that cannot be used for any further face modification like digital makeup.

6 EXPERIMENTAL RESULTS

In this section we will describe key observations, validate our hypothesis and show that declared goals are achieved. Bulat et al. [10] performed much worse than other methods. Also, there is a tremendous gap between state-of-the-art methods that use complicated Deep Learning approaches and old-fashioned regression methods, decision trees methods. This observation was already noted in [4]. We compared all approaches with first types of normalization and found out that the ranking is almost the same for different face regions (Figure 4 for all landmarks, Figure 5 for all landmarks except contour, Figure 6 for mouth landmarks, Figure 7 for brows and eyes landmarks). The huge advantage of our approach compared to Menpo challenge [4] is much smaller deviation at the much smaller size (300 vs 5335). Evidently, Kowalski et al. [7] showed the best result on our dataset: deviation on our dataset is 1/3 of a mean value, but in Menpo dataset deviation is more than a mean value. Since that ranking of results in Menpo dataset is not reliable and our approach allows to compare algorithms more consistently.

	Bulat[10]	Tzim.[13]	Kow.[7]	dlib	Trig.[17]	Shao[9]	Wu[6]
Tzim.[13]	1e-28	—	—	—	—	—	—
Kow.[7]	6e-51	6e-51	—	—	—	—	—
dlib	4e-36	6e-11	6e-51	—	—	—	—
Trig.[17]	5e-50	3e-44	2e-50	1e-26	—	—	—
Shao[9]	3e-49	1e-45	6e-51	5e-22	1e-03	—	—
Wu[6]	6e-51	6e-51	7e-36	7e-51	1e-48	2e-49	—
He[8]	6e-51	6e-51	9e-20	8e-51	2e-48	4e-50	1e-11

Table 1: Wilcoxon test for all 68 keypoints with first normalization

Surprisingly, mean error and a standard deviation are very similar (Figure 4, Figure 5, Table 2, Table 4) on 68 key points (entire face) and 41 key points (without contour). Our initial hypothesis was that it is difficult to make labeling of contour landmarks consistent. Therefore we expected that error on entire face without contour region would be much less. It turned out to be false.

The only region that suits for comparing keypoint localization algorithm employed in iOS is mouth region (Figure 6, Table 6). Anyway, the quality of that algorithm is clearly very poor. In fact, failure rate of iOS algorithm is more than 10 percents, when failure rate of other algorithms is less than 1 percent. Additionally, the only method with deviation more than mean value is the iOS algorithm (Table 6, Table 7). Also, Bulat et al. [10] bypass Tzimiropoulos et al. [13].

Another region that has slightly different ranking is eyes and brows region (Figure 7, Table 8). There is almost indistinguishable difference between leader He et al. [8] and runner-up Kowalski et al. [7].

Moreover, we compared all algorithms to each other using Wilcoxon signed-rank test to assure that our method produces reliable results and allows to compare algorithms. For each image, we computed error rate on 68 key points (first type of normalization by diagonal of face rectangle) and ran the test on 300 pairs of values (see Table 1).

Another part of our research consists of comparing two types of normalization. That second type is geometric normalization by taking a square root of sides of a corresponding face region. This almost doesn't affect the ranking of all face regions, but relative deviation becomes much smaller for mouth region (Table 6, Table 7) and remains the same for other regions.

7 CONCLUSION

We achieved our goal to create a small dataset that allows to efficiently and robustly rank and differentiate current state-of-the-art face alignment approaches. From our best knowledge it is the only such dataset. Summing up, the quality of face tracking of popular proprietory systems is far worse than top-level academic approaches. The quality of method by M.Kowalski et al. [7] shows excellent results from qualitative and quantitative points.

In our dataset, the overall mean error is smaller than in [4] that is an implication of nature of photos (well lighting, not extreme head rotation poses). The important observation is that quality of key points localization of different face regions (eyes, mouth, contour) highly correlates with quality on entire face. Another significant comment is that we achieved much smaller deviation without artificial clipping of photos with large head rotations. We believe that there is still a room for research to create a relevant small dataset with accurate labeling that represents the full diversity of face poses not limited to selfies. Our goal for further research is to create openly available benchmark for 3D landmark tracking on "in-the-wild" selfies.

8 ACKNOWLEDGMENT

Thanks to Adorable Inc. for providing access to data with face images.

9 REFERENCES

[1] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, IEEE Transactions on Pattern Analysis and Machine Intelligence (T- PAMI), 35(12), 2930-2940, 2013.

[2] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, In Proceedings of European conference on computer vision (ECCV) (pp. 679-692) Springer, 2012.

[3] X. Zhu, D. Ramanan, Face Detection, Pose Estimation, and Landmark Localization in the Wild, In CVPR 2012, 2012.

[4] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen, The Menpo Facial Landmark Localisation Challenge: A step towards the solution', In CVPRW 2017, 2017.

[5] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, 300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge, In ICCV 2013, 2013.

[6] W. Wu, S. Yang, Leveraging Intra and Inter-Dataset Variations for Robust Face Alignment,

In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge, 2017.

[7] 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.

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

[9] 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 and Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge, 2017.

[10] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), ICCV 2017, 2017.

[11] G. Chrysos, E. Antonakos. P. Snape, A. Asthana, S. Zafeiriou, A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild", International Journal of Computer Vision, 2017.

[12] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, S. Zafeiriou, Menpo: A comprehensive platform for parametric image alignment and visual deformable models, In Proceedings of ACM international conference on multimedia, (ACM'MM) (pp. 679-682). ACM, 2016.

[13] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, M. Pantic, Active orientation models for face alignment in-the-wild, IEEE Transactions on Information Forensics and Security, 9(12), 2024-2034, 2014.

[14] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, In IEEE proceedings of international conference on computer vision and pattern recognition (CVPR), (pp. 532-539), 2013.

[15] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D Face Model for Pose and Illumination Invariant Face Recognition, In Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments, Genova (Italy) - September 2-4, 2009.

[16] J. Shlens, A Tutorial on Principal Component Analysis, `https://www.cs.cmu.edu/~elaw/papers/pca.pdf`, 2004.

[17] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, S. Zafeiriou, Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment, Proceedings of IEEE International Conference on Computer Vision Pattern Recognition (CVPR 16), Las Vegas, NV, USA, June 2016.

[18] P. Huber, Z. Feng, W. Christmas, J. Kittler, M. Ratsch, Fitting 3D Morphable Models using Local Features, IEEE International Conference on Image Processing (ICIP 15), Quebec City, Canada, 2015.

[19] M. Yu, B. P. Tiddeman, Facial Feature Detection and Tracking with a 3D Constrained Local Model, WSCG 2010 conference proceedings, Copyright UNION Agency Science Press, pp: 181-188, WSCG 2010.

[20] H. Yang, X. Jia, C. C. Loy, P. Robinson, An Empirical Study of Recent Face Alignment Methods, arXiv preprint arXiv:1511.05049, 2015.

	Mean	Std	Median	MAD	Max Error	$AUC_{0.03}$
M. Kowalski *et al.* [7]	0.0039	0.0012	0.0038	0.0008	0.0091	0.8706
Z. He *et al.* [8]	0.0043	0.0014	0.0040	0.0007	0.0150	0.8571
W. Wu *et al.* [6]	0.0045	0.0013	0.0043	0.0007	0.0097	0.8511
G. Trigeorgis *et al.* [17]	0.0058	0.0016	0.0055	0.0009	0.0121	0.8083
X.-H. Shao *et al.* [9]	0.0059	0.0020	0.0055	0.0012	0.0200	0.8025
dlib	0.0071	0.0032	0.0063	0.0011	0.0271	0.7647
G. Tzimiropoulos *et al.* [13]	0.0076	0.0024	0.0072	0.0012	0.0237	0.7453
A. Bulat *et al.* [10]	0.0091	0.0025	0.0086	0.0011	0.0242	0.6982

Table 2: Entire face, diagonal normalization

	Mean	Std	Median	MAD	Max Error	$AUC_{0.03}$
M. Kowalski *et al.* [7]	0.0056	0.0018	0.0054	0.0011	0.0131	0.8146
Z. He *et al.* [8]	0.0061	0.0020	0.0058	0.0010	0.0214	0.7951
W. Wu *et al.* [6]	0.0064	0.0018	0.0061	0.0010	0.0139	0.7862
G. Trigeorgis *et al.* [17]	0.0082	0.0023	0.0078	0.0013	0.0171	0.7251
X.-H. Shao *et al.* [9]	0.0085	0.0028	0.0080	0.0017	0.0286	0.7168
dlib	0.0101	0.0045	0.0090	0.0016	0.0387	0.6650
G. Tzimiropoulos *et al.* [13]	0.0110	0.0034	0.0104	0.0018	0.0338	0.6349
A. Bulat *et al.* [10]	0.0130	0.0035	0.0123	0.0015	0.0346	0.5677

Table 3: Entire face, geometric normalization

	Mean	Std	Median	MAD	Max Error	$AUC_{0.03}$
M. Kowalski *et al.* [7]	0.0038	0.0013	0.0037	0.0008	0.0104	0.8739
Z. He *et al.* [8]	0.0039	0.0014	0.0037	0.0006	0.0180	0.8699
W. Wu *et al.* [6]	0.0040	0.0012	0.0039	0.0007	0.0109	0.8650
G. Trigeorgis *et al.* [17]	0.0051	0.0014	0.0049	0.0009	0.0110	0.8301
X.-H. Shao *et al.* [9]	0.0053	0.0019	0.0050	0.0011	0.0214	0.8225
dlib	0.0057	0.0031	0.0051	0.0011	0.0304	0.8113
G. Tzimiropoulos *et al.* [13]	0.0064	0.0021	0.0061	0.0011	0.0237	0.7859
A. Bulat *et al.* [10]	0.0078	0.0018	0.0076	0.0011	0.0156	0.7414

Table 4: Entire face without contour, diagonal normalization

	Mean	Std	Median	MAD	Max Error	$AUC_{0.30}$
M. Kowalski *et al.* [7]	0.0140	0.0041	0.0136	0.0022	0.0320	0.9534
Z. He *et al.* [8]	0.0146	0.0044	0.0140	0.0019	0.0619	0.9515
W. Wu *et al.* [6]	0.0151	0.0034	0.0149	0.0019	0.0376	0.9497
G. Trigeorgis *et al.* [17]	0.0191	0.0046	0.0184	0.0024	0.0405	0.9365
X.-H. Shao *et al.* [9]	0.0199	0.0064	0.0188	0.0033	0.0737	0.9336
dlib	0.0210	0.0102	0.0189	0.0033	0.1066	0.9300
G. Tzimiropoulos *et al.* [13]	0.0240	0.0066	0.0231	0.0031	0.0808	0.9202
A. Bulat *et al.* [10]	0.0290	0.0053	0.0281	0.0027	0.0667	0.9034

Table 5: Entire face without contour, geometric normalization

	Mean	Std	Median	MAD	Max Error	$AUC_{0.03}$
Z. He *et al.* [8]	0.0043	0.0023	0.0039	0.0012	0.0224	0.8566
M. Kowalski *et al.* [7]	0.0045	0.0026	0.0039	0.0014	0.0204	0.8512
W. Wu *et al.* [6]	0.0046	0.0020	0.0042	0.0012	0.0154	0.8468
G. Trigeorgis *et al.* [17]	0.0054	0.0024	0.0050	0.0013	0.0193	0.8204
X.-H. Shao *et al.* [9]	0.0055	0.0028	0.0049	0.0014	0.0253	0.8156
dlib	0.0060	0.0037	0.0052	0.0016	0.0278	0.7988
A. Bulat *et al.* [10]	0.0065	0.0023	0.0061	0.0012	0.0213	0.7831
G. Tzimiropoulos *et al.* [13]	0.0068	0.0037	0.0061	0.0015	0.0412	0.7748
iOS	0.0193	0.0682	0.0074	0.0020	0.6440	0.6572

Table 6: Mouth landmark region, diagonal normalization

	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.30}$
Z. He *et al.* [8]	0.0550	0.0214	0.0522	0.0123	0.1984	0.8167
M. Kowalski *et al.* [7]	0.0565	0.0244	0.0545	0.0150	0.1869	0.8118
W. Wu *et al.* [6]	0.0591	0.0193	0.0571	0.0125	0.1367	0.8029
G. Trigeorgis *et al.* [17]	0.0696	0.0232	0.0671	0.0128	0.1627	0.7679
X.-H. Shao *et al.* [9]	0.0711	0.0268	0.0665	0.0145	0.2243	0.7631
dlib	0.0769	0.0362	0.0678	0.0149	0.2573	0.7437
A. Bulat *et al.* [10]	0.0845	0.0209	0.0824	0.0116	0.2503	0.7185
G. Tzimiropoulos *et al.* [13]	0.0879	0.0428	0.0825	0.0127	0.6046	0.7103
iOS	0.2575	1.0055	0.0989	0.0232	12.5684	0.5756

Table 7: Mouth landmark region, geometric normalization

	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.03}$
Z. He *et al.* [8]	0.0038	0.0012	0.0037	0.0006	0.0125	0.8734
M. Kowalski *et al.* [7]	0.0038	0.0014	0.0036	0.0008	0.0078	0.8731
W. Wu *et al.* [6]	0.0039	0.0011	0.0037	0.0007	0.0095	0.8698
G. Trigeorgis *et al.* [17]	0.0053	0.0016	0.0051	0.0010	0.0132	0.8230
X.-H. Shao *et al.* [9]	0.0059	0.0021	0.0055	0.0012	0.0174	0.8041
dlib	0.0059	0.0040	0.0052	0.0011	0.0484	0.8039
G. Tzimiropoulos *et al.* [13]	0.0066	0.0020	0.0062	0.0011	0.0145	0.7816
A. Bulat *et al.* [10]	0.0081	0.0023	0.0078	0.0013	0.0256	0.7307

Table 8: Eyes and Brows landmark region, diagonal normalization

	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.30}$
M. Kowalski *et al.* [7]	0.0244	0.0075	0.0237	0.0052	0.0509	0.9188
Z. He *et al.* [8]	0.0246	0.0071	0.0240	0.0036	0.0793	0.9181
W. Wu *et al.* [6]	0.0251	0.0059	0.0247	0.0039	0.0531	0.9162
G. Trigeorgis *et al.* [17]	0.0343	0.0091	0.0328	0.0050	0.0862	0.8856
X.-H. Shao *et al.* [9]	0.0379	0.0123	0.0354	0.0065	0.1103	0.8738
dlib	0.0383	0.0241	0.0333	0.0066	0.2949	0.8725
G. Tzimiropoulos *et al.* [13]	0.0421	0.0099	0.0407	0.0056	0.0856	0.8596
A. Bulat *et al.* [10]	0.0522	0.0129	0.0504	0.0057	0.1676	0.8261

Table 9: Eyes and Brows landmark region, geometric normalization