Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild

Zhen-Hua Feng[1]   Patrik Huber[1]   Josef Kittler[1]   Peter Hancock[2]   Xiao-Jun Wu[3]
Qijun Zhao[4]   Paul Koppen[1]   Matthias Rätsch[5]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
[2]Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK
[3]School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
[4]Biometrics Research Lab, College of Computer Science, Sichuan University, Chengdu 610065, China
[5]Image Understanding and Interactive Robotics, Reutlingen University, 72762 Reutlingen, Germany

{z.feng, j.kittler, p.koppen}@surrey.ac.uk, patrikhuber@gmail.com,
wu_xiaojun@jiangnan.edu.cn, p.j.b.hancock@stir.ac.uk, qjzhao@scu.edu.cn

Abstract—This paper investigates the evaluation of dense 3D face reconstruction from a single 2D image in the wild. To this end, we organise a competition that provides a new benchmark dataset that contains 2000 2D facial images of 135 subjects as well as their 3D ground truth face scans. In contrast to previous competitions or challenges, the aim of this new benchmark dataset is to evaluate the accuracy of a 3D dense face reconstruction algorithm using *real*, accurate and high-resolution 3D ground truth face scans. In addition to the dataset, we provide a standard protocol as well as a Python script for the evaluation. Last, we report the results obtained by three state-of-the-art 3D face reconstruction systems on the new benchmark dataset. The competition is organised along with the 2018 13th IEEE Conference on Automatic Face & Gesture Recognition.

I. Introduction

3D face reconstruction from 2D images is a very active topic in many research areas such as computer vision, pattern recognition and computer graphics [1], [2], [3], [4], [5], [6]. While the topic has been researched for nearly two decades (with one of the seminal papers being Blanz & Vetter [1]), over the last two years, these methods have been growing out of laboratory-applications and have become applicable to in-the-wild images, containing larger pose variations, difficult illumination conditions, facial expressions, or different ethnicity groups and age ranges [7], [8], [9]. However, it is currently an ongoing challenge to quantitatively evaluate such algorithms: For 2D images captured in the wild, there is usually no 3D ground truth available. And, vice-versa, for 3D data, it is usually captured with a 3D scanner in a laboratory, and no in-the-wild 2D images of the same subject exist.

Thus, researchers often publish qualitative results, alongside some effort of quantitative evaluation, which is often not ideal, for the lack of 3D ground truth. Also, people resort to using a proxy task, for example face recognition. For an example of the former, people are using the Florence 2D/3D hybrid face dataset (MICC) [10], which contains 3D data with 2D videos (*e.g.* [11], [12]). However there is no standard evaluation protocol, and very often synthetic renderings of the 3D scans are used for evaluation, which do not contain background (*e.g.* they are rendered on black background) or any natural illumination variations (*e.g.* [13]), for the lack of better data. Other methods (*e.g.* 3DDFA [7]) compare their results against a 'ground truth' which is created by another fitting algorithm, which is itself problematic as these fitting algorithms have not yet been shown to be effective on in-the-wild data, even after manual correction. There have been a limited number of previous competitions which aimed to improve the situation, but they only solved the problem partially. For example the workshop organised by Jeni et al. [14] used their own algorithm as 'ground truth' (see also Section I-C). Other datasets have been recently proposed like KF-ITW [15] but therein Kinect Fusion is used as 3D ground truth, which does not consist of very high resolution meshes, and also the videos are recorded in rather controlled and similar scenarios (*i.e.* rotating around a chair in a lab).

In this paper, we report the results of a competition on 3D dense face reconstruction of in-the-wild 2D images, evaluated with accurate and high-resolution 3D ground truth, obtained from a 3D structured-light system. The competition is co-located with a workshop[1] of the 13th IEEE Conference on Automatic Face & Gesture Recognition (FG 2018).

A. Outcomes

- The competition provides a benchmark dataset with 2000 2D images of 135 subjects as well as their high-resolution 3D ground truth face scans. Alongside the dataset we supply a standard benchmark protocol to be used on the dataset, for future evaluations and comparisons, beyond the competition.
- An independent, objective evaluation and comparison of state-of-the art 3D face reconstruction algorithms. The plan is to perform two sets of evaluations: One set for single-image reconstruction, and another set where

[1]https://facer2vm.org/fg2018/

it is allowed to use all images of one particular person to reconstruct the 3D shape, allowing algorithms to leverage information from multiple images. Note that, in this paper, we only report results of the single image fitting protocol.

B. Impact

This is the first challenge in 3D face reconstruction from single 2D in-the-wild images with *real*, accurate and high-resolution 3D ground truth.

The provided benchmark dataset is publicly available, so that it can become a benchmark and reference point for future evaluations in the community.

The multi-image challenge allows to test algorithms that can work with multiple videos as well, having far-reaching impact, for example also in the face recognition community (*e.g.* for set-to-set matching, and recent 2D face recognition benchmarks such as the IARPA Janus Benchmark face challenge[2]).

In addition to that, one of the baseline 3D reconstruction algorithms and the *Surrey Face Model* (SFM) is publicly available too [16].

C. Relationship to previous workshops (competitions)

The topic of evaluating 3D face reconstruction algorithms on 2D in-the-wild data has gained much traction recently. The *1st Workshop on 3D Face Alignment in the Wild (3DFAW) Challenge*[3] [14] was held at ECCV 2016. The benchmark consisted of images from Multi-PIE, synthetically rendered images, and some in-the-wild images from the internet. The 3D 'ground truth' was generated by an automatic algorithm provided by the organisers.

As part of ICCV 2017, the iBUG group from Imperial College, UK, held a workshop *1st 3D Face Tracking in-the-wild Competition*[4]. It improved upon the ECCV 2016 challenge in some respects, but the 'ground truth' used was still from an automatic fitting algorithm, introducing bias, and resulting in the other algorithms being evaluated against the performance of another algorithm, and not against real 3D ground truth. Also, the evaluation is only done on a set of sparse 2D and 3D landmarks and not over a dense 3D mesh, leaving much room for further improvements on the benchmarking methodology.

The remaining of this paper outlines the data, protocol, evaluation metrics and results of the competition. The aim of the competition is to evaluate 3D face shape reconstruction performance of participants on true 2D in-the-wild images, with actual 3D ground truth available from 3D face scanners. The data is released to the public, together with a well-defined protocol, to provide a standard and public benchmark to the 3D face reconstruction community.

(a) high quality images

(b) low quality images

Figure 1.   Some examples of the 2D images in the test set, selected from the Stirling ESRC 3D face dataset.

II. Datasets

In general, the data used for the evaluation of a 3D face reconstruction algorithm should consist of a number of high-resolution 3D face scans, obtained from a 3D face imaging device, like for example the 3dMDface[5] system. Together with these 3D face ground truth, associated with each subject are multiple 2D images that have been captured in-the-wild, with a variety of appearance variations in pose, expression, illumination and occlusion. The aim is to measure the accuracy of an algorithm in reconstructing a subject's neutral 3D face mesh from unconstrained 2D images. To this end, the Stirling ESRC 3D face dataset[6] is used to create the test set and the JNU [6] 3D face dataset is used to form the validation set. For training, any 2D or 3D face dataset is allowed except for the Stirling ESRC and JNU datasets.

A. Test set

The test set is a subset of the Stirling ESRC face database that contains a number of 2D images, video sequences as well as 3D face scans of more than 100 subjects. The 2D and 3D faces in the Stirling ESRC dataset were captured under 7 different expression variations. To create the test set, 2000 2D neutral face images, including 656 high-quality and 1344 low-quality images, of 135 subjects were selected from the Stirling ESRC 3D face database. The high quality images were captured in constrained scenarios with good lighting conditions. The resolution of a high quality face image is higher than 1000×1000. In contrast, the low quality

images or video frames were captured with a variety of image degradation types such as image blur, low resolution, poor lighting, large scale pose rotation etc. Some examples of the selected 2D face images are shown in Fig. 1.

B. Validation set

Participants are allowed to use the validation set to fine-tune the hyper-parameters of their 3D face reconstruction systems, if required. The validation set contains 161 2D in-the-wild images of 10 subjects and their ground-truth 3D face scans. The validation set is a part of the JNU 3D face dataset collected by the Jiangnan University using a 3dMDface system. The full JNU 3D face dataset has the high resolution 3D face scans of 774 Asian subjects. For more details of the JNU 3D face dataset, please refer to [6].

III. PROTOCOL & EVALUATION METRICS

This section details the exact protocol, rules, and evaluation metrics to be used for the competition.

A. Protocol

The task of the competition is to reconstruct a subject's neutral 3D face shape from a single 2D input image. Multi-image fitting that allows the use of multiple input images of the same subject for the task of reconstruction of the subject's neutral face is not included in this competition. However, we leave this potential evaluation protocol as our future work.

For single image reconstruction, an algorithm is expected to take as input a single 2D image and output a neutral 3D face shape of the subject's identity. An algorithm should be run on each of the images of each subject individually. In addition, one set of parameters has to be used for all images of all subjects. No fine-tuning is allowed on a per-image or per-subject basis.

B. Evaluation metrics

Given an input 2D image or a set of 2D images, we use the 3D Root-Mean-Square Error (3D-RMSE) between the reconstructed 3D face shape and the ground truth 3D face scan calculated over an area consisting of the inner face as the evaluation metric. The area is defined as each vertex in the ground truth scan that is inside the radius from the face centre. The face centre is computed as the point between the annotated *nose bottom* point and the *nose bridge* (which is computed as the middle of the two annotated eye landmarks): face_centre = nose_bottom + $0.3 \times$ (nose_bridge − nose_bottom). The radius is computed as the average of the outer-eye-distance and the distance between nose bridge and nose bottom, times a factor of 1.2: radius = $1.2 \times$ (outer_eye_dist + nose_dist)$/2$. The radius is defined in a relative way because it is desired that the radius covers roughly the same (semantic) area on each scan, and we would like to avoid that e.g. with a very wide face, the
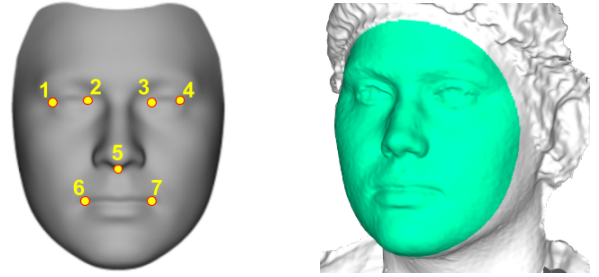

Figure 2. **Left**: The pre-defined seven landmarks used for the rigid alignment of the predicted face mesh with its ground-truth. In order: 1) right eye outer corner, 2) right eye inner corner, 3) left eye inner corner, 4) left eye outer corner, 5) nose bottom, 6) right mouth corner, and 7) left mouth corner. **Right**: The area over which face reconstruction is evaluated is defined for each ground-truth 3D scan by a radius around the face centre. This radius is relative to the subject's inter-ocular and eye-nose distance (see Section III-B for details).

evaluated area would cover a smaller part of that particular face. Typically, the resulting radius is around 80 mm. The area is depicted for an example scan in Figure 2.

The following steps are performed to compute the 3D-RMSE between two meshes:

1) The predicted and ground truth meshes will be rigidly aligned (by translation, rotation, and scaling). Scaling is compensated for because participants' resulting meshes might be in a different coordinate system, whereas the ground truth scans are in the unit of millimetres. The rigid alignment is based on seven points: both inner and outer eye corners, the nose bottom and the mouth corners (see Fig. 2). The ground truth scans have been annotated with these seven points, whereas participants are expected to specify those on their resulting meshes.

2) For each vertex in the ground truth 3D face scan, the distance is computed to the closest point on the surface of the predicted mesh. These distances are used to compute the 3D-RMSE as well as more specific analysis of a 3D face reconstruction system, *e.g.* the distribution of errors across different face regions.

A Python script is provided[7], performing the alignment and distance computations. The output of the script is an ordered list of distances.

IV. SUMMARY OF APPROACHES

Three 3D face reconstruction systems have been evaluated for the competition, including the system submitted by the Biometrics Research Lab at the Sichuan University (SCU-BRL) [17] and two baseline systems implemented by the competition organisers from the University of Surrey. Results were provided in the form of text files with per-vertex errors. In addition, participants were asked to provide

---

[7]https://github.com/patrikhuber/fg2018-competition

a brief summary of their approach. The descriptions below are based upon these provided summaries. A thorough comparison will then be presented in Section V.

A. University of Surrey

The baseline systems developed by the Centre for Vision, Speech and Signal Processing (CVSSP) from the University of Surrey have three main stages: face detection, facial landmark localisation and 3D face reconstruction.

1) Face detection: For face detection, the Multi-Task CNN (MTCNN) face detector was adopted to obtain the bounding box for each input 2D face image [18]. However, the faces of some low resolution images with extreme pose variations were missed by MTCNN. For those face images, a bounding box regression approach was used to obtain the face bounding box, as described in [19], [20].

2) Facial landmark detection: To perform facial landmark localisation, the CNN6 model [21], a simple CNN-based facial landmark localisation algorithm, was adopted. The model was trained on multiple in-the-wild face datasets, including the HELEN [22], LFPW [23], ibug [24] and AFW [25] datasets. These datasets have 3837 2D face images and each image has 68 facial landmarks annotated by the iBUG group from Imperial College London. In addition, a subset of the Multi-PIE [26] face dataset was also used for the CNN6 model training. This Multi-PIE subset has 25,200 2D face images and each face image was manually annotated using 68 facial landmarks [27]. Some examples from the Stirling low quality subset with the detected 68 facial landmarks using the CNN6 are shown in Fig. 3.

3) 3D face reconstruction: Given an input 2D image as well as its 2D facial landmarks, the eos fitting algorithm is used to recover the 3D face of the input [16]. The eos fitting algorithm reconstructs the 3D face shape based on the landmarks, using a 3D morphable shape and expression model. It consists of shape identity and blendshapes fitting, a scaled orthographic projection camera model, and a dynamic face contour fitting. For this evaluation, the SFM_3448 shape-only model was used, with the 6 standard Surrey expression blendshapes. The fitting was run for 5 iterations, fitting all shape coefficients of the model, and with a shape regularisation parameter of $\lambda = 30$. We use the term 'MTCNN-CNN6-eos' for this system.

The eos fitting algorithm is tailored for real-time 3D face fitting applications with a speed of more than 100 fps on a single CPU. It only relies on 2D facial landmarks, and does not use any texture information. Therefore, it would be interesting to explore more complicated 3D face model fitting algorithms exploiting textural information. To this end, we also evaluated the 3DDFA 3D face model fitting algorithm [7]. To be more specific, we first use the MTCNN and CNN6 to obtain the same 68 facial landmark for an input 2D image. Then the 68 landmarks are used to initialise
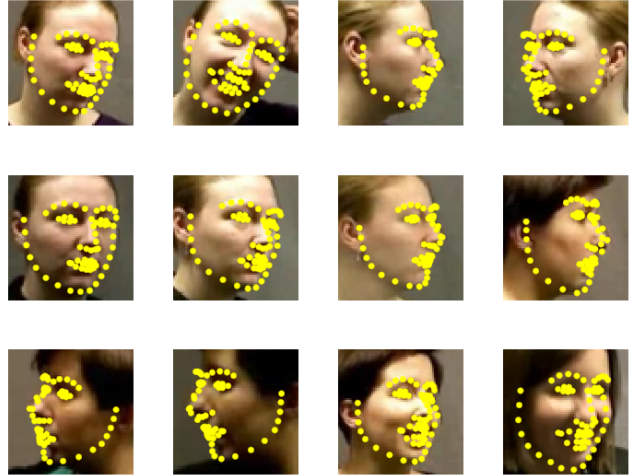

Figure 3. Some examples of the detected 68 landmarks by CNN6.

the 3DDFA fitting algorithm provided by its authors[8]. As face model, the 3DDFA fitting uses a modified Basel Face Model [28] with expressions. We use the term 'MTCNN-CNN6-3DDFA' for this system.

B. Sichuan University (SCU-BRL)

The system developed by the Biometrics Research Lab at the Sichuan University is based on a novel method that is able to reconstruct 3D faces from arbitrary number of 2D images using 2D facial landmarks. The method is implemented via cascaded regression in shape space. It can effectively exploit complementary information in unconstrained images of varying poses and expressions. It begins with extracting 2D facial landmarks on the images, and then progressively updates the estimated 3D face shape for the input subject via a set of cascaded regressors, which are off-line learned based on a training set of pairing 3D face shapes and unconstrained face images.

1) Facial landmark detection: For 2D facial landmark detection, the state-of-the-art Face Alignment Network (FAN) [8] was adopted by in the system developed by SCU-BRL. For some low quality images, FAN failed, and they manually annotated the 68 landmarks for them.

2) 3D face reconstruction: Given an arbitrary number of unconstrained face images $\{I_i\}_{i=1}^{p}, 1 \leq p \leq N$ of a subject, the goal is to reconstruct the person-specific frontal and neutral 3D face shape of the subject. We represent the 3D face shape by $\mathbf{S} \in \mathbb{R}^{3 \times q}$ based on the 3D coordinates of its q vertices, and denote a subset of $\mathbf{S}$ with columns corresponding to l annotated landmarks ($l = 68$ in our implementation) as $\mathbf{S}_L$. The projection of $\mathbf{S}_L$ on 2D planes are represented by $\mathbf{U}_i \in \mathbb{R}^{2 \times l}$. The relationship between 2D facial landmarks $\mathbf{U}_i$ and its corresponding 3D landmarks $\mathbf{S}_L$

---

[8]http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm

Table I
A COMPARISON OF DIFFERENT SYSTEMS IN TERMS OF 3D-RMSE.

Method	HQ	LQ	Full
MTCNN-CNN6-eos	2.70±0.88	2.78±0.95	2.75±0.93
MTCNN-CNN6-3DDFA	**2.04±0.67**	**2.19±0.70**	**2.14±0.69**
SCU-BRL	2.65±0.67	2.87±0.81	2.81±0.80

can be described as:

$$\mathbf{U}_i \approx f_i P_i R_i(\mathbf{S}_L + t_i), \qquad (1)$$

where f_i is the scale factor, P_i is the orthographic projection matrix, R_i is the 3×3 rotation matrix and t_i is the translation vector. Here, we employ weak perspective projection M_i to approximate the 3D-to-2D mapping. To fully utilize the correlation between the landmarks on all the images, we concatenate them to form a unified 2D facial landmark vector $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_p, \mathbf{U}_{p+1}, \cdots, \mathbf{U}_N)$, where $\mathbf{U}_i$ are zero vectors for $(p+1) \leq i \leq N$.

We reconstruct $\mathbf{S}$ from the given 'ground truth' landmarks $\mathbf{U}^*$ (either manually marked or automatically detected by a standalone method) for the unconstrained image set $\{I_i\}_{i=1}^p$. Let $\mathbf{S}^{k-1}$ be the currently reconstructed 3D shape after $k-1$ iterations. The corresponding landmarks $\mathbf{U}^{k-1}$ can be obtained by projecting $\mathbf{S}^{k-1}$ onto the image according to Eqn. (1). Then the updated 3D shape $\mathbf{S}^k$ can be computed by

$$\mathbf{S}^k = \mathbf{S}^{k-1} + \mathbf{W}^k(\mathbf{U}^* - \mathbf{U}^{k-1}), \qquad (2)$$

where $\mathbf{W}^k$ is the regressor in k^{th} iteration.

The K regressors $\{\mathbf{W}^k\}_1^K$ involved in the reconstruction process can be learned via optimising the following objective function over the m training samples (each sample contains up to N annotated 2D images and one ground truth 3D face shape):

$$\arg\min_{\mathbf{W}^k} \sum_{j=1}^m \| (\mathbf{S}_j^* - \mathbf{S}_j^{k-1}) - \mathbf{W}^k(\mathbf{U}_j^* - \mathbf{U}_j^{k-1}) \|_2^2, \quad (3)$$

where $\{\mathbf{S}_j^*, \mathbf{U}_j^*\}$ is one training sample consisting of ground truth landmarks $\mathbf{U}_j^*$ on the images of a subject and the subject's ground truth frontal and neutral 3D face shape $\mathbf{S}_j^*$.

A comprehensive description of the SCU-BRL system can be found in the paper Tian et al. [17].

V. EVALUATION RESULTS

We first compare the average RMSE of different systems on the benchmark dataset. The results are reported in Table I.

It is clear that the low quality subset is more challenging than the high quality subset. All the three methods have higher reconstruction errors on the low quality subset. However, this difference is minor for all the three systems. The reason is two-fold. *First*, the high quality subset contains many 2D face images with extreme pose variations, up to
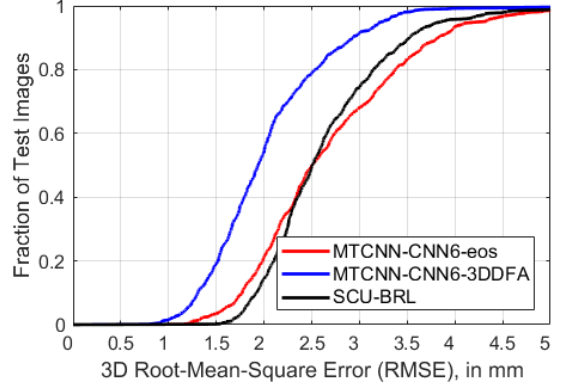
(a) High quality set
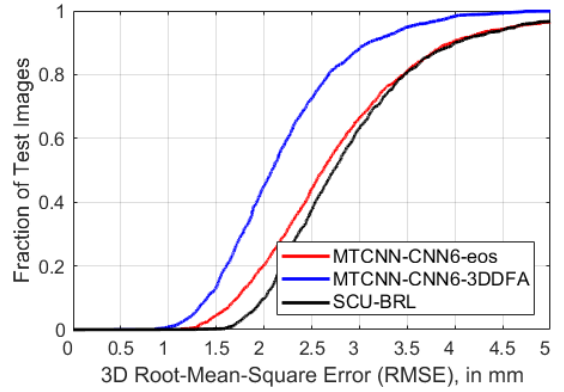
(b) Low quality set
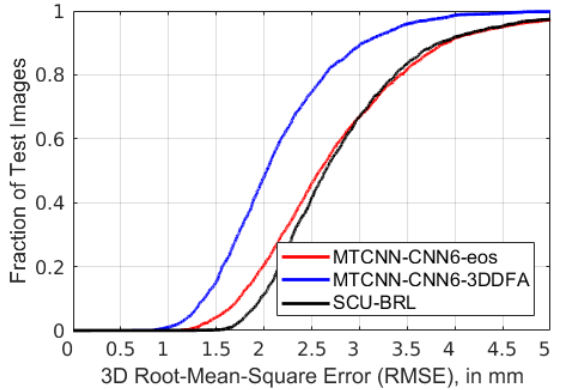
(c) Full set

Figure 4. A comparison of different systems in their CED curves. The results were evaluated on the (a) high quality subset, (b) low quality subset, and (c) full set of the test set selected from the Stirling ESRC dataset. The evaluation metric is RMSE in millimetre.

$\pm 90°$, as well as strong lighting changes thus makes the task more challenging. *Second*, both MTCNN+CNN6+eos and SCU-BRL systems are landmark-only 3D face reconstruction methods and their performance only relies on the accuracy of the detected 2D facial landmarks. The area of 2D facial landmark detection has already been very well developed for unconstrained facial images in the presence

of a variety of image degradation types. Thus the results of these two landmark-only fitting algorithms measured on the high quality and low quality subsets do not have very big difference. For the MTCNN+CNN6+3DDFA system, it uses the same state-of-the-art CNN-based face detector and landmark detector, but as initialisation of the 3D face fitting stage of 3DDFA, which is also CNN-based and trained on a large number of unconstrained faces. In addition, 3DDFA has multiple iterations cascaded for 3D face reconstruction using textural information. This is also why it performs significantly better than the other two landmark-only fitting systems. In this scenario, the 3DDFA algorithm benefits immensely from the detected 2D facial landmarks by the state-of-the-art CNN6 model.

To better understand the performance of different systems, we also plot the Cumulative Error Distribution (CED) curves of the three systems in Fig. 4. The MTCNN+CNN6+3DDFA outperforms the other two systems that only fit a 3D shape model to 2D facial landmarks. This is an interesting result for the community. It means that the textural information plays a very important role for high-performance 3D face reconstruction. However, it should be noted that the fitting of 3DDFA involves multiple iterations with cascaded CNN networks hence the speed of such a system cannot satisfy the requirement of a real-time application. The speed of the 3DDFA implementation provided by its authors is around 1.5 fps tested on a Intel Core i7 6700HQ CPU @ 3.2GHz. In contrast, the speed of eos is more than 100 fps, which is orders of magnitude faster than 3DDFA.

It should be noted that the SCU-BRL group also conducted multi-image fitting that used all the input 2D image of a subject for 3D face reconstruction. On average, their system reduces the 3D-RMSE from 2.81±0.80 to 2.26±0.72 on the benchmark dataset by fitting all the input 2D images together for a subject. This result shows that the SCU-BRL system can effectively utilise the complementary information in multiple images for 3D face reconstruction. For more details of their multi-image fitting method, please refer to [17].

VI. CONCLUSION

A new benchmark dataset was presented in this paper, used for the evaluation of 3D face reconstruction from single 2D face images in the wild. To this end, a subset of the Stirling ESRC 3D face dataset has been used to create the test set. The competition was conducted on the real 2D face images of 135 subjects and the evaluation was performed based on their real 3D face ground truth scans. To facilitate the competition, an evaluation protocol as well as a Python script were provided.

We have compared three state-of-the-art 3D face reconstruction systems on the proposed benchmark dataset, including a system submitted by the Sichuan University and two baseline approaches implemented by the organisers

from the University of Surrey. From the performance difference between purely landmark based and texture based reconstruction methods one main conclusion is that texture bolsters a significant amount of extra information about 3D shape. The exploitation of this, however, comes at a price of increased computational time.

The presented benchmark with evaluation data and protocol, together with a comprehensive analysis of different competing algorithms, support future evaluation in the community.

REFERENCES

[1] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," in *the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, W. N. Waggenspack, Ed., 1999, pp. 187–194.

[2] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.

[3] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3d morphable face models and their applications," in *9th International Conference on Articulated Motion and Deformable Objects (AMDO)*, vol. 9756, 2016, pp. 185–206.

[4] D. Zeng, Q. Zhao, S. Long, and J. Li, "Examplar coherent 3d face reconstruction from forensic mugshot database," *Image and Vision Computing*, vol. 58, pp. 193–203, 2017.

[5] F. Liu, J. Hu, J. Sun, Y. Wang, and Q. Zhao, "Multi-dim: A multi-dimensional face database towards the application of 3d technology in real-world scenarios," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 342–351.

[6] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin, "Gaussian mixture 3d morphable face model," *Pattern Recognition*, vol. 74, pp. 617–628, 2018.

[7] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face Alignment Across Large Poses: A 3D Solution," in *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2016, pp. 146–155.

[8] A. Bulat and G. Tzimiropoulos, "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230, 000 3D Facial Landmarks)," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[9] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," *arXiv preprint arXiv:1709.01442*, 2017.

[10] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The Florence 2D/3D Hybrid Face Dataset," in *the Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, p. 7980.

[11] M. Hernandez, T. Hassner, J. Choi, and G. G. Medioni, "Accurate 3D face reconstruction via prior constrained structure from motion," *Computers & Graphics*, vol. 66, pp. 14–22, 2017.

[12] A. T. Tran, T. Hassner, I. Masi, and G. G. Medioni, "Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1493–1502.

[13] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression," *arXiv*, 2017.

[14] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn, "The First 3D Face Alignment in the Wild (3DFAW) Challenge," in *Europearn Conference on Computer Vision Workshops - (ECCVW)*, G. Hua and H. Jégou, Eds., vol. 9914, 2016, pp. 511–520.

[15] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3D Face Morphable Models "In-the-Wild"," *arXiv*, 2017.

[16] P. Huber, G. Hu, J. R. Tena, P. Mortazavian, W. P. Koppen, W. J. Christmas, M. Rätsch, and J. Kittler, "A Multiresolution 3D Morphable Face Model and Fitting Framework," in *the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2016, pp. 79–86.

[17] W. Tian, F. Liu, and Q. Zhao, "Landmark-based 3D Face Reconstruction from an Arbitrary Number of Unconstrained Images," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2018.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[19] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Face Detection, Bounding Box Aggregation and Pose Estimation for Robust Facial Landmark Localisation in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 160–169.

[20] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2481–2490.

[21] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 679–692.

[23] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013, pp. 397–403.

[25] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2879–2886.

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[27] Z.-H. Feng, J. Kittler, W. Christmas, and X.-J. Wu, "A unified tensor-based active appearance face model," *arXiv*, 2016.

[28] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *IEEE International Conference on Advanced Video and Signal Based Surveillance AVSS*, 2009, pp. 296–301.