

The First 3D Face Alignment in the Wild (3DFAW) Challenge

László A. Jeni¹(✉), Sergey Tulyakov², Lijun Yin³, Nicu Sebe²,
and Jeffrey F. Cohn^{1,4}

¹ Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
laszlojeni@cmu.edu

² DISI, University of Trento, Trento, Italy

³ Department of Computer Science, State University of New York at Binghamton,
Binghamton, USA

⁴ Department of Psychology, The University of Pittsburgh, Pittsburgh, PA, USA

Abstract. 2D alignment of face images works well provided images are frontal or nearly so and pitch and yaw remain modest. In spontaneous facial behavior, these constraints often are violated by moderate to large head rotation. 3D alignment from 2D video has been proposed as a solution. A number of approaches have been explored, but comparisons among them have been hampered by the lack of common test data. To enable comparisons among alternative methods, The 3D Face Alignment in the Wild (3DFAW) Challenge, presented for the first time, created an annotated corpus of over 23,000 multi-view images from four sources together with 3D annotation, made training and validation sets available to investigators, and invited them to test their algorithms on an independent test-set. Eight teams accepted the challenge and submitted test results. We report results for four that provided necessary technical descriptions of their methods. The leading approach achieved prediction consistency error of 3.48%. Corresponding result for the lowest ranked approach was 5.9%. The results suggest that 3D alignment from 2D video is feasible on a wide range of face orientations. Differences among methods are considered and suggest directions for further research.

Keywords: 3D alignment from 2D video · Head rotation · Prediction consistency error · Faces in-the-wild

1 Introduction

Face alignment – the problem of automatically locating detailed facial landmarks across different subjects, illuminations, and viewpoints – is critical to face analysis applications, such as identification, facial expression analysis, robot-human interaction, affective computing, and multimedia.

Previous methods can be divided into two broad categories: 2D approaches and 3D approaches. 2D approaches treat the face as a 2D object. This assumption holds as long as the face is frontal and planar. As face orientation varies

from frontal, 2D annotated points lose correspondence. Pose variation results in self-occlusion that confounds landmark annotation. 2D approaches include Active Appearance Models [5, 16], Constrained Local Models [6, 21] and shape-regression-based methods [4, 8, 18, 24]). These approaches train a set of 2D models, each of which is intended to cope with shape or appearance variation within a small range of viewpoints.

3D approaches have strong advantages over 2D with respect to representational power and robustness to illumination and pose. 3D approaches [2, 7, 12, 27] accommodate a wide range of views. Depending on the 3D model, they easily can accommodate a full range of head rotation. Disadvantages are the need for 3D images and controlled illumination, as well as the need for special sensors or synchronized cameras in data acquisition.

Because these requirements are often difficult to meet, 3D alignment from 2D video or images has been proposed as a potential solution. A number of research groups have made advances in 3D alignment from 2D video [15, 17, 19, 20, 22]. How these various methods compare is relatively unknown. No commonly accepted evaluation protocol exists with which to compare them.

To enable comparisons among alternative methods of 3D alignment from 2D video, we created an annotated corpus of multi-view face images, partitioned training and hold-out test sets, and invited investigators to enter competition. The corpus includes images obtained under a range of conditions from highly controlled to in-the-wild. The resulting challenge provides a benchmark with which to evaluate 3D face alignment methods and enable researchers to identify new goals, challenges, and targets. This paper describes the 3D Face Alignment in the Wild Challenge and presents an overview of the results. The Challenge was held in conjunction with the 14th European Conference on Computer Vision.

2 Dataset

Four databases were used for the Challenge. They were the BU-4DFE [25], BP4D-Spontaneous [26], MultiPIE [11], and time-sliced videos from the internet. All four databases were annotated in a consistent way using a model-based structure-from-motion technique [14]. To increase variability in head rotation, we synthesized images across a range of pitch and yaw orientations as explained below.

2.1 BU-4DFE and BP-4D Spontaneous

BU-4DFE consists of approximately 60,600 3D frame models from 101 subjects (56% female, 44% male). Subjects ranged in age from 18 to 70 years and were ethnically and racially diverse (European-American, African-American, East-Asian, Middle-Eastern, Asian, Indian, and Hispanic Latino). Subjects were imaged individually using a Di3D (Dimensional Imaging¹) dynamic face capturing system while posing six prototypic emotion expressions (anger, disgust,

¹ <http://www.di3d.com>.

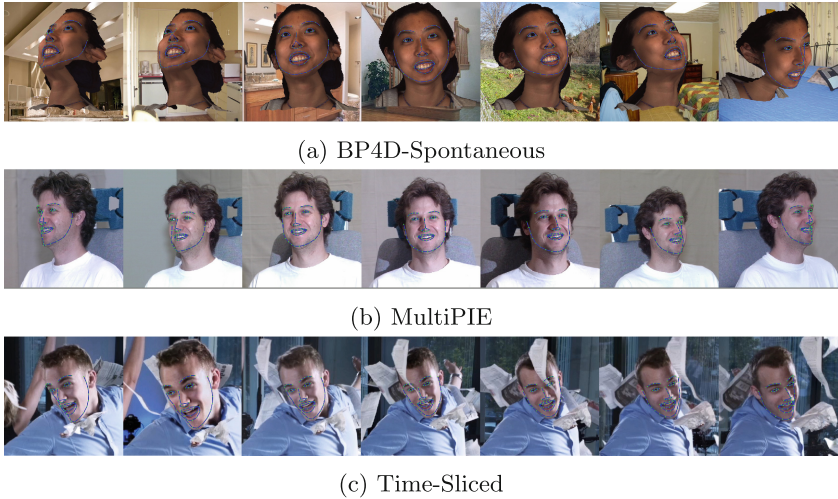


Fig. 1. Selected examples from the benchmark datasets. Selected views from the BP4D-Spontaneous (a), MultiPIE (b), and Time-Sliced (c) dataset. The contours of key facial parts are highlighted in blue for display purpose. (Color figure online)

happiness, fear, sadness, and surprise). The Di3D system consisted of two stereo cameras and a texture video camera arranged vertically. Both 3D model and 2D texture videos were obtained for each prototypic expression and subject. Given the arrangement of the stereo cameras, frontal looking faces have the most complete 3D information and smallest amount of texture distortion.

BP-4D-Spontaneous dataset [26] consists of over 300,000 frame models from 41 subjects (56 % female, 48.7 % European-American, average age 20.2 years) of similarly diverse backgrounds to BU-4DFE. Subjects were imaged using the same Di3D system while responding to a varied series of 8 emotion inductions; these were intended to elicit spontaneous facial expressions of amusement, surprise, fear, anxiety, embarrassment, pain, anger, and disgust. The 3D models range in resolution between 30,000 and 50,000 vertices. For each sequence, manual FACS coding [9] by highly experienced and reliable certified coders was obtained.

In BP-4DFE, 1365 uniformly distributed frames were sampled. In BP4D-Spontaneous, 930 frames were sampled based on FACS (Facial Action Coding System [9]) annotation to include a wide range of expressions.

The selected 3D meshes were manually annotated with 66 landmarks, referred to as facial fiducial points. The annotations were independently cross-checked by another annotator. Since the annotation was 3D, we can identify the self-occluded landmarks from every pose.

For each of the final 2295 annotated meshes, we synthesized 7 different views using a weak perspective camera model. These views span the range of $[-45, 45]$ degrees of yaw rotations in 15 degrees increments. The pitch rotation was randomly selected for each view from the range of $[-30, 30]$ degrees. Figure 1 shows

selected examples. In total 16,065 frames were synthesized. For each view we calculated the corresponding rotated 3D landmarks and their 2D projections with self-occlusion information. Since the 3D meshes lacked backgrounds, we added randomly selected non-face backgrounds from the SUN2012 dataset [23] in the final 2D images.

2.2 MultiPIE

Multi-PIE face database [11] contains images from 337 subjects acquired in a wide range of pose, illumination, and expression conditions. Images were captured in rapid order in a multi-camera, multi-flash recording. For the current database, we sampled 7000 frames from 336 subjects. For each frame, the visible portion of the face was annotated with 66 2D landmarks. Self-occluded landmarks were marked and excluded from the annotation.

2.3 Time-Sliced Videos

The above datasets were recorded in a laboratory under controlled conditions. To include uncontrolled (in-the-wild) images in the challenge, we collected time-sliced videos from the internet. In these videos subjects were surrounded by an array of still cameras. During the recording, the subjects displayed various expressions while the cameras fired simultaneously. Single frames from each camera were arranged consecutively to produce an orbiting viewpoint of the subject frozen in time.

We sampled 541 frames that correspond to several viewpoints from different subjects and expressions. Due to the unconstrained setting, the number of viewpoints per subjects varied between 3 and 7 views. For each frame, the visible portion of the face was annotated with 66 2D landmarks. Self-occluded landmarks were marked and excluded from the annotation.

2.4 Consistent 3D Landmark Annotation

Providing consistent 3D landmark annotation across viewpoints and across datasets was paramount for the challenge. In the case of BU4D and BP4D-Spontaneous data, we had 3D landmark annotation that is consistent across synthesized views of the same face. To provide the same consistency for the other two datasets, we employed a two-step procedure. First we built a deformable 3D face model from the annotated 3D meshes of BU4D and BP4D-Spontaneous. Second, we used a model-based structure-from-motion technique on the multi-view images [14].

Linear Face Models. A shape model is defined by a 3D mesh and, in particular, by the 3D vertex locations of the mesh, called landmark points. Consider the 3D shape as the coordinates of 3D vertices that make up the mesh:

$$\mathbf{x} = [x_1; y_1; z_1; \dots; x_M; y_M; z_M], \quad (1)$$

or, $\mathbf{x} = [\mathbf{x}_1; \dots; \mathbf{x}_M]$, where $\mathbf{x}_i = [x_i; y_i; z_i]$.

The 3D point distribution model (PDM) describes non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

$$\mathbf{x}_i = \mathbf{x}_i(\mathbf{p}, \mathbf{q}) = s\mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i\mathbf{q}) + \mathbf{t} \quad (i = 1, \dots, M), \quad (2)$$

where $\mathbf{x}_i(\mathbf{p}, \mathbf{q})$ denotes the 3D location of the i^{th} landmark and $\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{t}\}$ denotes the rigid parameters of the model, which consist of a global scaling s , angles of rotation in three dimensions ($\mathbf{R} = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma)$), a translation \mathbf{t} . The non-rigid transformation is denoted with \mathbf{q} . Here $\bar{\mathbf{x}}_i$ denotes the mean location of the i^{th} landmark (i.e. $\bar{\mathbf{x}}_i = [\bar{x}_i; \bar{y}_i; \bar{z}_i]$ and $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1; \dots; \bar{\mathbf{x}}_M]$). The d pieces of $3M$ dimensional basis vectors are denoted with $\Phi = [\Phi_1; \dots; \Phi_M] \in \mathbb{R}^{3M \times d}$. Vector \mathbf{q} represents the 3D distortion of the face in the $3M \times d$ dimensional linear subspace.

To build this model we used the 3D annotation from the selected BU-4DFE [25] and BP4D-Spontaneous [26] frames.

3D Model Fitting. To reconstruct the 3D shape from the annotated 2D shapes (\mathbf{z}) we need to minimize the reconstruction error using Eq. (2):

$$\arg \min_{\mathbf{p}, \mathbf{q}} \sum_{i=1}^M \|\mathbf{P}\mathbf{x}_i(\mathbf{p}, \mathbf{q}) - \mathbf{z}_i\|_2^2 \quad (3)$$

Here \mathbf{P} denotes the projection matrix to 2D, and \mathbf{z} is the target 2D shape. An iterative method can be used to register 3D model on the 2D landmarks [12]. The algorithm iteratively refines the 3D shape and 3D pose until convergence, and estimates the rigid ($\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{t}\}$) and non-rigid transformations (\mathbf{q}).

Applying Eq. (3) on a single image frame from a monocular camera has a drawback of simply “hallucinating” a 3D representation from 2D. From a single viewpoint there are multiple solutions that satisfy Eq. (3). To avoid the problem of single frame 2D-3D hallucination we apply the method simultaneously across multiple image-frames of the same subject. Furthermore, we have partial landmark annotation in the MultiPIE and TimeSliced data due to self-occlusion. We can incorporate the visibility information of the landmarks in Eq. (3), by constraining the process to the visible landmarks.

Let $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(C)}$ denote the C number of 2D measurements from the different viewpoints of the same subject. The exact camera locations and camera calibration matrices are unknown. In this case all C measurements represent the same 3D face, but from a different point of view. We can extend Eq. (3) to this scenario by constraining the reconstruction to all the measurements:

$$\arg \min_{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(C)}, \mathbf{q}} \sum_{k=1}^C \sum_{i \in \xi^{(k)}} \left\| \mathbf{P}\mathbf{x}_i(\mathbf{p}^{(k)}, \mathbf{q}) - \mathbf{z}_i^{(k)} \right\|_2^2 \quad (4)$$

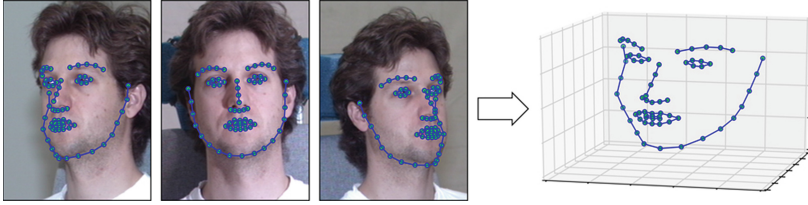


Fig. 2. The 3D shapes from the different views from the same subject and expression are consistent, they can be superimposed on each other in a canonical space.

where superscripts (k) denote the k^{th} measurement, with a visibility set of $\xi^{(k)}$. Minimizing Eq. (4) can be done by iteratively refining the 3D shape and 3D pose until convergence. For more details see [13, 14] (Fig. 2).

3 Evaluation Results

3.1 Data Distribution

Data were sorted into three subsets (training, validation, and test sets) and distributed in two phases using the CodaLab platform². In Phase-I, participants were granted access to the complete training set of images, ground truth 3D landmarks, and face bounding boxes and the validation set images and their bounding boxes. Participants became acquainted with the data and could train and perform initial evaluations of their algorithms. In Phase-II, they were granted access to the ground truth landmarks of the validation set and images and bounding boxes from the final test set. See Table 1 for more details.

3.2 Performance Measures

For comparative evaluation in the Challenge, we used the widely accepted evaluation matrices Ground Truth Error (GTE) and Cross View Ground Truth Consistency Error (CVGTCE). GTE is the average point-to-point Euclidean error

Table 1. Distribution of the different sets.

	Training	Validation	Test	Total
BP-4DFE	5677	1960	1918	9555
BP-4D-Spontaneous	3794	1365	1351	6510
MultiPIE	4200	1400	1400	7000
TimeSliced	298		243	541

² <https://competitions.codalab.org/>.

normalized by the outer corners of the eyes (inter-ocular). It is computed as:

$$GTE(\mathbf{x}^{pre}, \mathbf{x}^{gt}) = \frac{1}{M} \sum_{k=1}^M \frac{\|\mathbf{x}_k - \mathbf{y}_k\|_2}{d_i} \quad (5)$$

where M is the number of points, \mathbf{x}^{gt} is the ground truth 3D shape, \mathbf{x}^{pre} is the predicted shape and d_i is the inter-ocular distance for the i -th image.

CVGTCE evaluates cross-view consistency of the predicted landmarks from the 3D model. It is computed as:

$$CVGTCE(\mathbf{x}^{pre}, \mathbf{x}^{gt}, \mathbf{p}) = \frac{1}{M} \sum_{k=1}^M \frac{\|(s\mathbf{R}\mathbf{x}_k^{pre} + \mathbf{t}) - \mathbf{x}_k^{gt}\|_2}{d_i} \quad (6)$$

where the rigid transformation parameters $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}\}$ can be obtained in a similar fashion as in Eq. (3).

3.3 Participation

Eight teams submitted results. Of these, four completed the challenge by submitting a technical description of their methods. In the following we briefly describe their methods. More detail is provided in the respective papers. The final scores for all methods are available on the competition website³.

Zavan et al. [1] proposed a method that requires only the nose region for assessing the orientation of the face and the position of the landmarks. First, a Faster R-CNN was trained on the images to detect the nose. Second, a CNN variant was trained to categorize the face into several discretized head-pose categories. In the final step, the system imposes the average face landmarks onto the image using the previously estimated transformation parameters.

Zhao et al. [28] used a deep convolutional network based solution that maps the 2D image of a face to its 3D shape. They defined two criteria for the optimization: (i) learn facial landmark locations in 2D (ii) and then estimate the depth of the landmarks. Furthermore, a data augmentation approach was used to aid the learning. The latter involved applying 2D affine transformations to the training set and generating random occluding boxes to improve robustness to partial occlusion.

Gou et al. [10] utilized a regression-based 3D face alignment method that first estimates the location of a set of landmarks and then recovers 3D face shape by fitting a 3D morphable model. An alternative optimization method was employed for the 3D morphable model fitting to recover the depth information. The method incorporates shape and local appearance information in a cascade regression framework to capture the correspondence between pairs of points for 3D face alignment.

Bulat and Tzimiropoulos [3] proposed a two-stage alignment method. At the first stage, the method calculates heat-maps of 2D landmarks using convolutional

³ <https://competitions.codalab.org/competitions/10261>.

part heat-map regression. In the second stage, these heat-maps along with the original RGB image were used as an input to a very deep residual network to regress the depth information.

3.4 Results

Table 2 shows the Prediction Consistency Errors (CVGTCE) and Standard Errors (GTE) of the different methods on the final test set. Figure 3 shows the cumulative error distribution curves (CED) of the different methods.

Table 2. Prediction Consistency Error (CVGTCE) and Standard Error (GTE) of the different methods on the Test set.

Rank	Team	CVGTCE %	GTE %
1	Bulat and Tzimiropoulos [3]	3.4767	4.5623
2	Zhao et al. [28]	3.9700	5.8835
3	Gou et al. [10]	4.9488	6.2071
4	Zavan et al. [1]	5.9093	10.8001

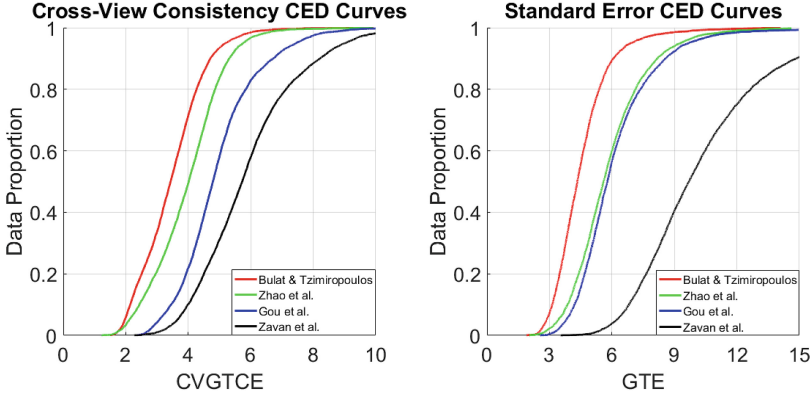


Fig. 3. Cumulative distribution curves (CED) of the different methods for Cross-View Consistency (left) and Standard Error (right).

4 Conclusion

This paper describes the First 3D Face Alignment in the Wild (3DFAW) Challenge held in conjunction with the 14th European Conference on Computer

Vision 2016, Amsterdam. The main challenge of the competition was to estimate a set of 3D facial landmarks from still images. The corpus includes images obtained under a range of conditions from highly controlled to in-the-wild. All image sources have been annotated in a consistent way, the depth information has been recovered using a model-based Structure from Motion technique. The resulting challenge provides a benchmark with which to evaluate 3D face alignment methods and enable researchers to identify new goals, challenges, and targets.

Acknowledgements. This work was supported in part by US National Institutes of Health grant MH096951 to the University of Pittsburgh and by US National Science Foundation grants CNS-1205664 and CNS-1205195 to the University of Pittsburgh and the University of Binghamton. Neither agency was involved in the planning or writing of the work.

References

1. de B. Zavan, F.H., Nascimento, A.C.P., e Silva, L.P., Bellon, O.R.P., Silva, L.: 3d face alignment in the wild: A landmark-free, nose-based approach. In: 2016 European Conference on Computer Vision Workshops (ECCVW) (2016)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194. SIGGRAPH (1999). <http://dx.doi.org/10.1145/311535.311556>
3. Bulat, A., Tzimiropoulos, G.: Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3DFAW) challenge. In: European Conference on Computer Vision Workshops (ECCVW) (2016)
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2887–2894, June 2012
5. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(6), 681–685 (2001)
6. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. *Pattern Recogn.* **41**(10), 3054–3067 (2008). <http://dx.doi.org/10.1016/j.patcog.2008.01.024>
7. Dimitrijevic, M., Ilic, S., Fua, P.: Accurate face models from uncalibrated and ill-lit video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. II-1034-II-1041, June 2004
8. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085, June 2010
9. Ekman, P., Friesen, W., Hager, J.: Facial Action Coding System (FACS): Manual. *A Human Face*, Salt Lake City (2002)
10. Gou, C., Wu, Y., Wang, F.Y., Ji, Q.: Shape augmented regression for 3d face alignment. In: 2016 European Conference on Computer Vision Workshops (ECCVW) (2016)
11. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. *Image Vis. Comput.* **28**(5), 807–813 (2010)
12. Gu, L., Kanade, T.: 3d alignment of face in a single image. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (CVPR)* **1**, 1305–1312 (2006)

13. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d videos in real-time. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2015). <http://zface.org>
14. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d video for real-time use. *Image Vis. Comput.* **28**(5), 807–813 (2016)
15. Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: CVPR (2016)
16. Matthews, I., Baker, S.: Active appearance models revisited. *Int. J. Comput. Vis.* **60**(2), 135–164 (2004)
17. Piotraschke, M., Blanz, V.: Automated 3d face reconstruction from multiple images using quality measures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3418–3427 (2016)
18. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1685–1692, June (2014)
19. Roth, J., Tong, Y., Liu, X.: Adaptive 3d face reconstruction from unconstrained photo collections. In: CVPR (2016)
20. Sánchez-Escobedo, D., Castelán, M., Smith, W.A.: Statistical 3d face shape estimation from occluding contours. *Comput. Vis. Image Underst.* **142**, 111–124 (2016)
21. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.* **91**(2), 200–215 (2011). <http://dx.doi.org/10.1007/s11263-010-0380-4>
22. Tulyakov, S., Sebe, N.: Regressing a 3d face shape from a single image. In: IEEE International Conference on Computer Vision (ICCV), pp. 3748–3755. IEEE (2015)
23. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
24. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539, June (2013)
25. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, 2008. FG 2008, pp. 1–6, Sept 2008
26. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp. 4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* **32**(10), 692–706 (2014). best of Automatic Face and Gesture Recognition 2013
27. Zhang, Z., Liu, Z., Adler, D., Cohen, M.F., Hanson, E., Shan, Y.: Robust and rapid generation of animated faces from video images: A model-based modeling approach. *Int. J. Comput. Vis.* **58**(2), 93–119 (2004). <http://dx.doi.org/10.1023/B:VISI.0000015915.50080.85>
28. Zhao, R., Wang, Y., Benitez-Quiroz, C.F., Liu, Y., Martinez, A.M.: Fast & precise face alignment and 3d shape recovery from a single image. In: 2016 European Conference on Computer Vision Workshops (ECCVW) (2016)