



300 Faces In-The-Wild Challenge: database and results☆☆☆



Christos Sagonas^{a,*}, Epameinondas Antonakos^a, Georgios Tzimiropoulos^b,
Stefanos Zafeiriou^a, Maja Pantic^{a,c}

^aImperial College London, Department of Computing, London, UK

^bUniversity of Nottingham, School of Computer Science, Nottingham, UK

^cFaculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands

ARTICLE INFO

Article history:

Received 19 March 2015

Received in revised form 2 October 2015

Accepted 4 January 2016

Available online 25 January 2016

Keywords:

Facial landmark localization

Challenge

Semi-automatic annotation tool

Facial database

ABSTRACT

Computer Vision has recently witnessed great research advance towards automatic facial points detection. Numerous methodologies have been proposed during the last few years that achieve accurate and efficient performance. However, fair comparison between these methodologies is infeasible mainly due to two issues. (a) Most existing databases, captured under both constrained and unconstrained (in-the-wild) conditions have been annotated using different mark-ups and, in most cases, the accuracy of the annotations is low. (b) Most published works report experimental results using different training/testing sets, different error metrics and, of course, landmark points with semantically different locations. In this paper, we aim to overcome the aforementioned problems by (a) proposing a semi-automatic annotation technique that was employed to re-annotate most existing facial databases under a unified protocol, and (b) presenting the 300 Faces In-The-Wild Challenge (300-W), the first facial landmark localization challenge that was organized twice, in 2013 and 2015. To the best of our knowledge, this is the first effort towards a unified annotation scheme of massive databases and a fair experimental comparison of existing facial landmark localization systems. The images and annotations of the new testing database that was used in the 300-W challenge are available from http://ibug.doc.ic.ac.uk/resources/300-W_IMAVIS/.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

During the last decades we notice a wealth of scientific research in computer vision for the problem of facial landmark points localization using visual deformable models. The main reason behind this are the countless applications that the problem has in human-computer interaction and facial expression recognition. Numerous methodologies have been proposed that are shown to achieve great accuracy and efficiency. They can be roughly divided into two categories: *generative* and *discriminative*. The generative techniques, which aim to find the parameters that maximize the probability of the test image being generated by the model, include Active Appearance Models (AAMs) [1,2], their improved extensions [3–10] and Pictorial

Structures [11–13]. The discriminative techniques can be further divided to those that use discriminative response map functions, such as Active Shape Models (ASMs) [14], Constrained Local Models (CLMs) [15–17] and Deformable Part Models (DPMs) [18], those that learn a cascade of regression functions, such as Supervised Descent Method (SDM) [19] and others [20–22], and, finally, those that employ random forests [23,24].

Arguably, the main reason why many researchers of the field focus on the problem of face alignment is the plethora of publicly available annotated facial databases. These databases can be separated in two major categories: (a) those captured under *controlled conditions*, e.g. Multi-PIE [25], XM2VTS [26], FRGC-V2 [27], and AR [28], and (b) those captured under totally *unconstrained conditions* (in-the-wild), e.g. LFPW [29], HELEN [30], AFW [18], AFLW [31], and IBUG [32]. All of them cover large variations, including different subjects, poses, illumination conditions, expressions and occlusions. However, for most of them, the provided annotations appear to have several limitations. Specifically:

- The majority of them provide annotations for a relatively small subset of images.

☆ The contribution of the first two authors on writing this paper is equal, with Christos Sagonas being the main responsible for the implementation and execution of various steps needed to run 300-W successfully including data annotation, annotation tool development, and running the experiments.

☆☆ This paper has been recommended for acceptance by Richard Bowden, PhD.

* Corresponding author.

E-mail address: c.sagonas@imperial.ac.uk (C. Sagonas).

- The annotation mark-up of each database consists of different number of landmark points with semantically different locations.
- The accuracy of the provided annotations in some cases is limited.

The above issues are due to the fact that manual annotation of large databases is a highly time consuming procedure that requires enormous workload and a trained expert. Moreover, factors like fatigue and lack of concentration are among the reasons why, in some cases, annotations are inaccurate. This highlights the need of creating a (semi-) automatic annotation tool.

Furthermore, by going through the published works of the last years, one can easily notice that the setup of the experiments is not always correct. Researchers employ different databases, experimental protocols and performance metrics, which lead to unfair comparisons between existing methods. Some characteristic such examples are the following:

- Authors compare their techniques against other state-of-the-art, but they do so by using, in many cases, completely different databases for training compared to the ones that the other methods were originally trained on.
- Authors compare their techniques on specific databases by replicating the originally presented curves and not the experiment.
- In some cases, authors report results on databases from which only a part can be used by the community, as some of the training/testing images are no longer publicly available.

Evidence shows that there is a lack of access to properly evaluate existing methods. Even though there exist open-source implementations of various state-of-the-art techniques (the most characteristic example is Menpo [33]), researchers still do not employ a unified benchmark. Since we are unaware of the achieved performances, it is impossible to investigate how far we are from attaining satisfactory performance. Therefore, a new evaluation needs to be carried out, using a unified experimental protocol.

Various methods have been proposed in the literature for the task of landmark localization under semi-supervised or weakly-supervised settings [34–37]. However, there are two major limitations of these methods. Firstly, most existing methodologies require additional information regarding the input images. Specifically, Jia et al. [34] employ the corresponding facial mask for each of the training images. The purpose of these masks is to indicate which pixels belong to the facial area and the only way to produce them is by manually annotating each image. In [35], the training procedure requires as input the orientation of each face depicted in the training images. Secondly, and most importantly, existing methods, such as [36] and [37], have only been applied on images that are captured under controlled conditions. The aforementioned issues, make the existing methods incapable for the task of semi-automatic annotation of large databases with in-the-wild images (most of the images are downloaded from the web with simple search queries), which is a much more challenging task.

Semi-automatic annotation systems can greatly benefit from the employment of generative models. Let us assume that we have cohorts of both annotated and non-annotated images. By training a generative model, such as AAMs, using the annotated images, we get a parametric model that describes the facial shape and appearance. Most importantly, the model can naturally generate novel instances of human face, by combining the shape and appearance variance of the training annotated images. This could enable the generation of instances that resemble accurately with the shape and appearance of the subjects in the non-annotated images. For instance, by training a model using images from one view (e.g. pose 15°) with neutral

expression and images from another view (e.g. pose 0°) with a non-neutral expression, one can fit the model to an instance that has the non-neutral expression with pose of 15°. However, the fitting procedure of a generative deformable model is a very tedious task, mainly because many of the models that have been proposed till now do not generalize well to unseen images. One of the AAM variants that has satisfactory generalization properties is Active Orientation Models (AOMs) [3,5]. AOMs are shown to be robust in cases with large variations, such as occlusions and extreme illumination conditions, and outperform discriminative methodologies, such as CLMs [16], DPMs [18] and SDM [19].

Motivated by the success of AOMs in generic face alignment, we propose, in this paper, a semi-automatic technique for annotating in a time efficient manner massive facial databases. We employed the proposed tool to re-annotate all the widely used databases, i.e. Multi-PIE [25], XM2VTS [26], FRGC-V2 [27], AR [28], LFPW [29], HELEN [30] and AFW [18]. The resulting annotations¹ are, in many cases, more accurate than the original ones and employ a unified mark-up scheme, thus overcome the limitations explained above.

Furthermore, in order to offer to the research community the ability to carry out rational comparisons between existing and future proposed methods, we organized two versions of the 300 Faces In-The-Wild Challenge (300-W), the first automatic facial landmark detection in-the-wild challenge. The first challenge² was organized in 2013 in conjunction with the IEEE International Conference on Computer Vision (ICCV'13) [32]. The second conduct³ of the challenge was completed in the beginning of 2015. In both conducts, the training set consisted of the XM2VTS, FRGC-V2, LFPW, HELEN, AFW and IBUG databases that were annotated using the proposed semi-automatic procedure. Additionally, we collected and annotated a new challenging in-the-wild database that was used for testing.⁴ The 300-W database consists of 300 *Indoor* and 300 *Outdoor* images downloaded from the web, thus captured under totally unconstrained conditions. The performance of the submitted methods was evaluated using the same fitting accuracy metric. The major difference between the two conducts of the challenge is that in the first version we provided the bounding boxes of the testing images to be used as initializations, while in the second version the participants were required to submit systems that performed both face detection and alignment. Additionally, contrary to the first version, in the second one the submitted methods were also compared with respect to their computational costs.

The contribution of this paper can be summarized as follows:

1. We propose a semi-automatic methodology for facial landmark points annotation. The proposed tool was employed in order to re-annotate large facial databases and overcome the major issues of the original annotations.
2. We present and analyze the results of the 300 Faces In-The-Wild Challenge (300-W), the first facial landmark localization challenge, that was conducted twice, in 2013 and 2015. The challenge is the first attempt towards a fair comparison of existing methods using a unified experimental protocol.
3. We make the very challenging 300-W dataset publicly available to the research community. It was employed as testing set in both conducts of the 300-W competition.

¹ The annotations of XM2VTS, FRGC-V2, LFPW, HELEN, AFW and IBUG are publicly available from <http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/>.

² The first conduct of the 300-W challenge (2013) is available in <http://ibug.doc.ic.ac.uk/resources/300-W/>.

³ The second conduct of the 300-W challenge (2015) is available in http://ibug.doc.ic.ac.uk/resources/300-W_IMAVIS/.

⁴ The 300-W database is publicly available from http://ibug.doc.ic.ac.uk/resources/300-W_IMAVIS/. We provide the images along with the corresponding bounding boxes.

The rest of the paper is organized as follows: Section 2 gives an overview of the available facial databases. Section 3 presents the proposed semi-automatic methodology for facial landmark points annotations along with the re-annotated databases. The 300-W challenge and the results are described in details in Section 4. Finally, Section 5 summarizes the results of this work and draws conclusions.

2. Overview of existing facial databases

There exist numerous facial databases which partially justifies the research advances for the task of face alignment. These databases exhibit large variations in resolution, image quality, identity, head pose, facial expression, lighting conditions and partial occlusion. As mentioned before, the existing databases can be split in two major categories. The first category includes databases that are captured under *controlled conditions*, normally within special indoor laboratories/studios in which the camera position and the lighting source and intensity can be controlled. In most of these databases, each subject is asked to perform a posed facial expression, thus we find more than one images per subject. The most popular such databases are Multi-PIE [25] (used for face recognition, expression recognition, landmark points localization), FRGC-V2 [27] (used for face recognition), XM2VTS [26] and AR [28] (both used for face recognition and landmark points localization). The facial databases of the second major category consist of images that are captured under totally *unconstrained conditions (in-the-wild)*. In most cases, these images are downloaded from the web by making face-related queries to various search engines. The most notable databases of this category are LFPW [29], HELEN [30], AFW [18], AFLW [31] and IBUG [32] (all used for facial landmark points localization).

The majority of the aforementioned databases provide annotations for a relatively small subset of images. Moreover, as shown in Fig. 1, they all have different annotation schemes between them, leading in different numbers of points with semantically different locations. There are also cases in which the accuracy of the provided annotations is limited. Sections 2.1 and 2.2 and Table 1 provide an overview of the characteristics of all the commonly-used existing databases.

2.1. Facial databases under controlled conditions

2.1.1. Multi-PIE

The CMU Multi-Pose Illumination, and Expression (Multi-PIE) Database [25] contains around 750,000 images of 337 subjects captured under laboratory conditions in four different sessions. For each subject, there are available images for 15 different poses, 19 illumination conditions and 6 different expressions (neutral, scream, smile, squint, surprise, disgust). The accompanying facial landmark annotations consist of a set of 68 points (Fig. 1a) for images in the range $[-45^\circ, 45^\circ]$.

2.1.2. XM2VTS

The Extended Multi Modal Verification for Teleservices and Security applications (XM2VTS) [26] database contains 2360 frontal images of 295 different subjects. Each subject has two available images for each of the four different sessions. All subjects are captured under the same illumination conditions and in the majority of images the subject has neutral expression. Facial landmark annotations of the whole database are available, where 68 points are provided for each image (Fig. 1b). However, the accuracy of the annotations in some cases is limited and the locations of the provided points do not correspond to ones of Multi-PIE.

2.1.3. FRGC-V2

The Face Recognition Grand Challenge Version 2.0 (FRGC-V2) database [27] consists of 4950 facial images of 466 different subjects. Each subject session consists of images captured under well-controlled conditions (i.e., uniform illumination, high resolution) and images captured under fairly uncontrolled conditions such as non-uniform illumination and poor quality. The provided annotations consist of 5 landmark points (Fig. 1c) only.

2.1.4. AR

The AR Face Database [28] contains over 4000 images corresponding to 126 subjects (70 males, 56 females). The images were captured in two sessions per subject and have frontal pose with variations in facial expressions, illumination conditions and occlusions (sunglasses and scarf). The images are annotated using 22 landmark points (Fig. 1d).

2.2. Facial databases under in-the-wild conditions

2.2.1. LFPW

The Labeled Face Parts in the Wild (LFPW) database [29] contains 1287 images downloaded from the internet (i.e., google.com, flickr.com, and yahoo.com). This database provides only the web URLs and not the actual images. We were therefore able to download only a subset of 811 out of 1100 training images and 224 out of 300 test images, due to broken links. These images contain large variations in pose, expressions, illumination conditions and occlusions. The provided ground truth annotations consist of 35 landmark points (Fig. 1e) and low accuracy is observed in several cases.

2.2.2. HELEN

The HELEN [30] database consists of 2330 images downloaded from flickr.com web service, that contain a broad range of appearance variation, including pose, illumination conditions, expression, occlusion and identity. The approximate face size of each image is 500×500 pixels. The provided annotations are very detailed and contain 194 landmark points (Fig. 1f), but the accuracy is limited.

2.2.3. AFW

The Annotated Faces in-the-wild (AFW) [18] database consists of 250 images with 468 faces, that is, more than one faces are annotated in each image. The images exhibit similar variations with those in the aforementioned in-the-wild databases. Facial landmark annotations are available for the whole database, but the annotation mark-up consists of only 6 points (Fig. 1g).

2.2.4. AFLW

The Annotated Facial Landmarks in the Wild (AFLW) [31] database consists of 25,993 images gathered from Flickr, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. However, the employed annotation scheme only includes 21 landmark points (Fig. 1h).

2.2.5. IBUG

The IBUG database was released as part of the first version of the 300-W challenge [32]. It consists of 135 images downloaded from the web, with large variations in expression, illumination conditions, and pose. The provided facial landmark annotations are produced by employing the annotation scheme of Multi-PIE (Fig. 1a).

3. Semi-automatic annotation tool

In this section, we propose a technique for semi-automatic annotation of large databases, which takes advantage of the good generalization properties of AOMs [3,5].

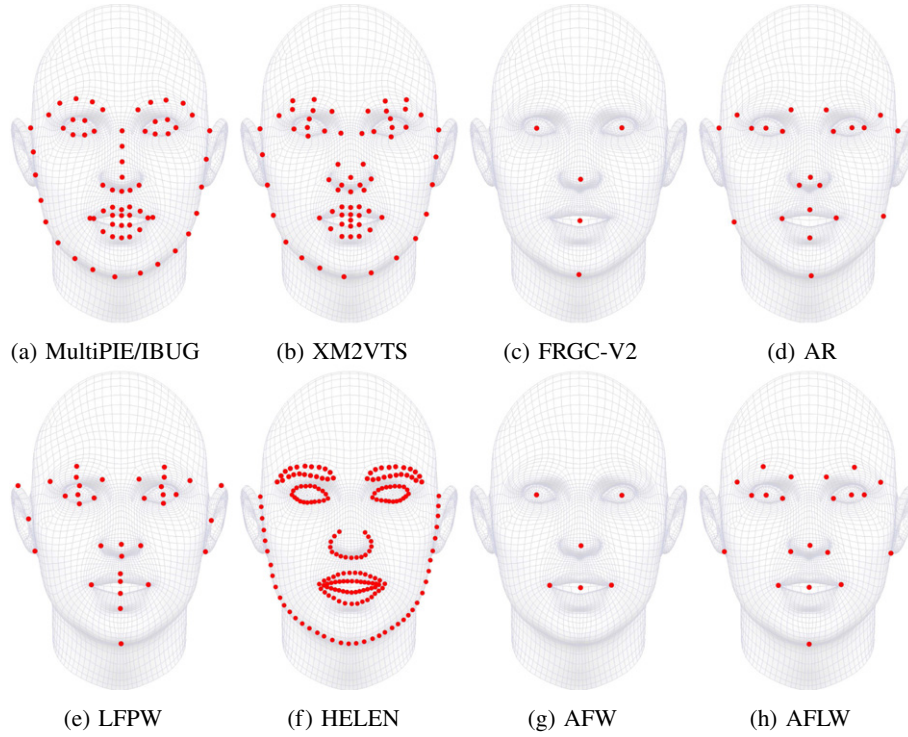


Fig. 1. Landmark configurations of existing databases. Note they all have different numbers of landmark points with semantically different locations.

3.1. Active orientation models

AOMs is a variant of AAMs [2]. Similar to AAMs, they consist of parametric statistical *shape* and *appearance models*, and a *deformation model*. However, the difference is that AOMs employ kernel PCA based on a similarity criterion that is robust to outliers. Specifically, the appearance model of AOMs consists of the principal components of image gradient orientations [38], which makes them generalize well to unseen face instances.

Let us assume that we have a set of D training images, $\{\mathbf{I}_1, \dots, \mathbf{I}_D\}$, annotated with N landmark points that represent the ground truth shape of each image. A shape instance is defined as the $2N \times 1$ vector $\mathbf{s} = [x_1, y_1, \dots, x_N, y_N]^T$, where (x_i, y_i) are the coordinates of the i -th fiducial point. The *shape model* is constructed by first aligning all training shapes using Generalized Procrustes Analysis in order to remove global similarity transformations and then applying Principal Component Analysis (PCA) on the aligned shapes to retrieve:

$$\{\bar{\mathbf{s}}, \mathbf{U}_S \in \mathbb{R}^{2N \times N_S}\}, \quad (1)$$

Table 1
Overview of the characteristics of existing facial databases.

Database	Conditions	# of faces	# of subjects	# of points	Pose
Multi-PIE	Controlled	~ 750,000	337	68	$[-45^\circ, 45^\circ]$
XM2VTS		2360	295	68	0°
FRGC-V2		4950	466	5	0°
AR		~ 4000	126	22	0°
LFPW	In-the-wild	1035		35	
HELEN		2330		194	
AFW		468	–	6	$[-45^\circ, 45^\circ]$
AFLW		25,993		21	
IBUG		135		68	

where $\bar{\mathbf{s}}$ is the mean shape and \mathbf{U}_S consists of the first N_S eigenvectors with the highest variance. A novel shape instance can be generated as:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{U}_S \mathbf{p}, \quad (2)$$

where $\mathbf{p} = [p_1, \dots, p_{N_S}]^T$ denotes the $N_S \times 1$ vector of shape parameters. The *deformation model* consists of a warp function, denoted as $\mathcal{W}(\mathbf{p})$, which maps all the pixels that belong into a shape instance generated from Eq. (2) with parameters \mathbf{p} to their corresponding locations in the mean shape $\bar{\mathbf{s}}$. We employ the Piecewise Affine Warp, which evaluates the mapping using the barycentric coordinates of the triangles extracted with Delaunay triangulation.

The *appearance model* of an AOM is based on normalized gradients [38]. Let us denote an image in vectorial form as \mathbf{i} with size $L \times 1$, thus L is the number of pixels. Moreover, we denote \mathbf{g}_x and \mathbf{g}_y to be the image gradients and $\phi = \arctan(\mathbf{g}_x/\mathbf{g}_y)$ the corresponding gradient orientation vector. The normalized gradient extraction function is defined as:

$$\mathcal{Z}(\mathbf{i}) = \frac{1}{\sqrt{L}} [\cos \phi^T, \sin \phi^T]^T, \quad (3)$$

where $\cos \phi = [\cos \phi(1), \dots, \cos \phi(L)]^T$ and $\sin \phi = [\sin \phi(1), \dots, \sin \phi(L)]^T$. By employing the deformation model, we can define the shape-free normalized gradients of an image \mathbf{i} as the $2L_Z \times 1$ vector:

$$\mathbf{z}(\mathbf{p}) \equiv \mathcal{Z}(\mathbf{i}(\mathcal{W}(\mathbf{p}))), \quad (4)$$

where L_Z is the number of pixels that belong to the mean shape $\bar{\mathbf{s}}$, which has the role of the reference shape. By applying PCA on the

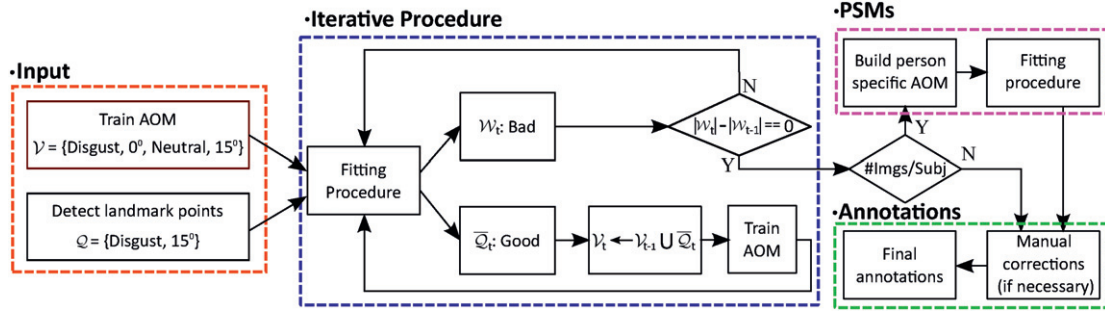


Fig. 2. Flowchart of the proposed tool. Given a set of landmarked images \mathcal{V} with various poses and expressions, we aim to annotate a set of non-annotated images \mathcal{Q} (1) with the same subjects and different poses and expressions, or (2) with different subjects but similar pose and expressions.

warped normalized gradients of the training images, i.e. $\{z_1, \dots, z_D\}$, we construct an *appearance model* of the form:

$$\mathbf{U}_Z \in \mathbb{R}^{2L_Z \times N_Z}, \quad (5)$$

where \mathbf{U}_Z stores the first N_Z eigenvectors with the highest variance. Note that in order to preserve the robust property of the normalized gradient kernel, we don't subtract the mean appearance vector from the training set, so it ends up as the first eigenvector. A novel appearance instance can be generated as:

$$\mathbf{z} = \mathbf{U}_Z \mathbf{c}, \quad (6)$$

where $\mathbf{c} = [c_1, \dots, c_{N_Z}]^T$ denotes the $N_Z \times 1$ vector of appearance parameters.

Given a testing image \mathbf{t} in vectorized form and the trained shape, appearance and deformation models, the fitting procedure aims to minimize:

$$\arg \min_{\mathbf{p}, \mathbf{c}} \|\mathbf{z}(\mathbf{p}) - \mathbf{U}_Z \mathbf{c}\|^2, \quad (7)$$

where $\mathbf{z}(\mathbf{p})$ denotes the normalized gradients of \mathbf{t} , as defined in Eq. (4). This optimization can be efficiently solved in an inverse compositional alternating manner, as shown in [3,5,39].

3.2. Method

The main idea behind the proposed tool is to take advantage of the generalization qualities of AOMs by building a model using annotated images with various poses and expressions and generate the annotations on images with different poses and expressions. Specifically, let us denote a database that consists of N_{subj} subjects as \mathcal{DB} . We assume that for each subject, images with different expressions $\{E_j\}$, $j \in \{1, 2, \dots, N_{exp}\}$, and poses $\{P_k\}$, $k \in \{1, 2, \dots, N_{pos}\}$ are available. Let \mathcal{V} be a subset of annotated images and \mathcal{Q} a subset of non-annotated images of \mathcal{DB} . The goal of our tool is to (1) generate annotations for the subjects in \mathcal{Q} which appear in \mathcal{V} with different expressions and poses, and (2) generate annotations for the subjects of \mathcal{Q} that are not included in \mathcal{V} . For example, in Multi-PIE, the annotations for subjects with expressions “disgust” at 0° and “neutral” at 15° are provided and we want to produce the annotations for subjects with expression “disgust” at 15° . In this case the annotated and non-annotated subsets are defined as $\mathcal{V} = \{\text{Disgust}, 0^\circ, \text{Neutral}, 15^\circ\}$ and $\mathcal{Q} = \{\text{Disgust}, 15^\circ\}$, respectively.

In order to annotate the images in \mathcal{Q} , we first train an AOM using the images in \mathcal{V} . The trained model is employed within an iterative fitting procedure which aims to augment the set of correctly annotated images in \mathcal{V} and build a more powerful AOM. Specifically, we fit the trained AOM to each image in \mathcal{Q} and manually classify the fitting results into two sets: *Good* denoted as \mathcal{Q} and *Bad* denoted as $\mathcal{W} = \mathcal{Q} \setminus \mathcal{Q}$. After this procedure is completed, the initial set of annotated images is augmented with \mathcal{Q} , i.e. $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{Q}$, a new AOM is built using

the updated \mathcal{V} and the fitting procedure is repeated. This iterative process is repeated until the cardinality of the subset \mathcal{W} has not changed between two consecutive iterations i.e., $|\mathcal{W}_t| - |\mathcal{W}_{t-1}| = 0$, thus we end up with fitting results for all the images in \mathcal{Q} . Note that we employ DPMs [18] to estimate the initial landmark locations for the first iteration of the above procedure.

In case \mathcal{Q} has multiple images per subject (e.g. Multi-PIE, XM2VTS, FRGC-V2, AR), the above method can be extended to further improve the generated annotations. Specifically, let us assume that we have a subset of images for each subject $\mathcal{Q}_p \subseteq \mathcal{Q}$ with N_p number of images each, where $p \in \{1, 2, \dots, N_{subj}\}$. For each such subset, we build and fit a Person Specific Model (PSM) [40] in an “one-vs-rest” manner, that is we fit each image $\mathbf{i} \in \mathcal{Q}_p$ using the PSM trained on the rest $N_p - 1$ images. This person-specific adaptation further improves the results, especially since we employ person-specific AOMs. The generated annotations of the images in \mathcal{Q} can be further manually improved, as a final step, although the above methodology ensures that minor corrections will be required. Fig. 2 and Algorithm 1 present the flowchart and pseudocode, respectively, of the proposed semi-automatic annotation technique. Finally, the above method can be readily applied to annotate a database \mathcal{DB}_1 using an already annotated database \mathcal{DB}_2 by setting $\mathcal{V} = \mathcal{DB}_2$ and $\mathcal{Q} = \mathcal{DB}_1$.

3.3. Annotations

In this section, we present how the proposed tool was used in order to re-annotate the databases presented in Section 2. The advantages of the generated annotations¹ are twofold: (1) They all have the same landmark configuration, i.e. the one employed in Multi-PIE (Fig. 1a), and (2) in many cases they are more accurate than the original ones.

Algorithm 1 Semi-automatic database annotation tool

Require: Annotated subset \mathcal{V} , Non-annotated subset \mathcal{Q}

Ensure: Annotations of \mathcal{Q}

- 1: Initialize landmarks locations of \mathcal{Q} .
- 2: Initialize $\mathcal{Q}_1 = \emptyset$, $\mathcal{V}_1 = \mathcal{V}$, and $\mathcal{W}_1 = \mathcal{Q}$.
- 3: $t = 1$.
- 4: **repeat**
- 5: Train an AOM using \mathcal{V}_t .
- 6: Fit the AOM to \mathcal{W}_t .
- 7: Manually classify the fittings to \mathcal{Q}_t (*Good*) and $\mathcal{W}_{t+1} = \mathcal{W}_t \setminus \mathcal{Q}_t$ (*Bad*).
- 8: Update $\mathcal{V}_{t+1} \leftarrow \mathcal{V}_t \cup \mathcal{Q}_t$.
- 9: $t \rightarrow t + 1$.
- 10: **until** $|\mathcal{W}_t| - |\mathcal{W}_{t-1}| = 0$
- 11: **if** multiple images per subject in \mathcal{Q} . **then**
- 12: **for** each subject $p = 1, 2, \dots, N_{subj}$ **do**
- 13: $\mathcal{Q}_p \subseteq \mathcal{Q}$ is the subset with the N_p images of the subject.
- 14: **for** each image $\mathbf{i} \in \mathcal{Q}_p$ **do**
- 15: Train a person-specific AOM using $\mathcal{Q}_p \setminus \{\mathbf{i}\}$.
- 16: Fit the person-specific AOM to the image \mathbf{i} .
- 17: **end for**
- 18: **end for**
- 19: **end if**
- 20: Check and manually correct, if necessary, the generated annotations of \mathcal{Q} .

3.3.1. Multi-PIE

The available Multi-PIE annotations cover only the neutral expression with pose $[-45^\circ, 45^\circ]$ and multiple non-neutral expressions with pose 0° . We employed the proposed tool to annotate 12,570 images for 6 expressions, all 337 subjects and poses in range $[-30^\circ, 30^\circ]$.

3.3.2. XM2VTS

The images of XM2VTS's first session were semi-automatically annotated by setting \mathcal{V} to be the subjects of Multi-PIE with neutral expression and $[-15^\circ, 15^\circ]$ poses. Subsequently, the annotated images of the first session were employed to annotate the images of the second session, and so on for all four available sessions. This procedure resulted in annotating 2360 images.

3.3.3. AR

A procedure similar to the one used for XM2VTS was used to generate annotations for the neutral images of AR. For images having a specific expression E_j , we used the annotated neutral images of AR

and the images with the corresponding expression and frontal pose from the Multi-PIE.

3.3.4. FRGC-V2

In the case of FRGC-V2, we first annotated a subset consisting of two images per subject with two different illumination conditions. This subset was annotated by employing images from Multi-PIE with six expressions and $[-15^\circ, 15^\circ]$ poses as \mathcal{V} . The rest of FRGC-V2 was annotated using this initial subset.

3.3.5. LFPW

Since LFPW database does not provide information regarding pose and expression characteristics for any image, we manually clustered the images in different poses $\{P_k\}$ in the range $[-30^\circ, 30^\circ]$. The images of each such pose cluster were semi-automatically annotated using images from Multi-PIE with the same pose.

3.3.6. HELEN, AFW, IBUG

The rest of in-the-wild databases were annotated using a common procedure. Specifically, \mathcal{Q} consisted of the non-annotated

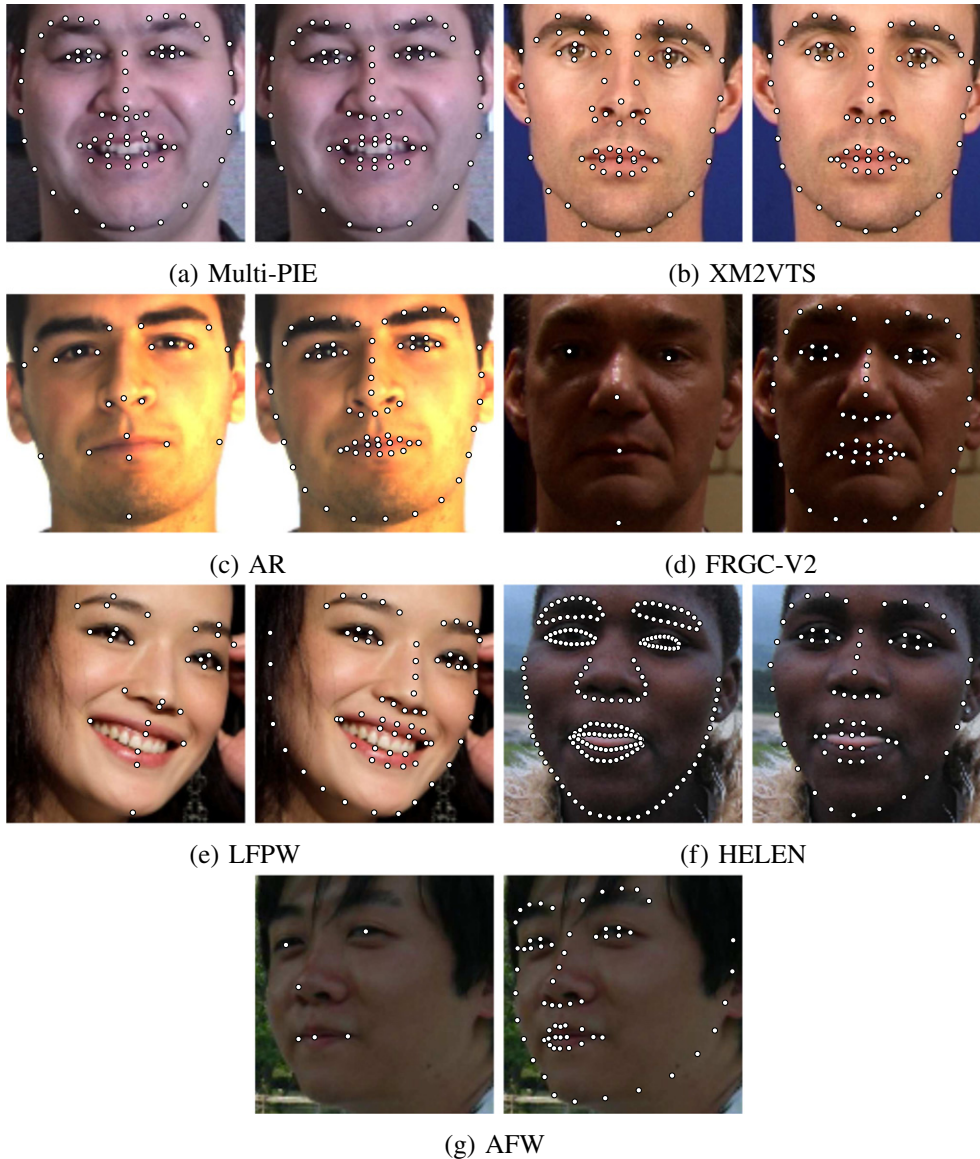


Fig. 3. Examples of the annotated images. For each database, the image on the *left* has the original annotations and the one on the *right* shows the annotations generated by the proposed tool. Note that in the case of Multi-PIE, even though the original and generated annotations have the same configuration, the generated ones are more accurate.

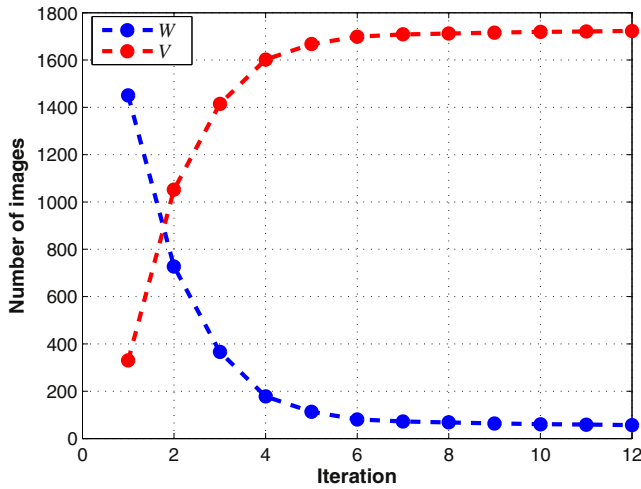


Fig. 4. The cardinality of \mathcal{W} and \mathcal{V} per iteration.

database \mathcal{DB}_i , and \mathcal{V} was set equal to all the rest annotated in-the-wild databases \mathcal{DB}_j , $j = \{1, 2, \dots, i-1\}$.

Fig. 3 shows examples for each database with the original annotations and the annotations produced using the proposed semi-automatic methodology.

3.4. Efficiency

In order to assess the efficiency of the semi-automatic tool, we conducted the following experiment. We used the testing set of Helen database (330 images) as the annotated subset \mathcal{V} while the non-annotated set \mathcal{Q} was formed by randomly selecting 1450 images from the training set of the same database. Note that the selected images exhibit significant variations in pose, illumination conditions, and occlusion. Then, we applied the proposed semi-automatic tool in order to generate the annotations. Fig. 4 visualizes the cardinality of \mathcal{W} and \mathcal{V} at each iteration until the termination of the procedure. The tool managed to generate annotations of good quality for 1393 out of 1450 images. The annotations for the rest 57 images were not adequately good mainly due to the existence of extreme poses, occlusions and illumination conditions. Given that an expert human annotator needs around 5 min to manually annotate from scratch one image, we have to spend 7250 min in order to annotate all the images. Instead, by using the proposed tool we dropped the requirement time for the creation of annotations in 1671 min. More specifically, we spent 820 min for the manual classification of fittings in *Good* and *Bad*, 549 min in order to refine the automatically created 1393 annotations, and 285 min for the manual annotation of the rest 57 images.

3.5. Discussion

In order to assess the variance of the manually annotated landmarks, we considered the simplest case of annotating images with frontal faces without any occlusion or expression. To this end, we selected such images of $N = 80$ different subjects with frontal pose from the Multi-PIE database. All these images were manually annotated by three expert annotators. Fig. 5 plots the variance of the manual annotations for each landmark point using an ellipse. Note that the ellipses are colored based on the standard deviation of the annotations, normalized by the size of the face.

This experiment shows that the agreement level among the annotators is high for the landmarks that correspond to the eyes and mouth. This is due to the fact that these landmarks are located to facial features which are very distinctive across all human faces.

Instead, the standard deviation is high for landmarks that do not have a clear semantic meaning. The chin is the most characteristic example of this category, as it demonstrates the highest variance. Finally, the result of this experiment suggests that it is more reliable to report the performance of landmark localization techniques using the 51-point mark-up (after removing the points of the face's boundary), as done in both 300-W competitions.

4. 300 Faces In-The-Wild Challenge

In this section, we present the 300 Faces In-The-Wild Challenge (300-W), the first facial landmark localization challenge that was held twice, in 2013 and 2015. The ultimate goal of the challenge is to provide a fair comparison between different automatic facial landmark detection methods. To this end, the 300-W database was collected and annotated using the same unified annotation scheme described in Section 3, in order to be used as testing set. Section 4.1 gives more details about the database and Sections 4.2 and 4.3 analyse the results of the two conducts of the competition.

4.1. 300-W database

The 300-W database⁴ is a newly-collected challenging dataset that consists of 300 *Indoor* and 300 *Outdoor* in-the-wild images. It covers a large variation of identity, expression, illumination conditions, pose, occlusion and face size. The images were downloaded from google.com by making queries such as “party”, “conference”, “protests”, “football” and “celebrities”. Compared to the rest of in-the-wild datasets, the 300-W database contains a larger percentage of partially-occluded images and covers more expressions than the common “neutral” or “smile”, such as “surprise” or “scream”. We annotated the images with the 68-point mark-up of Fig. 1a, using the semi-automatic methodology presented in Section 3. The images of the database were carefully selected so that they represent a characteristic sample of challenging but natural face instances under totally unconstrained conditions. Thus, methods that achieve accurate performance on the 300-W database can demonstrate the same accuracy in most realistic cases. Consequently, the experimental results on this database indicate how far the research community is from an adequately good solution to the problem of automatic facial landmark localization.

Table 2 summarizes the characteristics of the database. Many images of the database contain more than one annotated faces (293 images with 1 face, 53 images with 2 faces and 53 images with [3, 7] faces). Consequently, the database consists of 600 annotated face instances, but 399 unique images. Finally, there is a large variety of face sizes. Specifically, 49.3% of the faces have size in the range [48.6k, 2.0M] and the overall mean size is 85k (about 292×292) pixels.

4.2. 300-W challenge: first conduct (2013)

The first conduct of the 300-W challenge² was held in conjunction with IEEE International Conference on Computer Vision (ICCV) in 2013 [32].

4.2.1. Training

LFPW, AFW, HELEN, XM2VTS and FRGC-V2 were provided for training, along with the corrected annotations produced with the semi-automatic annotation tool (Section 3). The fact that only a very small proportion of images in LFPW and HELEN have expressions different than “smile” motivated us to collect and annotate

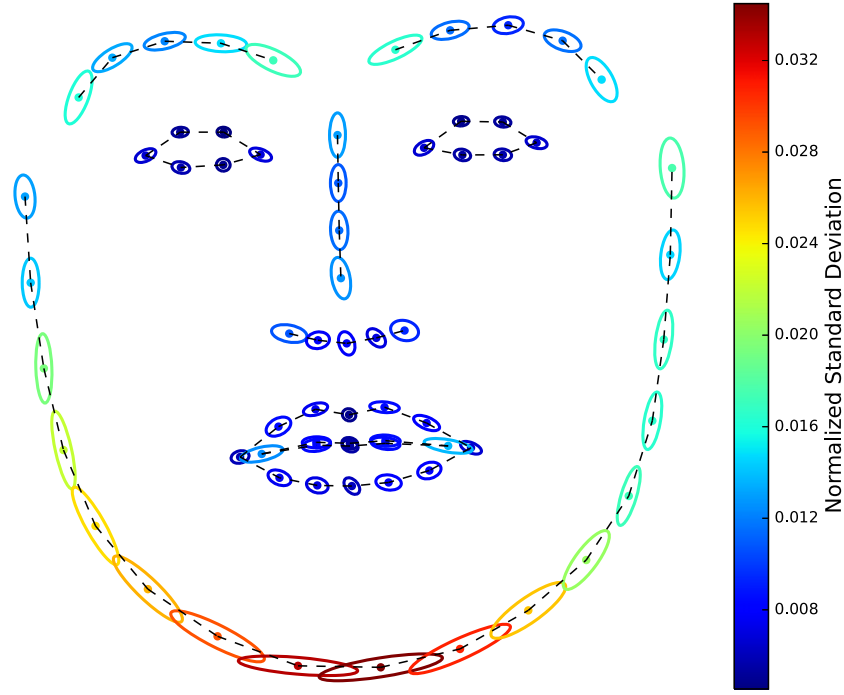


Fig. 5. Each ellipse denotes the variance of each landmark point with regard to three expert human annotators. The colors of the points rank them with respect to their standard deviation normalized by the face size. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

the IBUG database. It consists of 135 images with highly expressive faces under challenging poses and was provided to the participants as an additional option for training. Furthermore, we computed the bounding boxes of all the aforementioned databases by using our in-house face detector, the one that is also employed in [17], which is a variant of [18]. Both the annotations and the bounding boxes were made publicly available at the challenge's website². Note that the participants were encouraged but not restricted to use only the provided training sets and annotations.

4.2.2. Testing

To ensure a fair comparison between the submitted methodologies, participants did not have access to the 300-W testing database. They were requested to send us the compiled (binary) files of their pre-trained systems. On our behalf, we extracted the face's bounding box for each of the testing images using the same methodology as the one employed for the training images⁴. These bounding boxes were passed in to the submitted executables as initializations. The accuracy of the fitting results was measured by the point-to-point RMS error between each fitted shape and the ground truth annotations, normalized by the face's interocular distance, as proposed in [18]. Specifically, by denoting the fitted and ground truth shapes as $\mathbf{s}^f = [x_1^f, y_1^f, \dots, x_N^f, y_N^f]^T$ and $\mathbf{s}^g = [x_1^g, y_1^g, \dots, x_N^g, y_N^g]^T$ respectively, then the error between them is computed as:

$$\text{RMSE} = \frac{\sum_{i=1}^N \sqrt{(x_i^f - x_i^g)^2 + (y_i^f - y_i^g)^2}}{d_{\text{outer}} N}, \quad (8)$$

Table 2

Overview of the characteristics of the 300-W database.

	Indoor	Outdoor
# of faces	300	300
# of images	222	177
Image size (range in pixels)	[20.3k, 17.3M]	[27.2k, 21.0M]
Face size (range in pixels)	[5.0k, 0.8M]	[4.7k, 2.0M]
Interocular distance (range in pixels)	[42, 477]	[39, 805]

where d_{outer} is the interocular distance computed as the Euclidean distance between the outer points of each eye, as shown in Fig. 6. For the employed landmark configuration of Fig. 1a, the interocular distance is defined as $d_{\text{outer}} = \sqrt{(x_{37}^g - x_{46}^g)^2 + (y_{37}^g - y_{46}^g)^2}$.

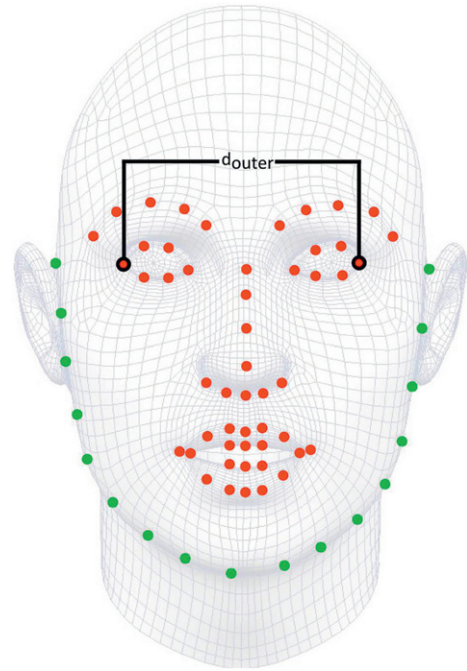


Fig. 6. The 51-point mark-up is a subset of the 68-points one after removing the 17 points of the face's boundary. The interocular distance is defined between the outer points of the eyes.

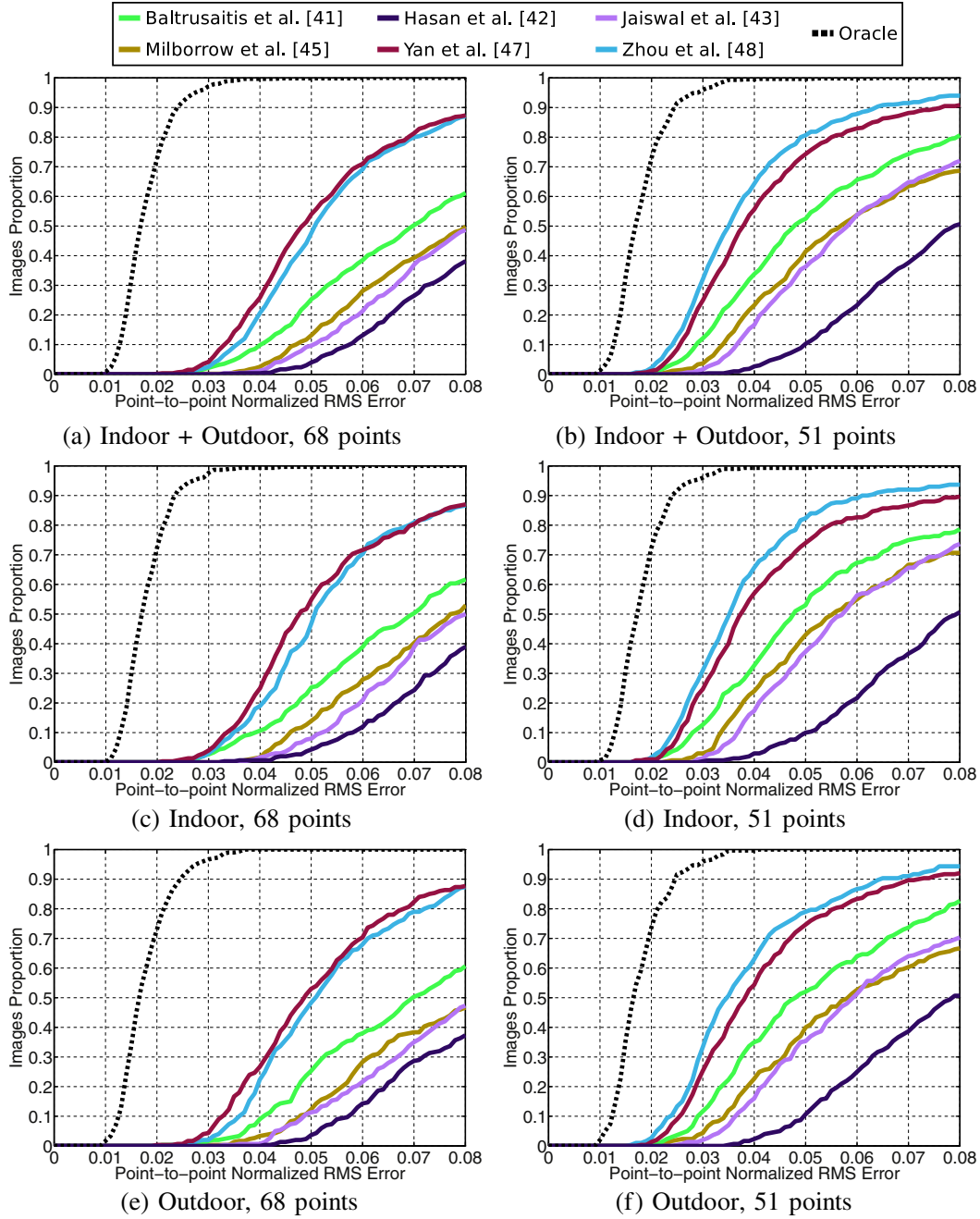


Fig. 7. Fitting results of the first conduct of the 300-W challenge in 2013. The plots show the Cumulative Error Distribution (CED) curves with respect to the landmarks (68 and 51 points) and the conditions (indoor, outdoor or both).

4.2.3. Participants

In total, there were six participants in this version of the challenge. Below is a brief description of the submitted methods:

- Baltrusaitis et al. [41] propose a probabilistic patch expert technique that learns non-linear and spatial relationships between the pixels and the probability of a landmark being aligned. To fit the model they propose a novel non-uniform regularized landmark mean-shift optimization technique which takes into account the reliabilities of each patch expert.
- Hasan et al. [42] first apply a nearest neighbor search using global descriptors and, then, aim to align local neighbors by dynamically fitting a locally linear model to the global key-point configurations of the returned neighbors. Neighbors are also used to define restricted areas of the input image in which

they apply local discriminative classifiers. Finally, an energy minimization approach is applied in order to combine the local classifier predictions with the dynamically estimated joint keypoint configuration model.

- Jaiswal et al. [43] use Local Evidence Aggregated Regression [44], in which local patches provide evidence of the location of the target facial point using Support Vector Regressors.
- Milborrow et al. [45] approach the problem with Active Shape Models (ASMs) that incorporate a modified version of SIFT descriptors [46]. They employ multiple ASMs and utilize the one that best estimates the face's yaw pose.
- Yan et al. [47] employ a cascade regression framework, where a series of regressors are utilized to progressively refine the shape initialized by the face detector. In order to handle inaccurate initializations from the face detector, they generate

multiple hypotheses and learn to rank or combine them in order to get the final results. They estimate the parameters in both “learn to rank” and “learn to combine” using a structural Support Vector Machine framework.

- Zhou et al. [48] propose a four-level Convolutional Network Cascade, where each level is trained to locally refine the outputs of the previous network levels. Moreover, each level predicts an explicit geometric constraint (face region and component position) to rectify the inputs of the next levels, which improves the accuracy and robustness of the whole network structure.

4.2.4. Results

The performance of the submitted systems was assessed based on both the 68 and 51 points. As shown in Fig. 6, the 51 points are a subset of the 68 points after removing the 17 points of the face’s boundary. Fig. 7 shows the Cumulative Error Distribution (CED) curves using the error metric of Eq. (8). The plots are divided based on the number of points (68 and 51) as well as the image subsets (*Indoor*, *Outdoor* and *Indoor + Outdoor*). Table 3 reports the median absolute deviation of the results and Fig. 10 shows some indicative fitting shapes.

All methodologies demonstrated a lower performance on *Outdoor* scenes. The main reason for this is the illumination condition variance which is much smaller within an *Indoor* environment. However, another factor affecting the performance is that the *Outdoor* images have larger variation in facial expressions compared to the *Indoor* ones. This is because we picked specific keywords for the selection of *Outdoor* images, such as “sports” and “protest”, which ended up in a big number of images with various expressions, such as “surprise” and “scream”, that are much more challenging than the expressions that are commonly seen in the *Indoor* ones, such as “smile” and “neutral”. We decided to announce two winners: one from an academic institution and one from industry. Based on the results, the winners were (a) Yan et al. [47] from The National Laboratory of Pattern Recognition at the Institute of Automation of the Chinese Academy of Sciences, and (b) Zhou et al. [48] from Megvii company. It is worth to mention that all groups achieved better results in the case of 51 points.

In order to show whether there is any room for further improvement on the performance, we also report an Oracle curve. We built a statistical shape model using the shapes of the training databases, as explained in Eq. (1), and kept the first 25 components. Using this model, we compute and plot the reconstruction error for each shape of the 300-W database. The reconstruction of a shape \mathbf{s} is achieved by first projecting as $\mathbf{p}_r = \mathbf{U}_r^T(\mathbf{s} - \bar{\mathbf{s}})$, and then reconstructing as $\mathbf{s}_r = \bar{\mathbf{s}} + \mathbf{U}_s \mathbf{p}_r$. The resulting curve shows that the 300-W dataset is not saturated and there is considerable room for further improvement.

4.3. 300-W challenge: second conduct (2015)

The second conduct of the 300-W challenge was completed in the beginning of 2015. The biggest difference compared to the previous conduct is that we were no longer providing the bounding boxes of the images to the fitting methods. On the contrary, the participants were required to submit systems that perform both face detection and landmark localization. The three main reasons that led us to this change are:

1. Various techniques perform differently when initialized with bounding boxes that cover different facial regions. For example, DPMs [18] tend to return bounding boxes that only include facial texture and not any of the subject’s hair, as usually done by the Viola–Jones detector [49].

Table 3

Median absolute deviation of the fitting results of the first conduct of 300-W challenge in 2013, reported for both 68 and 51 points.

Participant	68 points	51 points
Baltrusaitis et al. [41]	0.0486	0.0388
Hasan et al. [42]	0.0543	0.0551
Jaiswal et al. [43]	0.0527	0.0506
Milborrow et al. [45]	0.1126	0.1145
Yan et al. [47]	0.0211	0.0199
Zhou et al. [48]	0.0205	0.0182
Oracle	0.0038	0.0040

2. There are methods, like DPMs [18] and Pictorial Structures [11,12], that do not require any initialization.
3. There are algorithms for which the training is coupled with the face detector, such as SDM [19].

Of course, this change made the task even more challenging than before, since the search region of each image became much larger with a lot of background information.

4.3.1. 300-W images cropping

As mentioned in Section 4.1, many of the 300-W images contain more than one faces, which are not necessarily annotated. Consequently, we cropped the images so that they all included only one face. The cropping was performed in such a way to ensure that (1) only a single face is included in each image and (2) DPMs [18] and Viola–Jones [49] achieve the best true positive rate that they possibly can. Table 4 reports the characteristics of the cropped images. Naturally, the only thing that changes compared to the ones of the initial images in Table 2 is the image size (resolution). The mean size of the cropped images is 0.4 M pixels, which is much smaller than the 3.3 M pixels of the non-cropped images. Fig. 8 shows some representative examples of the way that the images were cropped. Note that the cropped images are provided along with the original images of the 300-W database⁴.

4.3.2. Training

The training instructions were the same as in the previous conduct. The authors were encouraged, but not restricted, to use LFPW, AFW, HELEN, IBUG, FRGC-V2 and XM2VTS databases with the provided annotations.

4.3.3. Testing

The testing procedure followed the same rules as in the previous version of the challenge. The participants were required to submit compiled pre-trained systems, the performance of which was evaluated using the metric of Eq. (8). The submitted systems could return nothing in case no face was detected or the detected face was estimated to be a false positive. Consequently, in order to facilitate the participants and make the competition less dependent to a face detector’s performance, we suggested them to use one of the face detection methods that took part in the Face Detection Data Set and Benchmark (FDDB) [50]. Finally, in this conduct of the competition, the submitted methods were also assessed with respect to their computational costs and a maximum limit of 2 min per image was typically set.

Table 4

Overview of the characteristics of the cropped images of the 300-W database.

	<i>Indoor</i>	<i>Outdoor</i>
# of faces	300	300
# of images	300	300
Image size (range in pixels)	[16.2k, 3.3M]	[11.2k, 4.5M]
Face size (range in pixels)	[5.0k, 0.8M]	[4.7k, 2.0M]
Interocular distance (range in pixels)	[42, 477]	[39, 805]



Fig. 8. Indicative examples of the way the images were cropped for the second conduct of the 300-W challenge.

4.3.4. Participants

In total, there were five participants in this version of the challenge. Below is a brief description of the submitted methods:

- Čech et al. [51] propose an algorithm where the sum of individual landmark scoring functions, that are trained by a structured output SVM classifier, is maximized with respect to the camera pose by fitting a parametric 3D shape model. They explicitly handle self-occlusions by excluding the corresponding contributions from the data term, which allows their algorithm to operate correctly for a large range of viewing angles. Their landmark localization procedure is initialized with a manually-engineered framework that involves a commercial face detector and the estimation of the initial 3D pose.
- Deng et al. [52] use a multi-view, multi-scale and multi-component cascade shape regression model. Their model learns view-specific cascaded shape regressors using multi-scale HOG features as shape-index features and is optimized with a multi-scale strategy that eliminates the risk of getting stuck on local minima.
- Fan et al. [53] approach the problem by proposing a deep learning system that consists of a cascade of multiple Convolutional Neural Networks (CNNs) that are optimized in a coarse-to-fine strategy in order to improve accuracy. Their fitting procedure is initialized with their implementation of the Viola–Jones face detector [49].
- Martinez et al. [54] employ a cascaded regression method that makes use of the $L_{2,1}$ norm in order to increase the robustness to poor initializations or partial occlusions compared to the commonly used least-squares regressor. They also attempt to improve the results by using multiple initializations with different spatial translations and four different head pose rotations.

- Uříčář et al. [55] propose a real-time multi-view methodology based on DPMs. They utilize different view-specific models in order to deal with the problem of self-occlusions and cover a large range of head poses. The model parameters are learned through structured output SVM. The dynamic programming optimization is performed in a coarse-to-fine search strategy that allows real-time performance.

4.3.5. Results

The number of participants in this version of the competition was 5. Fig. 9 shows the CED curves of the submitted methodologies. Table 5 reports the number of images for which an estimation of the landmarks was returned, the mean absolute deviation of the results, as well as the mean computational costs. The common subset of images for which all methods returned a detection consists of 519 images. Fig. 11 shows some indicative fitting results.

Similar to the first conduct of the competition, we selected to announce two winners: one from academia and one from industry. Based on the results, the winners of the second conduct of the competition are (a) Deng et al. [52] from the B-DAT Laboratory at the Nanjing University of Information and Technology, and (b) Fan et al. [53] from Megvii company. Even though the technique of Fan et al. [53] is slightly more accurate, it returns results for 526 images, as opposed to the one of Deng et al. [52] that detects the landmarks in 599 images and has a small mean absolute deviation. It is worth to notice that some systems employed an unreliable face detector. This seems to be the case with Čech et al. [51] and Uříčář et al. [55]. Their systems returned an output for about 98.5% of the images, however about 50% of the images resulted with very high fitting error partially because of false positive face detections. This is the reason why their mean absolute deviation is high. Moreover, only the submission by Martinez et al. [54] managed to return a detection for all 600 images and the results indicate that even though their methodologies are

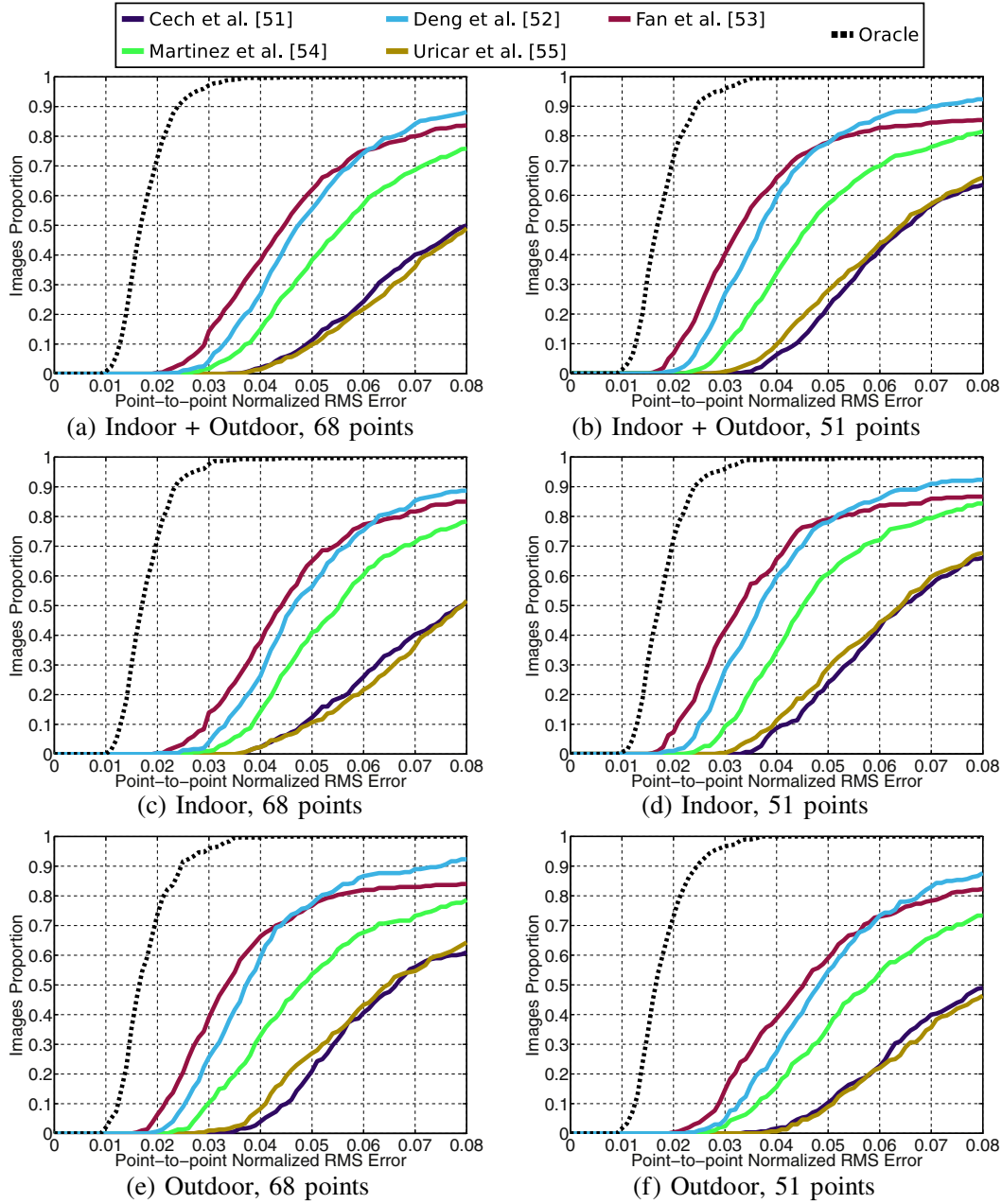


Fig. 9. Fitting results of the *second* conduct of the 300-W challenge in 2015. The plots show the Cumulative Error Distribution (CED) curves with respect to the landmarks (68 and 51 points) and the conditions (indoor, outdoor or both).

not the most accurate ones, they are very robust with small mean absolute deviations. Finally, the systems of Deng et al. [52] and Fan et al. [53] were also the fastest ones.

5. Discussion and conclusions

The results of both conducts of the 300-W challenge shown in Figs. 7 and 9 clearly prove that, even though much progress has been made during the last years, the research community is still far from accurately solving the problem of face alignment and that there is much room for further improvement. This is indicated by the gap between the participants' curves and the Oracle, which is the minimum error that can be achieved using the specific training databases. Table 6 reports the percentage of images with error less

than {0.02, 0.03, 0.04, 0.05, 0.06} for the top techniques of both competitions as well as the Oracle and makes it obvious that the gap is still huge, especially for small error values.

Table 6 also shows that there was a small improvement on the state-of-the-art performance between the first and the second conduct of the challenge. The top performing methodologies have relatively small differences and are close to each other. One of the main reasons behind this progress is the plethora of training data from which discriminative methods can greatly benefit. For example, techniques like Yan et al. [47] (cascade regression framework) and Zhou et al. [48] (convolutional network framework), can continually achieve better results with continuous rise in the amount of training data.

Additionally, the 300-W challenge was only focused on the task of *sparse* facial landmark points detection. Alignment using dense



Fig. 10. Fitting examples of the *first* conduct of the 300-W challenge in 2013. Each row shows the fitted landmarks for each participating method.

landmark mark-ups is much more difficult and the performance would get worse. This is because the more landmarks exist in the shape, there is more ambiguity about the semantic locations at which they are located. Consider for example the 41 boundary landmark points of the HELEN mark-up in Fig. 1f. Their locations have no special semantic discrimination. On the contrary they are just located with an approximately equal distance between them. Consequently, it is very hard to accurately detect such points since there is no

discriminative texture information that describes them and which could drive the fitting procedure. This highlights the need to further research how to select a relatively high number of landmark points that are capable to describe all the characteristic areas of an object.

Moreover, another factor that contributed towards creating more accurate and efficient alignment techniques is the great progress in the task of face detection. Most landmark localization methodologies are very sensitive to the initialization, thus the face detection

Table 5

Second conduct of the 300-W challenge. 2nd column: Number of images for which an estimation of the landmarks was returned. 3rd and 4th columns: The mean absolute deviation of the fitting results for both 68 and 51 points. 5th column: Mean computational cost per method.

Participant	# of images with detection	mad		Timings (secs)
		68 points	51 points	
Čech et al. [51]	591 (98.5%)	0.1047	0.0998	4.05
Deng et al. [52]	599 (99.8%)	0.0226	0.0213	1.97
Fan et al. [53]	526 (87.7%)	0.0309	0.0294	1.29
Martinez et al. [54]	600 (100%)	0.0514	0.0497	42.5
Uříčář et al. [55]	592 (98.7%)	0.0970	0.0945	3.46
Oracle	–	0.0038	0.0040	–

Table 6

Percentage of images with fitting error less than the specified values for the winners of the first (Yan et al. [47], Zhou et al. [48]) and second (Deng et al. [52], Fan et al. [53]) 300-W challenges, and Oracle. The error is based on 68 points using both indoor and outdoor images.

Method	<0.02	<0.03	<0.04	<0.05	<0.06
Yan et al. [47]	0.17%	4.17%	25.8%	54.0%	71.0%
Zhou et al. [48]	0%	2.50%	20.7%	47.7%	69.2%
Deng et al. [52]	0.17%	4.33%	26.8%	55.5%	74.3%
Fan et al. [53]	0.33%	14.3%	38.2%	62.0%	75.2%
Oracle	72.8%	97.2%	99.7%	99.8%	100%

performance. The results presented in the Face Detection Data Set and Benchmark (Fddb) [50] show that current state-of-the-art techniques achieve very good true positive rates. However, there is still

room for further improvement especially on images with in-the-wild conditions.

Finally, most current research effort focuses on detecting the facial landmarks and not tracking them within video sequences. We strongly believe that more attention should be given towards developing techniques that can track facial points in a robust manner,



Fig. 11. Fitting examples of the *second* conduct of the 300-W challenge in 2015. Each row shows the fitted landmarks for each participating method.

even under difficult conditions such as camera movement, disappearance and re-appearance of the face, challenging background and lighting, etc. Consequently, we believe that one promising step towards this direction would be the organization of a challenge, similar to the 300-W one, that focuses on facial landmark points tracking. The biggest difficulty of such a competition would be the annotation of the thousands of frames of the videos. However, using semi-automatic annotation tools as the one proposed in this paper, the task would be simplified and annotations could be efficiently generated.

Acknowledgments

This work is funded by the EPSRC project EP/J017787/1 (4D-FAB). The work of S. Zafeiriou is also partially supported by the EPSRC project EP/L026813/1 Adaptive Facial Deformable Models for Tracking (ADAManT). The work of G. Tzimiropoulos is also partially supported by EPSRC project EP/M02153X/1 Facial Deformable Models of Animals. The work of M. Pantic is further supported by the European Community Horizon 2020 [H2020/2014–2020] under grant agreement no. 645094 (SEWA).

References

- [1] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (6) (2001) 681–685.
- [2] I. Matthews, S. Baker, Active appearance models revisited, *Int. J. Comput. Vis.* 60 (2) (2004) 135–164.
- [3] G. Tzimiropoulos, J. Alabort-i-Medina, S. Zafeiriou, M. Pantic, Generic active appearance models revisited, *Proceedings of Asian Conference on Computer Vision (ACCV)*, Daejeon, Korea, 2012, pp. 650–663.
- [4] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 593–600.
- [5] 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*, Special Issue on Facial Biometrics in-the-wild 9, 2014, 2024–2034.
- [6] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, S. Zafeiriou, HOG active appearance models, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014, pp. 224–228.
- [7] J. Alabort-i-Medina, S. Zafeiriou, Bayesian active appearance models, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2014, pp. 3438–3445.
- [8] J. Alabort-i-Medina, S. Zafeiriou, Unifying holistic and parts-based deformable model fitting, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2015, pp. 3679–3688.
- [9] E. Antonakos, J.A. i medina, G. Tzimiropoulos, S. Zafeiriou, Feature-based Lucas-Kanade and active appearance models, *IEEE Trans. Image Process.* 24 (9) (2015) 2617–2632.
- [10] G. Tzimiropoulos, M. Pantic, Gauss-Newton deformable part models for face alignment in-the-wild, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2014, pp. 1851–1858.
- [11] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, *Int. J. Comput. Vis.* 61 (1) (2005) 55–79.
- [12] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: people detection and articulated pose estimation, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2009, pp. 1014–1021.
- [13] E. Antonakos, J. Alabort-i-Medina, S. Zafeiriou, Active pictorial structures, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2015, pp. 5435–5444.
- [14] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, *Comput. Vis. Image Underst.* 61 (1) (1995) 38–59.
- [15] D. Cristinacce, T. Cootes, Feature detection and tracking with constrained local models, *Proceedings of British Machine Vision Conference (BMVC)*, 3, 2006, pp. 929–938.
- [16] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, *Int. J. Comput. Vis.* 91 (2) (2011) 200–215.
- [17] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2013, pp. 3444–3451.
- [18] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2012, pp. 2879–2886.
- [19] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2013, pp. 532–539.
- [20] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, *Int. J. Comput. Vis.* 107 (2) (2014) 177–190.
- [21] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2014, pp. 1859–1866.
- [22] G. Tzimiropoulos, Project-out cascaded regression with an application to face alignment, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2015, pp. 3659–3667.
- [23] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2014, pp. 1685–1692.
- [24] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2014, pp. 1867–1874.
- [25] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, *Image Vis. Comput.* 28 (5) (2010) 807–813.
- [26] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Second international conference on audio and video-based biometric person authentication, 964, Citeseer 1999, pp. 965–966.
- [27] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2005, pp. 947–954.
- [28] A.M. Martinez, The AR Face Database, CVC Technical Report 24, 1998.
- [29] P.N. Belhumeur, D.W. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2011, pp. 545–552.
- [30] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, *Proceedings of European Conference on Computer Vision (ECCV)* Springer 2012, pp. 679–692.
- [31] M. Kostinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on Benchmarking Facial Image Analysis Technologies, IEEE, 2011, pp. 2144–2151.
- [32] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces in-the-Wild Challenge: the first facial landmark localization challenge, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on 300 Faces in-the-Wild Challenge (300-W), Sydney, Australia, 2013, pp. 397–403.
- [33] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, S. Zafeiriou, Menpo: a comprehensive platform for parametric image alignment and visual deformable models, *Proceedings of the ACM International Conference on Multimedia*, Open Source Software Competition, Orlando, FL, USA, 2014, pp. 679–682.
- [34] X. Jia, H. Yang, A. Lin, K.-P. Chan, I. Patras, Structured semi-supervised forest for facial landmarks localization with face mask reasoning, *Proceedings of British Machine Vision Conference (BMVC)*, 2014.
- [35] M. Pedersoli, T. Tuytelaars, L.V. Gool, Using a deformation field model for localizing faces and facial points under weak supervision, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2014, pp. 3694–3701.
- [36] Y. Tong, X. Liu, F.W. Wheeler, P.H. Tu, Semi-supervised facial landmark annotation, *Comput. Vis. Image Underst.* 116 (8) (2012) 922–935.
- [37] Y. Wu, Z. Wang, Q. Ji, A hierarchical probabilistic model for facial feature detection, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2014, pp. 1781–1788.
- [38] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, Subspace learning from image gradient orientations, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (12) (2012) 2454–2466.
- [39] G. Papandreou, P. Maragos, Adaptive and constrained algorithms for inverse compositional active appearance model fitting, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE, 2008, pp. 1–8.
- [40] R. Gross, I. Matthews, S. Baker, Generic vs. person specific active appearance models, *Image Vis. Comput.* 23 (12) (2005) 1080–1093.
- [41] T. Baltrusaitis, P. Robinson, L.-P. Morency, Constrained local neural fields for robust facial landmark detection in the wild, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on 300 Faces in-the-Wild Challenge (300-W), 2013, pp. 354–361.
- [42] M. Hasan, C. Pal, S. Moale, Localizing facial keypoints with global descriptor search, neighbour alignment and locally linear models, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on 300 Faces in-the-Wild Challenge (300-W), 2013, pp. 362–369.
- [43] S. Jaiswal, T. Almaev, M. Valstar, Guided unsupervised learning of mode specific models for facial point detection in the wild, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on 300 Faces in-the-Wild Challenge (300-W), 2013, pp. 370–377.
- [44] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (5) (2013) 1149–1163.
- [45] S. Milborrow, T. Bishop, F. Nicolls, Multiview active shape models with SIFT descriptors for the 300-W face landmark challenge, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W)*, Workshop on 300 Faces in-the-Wild Challenge (300-W), 2013, pp. 378–385.
- [46] D.G. Lowe, Object recognition from local scale-invariant features, *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2, IEEE 1999, pp. 1150–1157.

- [47] J. Yan, Z. Lei, D. Yi, S. Li, Learn to combine multiple hypotheses for accurate face alignment, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W), Workshop on 300 Faces in-the-Wild Challenge (300-W)*, 2013. pp. 392–396.
- [48] E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, Extensive facial landmark localization with coarse-to-fine convolutional network cascade, *Proceedings of IEEE International Conference on Computer Vision (ICCV-W), Workshop on 300 Faces in-the-Wild Challenge (300-W)*, 2013. pp. 386–391.
- [49] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, IEEE 2001, pp. 511–518.
- [50] V. Jain, E. Learned-Miller, FDDB: a benchmark for face detection in unconstrained settings, *Tech. Rep. UM-CS-2010-009*, University of Massachusetts, Amherst, 2010.
- [51] J. Čech, V. Franc, M. Uříčář, J. Matas, Multi-view facial landmark detection by using a 3D shape model, *Image Vis. Comput.* 47 (2016) 60–70.
- [52] J. Deng, Q. Liu, J. Yang, D. Tao, M³ CSR: multi-view, multi-scale and multi-component cascade shape regression, *Image Vis. Comput.* 47 (2016) 19–26.
- [53] H. Fan, E. Zhou, Approaching human level facial landmark localization by deep learning, *Image Vis. Comput.* 47 (2016) 27–35.
- [54] B. Martinez, M.F. Valstar, L2,1-based regression and prediction accumulation across views for robust facial landmark detection, *Image Vis. Comput.* 47 (2016) 36–44.
- [55] M. Uříčář, V. Franc, D. Thomas, A. Sugimoto, V. Hlaváč, Multi-view facial landmark detector learned by the structured output SVM, *Image Vis. Comput.* 47 (2016) 45–59.