300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge

Christos Sagonas[1], Georgios Tzimiropoulos[1,2], Stefanos Zafeiriou[1] and Maja Pantic[1,3]
[1] Comp. Dept., Imperial College London, UK
[2] School of Computer Science, University of Lincoln,U.K.
[3] EEMCS, University of Twente, The Netherlands
{c.sagonas, gt204, s.zafeiriou, m.pantic}@imperial.ac.uk

Abstract

Automatic facial point detection plays arguably the most important role in face analysis. Several methods have been proposed which reported their results on databases of both constrained and unconstrained conditions. Most of these databases provide annotations with different mark-ups and in some cases the are problems related to the accuracy of the fiducial points. The aforementioned issues as well as the lack of a evaluation protocol makes it difficult to compare performance between different systems. In this paper, we present the 300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge which is held in conjunction with the International Conference on Computer Vision 2013, Sydney, Australia. The main goal of this challenge is to compare the performance of different methods on a new-collected dataset using the same evaluation protocol and the same mark-up and hence to develop the first standardized benchmark for facial landmark localization.

1. Introduction

The problem of detecting a set of predefined facial fiducial points has been the focus in computer vision for more than two decades. Recent research efforts have focused on the collection and to some extend the annotation of real-world datasets of facial images captured in-the-wild, as well as on the development of algorithms that are capable of operating robustly on such imagery. However, a proper evaluation of what has been achieved so far and how far we are from attaining satisfactory performance is yet to be carried out.

The need for benchmarking the efforts towards automatic facial landmark detection is particularly evident from the fact that different researchers follow different testing approaches, different dataset and performance measures. Examples include the following.

- Authors compare their approaches against other previously published methods but they do so by using in many cases completely difference datasets for training compared to the dataset that the original method was trained on.
- Authors report comparison on specific datasets by replicating the originally presented curves and not the experiment.
- In some cases, authors report results on datasets that now only part of them can be used by the community because some of the training/testing images are no longer publicly available.

Additional challenges in benchmarking efforts in automatic facial landmark detection stem from the limitations of currently available databases/annotations. Although works in [19, 2, 8, 9] resulted in the very first annotated face databases collected in-the-wild, these datasets have a number of limitations like providing sparse annotations or, in some cases, annotations of limited accuracy but most importantly they all use different annotation schemes producing different fiducial points.

This paper, describes the First Automatic Facial Landmark Detection in-the-Wild Challenge, 300-W, which is held in conjunction with the International Conference on Computer Vision 2013, Sydney, Australia. The aim of this challenge is to provide a fair comparison between the different automatic facial landmark detection methods in a new in-the-wild dataset.

2. Existing in-the-wild databases

Annotated databases are extremely important in computer vision. Therefore, a number of databases containing faces with different facial expressions, poses, illumination and occlusion variations have been collected in the past [5], [13], [10]. However, the majority of these don't include

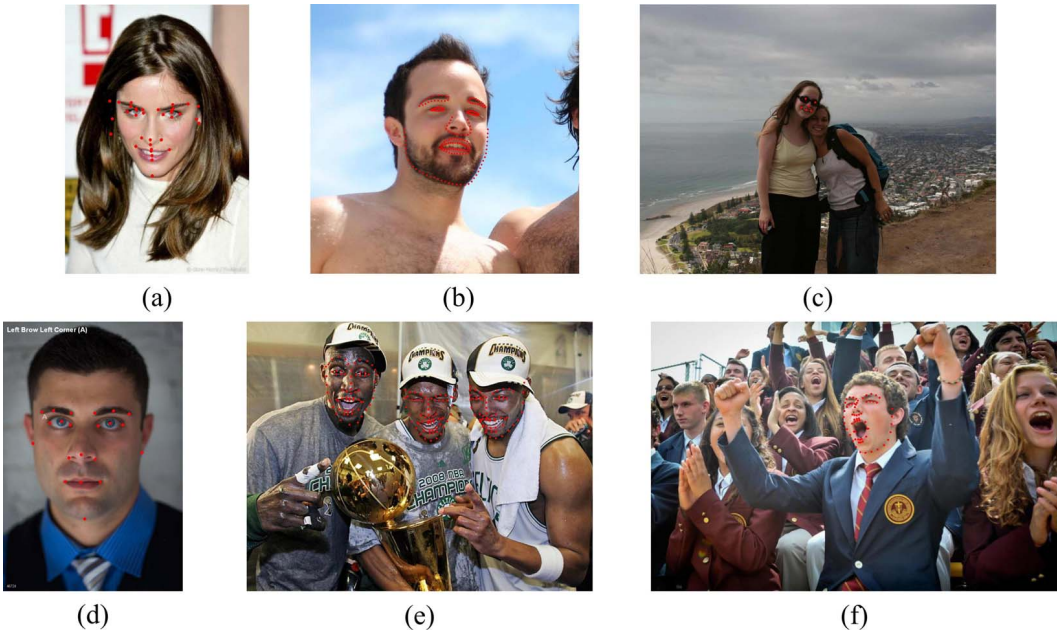


Figure 1. Annotated images from (a) LFPW, (b) HELEN, (c) AFW, (d) AFLW, (e) 300-W 'Indoor' and (f) 300-W 'Outdoor'.

images under unconstrained conditions. Hence, recently a number of databases containing faces in unconstrained, 'in-the-wild' conditions have been collected. The most well-known in-the-wild facial databases are: LFPW [2], HELEN [9], AFW [19] and AFLW [8]. In the following we provide an overview of the above datasets and comment on the different variations and the available mark-ups they provide.

LFPW: The Labeled Face Parts in-the-wild (LFPW) database contains 1,287 images downloaded from google.com, fickr.com, and yahoo.com. The images contain large variations including pose, expression, illumination and occlusion. The provided ground truth consists of 35 landmark points. An example of an image taken from the LFPW database along with the corresponding annotated landmarks is depicted in Figure 1 (a).

HELEN: The Helen database consists of 2,330 annotated images collected from the Flickr. The images are of high resolution containing faces of size sometimes greater than 500×500 pixels. The provided annotations are very detailed and contain 194 landmark points. Figure 1 (b) depicts an annotated image from HELEN.

AFW: The Annotated Faces in-the-wild (AFW) database contain 250 images with 468 faces. Six facial landmark points for each face are provided. Figure 1 (c) depicts an annotated image from AFW.

AFLW: The Annotated Facial Landmarks in-the-wild (AFLW) [8] contains 25,000 images of 24,686 subjects downloaded from Flickr. The images contain a wide range of natural face poses and occlusions. Facial landmark annotations are available for the whole database. Each annotation consists of 21 landmark points (Figure 1 (d)).

Table 2. Pose variations

Databases	Poses			
	$-30^0 : -15^0$	$-15^0 : 0^0$	$0^0 : 15^0$	$15^0 : 30^0$
LFPW	2.8%	44.25%	50.44%	2.51%
HELEN	2.15%	46.64%	47.9%	3.31%
AFW	6.23%	47.18%	41.25%	5.34%
300-W	5.83%	46%	41%	7.17%

The aforementioned databases, cover large variations including: different subjects, poses, illumination, occlusion etc. In order to make an analysis about the variations in expression we classified manually all images based to six different expressions: 'Neutral', 'Surprise', 'Squint', 'Smile', 'Disgust', 'Scream'. Additionally, information about occlusion is indicated as 'YES'/'NO' answer. As it can be seen in Table 1 for the majority of databases the most common expression is 'Neutral' and 'Smile'. More specifically, more than 80% of the images in each database capture these two expressions only.

3. The 300-W dataset

The 300-W test set is aimed to examine the ability of current systems to handle naturalistic, unconstrained face images. The test set should cover different variations like unseen subjects, pose, expression, illumination, background, occlusion, and image quality. Additionally, the test images should cover many different expressions instead of mainly

Table 1. Expression and occlusion variations of LFPW, HELEN, AFW, and 300-W databases.

Database	Expressions						Occluded		# Landmarks
	'Neutral'	'Surprise'	'Squint'	'Smile'	'Disgust'	'Scream'	Yes	No	
LFPW	48.66%	8.05%	1.34%	39.73%	0.44%	1.78%	18.31%	81.69%	35
HELEN	43.03%	2.12%	3.33%	49.09%	2.43%	0.00%	13.03%	86.97%	194
AFW	40.06%	10.09%	3.86%	43.62%	1.19%	1.18%	19.59%	80.41%	6
300-W	37.17%	12.34%	4.84%	29.83%	1.66%	14.16%	29.83%	70.17%	68

smiles which is the case in some of the existing databases (Table 1). Thus, we created a new dataset consisting of 300 'Indoor' and another 300 'Outdoor' images, respectively.

We were mainly interested in images with spontaneous expressions. Hence, the tags we used in order to download the image from google.com were simple keywords like "party", "conference", "protests", "football", "celebrities" etc. As it can be seen from Table 1, the collected test set covers all expressions. Furthermore, in our database faces are more frequently occluded than on other databases. Finally, a similar proportion of poses to AFW are included in our test set (Table 2).

The ground-truth for 300-W was created by using the semi-automatic methodology for facial landmark annotation proposed as in [15, 16] followed by additional manual correction. The resulting ground-truth for each image consists of 68 landmark points similar to well-established landmark configuration of MultiPIE [5]. Figures 2 (e) and (f) depict 'Indoor' and 'Outdoor' annotated images respectively.

4. The 300-W challenge

All participants had their algorithms run on the 300-W test set using the same face-bounding box initialization. To facilitate the training procedure, landmark annotations for LFPW [2], HELEN [9], AFW [19] and XM2VTS [13] became available from the 300-W challenge's website [1]. The testing procedure and the performance evaluation was carried out based on the same mark-up (i.e. set of facial landmarks) of provided annotated images (originally used in [5], Figure 2 (a)) . The exact procedures for training and testing stages are as follows.

- **Training:** The recently collected in-the-wild datasets LPFW, AFW, and Helen have been re-annotated using semi-supervised methodology in [15] and the well-established landmark configuration of MultiPIE [5] (68 points, Figure 2 (a)). For extra accuracy the final annotations were manually corrected by another annotator. In addition, XM2VTS collected in laboratory conditions, have been also re-annotated using the same mark-up. Since the LFPW and HELEN contain a small number of faces displaying an expression other than smile, for training purposes we collected another 135 images with highly expressive faces. All annotations and the bounding boxes as produced by our ibug-variation of the face detector proposed in [19] and were made publicly available through the website challenge. All the participants had the option to train their facial landmark detection systems using the aforementioned training sets and the provided annotations.

- **Testing:** Participants did not have access to the testing data. They sent binary code with their trained algorithms to the organisers, who run each algorithm on the 300-W test set using the same face-bounding box initialization for all algorithms. As baseline method we used the project-out inverse compositional Active Appearance Models algorithm described in [12], implemented using the edge-structure features described in [3].

[1] http://ibug.doc.ic.ac.uk/resources/300-W/

5. Evaluation results

5.1. Performance measure

In order to evaluate the accuracy of the submitted methods, we used as error measure the point-to-point Euclidean distance [4], normalized by the Euclidean distance between the outer corners of the eyes. Facial landmark detection performance was assessed on both the 68 fiducial points mark-up of Figure 2(a) and the 51 points which are the points without the boundary (Figure 2(b)). Finally, the cumulative error rates for the cases of 68 and 51 points were returned to the participants.

5.2. Participation

In total, six participants contributed to the challenge. In the following we describe very briefly the submitted methods.

Baltrusaitis et al. [1] proposed a probabilistic patch expert (landmark detector) that can learn non-linear and spatial relationships between the input pixels and the probability of a landmark being aligned. To fit the model a novel Non-uniform Regularised Landmark Mean-Shift optimisation technique which takes into account the reliabilities of each patch expert was used.

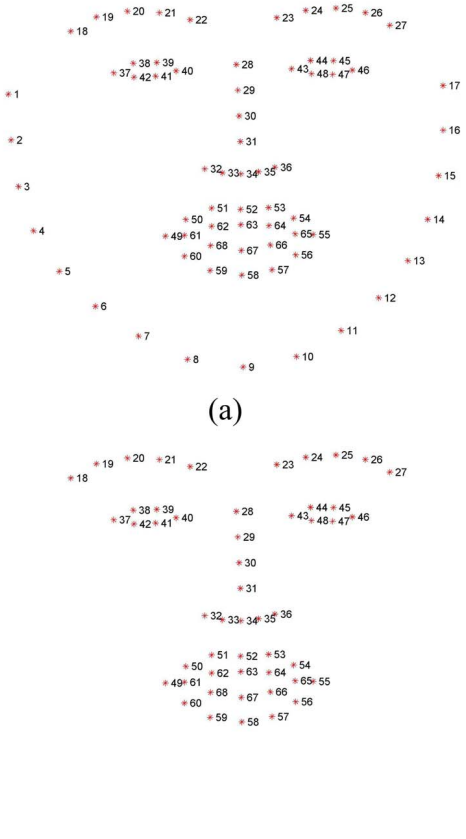


Figure 2. The 68 and 51 points mark-up used for provided annotations.

Milborrow et al. [14] approached the challenge with Active Shape Models (ASMs) that incorporated a modified version of SIFT descriptors. Multiple ASMs were used, searching for landmarks with the ASM that best matches the face's estimated yaw.

Yan et al. [17] build their method on a cascade regression framework, where a series of regressors were utilized to progressively refine the shape initialized by the face detector. In order to handle inaccurate initializations from the face detector, multiple hypotheses are generated and learned to rank or combine both in order to get the final results. The parameters in both 'learn to rank' and 'learn to combine' can be estimated in a structural SVM framework.

Zhou et al. [18] proposed a four-level convolutional network cascade, where each level was trained to locally refine the outputs of previous network levels. In addition, each level predicts an explicit geometric constraint (face region and component position) to rectify the inputs of the next levels. In that way improves the accuracy and robustness of the whole network structure.

Jaiswal et al. [7] employed Local Evidence Aggregated Regression [11], in which local patches provided evidence of the location of the target facial point using Support Vector Regressors.

Kamrul et al. [6] firstly applied a nearest neighbour search using global descriptors. Then, an alignment of local neighbours by dynamically fitting a locally linear model to the global keypoint configurations of the returned neighbours was employed. Neighbours are also used to define restricted areas of the input image in which they apply local discriminative classifiers. Finally, an energy function based minimization approach was applied in order to combine the local classifier predictions with the dynamically estimated joint keypoint configuration model.

5.3. Results

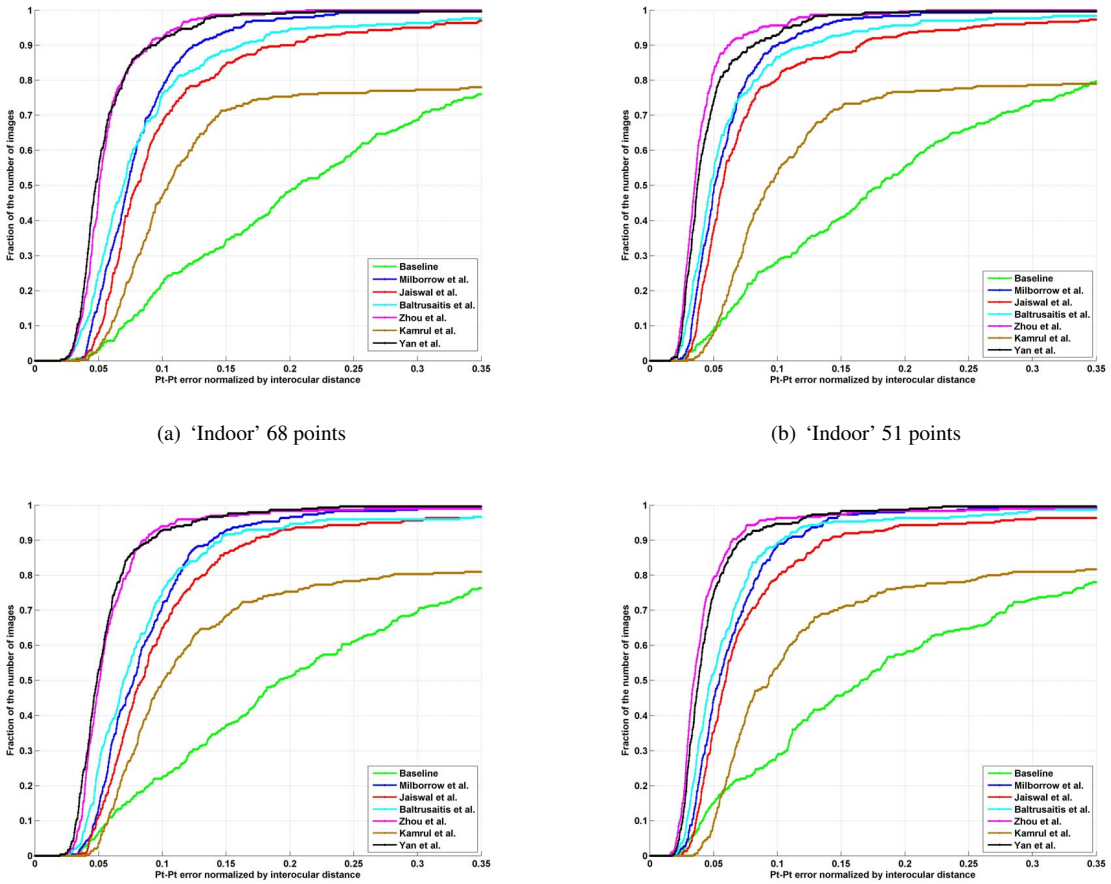
Figure 3 depicts the results of all participant. As it can be seen, all submitted methods outperform the baseline method both in cases of 'Indoor' and 'Outdoor' datasets. We decided to announce two winners one from an academic institution and one from industry. The basic criterion to select a winner team is based on the mean error of its performance in 'Indoor' and 'Outdoor' images. The winners are: a)Yan et al. [17] from The National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation of the Chinese Academy of Sciences (academia) and b) Zhou et al. [18] from Megvii company (industry). It is worth to mention that all groups achieved better results in the case of 51 points.

For all submissions, we observed a lower performance rate on 'Outdoor' scenes. A major reason for this is illumination. Another factor with an important effect is the variation of facial expressions. As we picked specific keywords for the selection of 'Outdoor' images such as 'sports' and 'protest', we were able to include different facial expressions like 'Surprise' and 'Scream'. The aforementioned expressions were more challenging compared to the common 'Indoor' ones such as 'Smile' and 'Neutral'.

In order to decide whether there is any further room for improvement we conducted the following experiment. All shapes from the given training databases were used to create a statistical shape model by applying Procrustes and Principal Component Analysis. We reconstructed the test shapes of 300-W by keeping only 25 eigen-shapes which correspond to the 98% of the total shape variance existing in testing set. The shape parameters were computed by projecting each test shape on the shape eigenspace. Finally, the reconstruction error was computed using the point-to-point Euclidean distance. Figure 4 depicts the cumulative error curves of the reconstruction error both in cases of 'Indoor' and 'Outdoor'. As it can be seen 300-W is not saturated and there is considerable room for further improvement.

6. Conclusion

This paper describes the 300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge held



(a) 'Indoor' 68 points

(b) 'Indoor' 51 points

(c) 'Outdoor' 68 points

(d) 'Outdoor' 51 points

Figure 3. The cumulative error rates produced by participants for (a) 'Indoor' with 68 points, (b) 'Indoor' with 51 points, (c) 'Outdoor' with 68 points and (d) 'Outdoor' with 51 points.

in conjunction with the International Conference on Computer Vision 2013, Sydney. The main challenge of the competition was to localize a set of 68 fiducial points in a newly collected test set with 2x300 facial images captured in real-world unconstrained settings (300 'Indoor' and 300 'Outdoor'). As a part of the challenge the most well-known databases XM2VTS, LFPW, HELEN, and AFW were re-annotated using the same mark-up and became available from the 300-W challenge's website. In total six participants submitted to the challenge. As can be seen the current technology is mature enough to produce very good results but there is considerable space for further improvement.

7. Acknowledgements

The work of Christos Sagonas and Stefanos Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4DFAB), while the work of Giorgios Tzimiropoulos by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG).

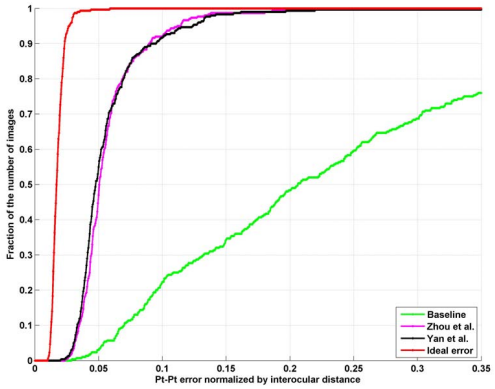
References

[1] T. Baltrusaitis, L.-P. Morency, and P. Robinson. Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

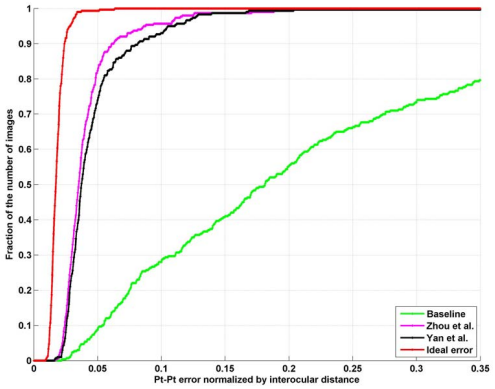
[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE, 2011.

[3] T. F. Cootes and C. J. Taylor. On representing edge structure for model matching. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–1114. IEEE, 2001.

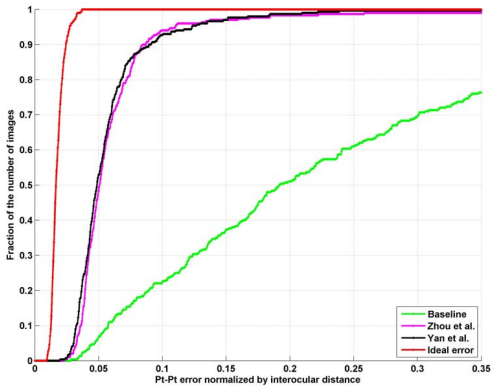
[4] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 17, pages 929–938, 2006.



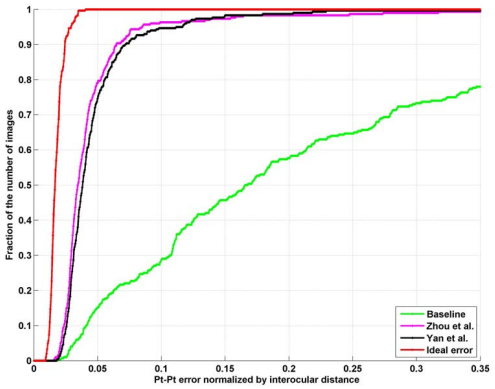
(a) 'Indoor' 68 points



(b) 'Indoor' 51 points



(c) 'Outdoor' 68 points



(d) 'Outdoor' 51 points

Figure 4. The best performing methods from academia and industry as well as, the reconstructed test-shapes for (a) 'Indoor' with 68 points, (b) 'Indoor' with 51 points, (c) 'Outdoor' with 68 points and (d) 'Outdoor' with 51 points.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[6] K. Hasan Md., S. Moalem, and C. Pal. Localizing facial keypoints with global descriptor search, neighbour alignment and locally linear models. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

[7] S. Jaiswal, T. Almaev, and M. Valstar. Guided unsupervised learning of mode specific models for facial point detection in the wild. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

[8] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[9] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[10] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[11] B. Martinez, M. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression based facial point detection. 2012.

[12] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[13] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.

[14] S. Milborrow, T. Bishop, and F. Nicolls. Multiview active shape models with sift descriptors for the 300-w face landmark challenge. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

[15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Computer Vision and Pattern Recognition Workshops*

(CVPRW), 2013 IEEE Conference on, pages 896–903. IEEE, 2013.

[16] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic. Generic active appearance models revisited. In *Computer Vision–ACCV 2012*, pages 650–663. Springer, 2013.

[17] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for face alignment. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

[18] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Facial landmark localization with coarse-to-fine convolutional network cascade. In *Computer Vision Workshops (ICCV-W), Sydney, Australia, 2013 IEEE Conference on*. IEEE, 2013.

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