The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results

Jie Shen[1], Stefanos Zafeiriou[1], Grigorios G. Chrysos[1], Jean Kossaifi[1],
Georgios Tzimiropoulos[2], Maja Pantic[1,3]

[1]Department of Computing, Imperial College London, U.K.
[2]School of Computer Science, University of Nottingham, U.K.
[3]EEMCS, University of Twente, N.L.

{jie.shen07, s.zafeiriou, g.chrysos, jean.kossaifi12}@imperial.ac.uk,
yorgos.tzimiropoulos@nottingham.ac.uk, m.pantic@imperial.ac.uk

Abstract

Detection and tracking of faces in image sequences is among the most well studied problems in the intersection of statistical machine learning and computer vision. Often, tracking and detection methodologies use a rigid representation to describe the facial region [1], hence they can neither capture nor exploit the non-rigid facial deformations, which are crucial for countless of applications (e.g., facial expression analysis, facial motion capture, high-performance face recognition etc.). Usually, the non-rigid deformations are captured by locating and tracking the position of a set of fiducial facial landmarks (e.g., eyes, nose, mouth etc.). Recently, we witnessed a burst of research in automatic facial landmark localisation in static imagery. This is partly attributed to the availability of large amount of annotated data, many of which have been provided by the first facial landmark localisation challenge (also known as 300-W challenge). Even though now well established benchmarks exist for facial landmark localisation in static imagery, to the best of our knowledge, there is no established benchmark for assessing the performance of facial landmark tracking methodologies, containing an adequate number of annotated face videos. In conjunction with ICCV'2015 we run the first competition/challenge on facial landmark tracking in long-term videos. In this paper, we present the first benchmark for long-term facial landmark tracking, containing currently over 110 annotated videos, and we summarise the results of the competition.

[1]Usually, this representation involves a rectangular or ellipse-shaped bounding-box

1. Introduction

Nowadays, face detection has matured enough so as to provide effective and efficient solutions in imagery captured in arbitrary conditions (referred to as "in-the-wild"). Some of the recent face detection systems are now fast enough to be integral parts of very popular electronic commodities, such as various kinds of cameras. The interested reader may refer to [67] for recent advances in face detection "in-the-wild". Usually, for efficiency purposes, face detection algorithms use a rigid representation in order to describe the facial region (e.g., using a rectangular or ellipse-like shape). Face tracking is another field of research that has received considerable attention in the past years. Similar to face detection, face tracking algorithms use a rigid rectangular representation of the face. This is mainly attributed to the fact that face tracking is modelled as a face detection or as a general object tracking problem [43, 23, 61, 24]. Even in the most recent generic object tracking benchmarks, which contain a considerable amount of short face videos, face is annotated using a rectangular bounding box [6, 3, 1].

Recently, it was shown that in order to achieve state-of-the-art results in a series of important computer vision applications, such as face recognition/verification and facial expression analysis, it is important to provide an enhanced representation of the face that contains the locations of several key facial landmarks [47, 52]. Hence, not surprisingly, facial landmark localization in static "in-the-wild" imagery is a problem that has received a lot of attention [8, 55, 11, 12, 9, 56, 57, 54, 15, 68, 42, 63]. Such methodologies achieving good performance have been presented and some of them have been integrated in certain devices (e.g., cameras installed in automobiles for monitoring the behaviour of the driver).

This progress would not be feasible without the efforts

made by the scientific community to design and develop both benchmarks with high-quality landmark annotations [16, 30, 28, 44, 46], as well as rigorous protocols for performance assessment. Arguably, the most comprehensive such benchmark was firstly presented in [46] and then continued in [44] (so-called 300-W benchmark). The annotated data from the 300-W benchmark are now used by the majority of scientific and industrial community for training and testing facial landmark localization algorithms [44, 46].

Even though a considerable amount of high-quality annotated data have been collected for benchmarking efforts regarding facial landmark localization, to the best of our knowledge there exists no benchmark for facial landmark tracking in long in-the-wild videos [2]. Currently, evaluation of facial landmark tracking algorithms in-the-wild is performed in the following two rather limited ways:

- In a pure qualitative manner. Facial landmark tracking is considered by many as a by-product of facial landmark detection. Hence, in order to demonstrate the effectiveness of their algorithm for facial landmark tracking, the authors often visualise the tracking results in a small number of short videos, which they subsequently upload on YouTube (as a kind of supplementary material) [63, 57, 42].

- In a quantitative manner using a small number of very short (2-3 secs) videos. Since annotating facial videos with regards to facial landmarks is a tedious, expensive and labour intensive procedure, researchers in the field often evaluate their algorithms using a very small number (around 4-5) of short videos [14, 45].

In this paper, we take a significant step further and present a new comprehensive benchmark, as well as summarise the results of the first challenge on landmark tracking/detection of a set of 60+ fiducial points in long-term facial videos (duration of each video is approximately 1 min). To the best of our knowledge, this is the first time that a comprehensive attempt to benchmark the efforts in the field is presented.

2. Existing Face Databases for Assessing Tracking Technologies

Rigid and non-rigid tracking of faces and facial features has been a very popular topic of research over the past twenty years [21, 29, 20, 17, 25, 37, 43, 23, 61, 19, 49, 14, 31, 34, 62, 63, 64, 10, 32, 40, 51, 39, 50, 35, 24, 26, 53, 56, 57, 54].

Rigid face tracking has been generally treated along the same lines as general object tracking [37, 23, 49, 32, 33, 43, 61, 31]. To this end, several short face sequences have been annotated with regards to the facial region (using a bounding box style annotation). One of the first sequences that has been annotated for this task is the so-called Dudek sequence [2] [3]. Nowadays several such sequences have been annotated and are publicly available [1, 3, 6].

Non-rigid tracking of faces can be further subdivided into tracking of certain facial landmarks [29, 17, 35, 40, 14, 63, 34, 62, 10, 51, 39, 53, 64] or tracking/estimation of dense facial motion [19, 21, 20, 65, 50, 26]. The performance of non-rigid dense facial tracking methodologies was usually assessed by using markers [19], simulated data [50], visual inspection [19, 21, 20, 65, 50, 26] or indirectly by the use of the dense facial motion for certain tasks, such as expression analysis [20, 65, 26]. Regarding tracking of facial landmarks, up until recently, the preferred method for assessing the performance was visual inspection in a number of selected facial videos [63, 53]. Other methods were assessed on a small number of short (few seconds in length) annotated facial videos [45, 14]. Until recently the longest annotated facial video sequence was the so-called talking face [4] which was used to evaluate many tracking methods [38, 10]. The talking face video comprises of 5000 frames (around 200 secs) taken from a video of a person engaged in a conversation [4]. The talking face video was initially tracked using an Active Appearance Model (AAM) that had a shape model of a total of 68 landmarks. The tracked landmarks were visually checked and manually corrected.

While these lines were written another annotated database was presented in [64] [4]. The database was built using videos from the Distracted Driver Face(DDF) and Naturalistic Driving Study (NDS)[5]. The DDF dataset contains 15 sequences, with a total of 10,882 frames. Each sequence displays a single subject posing as the distracted driver in a stationary vehicle or indoor environment. 12 out of 15 videos were recorded with subjects sitting inside of a vehicle. Five of them were recorded in the night under infrared (IR) light and the others were recorded during the daytime under natural lighting. The remaining three were recorded indoors. The NDS database contains 20 sub-sequences of driver faces recorded during a drive conducted between the Blacksburg, VA and Washington, DC areas (NDS is the more challenging than DDF since it's videos are of lower spatial and temporal resolution). Each sequence of NDS database consists of a one-minute video recorded at 15 frames per second (fps) with a resolution of 360×240. For both datasets one in every ten frames was annotated using either 49 landmarks, for near-frontal-faces, or 31 landmarks, for profile faces. The database contains

[2] While these lines were written, another considerably smaller benchmark appeared in [64].

[3] The Dudek sequence has been annotated with regards to certain facial landmarks only to be used for the estimation of an affine transformation

[4] In a private communication, the authors of [64] informed us that the annotated data, as described in [64], will not be made publicly available (at least not in the near future).

many extreme facial poses (90 yaw, 50 pitch) as well as many faces under extreme lighting condition (e.g., IR). In total the dataset presented in [64] contains between 2,000 to 3,000 annotated faces (please refer to [64] for example annotations).

The aim of the challenge presented in this paper is to go a significant step further and assess the performance of current and future facial landmark tracking technologies in long-term facial videos (with duration around or longer than 1 minute). To this end we have collected more than 300 videos, mostly from video-sharing websites (such as YouTube). Other sources that were used for collecting videos were the SEMAINE database [36] (22 video clips were taken). Until the submission of the binaries by the participants we have managed to annotate 114 videos (50 released for training and 64 for testing). The total and average duration of the videos was 7293 and 64 seconds, respectively. All videos were captured/encoded in 30 fps and the total number of frames was 218595 . All videos show only one person.

From the 114 videos, 86 were annotated using the semi-automatic procedure that is proposed in [18]. In brief, the procedure goes as follows: First a generic face and landmark localization scheme is applied to the video, then the generic deformable face detector is turned into a person specific one and re-applied to the video. The average recall (i.e., true positive) we achieved with the proposed procedure is more than 98% with almost 0% of false positives. The facial landmarks of a number of selected frames are detected and automatically corrected using a method similar to [13]. Next, a person-specific deformable model is trained and applied to all remaining frames. In the next step, annotators performed manual corrections to 1 in every 8 frames and a final person-specific model was trained and applied to the video. A visual inspection was performed in the final annotations and the frames. Annotations not deemed satisfactory were either corrected or removed (e.g., profile images). In total, annotating the 86 videos required 837 hours of manual labour. In order to evaluate the gain of using the above tool all frames of 6 videos were fully annotated manually. This task took around 260 hours. Hence, annotating all 86 videos by human annotators alone would have taken around 3727 hours (around 4.5 times gain by using the system). The pipeline and the annotation tool was build on top of the Menpo platform [7] and will be soon made publicly available.

The remaining 28 videos (14 for training and 14 for testing) were selected and annotated as follows: first, the state-of-the-art method Project-Out Cascaded Regression (PO-CR) of [54] was employed in a tracking-by-detection fashion where each frame is initialized by the bounding box of the previous frame. This provided per frame detection of the facial landmarks. Our tracking-by-detection framework was found to be much more robust than standard tracking; in fact, this way, all frames of the 28 sequences were tracked automatically without the need of re-initialization. Next, for each video, a person specific GN-DPM [57, 27] was trained using the fittings of PO-CR and re-fitting of the video was performed. Finally, all erroneously fitted frames were manually corrected and a final re-fitting was performed. Overall, this very simple pipeline resulted in remarkably accurate annotations. The annotations created by this pipeline can be visually inspected in training videos with IDs 112, 113, 115, 119, 120, 123, 138, 143, 144, 160, 204, 205, 223 and 225.

We separated the videos into three different categories.

- Category one: Contains videos of people recorded in well-lit conditions in various head poses (occlusions such as glasses and beards are possible but cases of occlusions by hand or another person are not be considered here). This scenario aims to evaluate algorithms that could be suitable for facial motion tracking in naturalistic well-lit conditions. Example frames with the corresponding annotation is shown in Figure 1.

- Category two: Contains videos of people recorded in unconstrained conditions (different illuminations, dark rooms, overexposed shots, etc.), displaying arbitrary expressions in various head poses but without large occlusions (similar to category one). This scenario aims to evaluate algorithms that could be suitable for facial motion analysis in real-world human-computer interaction applications. Example frames with the corresponding annotation is shown in Figure 2.

- Category three: Contains videos of people recorded in completely unconstrained conditions including the illumination conditions, occlusions, make-up, expression, head pose, etc. This scenario aims to assess the performance of facial landmark tracking in arbitrary recording conditions. Example frames with the corresponding annotation is shown in Figure 3.

All frames have been annotated with regards to the same mark-up (i.e. set of facial landmarks) used in the facial landmark localisation series of competitions [44, 46] (a total of 68 landmarks). Even though some videos contain frames of faces in profile views we did not used these frames in evaluation, since (a) currently there is no widely accepted mark-up for profile views and (b) to the best of our knowledge there are no publicly available samples of profile views in unconstrained conditions (profile view annotations can currently be found only in Multi-PIE [22]).

In July 2015 we made 50 videos (3063 seconds of annotated videos) publicly available for all three categories to be used as the training/development set. The contestants could use these videos to learn statistics in order to train

Figure 1: Representative frames of category one videos. Faces display expressions and have pose variations (but not extreme).

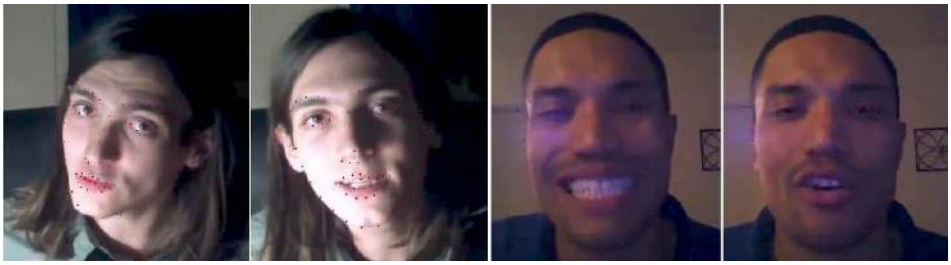


Figure 2: Representative frames of category one videos. The videos contain faces that undergo severe illumination changes or are captured under challenging illumination conditions.

Figure 3: Some frames of videos of category three. The faces are captured in very challenging conditions.

their trackers. The contestants were also allowed to use publicly or privately collected data to train their methods (e.g., they could use the datasets made available with the 300-W series of competitions [44, 46]).

Among the 64 test videos, 31 belong to category one, 19 to category two and the remaining 14 to category three. All categories contain videos of various spacial resolutions.

3. Experiments

3.1. Evaluation Methodology

To ensure a fair comparison between the submitted methodologies, the contestants did not have access to the testing dataset. All contestants had to submit a compiled (binary) file of their systems which could track 68 landmarks. The binaries submitted for the competition were all handled confidentially. They were used only for the scope of the competition and were subsequently erased after its completion. We did not provide face bounding boxes, hence all the submitted methodologies should have a face detection module in their pipelines. One of the pre-requisites was that the submitted trackers should track with a speed of at least 0.5 frame/sec. As a baseline we trained a Supervised Descent Method (SDM) [63] using annotation/images released with the 300 W competition [14] [5]. The SDM was coupled with the Matlab implementation of the Viola-Jones face detector [59].

As in 300-W competition, and since we had only one mark-up scheme, the accuracy of the fitting results was measured by the point-to-point Root-Mean-Square (RMS) error between each fitted shape and the ground truth annotations, normalized by the face's inter-ocular distance [46]. Specifically, by denoting the fitted and ground truth shapes as $\mathbf{s}^f = [x_1^f, y_1^f, \ldots, x_N^f, y_N^f]^T$ and $\mathbf{s}^g = [x_1^g, y_1^g, \ldots, x_N^g, y_N^g]^T$ respectively, the error is computed

[5] For the baseline we used only 49 landmarks out of 68

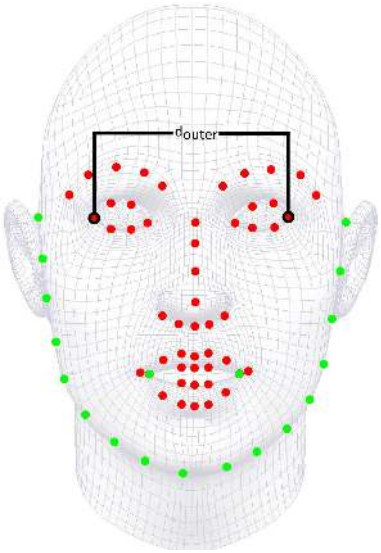


Figure 4: The data have been annotated using the 68-landmarks mark-up (both red and green landmarks). The performance of all submitted methodologies was assessed using both 68 and 49 landmarks (red subset). The inter-occular distance, used for normalization is defined between the outer points of the eyes.

as:

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

where d_{outer} is the inter-ocular distance computed as the Euclidean distance between the outer points of each eye, as shown in Figure 4. For the employed landmark configuration the inter-ocular distance is defined as $d_{outer} = \sqrt{(x_{37}^g - x_{46}^g)^2 + (y_{37}^g - y_{46}^g)^2}$.

3.2. Summary of Contestants

In total, five participants contributed to the challenge (plus another four binary file submissions not accompanied by an article describing the algorithm, hence were excluded from evaluation). A brief description of each participant's methodology (gathered after the end of the paper submission process) are given below. Each method is identified by the first author's surname.

- Method **Yang** [66]: The method uses a spatio-temporal cascade shape regression model for robust facial shape tracking. It's novelties lie in (a) the use of a multi-view cascade shape regression model that is employed to decrease the shape variance in shape regression model construction, (b) a time series regression model that is incorporated to enhance the temporal consecutiveness and (c) a novel re-initialization mechanism that is adopted to effectively and accurately locate the face when the face is misaligned or lost.

- Method **Uricar** [58]: The tracker is an extension of a well tuned tree-based Deformable Part Models (DPM) landmark detector originally developed for static images. The tracker is obtained by applying the static detector independently in each frame and using the Kalman filter to smooth estimates of the face positions as well as to compensate possible failures of the face detector.

- Method **Xiao** [48]: The method uses a multi-stage regression-based approach, which progressively initializes the shape from obvious landmarks with strong semantic meanings, e.g. eyes and mouth corners, to landmarks on face contour, eyebrows and nose bridge which have more challenging features. Compared with initialization based on mean shape and multiple random shapes, the proposed progressive initialization can very robustly handle challenging poses.

- Method **Rajamanoharan** [41]: The method uses a Multi-View Constrained Local Models which combines a global shape model with separate sets of response maps targeted at different head angles, indexed on the shape model parameters. The method explores shape- space division strategies to identify the optimal strategy.

- Method **Wu** [60]: The method applies a shape augmented regression method for face alignment, where the regression function is automatically chosen for different face shapes.

3.3. Summary of Results

The Cumulative Error Distribution (CED) curves using the 68 landmarks for all the five contestants is plotted in Figure 5 (a), (b) and (c) for videos in category one, two and three, respectively. Similarly, the CED curves using the 49 landmarks is plotted in Figure 6. We plot the curve up until 0.08 error (after that the tracking result is quite off). By performing a visual inspection we found that good tracking results can be found up until an error of 0.05.

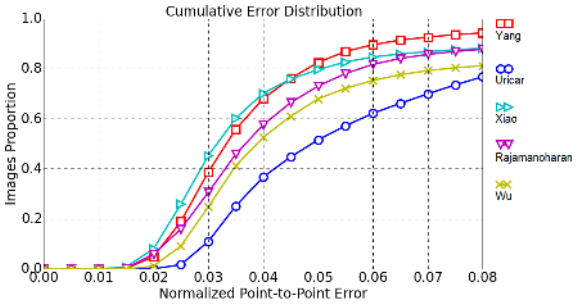
We found that the majority of the tested methods performed equally well in videos of category one and two. This may suggest that illumination changes may not be a significant factor in performance decrease. Of course more experiments are required in order to make a safe conclusion. We found that there is a significant performance drop in the videos of category three (which were the most challenging videos, containing occlusions etc.). Finally, as it has been verified in 300-W competition, the performance of all methods increases when the comparison is performed on the 49-landmark mark-up.

In order to declare the winners we measured the Area Under the Curve (AUC) until 0.08 in the three categories separately. The methods are then ranked according to ascending order AUC. Using this as the measure, the best performing method in category one is that of Yang [66] and in both category two and three is that of Xiao [48].

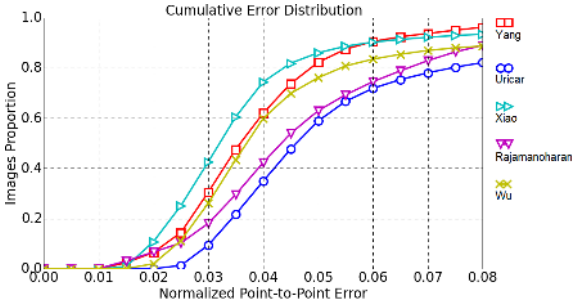
Finally, we compare the results of this competition with the results of the 300-W series of competitions. In the last run of the competition the best performing methods had 80% of the images with error less than 0.05. In comparison, the best performing methods in category one and two of this competition, which mainly contain videos which display faces that are not-occluded and do not show severe facial poses, have 90% of the images with error less than 0.05. However, the best performing methods in category three only achieved similar performance to the best performing methods in the last 300-W competition, indicating there is considerable space for improvement in this scenario.

4. Conclusions

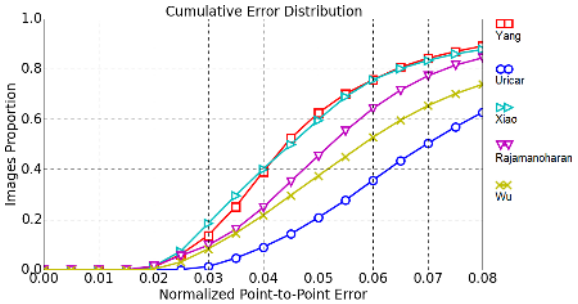
We have presented the first comprehensive benchmark for assessing the performance of facial tracking methodologies in long-term videos. The current version of the benchmark contains 114 annotated videos (around two hours of video and a total of 218,595 frames). Furthermore, we have ran the first challenge using the above data. We show that current methodologies achieve good performance in videos that display a single person and do not contain occlusions and severe head poses. Finally, we found that for fully unconstrained videos there is still a significant space for improvement.



(a) Category one



(b) Category two



(c) Category three

Figure 5: CED's for all nine contributions of the challenge using the 68 landmarks mark-up.

Acknowledgements

The work of G. Chrysos and S. Zafeiriou was supported by the EPSRC project EP/L026813/1 Adaptive Facial Deformable Models for Tracking (ADAManT). The work of J. Shen, M. Pantic and J. Kossaifi is supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA), as well as European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The annotation effort was supported by EPSRC project EP/J017787/1 (4DFAB). The work of G. Tzimiropoulos was funded by EPSRC project EP/M02153X/1 Facial Deformable Models of Animals.

References

[1] Annotated Face Videos. http://www.robots.ox.ac.uk/~stephan/dikt/, 2015. [Online; accessed 30-September-2015].

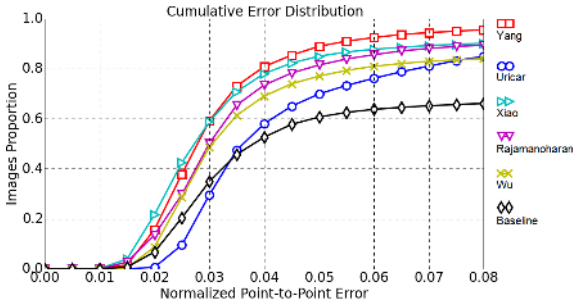
[2] Dudek Face Sequence. http://www.cs.toronto.edu/vis/projects/adaptiveAppearance.html, 2015. [Online; accessed 30-September-2015].

[3] NUS-PRO tracking challenge. http://www.lv-nus.org/pro/nus_pro.html, 2015. [Online; accessed 30-September-2015].

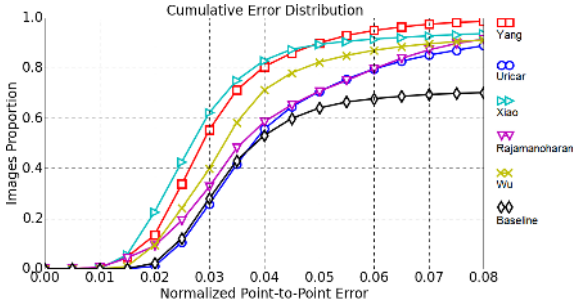
[4] Talking Face Video. http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html, 2015. [Online; accessed 30-September-2015].

[5] Transportation Research Board of the National Academies of Science. The 2nd strategic highway research program naturalistic driving study dataset. https://insight.shrp2nds.us/, 2015. [Online; accessed 30-September-2015].

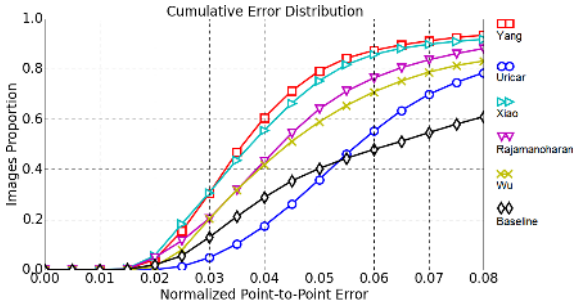
[6] Visual Tracking Benchmark. http://cvlab.hanyang.ac.kr/tracker_benchmark/, 2015. [Online; accessed 30-September-2015].



(a) Category one



(b) Category two



(c) Category three

Figure 6: CED's for all nine contributions of the challenge using the 49 landmarks mark-up.

[7] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *ACMMM*, pages 679–682. ACM, 2014.

[8] J. Alabort-i-Medina and S. Zafeiriou. Bayesian active appearance models. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3445, 2014.

[9] J. Alabort-i Medina and S. Zafeiriou. Unifying holistic and parts-based deformable model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3679–3688, 2015.

[10] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1714–1721. IEEE, 2009.

[11] E. Antonakos, J. Alabort-I-Medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 24(9):2617, 2015.

[12] E. Antonakos, J. Alabort-i Medina, and S. Zafeiriou. Active pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5435–5444, 2015.

[13] E. Antonakos and S. Zafeiriou. Automatic construction of deformable models in-the-wild. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1813–1820, 2014.

[14] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1859–1866, 2014.

[15] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1312–1320, 2015.

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

[17] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 374–381, 1995.

[18] G. Chrysos, E. Antonakos, and S. Zafeiriou. Offline deformable face tracking in arbitrary videos. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[19] D. Decarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Int. Journal of Computer Vision (IJCV)*, 38(2):99–127, 2000.

[20] I. Essa, S. Basu, T. Darrell, and A. Pentland. Modeling, tracking and interactive animation of faces and heads using input from video. In *Proceedings of Computer Animation'96*, pages 68–79, 1996.

[21] I. Essa, A. P. Pentland, et al. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):757–763, 1997.

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

[23] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(10):1296–1311, 2003.

[24] Z. Kalal, K. Mikolajczyk, and J. Matas. Face-tld: Tracking-learning-detection applied to faces. In *IEEE Proc. Conf. on Image Processing (ICIP)*, pages 3789–3792. IEEE, 2010.

[25] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[26] S. Koelstra, M. Pantic, and I. Y. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1940–1954, 2010.

[27] J. Kossaifi, G. Tzimiropoulos, and M. Pantic. Fast newton active appearance models. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1420–1424. IEEE, 2014.

[28] M. Köstinger, 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.

[29] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 368–373, 1995.

[30] 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.

[31] A. Li, M. Lin, Y. Wu, M.-H. Yang, and S. Yan. Nus-pro: A new visual tracking challenge. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*.

[32] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Euler principal component analysis. *Int. Journal of Computer Vision (IJCV)*, 101(3):498–518, 2013.

[33] S. Liwicki, S. Zafeiriou, G. Tzimiropoulos, and M. Pantic. Efficient online subspace learning with an indefinite kernel for visual tracking and recognition. *IEEE Trans. on Neural Networks and Learning Systems (TNNLS)*, 23(10):1624–1636, 2012.

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

[35] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):810–815, 2004.

[36] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[37] N. Oliver, A. P. Pentland, and F. Berard. Lafter: Lips and face real time tracker. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 123–129, 1997.

[38] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic. Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, 31(4):322–340, 2013.

[39] G. Papandreou and P. Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[40] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 97–102, 2004.

[41] G. Rajamanoharan and T. Cootes. Multi-view constrained local models for large head angle face tracking. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[42] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1692, 2014.

[43] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. Journal of Computer Vision (IJCV)*, 77(1-3):125–141, 2008.

[44] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. In *Image and Vision Computing*, 2015.

[45] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1789–1796. IEEE, 2014.

[46] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE Proc. Int. Conf. on Computer Vision Workshop (ICCV'W)*, pages 397–403, 2013.

[47] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[48] X. Shengtao, Y. Yan, Shuicheng, and A. Kassim. Facial landmark detection via progressive initialization. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[49] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: an experimental survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1442–1468, 2014.

[50] P. Snape, A. Roussos, Y. Panagakis, and S. Zafeiriou. Face flow. In *International Conference on Computer Vision (ICCV), 2015*, December.

[51] K. Sobottka and I. Pitas. Face localization and facial feature extraction based on shape and color information. In *Image Processing, 1996. Proceedings., International Conference on*, volume 3, pages 483–486. IEEE, 1996.

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[53] P. A. Tresadern, M. C. Ionita, and T. F. Cootes. Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision*, 96(3):280–289, 2012.

[54] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015.

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

[56] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 593–600. IEEE, 2013.

[57] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2014.

[58] M. Uricar and V. Franc. Real-time facial landmark tracking by tree-based deformable part model based detector. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[59] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[60] Y. Wu and Q. Ji. Shape augmented regression method for face alignment. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[61] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*.

[62] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+ 3d active appearance models. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 535–542, 2004.

[63] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[64] X. Xiong and F. De la Torre. Global supervised descent method. In *IEEE Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673, 2015.

[65] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *TPAMI*, 18(6):636–642, 1996.

[66] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *ICCV-W, First Facial Landmark Tracking in-the-Wild Challenge and Workshop*, 2015.

[67] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

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