Joint Face Detection and Alignment with a Deformable Hough Transform Model

John McDonagh and Georgios Tzimiropoulos(✉)

Computer Vision Laboratory, University of Nottingham, Nottingham, UK
yorgos.tzimiropoulos@nottingham.ac.uk

Abstract. We propose a method for joint face detection and alignment in unconstrained images and videos. Historically, these problems have been addressed disjointly in literature with the overall performance of the whole pipeline having been scantily assessed. We show that a pipeline built by combining state-of-the-art methods for both tasks produces unsatisfactory overall performance. To address this limitation, we propose an approach that addresses both tasks, which we call Deformable Hough Transform Model (DHTM). In particular, we make the following contributions: (a) Rather than scanning the image with discriminatively trained filters, we propose to employ cascaded regression in a *sliding window* fashion to fit a facial deformable model over the whole image/video. (b) We propose to capitalize on the large basin of attraction of cascaded regression to set up a Hough-Transform voting scheme for detecting faces and filtering out irrelevant background. (c) We report state-of-the-art performance on the most challenging and widely-used data sets for face detection, alignment and tracking.

Keywords: Face detection · Alignment · Tracking · Cascaded regression · Hough Transform

1 Introduction

From Viola and Jones [1] to Deformable Part Models [2–4] and from Active Appearance Models [5] to Cascaded Regression [6–9], face detection, alignment and tracking have all witnessed tremendous progress over the last years. Besides new methodologies, another notable development in the field has been the collection and annotation of large facial data sets captured in-the-wild [3,10–13], for which a number of newly developed methods have been shown to produce remarkable results.

Despite the progress in the field, the majority of prior work has disjointly considered the two problems: there is a large number of papers on face detection and perhaps even a larger number of papers on face alignment and tracking, but to the best of our knowledge there are only two papers [3,14] that study the combined problem of detection and alignment and no method that addresses and evaluates all three tasks jointly. However, for many subsequent, higher level

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 569–580, 2016.
DOI: 10.1007/978-3-319-48881-3_39

tasks, like face recognition, facial expression and attribute analysis, what matters is the *overall performance* in terms of accuracy in landmark localization. Notably, recent state-of-the-art methods for such tasks heavily rely on the accurate detection of landmarks (see for example [15,16]).

As we show hereafter, the overall performance in landmark localization accuracy might be unsatisfactory even by putting two recently proposed state-of-the-art methods (we used [4] for face detection and [9] for landmark localization) together. The reason for this is that face detection follows object detection in terms of measuring performance and, in particular, it uses the PASCAL VOC precision-recall protocol for object detection, thus requiring 50 % overlap between the ground truth and detected bounding boxes. As our results have shown, this accuracy is insufficient for initializing current landmark localization algorithms, even state-of-the-art methods like the one of [9] which is robust to poor initialization.

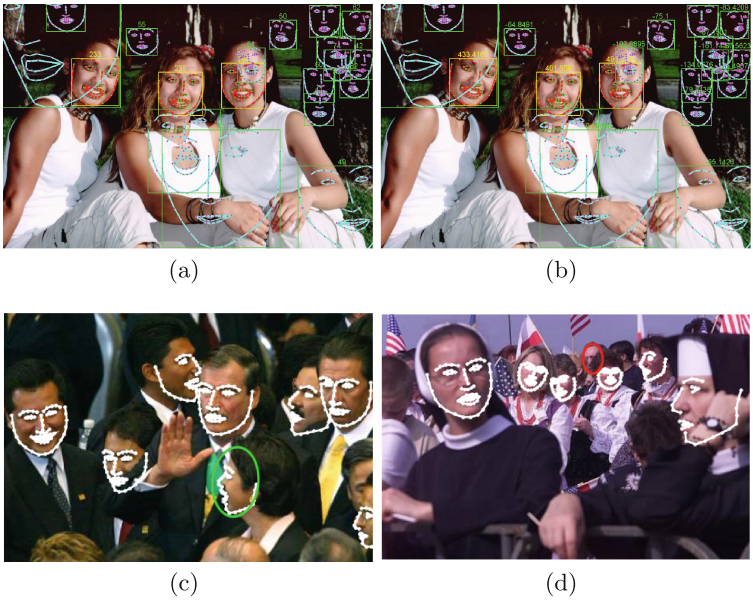

Fig. 1. (a), (b): Overview of DHTM. Our system scans an image in a sliding window fashion and for each candidate location fits a facial deformable model using PO-CR. Image locations that converge to the same location cast votes for that location in a Hough Transform fashion. Thresholding the voting surface and performing NMS results in a few candidate locations for which SVM scores are calculated by extracting SIFT and colour features. (a) System responses that received the highest number of votes. (b) Scores after applying SVM. (c, d) Output of our system on two challenging images from FDDB. The green ellipse shows a face that is not annotated. The red ellipse shows a missed face. (Color figure online)

1.1 Contributions

To address the aforementioned problem, we propose Deformable Hough Transform Model (DHTM). DHTM is largely motivated by the efficiency and robustness of cascaded regression methods for facial landmark localization. Essentially, rather than using a face detector to initialize them, we instead propose to employ them in order to jointly detect and track the location of faces and facial features in images/videos, too. Overall, our model jointly addresses face detection, alignment and tracking via scanning the image with Project-Out Cascaded Regression (PO-CR) [9] and aggregating the fitting results using a Hough-Transform (HT) voting scheme. In particular, we make the following contributions:

- Rather than exhaustively evaluating multiple templates as in [3,4] in order to cope with pose or other deformations, we propose to employ cascaded regression [6] in a *sliding window* fashion in order to evaluate the score of a *single* deformable template over a grid of image locations. For a deformable template, we choose one based on a parametric, densely connected shape model and an appearance model built from SIFT features. We fit this model using PO-CR [9], the complexity of which is only $O(nN)$ per iteration, where N is the number of features in the appearance model, and n is the number of parameters in the shape model.
- We propose to capitalize on the large basin of attraction of PO-CR and formulate a Hough-Transform voting scheme that filters out irrelevant objects and background areas, while at the same time "rewards" candidate image locations for which PO-CR converges to similar solutions. The main idea is that if the algorithm converges to the same solution for multiple initializations, then the converged solution "must" be a face.
- We report state-of-the-art results on challenging data sets for all 3 tasks: For face detection, DHTM is among the top performing methods on FDDB [10] and AFW [3] using the discrete score and sets a new state-of-the-art for the continuous score on FDDB. For face alignment, DHTM achieves state-of-the-art performance on the most challenging COFW [17] in terms of landmark localization error. For face tracking, DHTM achieves state-of-the-art performance on the 300-VW data set [18] in terms of landmark localization error.

2 Related Work

In this section, we review related work on face detection, alignment and performance measures.

Face detection. Face detection is one of the most popular and well-studied problems in computer vision with a multitude of approaches proposed over the last years reporting varying degrees of success. A comprehensive review of the topic is beyond the scope of this section, and we refer the reader to [4] for a recent survey. Interestingly, in the same paper, it is reported that a multi-channel, multi-view version of the Viola-Jones detector performs comparably

with a properly tuned vanilla Deformable Part Models (DPM) face detector, and that they both produce state-of-the-art performance on FDDB [10] and AFW [3] data sets. Hence, it is argued that part-based approaches are not always advantageous over standard approaches based on multi-view rigid templates, especially when a large amount of training data is available. A part-based approach to face detection is advocated in [3] and more recently in [14]. The Tree-Structure Model of [3] proposes a supervised way to train a DPM face detector based on manual annotations of parts, and a tree-based shape model that allows for a globally optimized model. An interesting extension of [3] that deals better with occlusion is described in [19]. The joint cascade detection and alignment algorithm of [14] proposes to use shape-indexed features for classification. [14] and [4] along with the more recent deep architectures of [20–22] are the state-of-the-art in face detection. Our work is similar to [3,14] in a sense that it produces the location of landmarks along with that of the face. However, both [3,14] are based on classification. In contrast, the main scoring scheme in the proposed DHTM is a novel voting scheme based on the large basin of attraction of cascaded regression that is used to cast votes for the location of candidate faces. A voting scheme for detecting faces is proposed in [23], however the voting is not based on deformable model fitting (as in our work), but on rigid image retrieval and is fundamentally different from the method presented herein. Finally, we note that although our method achieves state-of-the-art performance using standard SIFT and colour features, it could further benefit from region proposals and deep features as in [20–22].

Facial landmark localization. DHTM uses cascaded regression to fit a deformable template to each sub-window of a given image. Cascaded regression [6] is an iterative regression method in which the output of regression at iteration $k-1$ is used as input for iteration k, and each regressor uses image features that depend on the current pose estimate. DHTM is somewhat related to a number of regression-based face alignment methods [7–9,24–27] that have recently emerged as the state-of-the-art. Consensus methods for face alignment have been proposed in [17,28,29]. However, the aim of our work is not face alignment *given* a face detection initialization (as in all aforementioned algorithms) but joint face and facial landmark detection.

Performance measures. In face detection, performance is measured using the PASCAL VOC precision-recall protocol, requiring 50 % overlap between the ground truth and detected bounding boxes. In the FDDB benchmark, this is called "discrete" measure. FDDB also describes a "continuous" measure in which the detection score is weighted by the corresponding overlapping ratio. The continuous measure is thus more appropriate to reflect on the accuracy of the detected bounding box. The proposed DHTM has performance comparable to state-of-the-art when the discrete measure is considered and establishes a new state-of-the-art for the case of the continuous measure. In face alignment and tracking, performance is measured using the average normalized point-to-point (pt-pt) error between ground truth and detected landmarks. Performance strongly depends on the quality of initialization. DHTM produces state-of-the-art results when the joint problem of face and facial landmark

detection is considered, and performance is measured using the pt-pt error: DHTM largely outperforms the combination of [4] (for face detection) and [9] (for landmark localization).

3 Deformable Hough Transform Model

Our system scans an image in a sliding window fashion and for each candidate location $\mathbf{x}$, it fits a generative facial deformable model using PO-CR [9]. Image locations that converge to the same location cast votes for that location in a fashion similar to Hough Transform. Thresholding the surface of votes, obtained by our voting scheme, and performing non-maximal suppression, we end up with a few candidate locations per image. For these locations, multiple initializations are combined by taking the median and finally, SVM scores are calculated by extracting SIFT and colour features around the landmarks of each of the fitted shapes. Figure 1 aims to provide an overview of our system. The main components of the proposed Deformable Hough Transform Model (DHTM) are analyzed as follows.

3.1 Shape Model and Appearance

Shape model. DHTM uses cascaded regression to fit a deformable template to each sub-window of a given image. Our cascaded regression method of choice for this purpose is the recently proposed PO-CR [9] which has been shown to produce good fitting results for faces with large pose and expression variation. PO-CR uses parametric shape and appearance models both learned with PCA. Let us assume that we are given a set of training facial images $\mathbf{I}_i$ annotated with u fiducial points. For each image, the set of all points defines a vector $\in \mathcal{R}^{2u\times 1}$. The annotated shapes are firstly normalized by removing similarity transformations using Procrustes Analysis and the shape model is obtained by applying PCA on the normalized shapes. The model is defined by the mean shape $\mathbf{s}_0$ and n shape eigenvectors $\mathbf{s}_i$ represented as columns in $\mathbf{S} \in \mathcal{R}^{2u\times n}$. Finally, to model similarity transforms, $\mathbf{S}$ is appended with 4 additional bases [30]. Using this model a shape can be instantiated by:

$$\mathbf{s}(\mathbf{p}) = \mathbf{s}_0 + \mathbf{S}\mathbf{p}, \tag{1}$$

where $\mathbf{p} \in \mathcal{R}^{n\times 1}$ is the vector of the shape parameters.

Appearance. To represent appearance in facial images, an image is firstly warped to a reference frame so that similarity transformations are removed. Then, the local appearance around each landmark is encoded using SIFT [31] and all descriptors are stacked in a vector $\in \mathcal{R}^{N\times 1}$ which defines the part-based facial appearance. Finally, PCA is applied on all training facial images to obtain the appearance model defined by the mean appearance $\mathbf{A}_0$ and m appearance eigenvectors $\mathbf{A}_i$ represented as columns in $\mathbf{A} \in \mathcal{R}^{N\times m}$. Using this model a part-based facial representation can be instantiated by:

$$\mathbf{A}(\mathbf{c}) = \mathbf{A}_0 + \mathbf{A}\mathbf{c}, \tag{2}$$

where $\mathbf{c} \in \mathcal{R}^{m\times 1}$ is the vector of the appearance parameters.

3.2 Deformable Model Fitting with PO-CR

We assume that a sub-window of our original image contains a facial image. We also denote by $\mathbf{I}(\mathbf{s}(\mathbf{p})) \in \mathcal{R}^{N\times 1}$ the vector obtained by generating u landmarks from a shape instance $\mathbf{s}(\mathbf{p})$ and concatenating the SIFT descriptors for all landmarks. To localize the landmarks in the given sub-window, we fit the shape and appearance models (described in the previous section) by solving the following optimization problem:

$$\arg\min_{\mathbf{p},\mathbf{c}} ||\mathbf{I}(\mathbf{s}(\mathbf{p})) - \mathbf{A}(\mathbf{c})||^2. \tag{3}$$

As Eq. (3) is non-convex, a locally optimal solution can be readily provided in an iterative fashion using the Lucas-Kanade algorithm [30,32].

In particular, given an estimate of $\mathbf{p}$ and $\mathbf{c}$ at iteration k, linearisation of Eq. (3) is performed and updates, $\Delta\mathbf{p}, \Delta\mathbf{c}$ can be obtained in closed form. Notably, one can by-pass the calculation of $\Delta\mathbf{c}$ (for more details see [33]) by solving

$$\arg\min_{\mathbf{\Delta p}} ||\mathbf{I}(\mathbf{s}(\mathbf{p})) + \mathbf{J}_I\Delta\mathbf{p} - \mathbf{A}_0||^2_{\mathbf{P}}, \tag{4}$$

where $||\mathbf{x}||^2_{\mathbf{P}} = \mathbf{x}^T\mathbf{P}\mathbf{x}$ is the weighted ℓ_2-norm of a vector $\mathbf{x}$. The solution to the above problem is readily given by

$$\Delta\mathbf{p} = -\mathbf{H}_P^{-1}\mathbf{J}_P^T(\mathbf{I}(\mathbf{s}(\mathbf{p})) - \mathbf{A}_0), \tag{5}$$

where $\mathbf{J}_P = \mathbf{P}\mathbf{J}_I$ and $\mathbf{H}_P = \mathbf{J}_P^T\mathbf{J}_P$, $\mathbf{P} = \mathbf{E} - \mathbf{A}\mathbf{A}^T$ is a projection operator that projects out appearance variation from the image Jacobian $\mathbf{J}_I$, and $\mathbf{E}$ is the identity matrix.

Note that the above algorithm can be implemented in real-time for a single fitting, yet it is too slow to be employed for all sub-windows of a given image as the Jacobian, the Hessian and its inverse need to be re-computed per iteration. PO-CR by passes this computational burden by pre-computing a sequence of averaged projected-out Jacobians and Hessians (one per iteration) using regression. In particular, for iteration k, PO-CR pre-computes "averaged" matrices $\widehat{\mathbf{J}}_P(k)$, $\widehat{\mathbf{H}}_P(k) = \widehat{\mathbf{J}}_P(k)^T\widehat{\mathbf{J}}_P(k)$ and finally $\mathbf{R}(k) = \widehat{\mathbf{H}}_P(k)^{-1}\widehat{\mathbf{J}}_P(k)^T$. During testing, an update for iteration k can be obtained from $\Delta\mathbf{p}(k) = \mathbf{R}(k)(\mathbf{I}(\mathbf{s}(\mathbf{p}(k))) - \mathbf{A}_0)$ with cost $O(nN)$, only. Hence, fitting in PO-CR is very fast, with our parallel implementation running in a few thousand frames per second (one initialisation per frame).

3.3 Hough-Transform Voting

The proposed DHTM detects faces via a Hough-Transform voting scheme by capitalizing on the properties of the iterative optimization procedure employed by PO-CR. In particular, our system scans an image in a sliding window fashion and for each location $\mathbf{x}$ (we used a grid of equally spaced points, see Sect. 3.5), it fits our facial deformable model using the PO-CR described in the previous section. Voting in the proposed system is performed in a straightforward fashion. We simply

extract the translational component from $\mathbf{p}$ which represents the location of the fitted shape in the image. Then, for that location we cast a vote.

As with standard gradient descent fitting (PO-CR is a regression-based solution to Gauss-Newton optimization), we posit that when initialized in locations where no faces are present, PO-CR will converge to random locations/solutions. Examples of such cases are illustrated in Fig. 1 (a) as cyan "faces"'. The numbers in boxes indicate the number of times that the algorithm has converged to the nearby locations. As we may observe there are no more than 80 times that the algorithm converged to a similar solution. On the contrary, when initialized close to a face, because of the large basin of attraction of regression-based approaches, PO-CR is very likely to accurately recover both the face and its parts. Two examples of this idea are illustrated in Fig. 1(a) as red faces, with the numbers indicating that more than 150 times for both faces PO-CR has converged to the same solution. Thresholding the surface of votes, obtained by our voting scheme, and performing non-maximal suppression, our system removes most of the background clutter ending up with a few candidate locations per image. Finally, as our system is based on aggregating votes from different initializations, it comes naturally to consider how these initializations can be combined to produce a single fitting. We address this by simply taking the median of all fitted shapes that cast votes for the same peak in Hough space.

3.4 Final Re-scoring

Once the final fitted shape has been obtained, we perform re-scoring of the candidate face by evaluating an SVM trained on SIFT and colour features [4]. The overall detection process in DHTM is illustrated in Fig. 1(b).

3.5 Complexity and Implementation

Complexity. Assume that the PO-CR model has K levels of cascade. For each level, a regression matrix $\mathbf{R}(k)$ is learned having n regressors with N features each (columns of $\mathbf{R}(k)$). Recall that n is the number of parameters in our shape model. Hence, the complexity of fitting per sub-window is $O(K(nN))$. Because of the large basin of attraction of PO-CR, we perform fitting only on a grid of equally spaced points using a stride of 10 pixels. If there are L locations per image to perform fitting, the total complexity is $O(LK(nN))$ for a single level of the image pyramid. By making an analogy between the regressors in $\mathbf{R}(k)$ and the number of mixtures in [3] (the number of regressors ($n = 15 - 20$) is indeed similar to the number of mixtures in [3]), and assuming that [3] is also evaluated on L locations, we conclude that our model is slower than [3] only by a factor of K. However, L is smaller in DHTM because PO-CR optimizes for translation too, having very large basin of attraction. Additionally, by optimizing at the first level of the cascade only for scale, rotation and translation, and then casting votes in Hough space (as explained in the previous section), our method largely filters out most of the irrelevant background in the image leaving very few locations to evaluate in the subsequent levels of the cascade. Hence, in practice,

the total complexity is $O(LK(nN))$ with $K = 1$ and $n = 4$. For a VGA image, our parallel, but not entirely optimized implementation, runs at 1–2 Hz [1]. Note that we can readily attain much higher speeds, by applying any object/face proposal techniques to reduce the number of evaluations per image.

Training. Training in DHTM is very simple and includes learning $\mathbf{R}(k)$ as described in Sect. 3.2, and learning the SVM model for face re-scoring as described in Sect. 3.4. To learn $\mathbf{R}(k)$, we used the available landmark annotations of the 300-W challenge [13]. Our PO-CR model built from this data set is able to fit images with large yaw variation ($\pm 60°$) but not entirely profile images (yaw $\approx 90°$). Hence, we annotated more than 1000 profile images from the ALFW dataset and the internet, which we will make publicly available. For training the SVM model, we fitted our PO-CR approach to our training sets and used the fitted shapes as positive examples. This resulted in more realistic positive examples than using the ground truth shapes. Finally, negative examples were obtained by scanning background images and then recording all locations for which the number of votes obtained by HT voting was greater than 40.

4 Results

We report results on three tasks namely face detection, face detection followed by face alignment and face tracking.

4.1 Face Detection Experiments

To evaluate the performance of our method on face detection, two of the most popular in-the-wild datasets were used, namely AFW [3] and FDDB [10]. AFW is built from Flickr images. It consists of 205 images with a total of 474 annotated faces [4]. The images within this dataset tend to contain cluttered background and faces with large variations in both viewpoint and appearance. FDDB consists of 2845 images, with a total of 5171 ellipse face annotations. This dataset includes very challenging low resolution, out of focus and occluded faces. To report face detection performance, we generated the familiar precision-recall curve using the standard PASCAL protocol. In particular, faces are only considered detected if the intersection-over-union (IoU) ratio between the ground truth and the detected bounding box exceeds 50 %. For FDDB, we also report the value of IoU, known as "continuous score". In addition to the performance of DHTM, we report the performance of the top performing methods for each dataset.

Figure 2 summarizes our results on AFW. We compare with the methods recently reported in [4]. When the IoU overlap is set to 50 %, our detector is

[1] All tests were done using a NVIDIA GeForce GTX 980 GPU and an Intel I7–4790 k CPU, on a PC running Windows 8.1 64-bit with 16 GB of RAM. The proposed system was compiled for GPU devices of compute capability 3.5 and above, using the CUDA 7.0 development toolkit.

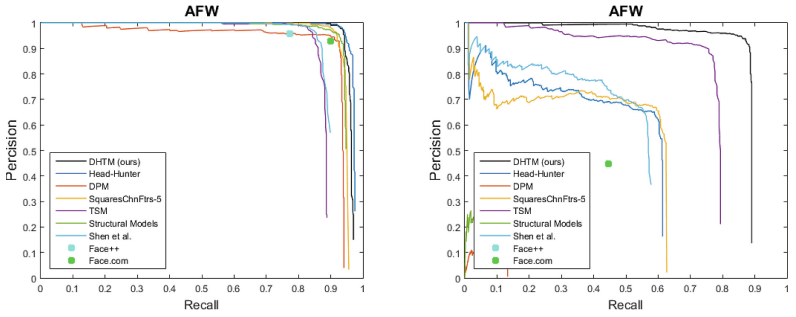

Fig. 2. Precision recall for AFW. (a) IoU ratio is set to the standard 50 %. (b) IoU ratio is set to 75 %.

comparable to both current commercial and published state-of-the-art methods. To further show the accuracy of our proposed detector, we increased the IoU overlap to 75 %, and as can be seen in Fig. 2(b), our detector clearly outperforms all commercial and published methods by a margin of over 10 % in detection accuracy.

Figure 3 summarizes our results on FDDB. We compare the performance of our proposed detector against the currently published state-of-the-art methods of [3,4,14,20–22,34,35]. For discrete scores, as shown in Fig. 3(a), our system is one of the top performing methods being outperformed only by [20–22]. All three methods are based on deep learning features. We have found that although PO-CR can fit some very difficult faces, the weakest component of our system is the SVM based on SIFT/colour features which for such faces yields low scores. Hence by incorporating deep learning features into our system, one can expect much better performance (this is left for future work). Notably, our system is the top performing method when using the continuous score, outperforming all [20–22] by a large margin.

4.2 Face Alignment and Tracking Experiments

For this experiment, we show localization performance of the *complete* DHTM system including face detection followed by facial feature localization. To measure landmark localization performance, we used the point-to-point Euclidean distance (pt-pt error) normalized by face size and report the cumulative curve corresponding to the fraction of images for which the error was less than a specific value [3]. We report performance on two very challenging datasets.

The first data set is COFW [17]. We chose this data set as it contains images with large amounts of occlusion. This not only affects face detection performance but also precise face localization which in turn affects facial feature localization accuracy. For comparison, we also report the performance of the combined system HeadHunter [4] followed by PO-CR. Figure 4(a) shows our results. Clearly, DHTM outperforms HeadHunter plus PO-CR by a large margin.

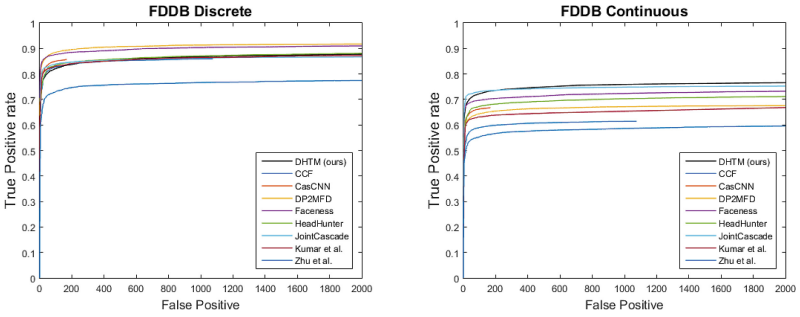

Fig. 3. Performance curves for FDDB. (a) Discrete score. (b) Continuous score.

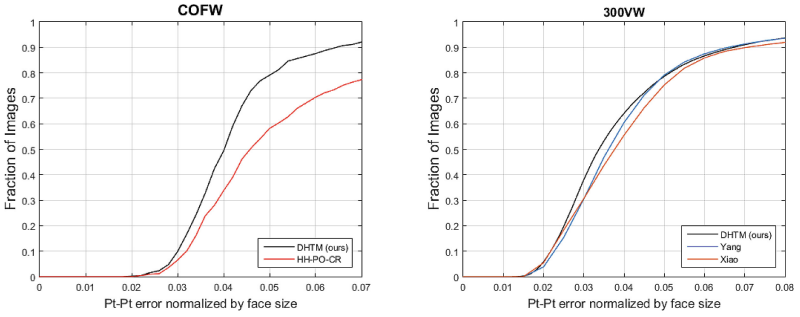

Fig. 4. Point-to-point error, relative to face size, for (a) COFW data set and (b) 300-VW (Category C) data set.

The second data set is the 300 videos in-the-wild (300-VW) data set recently released in [18]. We chose category C to report performance on as it is the most difficult category. This category contains 14 videos and more than 20,000 frames, therefore this is a very large scale experiment. Face localization in video is considered easier than in still images as one can exploit temporal coherency to improve performance, and indeed the top performing methods [22,36] do so. Instead, we considered each frame as a separate image and run our system to simultaneously detect the face and localize the landmarks. As Fig. 4(b) shows, even this case our system is outperforming the winners of the 300-VW competition.

5 Conclusions

We proposed a novel approach to face detection and landmark localization which we call Deformable Hough-Transform Model (DHTM). Our approach is largely motivated by the efficiency and robustness of recent cascaded regression approaches to facial landmark localization; essentially, rather than using a face

detector to initialize them, we instead propose to employ them in order to detect the location of faces in an image too. Rather than scanning the image with discriminatively trained filters, we propose to employ the PO-CR algorithm in a *sliding window* fashion to fit a facial deformable model and capitalize on the large basin of attraction of PO-CR to set up a Hough-Transform voting scheme. We report comparable performance to that of state-of-the-art face detection algorithms and significant improvement over the standard face detection/landmark localization pipeline when performance is measured in terms of landmark localization.

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE TPAMI **32**(9), 1627–1645 (2010)
3. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark estimation in the wild. In: CVPR (2012)
4. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 720–735. Springer, Heidelberg (2014)
5. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. TPAMI **23**(6), 681–685 (2001)
6. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR (2010)
7. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)
8. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR (2013)
9. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: CVPR (2015)
10. Jain, V., Learned-Miller, E.G.: FDDB: a benchmark for face detection in unconstrained settings. UMass Amherst Technical Report (2010)
11. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR (2011)
12. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012)
13. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: CVPR-W (2013)
14. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 109–122. Springer, Heidelberg (2014)
15. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: CVPR (2013)
16. Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., Matthews, I., Sridharan, S.: In the pursuit of effective affective computing: the relationship between features and registration. IEEE SMC-B **42**(4), 1006–1016 (2012)

17. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV (2013)
18. Shen, J., Zafeiriou, S., Chrysos, G., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: The first facial landmark tracking in-the-wild challenge: benchmark and results. In: ICCV-W (2015)
19. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: detecting and localizing occluded faces. In: CVPR (2014)
20. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: CVPR (2015)
21. Ranjan, R., Patel, V.M., Chellappa, R.: A deep pyramid deformable part model for face detection. arXiv preprint (2015). arXiv:1508.04389
22. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: a deep learning approach. In: ICCV (2015)
23. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Detecting and aligning faces by image retrieval. In: CVPR (2013)
24. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR (2013)
25. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 FPS via regressing local binary features. In: CVPR (2014)
26. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: CVPR (2014)
27. Kazemi, V., Josephine, S.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)
28. Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 278–291. Springer, Heidelberg (2012)
29. Yu, X., Lin, Z., Brandt, J., Metaxas, D.N.: Consensus of regression for occlusion-robust facial feature localization. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 105–118. Springer, Heidelberg (2014)
30. Matthews, I., Baker, S.: Active appearance models revisited. IJCV **60**(2), 135–164 (2004)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
32. Baker, S., Matthews, I.: Lucas-kanade 20 years on: a unifying framework. IJCV **56**(3), 221–255 (2004)
33. Tzimiropoulos, G., Pantic, M.: Gauss-Newton deformable part models for face alignment in-the-wild. In: CVPR (2014)
34. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: ICCV (2015)
35. Kumar, V., Namboodiri, A., Jawahar, C.: Visual phrases for exemplar face detection. In: ICCV (2015)
36. Xiao, S., Yan, S., Kassim, A.: Facial landmark detection via progressive initialization. In: ICCV-W (2015)