Face Sketch Landmarks Localization in the Wild

Heng Yang, *Student Member, IEEE*, Changqing Zou, and Ioannis Patras, *Senior Member, IEEE*

***Abstract*—In this letter, we propose a method for facial landmarks localization in face sketch images. As recent approaches and the corresponding datasets are designed for ordinary face photos, the performance of such models drop significantly when they are applied on face sketch images. We first propose a scheme to synthesize face sketches from face photos based on random-forests edge detection and local face region enhancement. Then we jointly train a Cascaded Pose Regression based method for facial landmarks localization for both face photos and sketches. We build an evaluation dataset, called Face Sketches in the Wild (FSW), with 450 face sketch images collected from the Internet and with the manual annotation of 68 facial landmark locations on each face sketch. The proposed multi-modality facial landmark localization method shows competitive performance on both face sketch images (the FSW dataset) and face photo images (the Labeled Face Parts in the Wild dataset), despite the fact that we do not use extra annotation of face sketches for model building.**

***Index Terms*—Cascaded pose regression, face sketch, facial landmark localization.**

I. Introduction

FACE sketches are frequently used as a means of visual representation of an individual's face. Such representation has been applied for digital entertainment like cartoon synthesis [20], [14], facial expression recognition [8], face retrieval [7] and face recognition in law enforcement [15], [22]. In the latter case, the photo of a suspect is not available and the face sketch is drawn based on the information collected from the witnesses. Taking the sketch retrieval and photo-to-sketch face recognition as an example, the challenge of using sketch representation mainly lies in the modality difference between the sketch and the photo. Several approaches [12], [15], [22], [7], [14] focus on bridging the gap of the two modalities. Similar to photo-to-photo face recognition, it is crucial to align the face sketch first into a canonical pose, where the face pose is always represented by a set of facial landmarks.

In recent years, facial landmarks localization (or face alignment) has made a significant progress on face images in the wild, using the holistic pose regression methods [4], [3], [16], [13], [17], [18], or local based methods [1], [11], [19], [23]. However, due to the modality difference, the performance drops significantly on face sketches. In this letter, we address this problem, in order to make applications like sketch-to-photo face recognition and face sketch retrieve more practical in real world.

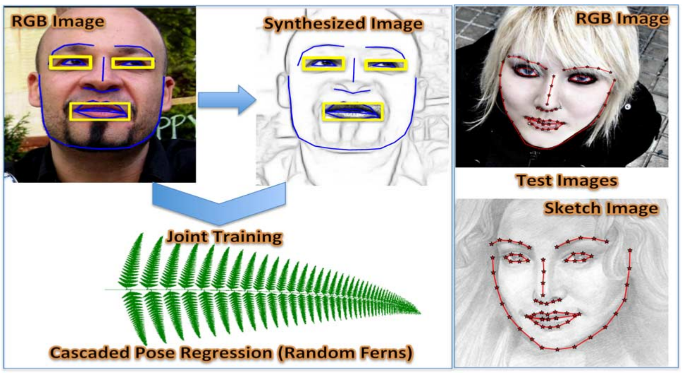

Fig. 1. Our approach trains a Cascaded Pose Regression model based on RGB face images and their synthesis (left), then estimates the facial landmarks locations in both face photos and face sketch images (right).

Only a few face sketch datasets are currently available. In most of them, like the CUFSF [15] and CUFS [22], the sketches are drawn by artists based on original face photos. Some sketches, like that in CUFSF are with shape exaggeration. These sketch images are not as challenging as those from the real world in two aspects: first, the original photos which the sketches synthesized from were taken from constrained frontal poses [15], [22] while the sketches in real world might be in arbitrary poses; second, the high quality in terms of facial details of the sketches in those datasets is difficult to be obtained in real world application. Due to these limitation of these datasets, it is difficult to train and to evaluate a general alignment model for face sketches. In order to deal with this, we propose to train a model for multi-modality facial landmark localization, by making full use of the publicly available face photo datasets collected in the wild with landmarks annotations. More specifically, we automatically generate sketches from images in those datasets by fusing local region enhancement and edge detection using structured random forests. We then train a Cascaded Pose Regression based on both the face photos and face sketches using the ground truth landmark annotation. The proposed method is illustrated in Fig. 1.

In order to evaluate the performance of the proposed method, we collect face sketches from the Internet and create the Face Sketches in the Wild (**FSW**[1]) dataset. We compare our method with the current state-of-the art facial landmarks localization methods. We achieve almost the same results to the Robust Cascaded Pose Regression [3] method trained on RGB images on

Manuscript received May 16, 2014; accepted June 23, 2014. Date of publication June 26, 2014; date of current version July 09, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen.

H. Yang and I. Patras are with the School of Electrical Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: heng.yang@qmul.ac.uk; i.patras@qmul.ac.uk).

C. Zou is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the Department of Physics and Electronic Information Science, Hengyang Normal University, Hengyang 421008, China (e-mail: aaronzou1125@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2333544

[1] Available at: https://sites.google.com/site/yanghengcv

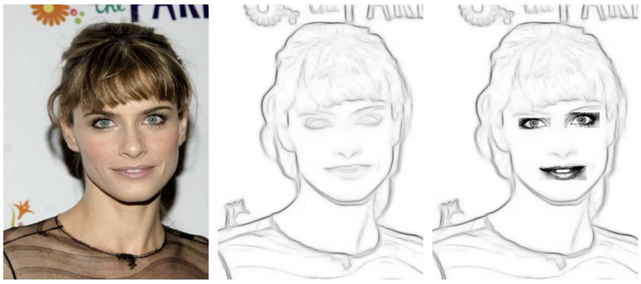

Fig. 2. An example image of face sketch synthesis. From left to right are the original RGB image, edge detection by [6] and our synthesized sketch image. In the synthesized image the eye regions and mouth region are enhanced and fused with the edge detection.

the face photo dataset and the best performance on the FSW dataset.

In summary, this letter 1) proposes a facial landmarks localization method for both face sketches and face photos showing competitive performance; 2) introduces a dataset with 450 face sketches collected in the wild with 68 facial landmarks annotation that can be used for future face sketch landmarks localization evaluation.

II. Method

In this section, we first present how we augment the training samples by synthesizing face sketches from face photos in Section II-A. Then we describe the model learning in Section II-B.

A. Face Sketch Synthesis

Most of the current face sketch synthesis approaches follow a supervised learning route, for instance the Markov Random Fields (MRF) in [15] thus they require a large number of *ground truth* face sketches that are often drawn by artists. It is quite expensive to acquire such training samples since face alignment model training often demands thousands of training instances. Moreover, using only drawing made by the artists limits the diversity of the face sketches. In applications such as face sketch retrieval, the sketch might be drawn by non-experts, that are key different from the drawing drawn by artists.

As opposed to a supervised learning synthesis, our face sketch synthesis scheme is based on fast edge detection using structured forests [6]. Note that though this edge detection method is learning based, it is trained for general edge detection. We define it as a non-learning based method because it is not necessary or desired to have face sketches at the training phrase. More specifically, we assume we are given a set of face photos $\mathcal{I} = \{I\}$, where for each I we have its annotation of facial landmarks locations. We denote the structured forests based edge detector by $\mathcal{F}$, thus given an image photo I, the edge detection result of I is:

$$I^e = \mathcal{F}(I). \tag{1}$$

I^e, as shown in Fig. 2 contains global shape information like the contour of the face but lacks details in local regions such as the eye shapes and mouth lips. However, the eye and mouth regions are important features on face thus they are often depicted in detail in sketch images. In order to synthesize the sketch with more details in these regions, we use their enhanced gray scale images. More specifically, we extract the rectangles around the two eye regions and the mouth region, based on the ground truth locations of their boundary landmarks. After converting the RGB image patches into gray scale, we further apply histogram equalization to increase the global contrast. Then the synthesized sketch is represented by:

$$I^s = I^e \oplus (I_{leye} \cup I_{reye} \cup I_{mouth}) \tag{2}$$

where I_{leye}, I_{reye} and I_{mouth} are the enhanced gray scale images from left eye region, right eye region and mouth region, respectively. The operator $\oplus$ works in a way of putting the layer of the right side on top of the layer of the left side, i.e. to replace the content of I^s at the corresponding pixels with the enhanced gray scale images. An example image is shown in Fig. 2. Though this procedure is very simple, the result looks very similar to sketch images. Its effectiveness in improving the landmark localization performance on sketch images will be demonstrated in the experimental section.

B. Joint Training of Cascaded Pose Regression

We use the Cascaded Pose Regression (CPR) [5] framework in this work given its efficiency and accurate performance for estimating face landmark locations [4], [3]. We follow the main steps of CPR evaluation procedure. A CPR consists of a cascade of T regressors $R^{1\cdots T}$. An estimation of a shape starts from a initial guess S^0, and progressively refine the estimation by an update in each iteration, until the final stage of regression is applied. As demonstrated in Algorithm 1, given the estimation of pose in the previous iteration S^{t-1}, image feature for the t-th iteration are calculated as $f^t = h^t(I, S^{t-1})$. Based on the feature f^t and the regressor R^t, an update ΔS is calculated, once is added on the previous estimation of the pose. Similar to [4] and [3], we use two stages of regression, i.e. at each iteration, multiple regressors are utilized and they share the same pose for feature calculation that is from the previous iteration. We also use the random fern as the primitive regressor and follow their training scheme that directly minimizes the alignment error. We use the interpolated shape-indexed features proposed in [3]. The latter uses a reference location between the locations of two landmarks thus is more robust against large pose variations and shape deformations.

Algorithm 1 Cascaded Pose Regression

```
Input: Image $I$, initial pose $S^0$, regressors $R^{1\cdots T}$
Output: Estimated pose $S^T$
for $t = 1$ to $T$ do
    $f^t = h^t(I, S^{t-1})$    ▷ Shape-indexed features
    $\Delta S = R^t(f^t)$    ▷ Apply regressor $R^t$
    $S^t = S^{t-1} + \Delta S$    ▷ update pose
end for
```

As discussed before, we assume we have a dataset with face photo images and their facial landmarks annotation

$\{(I_i, \hat{S}_i)\}_{i=1}^N$, where $\hat{S}_i$ is the vector of ground truth landmark locations. For each face photo I_i, we will generate a sketch synthesis as discussed in Eq. (2). Thus we have an additional set of training samples $\{(I_i^s, \hat{S}_i)\}_{i=1}^N$, based on the assumption that the synthesized face sketch image shares the same facial landmark annotation with the face photo. Similar to [4], [3], [5], we augment the training samples by initializing them with several random poses from other training samples. Like [4], each regressor is learnt by explicitly minimizing the sum of alignment errors. We adapt it by putting different weight on the error of sketch images and face photos, that is,

$$R^t = \arg\min_R \left(\alpha E_t(R) + (1-\alpha) E_t^s(R)\right) \tag{3}$$

where $E_t(R) = \sum_{i=1}^N \|\hat{S}_i - R(I_i, S_i^{t-1})\|$ is the sum of errors calculated over the face photo samples and $E_t^s(R) = \sum_{i=1}^N \|\hat{S}_i - R(I_i^s, S_i^{s,t-1})\|$ is the sum of errors calculated over the sketch samples. S_i^{t-1} is the shape of the i-th face photo sample estimated by the $t-1$ iteration and $S_i^{s,t-1}$ is that for the face sketch sample. By setting the values of α, we can adjust the relatively importance of face photos and face sketch images at the training stage. We note that, this parameter is not used during the testing stage once the regressor R^t is found. In this way, we can train the cascade of the regressors jointly for both face photos and face sketch images and the testing procedure is as described in Algorithm 1.

III. Experiment

A. Dataset and Implementation Details

We train our model using the training images of HELEN, a dataset that is widely used for evaluating facial landmarks localization in the wild. HELEN consists of 2510 training images and 330 test images, that are collected from the Internet, from search engine results or from Flickr. Most of those images exhibit a very large variability in pose, lighting, expression as well as general imaging conditions. Many images exhibit partial occlusions that are caused by head pose, objects (e.g., glasses, scarf, food), body parts (hair, hands) and shadows. We use the facial landmark annotations provided by the iBug challenge [10] for the following reasons: 1) most of the recent methods in facial landmark localization use the 68 facial landmark mark-up from Multi-PIE [9]; 2) it is a good benchmark and makes future comparisons more direct. We use the 2510 training images to build our model.

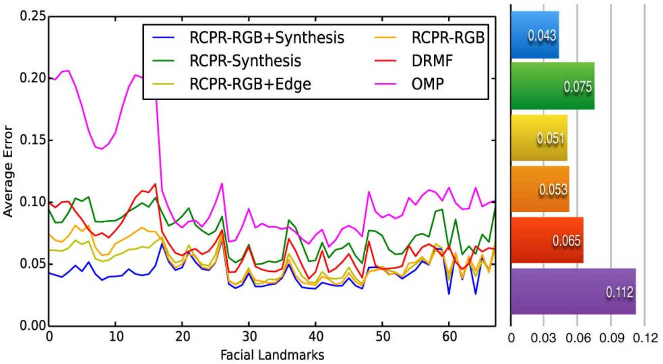

Fig. 3. Results on the FSW dataset. The left shows the landmark-wise average error. The right shows the overall mean error. The landmark ID number definition please refer to [10]. Roughly, from #1 to #7 are landmarks along the face contour while the remaining are inner facial landmarks. For DRMF and OMP method, the inner mouth corners are not detected and their errors are shown as the mean value of all the landmarks.

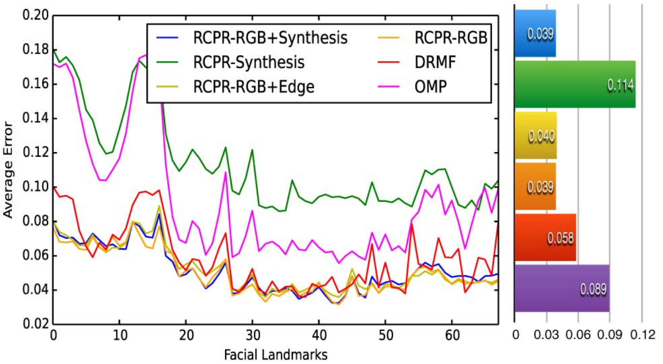

Fig. 4. Results on the test images (RGB) of LFPW dataset. The figure configuration is the same to Fig. 3.

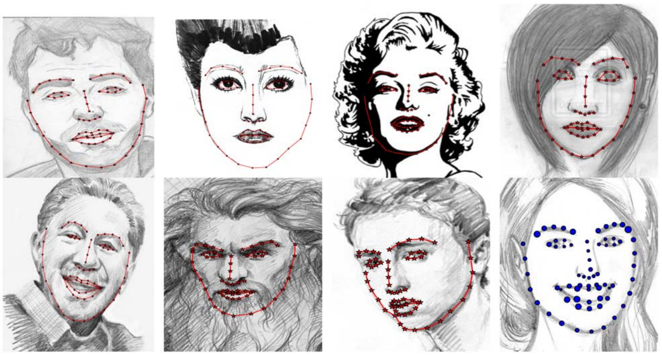

Fig. 5. FSW example results. Face sketch images in FSW show large variety of head pose and drawing styles. The last image in the second row shows the average FSW individual landmark error levels, represented by the point sizes.

The currently available face sketch datasets like CUFSF and CUFS are drawn by artists based on face images taken in very constrained environments and they exhibit very limited variability in terms of head poses, facial expressions and occlusions. Therefore, we produced a new and significantly more challenging dataset for evaluation, which we call Face Sketches in the Wild (FSW). We collect face sketch images from the Internet by searching using Google and Bing. The dataset is designed to present face sketches in real-world conditions. The sketches exhibit large head pose changes, resolution variability, occlusions and more importantly, different sketch styles. Some example images are shown in Fig. 5. We finally got 450 images for evaluation by excluding some non-face sketches such as the exaggerated cartoon face sketches. One face is detected in each sketch image by Viola-Jones face detector, followed by manual checking. Then we manually annotate the locations of 68 facial landmarks, with the mark-up used in Multi-PIE [9] and [10]. Note that we only use these images for evaluation but not for building the model.

Our implementation of the proposed method is based on the Robust Cascaded Pose Regression (RCPR) code provided by [3]. We use their default parameter setting, i.e., 50 boosted ferns at each iteration, 100 iterations in total, the number of features $F = 400$, depth 5 random ferns. When training the baseline method, the data augmentation factor is 20, i.e. 20 random initializations are used for each training sample. For our joint training, we set the data augmentation factor to 10 for a fair comparison since we double the number of traning examples by using the synthesized face sketches. We set the parameters $\alpha = 0.4$ in Eq. (3) by cross validation. We re-write the code using C++ on a standard 3.30 GHz CPU machine in order to get faster performance. An online demo is available on the FSW dataset web page where the user can upload images for testing. The face detection is carried out by OpenCV Viola-Jones face detector.

For better comparison, we also consider some other methods that are able to detect both inner landmarks and contour landmarks using the same mark-up of Multi-PIE. We consider two recent representative local based methods: the Discriminative Response Map Fitting (**DRMF**) in [1] and the Optimized Part Mixtures model (**OPM**) in [21]. For DRMF we run its model with given face detections. For OPM, which combines face detection and landmarks localization, we manually remove the false face detections when calculating the errors. This actually favours it since the face detection failure cases are often challenging images.

We report the error, i.e., the Euclidean distance between the ground truth location and the estimation, as a fraction of the inter-ocular distance, similar to [3], [4].

B. Results

1) Results on FSW: First we evaluate the performance of facial landmarks localization in sketch images, since this is the main aim of this letter, on the FSW dataset. We benchmark our method on the RCPR framework with the interpolated indexed feature in [3]. We do not use the full version since its training requires landmark visibility annotation. We do not use the re-start scheme of RCPR for a fair comparison since the re-start might vary from one to another and is also time-consuming. Different versions of such RCPR are trained including 1) **RCPR-RGB**, trained only on RGB face photo images (in practice, the RGB images are converted to gray scale images for model training); 2) $\mathbf{RCPR-RGB+Edge}$, jointly trained on RGB face photo images and the corresponding face edge images detected by [6]; 3) **RCPR-Synthesis** trained on the synthesized images only; 4) $\mathbf{RCPR-RGB+Synthesis}$, jointly trained on the RGB images and the corresponding synthesized images. We also compare to other facial landmarks localization method, that are trained for face landmarks localization for face photo images.

We report the average landmark-wise error of all the 68 facial landmarks of the test images of FSW, shown in Fig. 3. On this challenging dataset, the proposed method, $\mathrm{RCPR-RGB+Synthesis}$ significantly outperforms the others, both variations of the RCPR and the two local based models. The model learned using only the synthesized images for training (RCPR-Synthesis) has the worst performance among all RCPR variations. When comparing the results of $\mathrm{RCPR-RGB+Edge}$ to RCPR-RGB, we can observe the improvement for the landmarks along the contour but the performance drops for inner landmarks. This is very likely because the edge images captures very similar information to the face sketches along the contour but not the detail of the face inner parts. The local based methods, particularly the OMP, that were trained on RGB images, do not work well on the face sketch images, due to the change of the modality. The superior performance of $\mathrm{RCPR-RGB+Synthesis}$ over both the RCPR-RGB and $\mathrm{RCPR-RGB+Edge}$ validates the effectiveness of our proposed joint training scheme by using the RGB and synthesized images. It is worthy noting that, the improvement on the contour landmarks, which are generally regarded as more difficult parts, is more significant. We visualize the individual landmark error levels in the last image of Fig. 5, from which we can observe the high localization accuracy of most of the facial landmarks.

2) Results on LFPW Test Images: We also evaluated the generality of the proposed method by evaluating the facial landmarks localization accuracy on RGB images. We report the performance on the LFPW, a test set which is wildly used for evaluating facial landmarks localization in the wild [2], [23], [4], [3], [16]. The image in LFPW dataset which has much lower resolution than the HELEN dataset. The experiment is set in this way in a scenario the methods are trained on datasets different from the ones on which they are tested for fair comparison. The two local based methods, DRMF and OMP are trained on the Multi-PIE dataset while our RCPR variants are trained on the HELEN training images, all are tested on LFPW. We hereby note that the number of training instances in Multi-PIE is much larger than that of the HELEN training set. The result is shown is in Fig. 4. On RGB images, our proposed method performs on par with the other RCPR variants except the RCPR-Synthesis, which performed the worst on RGB images since it is only trained on synthesis images. All methods except RCPR-Synthesis perform better on RGB images than on the sketch images.

Though it is difficult for us to make exact comparison of the two modalities, we can observe our proposed method, and the DRMF method perform more consistently. However, the DRMF fails to achieve a high accuracy compared to our proposed method (6.5% vs. 4.3% on FSW and 6.8% vs. 3.9% on LFPW). For a conclusion, our proposed model, that is jointly trained on RGB images and their sketch synthesis, consistently performs better or very similar to the RCPR variants and the recent face landmark localization methods.

IV. Conclusion

In this letter, we propose a method for facial landmarks localization in 2D images of different modalities: face photos and face sketches. Based on the Cascaded Pose Regression framework, our model is jointly trained on both RGB images and synthesized sketch images, directly derived from the RGB images. The proposed method performs on par with the other RCPR variants and better than the other recent methods on RGB images. It shows significantly better results on sketch images from FSW dataset, collected in the wild, despite the fact that the model training is only based on the face photos and their synthesized sketches.

REFERENCES

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Conf. Computer Vision*, 2013.

[4] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[5] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[6] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2013.

[7] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, 2012.

[8] Y. Gao, M. K. Leung, S. C. Hui, and M. W. Tananda, "Facial expression recognition from line-based caricatures," *IEEE Trans. Syst., Man, Cybern. A: Syst. Humans*, vol. 33, no. 3, pp. 407–412, 2003.

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[10] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Computer Vision (300-W workshop)*, 2013.

[11] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang, "Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[12] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, 2004.

[13] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *Proc. IEEE Int. Conf. Computer Vision*, 2013.

[14] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.

[15] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, 2009.

[16] X. Xiong and F. De la Torre, "Supervised descent method, and its applications to face alignment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[17] H. Yang and I. Patras, "Face parts localization using structured-output regression forests," in *Proc. Asian Conf. Computer Vision*, 2012.

[18] H. Yang and I. Patras, "Privileged information-based conditional regression forests for facial feature detection," in *IEEE Conf. Automatic Face and Gesture Recognition*, 2013.

[19] H. Yang and I. Patras, "Sieving regression forests votes for facial feature detection in the wild," in *Proc. Int. Conf. Computer Vision*, 2013.

[20] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, 2012.

[21] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures, and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Computer Vision*, 2013.

[22] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 513–520.

[23] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Int. Conf. Computer Vision*, 2013.