Supervised Descent Method based on Appearance and Shape for Face Alignment

Yi Cheng
Institute of Automation Chinese Academy of Sciences, Beijing, China 100190
chengyi2014@ia.ac.cn

***Abstract*—Regression approaches have been recently shown to achieve state-of-the-art performance for face alignment. As a general optimization problem, face alignment is approximately solved by learning a series of mapping functions from local appearance to the coordinates increment of the pixels to detect. There have been extensive studies and continuous improvements have been made in recent years. However, most of the existing methods only rely on the current facial texture in every iteration. It is unreliable to only rely on local appearance information when facial landmarks are partially occluded in unconstrained scenarios. In this paper, a modified supervised descent method is proposed to settle the issue, utilizing both appearance and shape information in learning regression functions. Hence, we call it asSDM. The major contribution of our proposed method is to jointly capture shape and local appearance in cascade regression framework. We evaluate the performance of the proposed method on different data sets and the experimental results on benchmark databases demonstrate that our proposed method outperforms previous work for facial landmark detection.**

I. Introduction

In the field of computer vision, face alignment is among the most popular and well-studied problem. The purpose of face alignment is to detect facial key points on the facial images with large variations on face expression, head pose, illumination, and partial occlusions. The annotation models of each data set may consist of different number of facial key points. The information of the key points is essential for tasks like face recognition, head pose estimation, facial expression analysis and 3D face modeling. Over the past few decades, there have been extensive improvements of the face alignment algorithms. However, when it comes to the applications that require high precision and stability in unconstrained environments, face alignment is still a challenging task for current approaches.

Many existing methods for face alignment pose the problem as an optimization one. Alignment is achieved by finding the parameters that minimize the error function. A well-defined alignment error function can lead to a fast convergence rate and good performance of the method. In most of previous researches, the error function only considers about the appearance information. For example, the Active Appearance Model (AAM) [1] searches by using the texture residual between the model and the target image to predict improved model parameters to obtain the best possible match. And the Supervised Descent Method (SDM) [2]learns a sequence of regression functions that iteratively map images appearance to the target output.

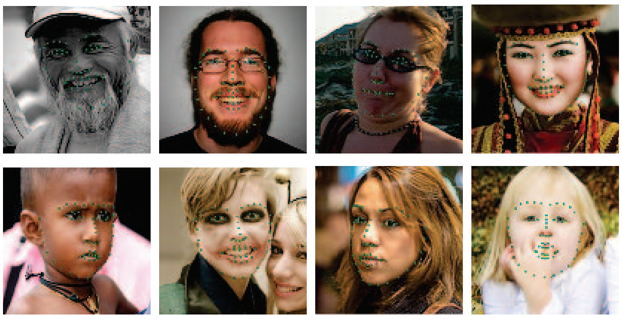

Fig. 1: Images are from the Helen database with 68 manually labeled landmarks.

In unconstraint conditions, however, the appearance features have limited expressive power to capture complex and subtle face image variations in pose, expression, illumination and occlusion. Specifically, the eyes may be occluded by glasses, sunglasses, or hair. And part of the face may be occluded by a hat, cigarette or hand. Besides, there may be heavy shadowing and the face may be made up theatrically. We notice that eye centers, eye corners and mouth corner are salient landmarks with special texture characteristics. However, points along face contour are significantly different from these salient landmarks, which need location information of other key points. To settle these issues, an elegant method combining appearance information and shape constraint is proposed in this work.

The rest of the paper is organized as follows. In section II, we review the related work. In section III, the proposed method is introduced. In section IV, we describe the data sets used in the experiment. And performance of the proposed method are shown. And the conclusions of the paper are given in section V.

II. Related Work

In the early research, AAM [3], first proposed in [4] , and the closely related concepts of Morphable Model and Active Blob represent the non-linear, generative, and parametric models. These models are direct optimization approaches that match shape and texture simultaneously, by learning correlation between errors in model parameters and the resulting residual

978-1-5090-2927-3/16/$31.00 ©2016 IEEE

texture errors. They use pattern of intensities or colors across an image patch as appearance features.

In the Constrained Local Method (CLM) [5] [6] [7] [8] [9], the joint model of shape and texture appearance has the same form as the AAM. However, the CLM appearance model takes the form of rectangular regions around each feature, instead of trying to approximate the image pixels directly. CLM is more robust and more accurate than the original AAM method, by using a different search algorithm.

Recently, regression approaches [10] [11] [2] have been shown to achieve significant better performance than AAM and CLM frameworks. Typically, regression-based methods wish to learn a function that maps the textual features to the landmark locations. They predict either the absolute landmark coordinates directly or the parameter update iteratively based on the current appearance information. In SDM, cascade regression models are used to learn a sequence of mapping functions from local appearance around the current landmark locations to the shape updates. It outperforms state-of-the-art approaches in facial key points detection and tracking in challenging databases.

Global Supervised Descent Method (GSDM) [12] is an extension of the SDM that divides the search space into domains of similar gradient direction. In SDM, a single generic descent direction is learnt in each iteration during training. Different with SDM, GSDM is a global optimization algorithm. It learns a set of generic descent directions for different domains of the objective function. Thus, GSDM is able to track the face from profile to profile with a better performance. However, partition strategies existing within the GSDM framework are still need to be improved. Besides, building models on different domains require larger amounts of training data, which leads to expensive computation and long time training.

Current shape information is not considered when SDM and GSDM learn the regression functions. Shape augmented regression method (SARM) [13] utilize the shape information by adding shape parameters in the regression function directly. The distances among pairs of landmarks or the differences of the coordinates are used as the shape features. Thus, the regression functions can change according to the current face shape.

The proposed method differs from the SARM by using the shape information in a more effective way. We use absolute locations of the facial key points as shape features instead of the distances or the differences. Thus, our method is better than SARM in terms of complexity, time consumption and model size. Moreover, we propose a new well-defined error function to build better regression functions. Different with the error functions used in the previous methods which ignore the shape restraint, the proposed error function utilize both appearance and shape information in optimization. With the additional shape parameters, the performance of the regression method is obviously improved in experiment. The result of comparison is given in section IV.

III. Supervised descent method based on appearance and shape

Face alignment can be solved as a nonlinear optimization problem. As a major optimization tool, the Newton's method can apply to many computer vision problems. Given an initial estimate $\mathbf{x}_0 \in \Re^{n\times 1}$, Newton's method can minimize $f(\mathbf{x})$ by computing a sequence of updates as,

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{H}_f^{-1}(\mathbf{x}_i)\mathbf{J}_f(\mathbf{x}_i)$$

where $\mathbf{H}_f(\mathbf{x}_i) \in \Re^{n\times n}$ and $\mathbf{J}_f(\mathbf{x}_i) \in \Re^{n\times 1}$ are the Hessian matrix and Jacobian matrix evaluated at $\mathbf{x}_i$.

Under conditions that the initial estimate is sufficiently close to the minimum, it is guaranteed to converge and the converge rate is quadratic. However, the Newton's method requires the function to be twice differentiable. Most image operators in computer vision applications do not meet the requirement. For instance, SIFT features extracted from patches are not differentiable. By learning a linear regression between the shape updates and the appearance information, SDM estimates the descent direction directly. Therefore, this method is not limit to the functions that second derivatives are available. And it is able to avoid expensive computation of the Jacobian matrix and Hessian matrix.

When applying SDM in face alignment, a main problem arise: it only uses appearance information to estimate the landmark locations, ignoring the shape information. However, in unconstraint conditions, the appearance information may become unstable with large difference in illumination and occlusion. To overcome the drawback of the original method, we modify the error function.

Suppose we are given an input image I that we wish to align. In particular, assuming the facial mark coordinates are denoted as

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_N \\ y_1 & y_2 & \dots & y_N \end{bmatrix}^{\mathrm{T}}$$

where N represents the number of landmarks. The face shape consisting of facial landmarks is defined as

$$S(\mathbf{x}) = [x_1, y_1, \dots, x_N, y_N]$$

We will assume that the correct N landmarks, denoted as $\mathbf{x}^*$, are known. The goal of the face alignment is to estimate a shape $S(\mathbf{x})$ that is as close as possible to the true shape $S(\mathbf{x}^*)$, minimizing

$$f(\mathbf{x}) = ||S(\mathbf{x}) - S(\mathbf{x}^*)||_2^2$$

The function is used to evaluate the performance of different methods. As $S(\mathbf{x}^*)$ is unknown during testing, it can not be used as error function directly. In previous methods, the error function is defined as follow,

$$f(\mathbf{x}) = \frac{1}{2}||A(\mathbf{x}, I) - A(\mathbf{x}^*, I)||_2^2$$

where $A(\mathbf{x}, I)$ represents the appearance features around the landmark locations $\mathbf{x}$ for image I. The feature extraction function can be SIFT or HOG.

It is notable that the facial texture features may be unreliable due to variations of illumination or occlusion in unconstraint conditions. However, the topology relationship of the facial landmarks remains unchanged. We redefined the error function as follows,

$$f(\mathbf{x}) = \frac{1}{2}||A(\mathbf{x}, I) - A(\mathbf{x}^*, I)||_2^2 + \frac{1}{2}||S(\mathbf{x}) - S(\mathbf{x}^*)||_2^2 \quad (1)$$

Then, face alignment is achieved by minimizing the error function

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad (2)$$

Assuming an initial face shape $\mathbf{x}_0$, we apply a second order Taylor expansion on the objective function. The function can be approximated as,

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}_0 + \triangle\mathbf{x}) \\ &\approx f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0)^{\mathrm{T}}\triangle\mathbf{x} + \frac{1}{2}\triangle\mathbf{x}^{\mathrm{T}}\mathbf{H}_f(\mathbf{x}_0)\triangle\mathbf{x} \end{aligned} \quad (3)$$

where $\mathbf{J}_f(\mathbf{x}_0)$ and $\mathbf{H}_f(\mathbf{x}_0)$ are the Jacobian and Hessian matrices of the function f evaluated at the current shape $\mathbf{x}_0$. In Equation 3, take the derivation of $f(\mathbf{x})$ with respect to $\triangle\mathbf{x}$ and set it to zero. We get the update for $\mathbf{x}$,

$$\begin{aligned} \triangle\mathbf{x} &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_f(\mathbf{x}_0) \\ &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}[\mathbf{J}_A(\mathbf{x}_0)(A(\mathbf{x}_0, I) - A(\mathbf{x}^*, I)) \\ &\qquad + \mathbf{J}_S(\mathbf{x}_0)(S(\mathbf{x}_0) - S(\mathbf{x}^*))] \\ &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_A(\mathbf{x}_0)A(\mathbf{x}_0, I) - \mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_S(\mathbf{x}_0)S(\mathbf{x}_0) \\ &\quad + \mathbf{H}_f(\mathbf{x}_0)^{-1}(\mathbf{J}_A(\mathbf{x}_0)A(\mathbf{x}^*, I) + \mathbf{J}_S(\mathbf{x}_0)S(\mathbf{x}^*)) \end{aligned} \quad (4)$$

To estimate the $\triangle\mathbf{x}$ in Equation 4, the Jacobian and Hessian matrices need to be recomputed for each iteration, which is computationally expensive. Another issue is that during testing, $\mathbf{x}^*$ is unknown but fixed. To solve these issues, some parameters are used in the optimization procedure.

$$\mathbf{R} = -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_A(\mathbf{x}_0) \quad (5)$$

$$\mathbf{Q} = -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_S(\mathbf{x}_0) \quad (6)$$

$$\mathbf{b} = \mathbf{H}_f(\mathbf{x}_0)^{-1}(\mathbf{J}_A(\mathbf{x}_0)A(\mathbf{x}^*, I) + \mathbf{J}_S(\mathbf{x}_0)S(\mathbf{x}^*)) \quad (7)$$

By introducing these regression parameters, we can rewrite the Equation 4 as follows,

$$\triangle\mathbf{x} = \mathbf{R}A(\mathbf{x}_0, I) + \mathbf{Q}S(\mathbf{x}_0) + \mathbf{b} \quad (8)$$

During training, the true shape updates of kth image in ith iteration $\triangle\mathbf{x}_i^{k,*} = \mathbf{x}_i^* - \mathbf{x}_i^k$ can be computed directly. Then, in each gradient iteration, parameter estimation can be solved by minimizing

$$\begin{aligned} &\tilde{\mathbf{R}}_i, \tilde{\mathbf{Q}}_i, \tilde{\mathbf{b}}_i \\ &= \arg\min_{\mathbf{R}_i, \mathbf{Q}_i, \mathbf{b}_i} \sum_k ||\triangle\mathbf{x}_{i-1}^{k,*} - \mathbf{R}_i A(\mathbf{x}_{i-1}^k, I^k) - \mathbf{Q}_i S(\mathbf{x}_{i-1}^k) - \mathbf{b}_i||_2^2 \end{aligned} \quad (9)$$

Minimizing the Equation 9 can be formulated as a least square problem. And the parameters can be obtained using a ridge-regression. When the regression parameters are learned, the face shape updates can be estimated. Adding the shape updates $\triangle\mathbf{x}_{i-1}$ to the current key points locations $\mathbf{x}_{i-1}$, we can get the new shape $\mathbf{x}_i$, which will be used in the next iteration.

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_{i-1} + \triangle\mathbf{x}_{i-1} \\ &= \mathbf{x}_{i-1} + \mathbf{R}_i A(\mathbf{x}_{i-1}, I) + \mathbf{Q}_i S(\mathbf{x}_{i-1}) + \mathbf{b}_i \end{aligned} \quad (10)$$

The shape updates monotonically decrease as a function of the number of regressors added. Thus, the succession of $\mathbf{x}_i$ converges to $\mathbf{x}^*$ gradually.

The overall framework of training and testing is summarized in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 Supervised descent method based on appearance and shape:Training

```
Normalize the images and landmark locations in training set
Calculate the mean face using ground truth
generate multiple initial face shapes for each training image by randomly re-scaling,rotating,and shifting the mean face
for $i = 1, 2, \ldots, \mathrm{N}$ do
    Extract feature descriptor
    Estimate regression parameters $\mathbf{R}_i, \mathbf{Q}_i, \mathbf{b}_i$ using Equation 9
    update the landmark location using Equation 10
end for
Output the regression parameters $\mathbf{R}, \mathbf{Q}, \mathbf{b}$
```

Algorithm 2 Supervised descent method based on appearance and shape:Testing

```
Initialize the landmark locations $\mathbf{x}_0$ using the mean face
for $i = 1, 2, \ldots, \mathrm{N}$ do
    Extract feature descriptor
    update the landmark location using Equation 10
end for
Output the estimated landmark location $\mathbf{x}_\mathrm{N}$
```

IV. Experiments

The past decades many research communities have collected a number of facial databases. We briefly introduce the data sets which are used in the experiments, namely Labeled Face Part in the Wild (LFPW) [14], Helen [15], Annotated Face in the wild (AFW) [9] and iBug [16].

LFPW was proposed by Belhumeur et al. in 2011. Different with Helen data set, they did not intentionally filtered out faces due to poor image quality. The images of human faces in LFPW are taken under a variety of acquisition conditions, used to test the face alignment methods in unconstraint conditions. Some images are no longer valid and we only download 811 of the 1,100 training images and 224 of the 300 testing images.

Helen was created by Le et al. in 2012 that consists of 2,000 training images and 330 test images. These high resolution images are gathered from Internet under a broad range of appearance variation, including pose, lighting, expression,

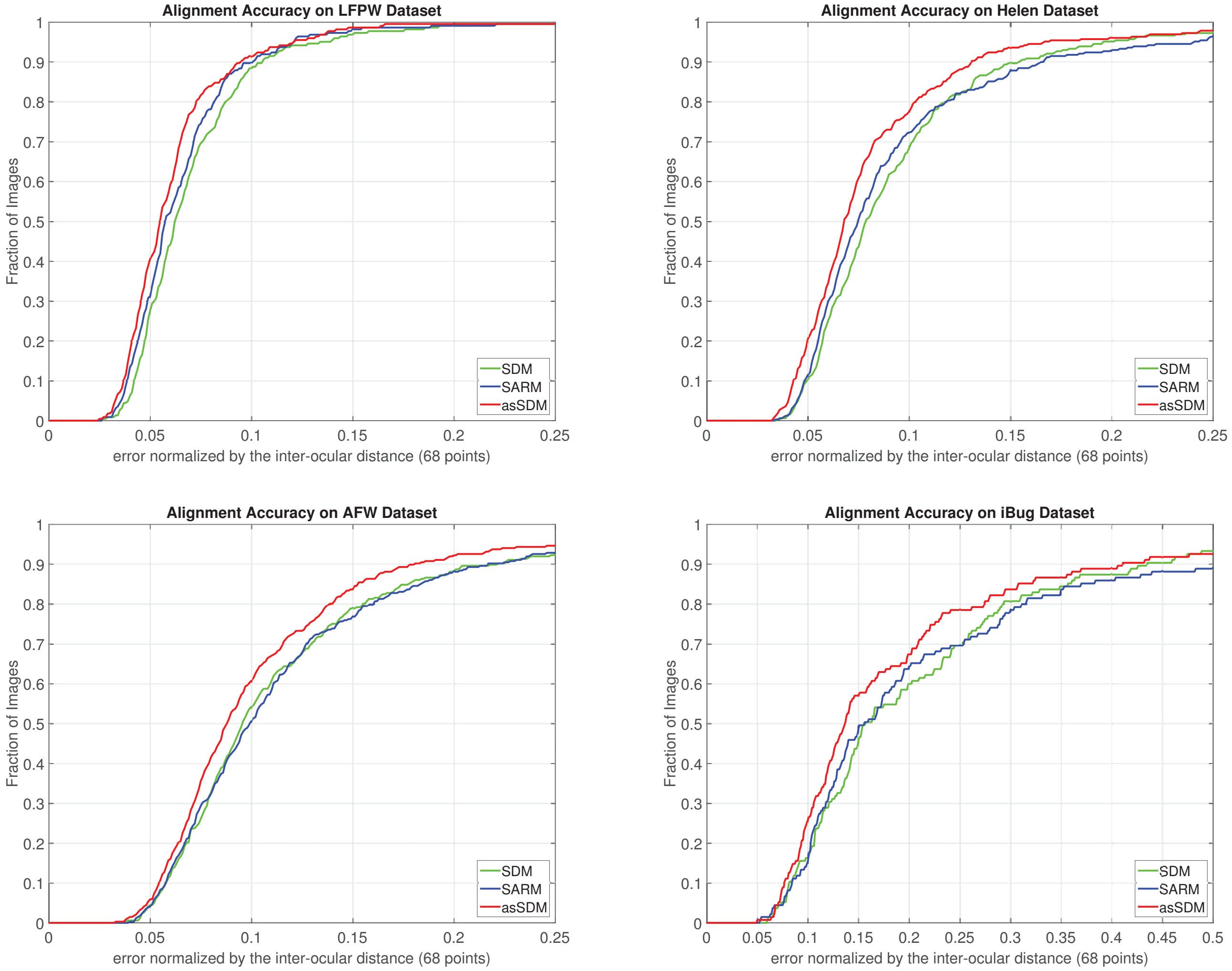

Fig. 2: Average pt-pt Euclidean error vs fraction of images for LFPW, Helen, AFW and iBug. We compare the performance of our method, SDM and SARM. The average error is computed over 68 points.

occlusion, and individual differences. All the images are hand-annotated to precisely locate the eyes, nose, mouth, eyebrows, and jawline.

AFW is also considered as general 'in-the-wild' database, and we use it for testing in the experiments. There are 337 images in AFW, which are more challenging.

iBug data set is the most challenging one with large shape and appearance variation. The data set contains only 135 images and we use all images from Helen and AFW as training set.

We compare the performance of our method with the similar work based on regression framework: SDM and SARM. All algorithms are implemented in Matlab and tested on Intel i-7 CPU with the same initialization and parameters.

In our experiment, the available landmark annotations of the 300-W challenge are used as ground truth points for training and evaluation. All facial images in the data sets are manually labeled with 68 landmarks.

For initialization, the ground truth points are used to compute the ground truth bounding box for each image. We normalize the images and landmark locations, and the width of the bounding box is 400 pixels. The mean face shape can be calculated with the normalized ground truth. For each training image, the facial landmark locations are estimated by the bounding box and mean face shape. By randomly re-scaling, rotating and shifting the mean face, we generate multiple initial samples for each image. The error is measured as the average Euclidean distance among the pairs of labeled and predicted landmark locations. Then we estimate the central locations of two eyes separately and the error is normalized by the distance between two centers.

TABLE I: Comparison of facial landmark detection errors, runtime (in FPS) and model size on LFPW database. The average error is computed over 68 points.

algorithm	SDM	SARM	asSDM
Error ($\times 10^{-2}$)	7.08	6.63	6.17
FPS/Hz	20.3	19.6	20.2
Model size/MB	26.0	29.8	26.4

The experimental results and their comparison with other similar methods are shown in Figure 2. In Figure 3, we present the fittings produced by our method on Helen data set. In Figure 4, the first two rows show the worst images of iBug data set measured by normalized mean error and the last four rows show some faces with reliable results. Table I reports the normalized errors, speeds (frames per second or FPS) and model size of the compared methods on LFPW data sets. From these results, there are some observations. First, AFW data sets are more challenge than LFPW and Helen data set and iBug data set is the most challenging. Our approach achieve error reduction in different data sets with respect to SDM and SARM. And it is the most robust one compared with the others. We believe this is due to the better utilization of the shape information. Second, as shown in Table I, the speed of our approach is almost the same as the speed of SDM. We can conclude that our approach outperforms SARM in both accuracy and efficiency. Third, the model size of our method is smaller than that of SARM, and just a little bit larger than the model size of SDM. The distances among pairs of landmarks are used as shape features in SARM, which has complexity of $O(N^2)$. In our approach, we use the absolute coordinates of landmarks as shape features without additional computation and the complexity of this part is linear. Thus, SARM is more time consuming than the proposed approach and the model size is bigger.

V. Conclusion

This paper presents a cascaded regression approach based on appearance and shape. The proposed method is able to overcome the limitations of the previous method, which ignore the shape constraint and utilize the appearance information only. An important contribution of this work is to redefine the error function by adding the shape information. The new error function can be applied to any face alignment method solved as an optimization problem. We conduct a large number of experiments on the most popular databases and our method outperforms the other similar methods. In future work, we would apply our method to facial feature tracking and other computer vision problem. Also, we will implement our method in C/CUDA to make it realtime.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[2] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[3] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[4] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 300–305.

[5] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." *BMVC*, vol. 1, no. 2, p. 3, 2006.

[6] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[7] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2729–2736.

[8] Y. Wu and Q. Ji, "Discriminative deep face shape model for facial point detection," *International Journal of Computer Vision*, vol. 113, no. 1, pp. 37–53, 2015.

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

[10] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[11] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1685–1692.

[12] X. Xiong and F. De la Torre, "Global supervised descent method," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2664–2673.

[13] Y. Wu and Q. Ji, "Shape augmented regression method for face alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 26–32.

[14] 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*, vol. 35, no. 12, pp. 2930–2940, 2013.

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

[16] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903.

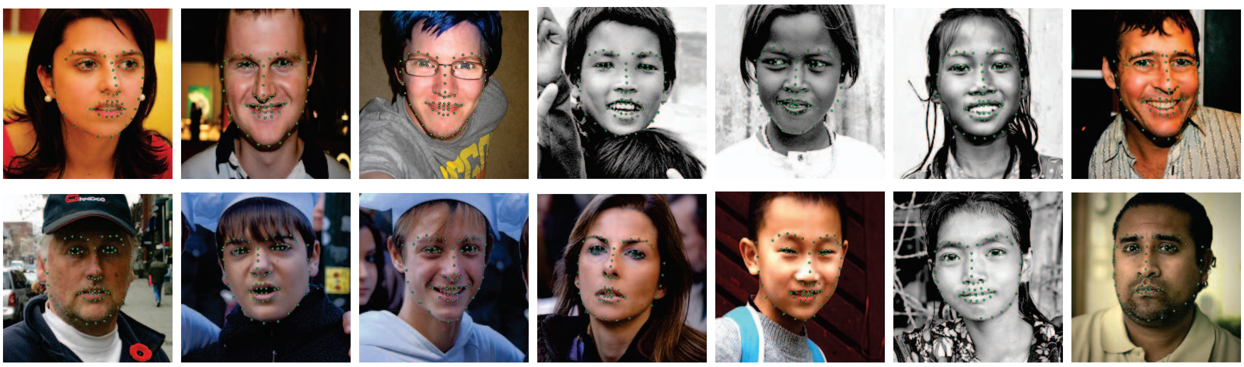

Fig. 3: Example results from our method on Helen data set.

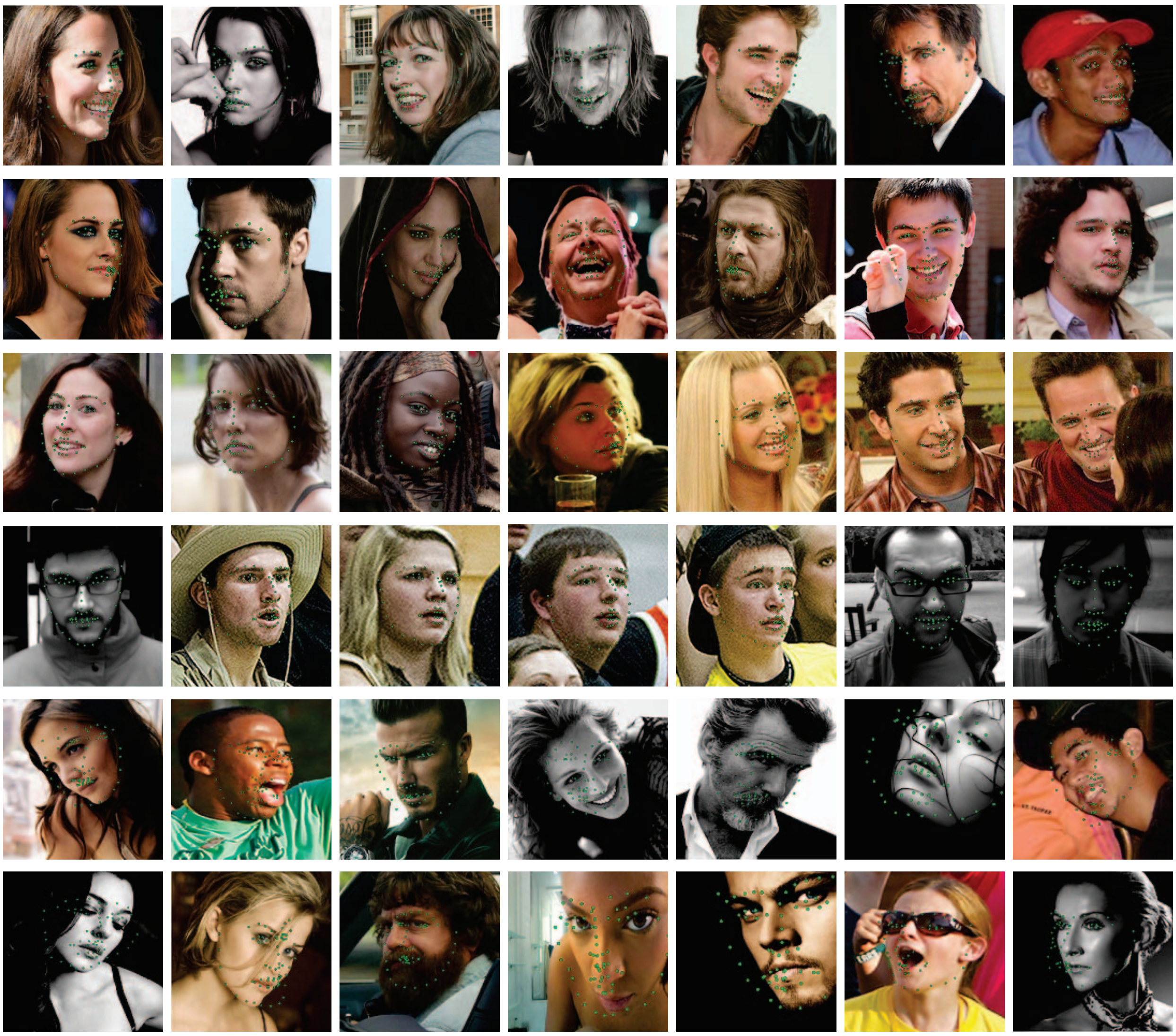

Fig. 4: Example results from our method on iBug data set. The first four rows show some faces with reliable results. The last two rows show the worst images measured by normalized mean error.