Shape Augmented Regression Method for Face Alignment

Yue Wu Qiang Ji
ECSE Department, Rensselaer Polytechnic Institute
110 8th street, Troy, NY, USA
{wuy9,jiq}@rpi.edu

Abstract

There have been tremendous improvements of the face alignment algorithms, among which the regression framework becomes the most popular one recently. The regression based works start from an initial face shape, and they learn regression models to predict the face shape updates based on the shape-indexed local appearance features. However, most of the regression methods ignore the fact that the regression function should directly rely on the current shape (e.g. regression function for frontal face should be different from that for the left profile face). To utilize this information and improve over the existing regression based methods, we propose the shape augmented regression method for face alignment where the regression function would automatically change for different face shapes. We evaluated the performance of the proposed method on both the general "in-the-wild" database and the 300 Video in the Wild (300-VW) challenge data set. The results show that the proposed method outperforms the state-of-the-art works.

1. Introduction

The goal of face alignment algorithm is to locate several facial key points and landmarks (e.g. eye corner, mouth corner) on the facial images (see Figure 1). The information of landmark locations is crucial for understanding and analyzing the human facial behavior. For example, the locations of facial landmarks can be used as features for facial expression analysis [27], head pose estimation [14], etc. However, face alignment is still a challenging task, even though there have been extensive studies over the past few decades.

The typical face alignment methods utilize the facial appearance information and the face shape patterns. The facial appearance information refers to the distinct intensity patterns around the facial landmarks or in the whole face region. The face shape patterns refer to the distinct patterns of the face shape defined by the landmark locations. For example, in the early study, the Active Appearance Model (AAM) [6] represents the appearance (texture) and shape information of human face with Principle Component techniques. In the Constrained Local Method (CLM) [8], local appearance information and the global face shape patterns are modeled. Recently, regression based works [26][15][4] show significant better performances than the AAM and CLM frameworks. They do not explicitly learn any shape model, and they directly predict the landmark locations from the facial appearance features.

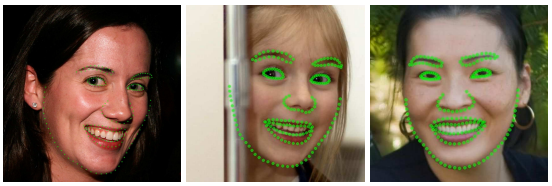


Figure 1. Facial images with the detected facial landmarks with the proposed method. Images are from the Helen database [11]

One major limitation of the current regression based methods is the ignorance or ineffective usage of the shape information. Specifically, the existing regression based methods [26][15][4] learn the regression functions without explicitly considering the current global shape, while, in fact, the regression function should change according to the current face shape (e.g. frontal face should have different regression function from that for the profile face). To tackle this problem and improve upon the regression based methods, in this work, we explicitly combine the shape information with the appearance information to build better regression functions.

The remaining part of the paper is organized as follows. In section 2, we review the related work. In section 3, we introduce the proposed method. In section 4, we show the experimental results. We conclude the paper in section 5.

2. Related Work

Face alignment methods can be classified into three major categories, including the holistic methods, the constrained local methods, and the regression based methods. The holistic methods model the facial appearance in the whole face region and the global face shape. The typical holistic model is the Active Appearance Model [6]. Most holistic methods [12][2][21] focus on designing new fitting algorithms that estimate the model parameters for the testing image. The Constrained Local Methods (CLM) [17][22][28][24] search each landmark point independently based on the local appearance information with the constraint that the estimated face shape satisfies the anthropological constraint embedded in the face shape model. It has been shown that the CLM based works outperform AAM based approaches, and they achieve better robustness in terms of illuminations, occlusions, etc.

The regression based methods directly learn the mapping from facial images to the landmark locations with regression models. They directly predict either the absolute landmark coordinates or the displacement vectors (the current landmark locations to the ground truth locations) based on the image appearance features. For example, in [19], deep convolutional neutral network model is used to learn the direct mapping from facial appearance of the whole face region to the landmark coordinates. In [4][19][26], regression models are used to learn the mapping from local appearance around the current landmark locations to the shape updates.

There are some regression based works that utilize the shape information, although in a limited sense. In [4], the shape indexed image features are proposed to embed the local facial appearance. It uses the pixel intensity differences as features while the pixel is indexed relative to the currently estimated shape instead of the image coordinate system. It is shown in [4] that the shape-indexed features achieve better geometric invariance and faster convergence. In the typical regression methods, since all the points are updated jointly, the shape pattern constraint is implicitly embedded in the model. In [9], the regression functions are learned for faces with different head poses which are then combined jointly in testing to handle varying poses.

The most similar work to our work is the Global Supervised Descent method (Global SDM) [25], which is developed independently and concurrently with the proposed method. It improves over the Supervised Descent Method (SDM) [26] by learning eight sets of shape dependent regression functions, where each set of the regression function is specifically learned for images with similar shape updates. During testing, one set of the regression function is selected based on the shape updates from the last iteration. The proposed method differs from the Global SDM. First, the proposed method can automatically and continuously adjust the regression model parameters based on the augmented shape features, and it does not need to select from a few regression functions as the Global SDM. Second, the proposed shape augmentation method can be applied to any face alignment method with regression framework. Third, in Global SDM, learning a few sets of regression functions may need a larger number of training data, and it is more time consuming than the proposed method.

3. Shape augmented regression method

In this section, we first discuss the general regression methods and then introduce the proposed shape augmented regression method.

3.1. General regression method

The Supervised Decent Method (SDM) [26] is one popular regression method that has been applied to face alignment. It formulates the face alignment task as a general optimization problem, which is then approximately solved by learning several sequential mapping functions from the local appearance to the shape updates with linear regression models. In particular, assuming the facial landmark coordinates are denoted as $\mathbf{x} = \{x_1, x_2, ..., x_D\}$, where D denotes the number of landmarks, face alignment problem can be formulated as a general optimization problem:

$$\begin{aligned} \widetilde{\mathbf{x}} &= \arg\min_{\mathbf{x}} f(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} \frac{1}{2}\|\Phi(\mathbf{x}, I) - \Phi(\mathbf{x}^*, I)\|_2^2. \end{aligned} \tag{1}$$

Here, $\Phi(\mathbf{x}, \mathbf{I})$ represents the local SIFT features around the landmark locations $\mathbf{x}$ for image $\mathbf{I}$. The ground truth landmark locations are denoted as $\mathbf{x}^*$. With Taylor expansion and assuming that we have an initial face shape $\mathbf{x}_0$, the objective function can be approximated as:

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}_0 + \Delta\mathbf{x}) \\ &\approx f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0)^T\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}_f(\mathbf{x}_0)\Delta\mathbf{x}, \end{aligned} \tag{2}$$

where $\mathbf{J}_f(\mathbf{x}_0)$ and $\mathbf{H}_f(\mathbf{x}_0)$ represent the Jacobian and Hessian matrices of function $f(.)$ evaluated at the current shape $\mathbf{x}_0$. Then, with gradient descent method and the Taylor expansion approximation, the problem becomes finding the shape updates $\Delta\mathbf{x}$ that minimize Equation 2 in each shape update step. The new face shape can be calculated as $\mathbf{x}_0 + \Delta\mathbf{x}$, which is then used in the next iteration until convergence. In Equation 2, take the derivation of $f(\mathbf{x})$ w.r.t. $\Delta\mathbf{x}$ and set it to zero, we get:

$$\begin{aligned} \Delta\mathbf{x} &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_f(\mathbf{x}_0) \\ &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi(\mathbf{x}_0)(\Phi(\mathbf{x}_0, \mathbf{I}) - \Phi(\mathbf{x}^*, \mathbf{I})) \\ &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi(\mathbf{x}_0)\Phi(\mathbf{x}_0, \mathbf{I}) + \mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi(\mathbf{x}_0)\Phi(\mathbf{x}^*, \mathbf{I}) \end{aligned} \tag{3}$$

There are two difficulties in the estimation of $\Delta\mathbf{x}$ in Equation 3. First, the Jacobian and Hessian matrices are generally difficult to estimate due to the involved complex features such as SIFT. Second, the ground truth landmark locations $\mathbf{x}^*$ are unknown in testing. To solve these issues, SDM uses some approximate representations:

$$\mathbf{R} = -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi(\mathbf{x}_0) \tag{4}$$

$$\mathbf{b} = \mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_\Phi(\mathbf{x}_0)\Phi(\mathbf{x}^*, \mathbf{I}) \tag{5}$$

Here, $\mathbf{R} \in \Re^{2D\times 128D}$ and $\mathbf{b} \in \Re^{2D}$ are the regression parameters. The SDM assumes that $\mathbf{R}$, $\mathbf{b}$ and the regression function only change according to iterations but stay the same, regardless of the current face shape $\mathbf{x}_0$. Then, in each gradient descent iteration t, the shape update can be calculated as:

$$\Delta\mathbf{x}^t = \mathbf{R}^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{b}^t \tag{6}$$

The shape update is added to the current shape for the estimation in the next iteration.

3.2. Shape augmented regression method

One major limitation of the Supervised Descent Method and some other general regression methods is that they usually assume constant regression parameters $\mathbf{R}$ and $\mathbf{b}$ for each gradient descent iteration, but, in fact, these regression parameters should change according to the current face shape as shown on the right hand side in Equations 4 and 5. To overcome this limitation, we propose to modify the regression function in Equation 6, so that the regression prediction would change explicitly according to the current face shape $\mathbf{x}^{t-1}$. To achieve this goal, we add an additional term in the regression function:

$$\Delta\mathbf{x}^t = \mathbf{R}^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{b}^t + \mathbf{Q}^t\Psi(\mathbf{x}^{t-1}), \tag{7}$$

where $\Psi(\mathbf{x}^{t-1})$ represents the shape features and $\mathbf{Q}^t$ represents additional parameters of the linear regression model. With the new term, the linear regression function would change directly according to different face shapes, which would better approximate the true prediction in Equation 3. The general framework of the shape augmented regression method is summarized in Algorithm 1.

One remaining issue is to figure out how to embed the shape information and calculate the shape features $\Psi(\mathbf{x}^{t-1})$ in Equation 7. There are a few options. The first approach is to learn a Point Distribution Model (PDM) beforehand and use the shape model parameters to represent the current shape. However, as shown in the prior arts [7][3][24], the face shape models may be difficult to construct. The second option is to use the distances among pairs of landmarks. The local distance information would provide useful features to the regression model. However, the distance metric would tend to stay unchange with in-plane rotation. The third option is to use the differences (both x and y coordinates) between pairs of landmark locations. With the shape features, the linear regression model would change according to different face shapes that vary with scale, rotation, translation, and non-ridge changes. Some details and experimental comparison are shown in section 4.

Algorithm 1: Shape augmented regression method

Initialize the landmark locations $\mathbf{x}^0$ using the mean face
for *t=1, 2, ..., T or convergence* **do**
  Predict the landmark location update

$$\Delta\mathbf{x}^t = \mathbf{R}^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{b}^t + \mathbf{Q}^t\Psi(\mathbf{x}^{t-1})$$

  Update the face shape

$$\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t$$

end
Output the estimated landmark locations $\mathbf{x}^T$

Model training is similar to the standard regression and SDM algorithm. Given the training data, including the facial image $\mathbf{I}_i$, the ground truth landmark locations $\mathbf{x}_i^*$, and the initial landmark locations $\mathbf{x}_i^0$ as the mean face, we first calculate the appearance and shape features, denoted as $\Phi(\mathbf{x}_i^0, \mathbf{I}_i)$ and $\Psi(\mathbf{x}_i^0)$. In addition, we can calculate the true shape updates $\Delta\mathbf{x}_i^{0,*} = \mathbf{x}_i^* - \mathbf{x}_i^0$. Then, parameter estimation in each iteration can be formulated as a least square problem with closed form solution:

$$\widetilde{\mathbf{R}}^t, \widetilde{\mathbf{Q}}^t, \widetilde{\mathbf{b}}^t = \arg\min_{\mathbf{R}^t,\mathbf{Q}^t,\mathbf{b}^t} \sum_i \|\Delta\mathbf{x}_i^{t,*} - \mathbf{R}^t\Phi(\mathbf{x}_i^{t-1}, \mathbf{I}_i) - \mathbf{b}^t - \mathbf{Q}^t\Psi(\mathbf{x}_i^{t-1})\|_2^2 \tag{8}$$

Given the learned regression parameters $\widetilde{\mathbf{R}}^t, \widetilde{\mathbf{Q}}^t, \widetilde{\mathbf{b}}^t$, the face shape updates for the training data can be estimated, and they are added to the current shape $\mathbf{x}_i^{t-1}$ to generate the shape $\mathbf{x}_i^t$ for the prediction in the next iteration.

4. Experimental results

In this section, we first discuss the implementation details. We then evaluate the proposed method and compare it to other state-of-the-arts on the general "in-the-wild" database and the 300-VW database.

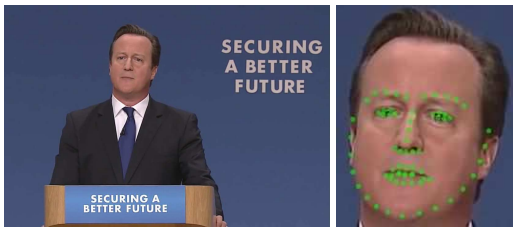
4.1. Implementation details

Databases: In the experiments, we used two kinds of databases. The first type of database refers to the general "in-the-wild" Helen database [11]. The Helen database contains general "in-the-wild" facial images collected from the web (Figure 3). It has 2000 training images and 330 testing images. All the images are annotated with 194 facial

landmarks. The Labeled Face Part in the Wild [3], and the Annotated Face in the Wild (AFW) databases [28] can also be considered as general "in-the-wild" databases, which we also used in the experiments.

The second database refers to the data provided by the 300 Videos in the Wild Challenge (300-VW) [18]. The training data includes both "in-the-wild" videos and images. Specifically, there are 50 high resolution video sequences with moderate expression, head pose, and illumination changes. Each video lasts around 1 minute. Facial landmark annotations (68 points) are provided for each frame with the semi-automatic annotation process [5][20]. There are also annotated "in-the-wild" images from the 300 Faces In-the-Wild Challenge (300-W) [16], which are collected from Helen[11], Labeled Face Part in the Wild (LFPW) [3], Annotated Face in the Wild (AFW) databases [28] etc.

There are three testing scenarios in the 300-VW test set, each contains 50 videos. Sample images and the detection results with the proposed method are shown in Figure 2. Videos in scenario 1 are with "well-lit" conditions displaying moderate facial expression, head poses, and illumination changes. Videos in this scenario could represent the facial motion in laboratory or naturalistic "well-lit" conditions. Videos in the second scenario are recorded in unconstrained conditions with varying illuminations, arbitrary facial expressions, and moderate head poses without significant illuminations. Videos in the second scenario would mimic the real world human computer interaction applications. The third scenario contains videos captured in completely unconstrained conditions, including significant occlusions, make-up, and head poses. The third scenario could be considered as the most challenging "in-the-wild" video sequences recorded.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

Figure 2. Left: sample images from three scenarios of 300-Video in the Wild (300-VW) test set. Right: detection results with the proposed method.

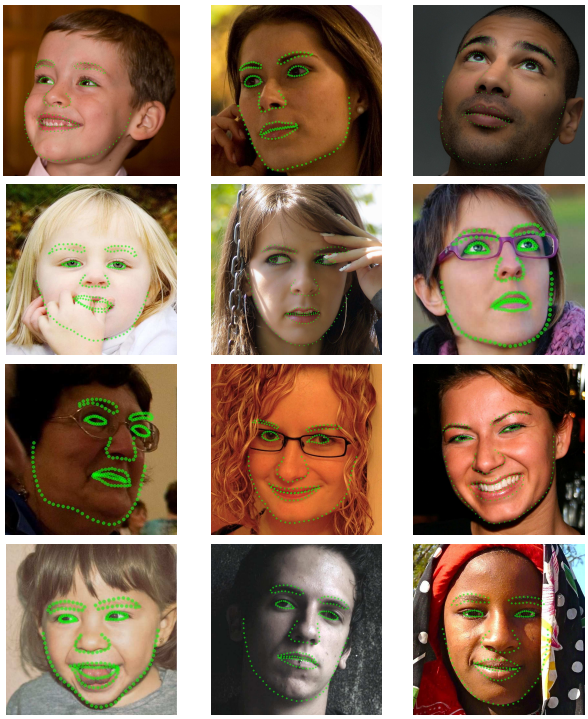


Figure 3. Facial landmark tracking results on sample images from Helen database [11] with the proposed methods.

Model parameters: Before model training, we first apply the Viola-Jones face detector [23] to training images. Both the images and landmark locations are normalized, so that the width of the face is approximated 200 pixels. Then, we calculate the mean face shape with the normalized training face shapes. To improve the robustness of the learned model, similar as the existing regression based methods [26], we generate multiple initial face shapes for each training image by randomly re-scaling, rotating, and shifting the mean face. To calculate the face shape features as discussed in section 3, we use every pair if only 68 landmarks are provided. If 194 landmarks are provided, to prevent over-fitting and improve the efficiency, we only use every pair of 64 key points.

Extend detection to tracking: To apply the shape-augmented regression methods to tracking, the only change

Table 1. Comparison of facial landmark detection errors (normalized by inter-ocular distance, E_{ocular}) on Helen database [11]. The reported results from the original papers are marked with "*".

algorithm	Helen (E_{ocular})
LBF [15]	**5.41**(fast 5.80)*
General regression method (SDM [26])	5.82
ESR [4]	5.70*
CompASM [11]	9.10*
STASM [13]	11.1
outs (distance shape feature)	**5.49**
ours (difference shape feature)	**5.53**

is to initialize the face shape based on the landmark locations in the last frame.

Evaluation criteria: There are two evaluation criteria in the experiments. In the first evaluation criteria, we calculate the detection and tracking error as the distance between the estimated landmark location and the ground truth location normalized by the inter-ocular distance (times 100). In the second evaluation criteria, we normalize the error by the distance between two outer eye corners. We use two criteria, because existing state-of-the-arts tend to use different criteria on different databases. Thus, we follow them for fair comparison. We denote the first and second criteria as E_{ocular} and E_{corner}. In the experiments, we show the mean errors across all landmarks for all testing images. In addition, we show the cumulative distribution curves.

4.2. General "in-the-wild" database

In this section, we show the experimental results on general "in-the-wild" Helen database [11]. We train the model with the training set and test it on the testing set. The training setting is similar to that used in the existing works [15]. In this section, we calculate the error by normalizing it with the inter-ocular distance (E_{ocular}) for fair comparison. We implemented two versions of the proposed method with different realizations of the shape features as discussed in section 3.2. The "distance shape features" refer to the distance between pairwise landmarks, while the "difference shape features" refer to the location differences.

The experimental results and their comparison with other state-of-the-art works are shown in Table 1 and Figure 3. There are a few observations. First, the proposed shape augmented regression method is better than the general SDM algorithm without the explicit shape information. Second, the proposed method outperforms the other state-of-the-art works, including the Explicit Shape Regression (ESR) method [4], Component based ASM (CompASM) [11], the STASM [13] method, and the fast LBF [15] algorithm. Third, the "distance shape features" and "difference shape features" achieve similar performances.

4.3. 300-VW database

In this section, we discuss the experimental results on 300-VW testing set with three scenarios. To train the model, we use the training set from Helen [11], LFPW [3], AFW [28], and some images from Multi-PIE database [10]. In total, there are 6222 training images. We randomly sampled some frames from the videos in the 300-VW training set. We use the "difference shape features" in this section and we calculate the error by normalizing it with distances between two outer eye-corners (E_{corner}). Chehra [1] with cascade framework is the baseline method.

The experimental results are shown in Figure 4, 5, and Table 2. There are a few observations. First, the proposed method significantly outperforms Chehra [1] as the baseline in all three scenarios. Second, the performances on scenario 2 are better than those on scenario 1 and 3. Scenario 3 is the most challenging set, and the performances of both algorithms drop significantly when evaluating on this set. Third, by considering the contour points when calculating the errors, the performances drop for all three scenarios. The most obvious drop happens on scenario 3. The visual results can be found in Figure 2, and Figure 6 shows the results on one sample sequence.

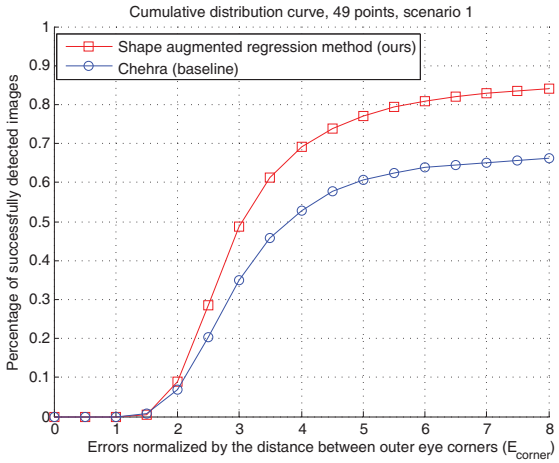
The proposed algorithm can achieve realtime tracking on a typical dual-core desktop machine with matlab implementation. It can track 68 points in about 20 frames per second. The speed is comparable to most of the existing algorithms [26][28][19], but is slower than very efficient algorithm, such as the LBP method in [15].

5. Conclusion

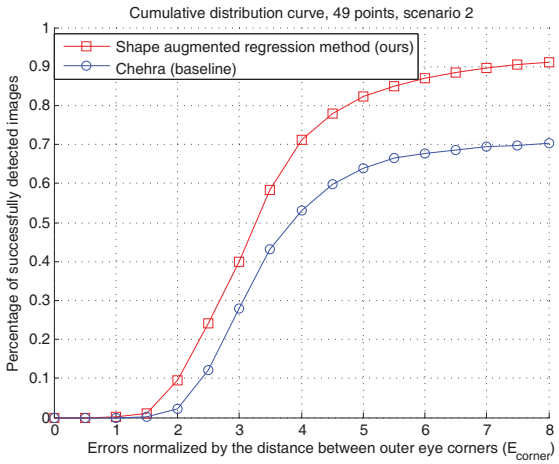
In this work, we proposed the shape augmented face alignment method. Different from the conventional regression based face alignment methods, which only use the appearance features, the proposed method explicitly adds the shape information. As a result, the regression prediction function would directly change according to different face shapes. We evaluated the proposed method on both

Table 2. Detection rates (percentage of successfully detected images with an error less than 6, E_{corner}) of the proposed method and the baseline (Chehra [1]) on 300-VW databases.

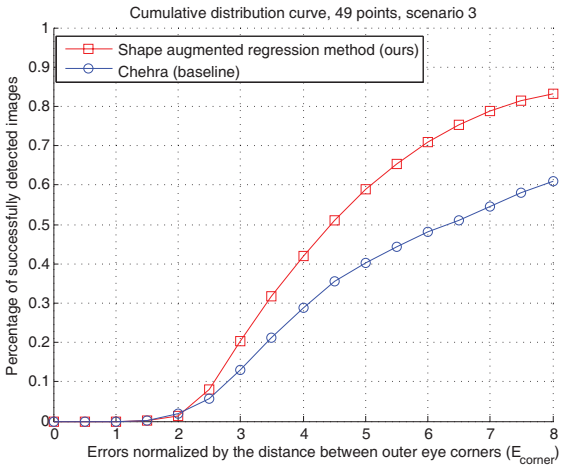
	# of points	Baseline [1]	Proposed method
scenario 1	49	63.75%	80.95%
	68	-	75.30%
scenario 2	49	67.75%	87.00%
	68	-	83.47%
scenario 3	49	47.96%	70.87%
	68	-	52.78%



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

Figure 4. Cumulative distribution curves of facial landmark detection and tracking results (**49 points**, E_{corner}) on the 300-VW databases with three scenarios.

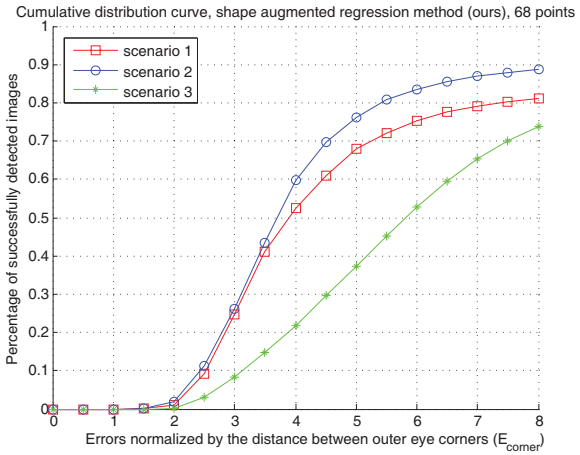


Figure 5. Performance of the proposed shape augmented facial landmark detection algorithm (**68 points**, E_{corner}) on the 300-VW database with three scenarios.

the general "in-the-wild" database and the 300 Video in the Wild (300-VW) challenge data set, on which the proposed method outperforms the state-of-the-art works.

In the future, we would extend the proposed work in a few directions. First, the evaluation results on the most challenging scenario in 300-VW indicate that the performances of the existing algorithms, including the proposed method, drop noticeably comparing to those on the easier scenarios. So, we should improve the robustness of the proposed algorithm to handle the difficult cases and compare it to similar work (e.g. Global SDM [25]) in those conditions. Second, contour point detection still remains challenging, as indicated by the fact that by adding the contour points, the performances of the proposed method drop noticeably. So, we should also improve the method to better handle the contour points. Third, we would further evaluate different shape features on more databases.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014. 5

[2] S. Baker, R. Gross, and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 3. *International Journal of Computer Vision*, 56:221–255, 2002. 2

[3] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, Dec 2013. 3, 4, 5

[4] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, pages 177–190, 2014. 1, 2, 5

[5] G. Chrysos, S. Zafeiriou, E. Antonakos, and P. Snape. Offline deformable face tracking in arbitrary videos. In *IEEE*

Figure 6. Facial landmark tracking results on one sample sequence from scenario 2 with the proposed method.

International Conference on Computer Vision Workshops. IEEE. 4

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001. 1, 2

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, Jan. 1995. 3

[8] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference*, pages 95.1–95.10. BMVA Press, 2006. 1

[9] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012. 2

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

[11] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 679–692, 2012. 1, 3, 4, 5

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

[13] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conference on Computer Vision: Part IV*, ECCV '08, pages 504–513, Berlin, Heidelberg, 2008. Springer-Verlag. 5

[14] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009. 1

[15] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, June 2014. 1, 5

[16] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, June 2013. 4

[17] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, Jan. 2011. 2

[18] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision Workshops*. IEEE, 2015. 4

[19] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013. 2, 5

[20] 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. 4

[21] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *IEEE International conference on Computer Vision*, pages 593–600. 2

[22] M. Valstar, B. Martinez, V. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 13–18, 2010. 2

[23] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004. 4

[24] Y. Wu and Q. Ji. Discriminative deep face shape model for facial point detection. *International Journal of Computer Vision*, 113(1):37–53, 2015. 2, 3

[25] X. Xiong and F. De la Torre. Global supervised descent method. June 2015. 2, 6

[26] X. Xiong and F. De la Torre Frade. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition*, May 2013. 1, 2, 4, 5

[27] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, January 2009. 1

[28] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. 2, 4, 5