3D-assisted Coarse-to-fine Extreme-pose Facial Landmark Detection

Shengtao Xiao[1] Jianshu Li[1] Yunpeng Chen[1] Zhecan Wang[3] Jiashi Feng[1] Shuicheng Yan[2,1]
Ashraf Kassim[1]
[1]National University of Singapore [2]Qihoo 360 [3]Olin College
{xiao_shengtao,jianshu,chenyunpeng}@u.nus.edu, olinzhecanwang@gmail.com
{elefjia,eleyans,ashraf}@nus.edu.sg

Abstract

We propose a novel 3D-assisted coarse-to-fine extreme-pose facial landmark detection system in this work. For a given face image, our system first refines the face bounding box with landmark locations inferred from a 3D face model generated by a Recurrent 3D Regressor at coarse level. Another R3R is then employed to fit a 3D face model onto the 2D face image cropped with the refined bounding box at fine-scale. 2D landmark locations inferred from the fitted 3D face are further adjusted with the popular 2D regression method, i.e. *LBF. The 3D-assisted coarse-to-fine strategy and the 2D adjustment process explicitly ensure both the robustness to extreme face poses and bounding box disturbance and the accuracy towards pixel-level landmark displacement. Extensive experiments on the Menpo Challenge test sets demonstrate the superior performance of our system.*

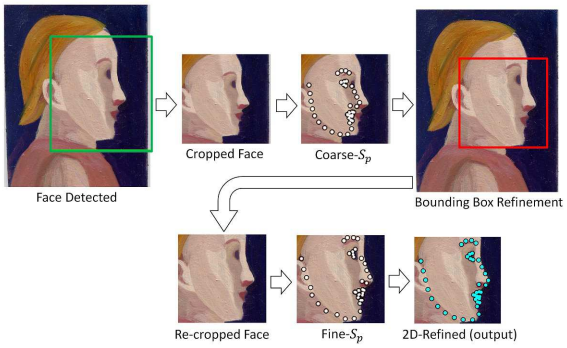


Figure 1: Flowchart of the proposed system. The detected face bounding box is first refined with landmark locations inferred from a coarse 3D face model. Based on the refined bounding box, our system fits a 3D face on re-cropped image at fine-level. The final landmark locations are obtained by adjusting the landmark locations, fine-S_p, with a 2D regression method.

1. Introduction

Facial landmark detection is a critical preprocessing step in many face related research works and applications since their performance is closely dependent on the accuracy of the predicted landmark locations. Recently 2D cascaded regression approaches [20, 14, 22] have demonstrated appealing facial landmark detection performance for face images under moderate conditions. However, robustness and accuracy of those approaches may drop significantly when the face images are with large poses and extreme expressions.

To solve the bottleneck of those 2D regression approaches, the 3D face model has been gradually employed in recent research [3, 10, 24, 11] to improve the robustness towards large poses and extreme expressions in facial landmark detection. The core idea is aligning a 3D morphable face model [2] for a given 2D face image. Since the 3D morphable face model (3DMM) can effectively capture the head pose and expressions, landmark detection inferred/initialized from a 3D face model inherits the merits of 3DMM and generally shows superior performance on large-pose face images.

In [3], binary shape-indexed pixel difference features are effectively extracted around the predicted landmark locations on the image and employed for 3D face regression. [3] shows desirable 3D face fitting performance for near-frontal face images. However, its performance is bottlenecked by the weak low-level features and drops for large-pose face images. In [24, 11], the 3D fitting parameters are predicted with cascaded convolutional neural networks. Those methods show remarkable performance enhancement of landmark detection on large-pose face images. They share a similar framework where the entire convolutional network is forward-propagated a few times to iteratively update the 3D fitting parameters. At each iteration, the input to the convolutional neural network is generated by concatenation

between the original face image and features generated with the predicted 3D face model. Extra computational resources are required for features preparation and forward propagation of the entire network. Therefore, we are motivated to seek for high-level deep features for 3D face regression without forward propagation of the entire network.

In [12], a deep shape-indexed feature extracting layer is introduced and utilized in [12, 18] which demonstrates much superior landmark detection performance over conventional features, *e.g.* SIFT, HOG and binary pixel difference features. Intuitively, such features can also be beneficial for the 3D face regression.

In order to enhance our system's robustness to face bounding boxes, we develop a 3D-assisted coarse-to-fine framework. 3D face model is first generated by utilizing deep shape-indexed features at coarse-level with the original face bounding box and used to provide a tighter bonding box. The refined bounding box is then used to re-crop the face image for further 3D face fitting and 2D landmark adjustment at fine-level. We observe that with the coarse-to-fine framework, the generated 3D face model can provide better initialization for 2D landmark refinement and hence leads to superior landmark detection performance.

Fig. 1 illustrates the flowchart of our landmark detection system. For a given image, our system first detects the face with a cascaded deep neural network-based face detector, 360-NUS[1], developed by our team. For the detected face, a coarse 3D face fitting process is performed with a Recurrent 3D Regressor (R3R) that takes deep shape-indexed features as input at each regression iteration. 2D landmark locations inferred from the coarse 3D face model are used to generate a tighter face bounding box. The original image is then re-cropped with the refined bounding box and passed into another R3R for fine-scale 3D face regression. The final landmark locations are obtained by adjusting the locations inferred from the fitted 3D face with LBF [14].

The main contributions of this paper are summarized as follows:

- A 3D-assisted coarse-to-fine facial landmark detection system is developed.
- A recurrent 3D regressor which requires only one-step forward propagation of the network is developed.
- Our system shows both strong robustness and high accuracy on face images under various challenging pose, expression and illumination conditions.

The remainder of the paper is as follows. We provide a review of related works in Section 2 before introducing our 3D-assisted coarse-to-fine facial landmark detector in Section 3. The experimental results are presented and discussed in Section 4. Then we conclude the paper in Section 5.

[1]Results of the "360-NUS" face detector can be found in the FDDB [9] website: http://vis-www.cs.umass.edu/fddb/results.html

2. Related Work

2.1. Deep Regression with Deep Shape-indexed Features

Deep features extracted around the predicted landmark locations from a feature layer are proposed recently [12] and utilized for iteratively refining landmark locations [12, 18]. The deep shape-indexed feature has shown significant performance enhancement as compared to the conventional hand-crafted features, *e.g.* HOG and SIFT [20, 23] and binary features [14, 5]. High-level discriminative shape information provided from deep shape-indexed feature makes it a good choice for 3D face regression as well. With the deep shape-indexed feature, our system only needs to forward propagate the entire network once during the refinement process.

2.2. 3D Morphable Face Model

3D Morphable Face Model (3DMM) [2] is used in this work to align the 3D face mesh to a given 2D face image. With 3DMM, a 3D face can be formally modelled by a set of expression and shape controlling parameters as follows:

$$\mathbf{A} = \bar{\mathbf{A}}_0 + \boldsymbol{\alpha}_{\mathrm{id}}\mathbf{A}_{id} + \boldsymbol{\alpha}_{\mathrm{exp}}\mathbf{A}_{exp}. \quad (1)$$

$\bar{\mathbf{A}}_0$ is the 3D mean face model. $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ are the PCA reconstruction basis for shape and expression accordingly. $\mathbf{A} = \{x_1, y_1, z_1; x_2, y_2, z_2, ..., x_N, y_N, z_N\} \in R^{3\times N}$ with N vertices. $\boldsymbol{\alpha}_{\mathrm{id}}$ and $\boldsymbol{\alpha}_{\mathrm{exp}}$ are the reconstruction coefficients for $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$. Following [24, 11], we use 29 expression bases and 199 shape bases to model $\mathbf{A}$. These bases are obtained with the Bessel Face Model [13] and Face Warehouse [4]. The 3D face $\mathbf{A}$ is fitted onto a 2D image via rotation, translation and projection. This is formulated as

$$\mathbf{M}(\mathbf{q}) = \mathbf{\Pi}_f\mathbf{R}_{\phi,\gamma,\theta}\mathbf{A} + \mathbf{t}_{2d}, \quad (2)$$

where $\mathbf{t}_{2d}$ is the translational vector in the 2D space and $\mathbf{R}_{\phi,\gamma,\theta}$ is the 3D rotation matrix formulated by the head pose, *i.e.* pitch, yaw and roll angles. $\mathbf{q} = \{\phi, \gamma, \theta, f, \mathbf{t}_{2d}, \boldsymbol{\alpha}_{\mathrm{id}}, \boldsymbol{\alpha}_{\mathrm{exp}}\}$ is the related parameters for 3D face fitting and 3D-to-2D projection. Assuming orthographic projection, the projection matrix can be described as

$$\mathbf{\Pi}_f = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \end{pmatrix},$$

where f is the scaling factor and $[c_x\ c_y]$ denotes the image center.

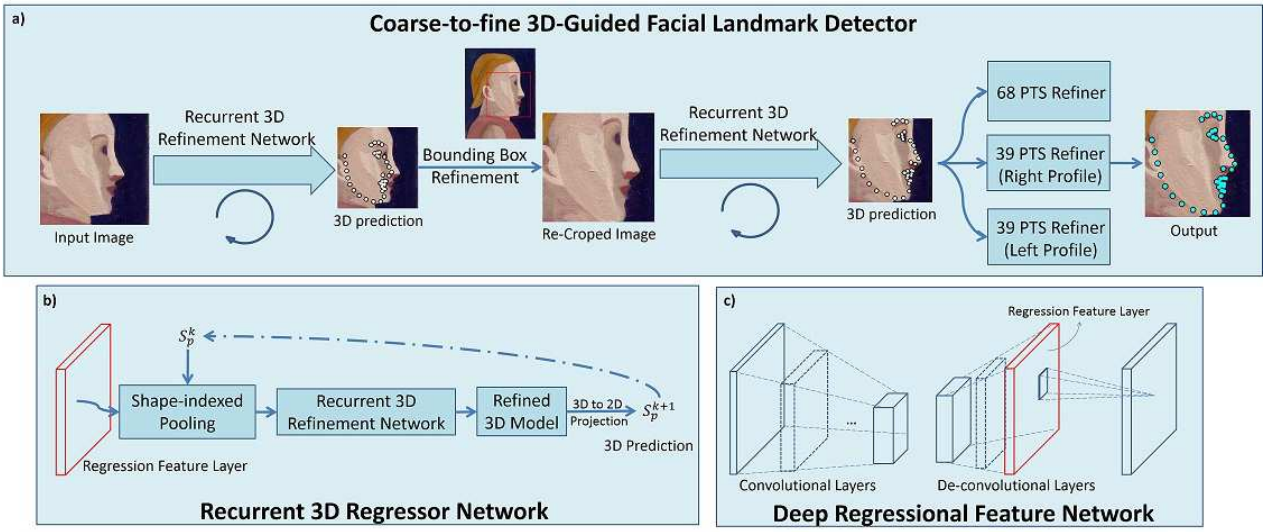


Figure 2: The flowchart of our 3D-assisted coarse-to-fine facial landmark detector. For the given face image (cropped with the bounding box given by our face detector), our system first predicts the 3D face model with a recurrent 3D regressor (R3R) network depicted as b) and estimates a tightly fitted face bounding box based on the 3D face model. The re-cropped image is then used to predict the 3D face model at fine-level. 2D refinement is employed to further adjust the landmark locations. The Deep Regressional Feature Network illustrated in c) prepares the feature layer for deep shape-indexed feature extraction [12] employed by R3R.

The facial landmark locations can be directly inferred from the fitted 3D face by extracting 2D locations of the key landmark vertices from $\mathbf{M}$ and formulated as

$$S_p = \mathbf{M}(\mathbf{q})^{\{\mathbf{v}\}} \in R^{2\times L} \tag{3}$$

where $\mathbf{M}(\mathbf{q})$ is the predicted 3D face model on the 2D plane with parameter set $\mathbf{q}$. $\mathbf{v}$ presents the correspondence between the 3D face model and 2D landmarks.

3. 3D-assisted Coarse-to-fine Landmark Detection

3.1. Overview

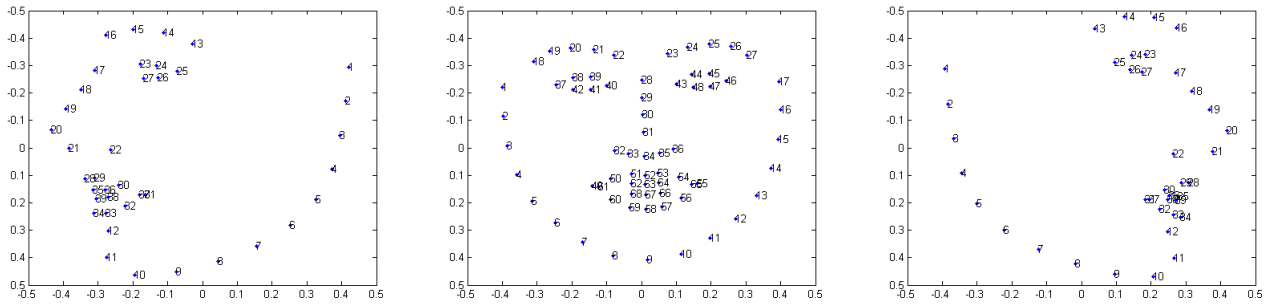
Fig. 2-a) shows the detailed framework of our landmark detection system. It takes a 2D image cropped with a coarse face bounding box provided by our face detector as input. The initial input image is passed into a recurrent 3D regressor (**R3R**) which generates a 3D face model and infers 2D landmark locations from the model. The estimated landmark locations are utilized to generate a face bounding box which tightly encloses the entire face region. Based on the refined bounding box, the original image is re-cropped and passed into another R3R which is designed for generating the 3D face model at fine-scale. 2D landmark locations inferred from the 3D model at fine-scale are further adjusted with the 2D refinement method to alleviate possible errors incurred due to inaccurate 3D fitting parameters and unperfect 3D-to-2D landmark correspondence. Details of our system are explained in this section.

3.2. Recurrent 3D Regression

We develop two networks to model the 3D face, Deep Regressional Feature Network (Fig. 2-c)) and Recurrent 3D Regressor Network (Fig. 2-b)). They are employed to iteratively and recurrently update the 3D face model and ensure that the fitted 3D face model captures the head pose and facial expression robustly even under extreme conditions. At each recurrent refinement iteration, the R3R extracts deep shape-indexed features [12] around landmark locations predicted at the previous iteration from Regression Feature Layer and updates 3D fitting parameters. This refinement process is formulated as

$$\mathbf{q}^k = \mathbf{q}^{k-1} + \mathrm{LSTM}(\mathbf{\Phi}^{k-1}, o^{k-1}, C^{k-1}) \tag{4}$$

where $\mathrm{LSTM}(\mathbf{\Phi}^{k-1}, o^{k-1}, C^{k-1})$ is the conventional long short-term memory unit [8] which takes features $\mathbf{\Phi}^{k-1}$ extracted around landmark locations S_p^{k-1} from the Regressional Feature Layer, previous memory output C^{k-1} and previous LSTM output o^{k-1} as inputs. $\mathrm{LSTM}(\mathbf{\Phi}^{k-1}, o_{k-1}, C_{k-1})$ outputs update to 3D fitting parameter $\Delta\mathbf{q}^k = o^k$ such that the following objective



(a) Left profile face (b) Frontal face (c) Right profile face

Figure 3: Landmark definitions by the Menpo benchmark

function is optimized:

$$\arg\min_{\mathbf{W}_s,\mathbf{b}_s} \|\mathbf{q}^* - (\mathbf{q}^{k-1} + \mathrm{LSTM}(\mathbf{\Phi}^{k-1}, o^{k-1}, C^{k-1})\|^2. \tag{5}$$

$\mathbf{W}_s, \mathbf{b}_s$ are weight and bias terms related to the LSTM unit and $\mathbf{q}^*$ is the ground-truth 3D fitting and projection parameters.

At each recurrent iteration, the landmark locations are directly inferred via

$$S_p^k = \mathbf{M}(\mathbf{q}^{k-1} + \Delta\mathbf{q}^k)^{\{\mathbf{v}\}} \tag{6}$$

where $\mathbf{v}$ represents the correspondence between the 3D face model and 2D landmarks. As the inner landmarks have fixed definitions to explicit facial components, *i.e.* eyes, eyebrows and mouth lips, the landmark vertices of these inner landmarks are fixed. However, 3D-to-2D correspondence of landmarks on the contour line is pose-variant. We follow [25] to find the corresponding contour landmark vertices on the 3D face. Once the correspondence between the 3D model and 2D landmark is established, the 2D landmark location can be easily inferred via Eqn. (3).

3.3. 2D Adjustment

Landmark locations inferred from 3D the face model may be inaccurate to small landmark displacement [3]. The inaccurate 3D-to-2D correspondence and prediction error of 3D fitting parameters directly deteriorates the locations of landmark inferred. The incurred localization error is at small scale and can hardly be captured by the 3D model.

To tackle this bottleneck related to the 3D face model, 2D refinement is employed. The objective of 2D refinement is formulated as

$$\arg\min_{\mathrm{R}_{2D}} \|S^* - (\hat{S}_p + \mathrm{R}_{2\mathrm{D}}(I, \hat{S}_p)\|^2 \tag{7}$$

where $\hat{S}_p = \mathbf{M}(\hat{\mathbf{q}})^{\{\hat{\mathbf{v}}\}}$ is the landmark locations obtained from the fitted 3D face defined by the estimated parameter $\hat{\mathbf{q}}$ via unperfect 3D-to-2D landmark correspondence $\hat{\mathbf{v}}$. R_{2D} represents the 2D refinement process which takes the original image as input and $\hat{S}_p$ as the initial prediction. Various 2D refinement methods can be used, *e.g.* SDM [20], ESR [5], LBF [14], etc. LBF is employed in this work.

3.4. 2D Landmark Refinement of Profile Face Images

The profile and semi-frontal categories are manually determined by the competition organizer and we do not need to classify a test image into semifrontal/profile categories. Unlike the images from the semi-frontal category[2], face images from the profile category are annotated with 39 landmarks defined in Fig. 3a and Fig. 3c. The contour landmarks have completely different physical meaning for the left and right profile faces. To ease the refinement process, we divide 2D refinement for faces from the profile category into two subtasks. To be more explicit, left and right profile face 2D refinement modules are designed to handle the two profile cases correspondingly and independently. Based on the estimated 3D pose of the face image, our system automatically selects the correct 2D profile landmark refinement task. The initial 39 landmark locations to be refined are obtained via one step linear projection of predicted 68 landmarks $\hat{S}_p$ by following the method in [7].

3.5. System Training

Our system consists of two recurrent 3D regressor (R3R) networks which are trained jointly. The coarse-R3R module is trained with a set of training images with large perturbation in face angles, scale and translation. The fine-R3R network is trained with face images re-cropped based on the landmark locations obtained from the 3D face

[2]The Menpo Challenge [21, 16] divides testing set into two categories, *i.e.* semi-frontal and profile.

Algorithm 1: Recurrent 3D Regression.

Inputs: Regressional Feature Layer L_{feat}; 3D parameters: $\mathbf{q}^*$; refinement steps: K
Initialization: $\mathbf{q}^0 = \bar{\mathbf{q}},\ S_p^0$.
while $k \leq K$ **do**
  Get deep shape-indexed feature $\mathbf{\Phi}(S_p^{k-1}, L_{feat})$
  Optimize recurrent 3D regression unit via Eqn. (4)
  Obtain 3D parameter update:
    $\Delta\mathbf{q}^k = \mathrm{LSTM}(\mathbf{\Phi}(S_p^{k-1}, L_{feat}), o^{k-1}, C^{k-1})$
  Update 3D parameters:
    $\mathbf{q}^k = \mathbf{q}^{k-1} + \Delta\mathbf{q}^k$
  Infer 2D landmark locations: S_p^{k-1} via Eqn. (3)
end
Outputs: $S_p^K, \mathbf{q}^K$

model generated with the coarse-R3R. The training process for both modules is same and described in Algorithm 1. We follow [12] to train the Regression Feature Layer. The Regressional Feature Layer is trained jointly with R3R and it is fed with cropped face images.

In this paper, LBF [14] is employed to refine the landmark locations directly inferred from 3D face model. We initialize the LBF regressor with S_p^K instead of the conventional mean shape $\bar{S}$. This ensures the LBF to have a good initial shape and focus on capturing pixel-level landmark displacement.

4. Experiment

4.1. Experimental Data

Training Data Preparation. We use the 300W [15], 300W-LP [24] and Menpo training set provided by the 2nd Facial Landmark Detection Challenge [21] for training. 300W-LP [24] is an extension of 300W with original 300W face images morphed into large poses. Menpo training set consists of 6,679 semi-frontal face images annotated with 68 landmarks and 2,300 profile face images annotated with 39 landmarks. We perform flipping, shifting, scaling and rotation operations with the face images. The ground-truth 3D reconstruction parameters $\mathbf{q}^*$ are generated by following [24, 11].

Evaluation Datasets. Our landmark detection performance is evaluated on both semi-frontal and profile categories.

Semi-frontal category: This category consists of 12,006 near-frontal face images in the wild. All images from the semi-frontal category are annotated with 68 landmarks defined as Fig. 3b. There is substantial number of challenging face images, *i.e.* large poses, extreme expression and heavy occlusion, within this category.

Profile Category: The profile category contains 4,253 profile face images which are annotated with 39 landmarks defined in Fig. 3a and Fig. 3c. All images are taken under unconstrained conditions.

4.2. Face Detection

Our system employs "360-NUS" face detector to localize the face inside a given image. For images with multiple detected faces, we select the one closer to the image center for landmark detection. We manually check the detected faces and find there are 405 false detections among the 16,239 testing images from both categories. The false detections are manually corrected. "360-NUS" ranks top in the FDDB [9] benchmark and can more effectively reduces false/multiple detections compared to Viola-Jones method.

4.3. Evaluation Metric

The accuracy of landmark detection is measured by the normalized point-to-point mean error. In this competition, inter-ocular distance is used to normalize the absolute distance error for the semi-frontal testing set. The normalized point-to-point landmark detection error for a face image can be formulated as

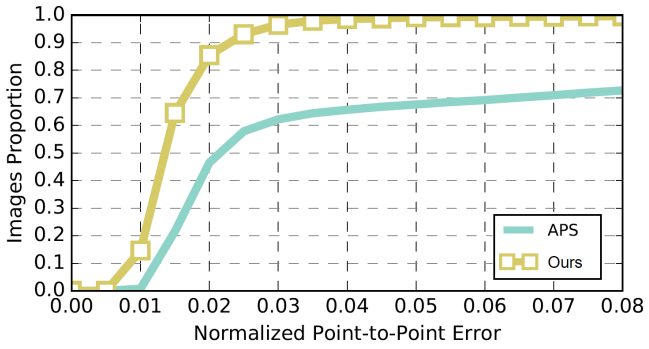
$$E_{p2p} = \frac{\sum_i^L \|S_i^* - \hat{S}_i\|_2}{LD_N} \tag{8}$$

where L is the number of landmarks and D_N represents the normalization factor. S_i^* and $\hat{S}_i$ represent the ground-truth and predicted locations of the i-th landmark. In the semi-frontal category, the normalization factor is defined by $D_N = \|S_{37}^* - S *_{46} \|_2$ with S_{37} and S_{46} being the outer corners of the left and right eyes. The normalization factor used for the profile category is unknown to the participants. For validation purpose, we define $D_N = \|S_{11} - S_{16}\|_2$. Readers may refer to Fig. 3 for the landmark definitions.

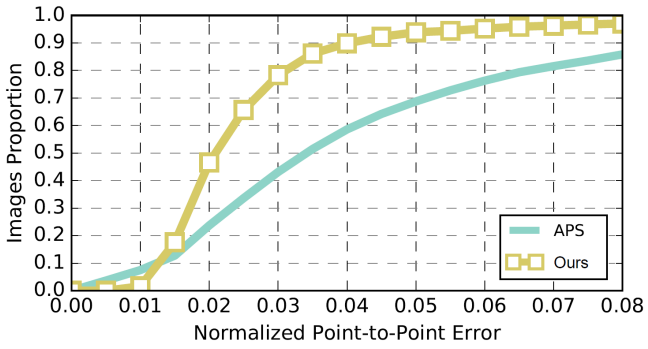
The Cumulative Error Distribution (CED) curves are given in Fig. 4 and 5. The CED curves are plotted up until 0.08. The X-axis represents the normalized error and the Y-axis represents the proportion of images.

4.4. Network Structure

We use the our customized ResNext network [19] as the initial model. The network structure is shown in Table 1 and it is pre-trained on the ImageNet dataset [6]. The input is RGB images with fixed 256×256 dimension. "de-conv4","de-conv5" and "Regressional Feature Layer" are de-convolutional layers which perform up-sampling. The output of the regressional feature layer has dimension of $256 \times 256 \times 64$. The Regressional Feature Layer is connected to a convolutional layer which predicts the locations of 68 landmarks directly via softmax regression



(a) Results on the semi-frontal category



(b) Results on the profile category

Figure 4: Evaluation on the semi-frontal and profile categories.

Table 1: The network structure used for regressional feature preparation.

stage	output	Operations
conv2	128x128	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, \text{R=32} \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3	64×64	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, \text{R=32} \\ 1\times1, 256 \end{bmatrix} \times 4$
conv4	32×32	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, \text{R=512} \\ 1\times1, 1024 \\ 1\times1, 256 \end{bmatrix} \times 6$
de-conv4	64×64	4x4,128, stride 2, de-conv
de-conv5	128×128	4x4,64 stride 2, de-conv
Regression Feature Layer	256×256	4x4, 64, stride 2, de-conv
Loss Layer	256×256	3x3, 68, conv

and each channel of the Loss Layer only responds to the location of a typical landmark. In the testing process, the loss layer is removed. The number of refinement stages within an R3R module is set to 5. The deep shape-indexed features extracted are reduced to dimension of 256 before passed into the LSTM unit.

Our model is trained via the standard stochastic gradient descent method with a momentum of 0.9, a mini-batch of 8 images and a weight decay parameter of 0.0001. The weights of LSTM are randomly initialized with a uniform distribution of $[-0.1,\ 0.1]$.

4.5. Results

Landmark locations predicted for both categories are submitted to the organizers. All predictions are evaluated by the organizers with ground-truth annotations which are unknown to the participants. The returned landmark detection errors are compared with the baseline method, Active Pictorial Structures (APS) [1] which employs a generative deformable model with pictorial structure and active appearance model. Viola Jones [17] face detector is used for face localization in the baseline method.

4.5.1 Semi-frontal Category

Fig. 4a shows that our system significantly outperforms the APS method. Over 95% of testing images are within 0.03 normalized point-to-point error for our system. This is about 50% performance enhancement as compared to the baseline. Our system has a median error of 0.013 which is around 56% of the median error of the baseline method.

4.5.2 Profile Category

Compared with the baseline method, our system shows consistently superior performance on the profile category. Over 75% of the testing images are within 0.03 normalized point-to-point error. Compared with 45% testing images within 0.03 error, ours achieves about 66% enhancement over the APS method. The median error of our system is 0.021 which is about 60% of the median error of the baseline method.

4.5.3 Menpo Validation Set

We randomly select 697 images from the semi-frontal training set and 200 images from the profile training set for validation. These images are not involved in the training process. Performance of the landmark detected with coarse-R3R+LBF and Coarse-to-fine-R3R + LBF is evaluated and shown in Fig. 5. It shows that the coarse-to-fine strategy performs consistently and significantly better than the coarse-R3R+LBF framework on both semi-frontal and profile categories.

4.6. Qualitative Analysis

In Fig. 7 and Fig. 8, sample face images from the semi-frontal and profile categories are shown with

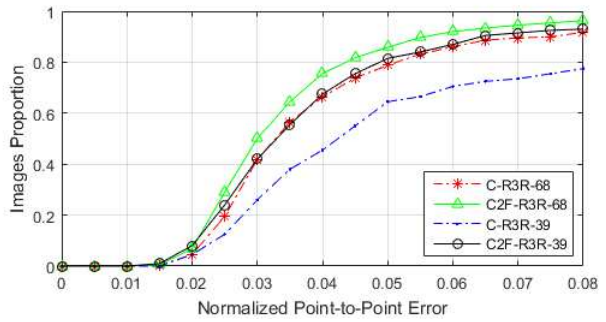


Figure 5: Results on the validation sets. "C-R3R" represents results of 2D adjustment on the the coarse-R3R module. "C2F-R3R" represents results of 2D adjustment on the Coarse-to-fine-R3R module. "-68" and "-39" are results on the semi-frontal and profile validation sets accordingly. (Best viewed in color)

landmarks detected by our system.

Fig. 7 shows face images under various challenging conditions, *i.e* heavy occlusion, large poses, extreme expressions, poor lighting and low image resolution. Our system can robustly overcome all these situations and demonstrates high accuracy. The first row of Fig. 7 shows a few sample face images under heavy occlusion. Desirable location accuracy of occluded landmarks is assured by the R3R which robustly fits a 3D face model and provides strong shape prior for 2D refinement, even though our system does not explicitly handle occlusion. The merits of R3R are also verified by the outstanding landmark detection performance for the faces with low resolution (last row of Fig. 7) and poor illumination (4-th row of Fig. 7). Robustness and accuracy of our system are further verified by the landmark detection performance on the profile face category in Fig. 8.

Failure Cases. Fig. 6 shows a few samples on which our system fails. These are images with extreme head roll angles and heavy occlusion on the profile faces. When the face is upside-down (1st image from Fig. 6) or the refined-bounding box is too large/small (3rd and 4th images from Fig. 6), our system hardly localizes the landmarks. This is possibly because the R3R fails at estimating the head pose and leads to large initial error for the 2D adjustment. For some heavily occluded profile faces, our system can roughly capture the head poses, but with much lower accuracy.

5. Conclusion

In this work, we developed a 3D-assisted coarse-to-fine extreme-pose facial landmark detection system which shows strong robustness and high accuracy for face images under various challenging conditions. A recurrent 3D regressor is developed to fit a 3DMM model to the face image in a coarse-to-fine manner. Landmark locations inferred from the 3D face model are further adjusted with 2D refinement to further reduce the estimation error inherited from 3D parametric error and imperfect 3D-to-2D landmark correspondence. The coarse-to-fine method shows superior performance over one stage coarse-3D regression framework and demonstrates significant enhancement over the baseline method.

Figure 6: Failure cases from both categories. (Best viewed in color)

6. Acknowledgement

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

References

[1] E. Antonakos, J. Alabort-i Medina, and S. Zafeiriou. Active pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5435–5444, 2015. 6

[2] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 1, 2

[3] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014. 1, 4

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 2, 4

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 5

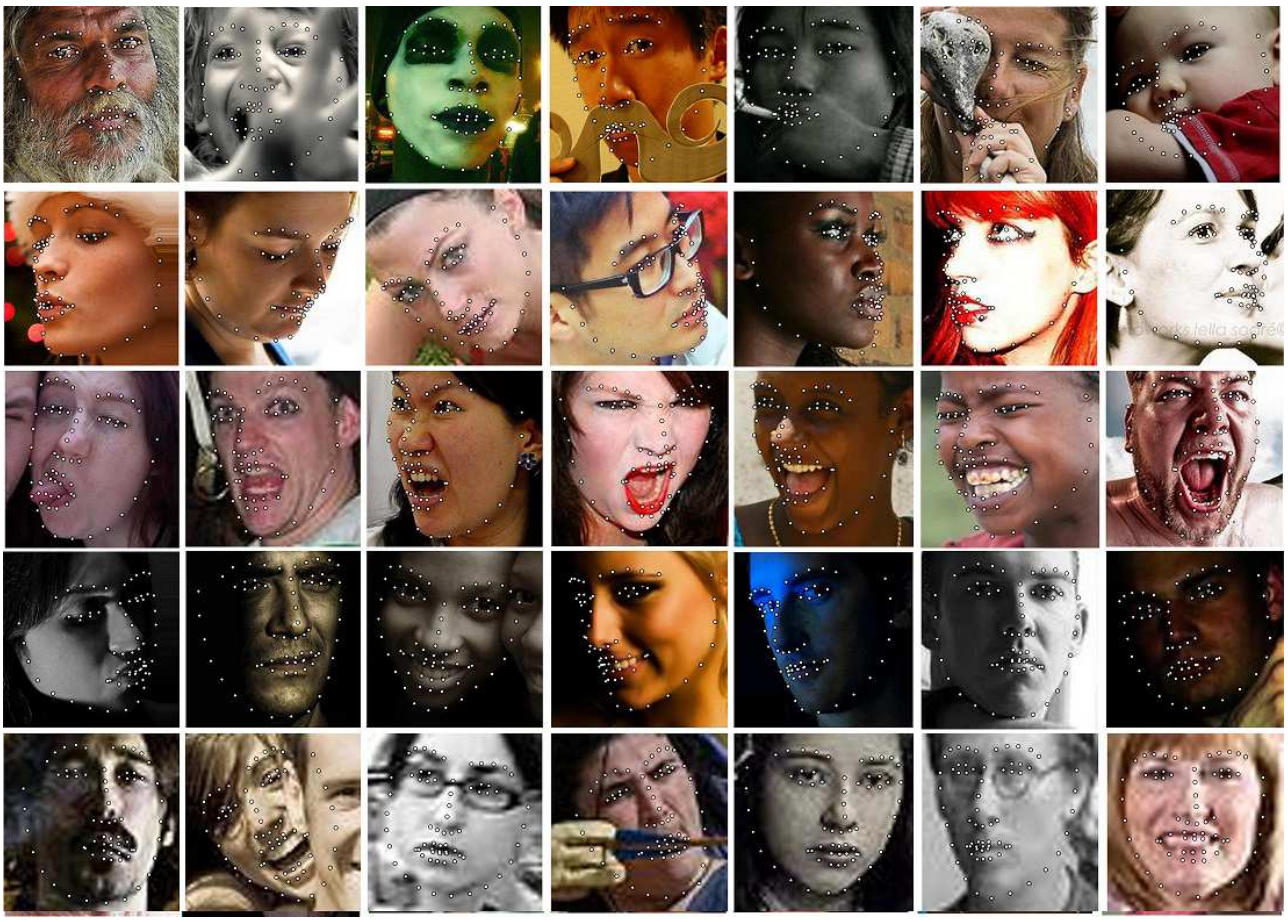


Figure 7: Some sample results on images from the semi-frontal benchmark are shown here. From the first to fifth row, face images with heave occlusion, large poses, extreme expression, challenging illumination and low resolution are shown accordingly. Our model can robustly and accurately handle all this challenging conditions. (Best viewed in color)

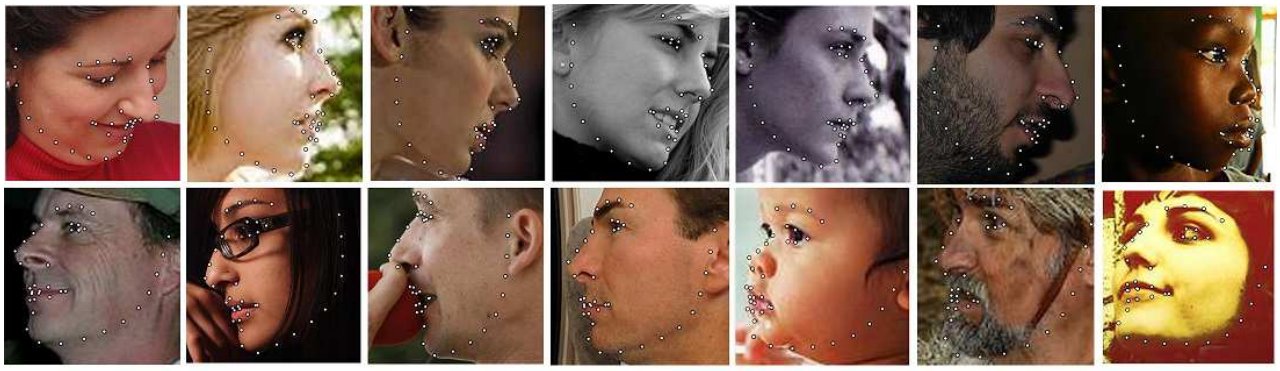


Figure 8: Sample results of profile category face images. (Best viewed in color)

[7] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2014. 4

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[9] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst,

2010. 2, 5

[10] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015. 1

[11] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 5

[12] H. Lai, S. Xiao, Z. Cui, Y. Pan, C. Xu, and S. Yan. Deep Cascaded Regression for Face Alignment. *ArXiv e-prints*, Oct. 2015. 2, 3, 5

[13] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009. 2

[14] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 1, 2, 4, 5

[15] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 5

[16] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016. 4

[17] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 6

[18] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision*, pages 57–72. Springer, 2016. 2

[19] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 5

[20] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 1, 2, 4

[21] S. Zafeiriou. The menpo facial landmark localisation challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 4, 5

[22] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proceedings of European Conference on Computer Vision*, pages 1–16. 2014. 1

[23] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 2

[24] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 5

[25] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3d morphable model fitting. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015. 4