Recurrent 3D-2D Dual Learning for Large-pose Facial Landmark Detection

Shengtao Xiao[1], Jiashi Feng[1], Luoqi Liu[2], Xuecheng Nie[1], Wei Wang[3], Shuicheng Yan[2,1], Ashraf Kassim[1]
[1]National University of Singapore, [2]Qihoo-360, [3]University of Trento
xiao_shengtao@u.nus.edu, elefjia@nus.edu.sg, liuluoqi@360.cn, niexuecheng@u.nus.edu
wei.wang@unitn.it {eleyans,ashraf}@nus.edu.sg

In this supplementary material, we provide details on how RDR works for large-pose 2D landmark detection and its network architecture.

1. 3D to 2D Landmark Projection

How accurate the following projection relationship can capture the true correspondence between 2D and 3D is important for performance of the 2D landmark detection,

$$S_p = \mathbf{M}(\mathbf{Q})^{\{\mathbf{v}\}} \in \mathbb{R}^{2\times L} \tag{1}$$

where $\mathbf{M}(\mathbf{Q})^{\{\mathbf{v}\}}$ are 2D landmark locations via direct 3D-to-2D projection from 3D face generated with 3D reconstruction parameters $\mathbf{Q}$. Here $\mathbf{Q} = [\boldsymbol{\alpha}_{\text{id}}, \mathbf{P}, \boldsymbol{\alpha}_{\text{exp}}]$. $\mathbf{P} = [\phi, \gamma, \theta, \mathbf{t}_{3d}, f]$ denotes the pitch, yaw and roll rotation angles, translation vector and focal length of the projection operation. We use $\boldsymbol{\alpha}_{\text{id}}$ to denote the identity reconstruction coefficients and $\boldsymbol{\alpha}_{\text{exp}}$ for the coefficients for expression blendshapes. Within Eqn. (1), $\mathbf{v} = [v_s^1, \ldots, v_s^L] \in \mathbb{R}^L$ indexes those vertices corresponding to L landmarks on the 3D face. 3D-to-2D landmark projection is to find the accurate correspondence between the 3D vertices and 2D landmarks. Correspondence between the 3D vertices of the interior facial components (*e.g.*, eyebrows, eyes, nose and mouth) and the 2D landmarks can be fixed as they are pose invariant (independent of $\mathbf{P}$).

However, the correspondence between the 3D contour vertices and the 2D contour landmarks may have much larger variance and would change dramatically with the face pose [2, 15]. To find the correspondence, Zhu *et al.* [15] proposed to use parallel lines to help locate the corresponding contour landmarks. When the pose varies, landmarks will move along their corresponding parallel lines to their visibility boundary. An example is given in Fig. 1. The visibility boundary is simply the extreme of x coordinates (the minimal x among the left contour points and the maximal x among the right contour points). This ensures the efficiency and quality of searching for correspondence points due to limited vertices on each parallel line.

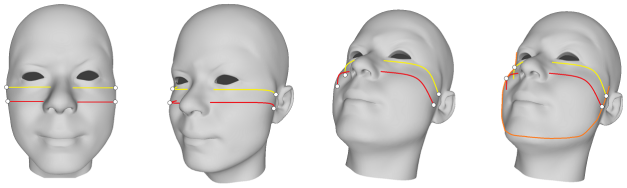

Figure 1. We improve the correspondence between 3D vertices and 2D landmarks by using both parallel lines and the exact contour line to locate corresponding vertices of contour landmark points. In this figure, parallel lines (red and yellow) of 4 contour points are drawn for illustration purpose. The key vertices are moving on their corresponding parallel lines. [15] directly use the extreme coordinates as the contour line (the first three figures). This may fail when the lower face region occludes the upper face region as shown in the third figure. To improve the robustness, we extract the contour line (the fourth figure) first. The contour landmarks are those vertices appear on both contour line and parallel lines. (Best viewed in color)

The above method relies on a critical assumption that extreme coordinates are always on the visibility boundary. However, the assumption may not hold when there is both large pitch and yaw angles, as illustrated in Fig. 1. In this work, we further improve the parallel line based method [15] by extracting the face contour line using a similar method as [10]. For a given 3D frontal face reconstructed with predicted $\boldsymbol{\alpha}_{\text{id}}$ and $\boldsymbol{\alpha}_{\text{exp}}$, we rotate the face based on the estimated pitch and yaw angles first. Roll angle is not considered here since the correspondence between the 3D vertices and the 2D landmarks is invariant to roll rotation. Contour line (orange line from the fourth figure from Fig. 1) is then extracted from the rotated 3D face. The parallel lines are employed to find the face contour landmarks by looking for the intersection between the extracted contour line (orange) and the parallel lines of contour landmarks (red and yellow). Since the rest landmarks for interior facial parts have fixed corresponding vertices, they are selected from the 3D model directly. Readers may also refer to [15] for details of how the parallel lines are defined.

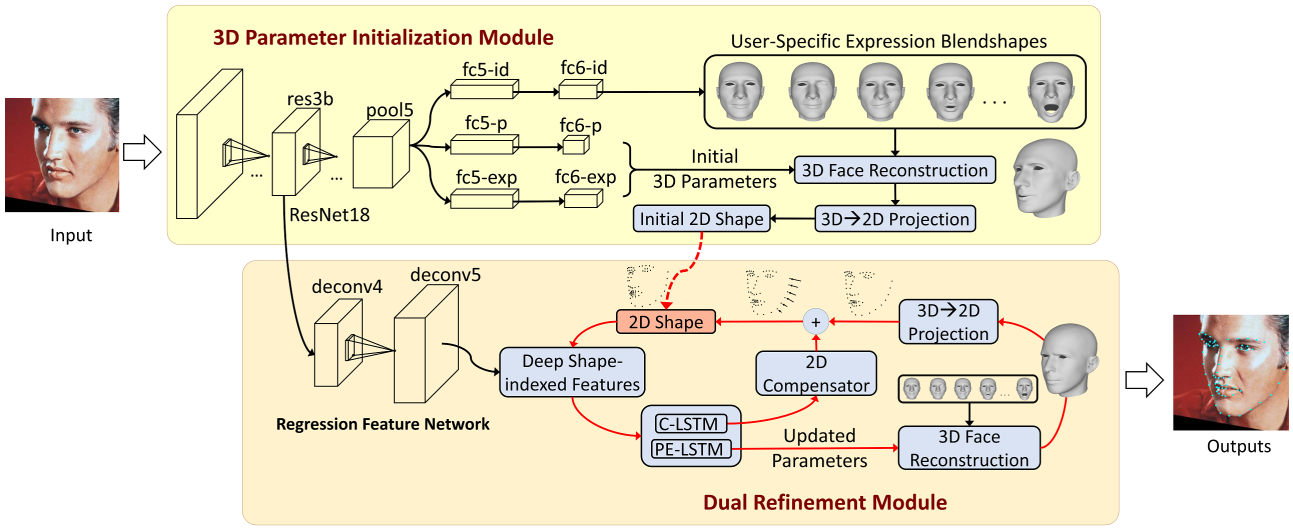

Figure 2. Overview of the proposed RDR model for large-pose facial landmark detection. Given an input 2D face image, RDR first directly predicts initial 3D face fitting parameters with the 3D parameter initialization module, generates an initial 3D face mesh and infers initial 2D landmark locations via 3D-to-2D projection. It then recurrently refines both the 3D face and 2D landmark locations with a dual refinement module consisting of a 3D face refinement component (PE-LSTM) and a 2D landmark refinement component (C-LSTM) with deep features directly extracted from the regression feature network.

2. Architecture Details of RDR

Architecture of RDR is given in Fig. 2. It consists of a 3D parameter initialization module, a regression feature network and a dual refinement module. The parameter initialization module is built on a variant of ResNet-18 [4]. All layers after pool5 from the original ResNet-18 are removed. The regression feature network is designed to provide features for dual refinement of both 3D face model and 2D landmark locations.

Within the 3D parameter initialization module, pool5 is connected to the last convolution layer of ResNet-18 [4] whose output dimension is $512 \times 4 \times 4$. Three independent fc5 layers, as shown in Fig. 2, for predicting three types of parameters, *i.e.* fc5-p, fc5-id and fc5-exp, are connected to pool5 and they have output size of 128, 256 and 512 respectively. Outputs of fc5 layers are directly fed into their corresponding fc6 layers. The fc6-p, fc6-id and fc6-exp have output dimension of 7, 50 and 46 and predict $\mathbf{P}$, $\boldsymbol{\alpha}_{\text{id}}$ and $\boldsymbol{\alpha}_{\text{exp}}$ respectively.

The regression feature nework takes output from Res3b as input and passes it through a de-convolutional layer (deconv4) with kernel size 4, stride 2 and output dimension 48. The resulted $48 \times 64 \times 64$ response map is then fed into another de-convolutional layer (deconv5), giving a response map of $80 \times 128 \times 128$. Within the architecture, conv6 (not shown in Fig. 2 since it only appears in the training process) is a softmax regression layer which takes outputs of deconv5 as input. It produces the feature map of $68 \times 128 \times 128$ — each channel is only responsive to the location of a specific landmark. By introducing the last convolutional layer, the features learnt in deconv5 can therefore provide effective high-level discriminative feature for dual refinement of both 3D face model and 2D landmark locations.

The deep shape-indexed features extracted around the previously predicted 2D landmark locations from deconv5, $\Phi(\mathrm{D}_{deconv5}, S^{k-1}) \in \mathbb{R}^w$, are passed through two fully connected layers to generate final inputs for the two refinement components (PE-LSTM and C-LSTM) within the dual refinement module. Here, we use $S^{k-1} \in \mathbb{R}^{2\times 68}$ to denote the 2D landmark locations predicted in previous iteration. $w = 80 \times 68$ is the dimension of the deep shape-indexed features since our model predicts locations of 68 landmarks and deconv5 has 80 channels. Each fully connected layer reduces the deep shape-indexed features to dimension of 256 during the dual refinement process. This effectively reduces the model size. PE-LSTM is designed to refine the 3D parameters $\mathbf{P}, \boldsymbol{\alpha}_{\text{exp}}$ and hence has output dimension of 53. C-LSTM has output dimension of 136 and is designed to refine the 2D landmark locations.

3. Additional Experimental Results

Due to limited space, landmark detection results of our model on AFLW-PIFA are presented here. AFLW-PIFA consists of 3,901 training images and 1,299 testing images. Face images from ALFW-PIFA [5] are originally annotated with 21 landmarks in [7] and extended to 34 landmarks in [6]. Our originally trained model is evaluated on both settings. Table 1 and Table 2 both show that our method outperforms the current state-of-the-art methods.

Table 1. Landmark detection results of AFLW-PIFA with 21 landmarks. Results are obtained from [13].

Method	NME
CDM [11]	8.59
RCPR [1]	7.15
CFSS [12]	6.75
ERT [3]	7.03
SDM [9]	6.96
PIFA [5]	6.52
CCL [13]	5.81
RDR (Ours)	**4.07**

Table 2. Landmark detection results of AFLW-PIFA with 34 landmarks. Results are obtained from [6].

Method	NME
RCPR [1]	6.26
PIFA [5]	8.04
D3PF [6]	4.72
RDR (Ours)	**4.11**

ALFW2000-3D [14] is another dataset formed with images from AFLW. Face images from AFLW2000-3D are annotated with 68 landmarks with the inner landmarks have same definition as that of 300-W [8]. The face contour landmarks from [14] are on the absolute jawline.

Our trained 3D model gives NME of 5.36 on the AFLW2000-3D dataset which is better than 3DDFA (5.42) [14]. Our 2D model (3D+2D) has NME of 4.93, better than 3DDFA+SDM (4.94) [14]. The performance enhancement is not very significant on the AFLW2000-3D dataset. This is possibly because our model is trained on the conventional 68 landmarks [8] which is different from AFLW2000-3D setting with contour landmarks annotated on the absolute jawline. Its performance can be further improved via fine-tuning on the AFLW2000 landmark setting. When only evaluating on the 51 landmarks (removing the contour landmarks), our model gives NMEs of 3.69 and 3.25 for 3D and 2D landmark predictions respectively. Such error is pretty low.

References

[1] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520, 2013. 3

[2] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 1

[3] 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. 3

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2

[5] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015. 2, 3

[6] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196, 2016. 2, 3

[7] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011. 2

[8] 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. 3

[9] 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 (CVPR)*, pages 532–539, 2013. 3

[10] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas. Expression flow for 3d-aware face component transfer. In *ACM Transactions on Graphics (TOG)*, volume 30, page 60. ACM, 2011. 1

[11] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1944–1951, 2013. 3

[12] 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 (CVPR)*, pages 4998–5006, 2015. 3

[13] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3409–3417, June 2016. 3

[14] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, June 2016. 3

[15] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015. 1