Cross-Cascading Regression for Simultaneous Head Pose Estimation and Facial Landmark Detection

Wei Zhang[1,2,3], Hongwen Zhang[1,2], Qi Li[1], Fei Liu[1], Zhenan Sun[1,2], Xin Li[4], and Xinxin Wan[4(✉)]

[1] Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{wei.zhang,hongwen.zhang}@cripac.ia.ac.cn, {qli,fei.liu,znsun}@nlpr.ia.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
[4] The National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
{lixin,wanxx}@cert.org.cn

Abstract. Head pose estimation and facial landmark localization are crucial problems which have a large amount of applications. We propose a cross-cascading regression network which simultaneously perform head pose estimation and facial landmark detection by integrating information embedded in both head poses and facial landmarks. The network consists of two sub-models, one responsible for head pose estimation and the other for facial landmark localization, and a convolutional layer (channel unification layer) which enables the communication of feature maps generated by both sub-models. To be specific, we adopt integral operation for both pose and landmark coordinate regression, and exploit expectation instead of maximum value to estimate head pose and locate facial landmarks. Results of extensive experiments demonstrate that our approach achieves state-of-the-art performance on the challenging AFLW dataset.

Keywords: Facial landmark detection · Head pose estimation
Cross-cascading regression · Integral regression
Deep convolutional network

1 Introduction

Head pose estimation and facial landmark localization have drawn much attention from computer vision community as they are of great significance and broad applications in problems such as face verification, face animation, and emotion recognition.

J. Zhou et al. (Eds.): CCBR 2018, LNCS 10996, pp. 148–156, 2018.
https://doi.org/10.1007/978-3-319-97909-0_16

Traditionally, head pose estimation and facial landmark localization are treated as independent problems and seldomly be studied jointly.

Thanks to the development of Deep Convolutional Neural Networks (DCNN), there has been significant progress on both head pose estimation and facial landmark localization [1–3] and recent methods generally adopt DCNN as their main building blocks. One of the major advantages of DCNN is its capability of performing end-to-end optimization, especially for multitask problems [4] where related tasks can benefit from each other. Facial landmark detection algorithms could be roughly classified into two categories, detection based methods and regression based methods. At present, most best performing methods are detection based, in which heatmaps indicating the probability of the precense of the facial landmarks are generated and the exact locations of landmarks are determined according to maximum likelihood. However, since the operation of taking maximum value is not differentiable, it breaks the back propagation chain required for end-to-end learning. Intuitively, head pose estimation and facial landmark detection are not isolated problems and low-level facial representations could be shared by the two objectives, thus they attract the attention of many researchers [5,6].

The motivation of this work is to integrate information from head pose and facial landmarks for improving the performance of both facial landmark detection and head pose estimation on arbitrary faces, taking advantages of DCNN. In this work, we propose a novel network architecture named Cross-Cascading Regression network which integrates information from both pose and landmarks, and simultaneously perform head pose estimation and facial landmark detection. Since our network structure is topological symmetric, we expand a single network module by consecutively appending multiple modules together at the end which achieves finer prediction.

To overcome the obstacle of non-differentiable operations, we adopt integral regression [7], and use expectation instead of maximum value to locate landmarks. The loss of the network consists of two components: classification and regression.

The proposed method achieves comparable or better results in comparison with state-of-the-art algorithms on the challenging dataset AFLW [8] for both head pose estimation and facial landmark detection. With more blocks stacked, the performance improves significantly.

2 Related Works

In this section, we introduce some related works in facial landmark localization and head pose estimation. Traditionally, these two problems are addressed as independent problems.

Facial Landmark Localization. There are two distinct families of methods for facial landmark localization: detection based and regression based methods. Detection based methods handle facial landmark detection as a heat map prediction problem, and many explorations have been made such as stacked

architectures, residual connections, and multiscale processing. Newell et al. [9] proposed the Stacked Hourglass Network, which incorporates multi-resolution features and improves scores on 2D pose estimation challenges significantly. On the other hand, facial landmark detection is essentially a regression problem. Typically, regression based methods use cascaded regressors to predict landmarks' coordinates directly from intensities of input images. Cao et al. [10] used a vectorial regression function to infer the whole facial shape from the input. Xiong et al. [11] proposed a Supervised Descent Method (SDM) for minimizing a Non-linear Least Squares (NLS) function to optimize the performance of facial feature detection. Although regression based methods have been widely used, the performance is still not satisfactory. The idea that using information from different tasks to constrain the solution space is also a optional approach to achieve better results. Zhang et al. [5] trained a multi-task network which optimizes facial landmark detection together with correlated tasks such as head pose estimation and facial attribute inference. Huang et al. [6] proposed a unified FCN framework named DenseBox to accomplish landmark localization and face detection simultaneously. Wu et al. [12] propose an iterative cascade method for simultaneous facial landmark detection, head pose estimation, and facial deformation analysis.

Head Pose Estimation. Head pose estimation usually serves as a by-product of facial landmark detection, which means the precision of head pose estimation is closely related to the accuracy of landmark detection. However, extremely relevant information can disturb prediction precision. It also fails to make the utmost of facial information. The research of independent head pose estimation is rare. Nataniel et al. [3] trained a multi-loss convolutional neural network on 300W-LP to estimate pose directly from input image through joint binned pose classification and regression.

3 Approach

In this section, we present the technical details of Cross-cascading Regression Network. The proposed model consists of two sub-networks, which performs head pose estimation and facial landmark localization simultaneously with intermediate facial feature sharing. Specifically, the network takes a face image as input, and outputs heatmaps where each per-pixel indicates the likelihood for locations of key points. Meanwhile, it outputs three float numbers which indicate the degrees of yaw, pitch and roll, and a combination of information maps for further processing.

3.1 Head Pose Estimation

The pose estimation sub-network aims at getting appraisals of three Euler angles Y, P and R (Y denotes yaw, P denotes pitch and R denotes roll). Since the range of head poses is divided into N classes, we adopt a N-way softmax layer at

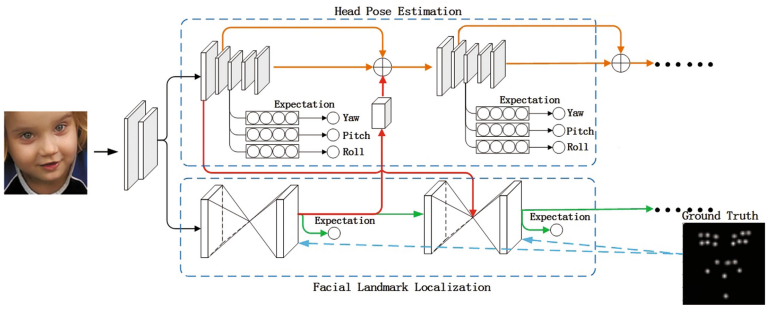

Fig. 1. Our cross-cascading regression network consists of head pose estimation sub-network and facial landmark localization sub-network

the top of the sub-network, generating the probability distribution of the head pose in the input image over N classes.

Instead of inferring head pose from the estimated landmarks, we directly predicted intrinsic Y, P, R from image intensities through joint binned pose classification and regression [3], which avoids irrelevant information damaging the prediction accuracy so that the module has greater robustness.

For network training, an cross-entropy loss is employed:

$$\mathcal{L}_{pc} = -\sum_{p} y_p \log \hat{y}_p \tag{1}$$

where y_p is the target probability distribution of head pose, and $\hat{y}_p$ is the predicted head pose probability distribution.

Inspired by [3], we also add a regression loss to improve the performance of head pose prediction, which is the Mean Square Error between the predicted pose and ground truth. The total loss of pose estimation sub-network is:

$$\mathcal{L}oss_p = \mathcal{L}_{pe} + \alpha_1 \mathcal{L}_{pc} = \sum_{k=1}^{3} \left\| Q_k - \hat{Q}_k \right\|_2 + \alpha_1 \mathcal{L}_{pc} \tag{2}$$

where α_1 is the balance factor, k indicates the k_{th} pose, Q_k and $\hat{Q}_k$ refers to the predicated and ground truth pose, respectively.

The pose estimation sub-network is built upon ResNet50 [13], with three fully-connected layers appended at the end to predict each angle independently. Pervious convolutional layers of the backbone network are shared by all of these fully-connected layers. By enabling back-propagation of the regression results of head pose angles, network learns to obtain fine-grained pose predictions.

3.2 Facial Landmark Localization

The design of the facial landmark localization sub-network is based on the Hourglass Networks [9] which has shown outstanding results on human pose estimation. We adapted the idea to the case of facial landmark localization. The output

of the sub-network are k heatmaps, and each heatmap H_k indicates the probability of the presence for the k_{th} key point.

Several convolutional and max pooling layers process the input image down to a very low resolution (4×4, for example). At the end of down-sampling operations, the network begins the top-down sequence of upsampling. In this procedure, features across different scales are combined together. After reaching the output resolution, we applied two 1-demention convolutions to get the final prediction, which is a set of heat maps.

In general, the final joint location coordinate is obtained as the location with the maximum value in a learnt heatmap. However, obtaining the location possessing the maximum value is non-differentiable, which breaks down the end-to-end training framework. On the other hand, since the size of heatmap is usually smaller than inputs, it also produces quantization error. We modifies the max operation to operation of taking expectation, formulated as

$$J_k = \sum_{p_y=1}^{H} \sum_{p_x=1}^{W} p \cdot \hat{H}_k(p) \tag{3}$$

where H and W are the height and width of predicted heatmap $\hat{H}_k$.

In addition, we adopt the Mean Square Error as a loss function $\mathcal{L}_{lc}$ to calculate the loss between predicted heat maps and ground truth, formulated as follows:

$$\mathcal{L}_{lc} = \sum_{k=1}^{M} \left\| H_k - \hat{H}_k \right\|_2 \tag{4}$$

where M indicates the number of landmarks, $\hat{H}_k$ is the predicted heatmap for the k_{th} landmark.

In a similar way, we added a regression loss to improve the performance of facial landmark estimation, which is the Mean Square Error of predicted landmarks and ground truth. The total loss of landmark sub-network is:

$$\mathcal{L}oss_l = \mathcal{L}_{le} + \alpha_2 \mathcal{L}_{lc} = \sum_{k=1}^{M} \left\| J_k - \hat{J}_k \right\|_2 + \alpha_2 \mathcal{L}_{lc} \tag{5}$$

where α_2 is the balance factor, M indicates the number of landmarks, J_k and $\hat{J}_k$ are the predicated and ground truth landmark coordinates, respectively.

3.3 Cross Cascading Regression

Inspired by [14], in order to make full use of the information of head pose and facial landmarks, we design Cross-cascading Regression Network which connects facial landmark localization and head pose estimation together.

Through several convolutional and max pooling layers, the input image is processed down to a lower resolution, which is applicable for facial landmark localization sub-network and head pose estimation sub-network to take as input.

At the same time, after obtaining the predicted head pose, two deconvolutional layers are added to compute the upsampling features. To match the number of channels of the facial landmark localization's output features and the upsampling features, we set a convolutional layer serves as channel unification layer. With the convolutional layer for channel unification, these feature maps are associated together, which enables the communication between head pose and facial landmark information. The output of head pose estimation sub-network is the summation of facial landmark heatmaps, the upsampling features and intermediate features of head pose estimation sub-network.

We adopt the coarse-to-fine strategy and extend network further by stacking a block at the end, feeding the combination of information maps achieved by former block as input into the following. Moreover, to facilitate the efficiency of information communication, we insert the pose information map into the intermediate structure of hourglass network in the next block. The structure of our network is shown in Fig. 1.

Since Cross-cascading Regression Network consists of two sub-networks completing the head pose estimation and facial landmark localization simultaneously, the loss function must give consideration to the information of both head pose and facial key points, which is formulated as follows:

$$\mathcal{L}oss = \mathcal{L}oss_p + \lambda \cdot \mathcal{L}oss_l \tag{6}$$

where λ indicates the relative importance of the two terms.

4 Experiment

4.1 Dataset

We train our network on AFLW datasets. AFLW is a challenging dataset which consists of 24386 images of human faces in the wild, with head pose ranging from 0° to 120° for yaw and up to 90° for pitch and roll. It also provides at most 21 key points for each face. In our experiments, we train on a subset of the dataset, which contains nearly 20000 images, and keep the rest for evaluation. For each sample image, the facial area is cropped out and then resized into 256×256 for normalization.

4.2 Implementation Details

The network is implemented using Pytorch framework. The variance σ of the 2D Guassians in heatmap is set to 1. For invisible landmarks, the ideal estimations are defined as 0. During training, the learning rate is fixed to 2.5e−4. Instead of taking the max activated location as the final prediction, we use the expectations of the output heatmaps to predict landmarks, and the predicted pose is the expectation of each output angle computed based on the output classification features.

4.3 Evalution Metric

To evaluate a facial landmark localization algorithm, we adopt the widely used Normalized Mean Error (NME) as the evaluation metric, which can be formulated as follows:

$$NME = \frac{1}{n}\sum_{i=1}^{n}\frac{\|x_i - x_i^*\|_2}{l} \tag{7}$$

where l denotes the normalized distance and n is the number of facial landmarks involved in the evaluation. In our experiment, l is the width (or height) of the face bounding box which is square for test samples in AFLW, and n indicates the number of visible landmarks.

4.4 Comparison with State of the Arts

We compare our Cross-cascading Regression network (CCR) with state-of-the-art head pose estimation and facial landmark detection approaches, results are shown in Tables 1 and 2. The result shows that our Cross-cascading regression Network achieves better or comparable performance when compared with state-of-the-art methods, which justifies the effectiveness of combining pose and landmark information explicitly.

Table 1. Mean Average Error (MAE) of Euler angles across different methods on AFLW.

Methods	Yaw	Pitch	Roll	MAE
Multi-loss ResNet50 [3] (α=1)	6.26	5.89	3.82	5.324
Multi-loss AlexNet [3] (α=1)	7.79	7.41	6.05	7.084
KEPLER [1]	6.45	**5.85**	8.75	7.017
Patacchiola, Cangelosi [15]	11.04	7.15	4.40	7.530
CCR (two blocks stacked)	**5.22**	**5.85**	**2.51**	**4.527**

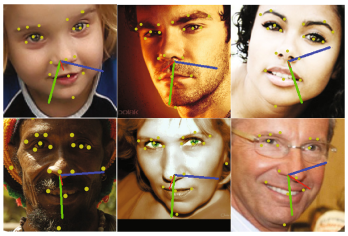

Fig. 2. Results of landmark detection and pose estimation generated from Cross-cascading Regression network. The red axis points towards the front of the face, green pointing downward and blue pointing to the side. (Color figure online)

Table 2. Normalized Mean Error (NME) of facial landmark detection across different methods on AFLW.

Methods	NME
CDM [16]	12.44
RCPR [17]	7.85
ESR [10]	8.24
Hyperface [18]	4.26
FRTFA [19]	4.23
PIFA [20]	6.80
CCL [21]	5.85
CCR (two blocks stacked)	5.72

5 Conclusion

In this work, we propose a novel network architecture named Cross-cascading Regression Network which consists of two sub-networks. The proposed model performs head pose estimation and facial landmark localization simultaneously with compact information communication. We extend our network architecture by stacking multiple blocks end-to-end, feeding the combination of information maps achieved by former block as input into the next, which achieves a coarse-to-fine prediction scheme. Our loss function consists of regression loss and classification loss, and the prediction of pose and landmarks are calculated by binned results. The proposed method achieves superior, or at least comparable performance in comparison with state-of-the-art methods on challenging datasets AFLW, which demonstrates the effectiveness of combining information from differenct tasks and the significance of cascading.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 61427811, 61273272, 61573360).

References

1. Kumar, A., Alavi, A., Chellappa, R.: KEPLER: keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2017)
2. Amador, E., Valle, R., Buenaposada, J.M., Baumela, L.: Benchmarking head pose estimation in-the-wild. In: Mendoza, M., Velastín, S. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (2018)
3. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshops) (2018)
4. Kokkinos, I.: UberNet: training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

5. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision (ECCV) (2014)
6. Huang, L., Yang, Y., Deng, Y., Yu, Y.: DenseBox: unifying landmark localization with end to end object detection, vol. abs/1509.04874 (2015)
7. Sun, X., Xiao, B., Liang, S., Wei, Y.: Integral human pose regression, volume arXiv:abs/1711.08229 (2017)
8. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2011)
9. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV) (2016)
10. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
11. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
12. Wu, Y., Gou, C., Ji, Q.: Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: dense human pose estimation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. In: Pattern Recognition (2017)
16. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: IEEE International Conference on Computer Vision (ICCV) (2013)
17. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision (ICCV) (2013)
18. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017)
19. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: a 3d spatial transformer network approach in unconstrained poses. In: International Conference on Computer Vision (ICCV) (2017)
20. Jourabloo, A., Liu, X.: Pose-invariant 3d face alignment. In: IEEE International Conference on Computer Vision (ICCV) (2016)
21. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)