https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-374

Robust Convolutional Neural Network Cascade for Facial Landmark Localization Exploiting Training Data Augmentation*

Ruiyi Mao [a]***, Qian Lin*** [b] ***and Jan P. Allebach*** [a]

[a] ***School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, U.S.A***
[b] ***HP Labs, Palo Alto, CA, U.S.A***

Abstract

Facial landmark localization plays a critical role in many face analysis tasks. In this paper, we present a coarse-to-fine cascaded convolutional neural network system for robust facial landmark localization of faces in the wild. The system consists of two cascaded convolutional neural network levels. The first level network generates an initial prediction of all facial landmarks. The second level networks are cascaded to implement facial component-wise local refinement of the landmark points. We also present a novel data augmentation method for facial landmark localization networks training. The experiment result shows our method outperforms state-of-the-art methods on 300W [18] common dataset.

1. Introduction

Facial landmark localization has numerous applications in face alignment, face recognition, facial emotion recognition, facial motion capture and 3D face reconstruction. In the past few years, facial landmark localization in unconstrained conditions has received a lot of attention. Many well-known facial landmark localization algorithms from early researches are mostly based on model-based approaches such as Active Appearance Models (AAMs) [1, 2], Constrained Local Models (CLMs) [3, 4], Active Shape Models (ASMs) [5, 6]. These algorithms work well in particular constrained conditions, but their performance may deteriorate in unconstrained scenarios.

With the development of descriptive features, regression-based algorithms such as cascaded regression becomes the mainstream in landmark localization. Recently, due to increasingly large amount of available training data and the breakthrough in deep learning, the current trend in landmark localization is to involve deep learning architectures into the solution [7, 8, 9, 10, 11, 12]. Sun et al. [7] first applied cascaded deep convolutional neural networks (DCNNs) to the regression framework. After that, a lot of works have been done to further exploit cascaded DCNNs framework [9, 12, 22] and achieved state-of-the-art performances. All these algorithms share the same strategy of level-wise coarse-to-fine refinement. At each level, the networks are trained to locally refine a subset of facial landmarks generated by networks from previous level. It has been demonstrated that by utilizing deep learning methodologies such as deep convolutional neural network cascade, superior robustness and accuracy have been achieved compared to the previous methods.

*This work is supported by HP Labs, Palo Alto, CA.

However, facial landmark localization still remains a very challenging problem. One challenge comes from the large variations of face appearance caused by different illuminations, different facial expressions, different yaw, pitch and roll angles of heads and different image qualities. In addition to more powerful networks, some other techniques are also used. Small image translations and rotations for training data augmentation are presented in [7]. Moreover, pre-alignments of face or facial component images by rotating them to a canonical direction are used ahead of some network levels by [9, 12, 22].

Another challenge is that the performance of the landmark localization algorithms highly relies on the consistency and accuracy of the detected face bounding boxes for the input images generated by a face detector. [9] claims that a large portion of its error can be attributed to the poor face detector it uses. To solve this problem, a bounding box aggregation technique is introduced in [27] in order to generate stable and accurate face bounding boxes. Multiple face detectors are used simultaneously in this method to provide input to a bounding box aggregation algorithm to generate an accurate final bounding box.

It is critical for a landmark localization method to be robust to both variant input face images and variant face detectors or detected face bounding boxes. Our works in this paper to address the challenges can be summarized as follows:

1. We designed a coarse-to-fine two-level convolutional neural network cascade for facial landmark localization. Different from some previous works [7, 9, 12, 22], our method doesn't use point-wise refinement level or image pre-alignments but still obtains higher accuracy.
2. We proposed a novel data augmentation method for facial landmark localization training. By using this method, our system has great robustness to both input face images and detected face bounding boxes. Very similar accuracy can be obtained by using different face detectors.
3. Experimental evaluations show that our method demonstrates both superior robustness and accuracy. It outperforms state-of-the-art methods on 300W common test dataset.

2. Related work

The early works in facial landmark localization mainly used model-based methods such as ASMs [1, 2], AAMs [5, 6] and CLMs [3, 4]. A prior generative shape model is generated and

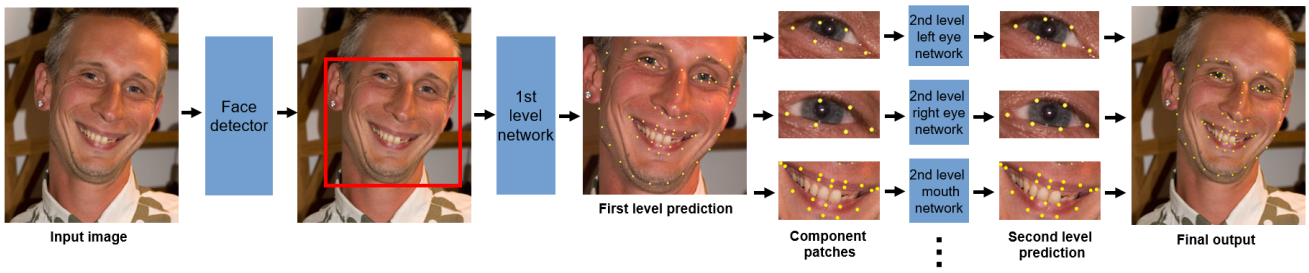


Figure 1. System overview. *A face bounding box is generated by a face detector. The face image is cropped by the face bounding box and fed to the first level network. An initial prediction of all facial landmark points are generated by the first level network. By using first level initial landmark prediction, regional images containing facial components are cropped and fed to the corresponding second level networks. Further component-wise landmark refinements are done by second level networks to generate final output.*

used. However these methods didn't work well for faces in the wild and the trend moved to regression based method as well as deep learning.

Regression based methods predict facial landmarks directly from appearance. Most regression based methods utilize a cascaded regression framework and coarse-to-fine manner, in which the landmark predictions are continually improved, for example Stochastic Descent Method (SDM) [19], LBF [20] and Coarse-to-Fine Shape Searching (CFSS) [21].

With the development of deep learning, this technique has recently be used in landmark localization. Sun et al. [7] first used a three-level cascaded convolutional neural network for robust facial landmark localization. By using the similar method of convolutional neural network cascade, more works have been done by cascading more levels [9, 23, 12]. Besides, Zhang et al. [22] developed a multi-task network to tackle both facial landmark localization and facial attribute classification. Trigeorgis et al. [26] first applied recurrent neural networks to facial landmark localization.

3. The Proposed method

3.1 Overview

Fig. 1 briefly illustrates the coarse-to-fine two-level cascaded facial landmark localization system. A CNN based face detector is trained and used in the system to generate face bounding boxes. The input to the first level network is the face region returned by the face bounding boxes. The first level of the network is used for generating robust initial prediction for all facial landmarks. The initial landmark prediction is robust and usually very close to the landmark true position. This is because the first level network takes the full face as its input to better use global information, context and structure of faces, avoiding the problem of local minimum and corrupted or ambiguous local features which may result in inaccurate predictions.

However, the capacity of a single network is limited by its size. Even though first level network is powerful enough to handle great variations of input face images, there is still room for growth of landmark localization accuracy, especially for some deformable facial components such as mouth and eyes. Since these components deform a lot with different expressions, making precise prediction for these components is difficult for a single network targeting at making an overall prediction. To tackle the problem, second level networks are trained and cascaded following the first level network. Second level networks can take a closer look at the details of each facial component and implement component-wise landmark local refinement. By doing so, the prediction burden is distributed across the networks in different levels, and good performance can be achieved by the cascaded networks of moderate size.

In order to make the landmark localization system more robust, we developed powerful data augmentation techniques for training the cascaded networks. This makes the landmark localization system robust to large variance of input face images as well as variant initial face bounding boxes given by different face detectors. The details of building and training the network are described in the following subsections.

3.2 Coarse-to-fine Convolutional Neural Network cascade

As shown in Fig. 1, the neural networks work in a coarse-to-fine manner. The input image is cropped according to the face bounding box generated by a face detector and then fed to the first level network. A VGG [13] style network is designed and serves as the first level convolutional network. The output layer of the network generates the predicted landmark coordinates relative to the input face region. Because input face images might be locally corrupted or partially occluded, the highest priority for the first level network is to generate robust prediction of all facial landmarks. By learning global features and structures, the first level network is capable to give reasonable predictions even if some landmark points are invisible.

With a robust initial prediction of all facial landmark points from first level, second level networks are mainly in charge of further refining predicted landmark positions. The input face image is segmented to smaller regions containing different facial components using landmark predictions from first level, e.g. eye, mouth, nose. These components' regional images are fed to the corresponding second level networks. Second level networks are similar VGG style networks as first level network but with a smaller size since second level networks only need to process regional information. The second level networks extract regional features and generate regional refined landmark predictions. These refined landmark points are combined with other landmark points from first stage to generate final landmark points prediction. And since both first level and second level networks are trained to be robust to variant input images, in our work face/facial component align-

ment such as canonical orientation transformation is not used. In Section 4 it is shown that we still obtained state-of-the-art performance without applying these input image alignments. It should be noted that there is no networks for point-wise landmark refinement involved in the system. This is because we believe that landmarks for facial components are the smallest landmark subsets that can be studied and processed separately since landmarks belonging to the same facial component have strong correlations with each other. If they are refined individually, the lack of global or regional information may cause severe failure.

(a) Sample original raw training image.

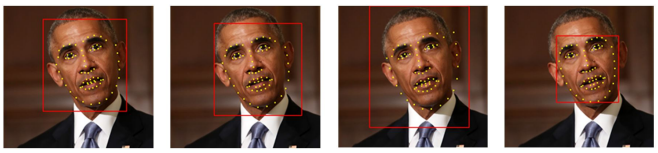
(b) Sample augmented training images with ground truth landmark points annotated. Bounding box random expansion, random rotation and random blurring are applied.

Figure 2. Example output images of proposed training data augmentation method.

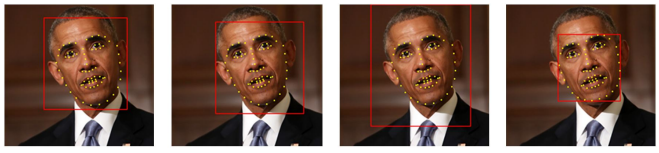
3.3 Training data augmentation for facial landmark localization

In order to train the networks to be robust and avoid overfitting, a set of data augmentation techniques are developed and used for the training of both first and second level networks. For the first stage network, to decouple network performance with particular face detector, no face detector is involved in generating face bounding boxes for training data. A novel bounding box random expansion data augmentation method is developed for training the networks. Firstly, a tightest bounding box containing all ground truth landmarks is generated as the initial face bounding box for one training image. Then the four boundary sides are randomly shifted towards left, right, top and bottom respectively. In this work the maximum shift for left/right sides and top/bottom sides is 0.3 of the width and height of the initial face bounding box. Finally the augmented training image sample is obtained by cropping with the expanded face bounding box and resizing to fit the network input dimension. By using bounding box expansion method, the same augmentation performance can be obtained comparing with the combination of conventional data augmentation methods, i.e. random cropping, random translation, random scaling and random stretching. However, bounding box expansion method is much simpler and easier to apply on landmark localization and similar problems than the conventional data augmentation methods set. This is because for landmark localization problems, data augmentation should be applied on image and landmarks consistently and simultaneously and need to be guaranteed that all the ground truth landmarks are always inside the augmented image.

Besides, for each training image, random rotation and random Gaussian blurring are applied before bounding box expansion for different augmentation purposes. At last, a horizontally flipped image is generated for each augmented training sample and added to the training dataset. The same data augmentation methods are used for the training of second level networks, the only difference is that the augmentation for second level is based on each facial component region instead of whole face. Fig. 2 shows the example augmented training images using the proposed method. Superior robustness of networks can be obtained by using this method for training data augmentation. Fig. 3 shows the comparison of networks trained using the proposed data augmentation method and trained using ground truth face bounding boxes. It can be seen that the networks trained with the proposed data augmentation method have great robustness to different detected face bounding boxes. It demonstrates that our networks can be easily cascaded with any reasonable face detector without being retrained but still obtain very similar performance.



(a) Example landmark predictions from network trained using ground truth bounding box (the left most one).



(b) Example landmark predictions from network trained using proposed data augmentation method.

Figure 3. Example landmark predictions with different detected face bounding boxes.

Since these data augmentation methods are independent and applied with random values, theoretically infinite number of different training images can be generated from one training sample. In this work, we manually divide the whole training dataset into two subsets: normal augmentation set and strong augmentation set. Normal augmentation set contains common training sam-

Figure 4. Examples from test dataset. *The data set contains great variations in pose, expressions and lighting conditions and our system is still able to give good facial landmark prediction.*

ples and strong augmentation set contains rarer and challenging training samples such as rare and challenging head poses, facial expressions and illuminations. Larger number of training samples are generated from each sample belonging to strong augmentation set than normal augmentation set. By doing so the networks can better learn the structure of facial landmarks, avoid over-fitting and have accurate predictions for uncommon faces.

4. Experiments

For training of the proposed networks, LFPW[14], HELEN[15], AFW[16] and Menpo benchmark[17] datasets with 68-point facial landmark annotation are used. To evaluate the performance of our cascaded networks, 300W[18] common test dataset (test set of LFPW, HELEN) is used. We also compare our methods with recent state-of-the-arts on 300W common dataset in Table. 1. It shows that our method has the highest accuracy among these methods. Fig. 4 gives some examples from test dataset. It can be seen that our system is very robust and can generate good landmark prediction for variant input face images.

5. Conclusion

In this paper we present a convolutional neural networks system for facial landmark localization. In our method, two CNN levels are carefully designed to form coarse-to-fine cascaded networks. Besides, a novel data augmentation method for facial landmark localization is presented. The experiment result shows the state-of-the-art performance of the proposed method which demonstrates its superiority.

Method	Normalized mean error
SDM [19]	5.57
CFAN [8]	5.50
LBF [20]	4.95
CFSS [21]	4.73
TCDCN [22]	4.80
Fan et al. [23]	4.76
Honari et al. [24]	4.67
Lai et al. [25]	4.07
Chen et al. [12]	3.73
Ours	**3.52**

* **Table 1.** Comparison of state-of-the-arts approaches on 300W common test dataset.

References

[1] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell.,23(6):681-685, 2001.

[2] F. Kahraman, M. Gokmen, S. Darkner, and R. Larsen. An active illumination and appearance (AIA) model for face alignment. In 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007.

[3] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In European Conference on Computer Vision, pages 340-353, 2008.

[4] 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, 2011.

[5] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2008.

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. Computer Vision and Image Understanding 61.1 (1995): 38-59.

[7] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade

for facial point detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3476-3483, 2013.

[8] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In European Conference on Computer Vision, pages 116. Springer, 2014.

[9] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 386-391, 2013.

[10] Z. Huang, E. Zhou, and Z. Cao. Coarse-to-fine face alignment with multi-scale local patch regression. arXiv preprint arXiv:1511.04901, 2015.

[11] S. Zhang, H. Yang and Z. Yin. Transferred deep convolutional neural network features for extensive facial landmark localization. IEEE Signal Processing Letters 23.4 (2016): 478-482.

[12] X. Chen, E. Zhou, J. Liu, and Y. Mo. Delving Deep into Coarse-to-fine Framework for Facial Landmark Localization. In Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge. 2017.

[13] K. Simonyan, and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[14] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 545-552, 2011.

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In European Conference on Computer Vision (ECCV), pages 679-692. Springer, 2012.

[16] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2879-2886. IEEE, 2012.

[17] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The Menpo Facial Landmark Localisation Challenge: A step closer to the solution. In IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW), 2017.

[18] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. Image and Vision Computing, 47:318, 2016.

[19] X. Xiong and F. D. L. Torre. Supervised descent method and its applications to face alignment. In Computer Vision and Pattern Recognition, pages 532-539, 2013.

[20] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pages 1685-1692, 2014.

[21] 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.

[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In European Conference on Computer Vision, pages 94-108. Springer, 2014.

[23] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. Image and Vision Computing, 47:27-35, 2016.

[24] S. Honari, J. Yosinski, P. Vincent, and C. Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[25] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan. Deep recurrent regression for facial landmark detection. IEEE Transactions on Circuits and Systems for Video Technology, 2016.

[26] Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., and Zafeiriou, S. 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.

[27] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu. Face Detection, Bounding Box Aggregation and Pose Estimation for Robust Facial Landmark Localisation in the Wild. In IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW), 2017.

Author Biography

Ruiyi Mao *received his B.Eng degree in Electrical and Computer Engineering from Huazhong University of Science and Technology, China, and University of Birmingham, UK, in May 2012. He is currently pursuing his Ph.D. degree in Electrical and Computer Engineering at Purdue University. His current research interests include machine learning, computer vision and digital image processing.*

Qian Lin *is a distinguished technologist working on computer vision and deep learning research in HP Labs. Dr. Lin joined the Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. Dr. Lin is inventor/co-inventor for 44 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, and Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013.*

Jan P. Allebach *is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IST), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IST, the highest award that IST bestows. He has received the IEEE Daniel E. Noble Award, and is a member of the National Academy of Engineering. He currently serves as an IEEE Signal Processing Society Distinguished Lecturer (2016-2017).*