

# Deep Convolutional Network Cascade for Facial Point Detection

Yi Sun<sup>1</sup>

Xiaogang Wang<sup>2,3</sup>

Xiaoou Tang<sup>1,3</sup>

<sup>1</sup>Department of Information Engineering, The Chinese University of Hong Kong

<sup>2</sup>Department of Electronic Engineering, The Chinese University of Hong Kong

<sup>3</sup>Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

sy011@i.e.cuhk.edu.hk

xgwang@ee.cuhk.edu.hk

xtang@i.e.cuhk.edu.hk

## Abstract

*We propose a new approach for estimation of the positions of facial keypoints with three-level carefully designed convolutional networks. At each level, the outputs of multiple networks are fused for robust and accurate estimation. Thanks to the deep structures of convolutional networks, global high-level features are extracted over the whole face region at the initialization stage, which help to locate high accuracy keypoints. There are two folds of advantage for this. First, the texture context information over the entire face is utilized to locate each keypoint. Second, since the networks are trained to predict all the keypoints simultaneously, the geometric constraints among keypoints are implicitly encoded. The method therefore can avoid local minimum caused by ambiguity and data corruption in difficult image samples due to occlusions, large pose variations, and extreme lightings. The networks at the following two levels are trained to locally refine initial predictions and their inputs are limited to small regions around the initial predictions. Several network structures critical for accurate and robust facial point detection are investigated. Extensive experiments show that our approach outperforms state-of-the-art methods in both detection accuracy and reliability<sup>1</sup>.*

## 1. Introduction

Facial keypoint detection is critical for face recognition and analysis, and has been studied extensively in recent years [3, 4, 5, 8, 9, 11, 20, 21, 23, 25, 26, 27, 28]. This problem is challenging when face images are taken with

positions of facial points iteratively and good initializations are critical. The mean shape or shapes sampled from the training set is often used as the initialization, which may far from the target position, and the update may end with a local minimum. In addition, many approaches face the problem that the visual features extracted are not as discriminative or not reliable enough to predict facial points, and context information becomes important. Most approaches employ shape constraints, which are relatively weak. It is desirable to directly extract texture context information over the whole face region, since they contain rich information. This requires much more powerful classifiers or regressors, since the visual complexity increases exponentially with the size of the image region.

To solve these problems, we propose a cascaded regression approach for facial point detection with three levels of convolutional networks. Different from existing approaches which roughly estimate the initial positions of facial points, our convolutional networks make accurate predictions at the first level, even on very challenging cases as shown in Figure 1. It effectively avoids the local minimum problem faced by other approaches. The convolutional networks take the full face as input to make the best use of texture context information, and extract global high-level features at higher layers of the deep structures, which can effectively predict keypoints even when low-level features from local regions are ambiguous or corrupted in challenging image examples. Our convolutional networks are trained to predict all the keypoints simultaneously and the constraints of keypoints are implicitly encoded.

The remaining two levels of convolutional networks refine the initial estimation of keypoints. Different from existing methods [5, 25, 26] which apply the same regressor at different cascade stages, we design different convolutional networks. The network structures at these two levels are shallower, since their tasks are low-level and their input is limited to small local regions around the initial positions. At each level, multiple convolutional networks are fused to improve the accuracy and reliability of estimation. Through detailed empirical investigation, we find that several factors regarding the network structures are critical for achieving good performance in facial point detection. Detailed experimental evaluations show that our approach outperforms state-of-the-art methods on both accuracies and reliability.

## 2. Related Work

Significant progress on facial keypoint detection has been achieved in recent years. Many used Adaboost [20], SVM [4, 28], or random forest [3] classifiers as component detectors and detection was based on local image features. Shape constraints are important to refine component detection results and much research has been focusing on this. The evidence given by local component detectors and the

shape constraints can be balanced by optimizing designed objective functions [4, 28]. Liang *et al.* [20] trained a set of direction classifiers to guide the search of good shape. Amberg and Vetter [3] employed a branch and bound algorithm to efficiently find optimal configurations from a large number of candidates proposed by component detectors.

Among regression-based approaches, Dantone *et al.* [9] and Valstar *et al.* [26] predicted facial points from local patches with random forests and support vector regressors respectively. To resolve the uncertainties in predictions, Valstar *et al.* [26] modelled the spatial relations of facial points with Markov random field and Dantone *et al.* [9] fused many predictions from patches densely sampled within the face region. Patrick *et al.* [25] updated the parameters of an active appearance model with regressors. Cao *et al.* [5] used the whole face region as input and random ferns as the regressor. Shapes to be predicted were expressed as linear combinations of training shapes.

Convolutional networks and other deep models have been successfully used in vision tasks such as face detection and pose estimation [24], face parsing [22], image classification [6, 17], and scene parsing [10]. The research works on convolutional networks mainly focus on two aspects: network structures and feature learning algorithms. Coates *et al.* [7] analyzed the performance of single-layer networks with different filter strides, filter sizes, and the numbers of feature maps. Jarrett *et al.* [14] introduced strong nonlinearities after convolution, including absolute value rectification and local contrast normalization, and also compared different combinations of nonlinearities and pooling strategies. Not until recently has the potential of convolutional networks truly been discovered, when it becomes big (with hundreds of maps per layer) and deep (with up to five convolutional stages). By using large-scale convolutional networks, Ciresan *et al.* [6] significantly improved the state-of-the-art on some standard classification datasets. Even larger convolutional network was introduced in [17], and it significantly improved image classification accuracies on the ImageNet. Examples of recently proposed feature learning algorithms include convolutional sparse coding [16] and topographic independent component analysis [18].

## 3. Cascaded convolutional networks

In this paper, we focus on the structural design of individual networks and their combining strategies. Figure 2 is an overview of our approach. There are five facial points to be detected: *left eye center* (LE), *right eye center* (RE), *nose tip* (N), *left mouth corner* (LM), and *right mouth corner* (RM). We cascade three levels of convolutional networks to make coarse-to-fine prediction. At the first level, we employ three deep convolutional networks, F1, EN1, and NM1, whose input regions cover the whole face (F1), eyes and nose (EN1), nose and mouth (NM1). Each network si-



sharing weights does not work well on images with fixed spatial layout, such as faces. For example, while eyes and mouth may share low-level features (*e.g.* edges), they are very different at high-level. So for networks whose inputs contain different semantic regions, locally sharing weights at high layers is more effective for learning different high-level features, *e.g.*, eyes, nose, and mouth. The idea of locally sharing weights was originally proposed for convolutional deep belief net for face recognition [12].

### **3.2. Multi-level regression**

We find several effective ways to combine multiple convolutional networks. The first is multi-level regression. The face bounding box is the only prior knowledge for networks at the first level. The relative position of a facial point to the bounding box could vary in a large range due to large pose variations and the instability of face detectors. So the input regions of networks at the first level should be large in order to cover many possible predictions. But large input region

|    | layer 0  | layer 1      | layer 2 | layer 3      | layer 4 | layer 5      | layer 6 | layer 7      | layer 8 | layer 9 |
|----|----------|--------------|---------|--------------|---------|--------------|---------|--------------|---------|---------|
| S0 | I(39,39) | CR(4,20,2,2) | P(2)    | CR(3,40,2,2) | P(2)    | CR(3,60,3,3) | P(2)    | CR(2,80,2,2) | F(120)  | F(10)   |
| S1 | I(31,39) | CR(4,20,1,1) | P(2)    | CR(3,40,2,2) | P(2)    | CR(3,60,2,3) | P(2)    | CR(2,80,1,2) | F(100)  | F(6)    |
| S2 | I(15,15) | CR(4,20,1,1) | P(2)    | CR(3,40,1,1) | P(2)    | F(60)        | F(2)    |              |         |         |
| S3 | I(39,39) | CR(4,20,2,2) | P(2)    | CR(3,40,2,2) | P(2)    | CR(3,60,3,3) | P(2)    | F(120)       | F(10)   |         |
| S4 | I(39,39) | CR(4,20,2,2) | P(2)    | CR(3,40,2,2) | P(2)    | F(120)       | F(10)   |              |         |         |
| S5 | I(39,39) | CR(4,20,2,2) | P(2)    | F(120)       | F(10)   |              |         |              |         |         |
| S6 | I(39,39) | C(4,20,2,2)  | P(2)    | C(3,40,2,2)  | P(2)    | C(3,60,3,3)  | P(2)    | C(2,80,2,2)  | F(120)  | F(10)   |
| S7 | I(39,39) | CR(4,20,1,1) | P(2)    | CR(3,40,1,1) | P(2)    | CR(3,60,1,1) | P(2)    | CR(2,80,1,1) | F(120)  | F(10)   |

Table 1: Summary of network structures. F1 adopts S0. Both EN1 and NM1 adopt S1. All the networks at the second and third levels share S2. To investigate different designs of network structures, we also compare different structures S3-S7 for F1 in experiments.

|    | net | left  | right | top   | bottom |
|----|-----|-------|-------|-------|--------|
| L1 | F1  | -0.05 | +1.05 | -0.05 | +1.05  |
|    | EN1 | -0.05 | +1.05 | -0.04 | +0.84  |
|    | NM1 | -0.05 | +1.05 | +0.18 | +1.05  |
| L2 | *21 | -0.16 | +0.16 | -0.16 | +0.16  |
|    | *22 | -0.18 | +0.18 | -0.18 | +0.18  |
| L3 | *31 | -0.11 | +0.11 | -0.11 | +0.11  |
|    | *32 | -0.12 | +0.12 | -0.12 | +0.12  |

Table 2: Summary of network input ranges, which are described by left, right, top, and bottom boundary positions. For networks at level 1 (L1), the four boundary positions are relative to the normalized face bounding box with boundary positions (0, 1, 0, 1). For networks at level 2 (L2) and level 3 (L3), the four boundary positions are relative to the predicted facial point position.

prediction. At each of the two levels, we use two regions of different sizes to predict each point. Regions at the third level are smaller than the second level. The precise input ranges of all the networks are listed in Table 2.

**Training.** At the first level, we take training patches according to the face bounding box, and augment them by small translation and rotation. At the following levels, we take training patches centered at positions randomly shifted from the ground truth position. The maximum shift in both horizontal and vertical directions is 0.05 at the second level, and 0.02 at the third level, where the distances are normalized with the face bounding box. Networks at the third level aim at more subtle adjustment to previous predictions than those at the second level. Learnable network parameters include the weight  $w$ , the gain  $g$ , and the bias  $b$ , which are initialized by small random numbers and learned by stochastic gradient descent. Levenberg-Marquardt method [19] is used to estimate the neurons’ learning rate individually. Training continues until converge.

## 5. Experiments

We first investigate different designs of network and cascade structures with a training set and a validation set col-

lected by ourselves. Then we compare with the state-of-the-art methods and commercial software on two public test sets without changing the training set. Our training and validation sets have no overlap with the two public test sets.

### 5.1. Investigate network and cascade structures

We created a dataset with 13,466 face images, a.963308.8, the gain

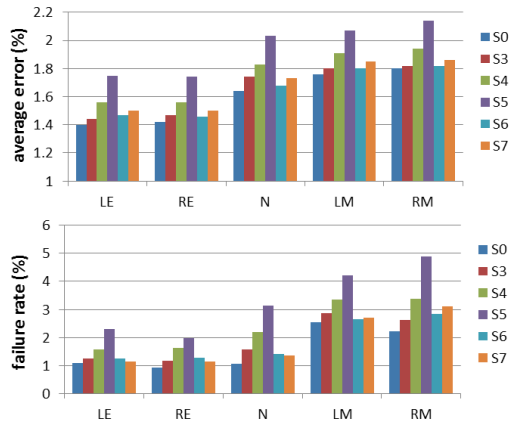


Figure 4: Average detection errors and failure rates of convolutional network F1 with different structures.

cally shared weights, but with different depths. The results show that the performance can be significantly improved by including more layers. Since the input face region is in size of  $39 \times 39$  and the network keeps downsampling images as going up to the top layers, S0 has reached the maximum number of possible layers. S6 and S7 have the same layer-

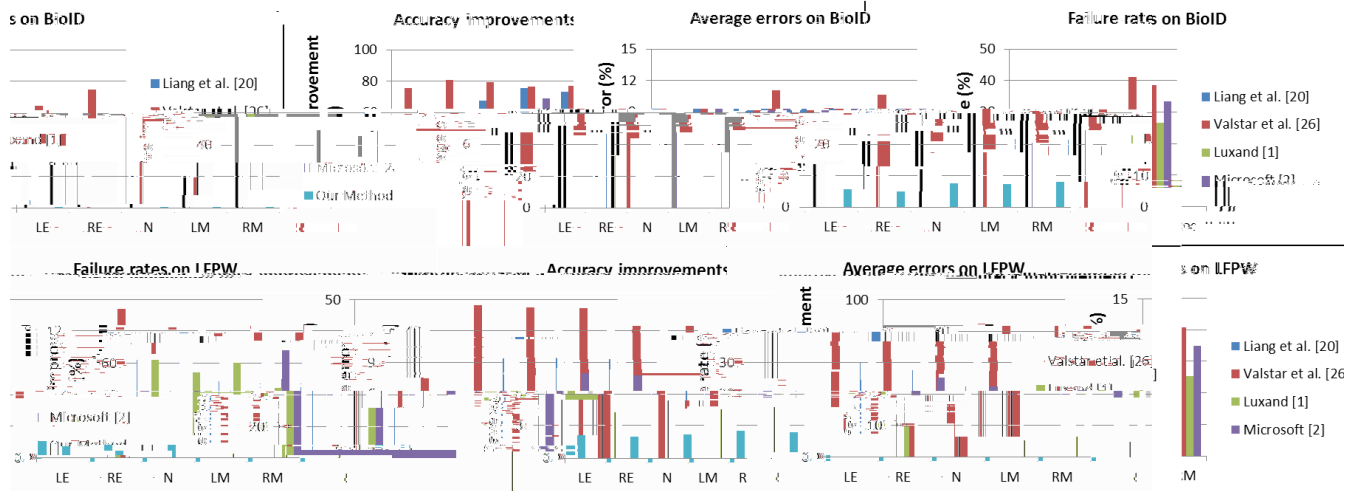


Figure 6: Comparison on BioID and LFPW. Since our failure rate approaches to zero on BioID, it may not be observable in the figure.  $\text{Relative improvement} = \frac{\text{reduced average error}}{\text{average error of the method in comparison}}$ . We achieved over 50% accuracy improvement on both datasets.

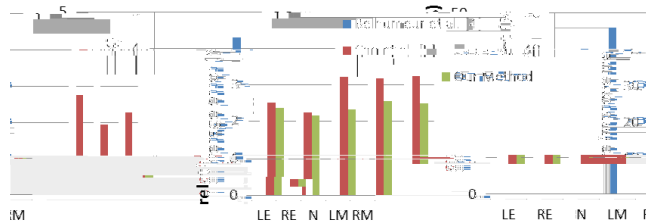


Figure 7: Compare with Belhumeur *et al.* [4] and Cao *et al.* [5] on LFPW test images.

errors and our relative accuracy improvements over the two methods. [4, 5] are very competitive methods, which perform significantly better than their contemporaries. Still, we improved their results with a large margin. More than 20% relative accuracy improvement is achieved for nose tip and two mouth corners. The C++ implementation of our algorithm takes 0.12 second to process one image on a 3.30GHz CPU<sup>3</sup>. The system can be easily parallelized since convolutional networks at each level are independent.

## 6. Conclusion

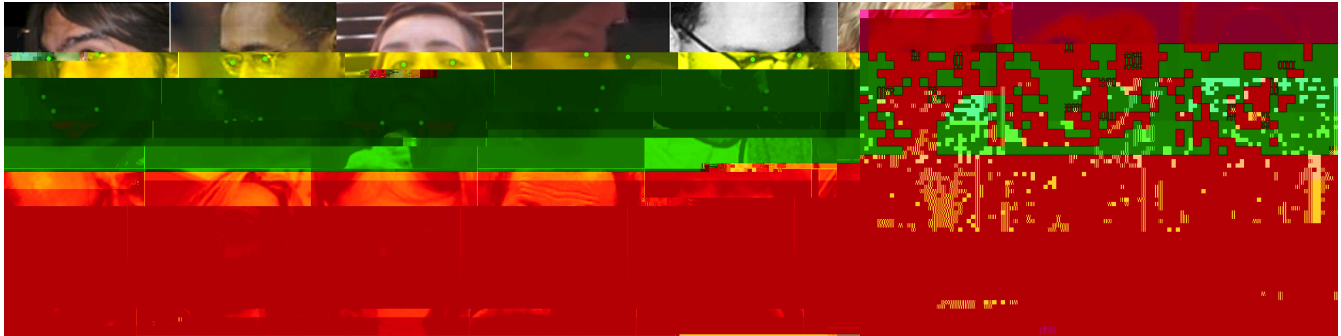
We proposed an effective convolutional network cascade for facial point detection. Deep convolutional networks at the first level provide highly robust initial estimations, while shallower convolutional networks at the following two levels finely tune the initial prediction to achieve high accuracy. By exploring a few key features of the network structure, we achieve high performance convolutional networks with a relatively small scale. Our method significantly improves the prediction accuracy of state-of-the-art methods and lat-

<sup>3</sup>The time preparing the input for our algorithm (face detection and image resizing) is excluded.

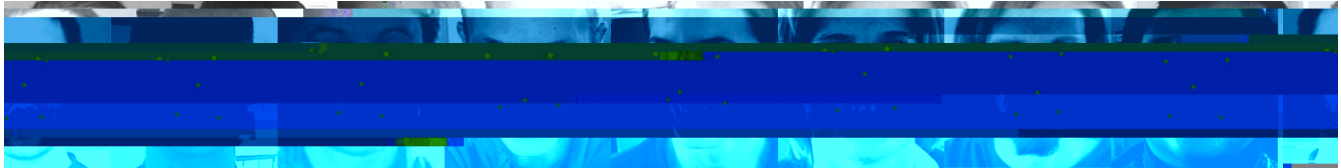
est commercial software.

## References

- [1] <http://www.luxand.com/facesdk/>. 6
- [2] <http://research.microsoft.com/en-us/projects/facesdk/>. 6
- [3] B. Amberg and T. Vetter. Optimal landmark detection using shape models and branch and bound. In *Proc. ICCV*, 2011. 1, 2
- [4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Proc. CVPR*, 2011. 1, 2, 6, 7
- [5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proc. CVPR*, 2012. 1, 2, 6, 7
- [6] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proc. CVPR*, 2012. 2
- [7] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research*, 2011. 2
- [8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. ECCV*, 1998. 1
- [9] M. Dantone, J. Gall, G. Fanelli, and L. J. V. Gool. Real-time facial feature detection using conditional regression forests. In *Proc. CVPR*, 2012. 1, 2
- [10] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 2013. 2
- [11] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *Proc. ECCV*, 2008. 1



(a) Validation images. Faces vary greatly on poses (column 1, 2, 3) and expressions (column 6), and may have occlusions (column 4, 5), artifacts (column 7) or blurring effect (column 8).



(b) BioID.



(c) LFPW. Faces vary greatly on poses (column 1, 2), expressions (column 5, 6) and illuminations (column 8), and may have occlusions (column 3, 4) or blurring effect (column 7).

Figure 8: Our results on validation images, BioID, and LFPW.

- [14] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Proc. ICCV*, 2009. 2, 3
- [15] O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *Proc. AVB-PA*, 2001. 6
- [16] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. In *Proc. NIPS*, 2010. 2
- [17] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012. 2
- [18] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *Proc. ICML*, 2012. 2
- [19] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and M. K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998. 5
- [20] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via a component-based discriminative search. In *Proc. ECCV*, 2008. 1, 2, 6
- [21] X. Liu. Generic face alignment using boosted appearance model. In *Proc. CVPR*, 2007. 1
- [22] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *Proc. CVPR*, 2012. 2
- [23] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *Proc. ECCV*, 2008. 1
- [24] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, 2007. 2
- [25] P. Sauer, T. Cootes, and C. Taylor. Accurate regression procedures for active appearance models. In *Proc. BMVC*, 2011. 1, 2
- [26] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Proc. CVPR*, 2010. 1, 2, 6
- [27] H. Wu, X. Liu, and G. Doretto. Face alignment via boosted ranking model. In *Proc. CVPR*, 2008. 1
- [28] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, 2012. 1, 2, 5