Image and Vision Computing 47 (2016) 27–35

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Approaching human level facial landmark localization by deep learning☆

Haoqiang Fan *, Erjin Zhou

Megvii Inc., Beijing, China

ARTICLE INFO

Article history:
Received 3 March 2015
Received in revised form 12 September 2015
Accepted 30 November 2015
Available online 12 December 2015

Keywords:
Facial landmark localization
Deep learning
Convolutional neural network

ABSTRACT

In this paper we present our solution to the 300 Faces in the Wild Facial Landmark Localization Challenge. We demonstrate how to achieve very competitive localization performance with a simple deep learning based system. Human study is conducted to show that the accuracy of our system has been very close to human performance. We discuss how this finding would affect our future direction to improve our system.

1. Introduction

Reliable face recognition crucially depends on accurate and robust face alignment. Good alignment enables the face recognizer to be robust against pose and expression change [1,2]. Facial landmark localization seeks to detect a set of predefined key points on a human face. It attracts intense interest from both the industry and the research community.

Despite rapid progress in this area, facial landmark localization in uncontrolled settings remains an unsolved problem. The major challenge comes from the large variations in the facial images (see Fig. 1 for example of images used in our experiments). The head can have large yaw or pitch angles. The lightening condition can be extreme. In addition, some of the images are of low quality. The resolution can be low. The image can be blurry or corrupted. When viewed locally, some of the facial landmarks are difficult to be recognized. Even humans have to use the global context to identify a point.

The major requirement of face alignment is robustness. The landmark localizer must return a set of reasonable point coordinates however the image is corrupted. Even if a point is not visible, it is typical for face recognizers to demand that the landmark localizer should "guess" a position for the point so that some global measurements (e.g. the rotation of the face) can be made. This poses great challenge to the landmark localizer.

The most straightforward way to solve the landmark localization problem is to view it as an image-based regression problem. The input is the RGB image. The outputs are the coordinate values. Any image-based regressor can be plugged in. This framework is capable of giving very promising result if the regressor is powerful enough. In our solution we use a CNN (Convolutional Neural Network) Cascade. There are two key ingredients in our solution:

- Deep network. To increase robustness, we used significantly deeper networks compared to previous works [3,4].
- Coarse-to-fine prediction. Using multiple CNNs in a coarse-to-fine manner improves accuracy.

However, obtaining good quality training data is difficult and expensive. We conduct human study to investigate human's ability to locate key points in an image. Then we discuss how our findings would influence our future direction to improve the system.

1.1. Related work

Face alignment is an indispensable step in modern face recognition system [5,6,7,and 8]. Generally, there are two schools of methods for facial landmark detection: model-based methods and regression-based methods. Model-based methods try to build models to fit input images. They can take into account the local texture appearance [9] or the part-based structure [10,11,12]. Their advantage is that human's prior knowledge can be easily incorporated into the system.

The regression-based methods are more straightforward. Deep neural networks [4,3,13,14] and boosted regressors [15] have been successfully employed. The regression-based method solely depends on the regressor's capacity to learn the extremely complex relation between pixel values and the appearance of facial features. Great care is needed in training these complex models. In our solution we aim at keeping the conceptual framework of the method as simple as possible. The major difference in our method is that our neural network's depths greatly exceed those used by existing works [4,3,13] which typically only has at most 4 convolutional layers (ours contains 8 convolutional

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.

* Corresponding author. Tel.: +86 13671270149.
E-mail addresses: fhq@megvii.com (H. Fan), zej@megvii.com (E. Zhou).

http://dx.doi.org/10.1016/j.imavis.2015.11.004
0262-8856/

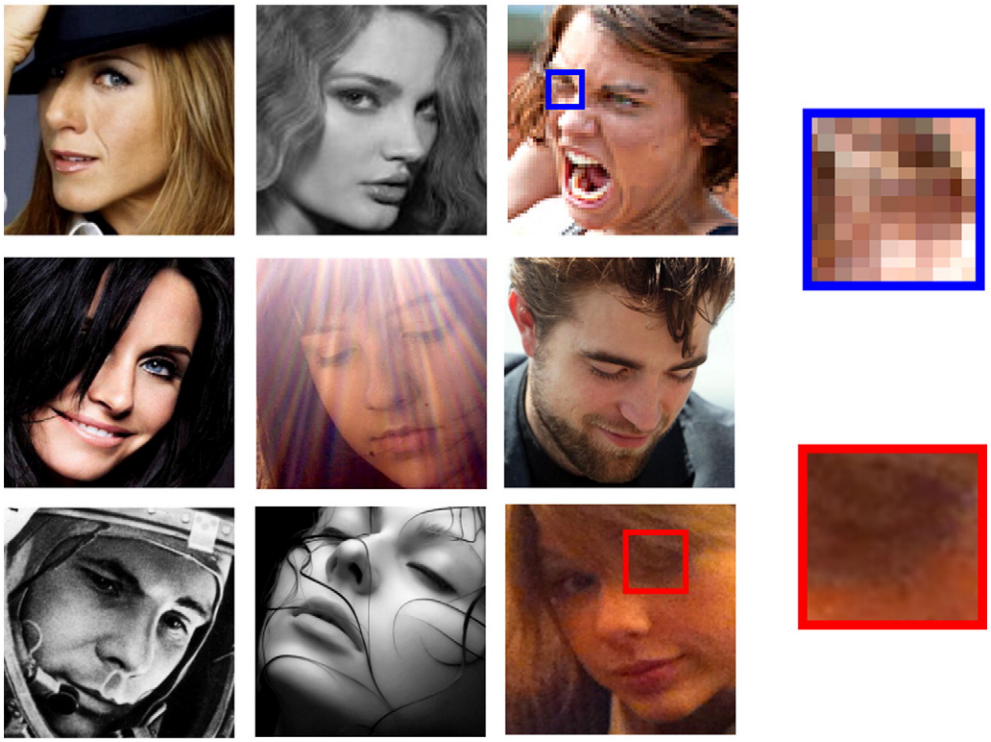

Fig. 1. Examples of pictures in the IBUG dataset which is used as the validation set in our experiments. This dataset contains very difficult pictures. On the right are two zoom-in view of two selected image regions. From the local patches we can hardly recognize the facial landmarks.

layers). Due to the extra power of the network, we can make the whole framework largely simplified.

2. Deep CNN cascaded for facial landmark localization

We formulate the landmark localization problem as learning a function that maps image pixel arrays to point coordinates. The input image $I_{h\times w}$ is a $h\times w$ three channel (RGB) image. It contains the face area found by the face detector. The output $P_{n\times 2}$ is landmark coordinates relative to the face's bounding box.

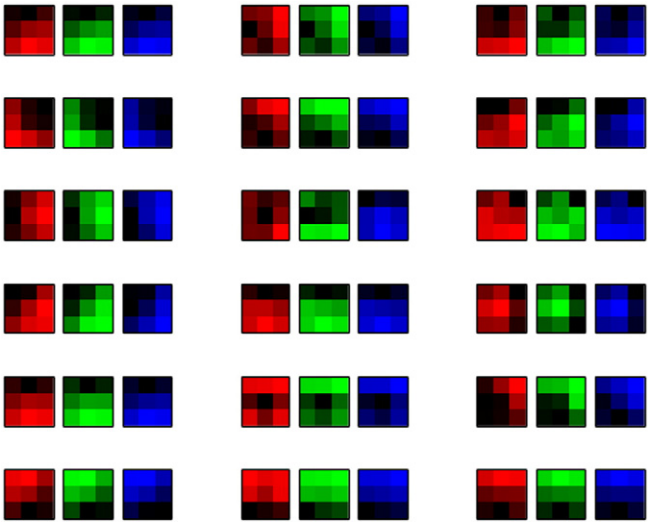

Fig. 2. Example of learned first layer filters. Each filter is a three channel weight map. Some of the filters can be recognized as line or corner detectors. The functions of others are not easily explainable.

In the following two subsections we present detailed descriptions of the two components of our framework: Convolutional Neural Network and coarse-to-find prediction.

2.1. Convolutional neural network

Mathematically, a Convolutional Neural Network is the composition of a series of non-linear function maps that operate on 3 dimensional arrays (height × width × # channels). We denote the input image as $I^0_{i,j,k}$ which is a three channel colored image. There are three types of functions: convolution, pooling and non-linear activation.

The convolution operation slides multi-channel "filters" on the image and computes the dot product of the filter's weights and the image pixels (See Fig. 2). In one network layer, many filters are applied to the image and their outputs are stacked to obtain the next layer's input channels. Mathematically, the function is defined as follows:

$$\mathrm{conv}_{W,B}\left(I^t\right)_{i,j,k} = \sum_{x=0}^{p\text{-}1}\sum_{y=0}^{q\text{-}1}\sum_{u=0}^{c\text{-}1} \mathrm{I}^t_{i\text{-}x,j\text{-}y,u} W^t_{k,u,x,y} + B^t_k.$$

The filter's size is $p\times q$ and the input image contains c channels. The four-dimensional array W stores the filter's weights, and B is a "bias" term which additively shifts the output for each output channel. The number of output channels does not need to match the input channels. We can think of the filters as pattern detectors that selectively response to certain input configurations. Fig. 1 visualizes the learned first layer filters in a network. Some of them are edge or corner detectors while the function of others is not easily explainable.

One special case of "convolution" is when the input image and output image are of size 1×1. In this case the operation reduces a matrix–vector multiplication. We call this kind of layers "fully connected" layers because each input value contributes to all output values.