Computer Standards & Interfaces 42 (2015) 105–112

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

Adaptive Cascade Deep Convolutional Neural Networks for face alignment☆

Yuan Dong *, Yue Wu *

Beijing University of Posts and Telecommunications, 100876, PR China

ARTICLE INFO

Article history:
Received 27 April 2015
Received in revised form 8 June 2015
Accepted 8 June 2015
Available online 16 June 2015

Keywords:
Face alignment
Adaptive cascade
Deep convolutional networks
Gaussian distribution

ABSTRACT

Deep convolutional network cascade has been successfully applied for face alignment. The configuration of each network, including the selecting strategy of local patches for training and the input range of local patches, is crucial for achieving desired performance. In this paper, we propose an adaptive cascade framework, termed Adaptive Cascade Deep Convolutional Neural Networks (ACDCNN) which adjusts the cascade structure adaptively. Gaussian distribution is utilized to bridge the successive networks. Extensive experiments demonstrate that our proposed ACDCNN achieves the state-of-the-art in accuracy, but with reduced model complexity and increased robustness.

1. Introduction

Face alignment or facial landmark localization plays a critical role in many visual applications such as face recognition, face tracking, facial expression recognition and 3D face modeling. Therefore, it has been extensively studied in recent years. However, robust facial landmark detection remains a challenging problem when face images are taken under the situation with extreme occlusion, lighting, expressions and pose. To address this issue, research explores the modeling of shape variation and appearance variation for improved performance. In general, this type of research can be categorized into three groups: constrained local model based methods [2–4], active appearance model based methods [5,6] and regression based methods [1,7–10].

Constrained local models build classifiers called component detectors to search for each facial feature point independently. These component detectors calculate response maps to present the appearance variance around facial feature points. Due to the problems of ambiguity and corruption in local features, facial points detected by the local experts may be far away from the ground truth positions. Then shape constraints are applied to adjust the initial positions for improved results [2,4]. However, the global contextual information is difficult to be embedded into these methods.

Instead of modeling the appearance with each facial point, active appearance models such as Active Appearance Model (AAM) [5] use a holistic perspective to model the appearance variance. An AAM model is composed of a linear shape model and a linear texture model. The Principal Component Analysis (PCA) is applied to bridge the relationship between the two models. Nevertheless, simple linear models can hardly present the nonlinear variations of facial appearance in the case of faces taken in complex environment (e.g., extreme lighting).

Regression based methods, on the other hand, directly learn a regression function from image appearance (features) to the target output (shapes) [11]. Cascade architecture is usually employed and explored in regression based models. In each stage of the cascade architecture, shape-index features [12] are extracted to predict the shape increment with linear regression [7], tree-based regression [8] where the mean shape is used as the initializations of the shapes. Coarse-to-Fine Auto-Encoder Networks (CFAN) [9] utilizes a Stacked Auto-encoder Network [13] to predict the face shape quickly by taking a whole face as input. DCNN [1] employs a deep CNN model to extract high-level features to make accurate predictions as the initialization. After the initialization, the DCNN designs two-level convolutional networks to refine each landmark separately by taking local regions as input. To train these networks, several factors are critical for achieving good performance. For example, Sun et al. [1] conduct extensive experiments to investigate different network structures which are the basic regression units. The input range of local regions and the selecting strategy of local patches for training are other main factors having great impacts on the accuracy and reliability. But these factors are set by intuition or empirically in traditional methods. Besides, the relationship between any two successive networks is less developed.

In this paper, we propose an Adaptive Cascade Deep Convolutional Neural Networks (ACDCNN) for facial point detection. After initializing the shape by a CNN model like DCNN, each landmark is refined by a series of networks. These networks take the output of previous networks as input and locate a new position of the landmark. Different from

☆ The work is sponsored by the Chinese NSFC project 61372169.
* Corresponding authors.
E-mail addresses: yuandong@bupt.edu.cn (Y. Dong), wuyuebupt@gmail.com (Y. Wu).

http://dx.doi.org/10.1016/j.csi.2015.06.004
0920-5489/

existing methods [1,9] which apply the same configuration of regression for each landmark or each facial component in a stage, we set the configurations according to different results of each landmark. In addition, a Gaussian distribution is used to model the output error of the previous network. The input range of the local region is related to the expectation and the standard deviation of this Gaussian distribution. After the input range is determined, patches centered at positions shifted from the ground truth position are taken for training. Instead of taking these patches randomly, they are fetched under that Gaussian distribution. Thus the most relevant image patches are selected for training the successive network. These better training samples lead to better performance. The comparison experiments show that the proposed ACDCNN outperforms or is comparable to the state-of-the-art methods on both robustness and accuracy.

The rest of the paper is organized as follows. Section 2 introduces related work followed by our proposed ACDCNN introduced in Section 3. The Implementation details are described in Section 4. Section 5 reports our experimental results followed by conclusion in Section 6.

2. Related work

Many approaches to face alignment have been reported in the past decades among which regression based methods show highly efficient and accurate performance thus have received increasing attentions. Valstar et al. [14] develop support vector regression to model the nonlinear transform from the input local features (Haar-like features) to target point locations. Dantone et al. [15] extend the regression forests [16] to conditional regression forests. Head poses are utilized in the framework as the prior probability. Cao et al. [17] use cascaded random ferns as regressors and take the shape-indexed features [12] extracted from the whole image as input. Ren et al. [8] propose the use of discriminative binary features with locality principle to further improve the accuracy and speed. Xiong et al. [7] develop the supervised descent method (SDM) with SIFT, and implement the regression procedure in a gradient descent view. Recently, an emerging field is deep models like convolutional networks widely used in computer vision applications such as image classification [18,19], object detection [20], scene recognition [21], face alignment [1,9,10] and face verification [22]. Sun et al. [1] propose the DCNN for point detection in a coarse-to-fine manner. Zhang et al. [9] develop Coarse-to-Fine Auto-Encoder Networks (CFAN), which utilizes several stacked auto-encoder networks to deal with the nonlinearity in inferring face shapes from face images. Zhang et al. [10] investigate the possibility of jointly optimizing facial landmark detection with a set of related task and propose a Tasks-Constrained Deep Convolutional Neural Network (TCDCNN).

We want to note that SDM [7] employs Normal distribution to generate training samples. However it only samples for initialization during training and the parameters of the Normal distribution capture the variance of a face detector. We argue while SDM may increase the robustness, the random sample in the whole process instead of only the initialization may further improve the robustness. Secondly, SDM uses the mean shape as the initialization to extract shape-indexed features, i.e. SIFT. The initialization is rough and may be far from the ground truth. In addition, SIFT is a hand-crafted feature which may be limited to complex shape with high nonlinearity. In DCNN [1], the configuration of input local patches and training parameters applied to each landmark is the same at the same level, which may ignore the different local appearance of each landmark. Secondly, DCNN selects the local training patches randomly by setting a maximum shift in both horizontal and vertical directions empirically, which lacks of intelligence guidelines. TCDCNN [10] jointly optimizes facial landmark detection with a set of related tasks requiring large number of labels of the data in training. CFAN [9] utilizes deep Auto-Encoder networks for the regression to model the complex nonlinearity between the SIFT features and the increment of the current shape. Deep Regression [23] makes use of a multi-stage structure based on linear regression and use back-propagation algorithm to jointly optimize the parameters. Both CFAN and Deep Regression use hand-crafted features which may suffer the same as that of SDM.

In this research, we propose a novel Adaptive Cascade Deep Convolutional Neural Networks (ACDCNN), which is based on DCNN, augmented with sampling strategies in every cascade of two successive networks. The advantages are two-fold. First, ACDCNN utilizes the whole face as input to predict the initial face shape, which is employed in DCNN. The global high-level feature extracted with a new deep network structure is capable to handle faces taken under complex environment. Secondly, ACDCNN models the error of the previous output with a Normal distribution and selects these patches under this distribution in any two consecutive networks (instead of only in the initiation). This sampling strategy can refine the network structure and its parameters adaptively thus improved robustness is expected.

3. Adaptive Cascade Deep Convolutional Neural Networks

In this section, we present a novel method termed ACDCNN for face alignment. The details of each component of ACDCNN, including the initialization with a Deep Convolutional Neural Network and the Local Adaptive Cascade Networks (LACN) are explained.

3.1. Method overview

As shown in Fig. 1, the proposed ACDCNN attempts to design a unified cascade pipeline for each facial point, with the regression in each stage modeled as a deep convolutional network. There are five facial points to detect: *left eye center* (LE), *right eye center* (RE), *nose tip* (N), *left mouth corner* (LM), and *right mouth corner* (RM). At the first level, a deep convolutional network is employed to predict all five facial landmarks simultaneously. The whole face is taken as input and high-level features learned from raw RGB pixels are utilized to make predictions. After getting an estimation of face shape from the first level, each facial landmark is refined by a set of networks. These networks take local patches centered at the predicted positions of the facial point from previous levels as input. Each set of networks for facial points is independent because local appearance of each landmark is different. And the error of each point in previous networks is distinguished from each other. To characterize the variations, each output error is modeled by Gaussian distribution. The following networks take the expectation and the standard deviation of this Gaussian distribution into consideration to estimate a better configuration for training. The iteration will stop once the error is no longer reduced.

3.2. Initialization with DCNN

The first network directly estimates the face shape by taking the whole face as input. Given a face image $x \in \mathrm{R}^{m \times 1}$ of m pixels, $\mathrm{d}(\mathrm{x}) \in \mathrm{R}^{2p \times 1}$ denotes p landmarks ($p = 5$) in the face image. The task of facial landmark detection is to learn a mapping function F from the image to the face shape as follows:

$$F : d(x) \leftarrow x.$$

Due to the complexity and nonlinearity of the mapping function, a deep convolutional network is developed. Raw RGB pixels are taken as input by the network and coordinates of the five points (one for each landmark) are the outputs. The architecture of the network contains six learned layers — three convolutional and three fully-connected. Each convolutional layer applies square convolution kernels (or filters) to the multichannel input feature maps and output the responses. Specifically, let the input layer be $I(h, w, l)$, where h, w and l are the height, width and depth of the input. Convolutional layer is denoted by $F(s, k, n)$, where s is the side length of the convolution kernel, k is the depth of the convolution kernel, n is the number of kernels in the