Pattern Recognition 48 (2015) 1277–1288

Contents lists available at ScienceDirect

Pattern Recognition

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

# Hierarchical facial landmark localization via cascaded random binary patterns

Zhanpeng Zhang [a,b], Wei Zhang [c,*], Huijun Ding [d], Jianzhuang Liu [a,c], Xiaoou Tang [a,b]

[a] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[b] Shenzhen Key Laboratory for Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P.R. China
[c] Media Technology Lab, Huawei Technologies Co. Ltd., P.R. China
[d] Institute of Biomedical Engineering, Medical College, Shenzhen University, P.R. China

ARTICLE INFO

*Article history:*
Received 13 September 2013
Received in revised form
22 July 2014
Accepted 8 September 2014
Available online 18 September 2014

*Keywords:*
Facial landmark localization
Random binary pattern
Hierarchical regression
Gradient boosting decision tree

ABSTRACT

The main challenge of facial landmark localization in real-world application is that the large changes of head pose and facial expressions cause substantial image appearance variations. To avoid high dimensional facial shape regression, we propose a hierarchical pose regression approach, estimating the head rotation, face components, and facial landmarks hierarchically. The regression process works in a unified cascaded fern framework with binary patterns. We present generalized gradient boosted ferns (GBFs) for the regression framework, which give better performance than ferns. The framework also achieves real time performance. We verify our method on the latest benchmark datasets and show that it achieves the state-of-the-art performance.



## 1. Introduction

Automatic facial landmark detection/localization is a long-standing problem in computer vision. It plays a key role in face recognition systems and many other face analysis applications. In [1], it has been shown that the performance of face recognition can be remarkably elevated when facial landmark locations can be utilized. In the application of facial attribute analysis [2,3], precise facial landmark locations need to be found for feature extraction. In [4], the facial landmarks are used as the input to drive the animation of a 3D avatar. For the above reasons, the problem of facial landmark localization has been extensively studied during the past decades, and great improvements have been achieved on the standard benchmarks, such as BioID [5], LFPW [6], AFLW [7] and 300-W [8]. However, the large variations of face appearance caused by illumination, expression, and out-of-plane rotation make the robust and accurate localization in real-world applications still a challenging task.

Recently, explicit regression based methods have achieved the state-of-the-art performance for accurate and robust face alignment. The basic framework of these methods is to treat the landmark localization as a regression task: Let $S$ be a parametric face shape. For a given input image $I$ with an initial shape estimation $S^0$, $S$ is progressively refined by cascaded regressors $\phi$ at stage $t$:

$$S^t = S^{t-1} \circ \phi^t(f^t(I, S^{t-1})), \tag{1}$$

where $f$ represents a feature extraction function, such as SIFT [9], HOG [10], and binary feature [11–14].

Compared with the generative model based methods, such as ASM [15] and AAM [16], this framework has the following advantages: (a) since it incorporates facial appearance in a reasonable coarse-to-fine manner, the regression strategy avoids large computation caused by local window search or model fitting; (b) global facial context is incorporated into the regression at the beginning; during the cascaded regression stages, the facial context is refined from coarse to fine so that it is constrained to a local region for precise landmark localization; (c) it is capable of handling a large amount of training data, which improves the generalization power when used in real world scenarios.

However, since the above approaches utilize global regressors for shape regression, they might suffer from the high dimensional regression problem when a large number of landmark points are required: Firstly, the high dimensional regression training cost might be unaffordable if we need to learn the features from large training data; Secondly, it can easily cause overfitting and hurt generalization ability during testing. In addition, it might not be the optimal strategy to use a global regression during the whole landmark localization process, because the face shape is refined in local regions during the latter stages of the regression. For example, it does not make sense that the local features in the

* Corresponding author. Tel.: +86 755 86392199; fax: +86 755 86392073.
*E-mail addresses:* zz013@ie.cuhk.edu.hk (Z. Zhang), zhangwei@siat.ac.cn (W. Zhang), hjding@szu.edu.cn (H. Ding), liu.jianzhuang@huawei.com (J. Liu), xtang@ie.cuhk.edu.hk (X. Tang).

http://dx.doi.org/10.1016/j.patcog.2014.09.007
0031-3203/

components of the eyes will influence the position of the mouth. In [11,6], a non-parametric shape prior is utilized to handle the high dimensional regression and it achieves the state-of-the-art performance.

In this paper, we propose a new regression framework to locate facial landmarks for real world applications. To handle the high dimensional face shape regression problem, we estimate facial landmarks in a hierarchical way, where the high dimensional shape is decoupled into a set of low dimensional parameters, which includes head rotation, facial component location and the whole facial landmark position. In the remaining parts of the paper, the head rotation and the locations of facial components and landmarks together are referred to as facial pose. Fig. 1 shows the overview of the framework. There are three levels in the hierarchical pose regression: head rotation, face components, and facial landmarks. In each level, we estimate the pose using generalized Gradient Boosted Ferns (GBFs). The motivation for our hierarchical structure is that the image appearance variations can be reduced in each level gradually. Besides, reducing the regression dimension also makes the learning process easier. Specifically, with the head rotation estimated in the top level, we obtain the conditional probability over the whole view space. Then we estimate the rest pose parameters with the view-based GBFs in level 2 and level 3. Also, in level 2, we estimate the locations of a few facial components, further constraining the regression space for level 3. The recent work [17] is especially related to our approach in its hierarchical strategy for shape regression. The high dimensional face shape input is decoupled into a set of facial components and the pose estimation is also performed in the final refinement stage. The deep convolutional neural network (CNN) [18] is used for the cascaded regression. Different from [17,18], our approach does not need the heavy computation used by CNN. Also, it works in a unified framework and does not need to crop the facial component patches in the cascaded stages for regression, which also saves substantial computation.

In the experimental section, we will show that using simple binary features with tree-based regression approaches can efficiently handle the high dimensional shape input. The proposed method is evaluated on the latest challenging datasets of [19,1,8] and achieves the state-of-the-art performance.

## 2. Related work

Early work on facial landmark detection is often treated as a component of face detection. Burl et al. [20] develop a bottom up approach for face detection where it needs to first detect candidate facial landmarks over the whole image. Gabor filters [21] have been applied to large-scale facial parts such as eyes, nose, and mouth. Without the global shape constraints, false alarm is the main challenge for these component based detection approaches, even for well-trained detectors.

To better handle larger pose variation, constraints can be built on the relative locations between facial components. It can be expressed as predicted locations of one facial component given another location [21]. In [7], the DPM [22] style detector is used for multi-view facial landmark detection and pose estimation simultaneously. Alternatively, the constraints can also be built on the joint distribution of all facial components. When such constraints are modeled as a multivariate normal distribution, it results in the well-known Active Shape Model (ASM) [15,23] and Active Appearance Model (AAM) [16,24,25]. ASM is extended in [26,27] by using a Gaussian Mixture Model for shape distribution whereas [28] utilizes a mixture of Gaussian trees to describe the relation between landmark positions and the face bounding box. Non-parametric shape constraint derived directly from training samples is used in [6].

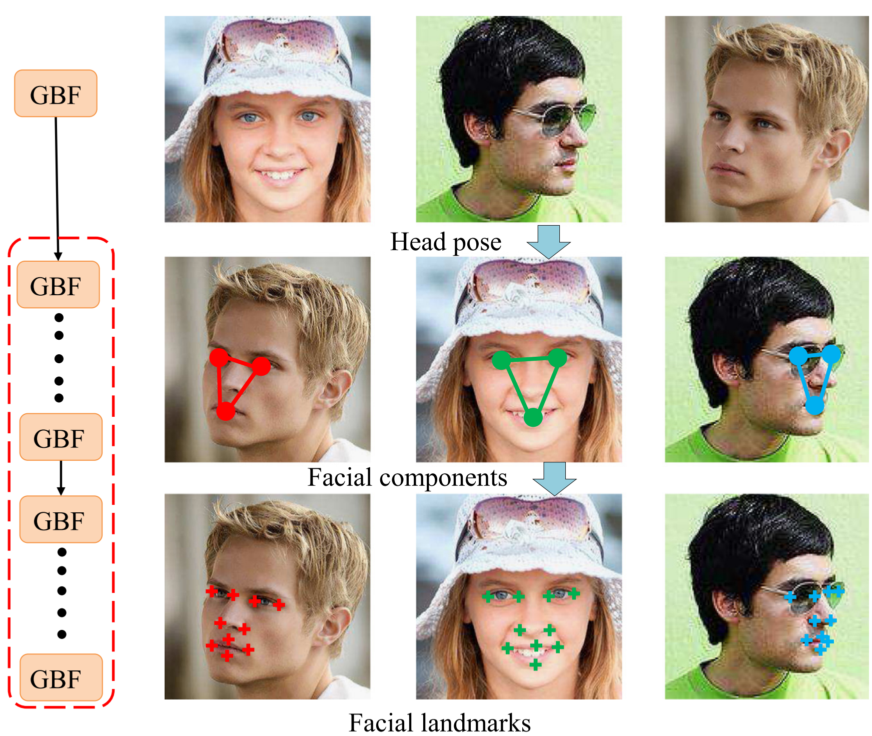


**Fig. 1.** Overview of our hierarchical pose regression approach, which is based on a unified framework with sequential groups of generalized gradient boosted ferns (GBFs). The conditional view-based GBFs are enclosed by the red rectangle on the left. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)