Image and Vision Computing 47 (2016) 19–26

Contents lists available at ScienceDirect

Image and Vision Computing

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

M^3 CSR: Multi-view, multi-scale and multi-component cascade shape regression☆

Jiankang Deng [a], Qingshan Liu [a,*], Jing Yang [a], Dacheng Tao [b]

[a] B-DAT Laboratory, School of Information and Control, Nanjing University of Information and Technology, Nanjing 210044, China
[b] QCIS Laboratory, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 17 August 2015
Accepted 30 November 2015
Available online 15 December 2015

Keywords:
Face alignment
Cascade shape regression
Multi-view
Multi-scale
Multi-component

ABSTRACT

Automatic face alignment is a fundamental step in facial image analysis. However, this problem continues to be challenging due to the large variability of expression, illumination, occlusion, pose, and detection drift in the real-world face images. In this paper, we present a multi-view, multi-scale and multi-component cascade shape regression (M^3CSR) model for robust face alignment. Firstly, face view is estimated according to the deformable facial parts for learning view specified CSR, which can decrease the shape variance, alleviate the drift of face detection and accelerate shape convergence. Secondly, multi-scale HoG features are used as the shape-index features to incorporate local structure information implicitly, and a multi-scale optimization strategy is adopted to avoid trapping in local optimum. Finally, a component-based shape refinement process is developed to further improve the performance of face alignment. Extensive experiments on the IBUG dataset and the 300-W challenge dataset demonstrate the superiority of the proposed method over the state-of-the-art methods.

1. Introduction

Automatic facial landmark localization plays an important role in facial image analysis [1–4]. A lot of methods [5–10] have been proposed, achieving remarkable improvements [11,12] on standard benchmarks in the past two decades. Existing methods can be roughly divided into three categories: generative methods, discriminative methods and statistical methods [13]. *Generative methods* attempt to optimize the shape parameter configuration by maximizing the probability of a face image being reconstructed by a facial deformable model. Active Shape Model (ASM) [14] and Active Appearance Model (AAM) [15–18] are two representative generative methods. *Discriminative methods* try to infer a face shape through a discriminative regression function, which directly maps a face image to the landmark coordinates [19–24]. There are two popular ways to learn such a regression function. One is based on deep neural network learning [25–28], the other is the well-known cascade shape regression model, which aims to learn a set of regressors to approximate complex nonlinear mapping between the initial shape and the ground truth [29–31]. The idea of *Statistical methods* is to combine both generative and discriminative methods, trying to fit the shape model on a statistical way after learning patch experts. The most notable example is probably the Constrained Local Model (CLM) [32–34] paradigm, which represents the face via a set of local image patches cropped around the landmark points. Recent research efforts have been made on the collection, annotation and alignment of face images captured in-the-wild [12]. However, face alignment is still challenging due to the large variability of expression, illumination, occlusion and pose in the real-world face images [11].

An automatic face alignment system also suffers from the performance of face detector, because its initialization is usually based on the output of face detector. Challenging factors such as pose, illumination, expression and occlusion also have great effects on the performance of face detection [35]. Moreover, face detection is often determined by the criterion [36] that the ratio of the intersection of a detected region with an annotated face region is greater than 0.5. As shown in Fig. 1, all of the faces are detected according to the criterion of 0.5 overlap, but there is more or less drift with the detection results. When the detection result is largely drifted from the ground truth, it is actually not accurate enough for the initialization of face landmark localization algorithm.

In this paper, we propose a robust face landmark localization algorithm. The proposed method is based on the popular cascade shape regression model, and we try to further improve its robustness from three aspects. Firstly, we develop a robust deformable parts model (DPM) [37,35] based face detector to provide a good shape initialization for face alignment. We also utilize the deformable parts information to predict the face view, so as to select the view-specific shape model. View based shape model is not only able to decrease the shape variance, but also can accelerate the shape convergence. Secondly, we develop a multi-scale cascade shape regression with multi-scale HOG features [38]. Multi-scale HOG features can incorporate local structure information implicitly, and multi-scale cascade shape regression helps to

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.
* Corresponding author.
E-mail addresses: jiankangdeng@gmail.com (J. Deng), qsliu@nuist.edu.cn (Q. Liu), yang.xiaojing00@gmail.com (J. Yang), Dacheng.Tao@uts.edu.au (D. Tao).

http://dx.doi.org/10.1016/j.imavis.2015.11.005
0262-8856/

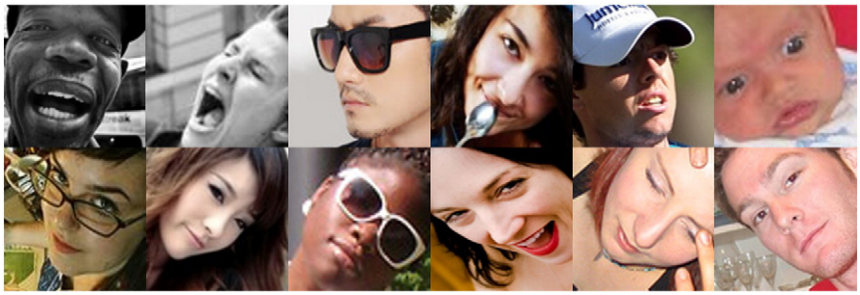

Fig. 1. Successful face detection results with more or less drift.

avoid trapping in local optimum. To further improve the performance of face alignment, a refinement process is conducted on facial components, such as mouth. The proposed methods achieve the state-of-the-art performance on the challenging benchmarks including the IBUG dataset and the 300-W dataset.

The rest of the paper is organized as follows. The related work is reviewed in Section 2. Cascade shape regression model with multi-view, multi-scale, and multi-component are presented in Section 3. Experimental results are shown in Section 4, and finally the conclusion is drawn in Section 5.

2. Related work

The cascade shape regression model (CSR) has attracted much attention in recent years, because it has achieved much success in face alignment under uncontrolled environment [13]. In [29], Cascade Pose Regression (CPR) is first proposed to estimate pose with pose-indexed features, which iteratively estimates object pose update from the features on current pose. Explicit Shape Regression (ESR) [30] improves CPR by using a two-level boosted regression and correlation-based feature selection. The Supervised Descent Method (SDM) [31] uses linear cascade shape regressions with fast SIFT features, and interprets the cascade shape regression procedure from a gradient descent view [39]. Global SDM (GSDM) extends SDM by dividing the search space into regions of similar gradient directions and obtains better and more efficient convergence [40], which indicates that decreasing shape variation is helpful for cascade shape regressions. Yan et al. [38] utilize the strategy of "learn to rank" and "learn to combine" from multiple hypotheses in a structural SVM framework to handle inaccurate initializations from the face detector. In [41], highly discriminative local binary features are used to jointly learn a linear regression. Because extracting and regressing local binary features is computationally cheap, this method achieves over 3000 fps on a desktop. [13] proposes an Incremental Parallel Cascade Linear Regression (iPar-CLR) method, which incrementally updates all the linear regressors in a parallel way instead of the traditional sequential manner. Each level is trained independently by using only the statistics of the former level, and the generative model is gradually turned to a person-specific model by the recursive linear least-squares method. [42] proposes an ℓ_1-induced Stagewise Relational Dictionary (SRD) model to learn consistent and coherent relationships between face appearance and shape for face images with large view variations. Yu et al. [43] propose an occlusion-robust regression method by forming a consensus estimation arising from a set of occlusion-specific regressors. Robust Cascade Pose Regression (RCPR) [44] reduces exposure to outliers by explicitly detecting occlusion on the training set marked with occlusion annotations. Substantially, CSR is a procedure of shape variance decreasing. In this paper, we develop a robust CSR for face alignment, in which multi-view, multi-scale and multi-component strategies are carefully designed to decrease shape variance.

3. M^3CSR model

Although the cascade shape regression model has achieved much success in face alignment [31], it is still sensitive to some large variations, such as illumination, pose, expression, and occlusion which often exist in real-world images, as well as shape initialization from face detector [38]. In this paper, we propose a new M^3CSR model to make CSR more robust to the real-world variations. Its work flow is illustrated in Fig. 2, in which we enrich the system from three steps. The first step is to develop a reliable face detection and view estimation algorithm to provide a view specified initialization and a view specified cascade shape regression. The second step is to design multi-scale cascade shape regressions with multi-scale HOG features. The last step is to refine facial components to obtain more accurate results.

3.1. Cascade shape regression

The main idea of CSR is to combine a sequence of regressors in an additive manner to approximate complex nonlinear mapping between the initial shape and the ground truth. Specifically, in [31,38], a linear regression function is iteratively used to minimize the mean square error:

$$\arg\min_{W^t} \sum_{i=1}^{N} \left\| \left(X_i^* - X_i^{t-1}\right) - W^t \Phi\left(I_i, X_i^{t-1}\right) \right\|_2^2,$$

where N is the number of training samples, $t = 1, \cdots, T$ is the iteration number, X_i^* is the ground truth shape, X_i^0 is the initialization of face shape, Φ is the shape-index feature descriptor, W^t is the linear transform matrix, which maps the shape-indexed features to the shape update. This is a linear least squares problem, and W^t has a close-form solution. During testing, the shape update is iteratively calculated by linear regressions based on the shape-indexed features.

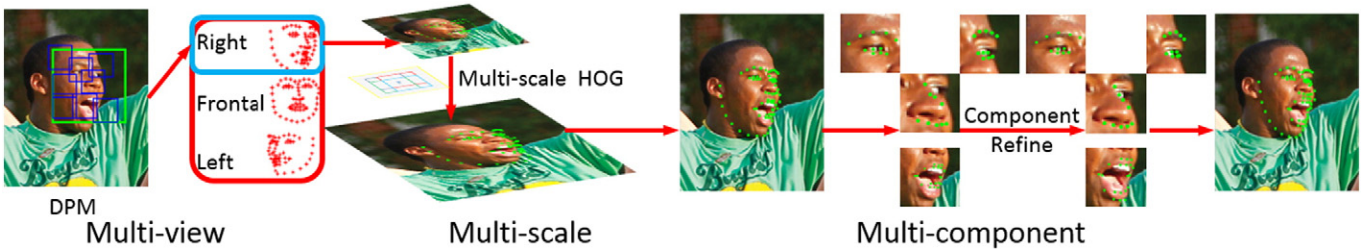

Fig. 2. The work-flow of M^3CSR.