Image and Vision Computing 47 (2016) 60–70

Contents lists available at ScienceDirect

Image and Vision Computing

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

Multi-view facial landmark detection by using a 3D shape model

Jan Čech *, Vojtěch Franc, Michal Uřičář, Jiří Matas

Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 21 December 2014
Received in revised form 16 September 2015
Accepted 30 November 2015
Available online 15 December 2015

Keywords:
Face
Landmarks
Detection
Localization
3D model
Shape
Occlusions

ABSTRACT

An algorithm for accurate localization of facial landmarks coupled with a head pose estimation from a single monocular image is proposed. The algorithm is formulated as an optimization problem where the sum of individual landmark scoring functions is maximized with respect to the camera pose by fitting a parametric 3D shape model. The landmark scoring functions are trained by a structured output SVM classifier that takes a distance to the true landmark position into account when learning. The optimization criterion is non-convex and we propose a robust initialization scheme which employs a global method to detect a raw but reliable initial landmark position. Self-occlusions causing landmarks invisibility are handled explicitly by excluding the corresponding contributions from the data term. This allows the algorithm to operate correctly for a large range of viewing angles. Experiments on standard "in-the-wild" datasets demonstrate that the proposed algorithm outperforms several state-of-the-art landmark detectors especially for non-frontal face images. The algorithm achieves the average relative landmark localization error below 10% of the interocular distance in 98.3% of the 300 W dataset test images.

1. Introduction

Facial landmarks refer to points on the face like the corners of the mouth, the corners of the eyes or the tip of the nose that can be annotated by a human. Detection of facial landmarks in images has many potential applications as e.g., animation, morphing, and it is an important step in most face image interpretation tasks. Face images usually need to be aligned and normalized with the help of landmarks prior to recognition of e.g., identity, age, gender, expression.

Most facial landmark detectors simultaneously model local appearance around the landmarks and their geometrical configuration. The local appearance is represented either by generative models (e.g. [1]) or by discriminatively trained detectors (e.g. [2]). The geometrical structure of the landmarks is usually modelled by a Point Distribution Model (PDM) [3], which describes the landmark positions on a face in canonical frame, and by a subsequent deviation from the canonical pose in 2D image coordinates. Both PDM of 2D shapes (e.g. [1,4]) and 3D shapes have been proposed (e.g. [5]).

Fitting the shape models into the image requires *optimization of a highly non-convex fitness function* typically carried out as gradient search sensitive to the initial estimate or by regression. The problem with local optima is mitigated either by re-initializing the optimization, by using global but expensive optimization methods (e.g. [6]), or by simplifying the shape model. A prominent example of a simplified 2D shape prior is the Deformable Part Model (DPM) [7] representing the shape by a pair-wise energy function whose *global optimum* can be found efficiently by dynamic programming. Excellent results of DPM based facial landmark detectors have been demonstrated e.g. in [8,9]. On the other hand, DPM detectors can describe only a limited range of face poses and thus a multi-view detector must be composed of several DPMs (e.g. [8]).

Recently, *regression* methods have been very successful. The methods avoid explicit local or global optimization and learn a cascade of regression functions that map the input image to the target output, which is either directly the landmark locations in the image [10,11] or indirectly a set of parameters of 2D [12] or 3D shape models [13]. Typically, pose-indexed (also known as shape-indexed) features are leveraged. In each stage of the cascade, a regressor extracts features from image locations relative to the current estimate of landmarks to predict a model update. Regression-based methods are usually much faster than optimization-based methods and often run faster than real-time [14,15]. However, like the local optimization methods, these methods require initialization. Another potential difficulty with the regression methods using pose-indexed features has to be overcome for training. Annotation of all landmarks is necessary for every image in the training set. Manual annotation is not always complete because of: (1) self-occlusions due to various pose changes – separate regressors need to be trained for poses with shared subsets of visible landmarks, or (2) occlusion by hairs, hands, or other objects – these images cannot be used, or simply (3) a different set of landmarks is often annotated for datasets – it is then typically not possible to combine multiple datasets for training.

Currently, the difference in performance of DPMs fitted by a global method, of the genuine 3D shape models fitted by local methods, and of the regression-based methods is not fully understood.

* Corresponding author.
E-mail addresses: cechj@cmp.felk.cvut.cz (J. Čech), xfrancv@cmp.felk.cvut.cz (V. Franc), uricamic@cmp.felk.cvut.cz (M. Uřičář), matas@cmp.felk.cvut.cz (J. Matas).

http://dx.doi.org/10.1016/j.imavis.2015.11.003
0262-8856/

Besides using the intensity image alone, there are approaches that work with RGB-D data (image + depth) to detect landmark and to align 3D faces, e.g. [16,17,18].

In this paper we show that a robust and precise landmark detector is obtained by fitting a simple 3D shape model into the image using a full perspective projection. The method jointly fits M shape parameters and the 6 DoF pose (position and orientation) of the 3D face model with respect to the camera, see Fig. 1.

In addition, we propose a novel method for discriminative learning of the local detectors that are used to guide the fitting of the model. We learn scoring functions whose value decreases approximately linearly with the Euclidean distance from the true landmark position. The method often produces unimodal peaks around the true landmark positions which helps to make the basin of attraction sufficiently large. The approach differs from the commonly used two-class classification methods, like the Support Vector Machines or AdaBoost, whose learning objective does not take the distance from the true landmark position into account.

Related to our approach is the work of [2], a local optimization-based method employing a 3D model, that has a single degree of freedom to capture the shape. However, a simpler camera model is used, weak perspective in [2] vs. full perspective, and a substantially different learning of the local landmark scoring functions is employed, a standard AdaBoost [2] vs. the novel learning method.

Another related work is a regression-based method [13] which uses a parametrization similar to ours, but again in conjunction with the weak perspective camera. Due to the nature of the regression that takes the entire face image as an input, this method is sensitive to self-occlusions. If the head is turned so that certain landmarks are not visible in the camera, the occluded landmarks cannot be easily disregarded. In the proposed method, a contribution of the occluded landmarks is easily switched off in the optimization data term. The property leads to accurate results on face images captured from arbitrary view-point and not only on near frontal images. To the best of our knowledge we are not aware of any landmark detector functioning reliably in a multi-view setup.

To summarize, the contributions of the paper are:

1. A novel precise local optimization-based algorithm coupled with a robust initialization scheme based on a global method is proposed. The initialization gives a raw estimate. Thanks to its global optimality it is likely to be free of outliers.
2. A novel method for learning local landmark detectors is introduced. It produces smooth unimodal score functions with a large basin of attraction.
3. Self-occlusions are explicitly taken into account, which results in an algorithm operating for a broad set of viewing angles, e.g. semi-profile and profile views.
4. A thorough comparison of the proposed method with state-of-the-art implementations of two different DPM based detectors [9,8], and with two recent regression-based method [13] and a multiview implementation of [11] is performed. The proposed method and [9] use the same local detectors, but differ in the used shape prior. The proposed method and [13] have a similar parametric model

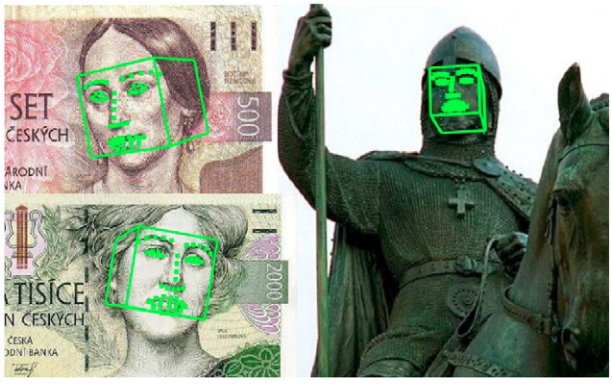

Fig. 1. The proposed method jointly estimates the position of 49 facial landmarks in the image and the head pose (position and orientation) with respect to the camera coordinate system. The landmarks are shown as crosses and the pose is visualized by projecting a virtual 3D cube around the head into the image. The method is robust, the landmark scoring function generalizes even to images of an artistic engraving or a bronze statue.

This paper is an extension of [19] with improved 3D model parametrization to capture a wide range of subjects and facial expressions. The optimization scheme is robustified by introducing a reliable initialization strategy. A range of applicable angles is extended by modelling the landmark visibility. We are now estimating up to 49 landmarks as opposed to 7 landmarks in [19].

The rest of the paper is structured as follows: The algorithm is presented in Section 2, the justification of the design choices is discussed in Section 3, and its implementation details are given in Section 4. Experimental validation, including both the introspection and comparison with several state-of-the-art methods, is presented in Section 5. Finally, Section 6 concludes the paper.

2. Facial landmarks and a head pose estimation

The estimation problem addressed entails: (1) localization of landmarks in the image, (2) estimation of the head pose, i.e., a position and orientation with respect to the camera coordinate system, (3) reconstruction of the landmark points in 3D. Quantities (1–3) are calculated from a single image.

The architecture of the proposed algorithm is depicted in Fig. 2. The pipeline consists of two stages: the initialization and the optimization. The three problems (the landmark detection, head pose estimation, and 3D reconstruction) are coupled and are solved as a single optimization problem, see Section 2.2. A parametric 3D shape model is fit to maximize the values of landmark scoring functions. Each landmark scoring function takes the image and for a query pixel returns a score proportional to how likely the landmark occurrence centred at the pixel is, see Section 2.1. The proposed criterion is non-convex, therefore a robust initialization procedure is needed, see Section 2.3. The initialization integrates a multi-view face detector and an implementation of a state-of-the-art DPM. We choose a simpler 2D landmark model working in a low image resolution, which provides a globally optimal solution.

2.1. Landmark scoring functions

Let us define the landmark score function $c_i(\mathbf{x}, I)$ which estimates the likelihood of the i-th landmark being at position $\mathbf{x}$ in the image I. The most likely position is $\hat{\mathbf{x}}_i = \text{argmax}_{\mathbf{x}\in\mathcal{X}_i} c_i(\mathbf{x}, I)$ where $\mathcal{X}_i$ denotes the searched positions. We consider a linearly parametrized score.

$$c_i(\mathbf{x}, I; w_i) = \langle \Psi(\mathbf{x}, I), w_i \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes a dot product, $\Psi(\mathbf{x}, I) \in \Re^n$ denotes a feature descriptor extracted from a patch cropped from the image I around the position $\mathbf{x}$ and $w_i \in \Re^n$ is a weight vector associated with the i-th landmark. We construct the descriptor $\Psi(\mathbf{x}, I)$ by concatenating the Local Binary Patterns (256 valued code assigned to a 3×3 patch) computed at all positions of the cropped patch rescaled to size 20×20, 10×10 and 5×5 pixels, respectively. By this process we obtain $256(18^2 + 8^2 + 3^2)$-dimensional sparse ($18^2 + 8^2 + 3^2$ non-zero elements) binary feature descriptor whose values are to some extent robust against scale and lighting conditions. The side of the cropped squared patch is 0.3 of the bounding box side returned by the face detector.