Image and Vision Computing 51 (2016) 69–83

Contents lists available at ScienceDirect

Image and Vision Computing

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

Robust face alignment and tracking by combining local search and global fitting ☆

Jongju Shin, Daijin Kim*

Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Gyeongbuk 790-784, South Korea

ARTICLE INFO

Article history:
Received 22 May 2015
Received in revised form 7 April 2016
Accepted 15 April 2016
Available online 3 May 2016

Keywords:
Face alignment
Face tracking
Local search method
Global fitting method
Combination of local search and global fitting methods

ABSTRACT

When a face in an image is considerably occluded, existing local search and global fitting methods often cannot find the facial features due to failures in the local facial feature detectors or the fitting limitations of appearance modeling. To solve these problems, we propose a new face alignment method that combines the local search and global fitting methods, where local misalignments in the local search method are restricted by holistic appearance fitting in the global fitting method and the divergent or shrinking alignments in the global fitting method are avoided by the restricting local movements in the local search method. The proposed alignment method consists of two stages: the initialization stage detects the face, estimates the facial pose and obtains the initial facial features by locating a pose-specific mean shape on the detected face; the optimization stage then obtains the facial features by updating the parameter set from the combined Hessian matrix and the combined gradient vector. We also extend the proposed face alignment to face tracking by adding a template image that is warped from the facial features obtained in the previous frame. In the experiments, the proposed method yields more accurate and stable face alignment or tracking under heavy occlusion and pose variation than the existing methods.

1. Introduction

Face alignment is an essential task in applications such as face recognition, facial expression recognition, head pose estimation, and head gesture understanding. The solution space of these applications can be reduced by locating facial features such as the eyes, eyebrows, nose, and mouth corners in the face images. Methods for solving this task are categorized into two approaches: local search and global fitting.

Local search method finds facial features by 1) detecting the face, 2) locating the mean shape on the detected face to obtain the initial facial features, 3) deploying a local facial feature detector within the search region of each facial feature, to estimate the facial features, 4) determining the current parameter set by projecting the estimated facial features onto the shape basis vectors, and 5) obtaining new facial features from the current parameter set. Steps 3), 4), and 5) are repeated until the facial features stop changing.

A typical example of the local search method is the active shape model (ASM) [1], which estimates the facial features by finding the smallest Mahalanobis distance between the input image and the trained template along the line perpendicular to each facial feature. A recent example of a popular local search method is the constrained local model (CLM), which uses the subregion as the search space for each facial feature and optimizes the parameter set to maximize the sum of the responses of each local facial feature detector.

The original CLM [2] used a patch-based appearance model and obtained the responses from the templates that were most similar to the current input face patches, then used the simplex method to optimize the parameter set. Because this CLM used the local patches around the facial features, it was robust to illumination changes. Wang et al. [3] proposed a similar method that used convex quadratic fitting (CQF) to approximate the parameter set as a 2D Gaussian model. The CQF provided better alignment of the facial features than the ELS method. Saragih et al. [4] represented each facial feature with a nonparametric kernel density estimation (KDE) model and used a mean-shift algorithm to optimize the facial features. The nonparametric representation using KDE aligned the facial features better than the parametric model using CQF. Zhu and Ramanan [5] used the histogram of gradients (HOG) as a feature descriptor and applied an efficient dynamic programming algorithm to a tree model to find the optimal facial features. Asthana et al. [6] used the HOG as a feature descriptor and used linear support vector regression to learn functions from the facial feature maps to update the parameter set. Belhumeur et al. [7] used the SIFT algorithm with the SVM to find the local facial feature candidates and used

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou, PhD.
* Corresponding author.
E-mail addresses: jjshin@postech.ac.kr (J. Shin), dkim@postech.ac.kr (D. Kim).

http://dx.doi.org/10.1016/j.imavis.2016.04.012
0262-8856/

exemplar-based optimization to find the optimal facial features; this method detected facial features well but was slow (∼1 s for each local facial feature detection). Baltrusaitis et al. [8] extended Saragih's alignment method [4] by using 3D depth information and combining the appearance model with head pose tracking. Valstar et al. [9] proposed a regression-based local facial feature detector and improved the alignment accuracy using Markov random fields. Wu et al. [10] proposed a novel face tracking method that obtained the shape model from the restricted Boltzmann machines (RBM), and used a simple Kalman filter and local image patches as face templates.

It is generally recognized that the local search method is less sensitive to variations in illumination, pose, and occlusion and achieves better face alignment than the global fitting method [11] when the current facial features are assumed to be close to the true facial features. However, this assumption is not always valid because the current facial features cannot be correctly detected in a face image when the true facial features are not within the search region. Furthermore, the local search method often fails to align the face under occlusion because it assumes that there is no significant occlusion, either explicitly or implicitly, yet occlusion can cause incorrectly-detected facial features to be involved in the optimization step.

The global fitting method finds facial features by 1) detecting the face, 2) locating the mean shape on the detected face to obtain the initial facial features, 3) computing the appearance error of the warped image by using the current parameter set and the appearance instance, 4) updating the current parameter set from the appearance error image, and 5) obtaining new facial features from the current parameter set. Steps 3), 4) and 5) are repeated until the facial features stop changing.

A typical example of the global fitting method is the active appearance model (AAM) [12], which uses least-squares regression to determine the facial features by training using the appearance error and the parameter displacements. Many variants of the AAM have been proposed. Matthews and Baker [13] applied the Lucas–Kanade algorithm [14] to face alignment and proposed an inverse compositional project out algorithm for optimization. The execution speed was very fast because the most computationally expensive parts, such as computing the inverse of the Hessian matrix and the steepest descent images, were pre-computed. Baker et al. [15] proposed a simultaneous algorithm that updated the Hessian matrix and the steepest descent images at each iteration, so that the approximated parameter set was close to the ground truth. This algorithm was more accurate but slower than the project out algorithm. Xiao et al. [16] used a 3D shape model obtained from the structure-from-motion algorithm [17], which generalized the 2D shape model. Cootes and Taylor [18] proposed a multi-channel AAM that used the edge and the gradient images instead of the intensity image. Their method was more accurate than using the intensity in cases where there were significant variations in lighting or expression. Navarathna et al. [19] combined the Gabor filter responses and achieved better alignment than the traditional approaches with variations in illumination.

The global fitting method can fail to converge to the true facial features because the solution space is high-dimensional and includes numerous local minima. Face alignment accuracy is reduced when the input image is dissimilar to the trained appearance model because it is almost impossible to span the entire space of the appearance by the trained appearance model [20].

To overcome the limitations of the local search and global fitting methods, we combine them to produce a novel method that achieves the accurate and stable alignment of facial features under occlusion. For a specific pose, the proposed face alignment method operates as follows. 1) A multi-view face detector for large pose variation is used to detect the face and to guess the initial facial features from the determined facial pose. 2) Pose-specific multi-block (MB) modified census transform (MCT)-based local facial feature detectors are used within each feature's search region to obtain the candidates for each facial feature that are approximated by their independent Gaussian models. 3) The best initial parameter set of the input image is estimated by generating multiple hypotheses, each of which is obtained from a pair of randomly selected facial feature candidates. Each hypothesis is evaluated by taking the median value of the errors between the facial features generated from the hypothesis and the nearest facial feature candidates. The best hypothesis, that is, the one that has the smallest median facial feature error, is adopted. 4) The visible facial features with errors not exceeding the minimum median error are identified. 5) The parameter set, including scale, rotation, translation, shape and appearance parameters, is updated using the combined Hessian matrix and the combined gradient vector, which are obtained from the visible facial features and appearance. Steps 4) and 5) are repeated until the facial features stop changing.

Recently, many other researchers try to combine the local search method and the global fitting method. Alabort-iMedina and Zafeiriou [21] combined the holistic model and the part-based model which is similar with our proposed method. However, our proposed method considers whether each facial feature is occluded or not in the optimization whereas their method does not consider this and our proposed method gives the best initial parameter from the facial feature candidates whereas their method gives initial parameter from the mean shape with Gaussian noise. Xiong and de la Torre proposed the global supervised descent method (GSDM) [22] which extends their previous supervised descent method (SDM) [23], GSDM learns several descent maps whereas SDM learns only one descent map. Therefore, GSDM gives more robust to pose variation than SDM but their method does not consider the occlusion and can not guarantee the good alignment performance under the occlusion. Tzimiropoulos [24] proposed to combine the cascased regression method with the project out method. Their method learns the shape parameter's update using the difference between the current facial feature's descriptor and the appearance model's descriptor but uses only local descriptor for face alignment whereas our proposed method uses the local and global descriptors simultaneously. Tzimiropoulos and Pantic [25] used the local appearance model which uses the SIFT descriptor as a feature and used Gauss–Newton-based optimization to find facial features. Their method also uses only local descriptor for face alignment whereas our proposed proposed method uses the local and global descriptors simultaneously.

The main contribution of this paper is to combine the local search and global fitting methods, where the local search method estimates the facial features and the global fitting method minimizes the appearance error. This combined method improves the accuracy of face alignment under severe occlusions and identifies whether each facial feature is occluded.

The remainder of this paper is organized as follows. Section 2 explains the face alignment process using the local search method. Section 3 explains the face alignment process using the global fitting method. Section 4 describes the proposed face alignment process using the combined local search and global fitting methods. Section 5 explains the face tracking method, which extends the proposed face alignment method. Section 6 reports extensive experimental results obtained using the proposed face alignment and tracking methods. Section 7 presents some conclusions and outlines future work.

2. Face alignment using local search method

We assume that the face has a non-rigid 3D shape $\mathbf{S}$ with N vertices as

$$\mathbf{S} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N] \in R^{3N}, \quad \mathbf{X}_i = [X_i,\ Y_i,\ Z_i]. \tag{1}$$