Neurocomputing 149 (2015) 1535–1543

Contents lists available at ScienceDirect

Neurocomputing

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

Monocular face reconstruction with global and local shape constraints ☆

Jian Zhang [a], Dapeng Tao [b,c,*], Xiangjuan Bian [a], Xiaosi Zhan [a]

[a] School of Science and Technology, Zhejiang International Studies University, Hangzhou 310012, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[c] The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form
30 June 2014
Accepted 12 August 2014
Communicated by Mingli Song
Available online 6 September 2014

Keywords:
3D reconstruction
Reference 3D face
Global shape constraint
Local shape constraint
Digital entertainment

ABSTRACT

To reconstruct 3D face from single monocular image, this paper proposes an approach which comprises three steps. First, a set of 3D facial features is recovered from 2D features extracted from the image. The features are recovered by solving equations derived from a regularized scaled orthogonal projection. The regularization is achieved by a global shape constraint exploiting a prior reference 3D facial shape. Second, we warp a high-resolution reference 3D face, using both recovered 3D features and local shape constraint at each model points. Last, realistic 3D face is obtained through texture synthesis. Compared with existing approach, the proposed feature recovery method has higher accuracy, and it is robust to facial pose variation appeared on the given image. Moreover, the model warping method based on local shape constraints can warp a high-resolution reference 3D face using few 3D features more reasonably and accurately. The proposed approach generates realistic 3D face with impressive visual effect.

1. Introduction

3D face reconstruction from single monocular image has received considerable attention from researchers in the field of computer vision and computer animation, and it has found its use in digital entertainment like film and game production for decades of years.

Recently, 3D face reconstruction gives light to some newly emerged applications such as human computer interaction [1,2], electronically mediated communication [3] and public security [4,5]. In [2], the authors developed an automatic face tracking and lip reading system through a reconstructed 3D face avatar for speech learning, emotional state monitoring and non-verbal human computer interfaces design. A real-time facial tracking system was developed in [3] to extract animation control parameters from videos. The system could translate these parameters to 3D facial expression and then retarget the expressions to reconstructed 3D faces for applications like teleconferencing. Also, in visual surveillance [5], face cues were combined with gait cues as biometrical features to achieve person identification.

Prior information is not indispensable for 3D reconstruction from multiple input views [6], but it is necessary for monocular reconstruction or pose estimation [7,8]. Specifically, monocular 3D face reconstruction is a highly ill-posed problem, the reconstruction process usually needs additional constraints derived from some prior knowledge. The most common constraint used in face reconstruction is shape constraint, which is usually concealed beneath 3D sample faces. The most favorable and related work is the approaches based on Morphable Models [9,10]. A Morphable Model refers to a statistical model constructed by linearly combining a set of 3D sample faces. The desired 3D face can be generated by tuning the parameters (combining coefficient) of the model. The optimal parameters are determined by fitting the Morphable Model to the given image to match the 2D projection of the model to the face appeared on the image. The sample face usually comprises tens of thousands of 3D points. Matching the 2D projection of these 3D points to image pixels incurs great computational cost. Therefore Morphable Model approaches usually have low computational efficiency. Furthermore, the shape constraint imposed by the sample faces is a kind of global constraint. This means that the constraint works simultaneously for all the points on the face model, and we cannot adjust any local face region by tuning the model parameters.

Instead of using all 3D points, some researchers compute the model parameters by fitting a few salient 3D facial feature points to corresponding image feature points [11–14]. The feature-based fitting speeds up the computation process enormously, but the parameter estimation of the above-mentioned works is unsatisfactory because it is based on alternating least square method which is not derived from the conditions driving the objective

☆This paper is supported by the National Natural Science Foundation of China (no. 61303143), Scientific Research Fund of Zhejiang Provincial Education Department (no. Y201326609).
* Corresponding author.
E-mail address: dtao.scut@gmail.com (D. Tao).

http://dx.doi.org/10.1016/j.neucom.2014.08.039
0925-2312/© 2014 Elsevier B.V. All rights reserved.

function to reach its optimal value. In addition, these improvements fail to consider some powerful methods such as sparse representation [15,16] and novel distance metric learning [17–20], which have been proved effective in solving linear approximation problems. Most importantly, the shape constraint here is also global shape constraint, and some approaches [13] can only recover a few 3D feature points, 3D face model should be generated by warping a high-resolution reference 3D face using the feature points. Prevalent warping method is scattered data interpolation [21] which relies on measuring Euclidean distances directly between feature points and landmark model points distributed on the model surface, therefore the warping is still conducted in a global way and rarely achieves ideal result with few feature points.

The aforementioned methods require a large number of sample face models and a detailed and accurate point-wise correspondence between all the models. This prevents these methods from wider usage when the sample face models are unavailable. In recent years, some researchers propose to regularize the 3D face reconstruction using shape constraint derived from a prior face model [22–24]. In [22], the authors detected the facial features around eyes, mouth, eyebrow and contour of face from a given face image, and adapted a generic 3D model into face specific 3D model using geometric transformations. Similar work can be found in [23], where a generic 3D face was projected onto image plane to fit the 2D projection to the input face. Shape regularization is implicitly used during the reconstruction procedure of these two methods, among which [22] uses local translation to achieve model fitting, and the depth information of the reconstructed 3D face in [23] totally comes from the generic face. Hence, the result of the regularization is not decent. In [24], the regularization was explicitly introduced in the formulation of the problem, which used the prior face model and albedo to extract illumination and pose information for 3D reconstruction. However, the result varies significantly depending on which prior model is used. This approach also requires plenty of manual work to register the image with the prior face.

To this end, we propose a novel approach to 3D face reconstruction. Unlike Morphable Model approaches, we reconstruct the unknown 3D face by fitting it directly to the given face image through a scaled orthogonal projection. To deal with the ill-posedness, we regularize the projection explicitly with a global shape constraint constructed using a reference 3D face. To ensure the efficiency, only few 3D facial features are fitted to a set of image feature points. Then, we obtain high-resolution 3D face by warping a reference 3D face model using the recovered 3D features. Unlike previous approaches, the warping is based on local shape constraint at each point of the face model[25]. The local shape constraints convey human characteristics of local face regions to the reconstructed 3D face. Realistic 3D face is generated after texture mapping. Our approach has following advantages: the feature recovery resorts to solving several equations, hence is very fast. In addition, the recovered features are very close to the optimal solution due to the global shape constraint, and the recovery is not sensitive to facial pose on the given image. Last, the warping based on local shape constraints is superior to scattered data interpolation in that it can achieve better result based on only few 3D features.

The rest part of this paper is organized as follows: In the next section, we describe our 3D feature recovery algorithm based on global shape constraint in detail. In Section 3, we discuss the high-resolution 3D face reconstruction method, as well as the local shape constraints. Texture mapping is introduced in Section 4. Section 5 shows some experimental results, and we conclude this paper in the last section.

2. 3D feature recovery

Twenty feature points of the input image are extracted by the Active Appearance Model (AAM) [26] for model fitting. Fig. 1 shows some examples selected from Pointing face database [28]. The result is quite robust when the horizontal viewpoint varies between $\pm 45^\circ$. Besides the 2D image features, we manually select corresponding 20 features from a prior reference human face to construct global shape constraint. See Fig. 2 for the reference 3D face with reference features (marked by red dots). The reference face reflects general facial characteristics and can be obtained by averaging several laser scanned human faces. Now we describe how to recover personalized 3D features from the image feature points.

Let $\{p_1, p_2, \ldots, p_n\}$ be the 2D feature points extracted from the input image, where $p_i = [x_i, y_i]^T$, $i=1,\ldots,n$, and let $\{\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_n\}$ be the corresponding 3D reference features. The unknown 3D feature points that will be recovered are denoted as $\{s_1, \ldots, s_n\}$. The 2D projection of $\{s_1, \ldots, s_n\}$ should be close to $\{p_1, p_2, \ldots, p_n\}$ as much as possible. Without loss of generality, a scaled orthogonal projection is used here, and the total errors

$$\sum_{i=1}^{n} \| p_i - \lambda(Rs_i + t) \|^2$$

should be small where λ is the scale factor, R is the rotation matrix and t is the translation vector. We assume that $t=0$ since we can rewrite t as $t = Rs_0$, and reuse the notation s_i for the shifted $s_i + s_0$. Equivalently, we assume that $\{p_i\}$ are centered. In addition, the unknown 3D features should match reference human features, meaning that $\sum_{i=1}^{n} \|s_i - \bar{s}_i\|^2$ should be also small. Of course, $\{\bar{s}_i\}$ are also centered. Taking the two concerns into account, we consider the following minimization model for recovering the unknown and personalized 3D features,

$$\min_{\lambda, \{s_i\}, R: RR^T = I} \sum_{i=1}^{n} \| p_i - \lambda R s_i \|^2 + \alpha \sum_{i=1}^{n} \|s_i - \bar{s}_i\|^2. \tag{1}$$

Fig. 1. Feature points extracted from images.