Pattern Recognition 56 (2016) 88–99

Contents lists available at ScienceDirect

Pattern Recognition

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

Shape-appearance-correlated active appearance model

Huiling Zhou [a,b], Kin-Man Lam [a,b,*], Xiangjian He [b]

[a] Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[b] School of Computing and Communications, University of Technology, Sydney, Australia

A R T I C L E I N F O

Article history:
Received 9 July 2015
Received in revised form
26 January 2016
Accepted 2 March 2016
Available online 11 March 2016

Keywords:
Facial-feature localization
Generic active appearance model
Canonical correlation analysis
Orthogonal CCA

A B S T R A C T

Among the challenges faced by current active shape or appearance models, facial-feature localization in the wild, with occlusion in a novel face image, i.e. in a generic environment, is regarded as one of the most difficult computer-vision tasks. In this paper, we propose an Active Appearance Model (AAM) to tackle the problem of generic environment. Firstly, a fast face-model initialization scheme is proposed, based on the idea that the local appearance of feature points can be accurately approximated with locality constraints. Nearest neighbors, which have similar poses and textures to a test face, are retrieved from a training set for constructing the initial face model. To further improve the fitting of the initial model to the test face, an orthogonal CCA (oCCA) is employed to increase the correlation between shape features and appearance features represented by Principal Component Analysis (PCA). With these two contributions, we propose a novel AAM, namely the shape-appearance-correlated AAM (SAC-AAM), and the optimization is solved by using the recently proposed fast simultaneous inverse compositional (Fast-SIC) algorithm. Experiment results demonstrate a 5–10% improvement on controlled and semi-controlled datasets, and with around 10% improvement on wild face datasets in terms of fitting accuracy compared to other state-of-the-art AAM models.

1. Introduction

Facial-feature detection and localization is a crucial process for various applications such as facial-expression recognition, face animation, 3D face reconstruction, etc. Among all competitive techniques, model-based algorithms have been proven to be most effective in automatic facial-information learning. The earliest work of such algorithms includes the deformable template [1] and the active contour model [2]. These approaches aim to extract facial features and locate face boundaries by studying feature points individually, and hence have limited robustness and accuracy. Most recently, more efficient methods, including the Active Shape Model (ASM) [3] and the Active Appearance Model (AAM) [4], have been proposed. ASM considers the facial-shape information (based on manually annotated facial-feature points) from a holistic perspective, while AAM also includes texture information (usually in terms of the pixel intensities within a face region). Due to these models' efficiency and accuracy, many variant ASM and AAM methods have been proposed in the past few decades, and they improve the localization performance. However, both ASM and AAM have problems in three different aspects, namely, insufficient robustness to variations, sensitivity to face-model initialization, and poor performance in generic situations. In the following, the challenges in these three aspects and those existing methods, which address these challenges, are discussed.

1.1. Insufficient robustness to variations. Since both ASM and AAM rely on global parametric models, they can work well for faces available in a training set with small variations in illumination, pose and expression. However, when these variations become greater, their performances usually degrade dramatically. One way to solve this problem is to integrate ASM and AAM [5,6] In [5] a texture-constrained shape model was used to prevent the local-minima problem, and it can achieve a robust performance under illumination variations. In [6], the profile-search step in ASM is changed into a gradient-based optimization problem to more accurately localize feature points. Recently, improved ASM models using 2-D profiles were proposed to achieve pose-adaptive localization [7–9] It has been proven that the 2-D profiles can capture more information around each landmark than the original 1-D profiles. By properly setting the initial face model and using an optimization method, these methods can achieve accurate results, and thus have become popular model-based localization methods.

1.2. Sensitivity to initialization. In the process of refining the feature-point locations, both ASM and AAM usually perform gradient-descent optimization over a whole face, so their

* Corresponding author.
E-mail addresses: hl.zhou@connect.polyu.hk (H. Zhou), enkmlam@polyu.edu.hk (K.-M. Lam), Xiangjian.He@uts.edu.au (X. He).

http://dx.doi.org/10.1016/j.patcog.2016.03.002
0031-3203/

performances are sensitive to the initial face model. This issue has drawn much attention, and can be improved in two major steps, namely, constructing a more representative initial face model and using a robust feature-point refining scheme. For the first step, several frameworks [10–12] reformulate the original AAM as a sparse representation problem [13] and approximate the local appearance of feature points with locality constraints. After the shape and appearance priors are learned, the K nearest neighbors with similar patterns to the test face in terms of pose, expression, etc. are searched from a training set, and are used to model the face in a locally linear sub-space. It has been shown that this pre-processing step helps to reach faster convergence and to obtain better fitting results. Similarly, [9,14] pre-define the number of face clusters and classify the test face into one of the clusters based on a statistical analysis. For the second step, in order to refine the face model, a stacking strategy is usually employed to search, in series, for a better location for each feature point in the face model iteratively [7,15,16].

1.3. Poor performance in generic situations. In the survey work of [17], statistical evaluation has shown that person-specific active models (i.e. images of a query also exist in the training set) are both easier to build and more robust to fitting than generic ones (i.e. no images of a query in the training set). To solve the generalization problem, frameworks [18,19] based on AAM were proposed to learn a discriminative fitting function and establish a mapping between the facial appearance and the face shape in order to improve the alignment accuracy. Unlike AAMs – which model a whole facial region – the family of Constrained Local Models (CLMs) [20–22] extracts templates around each landmark and matches them to new instances of an object using a shape-constrained search and iterative template generation. This process always relies on the response surfaces generated by fitting the current feature templates using normalized correlation at each point. Recently, an approach which can handle unseen faces and variations was proposed, and is known as the Active Orientation Model (AOM) [23] It establishes a generative deformable appearance model based on the principal components of images' gradient orientations, and it uses the project-out inverse compositional algorithm to optimize the results. An improved AAM model [24] using more efficient optimization algorithms was also proposed for generic situations.

As discussed in some survey papers [25–27] AAMs take advantage of all gray-level information across faces to build a convincing model with a relatively small number of landmarks, while ASM is just a special case of AAM. Therefore, in this paper, we focus on establishing a shape-appearance-correlated AAM (SAC-AAM) framework to tackle the above-mentioned three challenges at the same time, especially under a generic localization environment.

The contributions of this paper are given as follows. In order to fulfill the goals, we first propose a fast initialization scheme, which retrieves the most similar faces to a test face in terms of both poses and textures. Based on the idea of locality constraint, these nearest neighbors form a locally linear subspace. Then, the shape and appearance of the selected images are analyzed, and their correlation is maximized by applying Canonical Correlation Analysis (CCA) [28] (actually, the orthogonal CCA (oCCA) [29] is employed in our framework due to its superior data reconstruction property). We will show that our approach can increase the correlation between the principal components learned for face appearances and shapes, as well as the respective projection coefficients. This can improve the convergence speed and the fitting accuracy, while almost no additional computational cost will be added. By conducting experiments on different face datasets and comparing our proposed framework with state-of-the-art model-based methods, experimental results show that our framework can achieve a great improvement in terms of fitting accuracy, especially for faces under large pose, expression, and occlusion variations, as well as for unseen faces.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce the well-known AAM model and some of its latest improved models. The Canonical Correlation model and its orthogonal variant are also discussed there. In Section 3, our shape-appearance-correlated AAM (SAC-AAM) framework is presented, and the details of generating initial face models and obtaining more correlated principal components are described. Experimental results and analysis are given in Section 4. The conclusion is outlined in Section 5.

2. Related work

In this section, we will give a brief overview of the Active Appearance Model (AAM) and its latest variants. We will also introduce the concept of Canonical Correlation Analysis (CCA) and its extension to orthogonal CCA, together with its efficiency for various applications.

2.1. Active appearance model

As mentioned in the previous section, unlike ASM – which only deals with shape information – AAM also takes texture information into consideration. The shape vector is usually presented by concatenating the position coordinates of labeled landmarks, while texture is modeled in terms of the demeaned pixel intensities or colors within the convex hull of a facial shape. When given a training set of face images with corresponding labeled landmarks, the shape model is established from $2N$ fiducial points denoted as $\boldsymbol{s} = (x_1, y_1, x_2, y_2, \dots, x_N, y_N)^{\mathrm{T}}$. The shapes are normalized by using the Procrustes analysis [30], which is a commonly used method to align shapes to a common coordinate system (usually, the mean shape of the training objects). Then, the principal component analysis (PCA) is applied to project the normalized and aligned shapes onto the shape subspace. Thus, the shape instance $\boldsymbol{s}$ can be presented as a linear combination of principal shapes as follows:

$$\hat{\boldsymbol{s}} = \overline{\boldsymbol{s}} + \mathbf{P}_s \cdot \boldsymbol{\alpha}, \text{ and} \tag{1}$$

$$\boldsymbol{\alpha} = \mathbf{P}_s^{\mathrm{T}} (\boldsymbol{s} - \overline{\boldsymbol{s}}), \tag{2}$$

where $\overline{\boldsymbol{s}}$ is the mean shape, $\mathbf{P}_s$ is the matrix whose columns form a set of orthonormal base vectors, and the weight vector $\boldsymbol{\alpha}$ (also known as projection parameters) is used to control the shape variations.

The appearance model of a face image I is learned by first warping it into a "shape-free" model, usually the mean shape $\overline{\boldsymbol{s}}$. This is represented as a warping function $W(\boldsymbol{x}; \boldsymbol{\alpha})$, where $\boldsymbol{x}$ denotes a set of pixels inside the mean shape $\overline{\boldsymbol{s}}$. Then, PCA is again applied to project the "shape-free" appearance of the image $I(W(\boldsymbol{x}; \boldsymbol{\alpha}))$ on to the appearance subspace. The appearance instance $\boldsymbol{r}$ can be represented as a linear combination of principal appearances as follows:

$$\hat{\boldsymbol{r}} = \overline{\boldsymbol{r}} + \mathbf{P}_r \cdot \boldsymbol{\beta}, \text{ and} \tag{3}$$

$$\boldsymbol{\beta} = \mathbf{P}_r^{T} (\boldsymbol{r} - \overline{\boldsymbol{r}}), \tag{4}$$