Pattern Recognition 49 (2016) 115–128

Contents lists available at ScienceDirect

Pattern Recognition

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

Facial expression transfer method based on frequency analysis

Wei Wei [a,*], Chunna Tian [b], Stephen John Maybank [c], Yanning Zhang [a]

[a] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[b] VIPS Lab, School of Electronic Engineering, Xidian University, Xi'an, China
[c] Department of Computer Science and Information Systems, Birkbeck College, University of London, UK

ARTICLE INFO

Article history:
Received 21 January 2015
Received in revised form
6 August 2015
Accepted 10 August 2015
Available online 18 August 2015

Keywords:
Expression transfer
Warping technique
Facial feature location
Frequency domain analysis

ABSTRACT

We propose a novel expression transfer method based on an analysis of the frequency of multi-expression facial images. We locate the facial features automatically and describe the shape deformations between a neutral expression and non-neutral expressions. The subtle expression changes are important visual clues to distinguish different expressions. These changes are more salient in the frequency domain than in the image domain. We extract the subtle local expression deformations for the source subject, coded in the wavelet decomposition. This information about expressions is transferred to a target subject. The resulting synthesized image preserves both the facial appearance of the target subject and the expression details of the source subject. This method is extended to dynamic expression transfer to allow a more precise interpretation of facial expressions. Experiments on Japanese Female Facial Expression (JAFFE), the extended Cohn-Kanade (CK+) and PIE facial expression databases show the superiority of our method over the state-of-the-art methods.

1. Introduction

Emotions are often conveyed through body gestures or facial expressions rather than verbal communication [1,2]. Thus, automatic facial expression analysis is an interesting research topic. There have been many recent achievements in related research sub-areas such as facial landmark localization [3–8], tracking and recognition [9,10]. Realistic facial expression synthesis is useful for affective computing, human computer interaction [11,12], realistic computer animation [13] and facial surgery planning [14,15], etc. There are more than 20 groups of facial muscles innervated by facial nerves [16], which control actions of the face (e.g. opening or closing of eyes or mouth) and variations in local appearance (e.g. facial wrinkles and furrows).

Psychologists Ekman and Friesen developed a facial action coding system (FACS) based on the movements of facial muscles and their effects on the expression [17]. They divided the face into 44 action units (AU) and analyzed the motion characteristics and their effects on associated expressions. Many expression synthesis approaches had concentrated on capturing expressions through AUs. Platt and Badler proposed a model of the muscles to simulate FACS to synthesize facial expression [18]. Waters extended this model to a hierarchical one [19], in which facial muscles were divided into linear muscles and sphincter muscles to control the skin stretch and shrinkage, respectively. Based on Water's method, Koch et al. used a finite element method to simulate the physical structure of the human face [20]. These models emphasized the simulation of muscle movements. They are relatively simple compared with the one developed by Lee et al. [21]. Lee's model has a three-level structure of skin, bone and muscle, based on physiology. Thus, it can synthesize much more realistic expressions. But the complicated structure and heavy calculation load prevented its application in practice. FACS was used to define a non-continuous and non-uniform scale for scoring the strength of facial activities. It was hard to identify subtle facial activities. Thus, the valuable information contained in subtle expression changes can be lost [22]. The subtle local appearance variations are usually the main components of micro-expressions, which are important clues to distinguish different expressions. For example, without the micro-expressions, 'fear' and 'surprise' are hard to distinguish. In addition, micro-expressions reflect the emotion and inner working of people. As a result, photo-realistic facial expression synthesis is an active research area.

Darwin revealed the consistency of expressions among different races and genders [23]. This is because the shape deformations are similar for the same expression. However, the local appearance deformations, viz., the expression details, are quite person-specific. Therefore, a person-specific expression transfer, which clones the expression of a source subject to a target subject, has a wide range of applications. However, the subtle appearance deformations are difficult to synthesize. In this study, we locate

[*] Correspondence to: Box 886, School of Computer, Chang'an Campus, Northwestern Polytechnical University, 710129, China. Tel.: +86 137 7253 8134; fax: +86 29 8843 1533.
E-mail address: weiweinwpu@nwpu.edu.cn (W. Wei).

http://dx.doi.org/10.1016/j.patcog.2015.08.004
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

the facial landmarks automatically in order to describe the shape deformation, and extract and transfer the subtle local expressions using frequency analysis. Since dynamic variations are important in interpreting facial expression [24–26], we extend our work to dynamic expression transfer. The effectiveness of the proposed algorithm is verified on Japanese Female Facial Expression (JAFFE), the extended Cohn-Kanade (CK+) [27] and PIE databases.

The remainder of this paper is structured as follows. Section 2 provides an overview of the related work. Section 3 presents the facial landmark localization and face alignment methods. Section 4 proposes the static expression transfer method based on frequency analysis. We extend the expression transfer method to dynamic expression synthesis in Section 5. Section 6 details the evaluation of the proposed method as well as the experimental setup. Finally, Section 7 concludes the paper.

2. Related work

Our synthesis method involves local facial landmark detection and expression transfer. Below, we give a concise overview of relevant prior work on these two topics.

The Active Shape Model (ASM) [28] used a parametric deformable model to fit the shape of human face. It learned the variation modes of a shape from a set of training examples. Transformation and shape parameters were estimated iteratively to fit the mean shape of the observed object. The Active Appearance Model (AAM) [29,30] integrated facial texture with the ASM to fit the facial shape better. Matthews and Baker [31] proposed a computationally efficient AAM algorithm with rapid convergence to improve the fitting. The shapes of multi-view faces can be fitted through a gradient-descent search [32]. A vectorial regression function was learned from the training image with an Explicit Shape Regression (ESR) model to locate the facial landmarks. A two-level boosted regression, shape-indexed features and a correlation-based feature selection method were combined with an ESR model to locate the facial landmarks more accurately [5]. In [6], local binary features were learned using a regression tree to preserve the most discriminative information contained in the local texture around each facial landmark. The local binary features improved the efficiency of facial landmark detection.

Zhu and Ramanan [7] used a multi-tree model to handle different expressions. Each expression was modeled by a deformable model of the joint distribution of parts, consisting of local patches around facial landmarks. The Histogram of Oriented Gradient (HoG) features were used to describe the local patches. Each branch of the multi-tree model corresponds to one expression, but different trees share a pool of parts. The deformable model was trained by a latent support vector machine (LSVM). Since the human face is non-rigid and large nonlinear deformations occur in extreme expressions, [4] used a Haar-like feature based Adaboost face detector [33] to initialize the face location, then adopted a Supervised Descent Method (SDM) to refine the locations of facial landmarks. [8] presented a state-of-the-art method for facial landmark detection with super-real time performance, which used an ensemble of regression trees to estimate the positions of facial landmark accurately from a sparse subset of pixel intensities. The ensemble of regression trees was learned based on gradient boosting. The appropriate priors exploiting the structure of image data is helpful to efficient feature selection.

To transfer expression, the facial texture of the target subject was warped to the shape of the source subject, given the correspondence of the key facial landmarks [34]. The warped expression captures the shape motions of different expressions but ignores the micro-expressions. Liu et al. [35] represented the gray level of each point on the face image with the Lambertian model. Then, they calculated the ratio between the neutral and the non-neutral expressions at each point to obtain an expression ratio image. They used this expression ratio image to transfer the non-neutral expression to other neutral faces. In [36,37,38], multi-expression faces were arranged as a tensor data. In [36], a High Order Singular Value Decomposition (HOSVD) [39] was applied to the AAM [40] coefficients of the training faces to obtain a generative model. In this model, AAM coefficients of training images were factorized into identity and expression subspaces. The test face image was represented by the AAM coefficients, which were reconstructed by estimating its identity and expression coefficients in the generative model. The solved identity coefficients were combined with the remaining expression coefficients in the expression subspace of the generative model to synthesize multi-expression face images. In [37], a generative model of shape and texture is built in order to obtain the identity and expression coefficients, separately. The expression transfer was realized by swapping the expression coefficients of the source and target subjects. The authors in [38] proposed a tensor-based AAM, in which texture is aligned with the normalized shape of the AAM. The expression coefficients of the test face were synthesized by linearly combining the expression coefficients of training faces in the latent expression subspace. A texture variation ratio between the neutral and non-neutral expressions was used to transform the expression of the test face. However, the expression variations are often strongly nonlinear, thus the expression coefficient estimation for the test image may not be accurate. The expression variation ratio is not adaptive to the nonlinear variation of extreme expressions. The authors in [41] incorporated the expression manifold with the Tensor-AAM model to synthesize dynamic expressions of the training face. The Bilinear Kernel Reduced Rank Regression (BKRRR) method for static general expression synthesis was proposed in [15]. It synthesizes general expressions on the face of a target subject. Zhang and Wei [2] used TensorFace combined with an expression manifold to synthesize the dynamic expressions of a training face, then extracted and transferred the dynamic expression details of the training face to the target face. For extensive reviews on facial expression synthesis, we refer the reader to [42] and [43].

Since the facial landmarks around eyes, nose, mouth and eyebrows etc. are required to describe the shape correspondence between different expressions, we use state-of-the-art method proposed in [8] to detect the facial landmarks (See Fig. 1) in this study. We use AAM to separate and align the shapes and texture of multi-expression faces. Since the expression details are more easily observed using frequencies, we adopt the wavelet transform to divide images into four frequency bands and then transfer the details of expressions. We extend this work to dynamic expression transfer, to obtain highly realistic and natural looking facial animations. In summary, our main contributions include: 1) a proposal (See Fig. 1) for a static expression transfer method based on frequency analysis; 2) the extension of the static expression transfer to dynamic expression transfer; 3) a unified automatic dynamic expression transfer framework including facial landmark localization, expression alignment, dynamic shape synthesis, expression warping, expression detail extraction and transferring.

3. Automatic facial landmark extraction and face alignment

In this study, we use the method proposed in [8] to detect facial landmarks (See Fig. 1). Given the facial landmarks, we use AAM to separate and align the shapes and texture of multi-expression faces. Since human face is not homogeneous, we cannot use a global uniform transformation to warp the whole facial texture to the aligned shape. Thus, we divide the facial texture into small