Signal Processing: *Image Communication* 26 (2011) 550–566

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Smooth adaptive fitting of 3D face model for the estimation of rigid and nonrigid facial motion in video sequences

Yunshu Hou, Ping Fan, Ilse Ravyse *, Valentin Enescu, Hichem Sahli

Research Team on Audio Visual Signal Processing (AVSP), Vrije Universiteit Brussel (VUB), Electronics and Informatics Department, VUB-ETRO, Pleinlaan 2, 1050 Brussel, Belgium

ARTICLE INFO

Article history:
Received 30 August 2010
Accepted 9 May 2011
Available online 3 June 2011

Keywords:
Facial analysis
Model-based tracking
Motion-smoothness constraints

ABSTRACT

We propose a 3D wireframe face model alignment for the task of simultaneously tracking of rigid head motion and nonrigid facial expressions in video sequences. The system integrates two levels: (i) at the low level, automatic and accurate location of facial features are obtained via a cascaded optimization algorithm of a 2D shape model, (ii) at the high level, we recover, via minimizing an energy function, the optimal motion parameters of the 3D model, namely the 3D rigid motion parameters and seven nonrigid animation (Action Unit) parameters. In this latter inference, a 3D face shape model (Candide) is automatically fitted to the image sequence via a least squares minimization of the energy, defined as the residual between the projected 3D wireframe model and the 2D shape model, meanwhile imposing temporal and spatial motion-smoothness constraints over the 3D model points. Our proposed system tackles many disadvantages of the optimization and training associated with active appearance models. Extensive fitting and tracking experiments demonstrate the feasibility, accuracy and effectiveness of the developed methods. Qualitative and quantitative performance of the proposed system on several facial sequences, indicate its potential usefulness for multimedia applications, as well as facial expression analysis.

1. Introduction

3D facial model fitting allows recovering the face pose, its shape, and the facial appearance from an image sequence. The recovered information can then be used in many applications, among others: facial motion capture for animating a synthetic avatar, facial expression analysis and/or synthesis, realistic character animation in 3D computer games, very low bitrate model-based video coding and visual media production. Moreover, such information could also be used for the automatic recognition of the Action Units of Ekman and Friesen [1], which represent changes in facial appearance produced by the muscular activity. All these applications require that the facial model fitting algorithms are robust, efficient and accurate in the extraction of features, rigid head motion and nonrigid facial animation.

Automatic 3D rigid and nonrigid motion parameters estimation from monocular video sequences is a challenging problem in computer vision. The existing techniques can be divided into two classes, *facial features-based techniques* and *combined facial texture and features techniques*. The former techniques use local texture to track feature points, which are then used to guide the 3D motion estimation [2–11]. The latter techniques need to build up a statistical 3D face model including facial shape and texture information, and then align the model into the image sequence to obtain the proper shape, motion

* Corresponding author.
E-mail addresses: yhou@gmail.com (Y. Hou), pfan@etro.vub.ac.be (P. Fan), icravyse@etro.vub.ac.be, ilse.ravyse@lid.kviv.be (I. Ravyse), venescu@etro.vub.ac.be (V. Enescu), hsahli@etro.vub.ac.be (H. Sahli).

0923-5965/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2011.05.003

and texture parameters that best simulate the current face appearance [12–18].

Facial features-based techniques consider two steps: 2D facial feature points extraction and 3D motion inference based on the extracted feature points. Many methods have been proposed for feature points extraction, such as Kanade–Lucas–Tomasi Tracker (KLT) [19,20], heuristic facial dominant feature extraction [21,22], and active shape model (ASM) [23]. Given the tracked feature points, two typical ways, factorization-based structure from motion (SFM) [24–27] and recursive model-based SFM [4,28,5], have been often used in 3D motion parameters inference. Generally, factorization-based techniques work quite well for rigid motion models. However, for nonrigid motion the two steps factorization (first rigid followed by nonrigid) often has poor estimation performance [25,26]. Such approaches refine the estimation using bundle adjustment or random samplings techniques [27], and hence requires the entire sequence.

Combined facial texture and features techniques generally require a large database for the training of the shape and texture models. Because of the involvement of the statistical texture model, a large number of model parameters need to be estimated. Moreover, these approaches are computationally expensive [13] due to the required texture 'warping' while aligning the models. Additionally, direct texture-based motion parameters estimation may not produce precise point locations compared to *feature-based techniques*, especially at face contours and key facial components. However, it can achieve better overall robustness, and has gained much attention from the model based coding research groups [29,14].

Recently, methods such as active contour models (ACM) [30], active shape models (ASM) [23], Bayes tangent shape models (BTSM) [31], active appearance models (AAM) [32] and their extensions [33–45] gained interest. The ASM-like techniques (ACM, ASM and BTSM) fall into the *facial features-based techniques*, and the AAM-like methods fall into the *combined facial texture and features techniques*. BTSM [31] is proposed to greatly improve the facial feature extraction by modeling the 2D face shape variation due to the nonrigid motion and shape difference among people. The same idea has been generalized to multi-view face alignment [34] and to 3D face alignment [36]. More complicated sampling techniques is used to deeply improve the alignment performance [35]. A regularized ASM [43] imposes a natural smoothness constraint on the trained shape model to handle local noise. The nonlinear shape prior, geometric transformation and likelihood of multiple candidate landmarks constrain a multi-level generative ASM-like model for robust face alignment [45]. In [44] the initialization of the face alignment is alleviated by the optimization with a set of direction classifiers at the facial component level. Recently, 3D AAM has gained popularity [38–42,46]. 3D AAM algorithms have been introduced by Ahlberg in [38,39]. A bilinear active appearance model [40] is proposed to simultaneously model facial appearance variations caused by pose, identity and expression. The asymmetry issue is modeled into AAM to further improve the facial alignment preciseness [42]. In [46] the view-based AAM and an effective model selection method are used to estimate the pose of the face. The 3D anthropometric muscle-based AAM [41] is further integrated into an extended Kalman Filter (KF) to recover rigid and nonrigid facial motions.

As part of their use for realistic face modeling, nonrigid motion models play an important role in facial analysis. Two main classes of nonrigid models have been proposed in the literature, the pure *geometrically deformable models* and the *physics-based models*. The former class uses human knowledge to design nonrigid facial motion bases, such as the Candide wireframe model [47], or builds an anthropometric face model using variational techniques [48], or builds statistical facial shape models based on a training set, such as ASM [23,11] and AAM [32]. These techniques generally produce compact and semantic facial parameterizations to obtain over-constrained modeling systems and hence gain computational efficiency, however, the recovered model parameters may be noisy because of the direct dependency on the measurements. The physics-based model either defines a mass-spring-damper model [3,49–51] with physics approximations of explicit setup of the muscle topology and the extend functions for muscle control; either builds an elastic thin shell continuum mechanical finite element model (FEM) [52–54] with better physical approximations of natural facial nonrigid motions. These techniques provide physically-close models and are, therefore, expected to produce more smooth, natural and complex facial motions. Nevertheless, they deal with ill-posed un-constrained problems and require regularization terms to introduce additional information.

Even though the current facial analysis techniques have yielded significant results, their success is limited by the application conditions, such as precise tracking under scene variations (person-adapted facial expressions and illumination changes) while keeping computation efficiency. Having in mind that (i) *physics-based models* are computationally expensive [52], (ii) *combined facial texture and features techniques* may produce less precise point localization in key facial components [13], and (iii) the estimated 2D facial feature points are noisy, in this paper, we propose to use a *facial features-based technique* with a *geometrically deformable model*, augmented with temporal and spatial motion-smoothness constraints. More precisely, based on previous work described in [55–57], we start with a precise tracking of 2D facial features [55], which relies on a cascaded parameter prediction and optimization, and then fit a parameterized 3D wireframe model [56,57], the Candide-3 model [47]. The fitting consists of optimally adjusting the semantic and compact parameters of the 3D model and is cast as an energy minimization problem. The energy function includes an external energy which quantifies the matching error between the extracted 2D facial feature points and the projection of the 3D face model, and an internal energy imposing temporal and spatial smoothing constraints over the 3D model points. This enables the precise localization of the key facial feature points, while avoiding time-consuming texture 'warping' and noisy parameter estimates.

Our contributions are summarized as follows. Firstly we propose a novel 2D facial feature extraction method which relies on a cascaded parameter prediction and optimization [55]. Secondly we present a novel smooth