Image and Vision Computing 28 (2010) 1162–1172

Contents lists available at ScienceDirect

Image and Vision Computing

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

# Video-based face model fitting using Adaptive Active Appearance Model

Xiaoming Liu *

*Visualization and Computer Vision Lab, General Electric Global Research, Niskayuna, NY 12309, USA*

ARTICLE INFO

*Article history:*
Received 15 October 2008
Received in revised form 21 July 2009
Accepted 25 September 2009

*Keywords:*
Active Appearance Model
Model fitting
Subject-specific model
Generic model

ABSTRACT

Active Appearance Model (AAM) represents the shape and appearance of an object via two low-dimensional subspaces, one for shape and one for appearance. AAM for facial images is currently receiving considerable attention from the computer vision community. However, most existing work focuses on fitting an AAM to a single image. For many applications, effectively fitting an AAM to video sequences is of critical importance and challenging, especially considering the varying quality of real-world video content. This paper proposes an Adaptive Active Appearance Model (AAAM) to address this problem, where both a generic AAM component and a subject-specific appearance model component are employed simultaneously in the proposed fitting scheme. While the generic AAM component is held fixed, the subject-specific model component is updated during the fitting process by selecting the frames that can be best explained by the generic model. Experimental results from both indoor and outdoor representative video sequences demonstrate the faster fitting convergence and improved fitting accuracy.


## 1. Introduction

Model-based image registration/alignment is a fundamental topic in computer vision. *Active Appearance Model (AAM)* has been one of the most popular models for image alignment [10]. Face alignment using an AAM is receiving considerable attention from the computer vision community because it enables various capabilities such as facial feature detection, pose rectification, and gaze estimation. However, most existing work focuses on fitting the AAM to a single facial image. With the abundance of surveillance cameras and greater need for face recognition from video, methods to effectively fit an AAM to facial images in videos are of increasing importance. This paper addresses this problem and proposes a novel algorithm for it.

There are two basic components in face alignment using an AAM: *model learning* and *model fitting*. Given a set of facial images, *model learning* is the procedure of training the AAM, which is essentially two distinct linear subspaces modeling facial shape and appearance respectively. *Model fitting* refers to estimating the parameters of the resulting AAM on faces in an image or video frames by minimizing the distance measured between the face and the AAM.

In the context of fitting an AAM to video sequences, conventional methods directly fit the AAM to each frame by using the fitting results, i.e., the shape and appearance parameters, of the previous frame as the initialization of the current frame. However, as shown in the previous work [16], fitting to faces of an unseen subject can be hard due to the mismatch between the appearance of the facial images used for training the AAM and that of the video sequences, especially when the video sequences are captured in the outdoor environment. Also, the conventional method only registers each frame with respect to the AAM, without enforcing the frame-to-frame registration across video sequences, which is necessary for many practical applications, such as multi-frame super-resolution [28].

To address this problem, we propose a novel model learning and fitting approach to continuously fit a mesh-based face model to video sequences. Both a generic AAM component and a subject-specific appearance model component are employed simultaneously in the proposed model learning, where the subject-specific model is learned and updated in an online fashion by making use of the test video sequence. Hence, in our approach, the *training* (learn the subject-specific model) and *test* (fit the face model to a frame) phases take place simultaneously. The proposed fitting algorithm is an extension of the state-of-the-art image alignment algorithm – the Simultaneous Inverse Compositional (SIC) method [3], which minimizes the distance of the warped image observation and the generic AAM model during the fitting. We call our proposed approach as "*Adaptive Active Appearance Model (AAAM)*" algorithm, which not only minimizes the aforementioned distance measure, but also the distance between the warped image and the adaptive subject-specific model. Note that here "*Adaptive*" refers to the subject-specific appearance model component because the generic AAM component remains fixed throughout our algorithm. Extensive experimental results demonstrate that the AAAM algorithm improves both the fitting speed and the fitting accuracy compared to the conventional SIC method. The earlier version of this work was published at [23].

* Tel.: +1 518 3874211.
*E-mail address:* liux@research.ge.com

0262-8856/$ - see front matter 
doi:10.1016/j.imavis.2009.09.016

The proposed approach has three main contributions.

1. In terms of model learning, our AAAM is composed of a generic AAM and a subject-specific appearance model. By tailing toward the application of fitting face models to videos sequences, we study various strategies of adapting the subject-specific model in an online fashion using the video content at previous time instances, so that the AAAM can fully utilize the subject-specific information in face model fitting.
2. In terms of model fitting, this paper extends the conventional SIC method by allowing a hybrid appearance model, which includes both an eigenspace-based appearance model and a number of appearance templates. We provide the derivation of the fitting method using this novel appearance model, as well as the computation analysis.
3. In terms of applications, we improve the performance of fitting face models to video sequences compared to the state-of-the-art SIC method. We demonstrate that satisfying fitting performance can be observed when fitting a generic model to unseen subjects, in both indoor and outdoor scenarios.

This paper is organized as follows. After a brief description of the related work in Section 2, this paper presents the model learning and fitting methods of the conventional AAM in Section 3. Section 4 presents the proposed AAAM algorithm in detail. Section 5 provides experimental results, and conclusions are given in Section 6.

## 2. Prior work

Image alignment is a fundamental problem in computer vision. Since early 90 s, ASM [10] and AAM [11,24] have become one of the most popular model-based image alignment methods because of their elegant mathematical formulation and efficient computation. For the template representation, AAM's basic idea is to use two eigenspaces to model the object shape and shape-free appearance respectively. For the distance metric, the MSE between the appearance instance synthesized from the appearance eigenspace and the warped appearance from the image observation is minimized by iteratively updating the shape and/or appearance parameters. ASM and AAM have been applied extensively in many computer vision tasks, such as facial image processing [29,13,14], medical image analysis [6], image coding [4], industrial inspection [27], object appearance modeling [17], etc. Cootes and Taylor [12] have an extensive survey on this topic.

Due to the needs of many practical applications such as face recognition, expression analysis and pose estimation, extensive research has been conducted in face alignment, among which AAM [10,3] and their variations [5,14,8,15] are probably one of the most popular approach. Baker and Matthews [3] proposed the Inverse Compositional (IC) method and SIC method that greatly improves the fitting speed and performance. However, little work has been done in fitting an AAM to facial video sequences in particular. Ahlberg [1] utilized a simplified AAM to track facial features in videos. Koterba et al. [19] proposed to use a 3D face model as a constraint in fitting multiple video frames. Matthews et al. [25] also updated the generic AAM using the warped image observation, such that a subject-specific model can be obtained during the fitting process. Comparing to their approach, we will show that treating the previous frame information as an additional constraint can improve the fitting speed, not to mention saving the extra time needed to update the bulky eigenspace of the appearance model in an AAM. Bosch et al. [7] proposed an Active Appearance Motion Model that captures the motion pattern in video sequences by taking the concatenation of the landmarks from multiple frames as training samples. This approach takes advantage of the periodic motion pattern in medical image sequences. In contrast, our approach does not make any assumption on the object's motion. Batur and Hayes [5] proposed an extension of AAM fitting algorithm in that the gradient matrix can be adapted, rather than fixed, which offers improved fitting performance on static images. This is very different to our approach since we study video-based fitting and our appearance model contains both generic and subject-specific appearance information.

## 3. Active Appearance Model

This section will first introduce the model learning of the conventional Active Appearance Model, including the shape model and the appearance model. It will then briefly describe the method of fitting AAM to a static image.

### 3.1. Model learning

The shape model and appearance model part of an AAM are trained with a representative set of facial images. The distribution of facial landmarks are modeled as a multi-dimensional Gaussian distribution, which is regarded as the shape model. The procedure for training a shape model is as follows. Given a face database, each facial image is manually labeled with a set of 2D landmarks, $[x_i, y_i]$ $i = 1, 2, \ldots, v$. The collection of landmarks of one image is treated as one observation from the random process defined by the shape model, $\mathbf{s} = [x_1, y_1, x_2, y_2, \ldots, x_v, y_v]^{\mathrm{T}}$. Eigen-analysis is applied to the observation set and the resulting linear shape model represents a shape as,

$$\mathbf{s}(\mathbf{P}) = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i, \quad (1)$$

where $\mathbf{s}_0$ is the mean shape, $\mathbf{s}_i$ is the $i$th shape basis, and $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ are the shape parameters. By design, the first four shape basis vectors represent global rotation and translation. Together with other basis vectors, a mapping function from the model coordinate system to the coordinates in the image observation is defined as $\mathbf{W}(\mathbf{x};\mathbf{p})$, where $\mathbf{x}$ is a pixel coordinate defined by the mean shape $\mathbf{s}_0$.

After the shape model is trained, each facial image is warped into the mean shape using a piecewise affine transformation. These shape-normalized appearances from all training images are fed into an eigen-analysis and the resulting model represents an appearance as,

$$A(\mathbf{x};\boldsymbol{\lambda}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}), \quad (2)$$

where $A_0$ is the mean appearance, $A_i$ is the $i$th appearance basis, and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ are the appearance parameters. Fig. 1 shows an AAM trained using 534 images of 200 subjects from the ND1 3D face database [9]. In conclusion, the collection of the shape model and appearance model is conventionally treated as the AAM: $\Im = \{\mathbf{s}_i, A_j\}_{i\in[0,n],j\in[0,m]}$.

### 3.2. Model fitting

An AAM can synthesize facial images with arbitrary shape and appearance within the range expressed by the training population. Thus, the AAM can be used to *explain* a facial image by finding the optimal shape and appearance parameters such that the synthesized image is as similar to the image observation as possible. This leads to the cost function used for model fitting [11],