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

liux AT research.ge.com

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.

Key words: Active Appearance Model, Model fitting, Subject-specific model, Generic model

1 Introduction

Model-based image registration/alignment is a fundamental topic in computer vision. *Active Appearance Model (AAM)* have 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 image 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 appear-

ance 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 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].

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. Sections 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 90s, 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 [13, 14, 29], 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 [3,10] and their variations [5,8,14,15] 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] propose 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, $[x_i, y_i]$ $i = 1, 2, ..., v$. The collection of landmarks of one image is treated as one observation from the random process defined by the shape model, $\mathbf{s} = [x_1, y_1, x_2, y_2, ..., x_v, y_v]^\intercal$. Eigen-analysis is applied to the observation set and the resulting linear shape model represents a shape as,

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

where $\mathbf{s}_0$ is the mean shape, $\mathbf{s}_i$ is the i^{th} shape basis, and $\mathbf{p} = [p_1, p_2, ..., p_n]$ 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,

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

where A_0 is the mean appearance, A_i is the i^{th} appearance basis, and $\lambda = [\lambda_1, \lambda_2, ..., \lambda_m]$ are the appearance parameters. Figure 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: $\Im = \{\mathbf{s}_i, A_j\}_{i \in [0,n], j \in [0,m]}$.

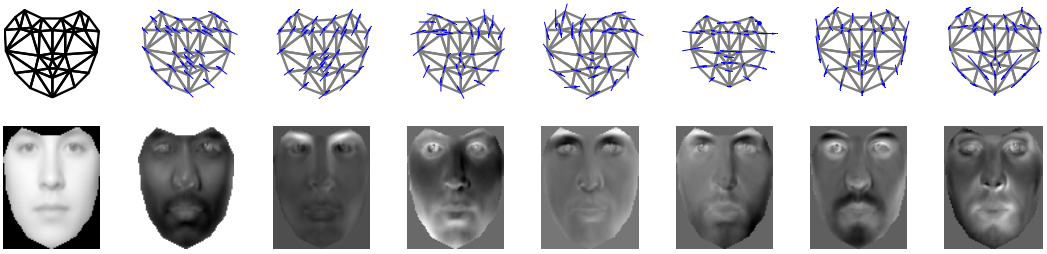

Fig. 1. The mean and first 7 basis vectors of the shape model (top) and the appearance model (bottom) trained from the ND1 database. The shape basis vectors are shown as arrows at the corresponding mean shape landmark locations.

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],

$$J(\mathbf{p}, \lambda) = \sum_{\mathbf{x} \in \mathbf{s}_0} [I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - A(\mathbf{x}; \lambda)]^2, \tag{3}$$

which is the mean-square-error (MSE) between the image warped from the observation $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ and the synthesized appearance model instance $A(\mathbf{x}; \lambda)$.

Traditionally this minimization problem is solved by iterative gradient-descent methods which estimate $\Delta\mathbf{p}$, $\Delta\lambda$ and add them to $\mathbf{p}$, λ. Baker and Matthews [3] proposed the compositional method to generate the new shape parameters based on $\Delta\mathbf{p}$ in their IC and SIC method. The key idea of IC and SIC is that the role of the appearance template and the input image is switched when computing $\Delta\mathbf{p}$. This enables the time-consuming steps of parameter estimation to be pre-computed and performed outside of the iteration loop. We will borrow this key idea in deriving the solution of our AAAM algorithm.

4 Adaptive Active Appearance Model

Having introduced the conventional AAM, in this section we will first present how to learn the Adaptive Active Appearance Model (AAAM) from a video sequence. Then we will describe the method of fitting an AAAM to a video frame in detail, as well as the computational cost.

4.1 Model Learning

Our approach studies the problem of fitting a statistical face model to faces contained in a video sequence. Conventionally this video-based fitting problem can be solved via learning an AAM in an offline fashion and treat the fitting of the video frame almost the same as that of the static image. The only difference between video-based fitting and image-based fitting might be that in the former the fitting result of frame $t-1$ can be used as the initialization of frame t, while in the latter the initialization might rely on the face detection. The AAM can be offline learned from the training images of one subject with manual labels of landmarks, which is normally considered as person-specific AAM [16]. This type of AAM can fit very reliably to video sequences of the same subject. However, the fitting performance degrades quickly when tested on video sequences of unseen subjects, i.e., subjects different to the one in the training data. One popular way to remedy this problem is to offline learn a generic AAM using training images from a large set of subjects, with the hope that the appearance variation of the test subject can be well explained by the generic AAM. However, as shown in [16], the fitting performance of the generic AAM on unseen subjects is still not satisfying. This is commonly

known as the generalization problem of the conventional AAM.

It might be worthwhile to understand the cause of this generalization problem. We assert this is mostly due to the representational power of the appearance model. Even the generic AAM is trained from a large set of images, the appearance model only learns the appearance variation retained in the training data. It is very likely that the appearance model can not represent the appearance information of the unseen test subject sufficiently well. However, using the MSE as the distance metric essentially employs an "interpretation through synthesis" approach. Hence, if the appearance model is unable to *synthesize* an appearance instance that is "similar" enough to the test image, the fitting process will have difficult to *estimate/interprate* the correct landmark locations of the test image. This problem can be even severe considering that the appearance parameters, which determines the synthesized appearance instance, and the shape parameters, which controls the landmark locations, have to be estimated simultaneously during the fitting.

Motivated by increasing the representational power of the appearance model, we propose an Adaptive Active Appearance Model (AAAM), which is learned and updated in an online fashion from a video sequence.

$$\bar{\Im} = \{\mathbf{s}_i, A_j, M_k\}_{i\in[0,n],j\in[0,m],k\in[1,K]} \tag{4}$$

Our AAAM has the same shape model $\{\mathbf{s}_i\}$ as the generic AAM. While the appearance model have two components: $\{A_j\}$ is the appearance model learned offline from a large set of labeled data, which is the same as the appearance model of the generic AAM; $\{M_k\}$ is the appearance model learned online from

the video sequence being tested. Hence, $\{M_k\}$ will contain the appearance information of the subject in the video sequence, which will help dramatically on the representational power of the appearance model. We consider $\{M_k\}$ as a subject-specific appearance model.

Before introducing how the subject-specific appearance model $\{M_k\}$ is learned, we first briefly mention the objective function of model fitting using our AAAM. Given an AAAM $\bar{\Im}$ and a test video frame I_t at time t, AAAM uses the following objective function to perform the face model fitting:

$$J_t(\mathbf{p}, \lambda) = \sum_{\mathbf{x}\in\mathbf{s}_0} [A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) - I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))]^2 + \frac{w}{K} \sum_{k=1}^{K} \sum_{\mathbf{x}\in\mathbf{s}_0} [M_k(\mathbf{x}) - I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))]^2, \tag{5}$$

which is composed of two terms weighted by a weighting parameter w. The first term is the same as Eq. (3), i.e., the MSE between the warped image and the synthesized appearance instance from the generic appearance model $\{A_j\}$. The second one is the MSE between the current warped image $I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))$ and the subject-specific appearance model $\{M_k\}$, which is obtained in an on-line fashion from video frame observations at the previous time instances. Thus in the fitting of each video frame, both MSE-based distance measures are served as constraints to guide the fitting process.

In general, from the aforementioned object function, it is obvious that $\{M_k\}$ will be learned from the warped images of the previous video frames. There are different strategies in learning and updating $\{M_k\}$.

Firstly, it can be the warped image of the video frame at time $t-1$:

$$M_0(\mathbf{x}) = I_{t-1}(\mathbf{W}(\mathbf{x};\mathbf{p}_{t-1})). \tag{6}$$

In this case, one image template serves as the subject-specific appearance

model and is updated when the fitting is completed for each frame.

Secondly, rather than updating the appearance model at every frame, we only retain a set of K warped images from previous frames, who have the smallest MSE with respect to the generic appearance model, i.e., the first term in Eq. (5). All K warped images (appearance templates) will be treated as the subject-specific appearance model and used in the fitting of future video frames. Once the fitting of the frame $t-1$ is finished, if the MSE of frame $t-1$ is less than the largest MSE among the K appearance templates, the warped image $I_{t-1}(\mathbf{W}(\mathbf{x};\mathbf{p}_{t-1}))$ will replace the template with the largest MSE. In this approach, since the K appearance templates $\{M_k\}_{k\in[1,K]}$ are the best fitting results based on the measurement of the generic AAM, they are most likely to benefit the fitting of the remaining frames in the video sequence.

Thirdly, the warped images of L previous video frames averaged by a decaying factor can also represent $M_k(\mathbf{x})$:

$$M_0(\mathbf{x}) = \frac{1-r}{r(1-r^L)} \sum_{l=1}^{L} r^l I_{t-l}(\mathbf{W}(\mathbf{x};\mathbf{p}_{t-l})), \tag{7}$$

where r is a decaying factor between 0 and 1. Similar to the first strategy, one image template serves as the subject-specific appearance model and is updated at every frame.

In Section 5, we will conduct experiments on AAAM fitting using the first two strategies. Of course, other learning strategies of $\{M_k\}$ are also possible, for example, the average of L previous warped images without decaying, and a dynamic eigenspace model of the previous warped images [21]. In the latter case, an efficient eigenspace updating method can be used to sequentially add the most recent warped image into the model [20], and additional appearance

parameters of this eigenspace model should be incorporated into the the second term of Eq. (5). In this paper, we view the various strategies as recipes of model adaption. In other words, just like each recipe might be appreciated by different people, various strategies are also applicable on different types of videos. There is rarely a strategy that performs the best for all types of facial videos. Hence, it is not the focus of this work to directly compare the performance of various strategies. In the experiments, we will show that both the first and second strategy can improve the conventional AAM for video-based model fitting.

There are clear benefits from using both generic and subject-specific appearance models during the face model fitting. First of all, in practical applications there is always mismatch between the imaging environment of the images used for training face models and the images to be fit, as well as the presence of the specific appearance information of the subject being fit that is not modeled by the generic appearance models. Thus the distance-to-subject-specific-model is employed to bridge such a gap. Secondly, if we only use the subject-specific model, the alignment error might propagate over time. The generic model is well suited for preventing the error propagation and correcting the drifting.

Finally we also want to point out that there is one key assumption underlying our adaptation framework. That is, the generic AAM model component $\{\mathbf{s}_i, A_j\}$ is capable of fitting to certain video frames reasonably well. It is clear that when fitting to the first frame of a video sequence, $\{M_k\}$ is empty and the generic AAM model needs to provide reasonable fitting in order to start the learning of $\{M_k\}$. Otherwise $\{M_k\}$ will be learned from noisy appearance information and the potential of adaptation will be greatly limited. In practical scenarios, we argue this assumption can be satisfied because in general

a video sequence is often a blend of "easy frames" and "difficult frames" in terms of their level of difficulty for fitting. For example, images with frontal view faces are easy ones, while images with profile view faces are difficult ones. In a way, our adaption framework seeks to enhance the fitting of "difficult frames" by taking advantage of the subject-specific information from the "easy frames". In addition, the slowly varying facial appearance, such as illumination, expression, and poses, also makes model adaptation necessary.

4.2 Model Fitting

Given an AAAM $\bar{\Im}$ learned online and a video frame I_t, model fitting seeks to minimize the objective function Eq. (5) by estimating the optimal shape parameters $\mathbf{p}$ and appearance parameters λ simultaneously.

Using an approach similar to the IC and SIC algorithms [3], the proposed AAAM fitting algorithm iteratively minimizes:

$$\sum_{\mathbf{x}} \left[A_0(\mathbf{W}(\mathbf{x};\triangle\mathbf{p})) + \sum_{i=1}^{m}(\lambda_i + \triangle\lambda_i) A_i(\mathbf{W}(\mathbf{x};\triangle\mathbf{p})) - I_t(\mathbf{W}(\mathbf{x};\mathbf{p})) \right]^2 + \frac{w}{K}\sum_{k=1}^{K}\sum_{\mathbf{x}} \left[M_k(\mathbf{W}(\mathbf{x};\triangle\mathbf{p})) - I_t(\mathbf{W}(\mathbf{x};\mathbf{p})) \right]^2 \tag{8}$$

with respect to $\triangle\mathbf{p}$ and $\triangle\lambda = (\triangle\lambda_1, ..., \triangle\lambda_m)^{\mathrm{T}}$ simultaneously, and then updates the warp $\mathbf{W}(\mathbf{x};\mathbf{p}) \leftarrow \mathbf{W}(\mathbf{x};\mathbf{p}) \circ \mathbf{W}(\mathbf{x};\triangle\mathbf{p})^{-1}$ and the appearance parameters $\lambda \leftarrow \lambda + \triangle\lambda$.

In order to solve for $\triangle\mathbf{p}$ and $\triangle\lambda$, the non-linear expression in Eq. (8) is linearized by performing a first order Taylor series expansion on $A_0(\mathbf{W}(\mathbf{x};\triangle\mathbf{p}))$, $A_i(\mathbf{W}(\mathbf{x};\triangle\mathbf{p}))$, and $M_k(\mathbf{W}(\mathbf{x};\triangle\mathbf{p}))$, and assuming that $\mathbf{W}(\mathbf{x};\mathbf{0})$ is the iden-

tity warp. This gives:

$$\sum_{\mathbf{x}}\left[A_0(\mathbf{x})+\nabla A_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\triangle\mathbf{p}+\sum_{i=1}^{m}(\lambda_i+\triangle\lambda_i)(A_i(\mathbf{x})+\nabla A_i\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\triangle\mathbf{p})-I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))\right]^2 + \frac{w}{K}\sum_{k=1}^{K}\sum_{\mathbf{x}}\left[M_k(\mathbf{x})+\nabla M_k\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\triangle\mathbf{p}-I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))\right]^2. \tag{9}$$

The first term in the above equation can be simplified as follows by neglecting the second order terms:

$$\sum_{\mathbf{x}}\left[A_0(\mathbf{x})+\sum_{i=1}^{m}\lambda_iA_i(\mathbf{x})-I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))+(\nabla A_0+\sum_{i=1}^{m}\lambda_i\nabla A_i)\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\triangle\mathbf{p}+\sum_{i=1}^{m}A_i(\mathbf{x})\triangle\lambda_i\right]^2. \tag{10}$$

To simplify the notation, firstly we denote $\mathbf{q}=(\mathbf{p}^\intercal\lambda^\intercal)^\intercal$ and similarly $\triangle\mathbf{q}=(\triangle\mathbf{p}^\intercal\triangle\lambda^\intercal)^\intercal$. Thus $\mathbf{q}$ is a $n+m$ dimensional vector including both the shape parameters $\mathbf{p}$ and the appearance parameters λ. Secondly, we denote $n+m$ dimensional steepest-descent images:

$$\mathbf{SD}_0(\mathbf{x})=\left[(\nabla A_0+\sum_{i=1}^{m}\lambda_i\nabla A_i)\frac{\partial\mathbf{W}}{\partial p_1},...,(\nabla A_0+\sum_{i=1}^{m}\lambda_i\nabla A_i)\frac{\partial\mathbf{W}}{\partial p_n},A_1(\mathbf{x}),...,A_m(\mathbf{x})\right], \tag{11}$$

and

$$\mathbf{SD}_k(\mathbf{x})=\left[\frac{w}{K}\nabla M_k\frac{\partial\mathbf{W}}{\partial p_1},...,\frac{w}{K}\nabla M_k\frac{\partial\mathbf{W}}{\partial p_n},\mathbf{0},...,\mathbf{0}\right]. \tag{12}$$

Thirdly, we denote the error images:

$$E_0(\mathbf{x})=A_0(\mathbf{x})+\sum_{i=1}^{m}\lambda_iA_i(\mathbf{x})-I_t(\mathbf{W}(\mathbf{x};\mathbf{p})), \tag{13}$$

and

$$E_k(\mathbf{x})=\frac{w}{K}(M_k(\mathbf{x})-I_t(\mathbf{W}(\mathbf{x};\mathbf{p}))). \tag{14}$$

Equation (9) is simplified to:

$$\sum_{k=0}^{K}\sum_{\mathbf{x}}[E_k(\mathbf{x})+\mathbf{SD}_k(\mathbf{x})\triangle\mathbf{q}]^2. \tag{15}$$

The partial derivative of Eq. (15) with respect to $\triangle\mathbf{q}$ is:

$$2\sum_{k=0}^{K}\sum_{\mathbf{x}}\mathbf{SD}_k^{\mathsf{T}}(\mathbf{x})[E_k(\mathbf{x})+\mathbf{SD}_k(\mathbf{x})\triangle\mathbf{q}]. \tag{16}$$

The closed form solution of Eq. (8) is obtained by setting Eq. (16) to equal zero:

$$\triangle\mathbf{q}=-\mathbf{H}^{-1}\sum_{k=0}^{K}\sum_{\mathbf{x}}\mathbf{SD}_k^{\mathsf{T}}(\mathbf{x})E_k(\mathbf{x}), \tag{17}$$

where $\mathbf{H}^{-1}$ is the inverse of the Hessian matrix:

$$\mathbf{H}=\sum_{k=0}^{K}\mathbf{H}_k=\sum_{k=0}^{K}\sum_{\mathbf{x}}\mathbf{SD}_k^{\mathsf{T}}(\mathbf{x})\mathbf{SD}_k(\mathbf{x}). \tag{18}$$

The algorithm is summarized in Figure 2. Note that even the Hessian matrix is the summation of $K+1$ $\mathbf{H}_k$ matrixes as indicated by Eq. (18), the summation of K matrixes, $\sum_{k=1}^{K}\mathbf{H}_k$, can be pre-computed, which is fixed as long as the subject-specific appearance model is not updated. Once the model is updated, for example one of the K appearance templates M_k, we need to re-compute the following: the gradient, the steepest descent image and the Hessian matrix for this particular M_k, and $\sum_{k=1}^{K}\mathbf{H}_k$.

The computation cost of the AAAM fitting algorithm is summarized in Table 1. It can be observed that the additional subject-specific appearance model results in slight more computation in Step (2) and Step (8). However, given the fact $n+m \gg K$, the per iteration computation cost of AAAM fitting, $O((n+m)^2N+(n+m)^3+KnN)$, is almost the same as that of the SIC algorithm, $O((n+m)^2N+(n+m)^3)$ [2].

Pre-compute:

(3) Evaluate the gradients ∇M_k, and ∇A_i for $i \in [0, m]$, $k \in [1, K]$

(4) Evaluate the Jacobian $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$ at $(\mathbf{x}; \mathbf{0})$

(5) Evaluate the steepest descent images and the Hessian matrixes using Eq. (12) and $\sum_{k=1}^{K} \mathbf{H}_k$

Iterate:

(1) Warp I with $\mathbf{W}(\mathbf{x}; \mathbf{p})$ to compute $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$

(2) Compute the error images $E_k(\mathbf{x})$ using Eq. (13) and Eq. (14)

(6) Compute the steepest descent image $\mathbf{SD}_0(\mathbf{x})$ using Eq. (11)

(7) Compute the Hessian matrix $\mathbf{H}$ using Eq. (18) and invert $\mathbf{H}$

(8) Compute $\sum_{k=0}^{K} \sum_{\mathbf{x}} \mathbf{SD}_k^{\mathrm{T}}(\mathbf{x}) E_k(\mathbf{x})$

(9) Compute $\triangle \mathbf{q}$ using Eq. (17)

(10) Update $\mathbf{W}(\mathbf{x}; \mathbf{p}) \leftarrow \mathbf{W}(\mathbf{x}; \mathbf{p}) \circ \mathbf{W}(\mathbf{x}; \triangle \mathbf{p})^{-1}$ and $\lambda \leftarrow \lambda + \triangle \lambda$

until $||\triangle \mathbf{p}|| \leq \epsilon$

Fig. 2. Summary of the AAAM fitting algorithm.

5 Experiments

In this section, we will demonstrate the effectiveness of our proposed algorithm in fitting facial video sequences. We will use the SIC algorithm as the baseline for performance comparison because on one hand our algorithm is a direct extension of the SIC algorithm, on the other hand SIC is also one of the most well known methods in generic face alignment.

To evaluate our algorithm, we collect a large set of images and videos for the experiments, as shown in Table 2. The training set is composed of three public available databases: the ND1 database [9], which contains 200 facial images with mostly frontal views from 200 subjects, and the FERET database [26],

Table 1

The computation cost of the AAAM fitting algorithm. The right column indicates the total cost for the pre-computation and each iteration. n is the number of shape bases, m is the number of appearance bases, K is the number of appearance templates and N is the number of pixels in the mean shape domain.

Pre-comp.	Step 3	$O(mN + KN)$	
	Step 4	$O(nN)$	
	Step 5	$O(nN + n^2N)$	$O((n^2 + m + K)N)$
Per Iteration	Step 1	$O(nN)$	
	Step 2	$O(mN + KN)$	
	Step 6	$O((n + m)N)$	
	Step 7	$O((n + m)^2N + (n + m)^3)$	
	Step 8	$O((n + m)N + KnN)$	
	Step 9	$O((n + m)^2)$	
	Step 10	$O(n^2 + m)$	$O((n + m)^2N + (n + m)^3 + KnN)$

which contains 200 images each from one subject with various poses, and Cohn-Kanade facial expression database [18], which contains 563 images from 100 subjects with various emotion-specified expressions. Figure 3 shows sample images from these three databases. The test set is composed of a number of video sequences captured at our lab, whose subjects are not seen in the training set. We have made efforts to ensure that the test set is representative for the practical application scenarios. For example, our test set includes videos from both indoor and outdoor, as well as different types of facial variations, such

Table 2

Summary of the dataset.

name	**# images**	**# subjects**	**environment**	**variation**
ND1 [9]	200	200	indoor	frontal
FERET [26]	200	200	indoor	pose
CMU Exp. [18]	563	100	indoor	expression
GRC1	980	1	outdoor	pose
GRC2	970	1	outdoor	resolution
GRC3-5	18885	3	indoor	expression

(a) (b) (c)

Fig. 3. Examples of the face dataset: (a) ND1 database, (b) FERET database, (c) CMU Cohn-Kanade facial expression database.

as pose, resolution and expression.

As we mentioned in the Section 4.1, there are different strategies in learning and updating our AAAM. In the following sections, we will evaluate the performance of our AAAM algorithm using the first and second learning strategy respectively.

5.1 The First Strategy

In the first strategy, the subject-specific appearance model $\{M_k\}$ is simply the warped image of the previous frame, as defined in Eq. (6). The generic appearance model $\{A_j\}$ and the shape model $\{\mathbf{s}_i\}$ need to be learned offline. In our experiment, this offline learning is conducted on a 400-image set, where 200 of them are from the ND1 database and another 200 images are from the FERET database. Each image in the 400-image set is manually labeled with 33 landmarks. Iterative model enhancement [22] is used in the training stage and results in a more compact model than the conventional approach. The resulting shape model $\{\mathbf{s}_i\}$ has 10 shape bases, the appearance model $\{A_j\}$ has 52 appearance bases, and the width of the mean shape is 62 pixels.

Two outdoor surveillance video sequences, whose subjects are not included in the training set, are used for test. Both sequences are captured at 30 frames per second (FPS). For comparison purpose, we have implemented both the SIC and AAAM algorithms in Matlab$^{\text{TM}}$. By manually placing the mean shape on the first video frame, AAAM and SIC algorithms are used to fit the face model to these test videos respectively. The weighting parameter for the AAAM algorithm, w, is set to be 1. Ideally w should be dynamically set according on the *correctness* of the appearance template M_0. The first video sequence (GRC1) contains 980 frames. The proposed AAAM algorithm successfully fits the face over the entire video sequence while the SIC algorithm loses the fitting starting from frame 780 due to a large pose change. In the case where there is no manual label for each frame of the test video sequences, visual inspection of the fitting results is one way of evaluating the performance. Figure 4 shows the comparison between two methods on 6 frames in this video. A visually

more accurately fitted mesh is observed when using the AAAM algorithm. Note that the inaccurate fitting results of SIC is mainly due to the fact that the imaging environment of the surveillance test video is very different from that of the training data, which are high quality images captured indoor. This is also the well-known generalization issue of the conventional AAM approach.

Other than visual inspection, an alternative way to evaluate the fitting performance is to quantitatively compute the registration consistency across frames, which is represented by the MSE of the warped image observations between consecutive frames. As shown in Figure 5, AAAM provides on average lower MSE for the entire sequence, especially when SIC has high MSE at certain frames due to the changing facial appearance. Hence this shows superior frame-to-frame registration using the AAAM algorithm. On one hand, this is a favorable property for many applications that requires accurate registration across time, such as super-resolution from video sequences. On the other hand, this is also an expected result since the frame-to-frame registration measure is a part of the AAAM's objective function.

Our proposed method can improve not only the fitting robustness and accuracy, but also the fitting speed. Figure 6 shows the number of iterations for fitting each frame using the SIC and AAAM algorithm. The lower curve of AAAM indicates that AAAM can converge much faster than SIC. This improvement is expected because the additional constraint in AAAM helps the minimization procedure. Given the fact that the computation cost per iteration in the fitting is almost the same as SIC, the average time for fitting one frame using AAAM is much lower because less iterations are needed for fitting to converge. Based on a Matlab$^{\text{TM}}$ implementation running on a conventional 2.13 GHz Pentium$^{\text{TM}}$4 computer, on average AAAM takes 0.1254 sec. to fit

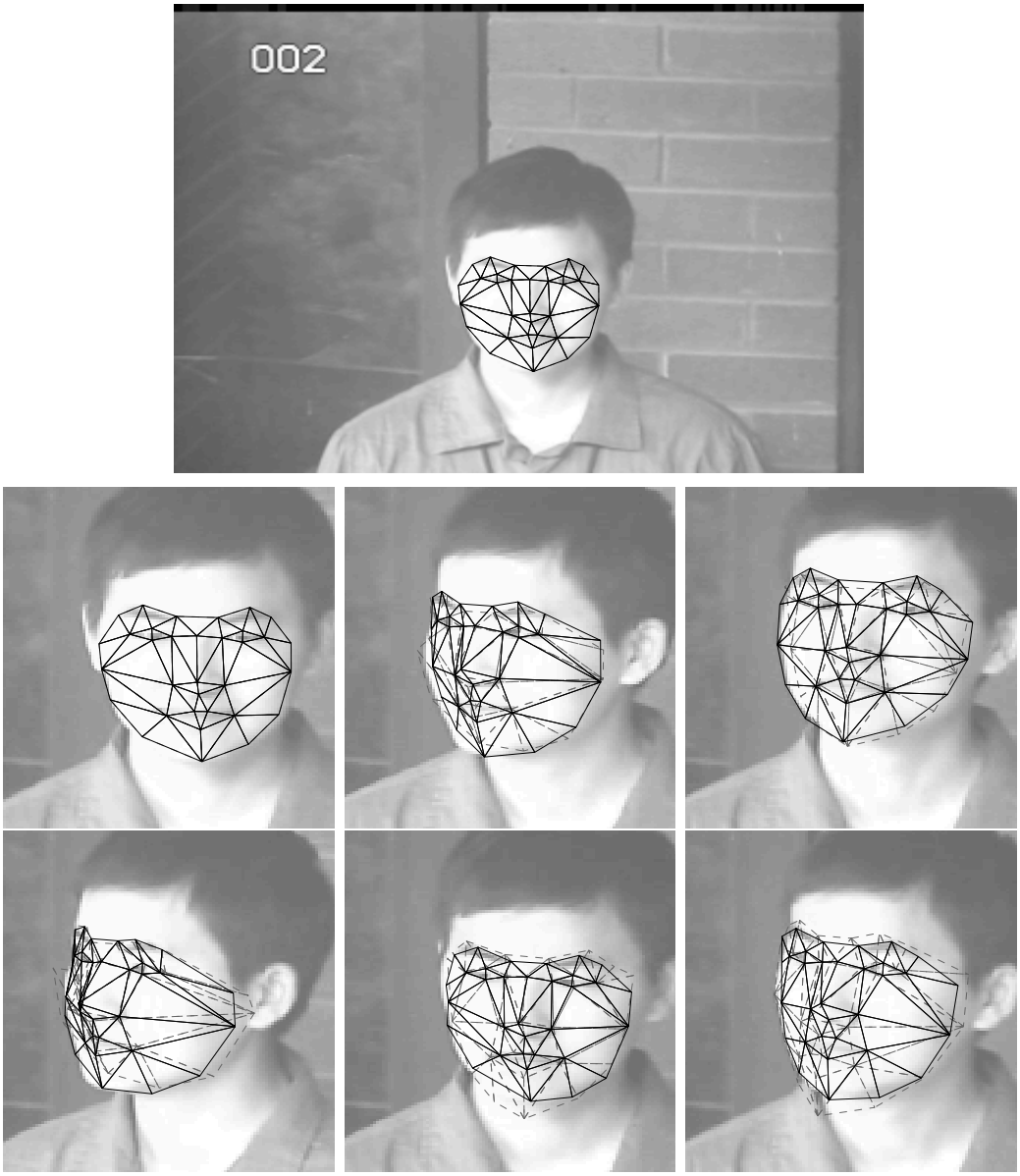

Fig. 4. Comparison of the fitted mesh using the AAAM algorithm (solid line) and the conventional SIC algorithm (dashed line) on 6 frames of video GRC3 (clipped views from frame 1, 40, 87, 287, 734 and 767). The top image is the un-clipped frame.

one frame compared to 0.2526 sec. by SIC.

Figure 7 shows the fitting results on another 970-frame-long video sequence (GRC2), where a Pan-Tilt-Zoom (PTZ) camera is pointing at three subjects

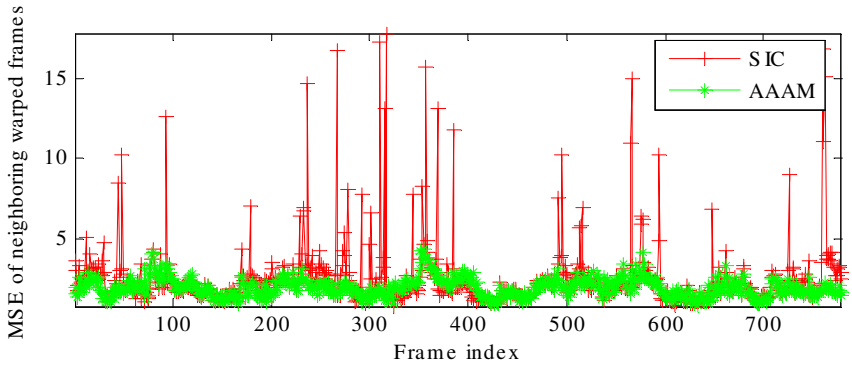

Fig. 5. The MSE of neighboring warped frames of a video sequence. Constant lower MSE indicates the improved frame-to-frame registration using the AAAM algorithm.

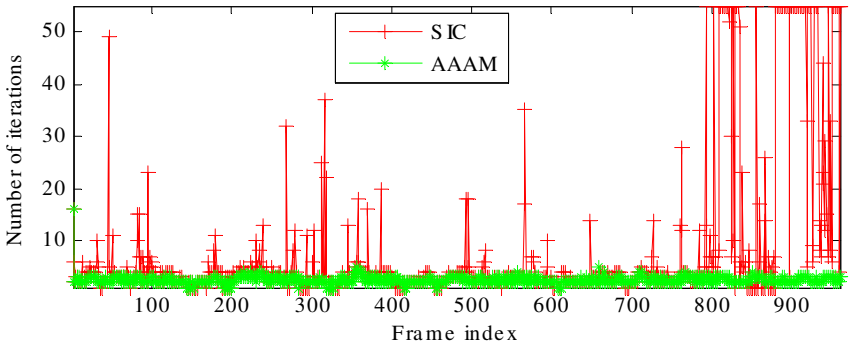

Fig. 6. The number of iterations in fitting each frame of a video sequence. Constant lower number of iterations is observed from the proposed AAAM algorithm.

and continuously zooming out [1] . This is to mimic the scenario where in surveillance applications the subjects can have various distance to a camera and the face image can be of low resolution. How to effectively fit a face model onto this type of challenging real-world video sequence receives relatively little attention in the vision community. The proposed AAAM algorithm successfully fits the entire video sequence, even when zooming happens and large scale change appears in consecutive frames. Note that the smallest face size in this video sequence only has the face width of 15 pixels. However, when applying the conventional SIC algorithm, the fitting diverges at frame 34 when the first

[1] Please play the supplementary video 1 for viewing this result.

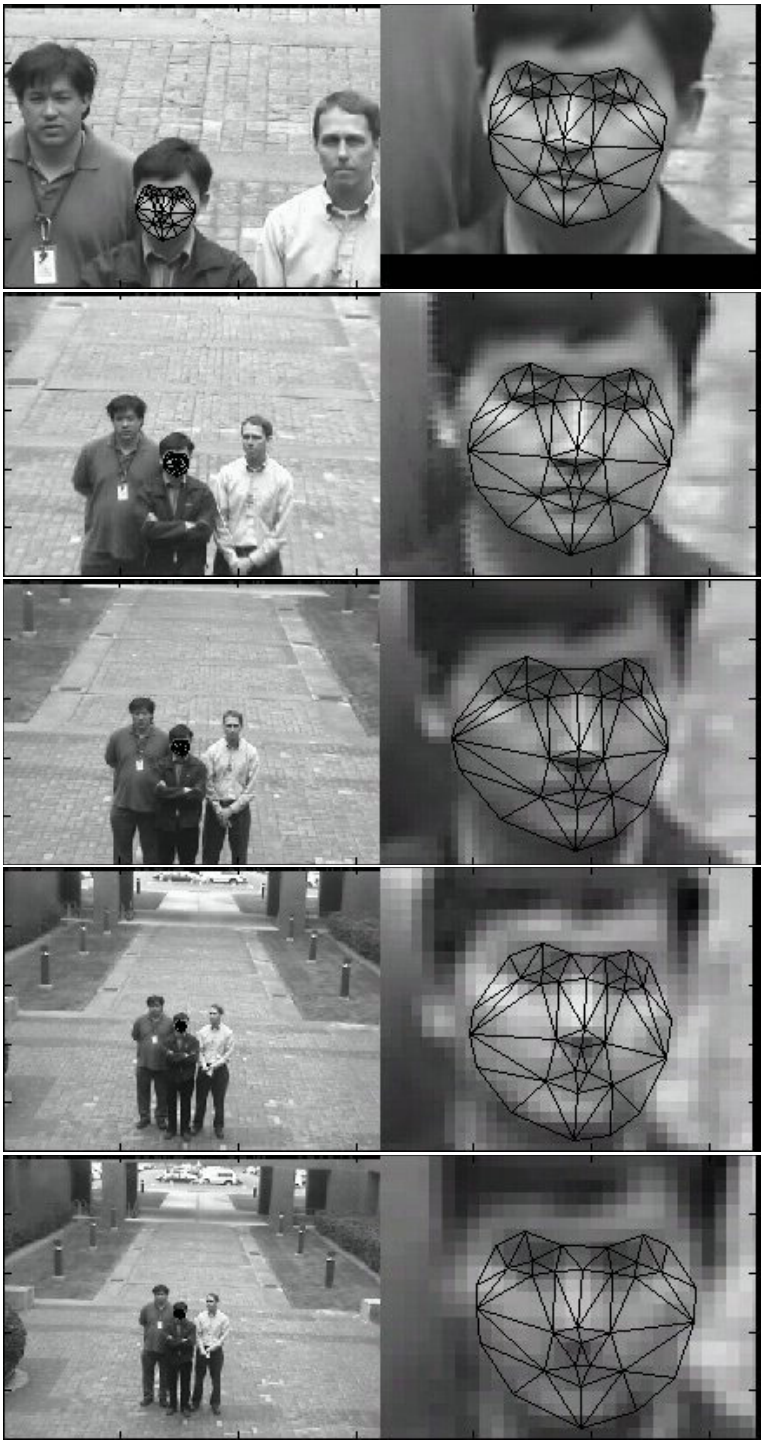

Fig. 7. Fitting results with zoom in facial area using AAAM. Reliable fitting is observed in dealing with zooming and low resolution, even for the facial area of 15 pixels wide (lower right).

zooming happens, and the fitting for the remaining frames are all failed.

In this section, we will present experimental results using the second learning strategy for the subject-specific appearance model. A subset of CMU Cohn-Kanade facial expression database including 563 images from 100 subjects is used as the training set, among which 97 images are with neutral expression and the remaining images are manually selected peak expression in each of the original expression session. Since we are interested in fitting face models to videos with subtle facial expression changes, each image in the training set is manually labeled with 72 landmarks, which is more than twice of the first strategy. Also, we apply iterative model enhancement [22] in the offline training stage. The resulting shape model $\{\mathbf{s}_i\}$ has 29 shape bases, the appearance model $\{A_j\}$ has 181 appearance bases, and the width of the mean shape is 126 pixels.

The test set includes three video sequences, one for each unseen subject. Each video is more than 7 minutes long and has 6295 frames in total. The size of the video frame is 960×720 pixels and the average face width is 150 pixels. These videos are captured via a web-cam while the subjects are asked to watch the same media content displayed on a LCD monitor. Since we choose a popular comics as the content, quite a lot facial expression, mostly smiling, can be observed from the captured videos.

Unlike the previous section, here we would like to have a quantitative analysis on the fitting performance. Hence, labeling the ground truth of landmark locations on the video sequences is required. Obvious manually labeling each frame of the entire sequence is practically impossible. We choose to rely on

the person-specific AAM, which is known to have excellent performance when the training and testing images are of the same subject [16], to provide the ground truth. That is, for each sequence, we randomly select 20 frames and manually label their landmarks. These 20 images and labels, together with the shape labels in the Cohn-Kanade database, are use to train an AAM for each sequence. Note that in the resulting AAM, the appearance model is learned from 20 images only, while the shape model attributes to both the 20 images and the labels in the Cohn-Kanade database. Finally the fitting results of each video sequence using its own AAM are treated as the ground truth. We visually go through the entire sequence and make sure that these ground truth landmarks are satisfying.

We first compare the performance of our AAAM algorithm with the conventional SIC algorithm. Given a video sequence, both algorithms initialize the first frame by placing the mean shape within the face detection window. For other frames, the fitted landmark locations from the previous frame will be used as the initialization. We evaluate the fitting performance by computing the MSE between the fitted landmark locations and the ground truth on a per-frame basis. For our AAAM algorithm, we choose K to be 4 and w to be 1.5. Figure 8 shows the comparison of two algorithms on all three test sequences. It can be observed that our algorithm not only achieves lower average MSE compared to SIC, but also much less variation in the estimation error. For example, the worst fitting error in AAAM is around 5 pixels, while for SIC a small set of frames have MSE well above 10 pixels, especially for the video GRC5. We show the fitting results of AAAM on the 8 frames of video GRC5 in Figure 9 [2]. Despite the large expression variation of the subject, our

[2] Due the large number of landmarks and triangles in the mesh representation, we

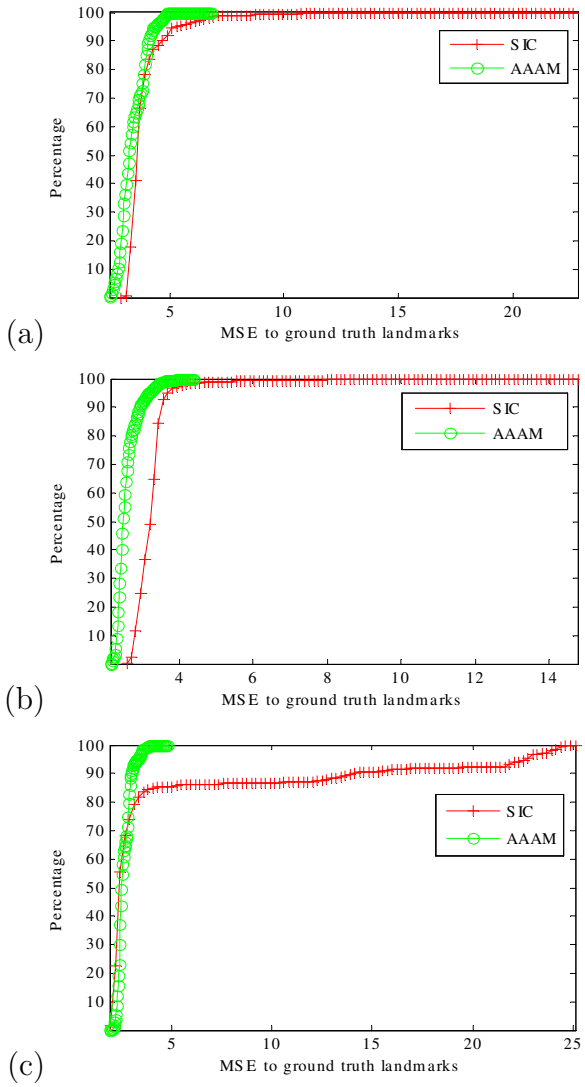

Fig. 8. Fitting performance comparison of AAAM and SIC on video GRC3, GRC4 and GRC5 from top to bottom. We plot the Cumulative Density Function (CDF) of the landmark estimation error (MSE between estimated landmarks and the truth), i.e., one point on the curve means that how much percentage of the frames have the MSE less than the horizontal axis.

approach still captures most of the expression changes.

The next experiment is to study the influence of the number of appearance

choose to plot the contour defined by the estimated landmarks for the visibility concern. Please play the supplementary video 2 for viewing this result.

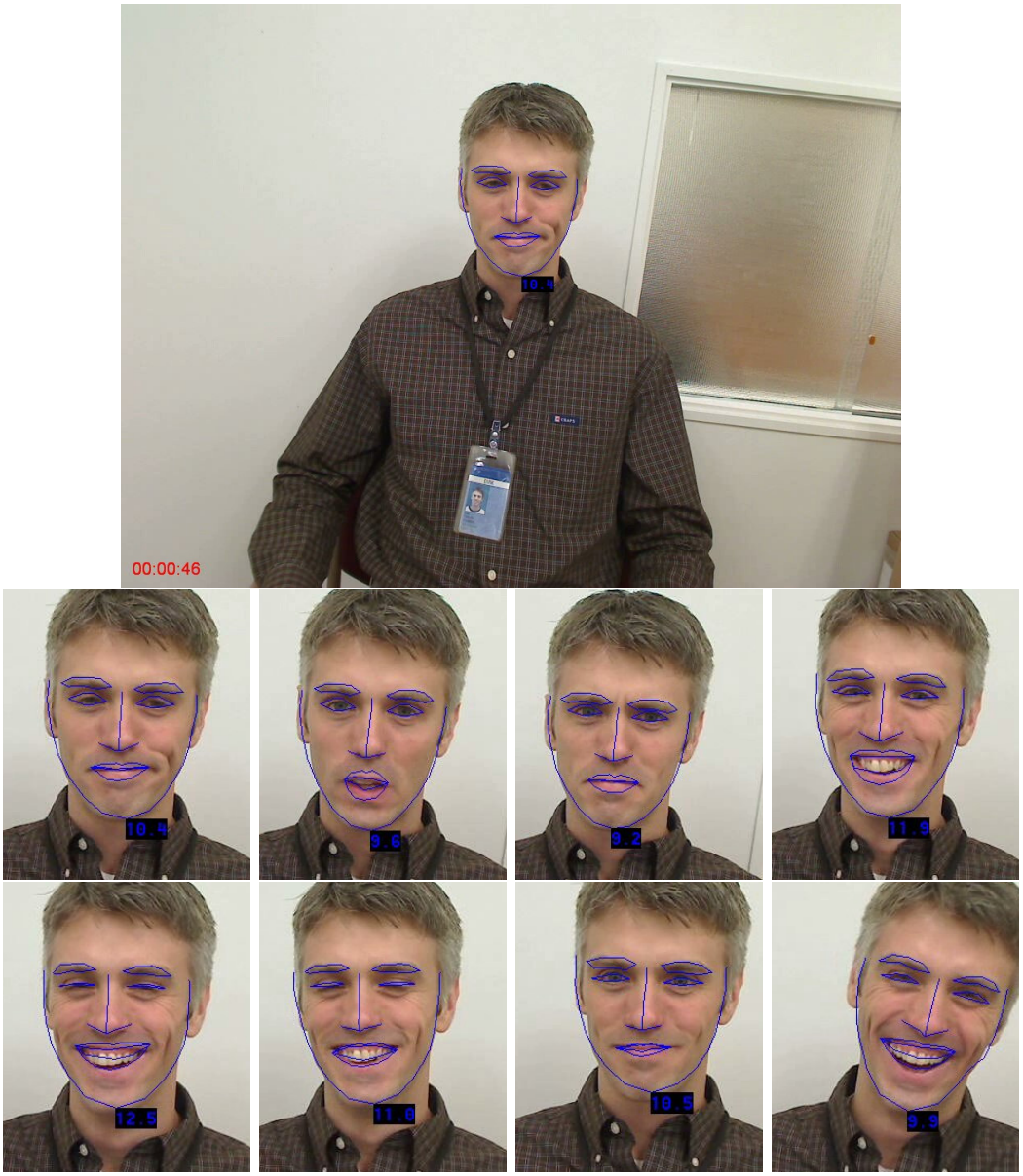

Fig. 9. One example un-clipped frame and fitting results on 8 frames (clipped view) of video GRC5 using the AAAM algorithm.

templates, i.e., K in the subject-specific appearance model $\{M_k\}$. We fix the weighting parameter w to be 1.5 and vary K to be 0, 1, 4, 8, and 12, respectively. Notice that when $K = 0$, our AAAM fitting algorithm is equivalent to the conventional SIC method since no subject-specific appearance template is used during the fitting. Table 3 illustrates the results. We first evaluate the mean of the MSE between the estimated landmarks and the ground truth for

Table 3

Performance with different number of subject-specific appearance templates K

	K	0	1	4	8	12
Mean(MSE) (pixel)	GRC3	3.92	3.72	3.38	3.43	3.53
	GRC4	3.31	3.29	2.59	3.67	3.47
	GRC5	4.88	2.79	2.64	2.54	2.77
Std(MSE) (pixel)	GRC3	1.08	0.68	0.56	0.50	0.55
	GRC4	0.66	0.35	0.29	0.25	0.24
	GRC5	6.00	0.63	0.33	0.35	0.47
Perc. of updates (%)	GRC3	0	0.1	0.6	1.8	3.7
	GRC4	0	0.6	2.4	4.4	2.9
	GRC5	0	0.3	5.1	9.7	12.9

all 6295 frames. It can be seen that the best performance is obtained when K is around 4. Also, the fact that $K = 4$ is always better than $K = 0$ shows that the AAAM approach outperforms the SIC method. In contrast, the larger K does not necessarily result in the better performance, as $K = 12$ always has worse performance than $K = 4$. Similar conclusion also holds for variance of the MSE for all frames. The smallest variance is obtained when K equals to 4 or 8, which is substantially smaller than that of the SIC method (when $K = 0$). The last measurement is the percentage of the frames being updated into the subject-specific appearance model. Note that in the second strategy, the updating happens only if we have a better fitting, i.e., the MSE of the cur-

Table 4

Performance with different weighting parameter w

	w	0.5	1	1.5	2	2.5
Mean(MSE) (pixel)	GRC3	3.76	4.01	3.38	3.64	3.51
	GRC4	3.15	4.14	2.59	3.21	3.00
	GRC5	2.82	2.60	2.64	2.76	2.63
Std(MSE) (pixel)	GRC3	1.14	0.62	0.56	0.78	0.55
	GRC4	0.85	0.43	0.29	0.33	0.31
	GRC5	0.65	0.44	0.33	0.38	0.32
Perc. of updates (%)	GRC3	0.5	0.7	0.6	3.8	52.4
	GRC4	1.3	0.1	2.4	3.5	63.3
	GRC5	12.9	3.2	5.1	4.5	46.8

rent warped image with respect to $\{A_j\}$ is less than the largest MSE among the K templates $\{M_k\}$. It is expected that the larger K is, the more frequent the subject-specific appearance model is updated. This is validated by our experimental results. Also notice that when $K = 4$, only a small percentage of the frames (up to 5.1%) among the entire video sequence is utilized for updating, which ensures that the updating would not hinder the efficiency of the fitting process.

We also compare the performance of our algorithm with respect to different weighting parameter w, as shown in Table 4. It can be seen that in general $w = 1.5$ is a good balance between the generic AAM and the subject-specific

appearance model. Too large weights will make the minimization process focus less on the first term, the MSE to the generic model, which is then unstable and results in a high percentage of model updating (last three rows in the table when $w = 2.5$). On the other hand, if too small weights are used, the subject-specific model contributes less to the fitting process. Hence our approach becomes similar to the conventional SIC method and high variances of the MSE are observed (Table 4 when $w = 0.5$).

6 Conclusions

This paper studies methods to effectively fit a mesh-based face representation to facial video sequences by using a novel statistical facial appearance and shape model. Motivated by improving the generalization ability of the conventional AAM, both a generic AAM and a subject-specific appearance model are employed simultaneously in the proposed model learning and fitting scheme. The subject-specific model is updated in an online fashion by making use of the test video sequence. Various online learning strategies are studied in this paper. By leveraging the idea of the SIC algorithm, we also introduce an efficient implementation of the fitting algorithm using the AAAM. Experimental results from various representative video sequences demonstrate the improved fitting robustness, accuracy and speed.

Future directions of this work can be experimenting with other learning strategies for the subject-specific model, such as Eq. (7), and as well as investigating the option of dynamically determining the weighting factor w based on the observed video frame.

Acknowledgement

This work was supported by awards #2005-IJ-CX-K060, #2006-IJ-CX-K045, and #2007-DE-BX-K191 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

References

[1] J. Ahlberg. An active model for facial feature tracking. *EURASIP Journal on Applied Signal Processing*, 2002(6):566–571, June 2002.

[2] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical Report CMU-RI-TR-03-35, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, November 2003.

[3] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Computer Vision*, 56(3):221–255, 2004.

[4] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(10):1380–1384, 2004.

[5] A.U. Batur and M.H. Hayes, III. Adaptive active appearance models. *IEEE Trans. Image Processing*, 14(11):1707–1721, 2005.

[6] R. Beichel, H. Bischof, F. Leberl, and M. Sonka. Robust active appearance models and their application to medical image analysis. *IEEE Trans. Medical Imaging*, 24(9):1151–1169, 2005.

[7] Johan G. Bosch, Steven C. Mitchell, Boudewijn P. F. Lelieveldt, Francisca Nijland, Otto Kamp, Milan Sonka, and Johan H. C. Reiber. Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Trans. Medical Imaging*, 21(11):1374–1383, 2002.

[8] C. Butakoff and A. Frangi. A framework for weighted fusion of multiple statistical models of shape and appearance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1847–1857, 2006.

[9] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *Proc. ACM Workshop on Multimodal User Authentication*, pages 25–32, December 2003.

[10] T. Cootes, D. Cooper, C. Tylor, and J. Graham. A trainable method of parametric shape description. In *Proc. 2nd British Machine Vision Conference, Glasgow, UK*, pages 54–61, September 1991.

[11] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[12] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, Imaging Science and Biomedical Engineering, University of Mancheste, March 2004.

[13] D. Cristinacce and T. Cootes. Facial feature detection and tracking with automatic template selection. In *Proc. 7th Int. Conf. on Automatic Face and Gesture Recognition, Southampton, UK*, pages 429–434, 2006.

[14] G. Dedeoglu, T. Kanade, and S. Baker. The asymmetry of image registration and its application to face tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):807–823, 2007.

[15] Rene Donner, Michael Reiter, Georg Langs, Philipp Peloschek, and Horst Bischof. Fast active appearance model search using canonical correlation

analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

[16] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, 2005.

[17] E. Jones and S. Soatto. Layered active appearance models. In *Proc. 10th Int. Conf. on Computer Vision, Beijing, China*, volume 2, pages 1097–1102, 2005.

[18] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. 4th Int. Conf. on Automatic Face and Gesture Recognition, Grenoble, France*, pages 46–53, 2000.

[19] Seth C. Koterba, Simon Baker, Iain Matthews, Changbo Hu, Jing Xiao, Jeffrey Cohn, and Takeo Kanade. Multi-view AAM fitting and camera calibration. In *Proc. 10th Int. Conf. on Computer Vision, Beijing, China*, volume 1, pages 511–518, October 2005.

[20] A. Levey and M. Lindenbaum. Sequential Karhunen-Loeve basis extraction and its application to images. *IEEE Trans. Image Processing*, 9(8):1371–1374, 2000.

[21] Xiaoming Liu, Tsuhan Chen, and Susan M. Thornton. Eigenspace updating for non-stationary process and its application to face recognition. *Pattern Recognition*, 36(9):1945–1959, 2003.

[22] Xiaoming Liu, Peter Tu, and Frederick Wheeler. Face model fitting on low resolution images. In *Proc. 17th British Machine Vision Conference, Edinburgh, UK*, volume 3, pages 1079–1088, 2006.

[23] Xiaoming Liu, Frederick Wheeler, and Peter Tu. Improved face model fitting on video sequences. In *Proc. 18th British Machine Vision Conference, University of Warwick, UK*, 2007.

[24] Iain Matthews and Simon Baker. Active appearance models revisited. *Int. J. Computer Vision*, 60(2):135–164, 2004.

[25] Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. In *Proc. 14th British Machine Vision Conference, Norwich, UK*, September 2003.

[26] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[27] B. Rolfe, M. Cardew-Hall, S. Abdallah, and G. West. Geometric shape errors in forging: developing a metric and an inverse model. *Proceedings of The Institution of Mechanical Engineers Part B- Journal of Engineering Manufature*, 215(9):1229–1240, 2001.

[28] Frederick W. Wheeler, Xiaoming Liu, Peter H. Tu, and Ralph Hoctor. Multi-frame image restoration for face recognition. In *Proc. IEEE Signal Processing Society Workshop on Signal Processing Applications for Public Security and Forensics (SAFE 2007), Washington, DC*, 2007.

[29] Shuicheng Yan, Ce Liu, Stan Z. Li, Hongjiang Zhang, Heung-Yeung Shum, and Qiansheng Cheng. Face alignment using texture-constrained active shape models. *Image and Vision Computing*, 21(1):69–75, 2003.