Face Model Fitting on Low Resolution Images †

Xiaoming Liu Peter H. Tu Frederick W. Wheeler

Visualization and Computer Vision Lab
General Electric Global Research Center
Niskayuna, NY, 12309, USA
{liux,tu,wheeler}@research.ge.com

Abstract

Active Appearance Models (AAMs) represent the shape and appearance of an object via two low-dimensional subspaces, one for shape and one for appearance. AAMs for facial images are currently receiving considerable attention from the vision community. However, most existing work focuses on fitting AAMs to high-quality facial images. For many applications, effectively fitting an AAM to low-resolution facial images is of critical importance. This paper addresses this challenge from two aspects. On the modeling side, we propose an iterative AAM enhancement scheme, which not only results in increased fitting speed, but also improves the fitting robustness. For fitting AAMs to low-resolution images, we build a multi-resolution AAM and show that the best fitting performance is obtained when the model resolution is slightly higher than the facial image resolution. Experimental results using both indoor video and outdoor surveillance video are presented.

1 Introduction

Active Appearance Models (AAMs) have been one of the most popular models for image registration [4]. Face registration using an AAM is receiving considerable attention because it enables facial feature detection, pose rectification, and gaze estimation. However, most existing work focuses on fitting the AAM to high-quality facial images. With the popularity of surveillance cameras and greater needs for face recognition at a distance, methods to effectively fit an AAM to low-resolution facial images are of increasing importance. This paper addresses this problem and proposes solutions for it.

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

†This project was supported by award #2005-IJ-CX-K060 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.

Conventional face modeling directly utilizes manual labels of facial landmarks when training the AAM. However, manual labels have various errors, which affect the shape and appearance models. To mitigate this problem, we propose a model enhancement scheme, where face modeling and model fitting are iteratively performed using the training image set. The iteration starts with the manual labels and stops when the fitted landmark locations do not change significantly. Compared to the initial AAM, the enhanced AAM has lower dimensionality, and the basis functions of the appearance model appear sharper visually, which is preferable for fitting. We experimentally show that the enhanced AAM not only increases the fitting speed, but also improves the fitting robustness.

In fitting an AAM to low-resolution images, there is a potential mismatch between the model resolution and the image resolution. To study how this mismatch affects the fitting performance, we build a multi-resolution AAM pyramid and use it in fitting facial images with various resolutions. We show that mismatched resolution can seriously degrade performance and conclude that the best fitting performance is obtained when the model resolution is slightly higher than the facial image resolution. We show that dynamically choosing the multi-resolution AAM at the right resolution improves performance when fitting to video sequences, where facial size varies with time.

Many approaches have been proposed for modeling faces via AAMs [4, 1]. Similar to our proposal, Gross *et al.* [9] suggested training AAMs using fitted landmarks. The primary difference is that our model enhancement is performed in an iterative fashion and results in greater improvements. A similar idea of iteratively adjusting landmarks was presented in [2, 8], where the initial landmarks are uniformly distributed and the focus is on the modeling, instead of model fitting, our present focus.

Little work has been done in fitting AAMs to low-resolution images. Cootes *et al.* [6] proposed a multi-resolution active shape model. However, its fitting strategy is very different compared to our approach. Recently Dedeoglu *et al.* [7] proposed integrating the image formulation process into the AAM fitting scheme. During the fitting, both image formulation parameters and model parameters are estimated in an united framework. The authors also showed the improvement of their method compared to fitting with a single high-resolution AAM. We will show that as an alternative fitting strategy, a multi-resolution AAM has far better fitting performance than using a high-resolution AAM.

2 Active Appearance Models and Model Fitting

The shape model and appearance model part of an AAM are trained with a representative set of facial images. The procedure for building 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, $\mathbf{s} = [x_1, y_1, x_2, y_2, ..., x_v, y_v]^T$. Eigen-analysis is applied to the observation set and the resulting 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 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

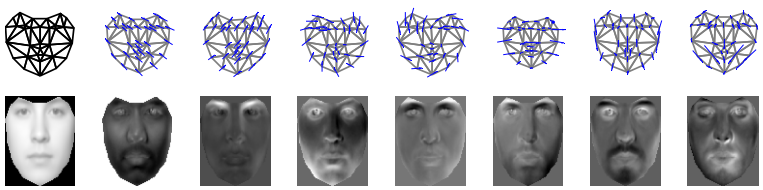

Figure 1: The mean and first 7 basis vectors of the shape model (top) and the appearance model (bottom) trained from the ND1 database. The shape basis vectors are shown as arrows at the corresponding mean shape landmark locations.

translation. Together with other basis vectors, a mapping function from the model coordination system to the coordinates in the image observation is defined as $\mathbf{W}(\mathbf{x};\mathbf{P})$, where $\mathbf{x}$ is a pixel coordinate within the face region $F(\mathbf{s}_0)$ defined by the mean shape $\mathbf{s}_0$.

Given the shape model, each facial image is warped into the mean shape via 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,

$$\mathbf{A}(\mathbf{x};\lambda) = \mathbf{A}_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \mathbf{A}_i(\mathbf{x}) \tag{2}$$

where $\mathbf{A}_0$ is the mean appearance, $\mathbf{A}_i$ is the appearance basis, and $\lambda = [\lambda_1, \lambda_2, ..., \lambda_n]$ are the appearance parameters. Note that the resolution of the appearance model here is the same as the resolution of the training images. Figure 1 shows an AAM trained using 534 images of 200 subjects from the ND1 3D face database [3].

An AAM can synthesize facial images with arbitrary shape and appearance within a 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 [5],

$$J(\mathbf{P},\lambda) = \frac{1}{N} \sum_{\mathbf{x}\in F(\mathbf{s}_0)} ||I(\mathbf{W}(\mathbf{x};\mathbf{P})) - \mathbf{A}(\mathbf{x};\lambda)||^2 \tag{3}$$

which is the mean-square-error (MSE) between the warped observation $I(\mathbf{W}(\mathbf{x};\mathbf{P}))$ and the synthesized appearance instance $\mathbf{A}(\mathbf{x};\lambda)$, and N is the total number of pixels in $F(\mathbf{s}_0)$.

Traditionally this minimization is solved by gradient-decent methods. Baker and Matthews [1] proposed the Inverse Compositional (IC) and Simultaneously Inverse Compositional (SIC) method that greatly improves the fitting performance. Their basic idea is that the role of appearance templates and the input image is switched when computing $\Delta\mathbf{P}$. Thus the time-consuming steps of parameter estimation can be pre-computed and remain constant during the iteration.

3 Face Model Enhancement

One requirement for AAM training is to manually position the facial landmarks for all training images. This is a time-consuming operation and is error-prone due both to the accuracy limitations of a manual operation, and also to different interpretations as to the

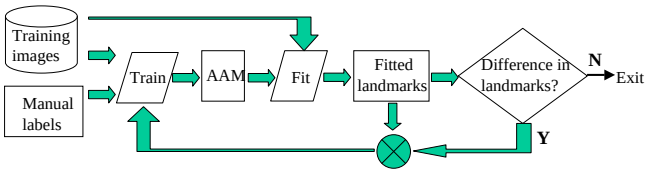

Figure 2: The diagram of AAM enhancement scheme. Iterative face modeling and model fitting are performed using the training images.

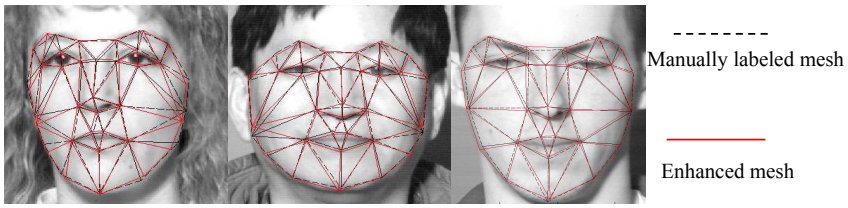

Figure 3: Comparison between the manual labels and the enhanced landmarks.

correct landmark locations. For example, when the same person processes a given image multiple times, noticeable differences in labeling can occur. Also, different people have differences for the same image because of different interpretations of the definition of certain landmarks, especially the ones along the outer boundary of the cheek. Since there is no distinct facial feature to rely on, it is not guaranteed that these landmarks correspond to the same physical position across multiple images.

The labeling error affects face modeling. First, the shape basis will model not only the inherent shape variation, but also the error of the labeling, which is not the goal of any modeling approach. Second, the appearance basis will contain less high-frequency detail due to poor alignment, which is an unfavorable property for model-based fitting.

To tackle the problem of labeling error, this paper proposes an AAM enhancement scheme, whose diagram is shown in Figure 2. Starting with a set of training images and manual labels, an AAM is trained using the above method. Then the AAM is fit to the same training images using the SIC algorithm, where the manual labels are used as the initial location for fitting. This fitting yields new landmark positions for the training images. This process is iterated. This new landmark set is used for face modeling again, followed by model fitting using the new AAM. The iteration continues until there is no significant difference between the landmark locations of the current iteration and the previous iteration. In the face modeling of each iteration, the basis vectors for both the appearance and shape models are chosen such that 98% and 99% of the energy are preserved, respectively.

Using a subset of 534 images from 200 subjects in the ND1 database, we have implemented this AAM enhancement scheme. After the enhancement process converges, we expect that the new set of landmark locations will deviate from the manual labels by a different amount for each image. Figure 3 shows two sets of landmarks for three images that are among the ones with the largest amount of deviation. A number of observations can be made. First, most of the landmarks with large deviation appear on the outer boundary of the cheek, which is consistent with the fact that they have inherent ambiguity in their definition. Second, most of the landmarks seem to migrate toward the correct position. This is expected given the assumption that people do not make consistent labeling errors.

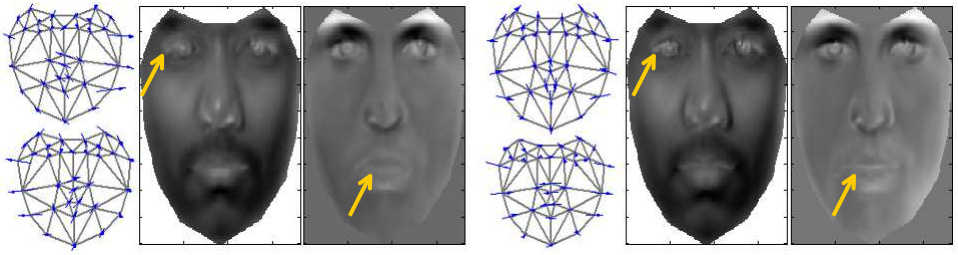

Figure 4: The 6^{th} and 7^{th} shape basis and the 1^{st} and 4^{th} appearance basis before enhancement (left) and after enhancement (right). After enhancement, more symmetric shape variation is observed, and certain facial areas appear sharper.

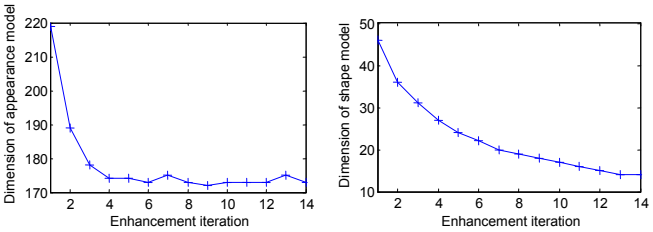

Figure 5: The improved compactness of the appearance model (left) and the shape model (right) during the iterative enhancement process, which converges in 14 iterations.

With the refined landmark locations, the AAM is expected to be improved as well. We show the enhancement effect in Figure 4. Note that the variation of landmarks around the outer boundary of the cheek becomes more symmetric after enhancement. Also, certain facial areas, such as the left eye boundary of the 1^{st} appearance basis and the lips of 4^{th} appearance basis, are visually sharper after enhancement. This is because the training images are better aligned thanks to improved landmark location accuracy.

In addition to the improved AAM basis, another benefit of this enhancement is improved compactness of the face model. As illustrated in Figure 5, both the appearance and shape models use fewer basis vectors to represent the same amount of variation. This is natural for the shape model because the variation due to labeling error is reduced during the enhancement. Thus fewer shape basis vectors are needed because only the inherent shape variation is modeled. For the appearance model, this is also consistent with the observed sharpness in the appearance basis.

There are at least two benefits of a more compact AAM. One is that fewer shape and appearance parameters need to be estimated during model fitting. Thus the minimization process is less likely to become trapped in a local minima, and fitting robustness is improved. The other is that the model fitting can be performed faster because the computation cost directly depends on the dimensionality of the shape and appearance models.

4 Multi-resolution AAM

This section introduces a method to construct a multi-resolution AAM. The terms *high-resolution* and *low-resolution* are used frequently below, in a *relative*, not absolute, sense.

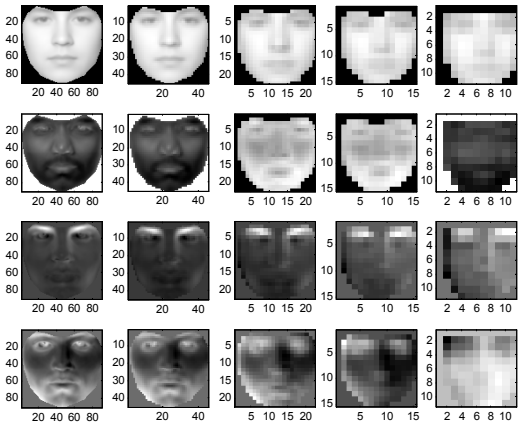

Figure 6: The appearance models of a multi-res AAM: Each column shows the mean and first 3 basis vectors at relative resolutions 1/2, 1/4, 1/8, 1/12 and 1/16 respectively.

The traditional AAM algorithm makes no distinction with respect to the resolution of the test images being fit. Normally the AAM is trained using the full resolution of the training images. That is, the number of pixels in the appearance basis is roughly equal to the number of pixels in the facial area of training images. We call this AAM a *high-res AAM*. When fitting a high-res AAM to a low-resolution image, an up-sampling step is involved in interpolating the observed image and generating a warped input image, $I(\mathbf{W}(\mathbf{x};\mathbf{P}))$. This can cause problems for the analysis by synthesis based fitting method. A high-res AAM has high frequency components that a low-resolution image observation does not contain. Thus, even with perfect estimation of the model parameters, the warped image will always have high frequency residual with respect to the high resolution model instance. At a certain point, this high frequency residual will overwhelm the residual due to the model parameter errors. Hence, fitting becomes problematic.

In a recent paper, Dedeoglu *et al.* [7] proposed to bridge the gap between the model resolution and the image resolution by integrating the image formulation process into the AAM fitting scheme. Our alternative approach separately trains the appearance model at each resolution leading to model compactness at lower resolution.

The basic idea of applying multi-resolution modeling to AAM is straightforward. Given a set of facial images, we down-sample them into low-resolution images at multiple scales. We then train an AAM using the down-sampled images at each resolution. We call the pyramid of AAMs a *multi-res AAM*. For example, Figure 6 shows the appearance models of a multi-res AAM at relative resolutions 1/2, 1/4, 1/8, 1/12 and 1/16. Comparing the AAMs at different resolutions within the multi-res AAM, we can see that the AAMs at lower resolutions have more blurring than the AAMs at higher resolutions. Also, the AAMs at lower resolutions have fewer appearance basis vectors compared to the AAMs at higher resolutions, which will benefit the fitting.

The landmarks used for training the AAM for the highest resolution are obtained using the enhancement scheme of Section 3. The landmarks for the other resolutions are found by appropriately scaling the landmark locations from the highest resolution AAM. So, the mean shapes of a multi-res AAM differ only by a scaling factor, while the shape basis vectors from different scales of the multiple-resolution AAM are exactly the same.

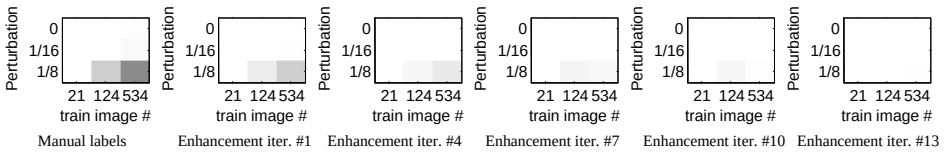

Figure 7: The convergence rate of fitting using an AAM trained from manual labels, and AAM after enhancement iteration number 1, 4, 7, 10 and 13. Continuing improvement of fitting performance is observed as the enhancement process progresses.

5 Experiments

It is well known that the fitting capability differs in a generic AAM and a person-specific AAM [9]. In order to study both cases, two facial datasets are used in the experiments. One is the ND1 face database [3], which contains 953 2D and 3D facial image pairs. We use a 2D image subset with 534 images from 200 subjects. The other dataset is collected with a PC camera and contains multiple video sequences of one subject.

In Section 3 we have shown the improvement of the AAM after model enhancement. However, the ultimate criterion of model enhancement is how much it improves the fitting performance. There are various measurements in evaluating the fitting performance. A popular one is the convergence rate with respect to different levels of perturbation on the initial landmark locations. The fitting is converged if the average MSE between the estimated landmarks and the true landmarks is less than a threshold. We adopt this measurement in this paper. Given the true landmarks of one image, we randomly deviate each landmark within a rectangular area up to a certain range, and the projection of the perturbed landmarks in the shape model is used as the initial shape parameters. Three different perturbation ranges, R, are used: 0, 1/16 and 1/8 of the facial height.

Another varying factor for the experiment is the number of images/subjects in the training set. When multiple images of one subject are used for training an AAM, the resulting AAM is considered as a person-specific AAM. When the number of subjects in the training set is large, the resultant AAM is a generic AAM. The more subjects used, the more generic the AAM is. Using the ND1 database, we test the modeling with three different population sizes, where the numbers of images are 21, 124, 534, and the corresponding numbers of subjects are 5, 25, 200 respectively.

Figure 7 shows the convergence rate of AAM fitting after a varying number of model enhancement iterations. The leftmost plot shows the convergence rate using an AAM trained from manual labels only, with varying population size and perturbation window size. Each element represents the convergence rate, which is computed using the same training set as test images being fit, between 0% to 100% via its brightness. There are some non-converged cases when more generic models are used with a larger perturbation window size. The rest of the plots show the convergence rate using the AAM trained after 1, 4, 7, 10 and 13 iterations of the enhancement algorithm. Continuing improvement of fitting performance is observed with additional enhancement iterations. After the model enhancement is completed, the fitting process converges for all test images, no matter how generic the model is and how large perturbation the initialization has.

Table 1 shows the computation cost for the fitting performed in Figure 7. To save space, only fitting using the model trained with manual labels, landmarks from enhance-

	Manual labels			Iteration #1			Iteration #13		
# of images	21	124	534	21	124	534	21	124	534
R=0	0.30	0.39	0.52	0.15	0.20	0.26	0.12	0.16	0.23
R=1/16	1.17	2.07	3.68	0.36	0.60	0.95	0.24	0.34	0.46
R=1/8	5.25	8.67	13.94	1.14	1.90	3.29	0.58	0.88	1.29

Table 1: Average fitting speed (sec.) before and after enhancement, verses perturbation range R, and the number of training images. Substantial improvement on fitting speed is observed with model enhancement.

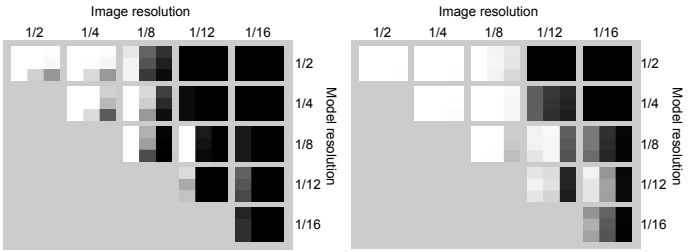

Figure 8: The convergence rate of fitting a multi-res AAM trained with manual labels (left) and enhanced landmarks (right) to images with different resolution. Each 3 by 3 block has the same axes as in Figure 7.

ment iteration number 1 and 13 are shown. The cost is averaged across converged fitting based on a Matlab implementation running on a conventional 2.13 GHz Pentium^{TM}4 computer. It can be seen that after model enhancement, the fitting speed is substantially faster than the one with manual labels, as well as the one with only a single iteration.

The second experiment is to test the fitting performance of a multi-res AAM on images with different resolutions. The same dataset and test scheme are used as in the previous experiment, except that the different resolutions of down-sampled training images are also used as test images for fitting. Model fitting is conducted with all combinations of AAM resolution and image resolution, where the model resolution no less than the image resolution. Note that lower resolution images have a proportionally lower threshold of convergence. For example, the convergence threshold of test images at 1/12 resolution is 1/6 of the images at 1/2 resolution. As shown in Figure 8, the AAM trained with enhanced landmarks performs much better than the AAM trained from manual labels. Also, for low-resolution images, the best fitting performance is obtained when the model resolution is slightly higher than the facial image resolution, which is far better than fitting using the AAM with the highest model resolution. This shows that the additional appearance detail in the higher resolution AAM does not help the model fitting. In fact, it seems to confuse the minimization process and results in degraded fitting performance.

In practice, as a person moves, facial size in a video sequence changes over time. The effective resolution of the face is dynamic. When a multi-res AAM is used to fit to a video sequence, the AAM model at the *right* resolution should be chosen to perform the fitting according to the actual facial size in each frame. We have performed an experiment using a facial video dataset. Eighty facial images from a single subject are used to train a multi-res AAM, where model enhancement is utilized to improve the landmark locations. Given

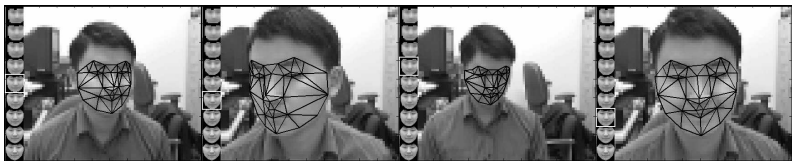

Figure 9: Fitting a multi-res AAM to a low resolution (60X80) video. The model resolution selected for fitting the current frame is highlighted with a white box.

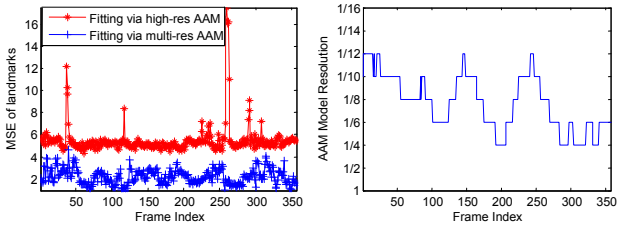

Figure 10: Fitting errors of the video (Figure 9) using a high-res AAM and a multi-res AAM (left), and the model resolution being chosen in multi-res AAM fitting (right).

a different (unseen) test video sequence with varying facial sizes of the same subject, AAM model fitting is performed on the original frame resolution (480X640) to obtain the ground truth of the landmarks for each frame. Then the test video is down-sampled to 1/8 of the original resolution and fit via a multi-res AAM. During the fitting of each frame, one particular AAM is chosen from the AAM pyramid such that the model resolution is slightly higher than the facial resolution in the current frame. Fitting results on four frames are shown in Figure 9, where obvious pose, expression variation and resolution changes are observed. The MSE between the estimated landmarks and the ground truth is used as the performance measurement. We compare the fitting performance with the one using a high-res AAM and plot the results in the left of Figure 10. Consistent reduced error in landmark estimation is observed with multi-res AAM fitting. The right part of Figure 10 shows the model resolution being chosen for each frame during the fitting. Using a multi-res AAM also greatly improves the fitting speed, where each frame takes on average 14.6 iterations (0.11 sec.) to converge, compared to 21.8 iterations (5.41 sec.) per frame using the high-res AAM based fitting.

The final experiment is model fitting on surveillance video. We collected an outdoor video sequence of the same subject as in Figure 9 using a PTZ camera. Sample fitting results using a multi-res AAM are shown in Figure 11. Although the frame size is 480 by 640 pixels, the facial area is not only at a low resolution, but also suffers from strong blurring, specular, and interlacing effects, which makes fitting a very challenging task. Our multi-res AAM continuously fits around 100 frames and provides reasonable results. However, the high-res AAM only successfully fits the first 4 frames in this sequence.

6 Conclusions

This paper studied methods to effectively fit an AAM to low-resolution images from two aspects. On the face modeling, we proposed an iterative AAM enhancement scheme,

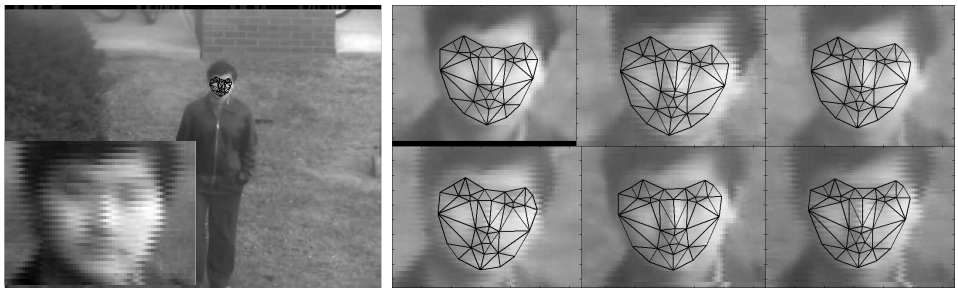

Figure 11: Fitting a multi-res AAM to an outdoor surveillance video: One sample frame with zoom in facial area (left) and six zoom in frames overlaid with fitting results (right).

which not only increases the fitting speed, but also improves the fitting robustness. For model fitting, we developed a multi-res AAM based on the finding that the best fitting performance is obtained when the model resolution is similar to the facial image resolution. Extensive experimental results showed the improved performance using our methods.

References

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

[2] 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, October 2004.

[3] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *In Proc. Multimodal User Authentication Workshop*, December 2003.

[4] 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. Springer, September 1991.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(1):681–684, 2001.

[6] T. Cootes, C. Taylor, and A. Lanitis. Active shape models: Evaluation of a multi-resolution method for improving image search. In *Proc. 5th British Machine Vision Conference, York, UK*, volume 1, pages 327–336. Springer, September 1994.

[7] G. Dedeoglu, S. Baker, and T. Kanade. Resolution-aware fitting of active appearance models to low-resolution images. In *Proceedings of the 9th European Conference on Computer Vision*. Springer, May 2006.

[8] G. Doretto. Modeling dynamic scenes with active appearance. In *IEEE Computer Vision and Pattern Recognition, San Diego, California*, volume 1, pages 66–73, 2005.

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