Resolution-Aware Fitting of Active Appearance Models to Low Resolution Images

Göksel Dedeoğlu, Simon Baker, and Takeo Kanade

The Robotics Institute, Carnegie Mellon University
{dedeoglu, simonb, tk}@cs.cmu.edu

Abstract. Active Appearance Models (AAM) are compact representations of the shape and appearance of objects. Fitting AAMs to images is a difficult, non-linear optimization task. Traditional approaches minimize the L2 norm error between the model instance and the input image warped onto the model coordinate frame. While this works well for high resolution data, the fitting accuracy degrades quickly at lower resolutions. In this paper, we show that a careful design of the fitting criterion can overcome many of the low resolution challenges. In our *resolution-aware formulation* (RAF), we explicitly account for the finite size sensing elements of digital cameras, and *simultaneously* model the processes of object appearance variation, geometric deformation, and image formation. As such, our Gauss-Newton gradient descent algorithm not only synthesizes model instances as a function of estimated parameters, but also simulates the formation of low resolution images in a digital camera. We compare the RAF algorithm against a state-of-the-art tracker across a variety of resolution and model complexity levels. Experimental results show that RAF considerably improves the estimation accuracy of both shape and appearance parameters when fitting to low resolution data.

1 Introduction

Image analysis at low resolution has its challenges. Due to camera blur, objects appear fuzzy, lose their boundaries, and start looking alike. This degradation makes detection, localization, and classification tasks increasingly more difficult, if not impractical.

In this paper, we focus on the tracking performance of Active Appearance Models (AAM) [5, 7] in low resolution regimes. Fitting AAMs is a non-trivial optimization task [10]. Traditional approaches minimize the L2 norm error between the model instance and the input image warped onto the model coordinate frame [5, 7, 10]. While this formulation works well for high resolution data, its accuracy degrades quickly at lower resolutions.

Any representation, model, and/or algorithm will perform poorly under conditions they are not built for, and the fitting of AAMs is no exception. In this paper, we diagnose why the traditional model fitting degrades, and propose a remedy. We show that a careful redesign of the AAM fitting criterion can indeed overcome accuracy degradation at low resolution.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 83–97, 2006.

2 Background

2.1 Active Appearance Models

An AAM [5, 7] consists of two models, namely the *shape* and *appearance* of an object. Each of these is a linear, Principal Components model learned from training data. The shape of an AAM is defined by a set of landmark locations

$$\mathbf{s} \; = \; (x_1, y_1, x_2, y_2, \ldots, x_v, y_v)^{\mathrm{T}}. \tag{1}$$

The shape model, parametrized with $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, expresses any shape as a linear combination of basis shapes added onto a base shape:

$$\mathbf{s}(\mathbf{p}) \; = \; \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i. \tag{2}$$

An AAM is defined in the coordinate system of the object being modeled. To express object instances in arbitrary poses, a global transform is needed. Following [10], we define four special shape bases to account for similarity transforms (scale, rotation, and two translations), and compose them with the shape model. We denote the combined geometric deformation by $\mathbf{W}(\mathbf{x}; \mathbf{p})$, where $\mathbf{x}$ is a model point coordinate being mapped onto an image coordinate.

The appearance model consists of the mean and basis images. These images are shape-normalized, *i.e.*, they are defined within the base shape $\mathbf{s}_0$. The appearance model is linear, and parametrized with $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ as

$$A(\mathbf{x}; \boldsymbol{\lambda}) \; = \; A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) \qquad \forall \; \mathbf{x} \in \mathbf{s}_0, \tag{3}$$

where $\mathbf{x}$ is a pixel coordinate in $\mathbf{s}_0$. The appearance basis images are usually defined at the same resolution as the training images.

In this paper, we consider the simpler case of *independent* AAMs [10], where the statistical dependence between the shape and appearance is ignored. While such couplings have been exploited in prior work, their advantages remain orthogonal to our discussion.

2.2 Traditional Fitting Formulation

Given a set of AAM parameters, the linear generative equations (2) and (3) can uniquely synthesize an object instance. Image analysis deals with the inverse of this process. It aims to recover those AAM parameters which *best* explain a given image. For this end, one needs to define a similarity metric to quantify what constitutes a good match, and a *fitting* algorithm for computing the parameter values which optimize the similarity metric. The choice of this fitting criterion is the main subject of this paper.

In the original AAM work by Cootes et al. [5,6,7], as well as its computationally efficient reformulation by Matthews and Baker [10], the fitting criterion was the sum of squared intensity differences between the synthesized model template and the *warped input image* I:

$$\sum_{\mathbf{x}\in\mathbf{s}_0}\Big[I\big(\mathbf{W}(\mathbf{x};\mathbf{p})\big)-A(\mathbf{x};\boldsymbol{\lambda})\Big]^2. \tag{4}$$

Note that the summation above is defined over $\mathbf{x}$, pixel coordinates in the shape-normalized template image. Since this objective function is highly nonlinear in its parameters, iterative gradient-descent methods were used to find its minimum: At each iteration, updates $\Delta\mathbf{p}$ and $\Delta\boldsymbol{\lambda}$ were computed and added to (or composed with) current estimates of $\mathbf{p}$ and $\boldsymbol{\lambda}$, respectively. Cootes et al. [5, 6, 7] assumed a constant, linear relationship between the error image and the additive updates. They learned this mapping through regression on perturbation-based training data. Matthews and Baker [10] explored linearizing the objective function just as in the Lucas-Kanade [2] registration algorithm, and achieved computational savings by switching the roles of the template and input images [9] in computing the warp update $\Delta\mathbf{p}$.

2.3 The Unsuspected Culprit in Low Resolution Problems

Any search method for optimizing the criterion (4) would suffer from a large number of local minima. In some cases, the solution might even be ambiguous. To make matters worse, these difficulties are only exacerbated when the available data is noisy and low in resolution, such as in surveillance imagery.

Let $\mathbf{u}$ denote the pixel coordinates of a low resolution observation I. As visualized in Fig. 1, the fitting criterion (4) prescribes *first warping and interpolating* the image I, and *then* comparing it against the synthesized template. Recall that the summation in (4) is defined over the pixels of the template. The latter is

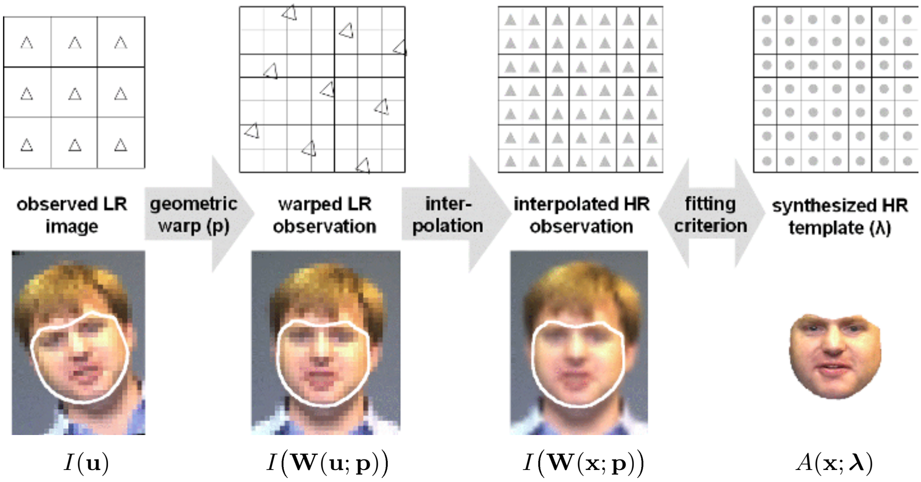

observed LR image
geometric warp (p)
warped LR observation
inter-polation
interpolated HR observation
fitting criterion
synthesized HR template (λ)
$I(\mathbf{u})$
$I\big(\mathbf{W}(\mathbf{u};\mathbf{p})\big)$
$I\big(\mathbf{W}(\mathbf{x};\mathbf{p})\big)$
$A(\mathbf{x};\boldsymbol{\lambda})$

Fig. 1. Graphical representation of the traditional fitting criterion of (4). From left to right, observed images get warped, interpolated, and finally compared against the synthesized model instance. When the input image is low in resolution, significant interpolation is needed to warp it onto the model coordinate frame.

normalized to shape $\mathbf{s}_0$ at the AAM's native resolution, and remains fixed in size. Consequently, when objects appear small in comparison to the AAM, they need to be enlarged through interpolation.

This reliance on interpolation used in the traditional formulation turns out to be its *Achilles' heel* in low resolution regimes. The fitting *criterion* itself becomes increasingly suboptimal (in accuracy) with higher scaling factors. This is an artifact of formulation. Using the same gradient-descent algorithm and low resolution data, but minimizing a more carefully designed fitting criterion, we will show that we can overcome low resolution challenges.

3 Resolution-Aware Fitting (RAF)

3.1 Formulation

We propose an alternative to the fitting criterion (4). In order to better account for low resolution data, our formulation takes a generative point of view and incorporates the image formation model of a typical CCD camera [1]. We feed the AAM and its current parameters into a camera model, and compare the outcome against the observed low resolution image. Mathematically, the proposed fitting criterion is

$$\sum_{\mathbf{u}\in I} \left[I(\mathbf{u}) - B\big(\mathbf{u}; A(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) \right]^2 , \qquad (5)$$

where the summation is now over pixel coordinates $\mathbf{u}$ of the observed image I. The operator B simulates a low resolution image of the object, believed to be what the camera would have captured under current AAM parameters. This

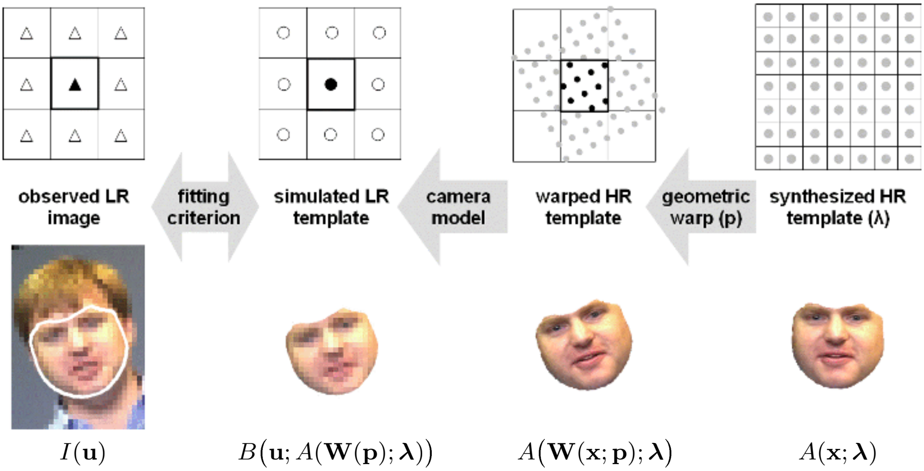

observed LR image
fitting criterion
simulated LR template
camera model
warped HR template
geometric warp (p)
synthesized HR template (λ)
$I(\mathbf{u})$
$B\big(\mathbf{u}; A(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big)$
$A\big(\mathbf{W}(\mathbf{x}; \mathbf{p}); \boldsymbol{\lambda}\big)$
$A(\mathbf{x}; \boldsymbol{\lambda})$

Fig. 2. The Resolution-Aware Fitting (RAF) algorithm simulates the formation of low resolution images in a digital camera. In contrast to the traditional formulation (Fig. 1), the fitting criterion is defined between observed and simulated image pixels.

formulation can accommodate arbitrary camera models and point spread functions. In this paper, we use the rectangular PSF

$$B\big(\mathbf{u}; A(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(\mathbf{u})} \int\limits_{\mathbf{u}' \in bin(\mathbf{u})} A\big(\mathbf{W}^{-1}(\mathbf{u}'; \mathbf{p}); \boldsymbol{\lambda}\big) d\mathbf{u}',$$

where the continuous integral is defined over $bin(\mathbf{u})$, the sensing area of the discrete pixel $\mathbf{u}$. As illustrated in Fig. 2, the blur operator itself is independent of AAM parameters. It simply averages out those template pixel intensities which map into a low resolution pixel's sensing area under the current warp $\mathbf{p}$. To express the integral above in the shape-normalized coordinate frame $\mathbf{s}_0$, we observe that $\mathbf{u}' = \mathbf{W}(\mathbf{x}; \mathbf{p})$, and consequently, $d\mathbf{u}' = \big|J\big(\mathbf{W}(\mathbf{p})\big)\big| d\mathbf{x}$,

$$B\big(\mathbf{u}; A(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(\mathbf{u})} \int\limits_{\substack{\mathbf{x} \in s_0 \text{ s.t.} \\ \mathbf{W}(\mathbf{x}; \mathbf{p}) \in bin(\mathbf{u})}} A(\mathbf{x}; \boldsymbol{\lambda}) \big|J\big(\mathbf{W}(\mathbf{p})\big)\big| d\mathbf{x}.$$

In practice, we implement this integration as a discrete, Jacobian-weigthed sum over template pixels,

$$B\big(\mathbf{u}; A(\mathbf{W}(\mathbf{p}); \boldsymbol{\lambda})\big) = \frac{1}{area(\mathbf{u})} \sum_{\substack{\mathbf{x} \in s_0 \text{ s.t.} \\ \mathbf{u} - \left[\begin{smallmatrix} .5 \\ .5 \end{smallmatrix}\right] < \mathbf{W}(\mathbf{x}; \mathbf{p}) < \mathbf{u} + \left[\begin{smallmatrix} .5 \\ .5 \end{smallmatrix}\right]}} A(\mathbf{x}; \boldsymbol{\lambda}) \big|J\big(\mathbf{W}(\mathbf{p})\big)\big|. \qquad (6)$$

Observe that our formulation avoids interpolating low resolution data, and models the object appearance, geometric deformation, and the image formation processes simultaneously.

3.2 RAF Algorithm

We now present a Gauss-Newton gradient-descent scheme for the minimization of the fitting criterion (5) with respect to $\mathbf{p}$ and $\boldsymbol{\lambda}$. Until convergence, updates $\Delta\mathbf{p}$ and $\Delta\boldsymbol{\lambda}$ will be iteratively computed and added to the current estimates. The derivation below closely follows that of the *simultaneous* algorithm in [8]. Expressing A as a sum of the mean and linearly weighted basis images, the fitting criterion is

$$\sum_{\mathbf{u} \in I} \left[I(\mathbf{u}) - B\Big(\mathbf{u}; A_0\big(\mathbf{W}(\mathbf{p})\big) + \sum_{i=1}^{m} \lambda_i A_i\big(\mathbf{W}(\mathbf{p})\big)\Big) \right]^2 .$$

Consider the Taylor expansion

$$\sum_{\mathbf{u} \in I} \left[I(\mathbf{u}) - B\Big(\mathbf{u}; A_0\big(\mathbf{W}(\mathbf{p} + \Delta\mathbf{p})\big) + \sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i) A_i\big(\mathbf{W}(\mathbf{p} + \Delta\mathbf{p})\big)\Big) \right]^2 .$$

Ignoring its second-order terms, the fitting criterion is approximately

$$\sum_{\mathbf{u}\in I}\left[I(\mathbf{u})-B\left(\mathbf{u};A_0(\mathbf{W}(\mathbf{p}))+\nabla A_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\Delta\mathbf{p}+\sum_{i=1}^{m}(\lambda_i+\Delta\lambda_i)\left(A_i(\mathbf{W}(\mathbf{p}))+\nabla A_i\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\Delta\mathbf{p}\right)\right)\right]^2.$$

For notational conciseness, denote $n+m$ steepest-descent images as

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

We can now compactly rewrite the fitting criterion as

$$\sum_{\mathbf{u}\in I}\left[I(\mathbf{u})-B\left(\mathbf{u};A_0(\mathbf{W}(\mathbf{p}))+\sum_{i=1}^{m}\lambda_i A_i(\mathbf{W}(\mathbf{p}))-\mathbf{SD}_{sim}\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}\right)\right]^2.$$

Observing that B is a linear operator, the objective function to be minimized is

$$\sum_{\mathbf{u}\in I}\left[I(\mathbf{u})-B\Big(\mathbf{u};A_0(\mathbf{W}(\mathbf{p}))\Big)+\sum_{i=1}^{m}\lambda_i B\Big(\mathbf{u};A_i(\mathbf{W}(\mathbf{p}))\Big)-B\big(\mathbf{u};\mathbf{SD}_{sim}\big)\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}\right]^2,$$

whose minimum is given by

$$\begin{pmatrix}\Delta\mathbf{p}\\ \Delta\lambda\end{pmatrix}=-H_{sim}^{-1}\sum_{\mathbf{u}\in I}B\big(\mathbf{u};\mathbf{SD}_{sim}^{\mathrm{T}}\big)\left[I(\mathbf{u})-B\Big(\mathbf{u};A_0(\mathbf{W}(\mathbf{p}))\Big)+\sum_{i=1}^{m}\lambda_i B\Big(\mathbf{u};A_i(\mathbf{W}(\mathbf{p}))\Big)\right],$$

where H_{sim} is the Hessian with appearance variation:

$$H_{sim}=\sum_{\mathbf{u}\in I}B\big(\mathbf{u};\mathbf{SD}_{sim}^{\mathrm{T}}\big)B\big(\mathbf{u};\mathbf{SD}_{sim}\big).$$

4 Quantifying the Benefits of RAF

We compared the RAF formulation (5) to the traditional formulation in (4). In particular, we compared the algorithm detailed in Section 3.2 with the simultaneous, inverse-compositional algorithm described in [11], which we refer to as AAMR-SIM. This represents a fair ground for comparison, since Matthews & Baker [10] "project out" the appearance variation. We artificially downscaled a variety of input test sequences by a range of scaling factors, and measured each algorithm's accuracy at lower input resolutions.

Independently of the resolution of a given test sequence, we initialized all algorithms with fitting results at the highest resolution. This allowed us to discard initialization quality as a confounding factor when comparing performances across resolution levels. While manual initialization is reasonable at higher resolutions, it becomes increasingly sub-optimal in lower resolutions, jeopardizing the fairness of comparisons across scales. Once in tracking mode, the fitting of each frame was initialized with the parameters of the preceding frame.

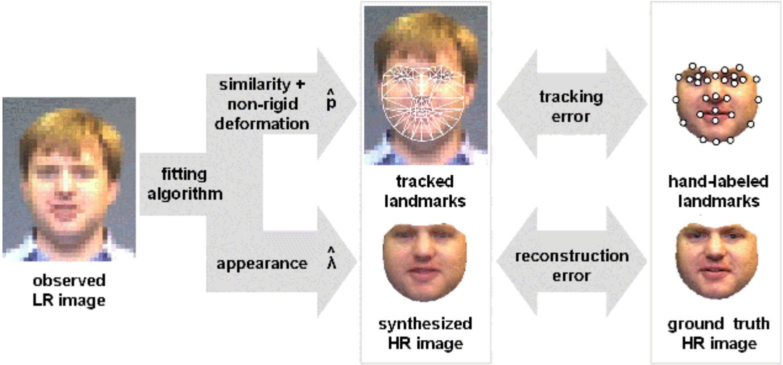


Fig. 3. We define two metrics to compare the fitting accuracy of algorithms. The average landmark tracking error combines the estimation accuracy of the similarity and non-rigid shape parameters. The reconstruction error quantifies how well the underlying high-resolution face could be inferred based only on low resolution data.

4.1 Metrics of Fit Quality

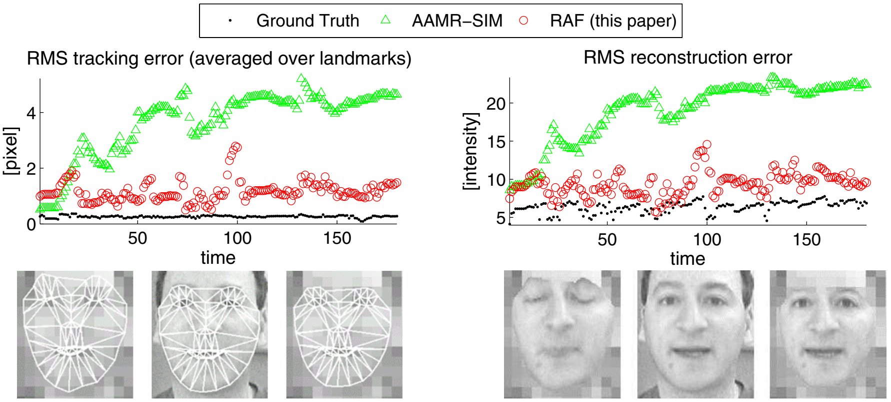
The most appropriate metric of an AAM's fit quality depends on the application at hand. For example, in an object tracking scenario, only the global pose (*i.e.*, the similarity transform parameters) may be of interest. For lip-reading, non-rigid deformations of a speaker's lips, encoded by a facial AAM's shape coefficients, may carry all the information. If the application requires synthesizing realistic face images, accurate appearance parameter estimates may be of importance.

In the lack of a specific application, we defined two metrics, illustrated in Fig. 3, to compare the fitting accuracy of the RAF and AAMR-SIM algorithms. The *tracking error* is the position error of landmarks (such as the corner of nostrils), averaged over the face: this is a combined effect of both similarity transform (scale, rotation, and translation) and non-rigid deformation parameters, as encoded by the estimate $\hat{\mathbf{p}}$. The *reconstruction error*, on the other hand, is computed by comparing the synthesized model instance, parametrized by $\hat{\boldsymbol{\lambda}}$, against the ground truth image. In addition, we report estimation errors for the coefficients of the top four principal shape and appearance modes.

For all test sequences included in this paper, only the landmark coordinates were available as hand-labeled, ground truth data. To infer the ground truth values for the similarity, non-rigid shape and appearance variables, we ran the AAMR-SIM tracker at the original resolution of the videos, and verified its convergence (each landmark's tracking error smaller than 1 high-resolution pixel). The resulting parameter estimates were then regarded as "ground truth" values.

4.2 Examples

Before presenting extensive quantitative results, some examples of our error metrics and their temporal behavior would be in order. In reporting Euclidian



AAMR-SIM Ground Truth RAF AAMR-SIM Ground Truth RAF

Fig. 4. The landmark tracking (upper left) and reconstruction (upper right) error metrics are plotted as a function of time for a 10-fold resolution degraded tracking experiment. Included images (bottom, captured at frame no. 102) display the mesh fits as well as synthesized model images (lower right). We overlay the latter onto pixel-replicated low resolution inputs (lower left) to demonstrate how well the underlying high-resolution image could be inferred.

distance metrics (as in translation parameters or landmark tracking error), we scale-normalize the estimates so that their numerical values are in high-resolution pixel units. Similarly, we normalize each shape and appearance coefficient according to its mode's variance, and report them in units of their standard deviation.

Fig. 4 plots error trajectories of a low resolution tracking experiment, where the subject's speaking and eye blinking were the major sources of motion. The input sequence was 10 times lower in resolution than the AAM. The error metrics indicate that RAF tracked the face consistently better than AAMR-SIM. To provide further evidence, Fig. 5 shows temporal trajectories of selected variables. Those estimated by AAMR-SIM do not follow the ground truth values, and remain mostly constant. In contrast, RAF can track the non-rigid deformations and appearance changes, amounting to a more accurate recovery of the facial expressions. We included this experiment and others in the supplemental video[1].

4.3 Test Set Statistics

It would be impractical to include time trajectories for all our experiments. In the following, we simply include the temporal mean and standard deviation of the Root Mean Squared (RMS) errors of selected variables. Note that lost trackers can easily corrupt these statistics with outliers. To prevent this, we required both trackers to produce valid results (*i.e.*, not have lost track of the face) for a fitting instance to be included in the comparison. This was achieved by visually inspecting all experiments and verifying that faces were tracked reasonably well.

[1] Demonstrations available at http://www.cs.cmu.edu/∼dedeoglu/eccv06

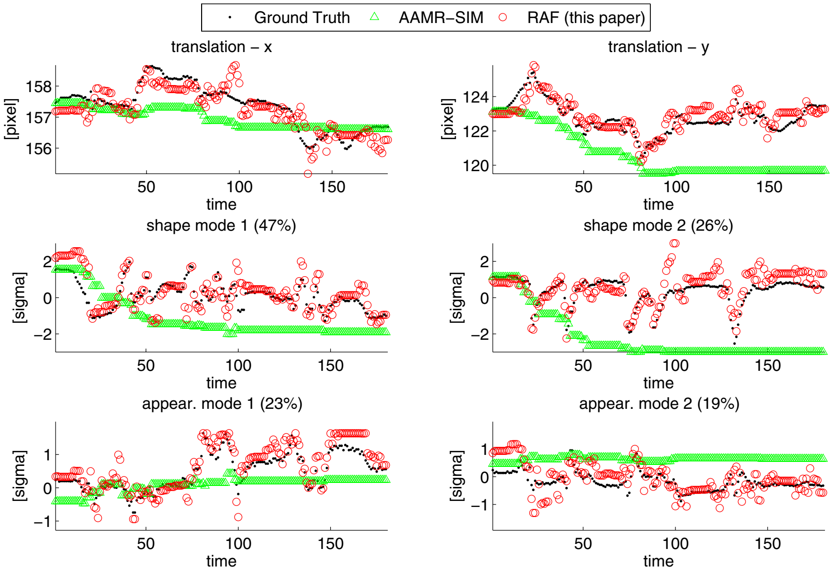


Fig. 5. Selected temporal trajectories are shown for a 10-fold resolution degraded face tracking experiment. As the supplemental video material shows, the main source of motion were the subject's speaking and eye blinking. See Fig. 4 for one example frame of this sequence. The estimates of AAMR-SIM do not follow the ground truth, and remain mostly constant. In contrast, RAF remains close to ground truth in all trajectories, indicating that it is able to extract the underlying facial expressions correctly.

Recall that each tracking experiment was initialized with the highest resolution fitting results. At lower input resolutions, such an optimistic initialization would cause the fitting performance to be overestimated at the beginning. To avoid this effect, we discarded the results of the first 20 frames of each sequence.

Fig. 6 compares the AAMR-SIM and RAF algorithms for fitting a single-person AAM. In the upper-left corner, we first provide a brief summary of experimental conditions. This AAM was built using 31 training images, and was tested on a set of 180. These were 8-bit grayscale images, and the AAM's native resolution was `100x104` pixels. We retained 95% of the total variation, yielding 11 shape and 23 appearance principal components.

The plots in Fig. 6 present extensive quantitative comparisons between the fitting algorithms. They are organized to show RMS error metrics as a function of downscaling factor. Observe how AAMR-SIM and RAF perform equally well at downsampling factor 2. This case corresponds to a minor degradation in resolution, but the fact that both algorithms perform similarly confirms the correctness of our derivations as well as implementations. Starting from downsampling factor 4, RAF brings substantial accuracy improvements across all metrics and variables of interest.

The performance of a model-based method ultimately depends on the quality of the available model. In order to investigate how the AAM fitting accuracy

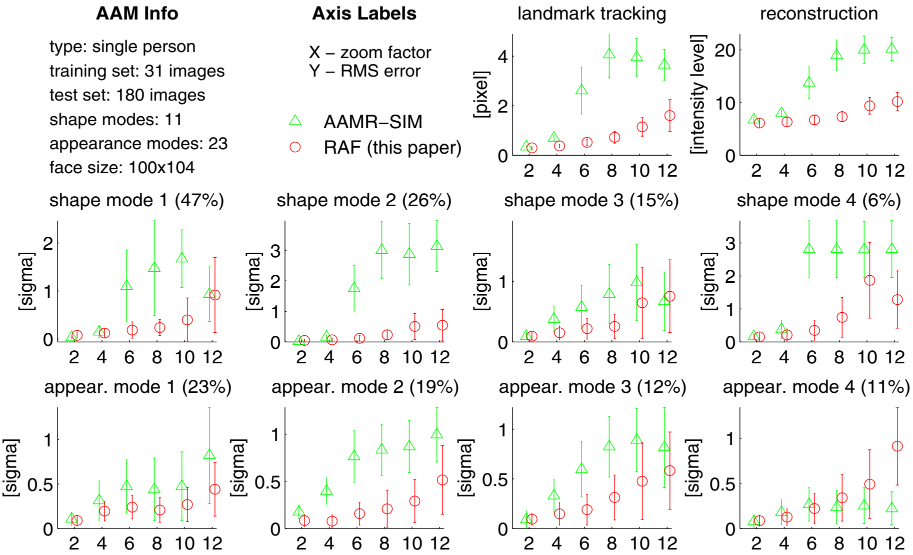


Fig. 6. Quantitative comparison between the AAMR-SIM and RAF algorithms for fitting the single-person AAM to a 180 frame-long sequence. Both algorithms perform well at half-resolution, validating the derivation and implementation of RAF. The latter brings substantial improvements across all metrics for downscaling factors 4 and higher. The principal modes are displayed in order of % energy (*i.e.*, variation) they capture.

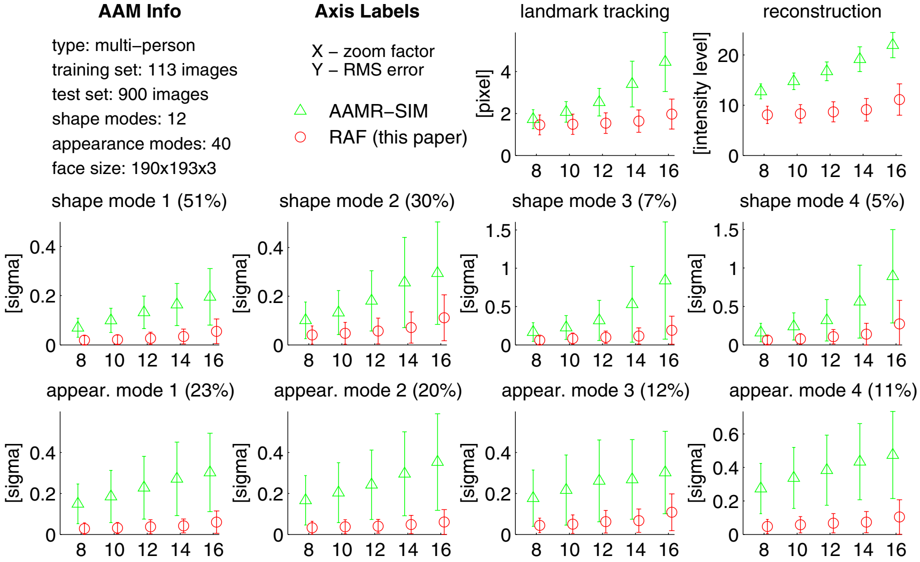


Fig. 7. Quantitative comparison between the AAMR-SIM and RAF algorithms for fitting the multi-person (5 subjects) AAM. Each reported mean and standard deviation is calculated over 900 frames, comprising 180 frames for each of 5 subjects. RAF improves the tracking, reconstruction, non-rigid shape, and appearance estimates considerably.

varies with model complexity, we also ran our experiments on a multi-person AAM, which we built using data from 5 subjects. Details of this AAM are provided in the upper-left corner of Fig. 7, organized in the same fashion as Fig. 6. The multi-person appearance model has almost twice the number of principal modes compared to the single-person case, indicating a richer sub-space being modeled. Again, RAF is observed to be consistently superior to AAMR-SIM in accuracy with regard to both tracking and reconstruction.

5 Qualitative Results

As a complementary method of comparison between the AAMR-SIM and RAF algorithms, we include a selection of synthesized model instances. For this end, we first pixel-replicated the original low resolution inputs, and then overlaid high-resolution reconstructions where the trackers thought the faces were. Many such reconstructions are included in the supplemental video.

Fig. 8 shows every second frame of a subsequence of the single-person AAM tracking experiment. Observe that RAF correctly extracts the eye blink and mouth opening, whereas AAMR-SIM does not. Fig. 9 offers a visual alternative for assessing how the trackers degrade with increased downscaling: it displays the single-person AAM results for frame no. 102 across various scales. While RAF can consistently recover the open eyes and mouth, AAMR-SIM's estimates degrade quickly: starting from downsampling factor 6, the eyes and mouth are first estimated to be half-open, and then totally closed. Similarly, Fig. 10 displays

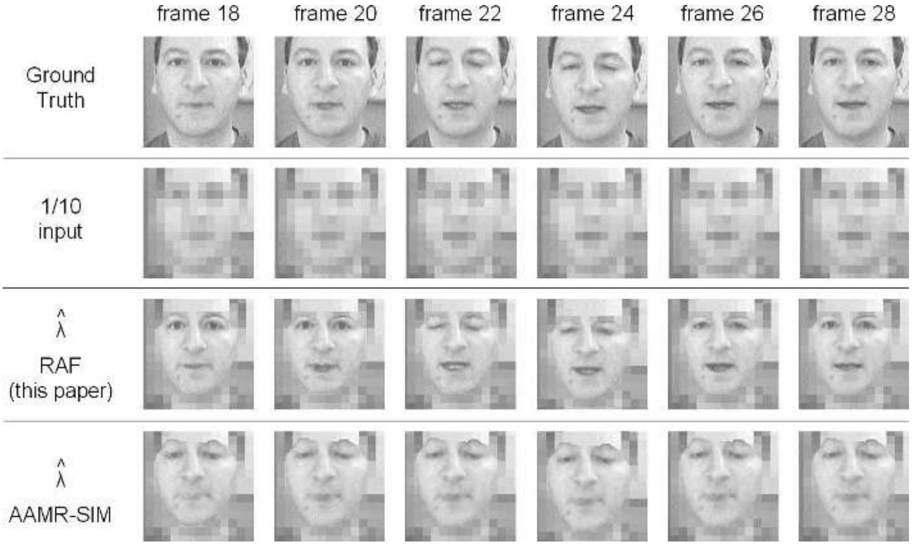


Fig. 8. Exemplar subsequence of high-resolution reconstructions, obtained by fitting the single-person AAM. Observe how RAF correctly extracts the eye blink and mouth opening, whereas AAMR-SIM does not. See complete videos at http://www.cs.cmu.edu/~dedeoglu/eccv06

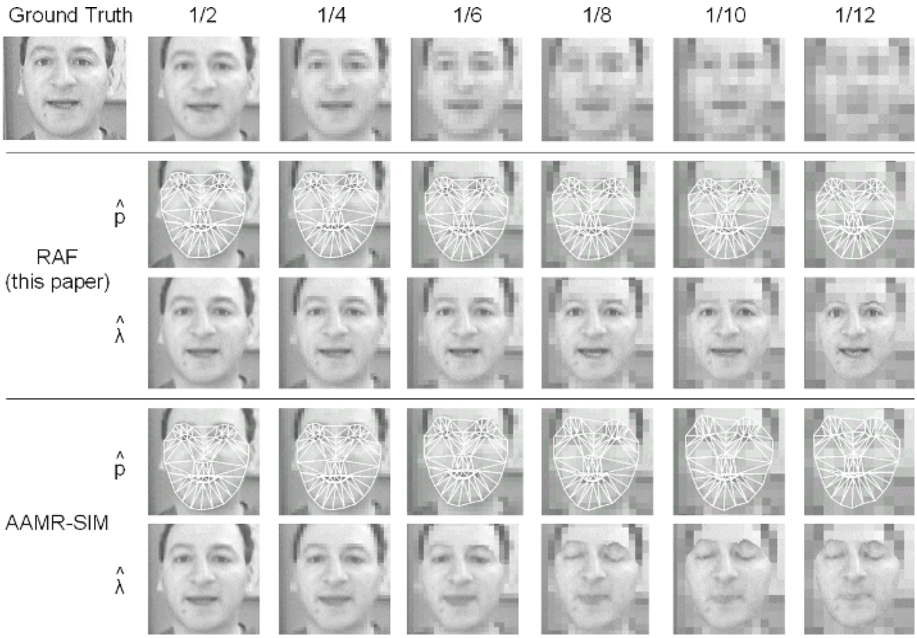


Fig. 9. We compared the AAMR-SIM and RAF algorithms over a range of scales. Increasingly lower resolution versions of input frame no. 102 are shown in the top row. While AAMR-SIM degrades quickly, RAF maintains a reasonable estimate of the face.

snapshots of different test subjects, all tracked using the multi-person AAM. For both AAMs, we find the visual reconstruction quality of RAF to be consistently superior to that of AAMR-SIM.

6 Discussion and Conclusions

In low resolution scenarios, there is significant scaling between the AAM and input images. In such cases, traditional fitting algorithms [5, 10] interpolate the observations when computing the fitting criterion. The essential novelty of our formulation is that it employs a camera model which mimics the image formation in digital cameras, and thereby avoids interpolation.

Throughout this paper, we focused on *accuracy* measures. Other factors such as robustness and computational efficiency may be as important. Indeed, in extremely low resolutions, we found the AAMR-SIM algorithm to be more robust than RAF. Given the smoothing effect of (bilinear) interpolation, this does not seem surprising. While RAF struggles among the many parameter settings which yield almost the same low resolution images, AAMR-SIM commits to an interpolated high-resolution observation, and pursues the fit.

We only fit nominal-resolution AAMs, independently of how much lower in resolution the observations were. This allowed us to reconstruct faces in

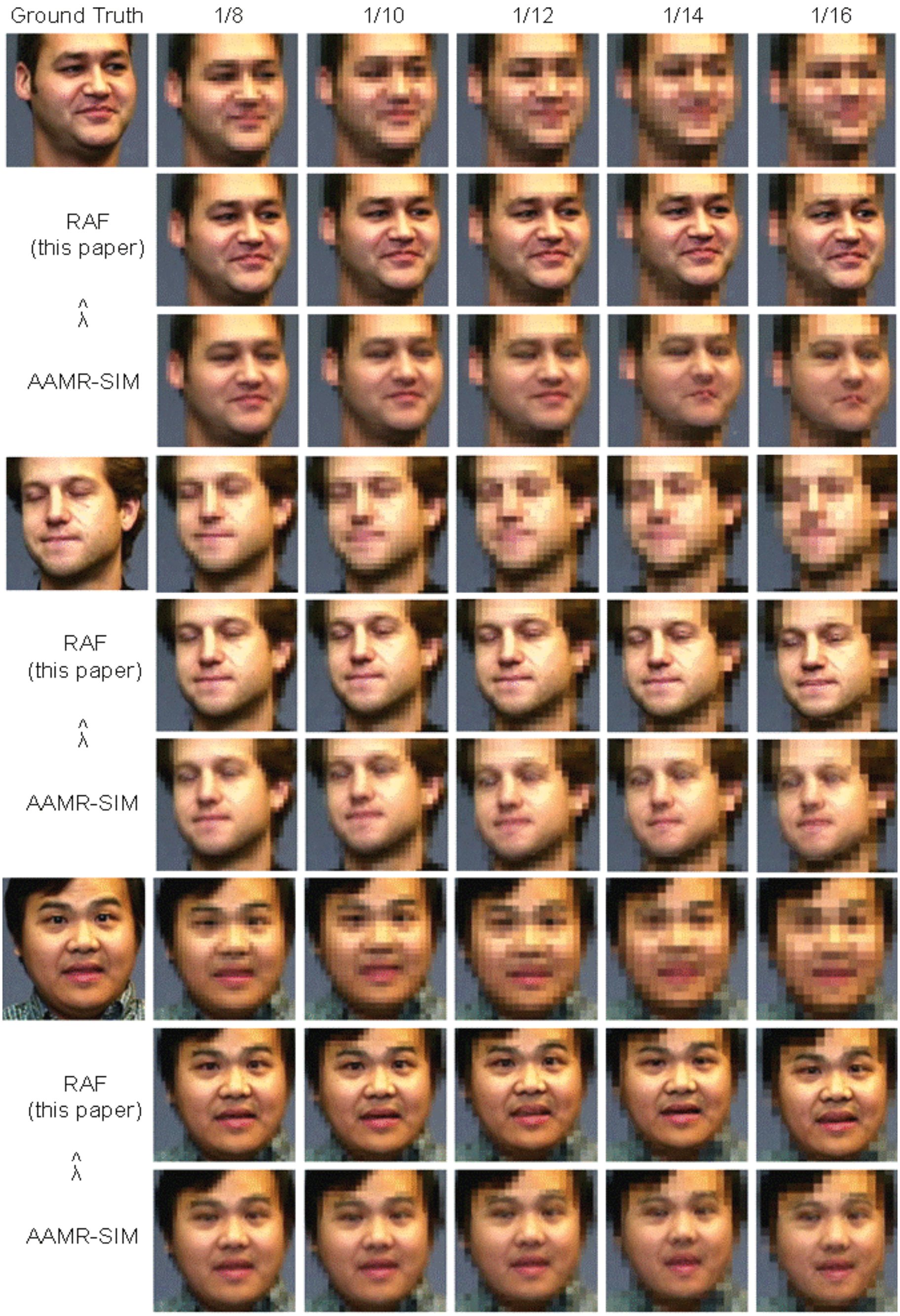


Fig. 10. Selected test frames are shown to visually compare the algorithms for fitting the multi-person AAM. The quantitative improvement in appearance estimates (Fig. 7) has visible effects. Mesh displays are omitted due to a lack of significant difference.

high-resolution. A related idea is to construct a scale-space pyramid of AAMs, and to model multiple resolutions in parallel. Due to blur, higher-level (*i.e.*, lower-resolution) AAMs would have more compact appearance models, and would therefore be easier to fit. Though this may seem to be an alternative to our approach, comparison *across models* is outside the scope of this paper. In comparing between fitting formulations across a range of resolution degradations, we used exactly the same AAMs. Our goal was to make a given fitting problem *more accurate*, rather than *finding an easier* fitting problem.

The fact that the summation in RAF's criterion is defined over observed image pixels has important consequences. Recall that the traditional fitting formulation had conveniently defined the summation over the model template pixels. Since the latter do not change as a function of the input, computational savings become possible: For instance, Matthews and Baker's [10] tracker considers the Taylor expansion for the warp parameters over the template, and pre-compute all associated Jacobians and Hessians. One area for future work is to incorporate such savings into the RAF formulation.

Our discussion remains orthogonal to practical search heuristics such as multiresolution, hierarchical and progressive [3, 4] methods. We can still exploit the advantages of these: for instance, a pyramid style algorithm would increase the robustness of RAF, complementing its accuracy at the bottom level.

In a more compherensive report [12], we argue that image-based warp estimation is an asymmetric problem: in the presence of relative scaling, the warp direction ought to be chosen such that the higher resolution image gets preblurred and warped onto the lower resolution one. As such, the AAM-based face tracking presented in this paper is an application of this general principle.

Acknowledgments. We are grateful to Iain Matthews for kindly providing the AAM Toolbox and sharing his expertise at various stages of implementation. We would also like to thank Jonas August and other members of the CMU misc-reading group for discussions. The research described in this paper was supported in part by U.S. Department of Defense contract N41756-03-C4024.

References

1. D.F. Barbe: Charge-Coupled Devices. Springer-Verlag, 1980.
2. B.D. Lucas and T. Kanade: An Iterative Image Registration Technique with an Application to Stereo Vision. *Proc. of the 7th Int. Joint Conference on Artificial Intelligence*, April, 1981, pp. 674-679.
3. P. Anandan: A Computational Framework and an Algorithm for the Measurement of Visual Motion. *International Journal of Computer Vision*, Vol. 2, No. 3, Jan., 1989, pp. 283-310.
4. J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani: Hierarchical Model-Based Motion Estimation. *Proc. of the European Conference on Computer Vision*, May, 1992, pp. 237-252.
5. T.F. Cootes, G.J. Edwards, and C.J. Taylor: Active Appearance Models. *Proc. of the European Conference on Computer Vision*, Vol. 2, 1998, pp. 484-498.

6. G.J. Edwards, C.J. Taylor, and T.F. Cootes: Interpreting Face Images Using Active Appearance Models. *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, June, 1998, pp. 300-305.
7. T.F. Cootes, G.J. Edwards, and C.J. Taylor: Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, June, 2001, pp. 681-685.
8. S. Baker, R. Gross, and I. Matthews: Lucas-Kanade 20 Years On: A Unifying Framework: Part 3. *Robotics Institute Technical Report CMU-RI-TR-03-35*, Carnegie Mellon University, November 2003.
9. S. Baker and I. Matthews: Lucas-Kanade 20 Years On: A Unifying Framework. *Int. Journal of Computer Vision*, Vol. 56, No. 3, March, 2004, pp. 221-255.
10. I. Matthews and S. Baker: Active Appearance Models Revisited. *International Journal of Computer Vision*, Vol. 60, No. 2, November, 2004, pp. 135-164.
11. R. Gross, I. Matthews, and S. Baker: Generic vs. Person Specific Active Appearance Models. *Image and Vision Computing*, Vol. 23, No. 11, Nov. 2005, pp. 1080-1093.
12. G. Dedeoglu, T. Kanade, and S. Baker: The Asymmetry of Image Registration and its Application to Face Tracking. *Robotics Institute Technical Report CMU-RI-TR-06-06, Carnegie Mellon University*, February, 2006.