Generic vs. Person Specific Active Appearance Models

Ralph Gross, Iain Matthews, and Simon Baker

The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{rgross, iainm, simonb}@cs.cmu.edu

Abstract

Active Appearance Models (AAMs) are generative parametric models that have been successfully used in the past to model faces. Anecdotal evidence, however, suggests that the performance of an AAM built to model the variation in appearance of a single person across pose, illumination, and expression (Person Specific AAM) is substantially better than the performance of an AAM built to model the variation in appearance of many faces, including unseen subjects not in the training set (Generic AAM). In this paper we present an empirical evaluation that shows that Person Specific AAMs are, as expected, both easier to build and more robust to fit than Generic AAMs. Moreover, we show that: (1) building a generic shape model is far easier than building a generic appearance model, and (2) the shape component is the main cause of the reduced fitting robustness of Generic AAMs. We then proceed to describe two refinements to Generic AAMs to improve their performance: (1) a refitting procedure to improve the quality of the ground-truth data used to build the AAM and (2) a new fitting algorithm. For both refinements we demonstrate vastly improved fitting performance.

1 Introduction

Active Appearance Models (AAMs) [3] are generative parametric models commonly used to model faces. Depending on the task at hand AAMs can be constructed in different ways. For example, we might build an AAM to model the variation in appearance of a single person across pose, illumination and expression. Such a *Person Specific AAM* might be useful for interactive user interface applications that involve head pose estimation, gaze estimation, or expression recognition. Alternatively, we might attempt to build an AAM to model any face, including unseen subjects not in the training set. The most common use of such a *Generic AAM* would be face recognition.

Anecdotal evidence suggests that Person Specific AAMs perform substantially better than Generic AAMs. The performance of an AAM depends on two steps: (1) Modelling: How well is the AAM able to model (or generate) images in the class under consideration and (2) Fitting: How robustly can the AAM be fit to a novel input image? AAMs consist of two components: (1) a shape component, and (2) an appearance component. In the first part of this paper (Section 3) we present an empirical evaluation that shows that

BMVC 2004 doi:10.5244/C.18.48

Person Specific AAMs are indeed both easier to build and far more robust to fit than Generic AAMs. Besides validating the anecdotal evidence, we also attempt to determine the reason for the inferior performance of Generic AAMs. In particular, we attempt to answer the following questions: Is it harder to build a generic shape model that models the shape of any face well, or is building a generic appearance model harder? What makes fitting harder, a large generic shape model, or a large generic appearance model?

In the second part of this paper (Section 4) we proceed to describe two refinements to Generic AAMs to improve their performance. We first propose a refitting procedure to improve the quality of the ground-truth data used to build the AAM (Section 4.1). We then introduce a new fitting algorithm (Section 4.2). For both refinements we demonstrate vastly improved fitting performance.

2 Background: Active Appearance Models

We begin with a brief review of Active Appearance Models [3, 9]. We explain how AAMs are constructed from training data and describe an algorithm for fitting AAMs to an image.

2.1 Model Construction

The *2D shape* of an AAM is defined by a 2D triangulated mesh and in particular the vertex locations of the mesh. Mathematically, we define the shape $\mathbf{s}$ of an AAM as the 2D coordinates of the n vertices that make up the mesh: $\mathbf{s} = (x_1, y_1, x_2, y_2, \ldots, x_n, y_n)^{\mathrm{T}}$. AAMs allow linear shape variation. This means that the shape matrix $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape matrices $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i \tag{1}$$

where the coefficients p_i are the shape parameters. AAMs are normally computed from training data consisting of a set of images with the shape mesh (usually hand) marked on them [3]. The training shapes are then geometrically aligned using the *Procrustes* algorithm [5]. Principal Component Analysis (PCA) [7] is then applied to the aligned training meshes. The base shape $\mathbf{s}_0$ is the mean shape and the matrices $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues.

The *appearance* of the AAM is defined within the base mesh $\mathbf{s}_0$. Let $\mathbf{s}_0$ also denote the set of pixels $\mathbf{u} = (u, v)^{\mathrm{T}}$ that lie inside the base mesh $\mathbf{s}_0$, a convenient abuse of terminology. The appearance of the AAM is then an image $A(\mathbf{u})$ defined over the pixels $\mathbf{u} \in \mathbf{s}_0$. AAMs allow linear appearance variation. This means that the appearance $A(\mathbf{u})$ can be expressed as a base appearance $A_0(\mathbf{u})$ plus a linear combination of l appearance images $A_i(\mathbf{u})$:

$$A(\mathbf{u}) = A_0(\mathbf{u}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{u}) \tag{2}$$

where the coefficients λ_i are the appearance parameters. The appearance images A_i are usually computed by applying PCA to the shape normalized training images [3, 9].

2.2 Model Fitting

Fitting an AAM may be formulated as minimizing the sum of squares difference between the appearance $A(\mathbf{u}) = A_0(\mathbf{u}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{u})$ and the input image warped back onto the

base mesh $I(\mathbf{N}(\mathbf{W}(\mathbf{u};\mathbf{p});\mathbf{q}))$ [9]:

$$\sum_{\mathbf{u}\in\mathbf{s}_0}\left[A_0(\mathbf{u})+\sum_{i=1}^{l}\lambda_i A_i(\mathbf{u})-I(\mathbf{N}(\mathbf{W}(\mathbf{u};\mathbf{p});\mathbf{q}))\right]^2 \tag{3}$$

In this equation, the warp $\mathbf{W}$ is the piecewise affine warp defined by the mesh triangulation from the base mesh $\mathbf{s}_0$ to the current AAM shape $\mathbf{s}$ and $\mathbf{N}$ is a 2D similarity transformation used to normalize the shape of the AAM. The goal of AAM fitting is to minimize the expression in Equation (3) simultaneously with respect to the appearance parameters λ, the linear shape parameters $\mathbf{p}$, and the similarity transform parameters $\mathbf{q}$.

One algorithm for fitting an AAM to an image is the "project-out" inverse compositional algorithm proposed in [9]. This algorithm performs the non-linear optimization of Equation (3) in two steps (similar to Hager and Belhumeur [6]). First the shape and linear transformation parameters $\mathbf{p}$ and $\mathbf{q}$ are found through a non-linear optimization in a subspace in which the appearance variation can be ignored. The second step is then a closed form linear optimization with respect to the appearance parameters λ. The algorithm is very fast, running at over 230 frames per second on standard hardware [9].

3 Generic vs. Person Specific AAMs

In this section we present an empirical comparison of Generic and Person Specific AAMs. To support these experiments we assembled datasets for the construction and fitting of Generic and Person Specific AAMs which separate shape and appearance variation. See Section 3.1. We use these datasets to evaluate the model construction performance (Section 3.2) and the model fitting performance (Section 3.3) of both types of AAMs.

3.1 Datasets

We assembled three datasets which separately vary illumination, pose and identity. For the illumination dataset we recorded a single subject in a static frontal pose and neutral expression while smoothly changing the position of a lamp illuminating the face. We randomly selected 100 images from this sequence for the experiments. This data is used to construct Person Specific AAMs. It contains a large amount of appearance variation, but little or no shape variation. In the second set the same subject was recorded under constant illumination and neutral expression while smoothly changing head pose. We randomly selected 100 images from the sequence for evaluation. Again, this data is used to build Person Specific AAMs. Unlike the illumination set, the pose set contains a lot of shape variation. Due to the non-uniform lighting and the presence of specularities, a small amount of appearance variation is visible as well. For the identity dataset, we chose 100 different subjects from the **fa** set of the FERET database [10]. This data is used to build Generic AAMs. This set contains both shape and appearance variation due to the variation across face identities. Even though the images were taken from the same set, differences in facial expression, face pose and illumination are also present. Figure 1 shows three example images from each dataset. To ground-truth the data the vertex locations of the shape mesh for all 300 images were marked by hand.

3.2 Model Construction

In order to quantify model construction performance we determine how well an AAM can model *unseen* data based on a training set of the same type. In the experiments we

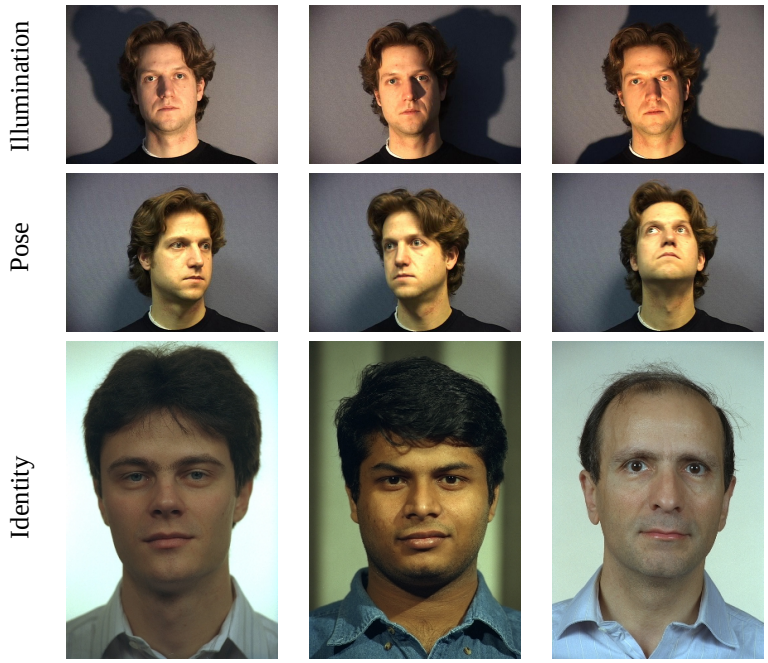


Figure 1: Datasets. Illumination set: The subject was recorded with constant frontal pose and neutral expression while smoothly changing the position of a lamp illuminating the face. Pose set: The same subject was recorded under constant illumination while smoothly changing head pose. Identity set: We selected 100 subjects from the **fa** set of the FERET database [10].

separately evaluate the shape and appearance components of the AAMs.

3.2.1 Experiment Description

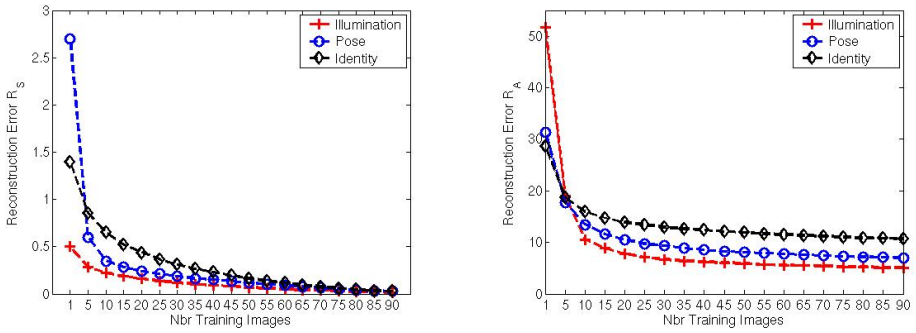
We randomly select a varying number of training images from the dataset to build shape and appearance models. For all models we retain enough variance to explain 95% of the training data. We then evaluate the reconstruction error of a fixed number of images from an *independent* test set. In order to calculate the reconstruction error for a test shape $\mathbf{s}$ and appearance A, we compute the shape parameters $p_1,\ldots,p_m$ and the appearance parameters $\lambda_1,\ldots,\lambda_l$ by projecting $\mathbf{s}$ and A into the shape and appearance eigenspaces. The reconstruction errors are then defined by:

$$R_S = \left\| \mathbf{s} - (\mathbf{s_0} + \sum_{i=1}^{m} p_i \mathbf{s_i}) \right\|_2 \qquad R_A = \left\| A - (A_0 + \sum_{i=1}^{l} \lambda_i A_i) \right\|_2 \tag{4}$$

where $\|\cdot\|_2$ is the Euclidean L2 Norm.

3.2.2 Experiment Results

Figure 2(a) plots the shape reconstruction errors for all three datasets against a varying number of training images. The illumination dataset theoretically has zero shape variation. Hence the reconstruction error for a single training image (0.5 pixels) can be attributed to errors in ground-truthing. We use this threshold to determine how many training images are needed for the pose and identity sets to model unseen data. The error



(a) Shape (b) Appearance

Figure 2: Shape and appearance reconstruction errors for the illumination, pose and identity datasets. We compute the reconstruction error by projecting independent test data into the eigenspace spanned by training sets of varying size, reconstructing the data using the eigenspace representation and measuring the Euclidean distance between original and reconstructed data.

	Illumination (90)	Pose (90)	Identity (90)	Identity (190)
R_A	5.06	6.99	10.67	9.14

Table 1: Appearance reconstruction error R_A for the illumination, pose and identity datasets for 90 and 190 training images. Even for 190 training images in the identity set the reconstruction error stays well above the error for the illumination and pose sets.

over the pose set falls below 0.5 after 5 training images, which is consistent with intuition and recent theoretical results showing that at most 6 2D shape vectors are needed to model a single rigid 3D face [11]. For the identity set 15 training images are needed to reach this level of modelling accuracy. We can therefore conclude that (1) it is possible to build a generic shape model and (2) as few as 15 training images are needed for a generic shape model. However, this only applies to generic shape models for frontal faces. More training images will be neccessary to build a generic shape model for faces under varying poses.

Figure 2(b) plots the appearance reconstruction error. Again we can use the reconstruction error over the illumination set as guideline to determine when the models only explain noise due to errors in (shape) ground-truthing and appearance model interpolation. This holds since faces under fixed pose but varying illumination can be modelled using a low dimensional subspace [8]. The reconstruction error over the pose set is actually slightly higher than over the illumination set, possibly due to the more difficult ground-truthing and the non-uniform illumination. The reconstruction error over the identity set always stays well above the level of either the illumination or the pose set. This observation holds even if we expand the training set to 190 images. See Table 1 for numerical results. Overall we can conclude that it is much more difficult to build a generic appearance model.

3.2.3 Experiment Conclusions

We empirically showed that it is relatively easy to build a generic shape model for frontal faces. With as few as 15 training images the shape model is able to model unseen faces with sufficient accuracy. However, the same does not hold for a generic appearance model.

3.3 Model Fitting

In order to quantify model fitting performance we determine how well an AAM can be fit to an image using the "project-out" algorithm described in Section 2.2. We again separately evaluate the shape and appearance components of both Generic and Person Specific AAMs following a similar evaluation methodology to the one in [9].

3.3.1 Experiment Description

For a given AAM and test image we randomly perturb the ground truth shape and similarity transform parameters by a large, fixed magnitude to generate the inital parameter estimates for the fitting algorithm. We then record the *average frequency of convergence* by measuring how often the algorithm converges after 20 iterations [1] to within 2.0 pixels [2] of the ground truth (RMS mesh point error). In order to quantify the influence of the shape model on the fitting performance we run the algorithm using shape models of varying size with a constant appearance model computed over the complete dataset. The shape models are computed by randomly choosing a fixed number n of training shapes and varying n between 5 and 100 while retaining the same fixed amount of variance (95%). The influence of the appearance model is determined in a similar fashion by running the fitting algorithm using a static shape model computed over the whole dataset and an appearance model of varying size. To separate the effects of model construction and fitting, the fitting algorithm is tested on the training images. We therefore know that the AAM is able to model the input image it is being fit to. If it fails to fit, it is due to the difficulty of the fitting process rather than the inability of the AAM to model the image.[3]

3.3.2 Experiment Results

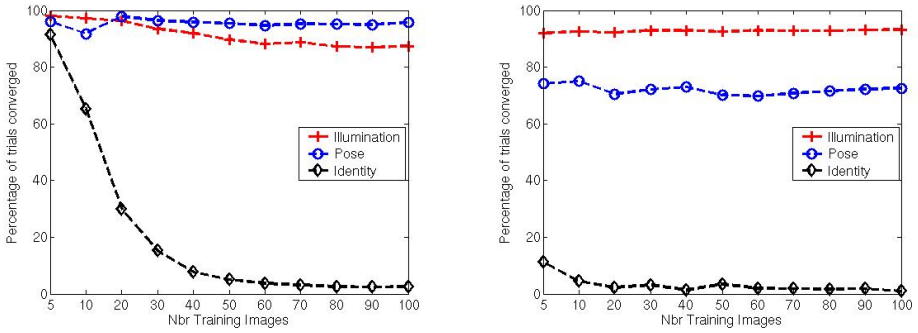
Figure 3 shows the results of the fitting experiments. It is immediately apparent that fitting the Generic AAM is significantly harder than fitting either of the Person Specific AAMs. Note that the relative performance is entirely due to the data. These results obtained using a theoretically sound algorithm should be interpreted as indicating that fitting the Generic AAM is an inherently far harder task than fitting the Person Specific AAM.

At this point, it is natural to ask what is the cause of this drastically different performance. Note the following results: (1) A Generic AAM built with 5 shape training images and 100 appearance training images operates about as well as similar Person Specific AAMs. See leftmost point in Figure 3(a). (2) A Generic AAM built with 100 shape training examples and 5 appearance training examples operates far, far worse than similar Person Specific AAMs. See leftmost point in Figure 3(b). From these results, we speculate that it is the shape component of the Generic AAM that is causing the problem. As further evidence of this argument, consider Figure 4 in which we plot the magnitudes

[1] In earlier experiments we found that the algorithm typically converges well within 20 iterations [9].

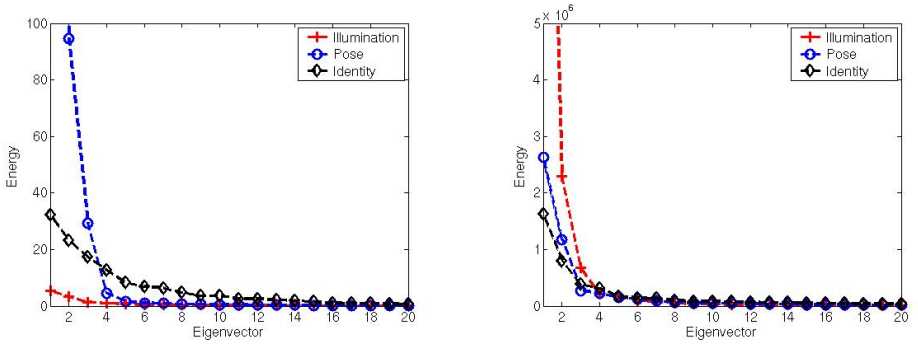
[2] There are no obvious or established choices for the selection of the convergence criterion. We simply choose a value and verified that it corresponds to "good" convergence. We plan to revisit this question in more detail in future work.

[3] In a future extended version of this paper we will investigate the combined difficulty of model construction and fitting by evaluating fitting performance on unseen images.



(a) Shape (b) Appearance

Figure 3: Average rate of convergence for the project-out algorithm [9]. (a) Results for a fixed appearance model computed over the whole dataset and a shape model computed using different numbers of training images. (b) Results for a fixed shape model and varying appearance model.



(a) Shape (b) Appearance

Figure 4: Energy distribution of AAM shape (a) and appearance (b) eigenvectors. Generic AAMs contain relatively more shape variation than Person Specific AAMs, but not much more appearance variation.

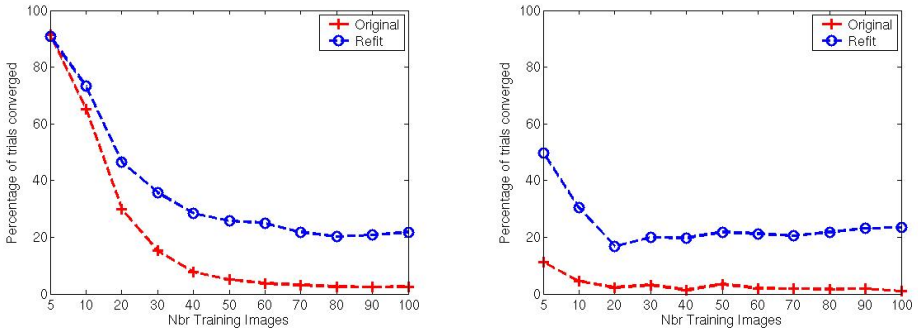
of the eigenvectors for the shape and appearance models each computed with all 100 training images. The appearance eigenvectors for the 3 databases are all very similar. If anything, there is less appearance variation in the Generic AAM than the Person Specific AAMs. On the other hand, the shape eigenvectors of the Generic AAM are substantially larger than those of the Person Specific AAMs (at least after the first 4). The "effective" dimensionality of the shape component of the Generic AAM is far higher than the dimensionality of the Person Specific shape models.

3.3.3 Experiment Conclusions

Although constructing a generic shape model is relatively easy, fitting a Generic AAM is far harder than fitting a Person Specific AAM because the effective dimensionality of the generic shape model is far higher than that of the Person Specific shape models.

4 Improvements to Generic AAMs

Perhaps the main reason for performing the evaluation in Section 3 was to suggest possible methods of improving the performance of Generic AAMs. In the remainder of this paper



(a) Shape (b) Appearance

Figure 5: Average rate of convergence for Generic AAMs computed using original and refitted labels. Results are shown for (a) AAMs with varying shape model sizes and fixed appearance model and (b) AAMs with varying appearance model sizes and fixed shape model.

we describe two such techniques: (1) refitting (Section 4.1) and (2) simultaneous fitting of shape and appearance (Section 4.2). These are by no means the only possibilities. We plan to cover several other possibilities in a future paper.

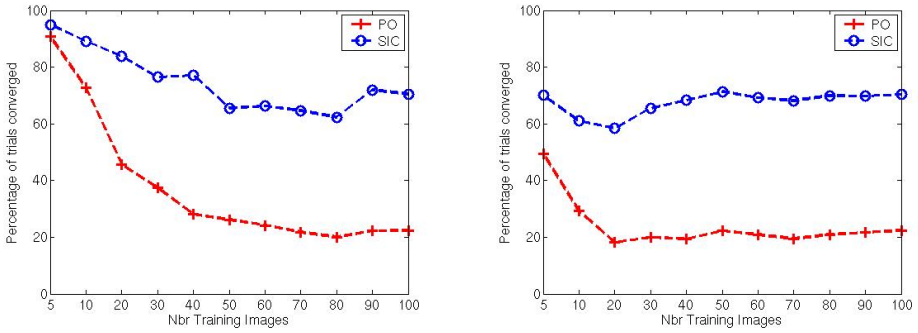
4.1 Data Refitting

The vertex locations of the shape mesh for all images (in total 16,800) used in the experiments were marked by hand. Due to the large amount of data involved and the difficulty of the task the quality of the labels is less than perfect. We use a two step algorithm to improve the ground truth data. First we construct a AAM with the original hand-marked labels. We then fit the AAM to the original training images and recover the vertex locations of the fitted shape mesh as new landmark data. We visually check that the fit is good for every image in the dataset, a process that may be automated using the model fitting error. We then compute new AAMs using only the refitted labels. Since we retain less than 100% of the variance in the shape and appearance model used to refit the data, outliers in the data are eliminated. Note that depending on the amount of variance retained refitting might remove signal along with the noise. Also, this procedure does not improve consistently misplaced labels.

The refitting procedure is evaluated by comparing the average rate of convergence of the project-out algorithm using Generic AAMs constructed from the original labels with the results obtained by fitting models constructed from the refitted labels. For the refitting procedure we used an AAM which retaines enough variance to explain 95% of both the shape and appearance training data. We follow the same evaluation methodology as used in Section 3.3. As shown in Figure 5 AAM fitting performance for the refitted labels is substantially better than fitting performance for the original labels. For AAMs with varying shape model sizes fitting performance improves on average from 20.8% of trials converged to 37.4% of trials converged (see Figure 5(a)). Similarly the fitting performance for AAMs with varying appearance model sizes improves on average from 3.1% of trials converged to 24.4% (see Figure 5(b)).

4.2 Simultaneous Fitting of Shape and Appearance

As stated in Section 2.2, the goal of AAM fitting is usually formulated as minimizing the sum of squares difference between the model instance and the input image warped back



(a) Shape (b) Appearance

Figure 6: Average rate of convergence for Generic AAMs using the simultaneous inverse compositional (SIC) and project-out (PO) algorithms.

onto the base mesh (see Equation (3)). Recently, Baker et al. [1] introduced the simultaneous inverse compositional algorithm which minimizes Equation (3) by performing Gauss-Newton gradient descent optimization simultaneously on the warp parameters **p**, the linear transformation parameters **q** and the appearance parameters λ [4]. In comparison to the project-out algorithm, the simultaneous algorithm is very slow (up to $30\times$ slower). See [1] for the details.

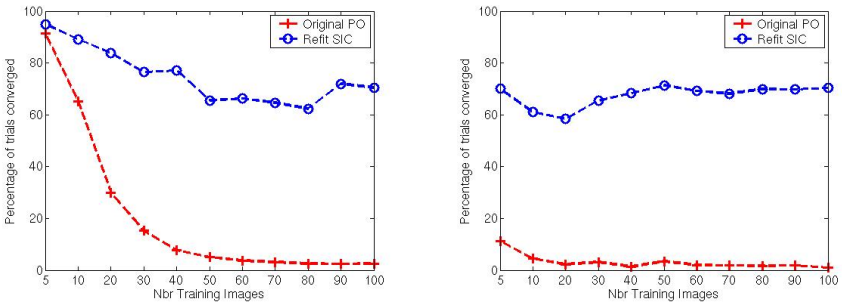
We compare the performance of the simultaneous inverse compositional and project-out algorithms by comparing the average rate of convergence for Generic AAMs for varying shape and appearance model sizes. Here we use AAMs constructed using refitted labels as described in Section 4.1. As shown in Figure 6, the simultaneous fitting algorithm performs significantly better than the project-out algorithm. For AAMs with varying shape model sizes the fitting performance improves on average from 37.4% of trials converged to 76.8% (Figure 6(a)). For AAMs with varying appearance model sizes the fitting performance improves on average from 24.4% to 67.5% (Figure 6(b)).

5 Discussion

In this paper we first empirically compared the performance of Generic and Person Specific Active Appearance Models (AAMs). In Section 3.2 we showed that building a generic shape model is comparatively easy, while building a generic appearance model is much harder. We then demonstrated in Section 3.3 that fitting a Generic AAM appears to be harder than fitting a Person Specific AAM due mainly to the higher effective dimensionality of the shape model. In Section 4 we then discussed two refinements to Generic AAMs: (1) label refitting and (2) the simultaneous inverse compositional fitting algorithm. The performance improvement due to these two refinements is summarized in Figure 7.

In Section 3.3 we showed that fitting an AAM with a complicated shape model is difficult. As a further refinement to address this problem we plan to incorporate shape priors into the fitting algorithm as suggested in [2, 4]. While the simultaneous fitting algorithm performs significantly better than the project-out algorithm (see Figure 6), it is also very

[4]The simultaneous inverse compositional algorithm is defined for *independent AAMs* [9], which separately parameterize shape and appearance . It is different from the original AAM fitting algorithm defined for *combined AAMs* which jointly parameterize shape and appearance [3].



(a) Shape (b) Appearance

Figure 7: Average rate of convergence for Generic AAMs using the simultaneous inverse compositional (SIC) algorithm on refitted labels and the conventional project-out (PO) algorithm on the original hand-marked labels.

slow [1]. We intend to combine the two algorithms by using the simultaneous algorithm on the first image of a sequence for a high quality fit, update the mean appearance image of the AAM with the extracted face appearance and continue to track the face with the efficient project-out algorithm. Preliminary results (omitted) show this technique to be promising.

6 Acknowledgments

The research described in this paper was supported by ONR contract N00014-00-1-0915 and in part by U.S. Department of Defense contract N41756-03-C4024.

References

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

[2] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 4. Technical Report CMU-RI-TR-04-14, Carnegie Mellon University Robotics Institute, 2004.

[3] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *PAMI*, 23(6):681–685, June 2001.

[4] T.F. Cootes and C.J. Taylor. Constrained active appearance models. In *Proc. ICCV*, pages 748–754, 2001.

[5] I.L. Dryden and K.V. Mardia. *Statistical Shape Analysis*. Wiley & Sons, 1998.

[6] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE PAMI*, 20(10), 1998.

[7] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE PAMI*, 12(1):103–108, 1990.

[8] K.-C. Lee, L. Ho, and D. Kriegman. Nine points of light: Acquiring subspaces for face recognition under variable lighting. In *IEEE CVPR*, pages 519–526, 2001.

[9] I. Matthews and S. Baker. Active Appearance Models revisited. *International Journal of Computer Vision*, 60(2), 2004.

[10] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[11] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. In *IEEE CVPR*, 2004.