Appeared in Proceedings of BMVC'07

Segmented AAMs Improve Person-Independent Face Fitting

Julien Peyras[1] Adrien Bartoli[2] Hugo Mercier[3] Patrice Dalle[3]

[1] Dipartimento di Scienze dell'Informazione, Milano, Italy
[2] LASMEA, Clermont-Ferrand, France
[3] IRIT, Toulouse, France

{Peyras,Mercier,Dalle}@irit.fr Adrien.Bartoli@gmail.com

Abstract

An Active Appearance Model (AAM) is a variable shape and appearance model built from annotated training images. It has been largely used to synthesize or fit face images. Person-independent face AAM fitting is a challenging open issue. For standard AAMs, fitting a face image for an individual which is not in the training set is often limited in accuracy, thereby restricting the range of application.

As a first contribution, we show that the limitation mainly comes from the inability of the AAM appearance counterpart to generalize, *i.e.* to accurately generate previously unseen visual data. As a second contribution, we propose an efficient person-independent face fitting framework based on what we call multi-level segmented AAMs. Each segment encodes a physically meaningful part of the face, such as an eye. A coarse-to-fine fitting strategy with a gradually increasing number of segments is used in order to ensure a large convergence basin.

Fitting accuracy is assessed by comparison with manual labelling statistics constructed from multiple data annotations. Experimental results support the claim that standard AAMs are well-adapted to person-specific fitting while segmented AAMs outperform the classical AAMs in a person-independent context in terms of accuracy, and ability to generate new faces.

1 Introduction

The Active Appearance Model (AAM) paradigm was introduced in 1998 by Cootes *et al.* [3] and since then it has had a great success. An AAM learns the shape and the appearance of a labelled set of images showing some class of objects. AAMs are widely used for face fitting, see *e.g.* [3, 8] and face synthesis, see *e.g.* [4]. Most of the applications – in the medical, psychological and linguistic fields, cognitive studies, expression transfer on an avatar, *etc.* – require highly accurate fitting. In other words, the AAM parameters must be recovered such that the synthesized image closely matches the input image.

Most of the previous work uses a single AAM modeling the face as a whole. Accurate fitting is achieved in a person-specific context. For instance, [8] uses AAMs for facial

deformation analysis. The standard AAM usually fails to achieve high accuracy for an image of a previously unseen face, *i.e.* for an individual not in the training set. Person-independent face fitting is however a very important problem since a training image set might not be available for an individual whose face needs to be accurately tracked in a video.

The closest work to ours is probably by Gross *et al.* [7]. They tackle the problem of constructing and fitting person-independent AAMs. They show that this is a difficult problem, even for frontal pose and neutral expression, and that the difficulties come from the inability of standard AAMs to generate new faces. A solution based on training, iteratively refitting the data with the AAM and re-training, is shown to improve the performances compared to traditional single step AAM training. Cristinacce *et al.* [5] recently proposed a paradigm called Constrained Local Model (CLM). It is shown to be effective at fitting a local face model based on measuring the image response around vertices and with a shape prior learnt from training images.

This paper tackles the important issue of person-independent face fitting with AAMs. We bring several statements and technical contributions:

- First, §3, we propose a means to assess fitting accuracy: the SSE (Statistical Shape Error). It is based on using several manual labellings of the input images by different users, from which gaussian statistics are computed for each label. The quality of an AAM fit is assessed by using the Mahalanobis distance with manual labelling statistics. This is an essential tool for the subsequent experimental analysis.

- Second, §4, we experimentally investigate the behavior of standard AAMs on unseen faces, and show that the lack of accuracy is mainly due to the inability of the appearance component to generate unseen faces. We state that standard AAMs are accurate in a person-specific context but not in a person-independent one.

- Third, §5, we show that segmented AAMs outperform standard ones in the person-independent context and achieve very accurate fitting, of the same order as the accuracy reached with manual labelling statistics. Segmented AAMs consist of several portions, each of which modeling a region of the face such as the mouth. Directly fitting each segment would reduce the convergence basin compared to fitting a standard AAM. As a remedy, we propose a coarse-to-fine fitting strategy which gradually splits a standard AAM into pre-defined segments. This *multi-level segmented AAM* we propose thus is able to generate new faces and can be effectively fit to images. Experimental results show that this outperforms the refitting solution of [7].

We give some background on AAMs below and our conclusions in §6.

2 Background on AAMs: Training and Fitting

An AAM combines two linear subspaces, one for the shape and one for the appearance, which are learnt from a previously labelled set of training images [3].

Principal Component Analysis (PCA) is applied on shape training data to retrieve a set of shape eigencomponents s_i expressing the shape model variation, and their associated eigenvalues proportional to the variance of the training data the s_i enclose. Four

extra components s_i^* are added to allow the 2D similarity transform, see [10]. Let $B_s = [s_1, \cdots, s_i, \cdots, s_1^*, \cdots, s_4^*]$ be the shape subspace basis. An instance of shape is defined as a linear expression: $s = B_s p_s$ with p_s the shape deformation parameters.

PCA is applied on the shape-corrected appearance data to retrieve a set of appearance eigencomponents A_i, allowing variations on the model appearance, and their associated eigenvalues proportional to the variance of the training data the A_i enclose. Two extra components A_0 for gain and A_I for bias are added, see [1]. Let $B_a = [A_1, \cdots, A_i, \cdots, A_0, A_I]$ be the appearance subspace basis. An instance of appearance is defined as a linear expression: $A = B_a p_a$ with p_a the appearance variation parameters.

Fitting an AAM consists to find the shape and appearance parameters that make it match the input image as best as possible. This is done by an iterative, nonlinear optimization process. We use the inverse compositional optimization scheme presented by Baker and Matthews in [10]. The Jacobian and Hessian matrices are derived analytically. Two versions of this algorithm were proposed and compared in [7]. Our implementation relies on the most accurate one called the *simultaneous inverse compositional algorithm*, originally described in [1].

We adapt this algorithm to our multi-level segmented AAM.

3 Assessing Fitting Accuracy

Fitting accuracy on unseen face images is generally assessed based on a single manual annotation of each image, considered as the absolute shape reference. The assumption behind this accuracy evaluation method is that the manual label is correct at the pixel level.

This assumption is often violated in practice: a vertex on a face image gets significantly different manual annotations, even from the same user. It is also incorrect to consider that one manual annotation is better than the others. It might also happen that a well performing automatic process is more accurate than manual labellers.

To address this improper accuracy assessment problem, Mercier *et al.* [11] suggest to annotate a face several times and build statistics for each vertex. It is then possible to set up a fitting error measure that takes the imprecision of manual annotation into account. The fitting accuracy score is given strong weight for those vertices that manual labellers have localised accurately, and light weight for badly localised vertices.

We use the multiple label data available from [11] to define the ground truth shape and the fitting error function. A set of $n_I = 40$ images were labelled $n_L = 10$ times each (labels describe the $n_V = 68$ vertices of the model mesh used in [10]). These frontal pose, neutral expression, homogeneous illumination, face images are extracted from the AR database [9]. Each image shows a different individual. A probability distribution is computed for each image i and vertex v, as the mean $\mu_{i,v}$ over its n_L labels $x_{i,v,l}$ and a (2×2) covariance matrix $\Sigma_{i,v}$ as:

$$\mu_{i,v} = \frac{1}{n_L} \sum_{l=1}^{n_L} x_{i,v,l} \quad \text{and} \quad \Sigma_{i,v} = \frac{1}{n_L - 1} \sum_{l=1}^{n_L} (x_{i,v,l} - \mu_{i,v})^T (x_{i,v,l} - \mu_{i,v}).$$

These define what we dub ‘manual labelling statistics’. Figure 1 shows face images overlaid with their manual labelling statistics, with each vertex represented by an ellipse showing its mean position and uncertainty. This methodology is in contrast to [11] in which a

single covariance matrix is computed for each vertex over all the n_I images. We believe that keeping a single covariance matrix for each vertex in each image makes sense since the visibility conditions may substantially differ from one image to the others for the same vertex. We want to preserve this information in the statitics.

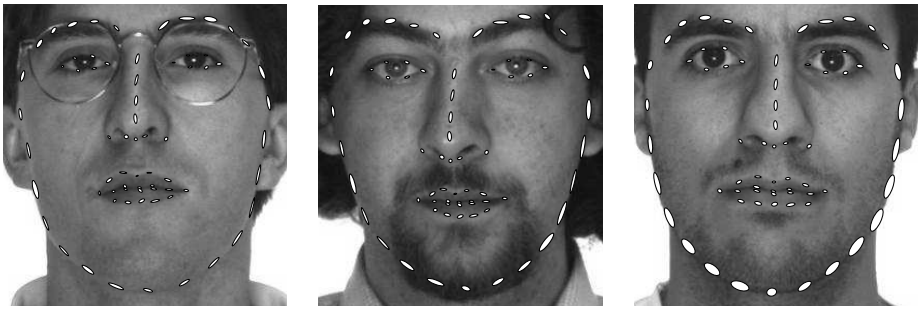


Figure 1: Faces number 1, 3 and 6 from the 40 faces that were annotated 10 times. Covariance ellipses represent the distribution of the 10 labels around mean vertices.

We propose the *Statistical Shape Error* (SSE) for a shape s on an image i that we define by the average of the Mahalanobis distances:

$$SSE_i(s) = \frac{1}{n_V} \sum_{v=1}^{n_V} \sqrt{(s_v - \mu_{i,v})^T \Sigma_{i,v}^{-1} (s_v - \mu_{i,v})}, \tag{1}$$

where s_v is the v-th vertex of shape s. The lower the SSE, the better the fitting accuracy. This error is strongly related to the negative log-likelihood of the parameters with respect to gaussian noise contamined labels. It scores automatic fits and can also be used to score manual fitting accuracy. In particular, we compute the SSE obtained by the 10 labellings on each image. From the 10 error scores on each image we retain the maximum and minimum scores, and compute the average score. This allows us to compare automatic fitting accuracy to manual fitting accuracy in §5.3, which gives a concrete idea of the accuracy that is reached.

4 Issues in Fitting AAMs to Unseen Faces

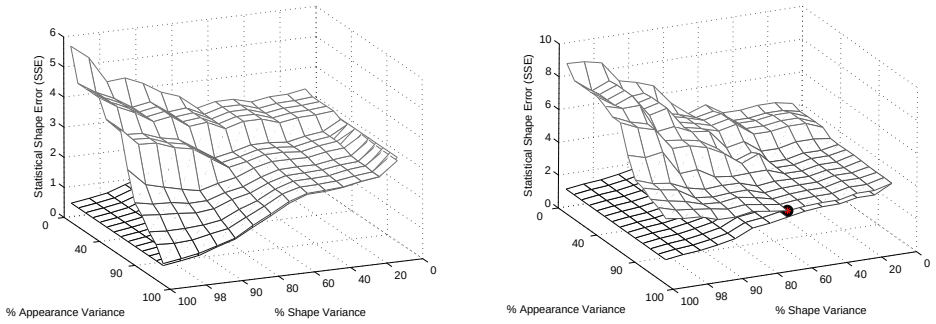
In the literature, AAMs are usually built by retaining 95 to 98% of the total shape and appearance variance contained in the training data without justifying this choice. Few works study the influence of the quantity of variance on the fitting performances. Gross *et al.* [7] recently investigated the effect of shape and appearance variance on the convergence of a fitting algorithm for unseen faces. They estimate the quantities of shape and appearance variance that maximize the number of successful trials. However, they do not explain why the convergence is limited for certain faces.

The experiment we report allows to highlight the fitting accuracy behavior for a range of shape and appearance variances. We identify the combination that maximizes the overall fitting accuracy on all trials and explain why this accuracy is limited and the fitting behaviour for various shape and appearance variance combinations. The experimental setup has similarities with the one in [7].

4.1 Fitting Seen Faces, *i.e.* Images in the Training Set

We use all the 40 images to train the AAM with different amounts of shape and appearance variance. We fit it to the 40 face images on turn. Each fitting trial lasts a number of iterations that allows to reach a clear final state, should it be convergence or divergence.

As in [7], we initialize the fitting process as close as possible to the optimal parameterization: we project the test face shape into the shape subspace to retrieve the initial shape parameters, and its appearance into the appearance subspace to retrieve the appearance parameters. In this way, we ensure that if the model diverges, it is not due to potential local minima but to the model inability to fit the test image. Figure 2 (a) shows the average SSE on all the 40 face images. The bottom curve shows the model SSE in its initial position, which is also the lowest error it can reach.



(a) Accuracy for **seen faces** (images in the training set)

(b) Accuracy for **unseen faces** (individuals not in the training set)

Figure 2: *Seen* and *unseen* contexts analysis. The bottom curves represent the initial SSE of equation (1) averaged over all the 40 images. The top curves show the average SSE after the algorithm has ran. Various amounts of shape and appearance variances are tested. The black dot in (b) represents the point of best average fitting accuracy on unseen faces that stands for 60% of the shape variance and 100% of the appearance variance.

We observe that for full appearance (100% of the variance retained), the fit remains in the best, initial position for any amount of shape variance retained. The characteristics of the full appearance AAM is that, up to appearance sampling artefacts, the test image can be completely reconstructed in appearance.

The second observation holds for any given fixed amount of shape variance: when less than 100% of the appearance variance is retained, the fitting accuracy decreases. The less the appearance variance, the worse the accuracy. For less than 100% of the training data, the AAM appearance space cannot totally reproduce the face appearance of the input image although this face is in the training set. This reluctant intensity discrepancy between the test image and the model causes the drop in the fitting accuracy. This test highlights the following property: the ability of the AAM appearance component to fully generate the face appearance of the test image is a necessary condition to obtain the best possible fitting accuracy. A natural question to answer is whether this also holds when the test image is not in the training set.

4.2 Fitting Unseen Faces, *i.e.* Individuals Not in the Training Set

This test is different from the one in §4.1 in that it is done in a leave-one-out manner: we train the AAM on 39 images and use the 40-th image as a test image of an unseen face. The test is performed for all the 40 face pictures, and for variable amounts of shape and appearance variance.

Figure 2 (b) shows the average SSE over the 40 face images. It is pretty similar to figure 2 (a) but a main difference is however observed. There is no combination of shape and appearance variances that makes the AAM remains into the initial, best position. Indeed, there is no junction between the SSE curves for the initial and fit curves. In contrast with the test on seen faces, even for full appearance AAMs, the fitting process shifts the AAM away from the initial solution. The AAM never remains on the best possible accuracy position, which results in limited accuracy capabilities.

4.3 Discussion

As an observation on the test for seen faces in §4.1, we saw that when the model can fully express the image in terms of appearance (the error in intensity between the model and the image are due to the model misplacement and/or non-optimal appearance parameterization). The fitting optimization process uses the error in intensity to iteratively update the model to a position where this error is minimized, and ideally equals zero. It is assumed that the model parameterization that minimizes the error in intensity correctly aligns the model to the face image. In practice, this is what happens when the model explicitly learnt the image it fits (and when the global minimum is reached). This explains the high fitting accuracy obtained in this context.

When the model appearance cannot fully express the face on the test image, the error in intensity due to this lack of expressivity is considered as being due to the model misplacement. The optimization process tunes the model parameters to minimize the residual error though it does not come from a misplacement. In this case, the minimum error usually does not correspond to the best placement of the model vertices. Indeed, the process bends the model in order to spread out the remaining error in intensity as much as it can to minimize the global error. This makes the model drift away from the sought after shape used as its initial position, *i.e.* fitting accuracy is spoiled. The more deformable the model the more the fitting process can bend it to further minimize error in intensity. For very high deformability the model can even diverge. In the same way for a given deformability (fixed amount of shape variance), the less the appearance variance, the less the model can express the test image data and the worse the fitting accuracy. In the case of fitting on seen faces, this happens when appearance variance is not fully retained (less than 100% of the variance is retained). In the case of fitting on unseen faces, a new face always presents visual aspects that are unknown from the model appearance component and the model always drifts away from the best possible position even when appearance is fully retained. As seen on the curves of figure 2 (b), the best overall accuracy for fitting on unseen faces is obtained for full appearance and 60% shape variance, making the model rigid enough not to bend too much, then minimizing the loss in fitting accuracy.

5 Segmented AAMs and Unseen Individuals

5.1 Motivations for Using Segmented AAMs

The AAM appearance space is unable to completely generate the appearance information. In other words, an unseen face added to the training set would bring new visual information. We saw that the limited ability to generalize the appearance component limits the fitting performance in terms of accuracy. One obvious solution to better generalize to any new face appearance would be to train the AAM on thousands of training images. This is difficult in practice for two reasons: first, this number of training data is hard to gather up, and second, this implies to retain a very high number of appearance components to explain as much of the variance as possible, which makes the optimization process computationally heavy and increases the possibility of getting stuck into local minima.

The solution we propose is to reduce the appearance space dimensionality. This makes more expressive the data coming from our reasonable size training set. To achieve a better fitting accuracy we rely on local models defined over a smaller face area. This approach is somehow similar to the concept of *segmented morphable models* briefly presented by Blanz and Vetter in [2].

5.2 Multi-Level Segmented AAMs and Coarse-to-Fine Fitting

The ability of local models to generalize their shape and appearance is better than for larger models. This makes them potentially more accurate for the same amount of training data. However, their reduced dimension penalizes their robustness to bad initialization: local models must be well initialized. To ensure this, we use a three stage coarse-to-fine strategy, illustrated on figure 3, where a global AAM is used to initialize intermediary AAMs, themselves used to initialize local AAMs.

Intermediary and local models represent a subgroup of the global model vertices. Models concerned with eyebrows also describe some extra vertices on lower eyebrows. A layer of *supporting points* is added to local and intermediary models in order to define visual gradients at 360° around all vertices.

The model is automatically initialized. We use a face and eye center detector available online[1] [6]. The rigid global model is transformed with a 2D similarity and is placed on the image such that its eye centers match the corresponding estimate given by the detector. From this initial position the fitting is launched until it converges.

Global model position is used to initialize each intermediary model: we keep the vertices the global model has in common with an intermediary model, and we find the intermediary model instance that best matches those vertices, as follows.

Let B_{succ} be the shape generating matrix of one intermediary model: the columns of B_{succ} are the n_C long deformation vectors s_i plus the four similarity transform vectors. Vector s_{curr} represents the vertex coordinates of the global model that are in common with the intermediary model. To this vector we add extra null coordinates for intermediary model vertices that are not in common with global model. s_{curr} thus becomes n_C long. We sort the coordinates in s_{curr} in a way such that they correspond to vertex coordinates defined in the vectors s_i. The instance of intermediary model that best matches its common vertices to those of the global model is found by solving the following optimization

[1] http://kolmogorov.sourceforge.net

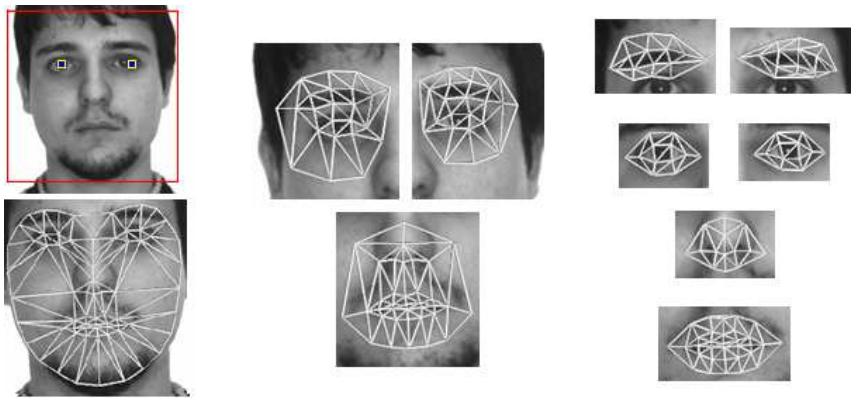


Figure 3: Illustration of the models used to fit the face. A global model is initialized with help of an eye center detector and is fitted on the image giving a first fitting result (left column). From this initialization a set of intermediary models are launched to further refine the fitting (center column). Eventually, the local models dedicated to each facial feature are launched to fit these features more accurately (right column).

problem:

$$\arg\min_{p} \sum_{c=1}^{n_C} Q(c)\,(s_{curr}(c) - B^c_{succ} p)^2\,, \tag{2}$$

where B^c_{succ} is the c^{th} row of matrix B_{succ}. Q is an n_C long vector of weights set to one for the coordinates of vertices that are common between the models, and to zero for the others. A closed form solution can be computed to find the optimal $p^\dagger$ (details are omitted due to lack of space):

$$p^\dagger = (K^T K)^{-1} K^T B^T_{succ} diag(Q) s_{curr}, \tag{3}$$

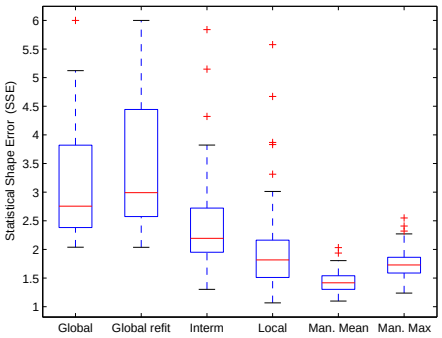
where $K = B^T_{succ} diag(Q) B_{succ}$ and $diag(Q)$ is a diagonal matrix, null everywhere excepted on diagonal where the Q vector coefficients are represented. The result $p^\dagger$ of this minimization can be used to instantiate the shape of the intermediary model: $s_{succ} = B_{succ} p^\dagger$. The process is applied to initialize all intermediary models that are then fitted to the image. Following the same strategy, we use (converged) intermediary model vertices to initialize local models that are in turn fitted to the image. Once each model is initialised in position, its appearance component is initialised by projection of the area underlying on the image onto the appearance subspace, in order to retrieve the initial appearance parameters.

5.3 Experimental Results

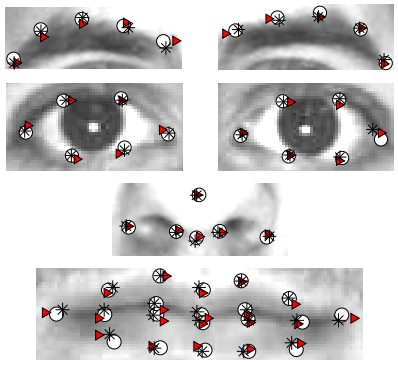
Fitting with intermediary and local models leads to improved fitting accuracy. On figure 4 (a), boxes of whiskers compare the SSE on 40 trials obtained respectively with global, global refitted, intermediary, local models and mean manual error as well as maximum manual error (see §3). The global refitted model is obtained by training the global model

onto refitted data: the used face data is learnt by an AAM retaining 99% variance of shape and appearance, and is fitted again on the same faces in order to increase the vertices correspondences among training data. Introduced in [7], this operation seems to improve the fitting results on unseen faces with respect to the results obtained when training the AAM on once-labelled data. Since we use multiply labelled data for training (the mean of 10 labels to define each vertex), their semantical position on the face is high and should naturally improve the correspondence among data.

A leave-one-out procedure is used to train and fit the global, global-refitted, intermediary and local models. Models are built with the shape and appearance variances that maximize their overall accuracy on unseen faces (*e.g.* 60% shape and 100% appearance for the global model). All intermediary models are gathered. The same is done for the local models. The SSE is computed using equation (1) only on vertices that are common between the models. We see the accuracy improvement allowed by intermediary and local models with respect to the global model, and we see that the accuracy is globally comparable to manual label accuracy evaluated with the statistics. The relative improvement obtained from global to local model fitting is 36% on average. The SSE obtained for refitted data is higher than for the global model trained with multiply labelled data. We believe that the higher semantical meaning obtained with label statistics is mainly responsible for the improvement. Indeed, these labels have higher semantical meaning since human labellers attempted several times to accurately set them into a given position on each face. The refitting process will displace once-labelled vertices to maximize their cohesion, but it is improbable that these new positions are semantically the very desired ones (although they might usually be improved). Multiply labelled data then constitute a maximum bound to accuracy, which explains the improved results obtained when we train an AAM with these data.



(a) Comparison results



(b) Example of fitting results

Figure 4: (a) Comparison between the 40 fitting scores obtained with a global model, a refitted global model, intermediary models, local models, and manual labellings, both maximum and average SSE. Local models often reach a SSE comparable to manual labellings. (b) Example of fitting results on face number 6. The circles represent the ground truth shape vertices (centers of the covariance ellipses), the triangles the vertices of the fitted global model (the SSE equals 2.46), and the stars represent the vertices of the fitted local models (the SSE equals 1.46).

6 Conclusion and Future Work

The AAM paradigm is often used without precisely understanding the influence of the *quantity and nature of training data* and of the *retained quantity of shape and appearance variance* on fitting performances. As a step towards such an understanding this work studies fitting accuracy on unseen frontal and neutral face data through the *Statistical Shape Error* we propose. We showed and explained the fitting accuracy limitations in this case. We propose a solution based on local models, namely the *multi-level segmented AAM*, that overcomes this limitation and reaches very high accuracy benchmarked by manual fitting accuracy with a large convergence basin. To summarize, standard and segmented AAMs are respectively well-adapted to person-specific and person independent face fitting.

We wish to extend these results to varying pose and expression: we will train one set of global, intermediary and local models for each possible pose and expression and set up a strategy to select the set that best suits for fitting the current face image.

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, November 2003.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of* SIGGRAPH, 1999.

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

[4] D. Cosker, D. Marshall, P. Rosin, and Y.A. Hicks. Speech driven facial animation using a hierarchical model. IEE *VISP*, 151(4):314–321, 2004.

[5] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference*, 2006.

[6] I. Fasel, B. Fortenberry, and J. R. Movellan. Generative framework for real-time object detection and classification. *CVIU*, 98(1):182–210, April 2005.

[7] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, 2006.

[8] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn. AAM derived face representation for robust facial action recognition. *FGR*, 2006.

[9] A.M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, June 1998.

[10] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, November 2004.

[11] H. Mercier, J. Peyras, and P. Dalle. Toward an efficient and accurate AAM fitting on appearance varying faces. *FGR*, 2006.