Generic Active Appearance Models Revisited

Georgios Tzimiropoulos[1,2], Joan Alabort-i-Medina[1], Stefanos Zafeiriou[1], and Maja Pantic[1,3]

[1] Department of Computing, Imperial College London, United Kingdom
[2] School of Computer Science, University of Lincoln, United Kingdom
[3] Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands
{gt204, ja310, s.zafeiriou, m.pantic}@imperial.ac.uk

Abstract. The proposed Active Orientation Models (AOMs) are generative models of facial shape and appearance. Their main differences with the well-known paradigm of Active Appearance Models (AAMs) are (i) they use a different statistical model of appearance, (ii) they are accompanied by a robust algorithm for model fitting and parameter estimation and (iii) and, most importantly, they generalize well to unseen faces and variations. Their main similarity is computational complexity. The project-out version of AOMs is as computationally efficient as the standard project-out inverse compositional algorithm which is admittedly the fastest algorithm for fitting AAMs. We show that not only does the AOM generalize well to unseen identities, but also it outperforms state-of-the-art algorithms for the same task by a large margin. Finally, we prove our claims by providing Matlab code for reproducing our experiments (`http://ibug.doc.ic.ac.uk/resources`).

1 Introduction

Because of their numerous applications in HCI, face analysis/recognition and medical imaging, the problems of learning and fitting deformable models have been the focus of cutting edge research in computer vision and machine learning for more than two decades. Put in simple terms, these problems can be summarized as follows: Learning a deformable model consists of (a) annotating (typically manually) a set of points (or landmarks) over a set of training images capturing an object of interest (e.g. faces), (b) learning a shape model (or point distribution model) which effectively represents the structure and variations among the annotated points and (c) learning appearance models from the image texture associated with the learned shape. Fitting a deformable model utilizes the learned shape and appearance models to detect the location of landmarks in new images; this can be done using regression, classification or could be formulated as a non-linear optimization problem.

Depending on the application and/or approach many terms have been used to coin this research: deformable model fitting, Active Shape Models (ASMs) [1], Constrained Local Models (CLMs) [2, 3] landmark localization, point detection,

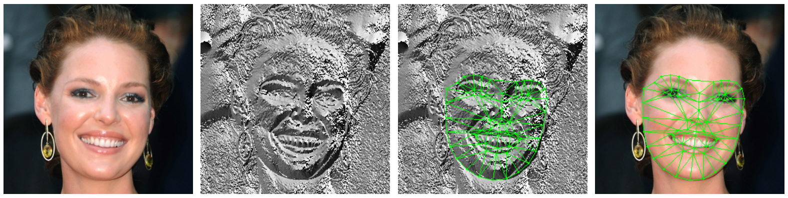


Fig. 1. The proposed Active Orientation Models (AOMs) are generative deformable models that: (i) use a statistically robust appearance model based on images gradient orientations, (ii) are accompanied by a robust algorithm for model fitting and parameter estimation and (iii) and, most importantly, generalize well to unseen faces and variations.

and Active Appearance Models (AAMs) [4, 5] to name a few. The latter approach and the problem of deformable face alignment are of particular interest to this work along with the seminal work [5] for fitting AAMs in face images. AAMs are generative models of shape and appearance typically learned by applying Principal Component Analysis to both shape and texture. In [5], fitting was formulated as a non-linear minimization problem which consists of minimizing the error between the model instance and the given image with respect to the model parameters which control the shape and appearance variation of faces. This problem was solved using the project-out inverse compositional algorithm, which decouples shape from appearance and results in a computationally efficient algorithm. Owing to its efficiency and accuracy, the algorithm for fitting AAMs proposed in [5] has become the de facto choice for building and fitting person-specific AAMs (i.e. AAMs trained to fit face images of a specific subject which is known in advanced).

Despite their efficiency and accuracy, AAMs in general, and the project out algorithm of [5] in particular, have been heavily criticized for their inability to generalize well in unseen variations of illumination, expression and most importantly identity. Various algorithms have been proposed to address these challenges. Recent research has suggested the use of simpler (often local) and thus easier to optimize appearance models and the application of discriminative methods for model fitting. Discriminative methods have been already shown to improve the ability of AAMs to fit new faces in [6–9]. The family of methods termed CLMs combine patch-based image representations, discriminatively trained point detectors and global shape constraints to localize landmarks in new images [2, 10, 3]. Among these methods, the non-parametric approach for approximating the response maps of the local detectors has been shown to produce state-of-the-art results [11]. A recent approach that combines the output of SVM-based local detectors with a non-parametric set of global models has been shown to produce good results on unconstrained images in [12]. Finally, because sliding-window landmark detectors may be slow, regression-based techniques have been proposed to learn a mapping between local patches and landmarks [12–14]. Not only do these methods enjoy a high degree of computational efficiency, but also it is often claimed that they achieve state-of-the-art performance for difficult experiments with unconstrained images.

Our motivation to pursue a generative AAM-based approach to generic face alignment is three fold. (a) It has not been shown yet that the accuracy of (generative) person-specific AAMs can be matched by any discriminative-based alignment method. (b) For specific scenarios of interest [15], generative models have been shown to model sufficiently accurately unseen variations, i.e. it is the fitting algorithm which fails to fit the model to unseen images. (c) New tools and insights on how to solve (b) have been recently suggested in [16, 17].

Main results. (a) **Models**: We propose Active Orientation Model, a generative deformable model that uses a statistically robust appearance model based on the principal components of images gradient orientations (see Fig. 1). We show how to use gradient ascent in order to maximize the correlation of this non-linear appearance model and a new image with respect to the model shape parameters. (b) **Complexity**: Similarly to the AAM formulation of [5], we show that AOMs can be optimized using the project-out inverse compositional algorithm which is admittedly the fastest algorithm for fitting deformable models in images. (c) **Robustness/Accuracy**: To the best of our knowledge, we demonstrate for the first time that this algorithm can be used to fit the learned models to faces not seen in the training set. While other variations of AAMs have been somewhat shown to fit unseen faces, the reported results are always significantly inferior to what is considered the state-of-the-art. On the contrary, we show that the AOM proposed here outperforms state-of-the-art algorithms including the best version of the recently proposed method of [18] by a large margin. (d) **Reproducible research**: Similarly to [13, 11, 18] we prove our claims by providing Matlab code which reproduces our results (`http://ibug.doc.ic.ac.uk/resources`).

2 Active Appearance Models

Prior work. AAMs able to fit unseen faces do exist, however the reported fitting performances have been always outperformed (in many cases by a large margin) by what it was considered the state-of-the-art at that time. One of the first attempts is the AAM of Cootes and Taylor [19] which is based on regression and features somewhat similar to the ones that the AOM is based on. However, this AAM was later shown to perform worse than the baseline Constrained Local Model (CLM) algorithm [2]. Also these features have been shown to be significantly inferior to the optimization strategy of [16] which has been adopted here for the case of AOMs. Other "generic" AAMs learn a fitting function through maximizing the score of a two-class classifier (aligned or not aligned) or ranking [6, 7]. Boosting a huge number of Haar features is very inefficient, and results are reported only for low resolution images. This immediately rules out the possibility of accurate landmark localization in high resolution images and it is clearly unsatisfactory. Other discriminative approaches include learning non-linear regressors from features to model parameters through boosting and simulation [8, 9]. Again, all these approaches seem to produce inferior results compared to the family of methods coined CLMs [2, 10, 3], which build upon the Active Shape Model [1]. Very recently, a globally optimized part-based model has been shown to produce state-of-the-art results [18]. The AOM proposed in this work outperforms both the state-of-the-art CLM method of [3] and the best version of [18] by a large margin.

Background. An AAM is defined by a shape, appearance and motion model. The shape model is typically learned by annotating N fiducial points $\mathbf{s}_i = [x_1, y_1, x_2, y_2, \ldots x_N, y_N]$ to each of a set of training images $\{\mathbf{I}_i\}$ and then applying PCA on $\mathbf{s}_i$. The resulting model $\{\bar{\mathbf{s}}, \mathbf{\Phi}_S \in \mathcal{R}^{\{2N,p\}}\}$ can be used to represent a test shape $\mathbf{s}_y$ as

$$\hat{\mathbf{s}}_y = \bar{\mathbf{s}} + \mathbf{\Phi}_S \mathbf{p}, \quad \mathbf{p} = \mathbf{\Phi}_S^T (\mathbf{s}_y - \bar{\mathbf{s}}). \tag{1}$$

The appearance model is learned by first warping each of the training images $\mathbf{I}_i(\mathbf{x}) \in \mathcal{R}^K$ to the canonical reference frame defined by the mean shape $\mathbf{s}_i$ using motion model $\mathbf{W}(\mathbf{x}; \mathbf{p})$ and then applying PCA on the shape-free textures. We choose piecewise affine warps as the motion model in this work. The resulting model $\{\bar{\mathbf{a}}, \mathbf{\Phi}_A \in \mathcal{R}^{\{K,q\}}\}$ can be used to represent a shape-free test texture $\mathbf{a}_y$ as

$$\hat{\mathbf{a}}_y = \bar{\mathbf{a}} + \mathbf{\Phi}_A \mathbf{c}, \quad \mathbf{c} = \mathbf{\Phi}_A^T (\mathbf{a}_y - \bar{\mathbf{a}}). \tag{2}$$

Given a test image $\mathbf{I}_y$, inference in AAMs entails estimating $\mathbf{p}$ and $\mathbf{c}$ assuming "reasonable" initialization of the fitting process. This initialization is typically performed by placing the mean shape according to the output of an object (in this work face) detector. Note that only $\mathbf{p}$ needs to be estimated for deformable model fitting. Estimating $\mathbf{c}$ is a by-product of the fitting algorithm. Various algorithms and cost functions have been proposed to estimate $\mathbf{p}$ and $\mathbf{c}$ including regression, classification and non-linear optimization methods. The latter approach is of particular interest in this work. It minimizes the ℓ_2-norm of the error between the model instance and the given image with respect to the model parameters as follows

$$\{\mathbf{p}_o, \mathbf{c}_o\} = \arg \min_{\{\mathbf{p},\mathbf{c}\}} ||\mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p})) - \bar{\mathbf{a}} - \mathbf{\Phi}_A \mathbf{c}||^2. \tag{3}$$

The project-out algorithm decouples shape and appearance by solving (3) in a subspace orthogonal to the appearance variation. This can be done by applying the projector operator $\mathbf{P} = \mathbf{E} - \mathbf{\Phi}_A \mathbf{\Phi}_A^T$ ($\mathbf{E}$ is the identity matrix) to any of $\mathbf{I}$, $\bar{\mathbf{a}}$. The resulting optimization problem can be then solved very efficiently using a variation of the Gauss-Newton algorithm coined inverse compositional algorithm [5].

3 Active Orientation Models

The deformable model fitting framework of the previous section has been highly criticized as difficult to optimize mainly due to the high-dimensional parameter space and the existence of numerous undesirable local minima in the cost function of (3). Therefore, the problem in hand is how to avoid these local minima during optimization. We propose to address this problem by using a similarity criterion robust to outliers. The Active Orientation Models proposed in this work are designed to use the same shape and motion model as the ones used by AAMs but a different appearance model and a different cost function to fit this model.

At the heart of AOMs there exists a robust kernel for measuring similarity. We define outliers to be anything that the learned appearance model cannot reconstruct because (a) it was not seen in the training set (e.g. appearance variation due to different identity, expression or illumination) (b) it does not belong to the face space at all (e.g. glasses) and (c) it was excluded from $\mathbf{\Phi}_A$ as noise because in any case the number of principal components in $\mathbf{\Phi}_A$ should be kept as small as possible so that the model is easier to optimize and cannot generate appearance which is unrelated to faces. Note that as it was shown in [15] for some cases of interest (e.g. appearance variation in frontal views), a very compact appearance space, learned from a training set with a few persons only, in general, results in relatively small reconstruction errors of unseen faces. This illustrates that a generative model is not an unreasonable choice for generic deformable model fitting. All that is needed is a robust cost function to fit this model.

A general framework for robust estimation is weighted least squares [20]. Let us define $\mathbf{e} = \mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p})) - \bar{\mathbf{a}} - \mathbf{\Phi}_A\mathbf{c}$. Then, weighted least squares methods optimize

$$\{\mathbf{p}_o, \mathbf{c}_o\} = \arg\min_{\{\mathbf{p},\mathbf{c}\}} \mathbf{e}^T\mathbf{Q}\mathbf{e}, \tag{4}$$

where $\mathbf{Q} \in R^{\{K,K\}}$ is a diagonal weighting matrix which down-weighs pixels corrupted by outliers. An ideal case would be $\mathbf{Q}_k = 0$ if pixel k is an outlier and $\mathbf{Q}_k = 1$ otherwise. The estimation of $\mathbf{Q}$ along with the optimal model parameters have been extensively studied in the literature of robust statistics (please see [20] for a review). However, none of these out-of-the-box approaches has been proven successful so far in AAM fitting because (a) the noise model for outliers in our case is very hard to define and (b) the estimation process is also very prone to local minima.

We propose to address this problem in AAMs by using a robust similarity criterion based on image gradient orientations [16, 17]. Suppose that we wish to measure the similarity between two images $\mathbf{I}_i$, $i = 1, 2$. For each image, we extract image gradients $\mathbf{g}_{i,x}, \mathbf{g}_{i,y}$ and the corresponding estimates of gradient orientation $\boldsymbol{\phi}_i$. Let us denote by $\mathbf{z}_i$ the the so-called normalized gradients

$$\mathbf{z}_i = \frac{1}{\sqrt{K}}[\cos(\boldsymbol{\phi}_i)^T,\ \sin(\boldsymbol{\phi}_i)^T]^T, \tag{5}$$

where $\cos(\boldsymbol{\phi}_i) = [\cos(\boldsymbol{\phi}_i(1)), \ldots, \cos(\boldsymbol{\phi}_i(K))]^T$ and $\sin(\boldsymbol{\phi}_i)$ is similarly defined. Then, the following kernel can be used to measure image similarity

$$\begin{aligned} s &= \mathbf{z}_1^T\mathbf{z}_2 \\ &= \frac{1}{K}\sum_{k\in\Omega}\cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)), \end{aligned} \tag{6}$$

where Ω denotes the image support.

Let us also denote by Ω_1 the image support that is outlier-free and Ω_2 the image support that is corrupted by outliers ($\Omega = \Omega_1 \cup \Omega_2$). Then, as it was shown in [17], under some assumptions, it holds

$$\sum_{k \in \Omega_2} \cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)) \approx 0. \tag{7}$$

Note that (a) in contrary to [21], no assumption about the structure of outliers is made and (b) no actual knowledge of Ω is required. Based on (7), we can re-write (6) as follows

$$\begin{aligned} s &= \sum_{k \in \Omega_1} \cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)) + \sum_{k \in \Omega_2} \cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)) \\ &= \sum_{k \in \Omega_1} 1 \cdot \cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)) + \sum_{k \in \Omega_2} \epsilon \cdot \cos(\boldsymbol{\phi}_1(k) - \boldsymbol{\phi}_2(k)) \\ &\approx \mathbf{z}_1^T \mathbf{Q}_{\text{ideal}} \mathbf{z}_2, \end{aligned} \tag{8}$$

where $\epsilon \to 0$ and $\mathbf{Q}_{\text{ideal}}$ is the "ideal" weighting matrix defined above. Note that $\mathbf{Q}_{\text{ideal}}$ in (8) is never calculated explicitly. We can write (8) only because outliers are approximately "canceled out" when the above kernel is used to measure image similarity.

3.1 Appearance Model

The robust kernel of (6) can be used to define a kernel PCA [17]. The appearance model in AOMs is learned using this robust PCA. Note that the kernel can be written using the explicit mapping of (5) and therefore no pre-image computation is required. All that is needed is to compute the normalized image gradients of (5), define the data matrix $\mathbf{Z}$ the columns of which are the shape-free normalized gradients of the training faces and then apply standard PCA on $\mathbf{Z}$. Note that to preserve the kernel properties no subtraction of the "mean" normalized gradient is needed and the first eigenvector is treated as the mean where it is required. We denote by $\mathbf{\Phi}_Z \in \Re^{2K \times q}$ the learned appearance model.

3.2 Inference

We perform inference in AOMs by maximizing the correlation of a test image with the learned appearance model

$$\{\mathbf{p}_o, \mathbf{c}_o\} = \arg\max_{\{\mathbf{p}, \mathbf{c}\}} \mathbf{z}[\mathbf{p}]^T \mathbf{\Phi}_Z \mathbf{c}, \tag{9}$$

where $\mathbf{z}[\mathbf{p}]$ denotes the normalized gradients of $\mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p}))$. We used a formulation similar to the one proposed in [16, 22] which does not require the calculation of second-order derivatives. This entails the maximization of the normalized correlation

$$\{\mathbf{p}_o, \mathbf{c}_o\} = \arg\max_{\{\mathbf{p},\mathbf{c}\}} \frac{\mathbf{z}[\mathbf{p}]^T \mathbf{\Phi}_Z \mathbf{c}}{||\mathbf{\Phi}_Z \mathbf{c}||}, \tag{10}$$

($\mathbf{z}[\mathbf{p}]$ has a unit norm by definition) using the inverse compositional framework of [5]. This is an efficient Gauss-Newton algorithm in which the Hessian is pre-computed and remains constant across iterations when appearance variation is known ($\mathbf{c}$ is known) or can be ignored. In brief, in the inverse compositional framework, each principal component of $\mathbf{\Phi}_Z$ is linearized with respect to the shape parameters around $\mathbf{0}$, an increment $\mathit{\Delta}\mathbf{p}$ is estimated at each iteration (using the closed-form expression obtained by setting the derivatives equal to 0) and then $\mathbf{W}(\mathbf{x};\mathbf{p})$ is composed with $-\mathit{\Delta}\mathbf{p}$. The updated $\mathbf{W}$ is finally used to warp the test image and move on to the next iteration. Note that because at each step $\mathbf{\Phi}_Z$ is linearized about $\mathbf{0}$, the Hessian can be pre-computed and stored along with the model. This results in significant computational savings. For more details on the inverse compositional algorithm and how to implement the compositional approach for AAMs please see [5]. Our optimization strategies are summarized below:

Simultaneous. The simultaneous AOM maximizes (10) with respect to both $\{\mathbf{p}, \mathbf{c}\}$. This requires the computation of the Hessian ($O((p+q)^2K)$) and its inverse ($O((p+q)^3)$) and it is very slow. For this reason, we did not look into this algorithm further.

Project-out. The project-out algorithm decouples shape from appearance parameters by maximizing the correlation of the test image with the "mean" appearance $\mathbf{t}$ in the subspace orthogonal to $\mathbf{\Phi}_Z$. Because we choose $\mathbf{t}$ to be the first eigenvector of $\mathbf{\Phi}_Z$, this is the case by construction. So projecting-out will be effective only after linearization. Let us denote by $\mathbf{\Phi}_{Z'}$ the appearance subspace obtained by removing the first column of $\mathbf{\Phi}_Z$ and define the projecting operator $\mathbf{P} = \mathbf{E} - \mathbf{\Phi}_{Z'}\mathbf{\Phi}_{Z'}^T$. We linearize (around $\mathbf{0}$) $\mathbf{t} = \mathbf{t}[\mathbf{0}] + \mathbf{J}\mathit{\Delta}\mathbf{p}$, and project-out. Then, the cost function of (10) takes the form

$$\mathbf{p}_o = \arg\max_{\mathbf{p}} \frac{\mathbf{z}[\mathbf{p}]^T\mathbf{t}[\mathbf{0}] + \mathbf{z}[\mathbf{p}]^T\mathbf{P}\mathbf{J}\mathit{\Delta}\mathbf{p}}{1 + 2\mathbf{t}[\mathbf{0}]^T\mathbf{J}\mathit{\Delta}\mathbf{p} + \mathit{\Delta}\mathbf{p}^T\mathbf{J}^T\mathbf{P}\mathbf{J}\mathit{\Delta}\mathbf{p}}. \tag{11}$$

The update is given by

$$\mathit{\Delta}\mathbf{p} = (\mathbf{J}^T\mathbf{P}\mathbf{J})^{-1}\mathbf{J}^T(\lambda\mathbf{z}[\mathbf{p}] - \mathbf{t}[\mathbf{0}]), \tag{12}$$

where λ is equal to

$$\lambda = \frac{1 - \mathbf{t}[\mathbf{0}]^T\mathbf{J}(\mathbf{J}^T\mathbf{P}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{t}[\mathbf{0}]}{\mathbf{z}[\mathbf{p}]^T\mathbf{t}[\mathbf{0}] - \mathbf{z}[\mathbf{p}]^T\mathbf{J}(\mathbf{J}^T\mathbf{P}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{t}[\mathbf{0}]}. \tag{13}$$

Because the Hessian $\mathbf{J}^T\mathbf{P}\mathbf{J}$ and its inverse can be pre-computed, this algorithm has complexity $O(pK)$ (per iteration) and has been shown to track faces at 300 fps [15].

Aternating. We propose to use alternating optimization to maximize (10) with respect to both $\{\mathbf{p}, \mathbf{c}\}$, which to our surprise has been barely used in AAM literature. This entails estimating the appearance of $\mathbf{z}[\mathbf{p}]$ at each iteration from $\tilde{\mathbf{z}} = \mathbf{\Phi}_Z \mathbf{\Phi}_Z^T \mathbf{z}[\mathbf{p}]$ and then maximizing the correlation of $\mathbf{z}[\mathbf{p}]$ with $\widetilde{\mathbf{z}}$. The cost function and the update rules are exactly as above with the only difference being that $\mathbf{t}$ is replace by $\widetilde{\mathbf{z}}$ and the projection operator is removed ($\mathbf{P} = \mathbf{E}$). Note however that the Hessian and its inverse must be computed at each iteration. The costs for these computations are $O(p^2K)$ and $O(p^3)$, respectively, while there is also a cost of $O(qK)$ associated with the update of appearance. This algorithm is slower than the project-out algorithm but still very fast probably allowing a real-time implementation.

4 Experiments

We assessed the performance of the proposed AOMs on generic face fitting under variations of identity, expression and pose both quantitatively and qualitatively. We also compared their performance with that of two state-of-the-art methods: the best available variation of CLMs [3] and the best possible version of the very recently proposed parts-based tree shape model of [18]. Where possible, we used the code/implementations kindly provided by the authors. Matlab code reproducing our results is also available at `http://ibug.doc.ic.ac.uk/resources`. We also provide the performance of the project-out AAM (AAM-PO) [5] as baseline. Finally, we experimented with the Fourier-Gabor AAMs [23], but our implementation did not produce any improvement over [5] for the experiments reported here, and these results were omitted for clarity of presentation.

Similarly to standard AAMs, our models work best when the shape model is constructed using a large number of annotated points per face. Therefore, for all experiments, our shape models were learned from 68-point markups which are the most dense annotations we had access to (but still quite sparse). We report the fitting accuracy achieved by two optimization strategies: AOM fitted using the project-out algorithm (AOM-PO) and AOM using alternating optimization (AOM-A). For both versions we used a standard multi-scale implementation and a simple Gaussian shape prior (only at the composition step, as in [24]) for enhancing convergence, while both algorithms were initialized by a face detector. In this sense, we report here only the baseline performance of our algorithms. In fact, an additional advantage of AOMs is the fact that they can benefit from more than 15 years research on AAMs. Sophisticated enhancements such as 3D constraints [25], densification [26] and other sophisticated shape priors [27] are left as future work.

For assessing the performance of AOMs on generic face alignment quantitatively, we used two popular databases: (a) the XM2VTS database [28], and (b) the CMU Multi-PIE [29]. Because the markups for these databases are different, we did not perform out-of-database quantitative experiments. We also provide illustrative examples of fitting AOMs trained on Multi-PIE to some of the in-the-wild images of [12]. Finally, a quantitative evaluation of the performance of

AOMs on this data set is not feasible mainly because (a) only 29 (after excluding the ears) landmarks are provided while our AOMs require dense annotation (b) a significant number of both training/testing images is no longer available or was never made publicly available.

Overview of results. (a) For the quantitative experiments, we show that our AOMs outperform both [3] and [18] in terms of robustness and fitting accuracy by a large margin. We believe that this a notable achievement given that AAMs have been in general considered successful only for person-specific face alignment. (b) Our qualitative results, although limited, indicate that AOMs are able to fit unseen in-the-wild face images. This result is surprising given that we trained our models using only 54 different subjects taken under the controlled conditions of Multi-PIE.

4.1 XM2VTS

The XM2VTS database [28] contains 2,360 frontal images of 295 different subjects displaying neutral expression. Experiments on this database are interesting mainly because of the large variation in facial shape and appearance due to facial hair, glasses, ethnicity etc. To compare directly with [3] (the result for [3] has been taken directly from the paper), we reproduced the experiment reported therein by dividing the database into four different sets with no identity overlap and perform four-fold cross validation experiments, using three parts for training and one for testing in every trial. Results for this experiment are shown in Fig. 2 and Table 4.1. Unfortunately, because the trees provided in [18] were hard coded for Multi-PIE, fair comparison with [18] is not feasible. For the remaining of methods considered, the graph shows the cumulative curve obtained by computing the percentage of images for which the fitting error was less than a specific value. We used the same shape RMSE as the one used in [3]. Table 4.1 states the exact proportion of images that were fitted with RMSE ≤ 4 and RMSE ≤ 5 pixels accuracy. For these two cases of interest, both versions of AOMs outperform the CLM algorithm of [3] for about 40% and 30% (in absolute terms) fitting accuracy. Illustrative examples of fittings for XM2VTS are shown in Fig. 4.

Shape RMSE	<3 pixels	<4 pixels	<5 pixels
AOMs - A	0.19	0.64	**0.89**
AOMs - PO	**0.25**	**0.66**	0.84
CLMs	0.03	0.25	0.55
AAMs - PO	0.05	0.19	0.42

Table 1. Proportion of images that were fitted with a shape RMSE < 4 and < 5 pixels accuracy in the XM2VTS database experiment.

4.2 Multi-PIE

The Multi-PIE face database contains around 750,000 images of 337 subjects under 15 view points, 19 illumination conditions and displaying 6 different facial expressions. We had access to markup annotations for a small subset of the database. We trained our models using 432 images from 54 different subjects. For each subject we used 8 images in total as follows: 1 image for frontal (0 degrees) neutral expression, 2 images for 2 different viewpoints (-15 and 15 degrees) displaying neutral expression; and 5 frontal images (0 degrees) displaying the remaining 5 expressions. For testing, we used the remaining annotated images (more than 1000 images in total, all faces of subjects not seen in our training set). We report the fitting accuracies for the 3 above cases in Fig. 3 and Table 4.2. For comparison purposes, we used the CLM code provided by [3]. This was trained only for frontal images so the results we report for the pose experiment are not representative of the full capabilities of the method. For the method in [18] we had two options: the first was to use the (fully shared) pre-trained models provided by the authors or to train the so-called independent model using our own data set and the training code provided by the authors. We followed the latter approach in order to report here the best possible results for [18]. Note that the independent model roughly corresponds to training different models for each of the variations being present in our data set and requires large fitting times. On the contrary, in our case, we trained a single combined AOM using the whole training set. As before, for all methods considered, each graph shows the cumulative curve obtained by computing the percentage of images for which the fitting error was less than a specific value. For this experiment, we used the same point-to-point RMSE normalized by the face size as the one used in [18] which we believe that it is the best measure for reporting fitting accuracy. Our results show that the AOM-A performs by far the best. The AOM-PO outperforms [18] for the frontal-neutral experiment and is able to fit 60% of the expression images more accurately than [18], while for the pose experiment both methods perform similarly. Finally, the CLM performs consistently much worse. Illustrative examples of fittings for Multi-PIE are shown in Fig. 5.

Experiments	Frontal Neutral		Frontal Expressions		Pose	
Norm. pp. RMSE	< 0.02	< 0.03	< 0.02	< 0.03	< 0.02	< 0.03
AOMs - A	**0.67**	**0.98**	**0.53**	**0.90**	**0.63**	**0.96**
AOMs - PO	0.46	0.91	0.26	0.69	0.20	0.87
Indep.	0.12	0.81	0.09	0.73	0.23	0.90
CLMs	0.37	0.81	0.12	0.50	0.00	0.12
AAMs - PO	0.16	0.58	0.08	0.38	0.02	0.19

Table 2. Proportion of images that were fitted with normalized point-to-point RMSE < 0.02 and < 0.03 for the frontal neutral expression, frontal remaining expression (no neutral) and pose experiments in the Multi-PIE database.

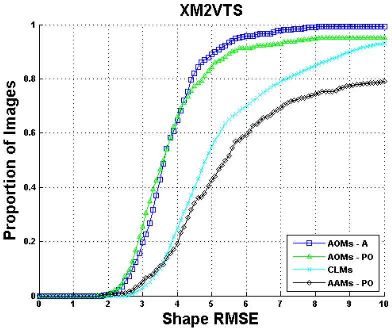
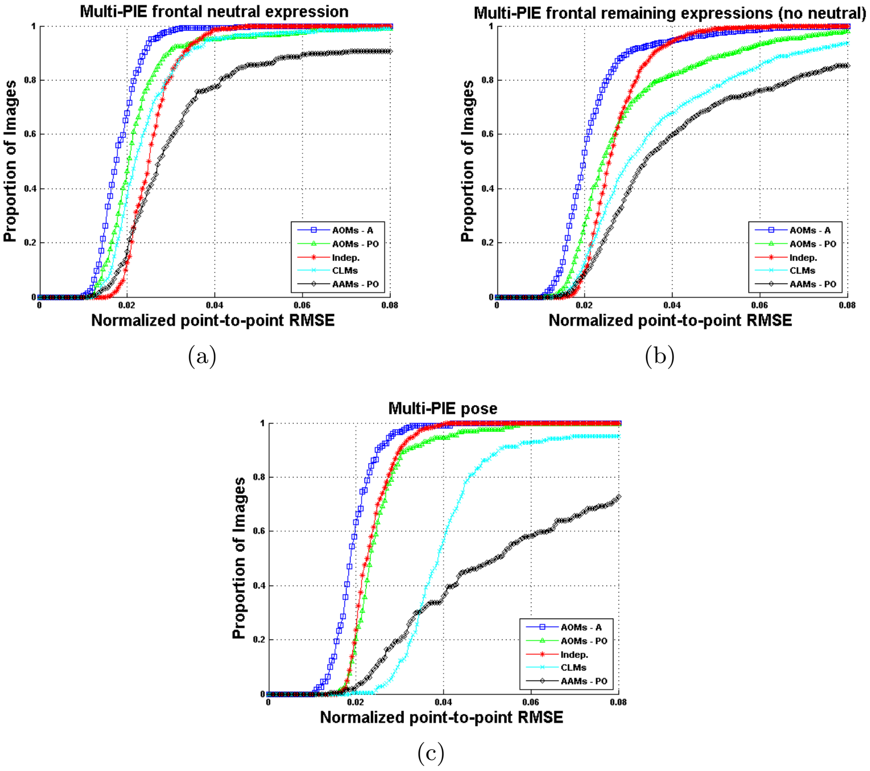


Fig. 2. Cumulative error curves for the XM2VTS database experiment.



(a)

(b)

(c)

Fig. 3. Cumulative error curves for the (a) frontal neutral expression, (b) frontal remaining expressions (no neutral) and (c) pose experiments in the Multi-PIE database.

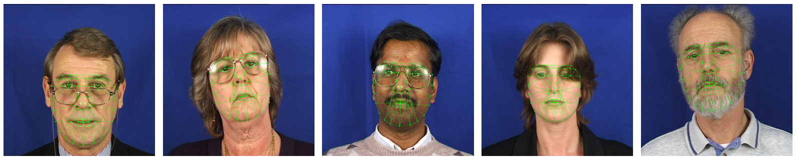


Fig. 4. Example of fitting results in the XM2VTS database.

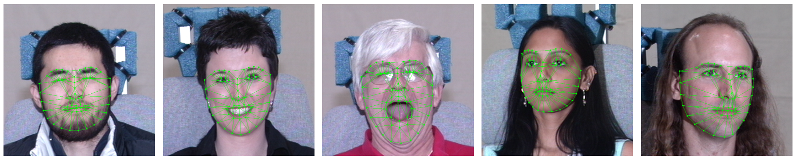


Fig. 5. Example of fitting results in the MultiPIE database.

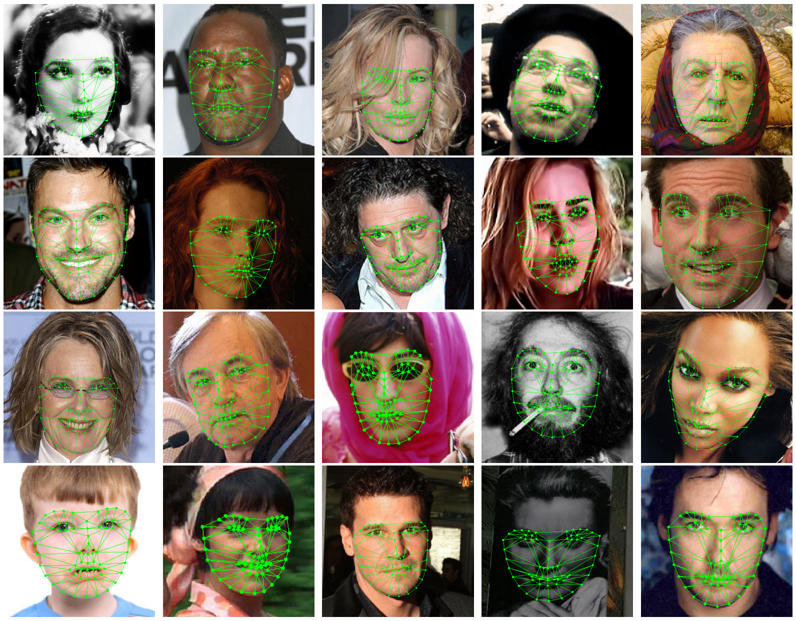


Fig. 6. Example of fitting results in the LFPW database.

4.3 In-the-wild

Although we did not assess the performance of AOMs for the case of in-the-wild images quantitatively, for the reasons mentioned above, we do provide illustrative examples of fittings in Fig. 6. We find these results surprising given that we trained our models using only 54 different subjects taken under the controlled conditions of Multi-PIE and given that only a baseline implementation of our algorithm was used.

5 Conclusions

We introduced Active Orientation Models, generative models for generic deformable face alignment that generalize well to unseen faces and variations. We show that not only does the AOM generalize well to unseen examples, but also it outperforms state-of-the-art algorithms for the same task by a large margin. Finally, we prove our claims by providing Matlab code for reproducing our experiments (`http://ibug.doc.ic.ac.uk/resources`).

Acknowledgement. The research presented in this paper has been funded by the European Communitys 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG).). The work of Joan Alabort-i-Medina is funded by a European DTA from Imperial College London. The work of Maja Pantic is funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007- StG-203143 (MAHNOB).

References

1. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. CVIU **61** (1995) 38–59
2. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition **41** (2008) 3054–3067
3. Saragih, J., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: ICCV. (2009)
4. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. TPAMI **23** (2001) 681–685
5. Matthews, I., Baker, S.: Active appearance models revisited. IJCV **60** (2004) 135–164
6. Liu, X.: Generic face alignment using boosted appearance model. In: CVPR. (2007)
7. Wu, H., Liu, X., Doretto, G.: Face alignment via boosted ranking model. In: CVPR. (2008)
8. Saragih, J., Gocke, R.: Learning aam fitting through simulation. Pattern Recognition **42** (2009) 2628–2636
9. Saragih, J., Goecke, R.: A nonlinear discriminative approach to aam fitting. In: ICCV. (2007)
10. Lucey, S., Wang, Y., Cox, M., Sridharan, S., Cohn, J.: Efficient constrained local model fitting for non-rigid face alignment. Image and Vision Computing **27** (2009) 1804–1813

11. Saragih, J., Lucey, S., Cohn, J.: Deformable model fitting by regularized landmark mean-shift. IJCV **91** (2011) 200–215
12. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR. (2011)
13. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: CVPR. (2010)
14. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. (2012) 2887 –2894
15. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing **23** (2005) 1080–1093
16. Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: Robust and efficient parametric face alignment. In: ICCV. (2011)
17. Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: Subspace learning from image gradient orientations. IEEE TPAMI (to appear)
18. Zhu, X., , Ramanan, D.: Face detection, pose estimation, and landmark estimation in the wild. In: CVPR. (2012)
19. Cootes, T., Taylor, C.: On representing edge structure for model matching. In: CVPR. (2001)
20. De La Torre, F., Black, M.: A framework for robust subspace learning. IJCV **54** (2003) 117–142
21. Baker, S., Gross, R., Matthews, I.: Lucas-kanade 20 years on: Part 3. Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-03-35 (2003)
22. Evangelidis, G., Psarakis, E.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE TPAMI **30** (2008) 1858–1865
23. Navarathna, R., Sridharan, S., Lucey, S.: Fourier active appearance models. In: ICCV. (2011)
24. Basso, C., Vetter, T., Blanz, V.: Regularized 3d morphable models. In: Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. First IEEE International Workshop on. (2003)
25. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. In: CVPR. (2004)
26. Ramnath, K., Baker, S., Matthews, I., Ramanan, D.: Increasing the density of active appearance models. In: CVPR. (2008)
27. Patel, A., Smith, W.: 3d morphable face models revisited. In: CVPR. (2009)
28. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: 2nd international conference on audio and video-based biometric person authentication. Volume 964. (1999) 965–966
29. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28** (2010) 807–813