Discriminative Bayesian Active Shape Models

Pedro Martins, Rui Caseiro, João F. Henriques, and Jorge Batista

Institute of Systems and Robotics, University of Coimbra, Portugal
{pedromartins,ruicaseiro,henriques,batista}@isr.uc.pt

Abstract. This work presents a simple and very efficient solution to align facial parts in unseen images. Our solution relies on a Point Distribution Model (PDM) face model and a set of discriminant local detectors, one for each facial landmark. The patch responses can be embedded into a Bayesian inference problem, where the posterior distribution of the global warp is inferred in a *maximum a posteriori* (MAP) sense. However, previous formulations do not model explicitly the covariance of the latent variables, which represents the confidence in the current solution. In our Discriminative Bayesian Active Shape Model (DBASM) formulation, the MAP global alignment is inferred by a Linear Dynamical System (LDS) that takes this information into account. The Bayesian paradigm provides an effective fitting strategy, since it combines in the same framework both the shape prior and multiple sets of patch alignment classifiers to further improve the accuracy. Extensive evaluations were performed on several datasets including the challenging Labeled Faces in the Wild (LFW). Face parts descriptors were also evaluated, including the recently proposed Minimum Output Sum of Squared Error (MOSSE) filter. The proposed Bayesian optimization strategy improves on the state-of-the-art while using the same local detectors. We also show that MOSSE filters further improve on these results.

1 Introduction

Deformable model fitting aims to find the parameters of a Point Distribution Model (PDM) that best describe the object of interest in an image. Several fitting strategies have been proposed, most of which can be categorized as being either holistic (generative) or patch-based (discriminative). The holistic representations [1][2] model the appearance of all image pixels describing the object. By synthesizing the expected appearance template, a high registration accuracy can be achieved. However, such representation generalizes poorly when large amounts of variability are involved, such as the human face under variations of identity, expression, pose, lighting or non-rigid motion, due to the huge dimensional representation of the appearance (learn from limited data).

Recently, discriminative-based methods, such as the Constrained Local Model (CLM) [4][5][6][7][8], have been proposed. These approaches can improve the model's representation power, as it accounts only for local correlations between pixel values. In this paradigm, both shape and appearance are combined by constraining an ensemble of local feature detectors to lie within the subspace

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 57–70, 2012.

Fig. 1. Examples of the DBASM global alignment on LFW dataset [3]

spanned by the PDM. The CLM implements a two step fitting strategy: a local search and global optimization. The first step performs an exhaustive local search using a feature detector, obtaining response maps for each landmark. Then, the global optimization finds the PDM parameters that jointly maximize the detection responses. Each landmark detector generates a likelihood map by applying local detectors to the neighborhood regions around the current estimate.

Some of the most popular optimization strategies propose to replace the true response maps by simple parametric forms (Weighted Peak Responses [4], Gaussians Responses [8], Mixture of Gaussians [9]) and perform the global optimization over these forms instead of the original response maps. The detectors are learned from training images of each of the object's landmarks. However, due to their small local support and large appearance variation, they can suffer from detection ambiguities. In [10] the authors attempt to deal with these ambiguities by nonparametrically approximating the response maps using the mean-shift algorithm, constrained to the PDM subspace (Subspace Constrained Mean-Shift - SCMS). However, in the SCMS global optimization the PDM parameters update is essentially a regularized projection of the mean-shift vector for each landmark onto the subspace of plausible shape variations. Since a least squares projection is used, the optimization is very sensitive to outliers (when the mean-shift output is very far away from the correct landmark location). The patch responses can be embedded into a Bayesian inference problem, where the posterior distribution of the global warp can be inferred in a *maximum a posteriori* (MAP) sense. The Bayesian paradigm provides an effective fitting strategy, since it combines in the same framework both the shape prior (the PDM) and multiple sets of patch alignment classifiers to further improve the accuracy.

1.1 Main Contributions

1. We present a novel and efficient Bayesian formulation to solve the MAP global alignment problem (Discriminative Bayesian Active Shape Model - DBASM). The main advantage of the proposed DBASM with respect to the previous Bayesian formulations is that we model the covariance of the latent variables, which represents the confidence in the current parameters estimate i.e. DBASM explicitly maintains 2^{nd} order statistics of the shape and pose

parameters, instead of assuming them to be constant. We show that the posterior distribution of the global warp can be efficiently inferred using a Linear Dynamical System (LDS) taking this information into account.
2. We aim to prove that solving the PDM using a Bayesian paradigm offers superior performance versus the traditional first order forwards additive update [4][8][9][10]. We confirm experimentally that the MAP parameter update outperforms the standard optimization strategies, based on maximum likelihood solutions (least squares). See figures 5 and 6.
3. We present a comparison between several face parts descriptors, including the recently proposed Minimum Output Sum of Squared Error (MOSSE) filters [11]. The MOSSE maps aligned training patch examples to a desired output, producing correlation filters that are notably stable. These filters exhibit a high invariance to illumination, due to their null DC component. Results show that the MOSSE outperforms the others detectors, being particularly well-suited to the task of generic face alignment (figures 2 and 4).

The remaining paper is organized as follows: **Section** 2 briefly explains the shape model PDM. **Section** 3 presents our Bayesian global optimization approach. Experimental results comparing the fitting performances of several local detectors (including the MOSSE filters) and several global optimizations strategies are shown in **Section** 4. Finally, **Section** 5 provides some conclusions.

2 The Shape Model – PDM

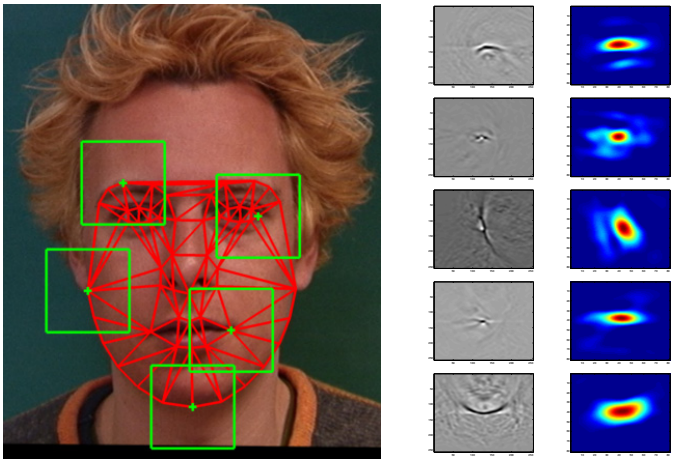
The shape $\mathbf{s}$ of a Point Distribution Model (PDM) is represented by the 2D vertex locations of a mesh, with a $2v$ dimensional vector $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$. The traditional way of building a PDM requires a set of shape annotated images that are previously aligned in scale, rotation and translation by Procrustes Analysis. Applying a PCA to a set of aligned training examples, the shape can be expressed by the linear parametric model

$$\mathbf{s} = \mathbf{s}_0 + \Phi \mathbf{b}_s + \Psi \mathbf{q} \tag{1}$$

where $\mathbf{s}_0$ is the mean shape (also referred to as the base mesh), Φ is the shape subspace matrix holding n eigenvectors (retaining a user defined variance, e.g. 95%), $\mathbf{b}_s$ is a vector of shape parameters, $\mathbf{q}$ contains the similarity pose parameters and Ψ is a matrix holding four special eigenvectors [2] that linearly model the 2D pose. From the probabilistic point of view, $\mathbf{b}_s$ follows a multivariate Gaussian distribution $\mathbf{b}_s \propto \mathcal{N}(\mathbf{b}_s|\mathbf{0}, \Lambda)$, with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where λ_i denotes the PCA eigenvalue of the i^{th} mode of deformation.

3 Global PDM Optimization – DBASM

This section describes the proposed global optimization method (Discriminative Bayesian Active Shape Models -DBASM). The deformable model fitting goal (that follows the parametric form eq.1) is formulated as a global shape alignment problem in a *maximum a posteriori* (MAP) sense.



(a) Example image and search regions. (b) $\mathcal{F}^{-1}\{\mathbf{H}_i^*\}$ (c) $p_i(\mathbf{z}_i)$

Fig. 2. The DBASM combines a Point Distribution Model (PDM) and a set of discriminant local detectors, one for each landmark. a) Image with the current mesh showing the search region for some landmarks. b) The local detector (the MOSSE filter [11] itself). c) Response maps for the correspondent highlighted landmarks. The DBASM global optimization jointly combines all landmark response maps, in a MAP sense, using 2^{nd} order statistics of the shape and pose parameters.

3.1 MAP Formulation

Given a $2v$ vector of observed positions $\mathbf{y}$, the goal is to find the optimal set of parameters $\mathbf{b}_s^*$ that maximizes the posterior probability of being its true position. Using a Bayesian approach, the optimal shape parameters are defined as

$$\mathbf{b}_s^* = \arg\max_{\mathbf{b}_s} p(\mathbf{b}_s|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b}_s)p(\mathbf{b}_s) \tag{2}$$

where $\mathbf{y}$ is the observed shape, $p(\mathbf{y}|\mathbf{b}_s)$ is the likelihood term and $p(\mathbf{b}_s)$ is a prior distribution over all possible configurations. The section 3.2 describe some possible strategies to set the observed shape vector $\mathbf{y}$.

The complexity of the problem, in eq.2, can be reduced by making some simple assumptions. Firstly, conditional independence between landmarks can be assumed simply by sampling each landmark independently. Secondly, it can also be considered that we have an approximate solution to the true parameters ($\mathbf{b} \approx \mathbf{b}_s^*$). Combining these approximations, the eq.2 can be rewritten as

$$p(\mathbf{b}|\mathbf{y}) \propto \left(\prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b}) \right) p(\mathbf{b}|\mathbf{b}_{k-1}^*) \tag{3}$$

where $\mathbf{y}_i$ is the i^{th} landmark coordinates and $\mathbf{b}_{k-1}^*$ is the previous optimal estimate of $\mathbf{b}$.

The likelihood term, including the PDM model (in eq.1), becomes the following convex energy function:

$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\underbrace{\mathbf{y}-(\mathbf{s}_0}_{\Delta\mathbf{y}}+\Phi\mathbf{b}))^T\Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-(\mathbf{s}_0+\Phi\mathbf{b}))\right) \tag{4}$$

where $\Delta\mathbf{y}$ is the difference between the observed and the mean shape and $\Sigma_{\mathbf{y}}$ is the uncertainty of the spatial localization of the landmarks ($2v \times 2v$ block diagonal covariance matrix). From the probabilistic point of view, the likelihood term follows a Gaussian distribution given by

$$p(\mathbf{y}|\mathbf{b}) \propto \mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_{\mathbf{y}}). \tag{5}$$

The prior term, according to the approximations taken, can be written as

$$p(\mathbf{b}_k|\mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \tag{6}$$

where $\mu_{\mathbf{b}} = \mathbf{b}_{k-1}$ and $\Sigma_{\mathbf{b}} = \Lambda + \Xi$. The Λ is the shape parameters covariance (diagonal matrix with PCA eigenvalues) and Ξ is an additive dynamic noise covariance (that can be estimated offline).

An important property of Bayesian inference is that, when the likelihood and the prior are Gaussian distributions the posterior is also Gaussian [12]. Following the Bayes' theorem for Gaussian variables, and considering $p(\mathbf{b}_k|\mathbf{b}_{k-1})$ a prior Gaussian distribution for $\mathbf{b}_k$ and $p(\mathbf{y}|\mathbf{b}_k)$ a likelihood Gaussian distribution, the posterior distribution takes the form ([12], pag 90).

$$p(\mathbf{b}_k|\mathbf{y}) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7}$$

$$\boldsymbol{\Sigma} = (\Sigma_{\mathbf{b}}^{-1} + \Phi^T\Sigma_{\mathbf{y}}^{-1}\Phi)^{-1} \tag{8}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}(\Phi^T\Sigma_{\mathbf{y}}^{-1}\mathbf{y} + \Sigma_{\mathbf{b}}^{-1}\mu_{\mathbf{b}}). \tag{9}$$

Note that, the conditional distribution $p(\mathbf{y}|\mathbf{b}_k)$ has a mean that is a linear function of $\mathbf{b}_k$ and a covariance which is independent of $\mathbf{b}_k$. This could be a possible solution to the global alignment optimization [13]. However, in practice, this is a naive approach because it does not model the covariance of the latent variables, $\mathbf{b}_k$, which is crucial to account for the confidence in the current parameters estimate.

The MAP global alignment solution can be inferred by a Linear Dynamical System (LDS). The LDS is the ideal technique to model the covariance of the latent variables and solve the naive approach limitations. The LDS is a simple approach that recursively computes the posterior probability using incoming Gaussian measurements and a linear model process, taking into account all the available measures (same requirements as our alignment problem). The state and measurement equations of the LDS, according to the PDM alignment problem, can be written as

$$\mathbf{b}_k = \mathbf{A}\mathbf{b}_{k-1} + q \tag{10}$$

$$\Delta\mathbf{y} = \Phi\mathbf{b}_k + r \tag{11}$$

where the current shape parameters $\mathbf{b}_k$ are the hidden state vector, $q \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{b})$ is the additive dynamic noise, $\Delta\mathbf{y}$ is the observed shape deviation that are related to the shape parameters by the linear relation Φ (eq.1) and r is the additive measurement noise following $r \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{y})$. The previous shape estimated parameters $\mathbf{b}_{k-1}$ are connected to the current parameters $\mathbf{b}_k$ by an identity relation plus noise ($\mathbf{A} = \mathbf{I}_n$).

We highlight that the final step of the LDS derivation consists of a Bayesian inference step [12] (using the Bayes' theorem for Gaussian variables), where the likelihood term is given by eq.5 and the prior follows $\mathcal{N}(\mathbf{A}\mu^\mathbf{F}_{k-1}, \mathbf{P}_{k-1})$ where

$$\mathbf{P}_{k-1} = (\Lambda + \Xi) + \mathbf{A}\boldsymbol{\Sigma}^\mathbf{F}_{k-1}\mathbf{A}^T. \tag{12}$$

From these equations we can see that the LDS keep up to date the uncertainty on the current estimate of the shape parameters. The LDS recursively computes the mean and covariance of the posterior distributions of the form

$$p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}^\mathbf{F}_k, \boldsymbol{\Sigma}^\mathbf{F}_k) \tag{13}$$

with the posterior mean $\boldsymbol{\mu}^\mathbf{F}_k$ and covariance $\boldsymbol{\Sigma}^\mathbf{F}_k$ given by the LDS formulas:

$$\mathbf{K} = \mathbf{P}_{k-1}\Phi^T(\Phi\mathbf{P}_{k-1}\Phi^T + \Sigma_\mathbf{y})^{-1} \tag{14}$$

$$\boldsymbol{\mu}^\mathbf{F}_k = \mathbf{A}\boldsymbol{\mu}^\mathbf{F}_{k-1} + \mathbf{K}(\mathbf{y} - \Phi\mathbf{A}\boldsymbol{\mu}^\mathbf{F}_{k-1}), \quad \boldsymbol{\Sigma}^\mathbf{F}_k = (\mathbf{I}_n - \mathbf{K}\Phi)\mathbf{P}_{k-1}. \tag{15}$$

Finally, the optimal shape parameters that maximize eq.2 are given by $\boldsymbol{\mu}^\mathbf{F}_k$. In order to estimate the pose parameters, we also apply the LDS paradigm. The difference is that, in this case, the state vector is given by $\mathbf{q}$ and the observation matrix is Ψ. The algorithm 1 summarizes the proposed DBASM global optimization.

3.2 Local Optimization Strategies

This section briefly describes several local strategies to represent the true response maps by a probabilistic model (parametric and nonparametric). We also describe how to extract from each probabilistic model the likelihood term of the MAP formulation (observed shape $\mathbf{y}$ and the uncertainty covariance $\Sigma_\mathbf{y}$).

Let $\mathbf{z}_i = (x_i, y_i)$ be a candidate to the i^{th} landmark, being $\mathbf{y}^c_i$ the current landmark estimate, $\boldsymbol{\Omega}_{\mathbf{y}^c_i}$ a $L \times L$ patch centered at $\mathbf{y}^c_i$, a_i a binary variable that denotes correct landmark alignment, $\mathcal{D}_i$ the score of a generic local detector and $\mathbf{I}$ the target image up to a similarity transformation (typically the detector is designed to operate at a given scale). The probability of pixel $\mathbf{z}_i$ to be aligned is given by

$$p_i(\mathbf{z}_i) = p(a_i = 1|\mathbf{I}(\mathbf{z}_i), \mathcal{D}_i) = \frac{1}{1 + e^{-a_i\mathcal{D}_i(\mathbf{I}(\mathbf{z}_i))}} \tag{16}$$

where the detector score is converted to probability using the logistic function.

The parameters $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$ can be found by minimizing the expression [13]

$$\arg\min_{\mathbf{y}_i, \Sigma_{\mathbf{y}_i}} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}^c_i}} p_i(\mathbf{z}_i)\mathcal{N}(\mathbf{z}_i|\mathbf{y}_i, \Sigma_{\mathbf{y}_i}) \tag{17}$$

where several strategies can be used to do this optimization.

Weighted Peak Response (WPR): The simplest solution is to take the spatial location where the response map has a higher score [4]. The new landmark position is then weighted by a factor that reflects the peak confidence. Formally, the WPR solution is given by

$$\mathbf{y}_i^{\mathrm{WPR}} = \max_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} \left(p_i(\mathbf{z}_i)\right), \quad \Sigma_{\mathbf{y}_i}^{\mathrm{WPR}} = diag(p_i(\mathbf{y}_i^{\mathrm{WPR}})^{-1}) \tag{18}$$

that is equivalent to approximate each response map by an isotropic Gaussian $\mathcal{N}(\mathbf{z}_i|\mathbf{y}_i^{\mathrm{WPR}}, \Sigma_{\mathbf{y}_i}^{\mathrm{WPR}})$.

Gaussian Response (GR): The previous approach was extended in [8] to approximate the response maps by a full Gaussian distribution $\mathcal{N}(\mathbf{z}_i|\mathbf{y}_i^{\mathrm{GR}}, \Sigma_{\mathbf{y}_i}^{\mathrm{GR}})$. This is equivalent to fit a Gaussian density to weighted data.

Let $d = \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)$, the solution is given by

$$\mathbf{y}_i^{\mathrm{GR}} = \frac{1}{d} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\mathbf{z}_i, \quad \Sigma_{\mathbf{y}_i}^{\mathrm{GR}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})^T. \tag{19}$$

Kernel Density Estimator (KDE): The response maps can also be approximated by a nonparametric representation, namely using a Kernel Density Estimator (KDE) (isotropic Gaussian kernel with a bandwidth σ_h^2). Maximizing over the KDE is typically performed by using the well-known mean-shift algorithm [10]. The kernel bandwidth σ_h^2 is a free parameter that exhibits a strong influence on the resulting estimate. This problem can be addressed by an annealing bandwidth schedule [14]. It can be shown that there exists a σ_h^2 value such that the KDE is unimodal. As σ_h^2 is reduced, the modes divide and the smoothness of KDE decreases, guiding the optimization towards the true objective. Formally, the i^{th} annealed mean-shift landmark update is given by

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)} \tag{20}$$

where $\mathbf{I}_2$ is a two-dimensional identity matrix and $\sigma_{h_j}^2$ represents the decreasing annealed bandwidth. The KDE uncertainty error consists on computing the weighted covariance using the mean-shift results as mean

$$\Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T. \tag{21}$$

Figure 3 highlights the differences between the three local optimization strategies (WPR, GR and KDE). Notice that DBASM deals with mild occlusions. When a landmark is under occlusion typically the response map is multi-modal. If a KDE local strategy is used (DBASM-KDE), the landmark update will select the nearest mode (eq.20) and the covariance of that landmark (eq.21) will

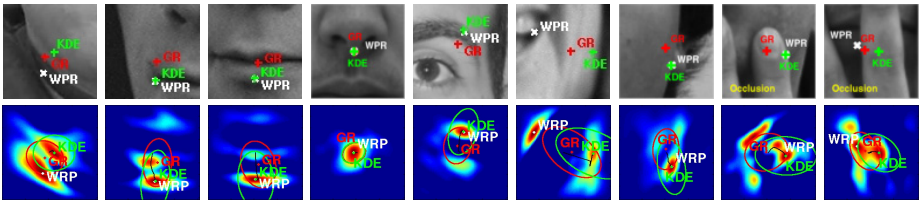


Fig. 3. Qualitative comparison between the three local optimization strategies. The WPR simply chooses the maximum detector response. GR approximates the response map by a full Gaussian distribution. KDE uses the mean-shift algorithm to move to the nearest mode of the density. Its uncertainty is centered at the found mode. The two examples in the right show patches under occlusion (typically multimodal responses).

be inherently large, modeling a high localization uncertainty. Then, the global optimization stage jointly combines all uncertainties (MAP sense) handling occlusions. Similarly, to deal with large occlusions, a minor tweak is required. One can simply set a large covariance for the occluded landmarks.

Precompute:
The parametric models ($\mathbf{s}_0$, Φ, Ψ) and the MOSSE filters in the Fourier domain $\mathbf{H}_i^*$
Initial estimate of the shape/pose parameters and covariances $(\mathbf{b}_0, \mathbf{P}_0)$ / $(\mathbf{q}_0, \mathbf{Q}_0)$.
repeat
  Warp image $\mathbf{I}$ to the base mesh using the current pose parameters $\mathbf{q}$ [0.5ms]
  Generate current shape $\mathbf{s} = \mathbf{s}_0 + \Phi\mathbf{b} + \Psi\mathbf{q}$
  for *Landmark* $i = 1$ **to** v **do**
    Evaluate the detectors response (MOSSE correlation $\mathcal{F}^{-1}\{\mathcal{F}\{(\mathbf{I})\} \odot \mathbf{H}_i^*\}$) [3ms]
    Find $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$ using a local strategy (sec. 3.2),e.g. if using KDE, eqs.20 and 21, respectively.
  end
  Update the pose parameters and their covariance [0.1ms]:
    $\mathbf{Q}_{k-1} = (\Lambda_q + \Xi_q + \mathbf{Q}_{k-1}), \quad \mathbf{K}_q = \mathbf{Q}_{k-1}\Psi^T(\Psi\mathbf{Q}_{k-1}\Psi^T + \Sigma_\mathbf{y})^{-1}$
    $\mathbf{q}_k = \mathbf{q}_{k-1} + \mathbf{K}_q(\mathbf{y} - \Psi\mathbf{q}_{k-1}), \quad \mathbf{Q}_k = (\mathbf{I}_4 - \mathbf{K}_q\Psi)\mathbf{Q}_{k-1}$
  Update the shape parameters (with pose correction) and their covariance [0.2ms]:
    $\mathbf{P}_{k-1} = (\Lambda + \Xi + \mathbf{P}_{k-1}), \qquad \mathbf{K}_b = \mathbf{P}_{k-1}\Phi^T(\Phi\mathbf{P}_{k-1}\Phi^T + \Sigma_\mathbf{y})^{-1}$
    $\mathbf{b}_k = \mathbf{b}_{k-1} + \mathbf{K}_b(\mathbf{y} - \Phi\mathbf{b}_{k-1} - \Psi\mathbf{q}_k), \quad \mathbf{P}_k = (\mathbf{I}_n - \mathbf{K}_b\Phi)\mathbf{P}_{k-1}$
until $||b_k - b_{k-1}|| \leq \varepsilon$ *or maximum number of iterations reached* ;

Algorithm 1: Overview of the DBASM method. The performance of DBASM is comparable to ASM [4], CQF [8] or SCMS [10] depending of the local strategy DBASM-WPR, DBASM-GR or DBASM-KDE, respectively. It achieves near real-time performance. The bottleneck is always obtaining the response maps (3ms x number landmarks), although it can be done in parallel.

3.3 Hierarchical Search (DBASM-KDE-H)

A slightly different annealing approach is proposed in this section. Instead of using the mean-shift with an iterative kernel bandwidth relaxation and then optimize using the LDS MAP formulation, a hierarchical search can be used instead. This solution is composed by multiple levels of fixed kernel bandwidth mean-shifts followed by LDS optimization steps. The annealing is performed between hierarchical levels.

4 Evaluation Results

The experiments in this paper were designed to evaluate the local detector (MOSSE [11]) and the new Bayesian global optimization (DBASM). All the experiments were conducted on several databases with publicly available ground truth. **(1)** The IMM [15] database that consists on 240 annotated images of 40 different human faces presenting different head pose, illumination, and facial expression (58 landmarks). **(2)** The BioID [16] dataset contains 1521 images, each showing a near frontal view of a face of one of 23 different subjects (20 landmarks). **(3)** The XM2VTS [17] database has 2360 images frontal faces of 295 subjects (68 landmarks). **(4)** The tracking performance is evaluated on the FGNet Talking Face (TF) [18] video sequence that holds 5000 frames of video of an individual engaged in a conversation (68 landmarks). **(5)** Finally, a qualitative evaluation was also performed using the Labeled Faces in the Wild (LFW) [3] database that contains images taken under variability in pose, lighting, focus, facial expression, occlusions, different backgrounds, etc.

4.1 Local Detector – MOSSE Filter

The Minimum Output Sum of Squared Error (MOSSE) filter, recently proposed in [11], finds the optimal filter that minimizes the Sum of Squared Differences (SSD) to a desired correlation output. Briefly, correlation can be computed in the frequency domain as the element-wise multiplication of the 2D Fourier transform ($\mathcal{F}$) of an input image $\mathbf{I}$ with a filter $\mathbf{H}$, also defined in the Fourier domain as $\mathbf{G} = \mathcal{F}\{\mathbf{I}\} \odot \mathbf{H}^*$, where the $\odot$ symbol represents the Hadamard product and $*$ is the complex conjugate. The correlation value is given by $\mathcal{F}^{-1}\{\mathbf{G}\}$, the inverse Fourier transform of $\mathbf{G}$.

MOSSE finds the filter $\mathbf{H}$, in the Fourier domain, that minimizes the SSD between the actual output of the correlation and the desired output of the correlation, across a set of N training images, $\min_{\mathbf{H}^*} \sum_{j=1}^{N} \left(\mathcal{F}\{\mathbf{I}_j\} \odot \mathbf{H}^* - \mathbf{G}_j\right)^2$, where $\mathbf{G}$ is obtained by sampling a 2D Gaussian uniformly. Solving for the filter $\mathbf{H}^*$ yields the closed form solution

$$\mathbf{H}^* = \frac{\sum_{j=1}^{N} \mathbf{G}_j \odot \mathcal{F}\{\mathbf{I}_j\}^*}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^*}. \tag{22}$$

The MOSSE filter maps all aligned training patch examples to an output, $\mathbf{G}$, centered at the feature location, producing notably stable correlation filters.

At the training stage, each patch example is normalized to have zero mean and a unitary norm, and is multiplied by a cosine window (required to solve the Fourier Transform periodicity problem). This also has the benefit of emphasizing the target center. These filters have a high invariance to illumination changes, due to their null DC component and revealed to be highly suitable to the task of generic face alignment.

4.2 Evaluating Local Detectors

Three landmark expert detectors were evaluated. The most used detector [8][10] is based on a linear classifier built from aligned (positive) and misaligned (negative) grey level patch examples. The score of the i^{th} linear detector is given by

$$\mathcal{D}_i^{\text{linear}}(\mathbf{I}(\mathbf{y}_i)) = \mathbf{w}_i^T \mathbf{I}(\mathbf{y}_i) + b_i, \tag{23}$$

with $\mathbf{w}_i$ being the linear weight, b_i the bias constant and $\mathbf{I}(\mathbf{y}_i)$ a vectorized patch of pixel values sampled at $\mathbf{y}_i$. Similarly, a quadratic classifier can be used

$$\mathcal{D}_i^{\text{quadratic}}(\mathbf{I}(\mathbf{y}_i)) = \mathbf{I}(\mathbf{y}_i)^T \mathbf{Q}_i \mathbf{I}(\mathbf{y}_i) + \mathbf{L}_i^T \mathbf{I}(\mathbf{y}_i) + b_i \tag{24}$$

with $\mathbf{Q}_i$ and $\mathbf{L}_i$ being the quadratic and linear terms, respectively. Finally, the MOSSE filter correlation gives

$$\mathcal{D}_i^{\text{MOSSE}}(\mathbf{I}(\mathbf{y}_i)) = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{I}(\mathbf{y}_i)\} \odot \mathbf{H}_i^*\} \tag{25}$$

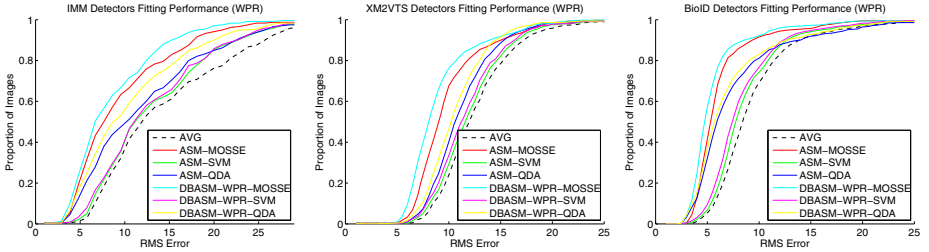
where $\mathbf{H}_i^*$ is the MOSSE filter from eq.22. Both linear and quadratic classifiers (linear-SVM [19] and Quadratic Discriminant Analysis) were trained using images from the IMM [15] dataset with 144 negative patch examples (for each landmark and each image) being misaligned up to 12 pixels in x and y translation.

The MOSSE filters were built using aligned patch samples with size 128×128. A power of two patch size is used to speed up the FFT computation, however only a 40×40 subwindow of the output is considered. During the MOSSE filter building, each training patch requires a normalization step. Each example is normalized to have a zero mean and a unitary norm and is multiplied by a cosine window. The desired output $\mathbf{G}$ (eq.22) is set to be a 2D Gaussian function centered at the landmark with 3 pixels of standard deviation.

The global optimization method that best evaluates the detectors performance is the approach that relies the most on the output of the detector, i.e., the Active Shape Models (ASM) [4]. The results are present in the form of fitting performance curves, which were also adopted by [20][5][21][8][10]. These curves show the percentage of faces that achieved a given Root Mean Square (RMS) error amount. The figure 4 shows fitting performance curves that compare the three kinds of detectors using the ASM [4] optimization[1] and the proposed global DBASM technique using a Weighted Peak Response strategy (DBASM-WPR). From the results we can highlight some conclusions: (1) the MOSSE filter always outperforms the others, specially when using simpler optimization methods; (2) the DBASM optimization improves the results even with simple detectors; (3) maximum performance can be achieved by using the MOSSE detector and the DBASM optimization.

The use of MOSSE filters is an interesting solution that works well in practice and is particularly suited to detection of facial parts. However we highlight that is not crucial for the performance of our Bayesian formulation. DBASM still improves performance when using standard detectors.

[1] The ASM [4], CQF [8] and SCMS [10] use as local optimizations the WPR, GR and KDE strategies, respectively.



(a) IMM [15] database (b) XM2VTS [17] database (c) BioID [16] database

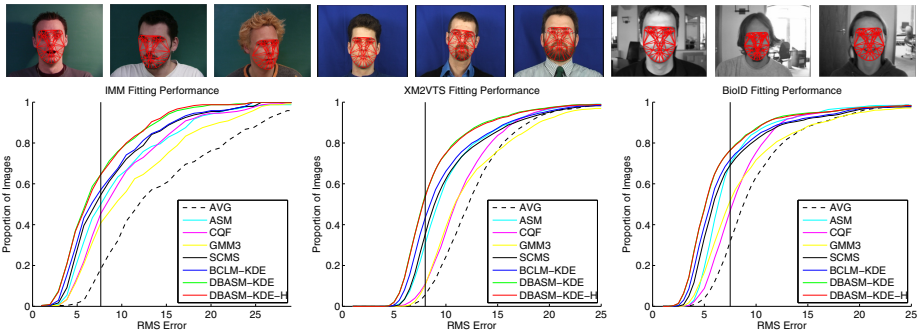
Fig. 4. Fitting performance curves comparing different detectors (Linear, Quadratic and MOSSE) on the IMM, XM2VTS and BioID database, respectively. The AVG means the average location provided by the initial estimate (Adaboost [22] face detector).

4.3 Evaluating Global Optimization Strategies

In this section the DBASM optimization strategy is evaluated w.r.t. state-of-the-art global alignment solutions. The proposed DBASM and DBASM-H methods are compared with (1) ASM [4], (2) CQF [8], (3) BCLM [13], (4) GMM [9] using 3 Gaussians (GMM3) and (5) SCMS [10]. Note that the DBASM can be used with different local strategies to approximate the response maps (e.g. WPR, GR or KDE as described in section 3.2). In these experiments we fixed the local strategy as a KDE (BCLM-KDE, SCMS-KDE, DBASM-KDE) in order to compare the global optimization approaches. The results from ASM, CQF and GMM3 are provided as a baseline. The same bandwidth schedule of $\sigma_h^2 = (15, 10, 5, 2)$ is always used for KDE. All the experiments, in this section, use MOSSE filters as local detectors (using the same settings as in section 4.2) built with only training images from the IMM [15] set and tested on the remaining datasets[2]. In all cases, the nonrigid parameters start from zero, the similarity parameters were initialized by a face detection [22] and the model was fitted until convergence (limited to a maximum of 20 iterations).

Figure 5 shows the fitting performance curves for the IMM, XM2VTS and BioID datasets, respectively. The CQF performs better than GMM3, mainly because GMM is very prone to local optimums due to its multimodal nature (it is worth mentioning that given a good initial estimate GMM offers a superior fitting quality). The main drawback of CQF is the limited accuracy due to the over-smoothness of the response map (see figure 3). The BCLM is slightly better than SCMS due to its improved parameter update (MAP update vs first order forwards additive). The SCMS improves the results when compared to CQF due to the high accuracy provided by the mean-shift. In some cases, the ASM achieves a comparable performance to the SCMS; the reason for this relies on the excellent performance of the MOSSE detector. The proposed Bayesian global optimization (DBASM) outperforms all previous methods, by modeling the covariance of the latent variables which represent the confidence in the

[2] The results on the IMM dataset use training images collected at our institution.



(a) IMM [15] database (b) XM2VTS [17] database (c) BioID [16] database

Reference 7.5 RMS	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)	
ASM	50.0		30.7		70.0	
DBASM-WPR* (our method)	**56.7**	(+6.7)	**45.1**	(+14.4)	**75.4**	(+5.4)
CQF	45.4		10.9		47.0	
GMM3	40.8	(-4.6)	10.4	(-0.5)	51.7	(+4.7)
BCLM-GR*	48.3	(+2.9)	15.9	(+5.0)	54.2	(+7.2)
DBASM-GR* (our method)	**50.4**	(+5.0)	**18.0**	(+7.1)	**62.2**	(+15.2)
SCMS-KDE	54.6		35.7		69.0	
BCLM-KDE	57.1	(+2.5)	43.4	(+7.7)	71.9	(+2.9)
DBASM-KDE (our method)	**64.6**	(+10.0)	**54.5**	(+18.8)	**76.5**	(+7.5)
DBASM-KDE-H (our method)	**64.6**	(+10.0)	53.5	(+17.8)	**76.5**	(+7.5)

Fig. 5. Fitting performance curves. The table shows quantitative values taken by setting a fixed RMS error amount (7.5 pixels - vertical line in the graphics). Each table entry show how many percentage of images converge with less or equal RMS error than the reference. The results show that our proposed methods outperform all the other (using all the local strategies WPR, GR and KDE). Top images show DBASM-KDE fitting examples from each database.

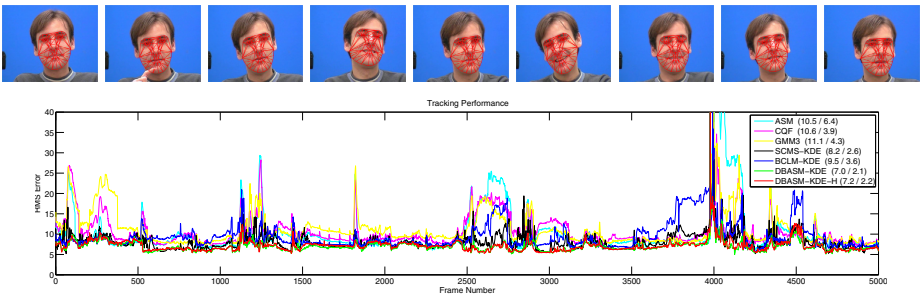
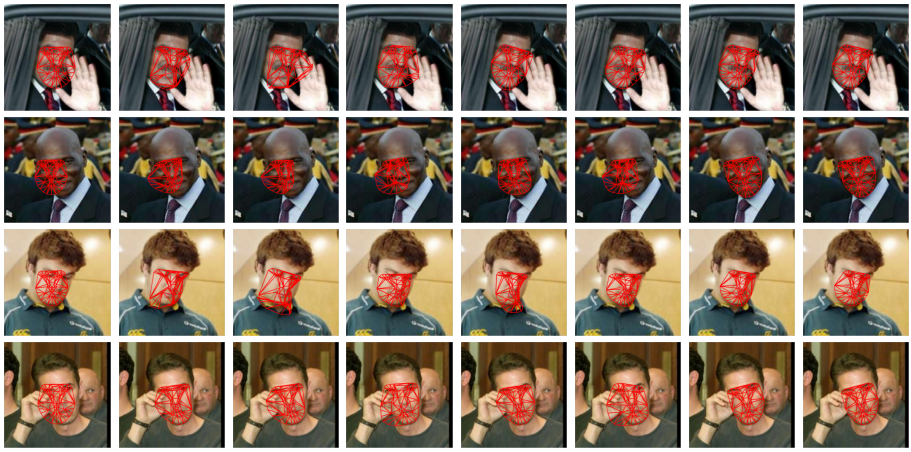


Fig. 6. Tracking performance evaluation of several fitting algorithms on the FGNET Talking Face [18] sequence. The values on legend box are the mean and standard deviation RMS errors, respectively. Top images show DBASM-KDE fitting examples.

current parameters estimate (see figure 5). The results show that the hierarchical annealing version of DBASM-KDE (DBASM-KDE-H) performs slightly better, but at the cost of more iterations. Tracking performance is also tested on the FGNET Talking Face video sequence (figure 6). Each frame is fitted using as



(a) AVG (b) ASM (c) CQF (d) GMM3 (e) BCLM - KDE (f) SCMS - KDE (g) DBASM -KDE (h) DBASM -KDE-H

Fig. 7. Qualitative fitting results on LFW database [3]. AVG is the initial estimate.

initial estimate the previously estimated shape and pose parameters. The relative performance between the global optimization approaches is similar to the previous experiments, where the DBASM technique yields the best performance. Qualitative evaluation is also performed using the Labeled Faces in the Wild (LFW) database [3], where some results can be seen on figure 7.

5 Conclusions

An efficient solution to align facial parts in unseen images is described in this work. We present a novel Bayesian paradigm (DBASM) to solve the global alignment problem, in a MAP sense, showing that the posterior distribution of the global warp can be efficiently inferred using a Linear Dynamical System (LDS). The main advantage w.r.t. previous Bayesian formulations is that DBASM model the covariance of the latent variables which represent the confidence in the current parameters estimate. Several performance evaluation results are presented, comparing both local detectors and global optimization strategies. Evaluating the local detectors show that the MOSSE correlation filters offer a superior performance in landmark local detection. Global optimizations evaluation were performed in several image publicly available datasets, namely, the IMM, the XM2VTS, the BioID, and the LFW. Tracking performance is also evaluated on the FGNET Talking Face video sequence. This new Bayesian paradigm is shown to significantly outperform other state-of-the-art fitting solutions.

Acknowledgments. This work was supported by the Portuguese Science Foundation (FCT) by the project "Dinâmica Facial 4D para Reconhecimento de

Identidade"(grant PTDC/EIA-CCO/108791/2008). Pedro Martins, Rui Caseiro and João Henriques acknowledge the FCT through the grants SFRH/BD/45178/2008, SFRH/BD74152/2010 and SFRH/BD/75459/2010, respectively.

References

1. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE TPAMI 23, 681–685 (2001)
2. Matthews, I., Baker, S.: Active appearance models revisited. IJCV 60, 135–164 (2004)
3. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
4. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. CVIU 61, 38–59 (1995)
5. Cristinacce, D., Cootes, T.F.: Boosted regression active shape models. In: BMVC (2007)
6. Tresadern, P., Bhaskar, H., Adeshina, S., Taylor, C., Cootes, T.F.: Combining local and global shape models for deformable object matching. In: BMVC (2009)
7. Cristinacce, D., Cootes, T.F.: Automatic feature localisation with constrained local models. Pattern Recognition 41, 3054–3067 (2008)
8. Wang, Y., Lucey, S., Cohn, J.: Enforcing convexity for improved alignment with constrained local models. In: IEEE CVPR (2008)
9. Gu, L., Kanade, T.: A Generative Shape Regularization Model for Robust Face Alignment. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 413–426. Springer, Heidelberg (2008)
10. Saragih, J., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: IEEE ICCV (2009)
11. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: IEEE CVPR (2010)
12. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
13. Paquet, U.: Convexity and bayesian constrained local models. In: IEEE CVPR (2009)
14. Shen, C., Brooks, M.J., Hengel, A.: Fast global kernel density mode seeking: Applications to localization and tracking. IEEE TIP 16, 1457–1469 (2007)
15. Nordstrom, M., Larsen, M., Sierakowski, J., Stegmann, M.: The IMM face database - an annotated dataset of 240 face images. Technical report, Technical University of Denmark, DTU (2004)
16. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust Face Detection Using the Hausdorff Distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
17. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended M2VTS database. In: AVBPA (1999)
18. FGNet: Talking face video (2004)
19. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. JMLR, 1871–1874 (2008)
20. Cristinacce, D., Cootes, T.F.: Facial feature detection using adaboost with shape constraints. In: BMVC (2003)
21. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: BMVC (2006)
22. Viola, P., Jones, M.: Robust real-time object detection. IJCV 57, 137–154 (2002)