Let the Shape Speak - Discriminative Face Alignment using Conjugate Priors

Pedro Martins
pedromartins@isr.uc.pt
Rui Caseiro
ruicaseiro@isr.uc.pt
João F. Henriques
henriques@isr.uc.pt
Jorge Batista
batista@isr.uc.pt

Institute of Systems and Robotics
University of Coimbra, Portugal

Abstract

This work presents a novel Bayesian formulation for aligning faces in unseen images. Our approach is closely related to Constrained Local Models (CLM) and Active Shape Models (ASM), where an ensemble of local feature detectors are constrained to lie within the subspace spanned by a Point Distribution Model (PDM). Fitting a model to an image typically involves two steps: a local search using a detector, obtaining response maps for each landmark (likelihood term) and a global optimization that finds the PDM parameters that jointly maximize all the detection responses. The global optimization can be seen as a Bayesian inference problem, where the posterior distribution of the PDM parameters (including pose) can be inferred in a *maximum a posteriori* (MAP) sense. Faces are nonrigid structures described by continuous dynamic transitions, so it is crucial to account for the underlying dynamics of the shape. We present a novel Bayesian global optimization strategy, where the prior is used to encode the dynamic transitions of the PDM parameters. Using recursive Bayesian estimation we model the prior distribution of the data as being Gaussian. The mean and covariance were assumed to be unknown and treated as random variables. This means that we estimate not only the mean and the covariance but also the probability distribution of the mean and the covariance (using conjugate priors). Extensive evaluations were performed on several standard datasets (IMM, BioID, XM2VTS and FGNET Talking Face) against state-of-the-art methods while using the same local detectors. Finally, qualitative results taken from the challenging Labeled Faces in the Wild (LFW) dataset are also shown.

1 Introduction

Non-rigid image registration of human faces in unconstrained environments, also known as facial alignment in the wild, is a central problem in Computer Vision with applications including tracking, motion estimation, model-based recognition (both identity and facial expression), etc. The goal of parametric deformable fitting is to find the Point Distribution Model (PDM) [19] parameters that best describe a face in a target image. Several strategies have been proposed, which can be categorized as being either holistic (generative) or

http://dx.doi.org/10.5244/C.26.130

patch-based (discriminative). The holistic representations [21][11] model the appearance of all image pixels describing a face. This representation synthesizes the expected appearance instance allowing a high registration accuracy. Although, poor performance is shown under variations of identity, expression, pose, lighting or non-rigid motion, due to the huge dimensional representation of the appearance. Typically, target individuals must be included in the training dataset otherwise the fitting quality will be very poor.

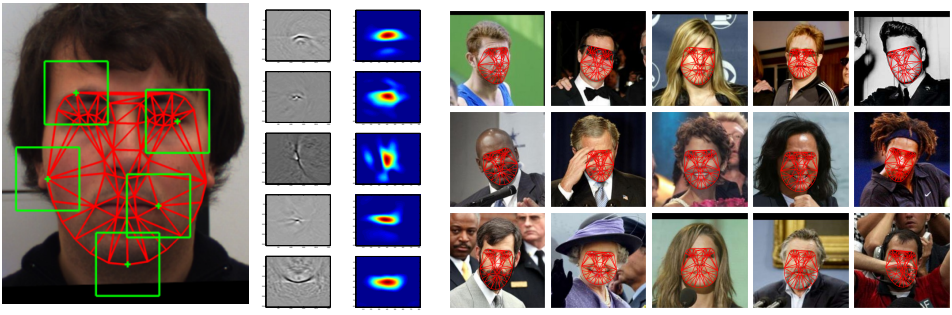
Recently, discriminative-based methods, such as the Constrained Local Model (CLM) [20][5][17][2][23], have been proposed. These approaches improve the generic face representation by accounting only for the local correlations between pixel values. Both shape and appearance are combined by constraining an ensemble of local feature detectors to lie within the subspace spanned by the PDM. The CLM implements a two step fitting strategy: a local search and a global optimization. The first step involves an exhaustive local search using an expert feature detector, obtaining response maps for each landmark (likelihood map). The second step (the global optimization) finds the PDM parameters that jointly maximize the detection responses. Most popular optimization strategies approximate the landmark response maps by simple parametric forms (Weighted Peak Responses [20], Gaussians Responses [23], Mixture of Gaussians [14]) and perform the global optimization over these forms instead of the original response maps. As the local detectors are designed to operate fast, having small local support and must cover a large appearance variation they can suffer from detection ambiguities. In SCMS [12] the authors attempt to deal with these ambiguities by nonparametrically approximating the response maps using the mean-shift algorithm. However, their global optimization is essentially a regularized projection of the mean-shift vector for each landmark onto the subspace of plausible shape variations, being 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 PDM parameters can be inferred in a *maximum a posteriori* (MAP) sense [22]. 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. Faces are nonrigid structures described by continuous dynamic transitions, so it is crucial to account for the underlying dynamics of the shape.

1.1 Main Contributions

1. We present a novel Bayesian global optimization strategy designed to infer both the PDM and the pose parameters, in a *maximum a posteriori* (MAP) sense, by explicitly modelling the prior distribution (encoding the dynamic transitions of the PDM parameters). Using recursive Bayesian estimation we model the prior distribution of the data as being Gaussian. The mean and covariance were assumed to be unknown and treated as random variables. This means that we estimate not only the mean and the covariance but also the probability distribution of the mean and the covariance (using conjugate priors).

2. We show that aligning the PDM using a Bayesian approach offers a significative increase in performance, in both fitting still images and video sequences, when compared with state-of-the-art first order forwards additive methods [20][23][12]. We confirm experimentally that the MAP parameter update outperforms the standard optimization strategies, based on maximum likelihood solutions (least squares).



(a) Local search regions. (b) $\mathcal{F}^{-1}\{\mathbf{H}_i^*\}$ (c) $p_i(\mathbf{z}_i)$ (d) Examples of the global alignment in the LFW datset [10].

Figure 1: A Point Distribution Model (PDM) is combined with an ensemble of local feature detectors. The novel Bayesian global optimization strategy jointly combines all detectors scores, in a MAP sense, by explicitly modelling the prior distribution. a) Image showing the search region for some landmarks. b) The local detector [8]. c) Detectors responses for the correspondent highlighted landmarks. d) Qualitative image alignment examples in the challenging Labeled Faces the Wild dataset [10].

3. Extensive evaluations were performed on several standard datasets (IMM [15], BioID [16], XM2VTS [13] and FGNET Talking Face [9]) against state-of-the-art methods while using the same local detectors. Qualitative results of the challenging Labeled Faces in the Wild (LFW) [10] dataset are also shown.

2 The Shape Model - PDM

The shape $\mathbf{s}$ of a Point Distribution Model (PDM) [19] 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 n dimensional vector of shape parameters, $\mathbf{q}$ is vector that contains the four similarity pose parameters and Ψ is a $2v \times 4$ matrix holding four special eigenvectors that linearly model the 2D pose [10].

3 Global PDM Optimization

We propose a global optimization method (Bayesian Active Shape Models - BASM) where 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.

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 an Bayesian

approach, the optimal shape parameters are

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

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 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 $\mathbf{b}$.

3.1 The Likelihood Term

The likelihood term, includes the PDM model (in eq.1), becoming 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 follow a Gaussian distribution given by $p(\mathbf{y}|\mathbf{b}) \propto \mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_{\mathbf{y}})$.

3.1.1 Finding the Likelihood Parameters

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 (the observed shape $\mathbf{y}$ and the landmark uncertainty covariance $\Sigma_{\mathbf{y}}$).

Let $\mathbf{z}_i = (x_i, y_i)$ be a candidate to the i^{th} landmark, being $\mathbf{y}_i^c$ the current landmark estimate, $\Omega_{\mathbf{y}_i^c}$ a $L \times L$ patch centered at $\mathbf{y}_i^c$, 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

$$p_i(\mathbf{z}_i) = p(a_i = 1|\mathbf{I}(\mathbf{z}_i), \mathcal{D}_i) = \frac{1}{1 + e^{-a_i \mathcal{D}_i(\mathbf{I}(\mathbf{z}_i))}} \tag{5}$$

where the detector score is converted to probability using the logistic function. The likelihood parameters $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$ can be found by minimizing [22]

$$\arg\min_{\mathbf{y}_i, \Sigma_{\mathbf{y}_i}} \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \mathcal{N}(\mathbf{z}_i|\mathbf{y}_i, \Sigma_{\mathbf{y}_i}) \tag{6}$$

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 [20]. The new landmark position is then weighted by a factor that reflects the peak confidence. Formally, the WPR solution is given by

$$\mathbf{y}_i^{\mathrm{WPR}} = \max_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} (p_i(\mathbf{z}_i)), \quad \Sigma_{\mathbf{y}_i}^{\mathrm{WPR}} = diag(p_i(\mathbf{y}_i^{\mathrm{WPR}})^{-1}) \tag{7}$$

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 [23] 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. Defining $d = \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)$, the solution is given by

$$\mathbf{y}_i^{\mathrm{GR}} = \frac{1}{d} \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\mathbf{z}_i, \quad \Sigma_{\mathbf{y}_i}^{\mathrm{GR}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})^T. \tag{8}$$

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 [12]. 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 [3]. 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.

The i^{th} annealed mean-shift landmark update and its uncertainty are given by

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)}|\mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}, \quad \Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T, \tag{9}$$

where $\mathbf{I}_2$ is a two-dimensional identity matrix and $\sigma_{h_j}^2$ represents the decreasing bandwidth.

3.2 The Prior Term

Faces are nonrigid structures described by continuous dynamic transitions. In the Bayesian paradigm the prior term can be used to encode the underlying dynamic of the shape. The prior term follows a Gaussian distribution with mean $\mu_{\mathbf{b}}$ and covariance $\Sigma_{\mathbf{b}}$

$$p(\mathbf{b}_k|\mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}). \tag{10}$$

Mean $\mu_{\mathbf{b}}$ and covariance $\Sigma_{\mathbf{b}}$ of the data are assumed to be unknown and modeled as random variables ([1] pag.87-88). Recursive Bayesian estimation can be applied to infer the parameters of the prior distribution in eq.10. Defining $\mathbf{b}$ as an observable vector, the Bayes theorem tells us that the joint posterior density can be written as

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b}) \propto p(\mathbf{b}|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}). \tag{11}$$

Performing recursive Bayesian estimation with new observations requires that joint prior density $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ should have the same functional form than the joint posterior density $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$. The joint prior density, conditioning on the covariance $\Sigma_{\mathbf{b}}$, can be written as

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) = p(\mu_{\mathbf{b}}|\Sigma_{\mathbf{b}}) p(\Sigma_{\mathbf{b}}). \tag{12}$$

The previous condition is true if we assume that the covariance follow an inverse-Wishart distribution and $\mu_{\mathbf{b}}|\Sigma_{\mathbf{b}}$ follow a normal distribution (the conjugate prior for a Gaussian with known mean is an inverse-Wishart distribution [11])

$$\Sigma_{\mathbf{b}} \sim \text{Inv-Wishart}_{\upsilon_{k-1}}(\Lambda_{\upsilon_{k-1}}^{-1}), \qquad \mu_{\mathbf{b}}|\Sigma_{\mathbf{b}} \sim \mathcal{N}(\theta_{k-1}, \frac{\Sigma_{\mathbf{b}}}{\kappa_{k-1}}) \tag{13}$$

where υ_{k-1} and Λ_{k-1} are the degrees of freedom and scale matrix for the inverse-Wishart distribution, respectively. θ_{k-1} is the prior mean and κ_{k-1} is the number of prior measurements. According with these assumptions, the joint prior density becomes

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \propto |\Sigma_{\mathbf{b}}|^{-(\upsilon_{k-1}+n)/2+1} \exp\left(-\frac{1}{2}\text{tr}(\Lambda_{k-1}\Sigma_{\mathbf{b}}^{-1}) - \frac{\kappa_{k-1}}{2}(\mu_{\mathbf{b}} - \theta_{k-1})^T\Sigma_{\mathbf{b}}^{-1}(\mu_{\mathbf{b}} - \theta_{k-1})\right), \tag{14}$$

a normal-inverse Wishart distribution (the product between a Gaussian and an inverse-Wishart). We recall that n is the number of shape parameters.

The inference step in eq.11 involves a Gaussian likelihood and the joint prior $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$, resulting in a joint posterior density of the same family (conjugate prior for a Gaussian with unknown mean and covariance), i.e. following a normal inverse-Wishart$(\theta_k, \Lambda_k/\kappa_k; \upsilon_k, \Lambda_k)$ distribution with the hyperparameters [11]:

$$\upsilon_k = \upsilon_{k-1} + m, \quad \kappa_k = \kappa_{k-1} + m \tag{15}$$

$$\theta_k = \frac{\kappa_{k-1}}{\kappa_{k-1}+m}\theta_{k-1} + \frac{m}{\kappa_{k-1}+m}\bar{\mathbf{b}} \tag{16}$$

$$\Lambda_k = \Lambda_{k-1} + \sum_{i=1}^{m}(\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})^T + \frac{\kappa_{k-1}m}{\kappa_{k-1}+m}(\bar{\mathbf{b}} - \theta_{k-1})(\bar{\mathbf{b}} - \theta_{k-1})^T \tag{17}$$

where $\bar{\mathbf{b}}$ is the mean of the new samples, m the number of samples used to update the model. The posterior mean θ_k is a weighted average between the prior mean θ_{k-1} and the sample mean $\bar{\mathbf{b}}$. The posterior degrees of freedom are equal to prior degrees of freedom plus the sample size. In our case, the second term in eq.17 ($\sum_{i=1}^{M}\cdots$) is null because the model is updated with one sample each time ($m = 1$).

Marginalizing over the joint posterior distribution $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ (eq.11) with respect to $\Sigma_{\mathbf{b}}$ gives the marginal posterior distribution for the mean of the form

$$p(\mu_{\mathbf{b}}|\mathbf{b}) \propto t_{\upsilon_k-n+1}(\mu_{\mathbf{b}}|\theta_k, \Lambda_k/(\kappa_k(\upsilon_k - n + 1))). \tag{18}$$

where t_{υ_k-n+1} is the multivariate Student-t distribution with $\upsilon_k - n + 1$ degrees of freedom.

Using the expectation of marginal posterior distribution $p(\mu_{\mathbf{b}}|\mathbf{b})$ as the model parameters at time k, we get (see table of expectation for multivariate t-distributions e.g. [11] pag.576).

$$\mu_{\mathbf{b}_k} = E(\mu_{\mathbf{b}}|\mathbf{b}) = \theta_k. \tag{19}$$

Similarly, marginalizing over the joint posterior distribution $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ with respect to $\mu_{\mathbf{b}}$ gives the marginal posterior distribution $p(\Sigma_{\mathbf{b}}|\mathbf{b})$ that follows an inverse Wishart distribution. The expectation for marginal posterior covariance is (see table of expectation for inverse Wishart distributions e.g. [11] pag.575)

$$\Sigma_{\mathbf{b}_k} = E(\Sigma_{\mathbf{b}}|\mathbf{b}) = (\upsilon_k - n - 1)^{-1}\Lambda_k. \tag{20}$$

.

3.3 Global Alignment Maximum a Posteriori (MAP)

In Bayesian inference, when the likelihood and the prior are Gaussian distributions the posterior is also a Gaussian. Consequently, a possible solution to the global alignment, can be given by the Bayes' theorem for Gaussian variables ([2], pag.90), 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. 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 $\mathbf{b}_k$ (eq.4). However, we further extend this result by adding two main components: **(1)** use a second order estimate of the latent variables (the covariance Σ_{k-1}). Using the covariance of the latent variables is a crucial issue, as it allows to account for the confidence on the current estimate (i.e. the amount of uncertainty in $\mathbf{b}_{k-1}$ should be considered in the estimate of $\mathbf{b}_k$). **(2)** Bayesian fusion of detectors. Allow to multiple (M) local detectors ($\sum_{m=1}^{M}\cdots$) to be seamlessly incorporated into the model, usually increase the fitting accuracy. The recursive posterior distribution takes the form of

$$p(\mathbf{b}_k|\mathbf{y}_k,\ldots,\mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\mu_k,\Sigma_k) \tag{21}$$

$$\Sigma_k = \left((\Sigma_{\mathbf{b}_k}+\Sigma_{k-1})^{-1}+\Phi^T\sum_{m=1}^{M}\left(\Sigma_{\mathbf{y}_{(m)}}^{-1}\right)\Phi\right)^{-1} \tag{22}$$

$$\mu_k = \Sigma_k\left(\Phi^T\sum_{m=1}^{M}\left(\Sigma_{\mathbf{y}_{(m)}}^{-1}\Delta\mathbf{y}_{(m)}\right)+(\Sigma_{\mathbf{b}_k}+\Sigma_{k-1})^{-1}\mu_{\mathbf{b}_k}\right) \tag{23}$$

where $\Delta\mathbf{y}_{(m)}$, $\Sigma_{\mathbf{y}_{(m)}}$ are the multiple likelihood observations.

The pose parameters $\mathbf{q}$ are estimated in the same way. The parameters of the normal inverse-Wishart distribution (eqs.15, 16 and 17) are kept up date and the global optimization step is used. However, the term Φ must be changed by Ψ in both eqs.22 and 23. See algorithm 1 where the overall global optimization is summarized.

Precompute: PDM $\mathbf{s}_0$, Φ, Ψ, $\Lambda_{PCA}=\text{diag}(\lambda_1,\ldots,\lambda_n)$, where λ_i is the i^{th} PCA eigenvalue and local detectors $\mathbf{H}_i^*$
Initial estimate of the shape/pose parameters and their covariances $(\mathbf{b}_0,\Sigma_0)$; $(\mathbf{q}_0,\Sigma_0^q)$
(**shape:** $\upsilon_0=2n$, $\kappa_0=1$, $\theta_0=\mathbf{b}_0$, $\Lambda_0=n\Lambda_{PCA}$) (**pose:** $\upsilon_0^q=8$, $\kappa_0^q=1$, $\theta_0^q=\mathbf{q}_0$, $\Lambda_0^q=4\times\text{diag}([0.05\ 0.005\ 5\ 5]^2)$)
repeat
  Warp image **I** to the base mesh using the current pose parameters $\mathbf{q}_k$ [0.5ms]
  Generate current shape $\mathbf{s}=\mathbf{s}_0+\Phi\mathbf{b}_k+\Psi\mathbf{q}_k$
  for *Landmark* $i=1$ **to** v **do**
    Evaluate the M detector(s) response(s), eq.24 [M x 3ms]
    Find the likelihood parameters $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$ using a local strategy (section 3.1.1, e.g. if KDE use eqs.9)
  end
  Estimate the pose parameters: (shape observation: $\Delta\mathbf{y}=\mathbf{y}-\mathbf{s}_0$)[0.15ms]
    - Update the parameters of the inverse Wishart distribution using eqs.15, 16 and 17
    - Expectation of the prior parameters $\mu_{\mathbf{q}_k}=\theta_k^q$ and $\Sigma_{\mathbf{q}_k}=(\upsilon_k^q-4-1)^{-1}\Lambda_k^q$
    - Evaluate the pose parameters $\mathbf{q}_k$ and the covariance $\Sigma_{\mathbf{q}_k}$ by eqs.23 and 22, (changing Φ by Ψ)
  Estimate the shape parameters: (shape observation: $\Delta\mathbf{y}=\mathbf{y}-\mathbf{s}_0-\Psi\mathbf{q}_k$) [0.25ms]
    - Update the parameters of the inverse Wishart distribution using eqs.15, 16 and 17
    - Expectation of the prior parameters $\mu_{\mathbf{b}_k}=\theta_k$ and $\Sigma_{\mathbf{b}_k}=(\upsilon_k-n-1)^{-1}\Lambda_k$
    - Evaluate the shape parameters $\mathbf{b}_k$ and the covariance $\Sigma_{\mathbf{b}_k}$ by eqs.23 and 22
until $||\boldsymbol{b}_k-\boldsymbol{b}_{k-1}||\leq\varepsilon$ *or maximum number of iterations reached* ;

Algorithm 1: Overview of the Bayesian Active Shape Models (BASM) method. The performance of BASM is comparable to ASM [20], CQF [23] or SCMS [12] depending of the local strategy BASM-WPR, BASM-GR or BASM-KDE, respectively. It achieves near real-time performance. The bottleneck is always obtaining the response maps (M x 3ms x number landmarks), although it can be done in parallel.

3.4 Hierarchical Search (BASM-KDE-H)

This section propose a slightly different annealing approach. When the local response maps are approximated by KDE representations, the global alignment can be done by a hierarchical search. The mean-shift bandwidth annealing schedule can be combined with additional global optimization steps. Bottom levels use highest KDE bandwidth and perform global optimization steps (section 3.3). Then the next level shrinks the bandwidth and repeats the process. Using this strategy the KDE annealing is performed between hierarchical levels.

4 Evaluation Results

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 23 subjects (20 landmarks). **(3)** The XM2VTS [13] database has 2360 images frontal faces of 295 subjects (68 landmarks). **(4)** The tracking performance is evaluated on the FGNet Talking Face (TF) [9] 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) [10] database that contains images taken under variability in pose, lighting, facial expression, occlusions, backgrounds, etc.

4.1 Local Detectors

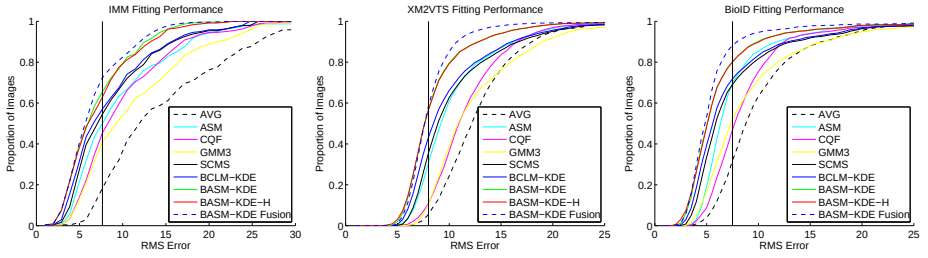
Performing a fair comparison requires that all the evaluated global optimization strategies use the same local detector. We experimentally found that the recently proposed MOSSE filter [8] perform better than the most used detector: the linear classifier build from aligned (positive) and misaligned (negative) examples [23][12]. As so, all the experiments use the MOSSE filter as local landmark detector. Briefly, the score of the i^{th} landmark detector, in eq.5, is given by

$$\mathcal{D}_i(\mathbf{I}(\mathbf{y}_i)) = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{I}(\mathbf{y}_i)\} \odot \mathbf{H}_i^*\}, \quad \text{with} \quad \mathbf{H}_i^* = \frac{\sum_{j=1}^{N} \mathbf{G}_j \odot \mathcal{F}\{\mathbf{I}_j\}^*}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^*}, \tag{24}$$

where $\mathbf{H}_i^*$ is the MOSSE filter [8], $\mathbf{I}(\mathbf{y}_i)$ a vectorized patch of pixel values sampled at $\mathbf{y}_i$, $(\mathcal{F})$ is the 2D Fourier transform, $(*)$ means the complex conjugate and $\odot$ the Hadamard product. $\mathbf{I}_j$ are aligned patch examples with size 128×128 (a power of two patch size to speed up the FFT computation, however only a 40×40 subwindow of the output is considered), N is the number of training images and $\mathbf{G}$ is the desired output which is set to be a 2D Gaussian function centered at the landmark with 3 pixels of standard deviation. In the following section, the performance of a Bayesian fusion of detections is also evaluated. The additional detector used is still a MOSSE filter but built from magnitude of image gradients $||\nabla \mathbf{I}_j||$.

4.2 Evaluating Global Optimization Strategies

In this section the BASM optimization strategy is evaluated w.r.t. state-of-the-art global alignment solutions. The proposed BASM and BASM-H methods are compared with (1) ASM [20], (2) CQF [23], (3) BCLM [22], (4) GMM [14] using three Gaussians (GMM3)



Reference 7.5 RMS	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)	
ASM	50.0		30.7		70.0	
BASM-WPR (our method)	**58.4**	(+8.4)	**47.4**	(+16.7)	**77.1**	(+7.1)
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)
BASM-GR (our method)	**51.8**	(+6.4)	**19.7**	(+8.8)	**63.5**	(+16.5)
SCMS-KDE	54.6		35.7		69.0	
BCLM-KDE	57.1	(+2.5)	43.4	(+7.7)	71.9	(+2.9)
BASM-KDE (our method)	**65.4**	(+10.8)	**57.0**	(+21.3)	**80.3**	(+11.3)
BASM-KDE-H (our method)	64.0	(+9.4)	56.6	(+20.9)	79.9	(+10.9)
BASM-KDE Fusion of 2 Detectors	**72.5**	(+17.9)	**58.7**	(+23.0)	**88.2**	(+19.2)

Figure 2: 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). AVG is the location provided by the initial estimate [18].

and (5) SCMS [12]. Note that the BASM can be used with different local strategies to approximate the response maps (e.g. WPR, GR or KDE as described in section 3.1.1 - Note that ASM, CQF and SCMS use as local strategy the WPR, GR and KDE, respectively). In these experiments we fixed the local strategy as a KDE (BCLM-KDE, SCMS-KDE, BASM-KDE) in order to compare the global optimization approaches. The same bandwidth schedule of $\sigma_h^2 = (15, 10, 5, 2)$ is always used. The results from ASM, CQF and GMM3 are provided as a baseline. In all cases, the nonrigid parameters start from zero, the similarity parameters were initialized by a face detection (Adaboost [18]) and the model was fitted until convergence (limited to a maximum of 20 iterations).

Figures 2 shows the fitting performance curves for the IMM, XM2VTS and BioID datasets, respectively. These fitting curves, also adopted by [4][5][6][23][12], show the percentage of faces that achieved convergence with a given Root Mean Square (RMS) error. The table, in the same figure 2, shows quantitative values taken by sampling the curves using a fixed RMS error amount (7.5 pixels, shown as a vertical line in graphics). To avoid confusion, the remainder local strategies (WPR and GR) appear only in the table. Results show that CQF performs better than GMM3, mainly because GMM is very prone to local optimums due to its multimodal nature. The main drawback of CQF is the limited accuracy due to the over-smoothness of the response map.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 (BASM) outperforms all previous methods. Explicitly modelling the prior distribution and using the covariance of the latent variables offers a significative in-

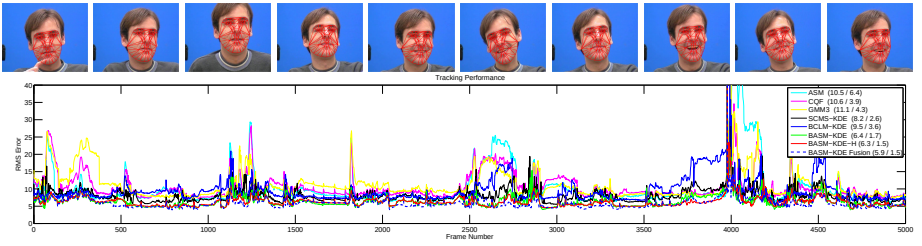


Figure 3: Evaluation of the tracking performance of several fitting algorithms on the FGNET Talking Face [9] sequence. The values on legend box are the mean and standard deviation RMS errors, respectively. Top images show BASM-KDE fitting examples.

crease in fitting performance. The Bayesian fusion of $(M = 2)$ local detectors was evaluated using the method that previously achieved the best performance (BASM-KDE). The results (BASM-KDE Fusion) show that including multiple sets of patch alignment classifiers further improve (a lot) the accuracy. In fact, this approach achieves the overall best results.

Tracking performance is evaluated in the FGNET Talking Face video sequence (fig. 3). Each frame is fitted using as 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 BASM techniques yields the best performance. Here, the hierarchical annealing version of BASM-KDE (BASM-KDE-H) performs slightly better, but at the cost of more iterations. The fusion of local detectors (BASM-KDE Fusion), as expected, improves even further the performance. Qualitative evaluation is also performed on the Labeled Faces in the Wild dataset [10], where some results can be seen in figure 1.

5 Conclusions

This work presents a novel Bayesian formulation for aligning faces in unseen images. Fitting a Point Distribution Model (PDM) to an image involves a global optimization step where the responses of an ensemble of local feature detectors are jointly maximized. The prior distribution models the dynamic transitions of the PDM parameters, being continuously kept up to date. The new global optimization strategy infers both the PDM and pose parameters, in a MAP sense, by explicitly modelling the prior distribution. Using recursive Bayesian estimation, a Gaussian prior distribution is modeled, treating the mean and covariance as random variables. This means that we estimate not only the mean and the covariance but also the probability distribution of the mean and the covariance (using conjugate priors). Extensive evaluations were performed on several standard datasets against state-of-the-art methods while using the same local detectors. We show that, generic image alignment by explicitly modelling the prior distribution offers a significant increase in performance.

Acknowledgements

This work was supported by the Portuguese Science Foundation (FCT) by the project “Dinâmica Facial 4D para Reconhecimento de Identidade“ with grant PTDC/EIA-CCO/108791/2008. Pedro Martins, Rui Caseiro and João F. Henriques acknowledge the FCT through the grants SFRH/BD/45178/2008, SFRH/BD74152/2010 and SFRH/BD/75459/2010, respectively.

References

[1] *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.

[2] C.M.Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] C.Shen, M.J.Brooks, and A.Hengel. Fast global kernel density mode seeking: Applications to localization and tracking. *IEEE TIP*, 16(5):1457–1469, May 2007.

[4] D.Cristinacce and T.F.Cootes. Facial feature detection using adaboost with shape constraints. In *BMVC*, 2003.

[5] D.Cristinacce and T.F.Cootes. Boosted regression active shape models. In *BMVC*, 2007.

[6] D.Cristinacce and T.F.Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[7] D.Cristinacce and T.F.Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[8] D.S.Bolme, J.R.Beveridge, B.A.Draper, and Y.M.Lui. Visual object tracking using adaptive correlation filters. In *IEEE CVPR*, 2010.

[9] FGNet. Talking face video, 2004. URL www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html.

[10] G.B.Huang, M.Ramesh, T.Berg, and E.L.-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[11] I.Matthews and S.Baker. Active appearance models revisited. *IJCV*, 60(1):135–164, 2004.

[12] J.Saragih, S.Lucey, and J.Cohn. Face alignment through subspace constrained mean-shifts. In *IEEE ICCV*, 2009.

[13] K.Messer, J.Matas, J.Kittler, J.Luettin, and G.Maitre. XM2VTSDB: The extended M2VTS database. In *AVBPA*, 1999.

[14] L.Gu and T.Kanade. A generative shape regularization model for robust face alignment. In *ECCV*, 2008.

[15] M.Nordstrom, M.Larsen, J.Sierakowski, and M.Stegmann. The IMM face database - an annotated dataset of 240 face images. Technical report, Technical University of Denmark, DTU, 2004. URL http://www2.imm.dtu.dk/pubdb/p.php?3160.

[16] O.Jesorsky, K.Kirchberg, and R.Frischholz. Robust face detection using the hausdorff distance. In *AVBPA*, 2001.

[17] P.Tresadern, H.Bhaskar, S.Adeshina, C.Taylor, and T.F.Cootes. Combining local and global shape models for deformable object matching. In *BMVC*, 2009.

[18] P.Viola and M.Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, July 2002.

[19] T.F.Cootes and C.J.Taylor. Statistical models of appearance for computer vision. Technical report, Imaging Science and Biomedical Engineering, University of Manchester, 2004.

[20] T.F.Cootes, C.J.Taylor, D.H.Cooper, and J.Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.

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

[22] U.Paquet. Convexity and bayesian constrained local models. In *IEEE CVPR*, 2009.

[23] Y.Wang, S.Lucey, and J.Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE CVPR*, 2008.