Bayesian Constrained Local Models Revisited

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

Abstract—This paper presents a novel Bayesian formulation for aligning faces in unseen images. Our approach revisits the Constrained Local Models (CLM) formulation where an ensemble of local feature detectors are constrained to lie within the subspace spanned by a Point Distribution Model (PDM). Fitting such a model to an image typically involves two main 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 detections at once. The so-called global optimization can be posed as a Bayesian inference problem, where the posterior distribution of the shape (and pose) parameters can be inferred in a *maximum a posteriori* (MAP) sense. This work introduces an extended Bayesian global optimization strategy that includes two novel additions: (1) to perform second order updates of the PDM parameters (accounting for their covariance) and (2) to model the underlying dynamics of the shape variations, encoded in the prior term, by using recursive Bayesian estimation. Extensive evaluations were performed against state-of-the-art methods on several standard datasets (IMM, BioID, XM2VTS, LFW and FGNET Talking Face). Results show that the proposed approach significantly increases the fitting performance.

Index Terms—Non-rigid face alignment, face registration, Constrained Local Models (CLM), Active Shape Models (ASM).

1 Introduction

Face alignment, or nonrigid face registration, is a fundamental problem in computer vision. Typically, this kind of registration step, usually taken at early stages, has a large impact on the robustness and quality of later processes, playing a central role in applications like: tracking, recognition (either identity or facial expression recognition), security, video compression, human computer interaction, beyond others. The main goal of face alignment consists of locating, with accuracy, the semantic structural facial landmarks such as eyes, nose, mouth, chin, eye brows, etc (usually between a target image and a reference frame). This problem has been studied for several years, being particularly difficult to track subjects with previously unseen appearance variations. These variations can be highly complex, typically caused by the rigid and non-rigid facial motion, changes in identity, expression, illumination and occlusion.

Face alignment has received an increasing attention over the years, specially since the introduction of the Active Shape Model (ASM) [1] and later with the Active Appearance Model (AAM) [2], becoming popular to describe faces by finding the parameters of a Point Distribution Model (PDM). The PDM is a simple, yet efficient, linear model where (facial) shapes are represented as a linear combination of 'eigenshapes' around the mean. Most important, given enough data, the PDM is able to generalize to unseen faces [3].

Several PDM fitting strategies have been proposed, most of which can be categorized as being either holistic (generative) or patch-based (discriminative). The holistic representations [2][4][5][6][7] model the appearance of all image pixels describing the object. In fact, matching a holistic model can be interpreted as building synthetic versions of the target face. By generating the expected appearance template, a high registration accuracy can be achieved (under favourable conditions). The AAMs [2][4] are probably the most widely used generative technique where parametric models of shape (the PDM) and appearance are matched into new images. However, such representation generalizes poorly when the object of interest exhibits large amounts of variability, such as the case of the human face under unseen variations, due to the high dimensional representation of the appearance (typically learned from limited data).

Recently, discriminative based methods, such as the Constrained Local Models (CLM) [1][8][9][10][11][12][13][14], have been proposed. These approaches improve the model's representation capacity, as they account 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 spanned by the PDM. The CLM implements a two step fitting stage: a local search and a global optimization. The first step performs an exhaustive local search using a feature detector, obtaining response maps for each landmark (likelihood term). Afterwards, a global optimization strategy finds the PDM parameters that jointly maximize all detection responses.

In this paper, we revisit the overall CLM framework, in particular, we show that the so-called global optimization can be posed as a Bayesian inference problem, where the posterior distribution of the PDM parameters can be inferred in a *maximum a posteri-*

• *P. Martins, J. Henriques, R. Caseiro and J. Batista are with the Institute of Systems and Robotics (ISR) at the University of Coimbra, Portugal. e-mails: {pedromartins, henriques, ruicaseiro, batista}@isr.uc.pt url: http://www.isr.uc.pt/~pedromartins*

ori (MAP) sense. An extended Bayesian global optimization strategy, that includes two main technical insights, is introduced. First, the overall PDM global alignment is formulated in terms of a Linear Dynamic System (LDS). Accordingly, the model is able to perform second order updates of the PDM parameters, accounting for the covariance of the shape and pose parameters (which represents the confidence in the current parameters estimate). Second, the underlying dynamics of the shape variations, encoded by the prior term, are explicitly modeled using recursive Bayesian estimation. This means that our CLM can tune its PDM to the new incoming data, becoming a more accurate representation. An extensive and thorough evaluation was performed in several standard datasets, demonstrating that our CLM formulation achieves a higher fitting performance. Additionally, a general evaluation of face parts descriptors (local landmark detectors) was also performed, where the recently proposed Minimum Output Sum of Squared Error (MOSSE) filters [15] stand out.

1.1 Related Work

Some of the most popular CLM optimization strategies propose to replace the true response maps by simple parametric forms (Weighted Peak Responses [1], Gaussians Responses [13], Mixture of Gaussians [16]) 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 [17], 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). This problem was mitigated later in [18] where a robust norm is used to select the most reliable landmarks.

Recently, a new paradigm has emerged. This new strategy suggests to formulate the global alignment as a Bayesian inference problem. The patch responses can be embedded into a Bayesian inference problem, where the posterior distribution of the global warp can be inferred in a MAP sense. The Bayesian approach provides an effective fitting solution as it combines in the same framework both the shape prior (the PDM) and multiple sets of patch alignment classifiers to further improve the accuracy. Previously, Bayesian extensions of the original ASM formulation have been proposed [8][19]. Likewise, the original CLM formulation [20] and consequently the Convex Quadratic Fitting approach [13] were extended into Bayesian formulations [21], by using a basic inference of the PDM parameters. Similarly, the later SCMS [18], described previously, include an enhanced PDM parameters update (maximum likelihood vs MAP).

Many other remarkable face alignment techniques have been recently proposed. These methods, among other features, include the use of fully non-parametric shape models [22], part-based tree structured models [23] (enabling face detection, pose estimation and parts localization in the same framework), shape regression updates [24][25], locally predict landmark updates by regression [9][26][27][28], regressing the PDM parameters from response maps [29], enhanced general nonlinear regression [30] (learning a set of steepest descent directions, not requiring the evaluation of Jacobian or Hessian afterwards), alignment on batches of images [31][32] and also discriminative versions of AAMs [10][33][34][35]. We remark that some of these methods rely on non-parametric shape models [22] [24][23][28] or aim to enhance local landmark detection [26][29] or even are based in graphical models inference [11][23] and therefore should not be compared with ours. This paper aims to extend the widely used CLM methodology, by using Bayesian inference techniques while maintaining its linear PDM regularization.

1.2 Contributions

This work presents a novel and efficient Bayesian CLM global alignment technique that includes two main additions: 1) to model the covariance of the latent variables, which represents the confidence in the current parameters estimate (explicitly maintaining second order statistics of the shape and pose parameters, instead of assuming them to be constant). It is shown that the posterior distribution of the global warp can be efficiently inferred by formulating the global alignment in terms of a Linear Dynamical System (LDS); 2) An extension that explicitly models the prior distribution, encoding the dynamic transitions of the PDM parameters, by using recursive Bayesian estimation. The prior distribution of the incoming new data is modeled as being Gaussian where the mean and covariance were assumed to be unknown and treated as random variables.

The paper shows that aligning the PDM using a Bayesian approach offers a significant increase in performance, in both fitting still images and video sequences, when compared with state-of-the-art first order forwards additive methods [1][13][17]. We confirm experimentally that the MAP parameter update outperforms the standard optimization strategies based on maximum likelihood solutions (least squares). A

comparison between several face parts descriptors is also presented, including the recently proposed Minimum Output Sum of Squared Error (MOSSE) filters [15]. The MOSSE maps aligned training patch examples into a desired output, producing correlation filters that are notably stable. Results show that the MOSSE outperforms others detectors, being particularly well-suited to the task of generic face alignment.

Finally, extensive evaluations were performed on several standard datasets (IMM [36], BioID [37], XM2VTS [38], FGNET Talking Face [39] and the Labeled Faces in the Wild (LFW) [40]) against state-of-the-art CLM methods while using the same likelihood source (local detectors).

Preliminary versions of this work were presented earlier in [41] and later in [42]. Here a unifying Bayesian CLM framework is described combining all previous contributions. The experimental results were largely extended, in particular, with the evaluation in the challenging LFW [40] dataset. The results have also been revised using more accurate error metrics.

1.3 Outline

The remainder of the paper is organized as follows: section 2 explains the basics in ASM/CLM design, in particular, the shape model and the local landmark detectors. Section 3 starts by describing the overall Bayesian alignment goal, then it revisits some likelihood extraction methods and finally it presents our Bayesian global strategy. The prior distribution in described in section 3.3 and the second order update strategy in section 3.5. Section 4 shows the main experimental results, including the local landmark detectors evaluation (section 4.2) and the fitting and tracking performances of several global strategies (section 4.3 and 4.4). Finally, section 5 summarizes the paper and provides the conclusions.

2 Background

2.1 Linear Shape Model

The shape $\mathbf{s}$ of a 2D Point Distribution Model (PDM) [43] is represented by the vertex locations of a mesh, with a $2v$ dimensional vector $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$. The usual way of building a PDM requires a set of shape annotated images that are previously aligned in scale, rotation and translation by a Generalized Procrustes Analysis. Afterwards, applying a Principal Components Analysis (PCA) to the set of aligned examples, each shape can be expressed by the following 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 known as the base mesh), Φ is the shape subspace matrix holding n eigenvectors (or the modes of deformation that retain a given amount of variance, e.g. 95%), $\mathbf{b}_s$ is a n dimensional vector of shape parameters representing the mixing weights and Ψ is a $2v \times 4$ matrix holding a special set of eigenvectors that linearly model the 2D pose [4], function of the $\mathbf{q} = [s\cos(\theta) - 1,\ s\sin(\theta),\ t_x,\ t_y]^T$ pose parameters (s, θ, t_x, t_y are the scale, rotation and translations w.r.t. the base mesh, respectively). Please refer to [43] for additional details in PDMs.

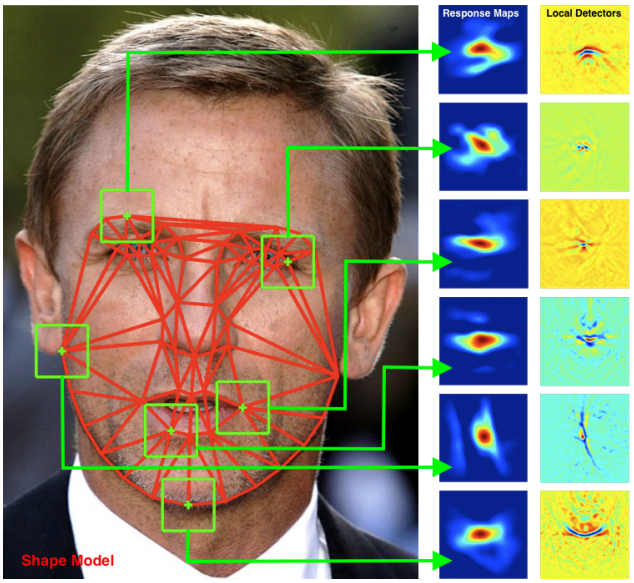


Fig. 1: The Constrained Local Model combines an ensemble of local feature detectors whose locations are constrained to be in a subspace spanned by a linear model. The novel Bayesian global optimization strategy (BCLM) jointly combines all detectors scores, in a MAP sense, using second order updates of the parameters and modelling the prior distribution. The image shows the local search regions for some highlighted landmarks, followed by a column with the detectors responses and their local detectors, respectively.

2.2 Local Detectors

The appearance model of a CLM consists of an ensemble of v expert local detectors [1][20][17][41][42] whose locations are regularized by the linear shape model as described in the previous section (figure 1).

The correlation of the i^{th} landmark detector, evaluated at the pixel location $\mathbf{x}_i = (x_i, y_i)$, is given by

$$\mathcal{D}_i(\mathbf{I}(\mathbf{x}_i)) = \mathbf{h}_i^T \mathbf{I}(\mathbf{x}_i) \tag{2}$$

where $\mathbf{h}_i$ is a linear detector and $\mathbf{I}(\mathbf{x}_i)$ is the surrounding $L \times L$ support region (i.e. the image patch, denoted by $\mathbf{\Omega}_{\mathbf{x}_i}$). Note that these landmark detectors are usually designed to operate at a given scale. The ASM/CLM framework deals with this by including a warp normalization step, in particular, a similarity transformation into the base mesh. At this stage the detector score must be converted into a probability value. The simplest solution is to use a logistic function. Defining a_i to be a binary variable that denotes

correct landmark alignment, the probability of pixel $\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{x}_i}$ being aligned is given by

$$p_i(\mathbf{z}_i) = p(a_i = 1|\mathcal{D}_i, \mathbf{I}(\mathbf{z}_i)) = \frac{1}{1 + e^{-a_i\beta_1\mathcal{D}_i(\mathbf{I}(\mathbf{z}_i))+\beta_0}} \tag{3}$$

where β_1 and β_0 are the regression coefficient and intercept, respectively (both β_1 and β_0 are usually found by cross-validation). In the previous, and for the sake of simplicity, $p_i(\mathbf{z}_i)$ is just used as a condensed representation for the response map. Note that a proper probability is used, always non-negative and $p(a_i = 1|\mathbf{I}(\mathbf{z}_i)) + p(a_i = -1|\mathbf{I}(\mathbf{z}_i)) = 1$.

3 Global PDM Optimization

The deformable model fitting goal is formulated as a global shape alignment problem using Bayesian inference techniques. This section describes the proposed (Bayesian) global optimization strategy where two main additions are included. The first formulates the global alignment in a *maximum a posteriori* (MAP) sense, by means of second order updates of the PDM parameters. The global optimization uses the covariance of the parameters, effectively accounting for previous uncertainty. The second novelty consists in explicitly update the prior distribution, i.e. to learn the way the PDM parameters change.

3.1 The Alignment Goal

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

$$\mathbf{b}_s^* = \arg\max_{\mathbf{b}_s} p(\mathbf{b}_s|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b}_s)p(\mathbf{b}_s) \tag{4}$$

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. 4, can be reduced by making some simple assumptions. Firstly, conditional independence between landmarks can be assumed by simply sampling each landmark independently. Secondly, it can also be considered that we are facing an approximate solution to the true parameters ($\mathbf{b} \approx \mathbf{b}_s^*$). Combining these approximations, the Eq. 4 can be rewritten as

$$p(\mathbf{b}|\mathbf{y}) \propto \left(\prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b})\right) p(\mathbf{b}|\mathbf{b}_{k-1}^*) \tag{5}$$

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

3.2 The Likelihood Term

The likelihood term, including the PDM model in Eq. 1, can be written by the following convex energy function:

$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\underbrace{\mathbf{y} - (\mathbf{s}_0}_{\Delta\mathbf{y}} + \Phi\mathbf{b}))^T\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))\right) \tag{6}$$

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). Summarizing, from the probabilistic point of view, the likelihood term follows a Gaussian distribution given by

$$p(\mathbf{y}|\mathbf{b}) \propto \mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_{\mathbf{y}}). \tag{7}$$

3.2.1 Local Optimization Strategies

Several local strategies can be used to represent the true response maps either by parametric or non-parametric probabilistic models. These local strategies consists of extracting the likelihood parameters (both the observed shape $\mathbf{y}$ and the landmark uncertainty covariance $\Sigma_{\mathbf{y}}$) from each probabilistic model representing the response map.

The parameters $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$ (candidates to the i^{th} landmark) can be found by maximizing the expression

$$\arg\max_{\mathbf{y}_i, \Sigma_{\mathbf{y}_i}} \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\mathcal{N}(\mathbf{z}_i|\mathbf{y}_i, \Sigma_{\mathbf{y}_i}) \tag{8}$$

where $p_i(\mathbf{z}_i)$, defined in Eq. 3, represents the probability of a candidate pixel $\mathbf{z}_i$ being aligned (i.e. the response map) and $\mathbf{\Omega}_{\mathbf{y}_i^c}$ is the patch support region centered at $\mathbf{y}_i^c$, which represents the current landmark estimate. Several strategies can be used to perform this optimization:

Weighted Peak Response (WPR) - The simplest solution is to take the spatial location where the response map has a higher score [1]. The new landmark position is then weighted by a factor that reflects the peak confidence (the uncertainty is inverse proportional to the peak value). Formally, the WPR solution is given by

$$\mathbf{y}_i^{\text{WPR}} = \max_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} (p_i(\mathbf{z}_i)), \quad \Sigma_{\mathbf{y}_i}^{\text{WPR}} = diag(p_i(\mathbf{y}_i^{\text{WPR}})^{-1}) \tag{9}$$

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

Gaussian Response (GR) - The previous approach was extended in [13] to approximate the response maps by a full Gaussian distribution $\mathcal{N}(\mathbf{z}_i|\mathbf{y}_i^{\text{GR}}, \Sigma_{\mathbf{y}_i}^{\text{GR}})$. This is equivalent to fit a Gaussian density to weighted data. Defining $d = \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)$, the solution is given by

$$\mathbf{y}_i^{\text{GR}} = \frac{1}{d} \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\ \mathbf{z}_i \tag{10}$$

$$\Sigma_{\mathbf{y}_i}^{\mathrm{GR}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})^T. \quad (11)$$

Kernel Density Estimator (KDE) - The response maps can 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 [44][17]. 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. It can be shown [45] 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

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)} \quad (12)$$

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

$$\Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T. \quad (13)$$

The qualitative differences between the three local optimization strategies (WPR, GR or KDE) are shown in figure 2. Notice that, by default, the global approach deals with mild occlusions, i.e. when a landmark region is under occlusion, typically the response map exhibits a multimodal behavior (see the examples in figure 2). Assuming that a KDE local strategy is used, the landmark update will select the nearest mode, according to the bandwidth (Eq. 12), and the covariance of that landmark (Eq. 13) will be inherently large, modeling a high localization uncertainty. The global optimization stage then jointly combines all uncertainties (MAP sense), dealing with the occlusion. Similarly, to deal with large occlusions, a minor tweak is required. One can simply set a large covariance $\Sigma_{\mathbf{y}_i}$ for the occluded landmarks.

3.3 The Prior Term

Faces are special nonrigid structures described by continuous dynamic transitions that deform continuously in time. In the Bayesian paradigm, the prior term can be used to encode this underlying dynamic of the shape. In following sections two different approaches are considered: defining the prior by the standard PDM belief (section 3.3.1) or effectively modeling it, keeping the prior distribution always up to date (section 3.3.2).

3.3.1 PDM based Prior Term

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

$$p(\mathbf{b}_k | \mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k | \mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \quad (14)$$

where $\mu_{\mathbf{b}} = \mathbf{b}_{k-1}$ and $\Sigma_{\mathbf{b}} = \Lambda$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, being the shape parameters covariance and λ_j denotes the PCA eigenvalue of the j^{th} mode of deformation. This form of prior assumption, simple enough for most cases, can be largely improved.

3.3.2 Modeling the Prior Term

The previous section defines the prior term to follow a Gaussian distribution with a given mean ($\mu_{\mathbf{b}}$) and covariance ($\Sigma_{\mathbf{b}}$). These parameters are established in the PDM building process and remain unchanged afterwards.

In this section let's consider the mean $\mu_{\mathbf{b}}$ and covariance $\Sigma_{\mathbf{b}}$ of the data to be unknown and modeled as random variables ([46] pag.87-88). Recursive Bayesian estimation can be applied to infer the parameters of the prior distribution in Eq. 14. Defining $\mathbf{b}$ as an observable vector, the Bayes theorem tells us that the joint posterior density can be written as

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}} | \mathbf{b}) \propto p(\mathbf{b} | \mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}). \quad (15)$$

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

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) = p(\mu_{\mathbf{b}} | \Sigma_{\mathbf{b}}) p(\Sigma_{\mathbf{b}}). \quad (16)$$

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 [46])

$$\Sigma_{\mathbf{b}} \sim \text{Inv-Wishart}_{\upsilon_{k-1}}(\Lambda_{\upsilon_{k-1}}^{-1}) \quad (17)$$

$$\mu_{\mathbf{b}} | \Sigma_{\mathbf{b}} \sim \mathcal{N}(\theta_{k-1}, \frac{\Sigma_{\mathbf{b}}}{\kappa_{k-1}}) \quad (18)$$

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

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \propto |\Sigma_{\mathbf{b}}|^{-(\upsilon_{k-1}+n)/2+1} \exp\left(-\tfrac{1}{2}\mathrm{tr}(\Lambda_{k-1}\Sigma_{\mathbf{b}}^{-1}) - \tfrac{\kappa_{k-1}}{2}(\mu_{\mathbf{b}} - \theta_{k-1})^T \Sigma_{\mathbf{b}}^{-1}(\mu_{\mathbf{b}} - \theta_{k-1})\right) \quad (19)$$

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. 15 involves a Gaussian likelihood and the joint prior $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$, resulting in

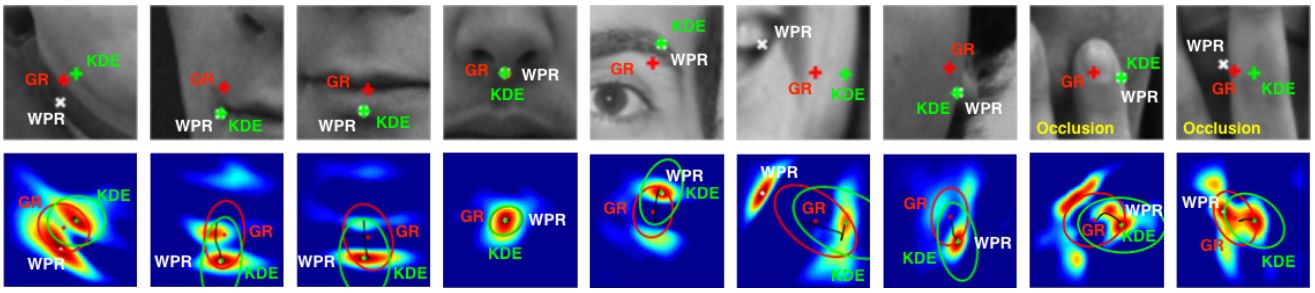


Fig. 2: Qualitative comparison between the three local optimization strategies. The WPR simply chooses the maximum detector response. The GR approximates the response map by a full Gaussian distribution and KDE uses the mean-shift algorithm to move to the nearest mode of the density. Its uncertainty covariance is found using the entire response map centered at the found mode. The two examples in the right show patches under occlusion (typically multimodal responses).

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 [46]:

$$\upsilon_k = \upsilon_{k-1} + m, \quad \kappa_k = \kappa_{k-1} + m \tag{20}$$

$$\theta_k = \frac{\kappa_{k-1}}{\kappa_{k-1}+m}\theta_{k-1} + \frac{m}{\kappa_{k-1}+m}\overline{\mathbf{b}} \tag{21}$$

$$\Lambda_k = \Lambda_{k-1} + \sum_{i=1}^{m}(\mathbf{b}_i - \overline{\mathbf{b}})(\mathbf{b}_i - \overline{\mathbf{b}})^T + \frac{\kappa_{k-1}m}{\kappa_{k-1}+m}(\overline{\mathbf{b}} - \theta_{k-1})(\overline{\mathbf{b}} - \theta_{k-1})^T \tag{22}$$

where $\overline{\mathbf{b}}$ 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 $\overline{\mathbf{b}}$. The posterior degrees of freedom are equal to prior degrees of freedom plus the sample size. In the present case, the second term in Eq. 22 ($\sum_{i=1}^{m}\cdots$) 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. 15) with respect to $\Sigma_\mathbf{b}$ gives the marginal posterior distribution for the mean of the form

$$p(\mu_\mathbf{b}|\mathbf{b}) \propto t_{\upsilon_k - n+1}(\mu_\mathbf{b}|\theta_k, \Lambda_k/(\kappa_k(\upsilon_k - n + 1))) \tag{23}$$

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. [46] pag.576)

$$\mu_{\mathbf{b}_k} = E(\mu_\mathbf{b}|\mathbf{b}) = \theta_k. \tag{24}$$

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. [46] pag.575)

$$\Sigma_{\mathbf{b}_k} = E(\Sigma_\mathbf{b}|\mathbf{b}) = (\upsilon_k - n - 1)^{-1}\Lambda_k. \tag{25}$$

A similar approach is used to estimate the pose parameters ($\mathbf{q}_k$). The parameters of the normal inverse-Wishart distribution (Eqs. 20, 21 and 22) are kept up date and the updates for the $\mu_{\mathbf{q}_k}$ and $\Sigma_{\mathbf{q}_k}$ are given by the according expectations.

3.4 The MAP Global Alignment

An important property of Bayesian inference is that, when the likelihood and the prior are Gaussian distributions the posterior is also Gaussian [47]. Additionally, the conditional distribution $p(\mathbf{y}|\mathbf{b}_k)$, in Eq. 6, has a mean that is a linear function of $\mathbf{b}_k$ and a covariance which is independent of $\mathbf{b}_k$. 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 ([47], pag 90)

$$p(\mathbf{b}_k|\mathbf{y}) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}^\mathbf{F}, \boldsymbol{\Sigma}^\mathbf{F}) \tag{26}$$

$$\boldsymbol{\Sigma}^\mathbf{F} = (\Sigma_\mathbf{b}^{-1} + \Phi^T\Sigma_\mathbf{y}^{-1}\Phi)^{-1} \tag{27}$$

$$\boldsymbol{\mu}^\mathbf{F} = \boldsymbol{\Sigma}^\mathbf{F}(\Phi^T\Sigma_\mathbf{y}^{-1}\Delta\mathbf{y} + \Sigma_\mathbf{b}^{-1}\mu_\mathbf{b}). \tag{28}$$

This could be a possible solution to the global alignment optimization ($\mathbf{b}_s^* = \boldsymbol{\mu}^\mathbf{F}$, solution proposed in [21]). In the mentioned technique, Eqs. 28 and 27 are iteratively reused, where the subscript k represents the iteration number, along with the response maps $p_i(\mathbf{z}_i)$ evaluated at the new updated locations, until converge. The previous approach can be largely improved by additionally modeling the covariance of the latent variables ($\mathbf{b}_k$) which allows to account for the amount of confidence in the current parameters estimate (i.e. the amount of uncertainty in $\mathbf{b}_{k-1}$ should be considered in the estimate of $\mathbf{b}_k$).

3.5 Second Order Global Alignment

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 basic 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. The state and measurement equations of the LDS, according to the PDM alignment problem, can be written as

$$\mathbf{b}_k = \mathbf{A}\mathbf{b}_{k-1} + q \quad (29)$$
$$\Delta\mathbf{y} = \Phi\mathbf{b}_k + r \quad (30)$$

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 is 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 ($\mathbf{A} = \mathbf{I}_n$).

3.5.1 Inference using the LDS

We highlight that the final step of the LDS derivation consists of a Bayesian inference step [47] (using Bayes' theorem for Gaussian variables, Eqs. 27 and 28), where the likelihood term is given by Eq. 7 and a prior that follows $\mathcal{N}(\mathbf{A}\mu^\mathbf{S}_{k-1}, \mathbf{\Sigma}^\mathbf{P}_{k-1})$ with

$$\mathbf{\Sigma}^\mathbf{P}_{k-1} = \Lambda + \mathbf{A}\mathbf{\Sigma}^\mathbf{S}_{k-1}\mathbf{A}^T. \quad (31)$$

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

$$p(\mathbf{b}_k|\Delta\mathbf{y}_k, \ldots, \Delta\mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}^\mathbf{S}_k, \mathbf{\Sigma}^\mathbf{S}_k) \quad (32)$$

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

$$\mathbf{K} = \mathbf{\Sigma}^\mathbf{P}_{k-1}\Phi^T(\Phi\mathbf{\Sigma}^\mathbf{P}_{k-1}\Phi^T + \Sigma_\mathbf{y})^{-1} \quad (33)$$
$$\boldsymbol{\mu}^\mathbf{S}_k = \mathbf{A}\boldsymbol{\mu}^\mathbf{S}_{k-1} + \mathbf{K}(\Delta\mathbf{y} - \Phi\mathbf{A}\boldsymbol{\mu}^\mathbf{S}_{k-1}) \quad (34)$$
$$\mathbf{\Sigma}^\mathbf{S}_k = (\mathbf{I}_n - \mathbf{K}\Phi)\mathbf{\Sigma}^\mathbf{P}_{k-1}. \quad (35)$$

3.5.2 Second Order Inference with Prior Modeling

Unfortunately, the previously LDS inference only applies to the simple prior assumption (section 3.3.1). When considering the more evolved prior form (section 3.3.2), Eq. 29 is no longer valid (where the adaptive prior term is now $\mathcal{N}(\mu_{\mathbf{b}_k}, \Sigma_{\mathbf{b}_k} + \mathbf{\Sigma}_{k-1})$ instead of $\mathcal{N}(\mu^\mathbf{S}_{k-1}, \mathbf{\Sigma}^\mathbf{P}_{k-1})$), therefore the inference step falls back to a more general approach given by

$$p(\mathbf{b}_k|\Delta\mathbf{y}_k, \ldots, \Delta\mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \quad (36)$$

with

$$\mathbf{\Sigma}_k = \left((\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} + \Phi^T\Sigma^{-1}_\mathbf{y}\Phi\right)^{-1} \quad (37)$$
$$\boldsymbol{\mu}_k = \mathbf{\Sigma}_k\left(\Phi^T\Sigma^{-1}_\mathbf{y}\Delta\mathbf{y} + (\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1}\mu_{\mathbf{b}_k}\right). \quad (38)$$

Precompute: The PDM $\mathbf{s}_0$, Φ, Λ and local detectors $\mathbf{H}^*_i$
Initial estimate of the shape/pose parameters and their covariances $(\mathbf{b}_0, \Sigma_{\mathbf{b}_0})/(\mathbf{q}_0, \Sigma_{\mathbf{q}_0})$
repeat
- Extract likelihood from image:
  - Backwarp input image to base mesh using $\mathbf{q}_k$
  - Generate shape at base mesh: $\mathbf{s} = \mathbf{s}_0 + \Phi\mathbf{b}_k$
  - **for** *Landmark* $i = 1$ **to** v **do**
    Evaluate M detector(s) response(s) (Eq. 3)
    Find the likelihood parameters $\mathbf{y}_i$ and $\Sigma_{\mathbf{y}_i}$
    - use WPR, GR or KDE local strategy (sec. 3.2.1)
  - **end**
- Estimate the pose parameters:
  - Update hyperparameters $\upsilon^q_0, \kappa^q_0, \theta^q_0, \Lambda^q_0$ (Eqs.20, 21, 22)
  - Expectation of the prior parameters $(\mu_{\mathbf{q}_k}, \Sigma_{\mathbf{q}_k})$
  - Pose observation: $\Delta\mathbf{y}^q = \mathbf{y} - \mathbf{s}_0$
  - Posterior parameters $\mathbf{q}_k$ and $\Sigma_{\mathbf{q}_k}$ (Eqs. 40, 39)
- Estimate the shape parameters:
  - Update hyperparameters $\upsilon^b_0, \kappa^b_0, \theta^b_0, \Lambda^b_0$ (Eqs.20, 21, 22)
  - Expectation of the prior parameters $(\mu_{\mathbf{b}_k}, \Sigma_{\mathbf{b}_k})$
  - Shape observation: $\Delta\mathbf{y}^b = \mathbf{y} - \mathbf{s}_0 - \Psi\mathbf{q}_k$
  - Posterior parameters $\mathbf{b}_k$ and $\Sigma_{\mathbf{b}_k}$ (Eqs. 40, 39)

until $||\boldsymbol{b}_k - \boldsymbol{b}_{k-1}|| \leq \varepsilon$ *or maximum number of iterations* ;

Algorithm 1: Overview of the Bayesian Constrained Local Models (BCLM) method.

The optimal shape parameters $\mathbf{b}^*_s$ that maximize the overall goal, in Eq. 4, are given by $\boldsymbol{\mu}_k$ (or alternatively by $\boldsymbol{\mu}^\mathbf{S}_k$ if the simple prior is used). Similarly, in order to estimate the pose parameters ($\mathbf{q}_k$), the same paradigm is also applied. The main difference in this case is that the observation matrix becomes Ψ (a linear representation for the 2D pose, in Eq. 1).

3.6 Multiple Local Detectors per Landmark

An useful strategy aimed to increase the overall fitting accuracy is to include multiple landmark detectors for each landmark [48][49][50]. In the Bayesian framework these different likelihood sources (a.k.a. Bayesian fusion of detectors), can be seamlessly incorporated into the same model. In general, by combining the usage of multiple (M) local detectors with the second order inference and prior term modelling (section 3.3.2), results in a Gaussian posterior distribution with the parameters given by

$$\mathbf{\Sigma}'_k = \left((\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} + \Phi^T\sum_{m=1}^{M}\left(\Sigma^{-1}_{\mathbf{y}_m}\right)\Phi\right)^{-1} \quad (39)$$

$$\boldsymbol{\mu}'_k = \mathbf{\Sigma}'_k\left(\Phi^T\sum_{m=1}^{M}\left(\Sigma^{-1}_{\mathbf{y}_m}\Delta\mathbf{y}_m\right) + (\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1}\mu_{\mathbf{b}_k}\right) \quad (40)$$

where $\Delta\mathbf{y}_m$, $\Sigma_{\mathbf{y}_m}$ are the multiple likelihood observations that are extracted from multiple response maps.

The algorithm 1 describes the overall global optimization here referred as Bayesian Constrained Local Models (BCLM) in its most generic form (i.e. second order inference while modelling the prior term and including multiple detections per landmark). Notice that: when the simple prior term is used, the hyperparameters updates steps are not required (Eqs. 20,

21, 22) and the posterior parameters in Eqs. 40 and 39 reduce to the LDS formulas (Eqs. 34 and 35).

The BCLM training stage consists of learning the PDM, which includes the base mesh $\mathbf{s}_0$, the linear shape subspace Φ, their eigen values diagonal matrix Λ, the linear pose subspace Ψ and the evaluation of all the local landmark detectors $\mathbf{H}_i^*$ ($i = 1, \ldots, v$ landmarks). The model requires an initial estimate for the shape, the pose parameters and their respective covariances. Typically, the mean shape is used ($\mathbf{b}_0 = \mathbf{0}$), the pose parameters $\mathbf{q}_0$ are given by a face detector [51] (scale and translation adjusting the base mesh to the detection), the covariance of the shape parameters is initialized as a diagonal matrix with the average PDM variance ($\overline{\lambda} = \frac{1}{n}\sum_j \lambda_j$) or $\Sigma_{\mathbf{b}_0} = \text{diag}(\overline{\lambda})$ and finally, the covariance of the pose parameters starts as $\Sigma_{\mathbf{q}_0} = \text{diag}([0.1\ 0.01\ 10\ 10]^2)$. Two different joint posterior densities are inferred ($p(\mu_\mathbf{b}, \Sigma_\mathbf{b}|\mathbf{b})$ and $p(\mu_\mathbf{q}, \Sigma_\mathbf{q}|\mathbf{q})$), according, two sets of hyperparameters of the normal inverse-Wishart distributions are used. They are initialized as $\upsilon_0^b = 2n$, $\kappa_0^b = 1$, $\theta_0^b = \mathbf{b}_0$, $\Lambda_0^b = n\Lambda$, and $\upsilon_0^q = 8$, $\kappa_0^q = 1$, $\theta_0^q = \mathbf{q}_0$, $\Lambda_0^q = 4\times\text{diag}([0.05\ 0.005\ 5\ 5]^2)$ where the upperscripts b and q refers to shape and pose, respectively.

The model fitting itself is similar, with some minor modifications, to a standard CLM search. For each iteration, the input image is backwarped using the current pose parameters estimate, then local detectors are used to evaluate every response map. One of the local strategies (WPR, GR or KDE) is selected to find a new set of landmark candidates (and their uncertainty), the hyperparameters of the normal inverse-Wishart distributions are updated with the incoming example, then the parameters of the current prior distribution are evaluated (by the expectation of the joint posterior distributions) and finally the second order global optimization is used to find the new set of shape and pose parameters for the next iteration (Eqs. 40 and 39). This process is repeated until convergence, when both shape and pose parameters do not change substantially.

The performance of our approach (BCLM) is comparable to ASM [1], CQF [13] or SCMS [17] depending of the local strategy used (refereed, from now on, as BCLM-WPR, BCLM-GR or BCLM-KDE, respectively). The bottleneck, like in any other CLM approach, lies on the evaluation of the response maps (M x 3ms x number landmarks), although keep in mind that it can be done in parallel.

3.7 Hierarchical Model Search

When the local response maps are approximated by KDE representations (section 3.2.1), the overall alignment can be done using a slightly different annealing approach that is described here as an hierarchical search strategy. The standard search iteratively uses the mean-shift algorithm with a kernel bandwidth relaxation, e.g. $\sigma_{h_j}^2 = [15, 10, 5, 2]$, followed by a global optimization step. However, the bandwidth annealing schedule can be combined within the global optimization steps. This solution consists of multiple levels of fixed bandwidth mean-shifts followed by global optimization steps (i.e. the annealing is performed between hierarchical levels).

TABLE 1: Comparative view between the standard first order, the second order (proposed) and the BCLM (proposed) inference techniques. Secs. 3.4, 3.5.1 and 3.5.2, respectively.

	1^{st} order	2^{nd} order	BCLM
Likelihood	$\mathcal{N}(\Phi\mathbf{b}, \Sigma_\mathbf{y})$	$\mathcal{N}(\Phi\mathbf{b}, \Sigma_\mathbf{y})$	$\mathcal{N}(\Phi\mathbf{b}, \Sigma_\mathbf{y})$
Prior	$\mathcal{N}(\mathbf{0}, \Sigma_\mathbf{b})$	$\mathcal{N}(\mu_\mathbf{b}, \Sigma_\mathbf{b} + \mathbf{\Sigma}_{k-1}^\mathbf{S})$	$\mathcal{N}(\mu_{\mathbf{b}_k}, \Sigma_{\mathbf{b}_k} + \mathbf{\Sigma}_{k-1})$
Posterior	$\mathcal{N}(\boldsymbol{\mu}^\mathbf{F}, \mathbf{\Sigma}^\mathbf{F})$	$\mathcal{N}(\boldsymbol{\mu}_k^\mathbf{S}, \mathbf{\Sigma}_k^\mathbf{S})$	$\mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$

3.8 Summary

Table 1 shows a comparative probabilistic view between all the described methods, namely: the first order, the second order and the BCLM (also second order) inference techniques. The likelihood term $\mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_\mathbf{y})$, which is extracted from the response maps, remains the same for all approaches. In general, the Bayes's theorem of Gaussian variables, ensures the all methods have Gaussian posterior distributions. The main diference between them comes down to the prior distribution. The most basic method (1^{st} order) only considers the usual PDM assumption (a constant prior). The second order method enhances the previous by accounting with an adaptive prior. Finally, the BCLM further improves the second order technique by continuously updating $\mu_{\mathbf{b}_k}$ and $\Sigma_{\mathbf{b}_k}$.

4 Evaluation Results

The experiments evaluate, in the first place, the performance of the local landmark detectors. Afterwards, the new Bayesian global optimization (BCLM) is put to test while using the best detector. All the experiments were conducted on several databases with publicly available ground truth: 1) The IMM [36] 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 [37] dataset that contains 1521 images, each showing a near frontal view of a face of one of 23 different subjects (20 landmarks). 3) The XM2VTS [38] database has 2360 images of frontal faces from 295 subjects (68 landmarks). 4) The tracking performance was evaluated on the FGNet Talking Face (TF) [39] video sequence that holds 5000 frames of video of an individual engaged in a conversation (68 landmarks). To the best of our knowledge this is the only dataset with fully landmark annotations in a video sequence. 5) Finally, evaluation was also performed using the Labeled Faces in the Wild (LFW) [40] database that contains images taken under

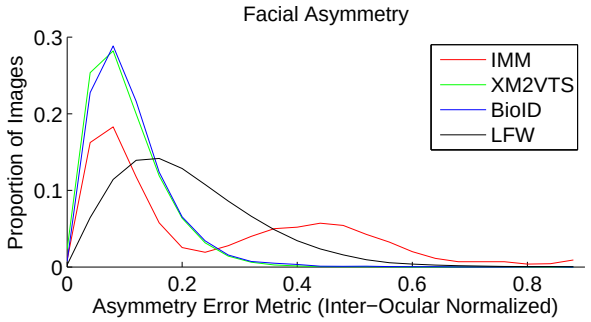


Fig. 3: Distribution of face asymmetry in the evaluated datasets.

variability in pose, lighting, focus, facial expression, occlusions, different backgrounds, etc. Qualitatively, both XM2VTS and BioID focuses mainly on variations in identity. Nevertheless, they exhibit large diversity in appearance due to facial hair, glasses, ethnicity and other subtle changes. The IMM is the smallest database, however it presents a large variation in head pose, illumination, and spontaneous facial expressions along several individuals. Unlike the previous, the LFW database is an extremely challenging database, completely taken in wild.

Inspired by [22], the overall face alignment challenge was evaluated in the different datasets, by assessing a measure of facial asymmetry for each image. Natural symmetric features such as the eyes out corners and mouth corners were reflected about a vertical line passing the nose center and the (normalized) average distances between them are computed. This metric holds a lower value (close to zero) in near frontal faces. Figure 3 shows this asymmetry measure over the evaluated datasets. We can see that both BioID and XM2VTS sets hold more symmetric images (more frontal), by other hand, the IMM and LFW have indeed more challenging images with a lot more 3D pose variability, therefore more difficult to align.

4.1 The Local Detector - MOSSE Filter

The Minimum Output Sum of Squared Error (MOSSE) filter, recently proposed in [15], 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

$$\mathbf{G} = \mathcal{F}\{\mathbf{I}\} \odot \mathbf{H}^* \tag{41}$$

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 $\mathbf{G}$. The 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,

$$\min_{\mathbf{H}^*} \sum_{j=1}^{N} \left(\mathcal{F}\{\mathbf{I}_j\} \odot \mathbf{H}^* - \mathbf{G}_j\right)^2 \tag{42}$$

where $\mathbf{G}$ is obtained by sampling a 2D Gaussian uniformly. Solving for the filter $\mathbf{H}^*$ yields the closed form solution

$$\mathbf{H}^* = \frac{\sum_{j=1}^{N} \mathbf{G}_j \odot \mathcal{F}\{\mathbf{I}_j\}^*}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^* + \epsilon} \tag{43}$$

where ϵ is a regularization parameter introduced to prevent divisions by zero.

The MOSSE filter maps all aligned training patch examples to an output, $\mathbf{G}$, centered at the feature location, producing notably stable correlation filters. Each sample $\mathbf{I}_j$ is normalized to have zero mean and a unitary norm and it is multiplied by a cosine window (required to solve the Fourier Transform periodicity problem) which 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 [13][17] 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

$$\mathcal{D}_i^{\text{linear}}(\mathbf{I}(\mathbf{y}_i)) = \mathbf{w}_i^T \mathbf{I}(\mathbf{y}_i) + b_i, \tag{44}$$

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

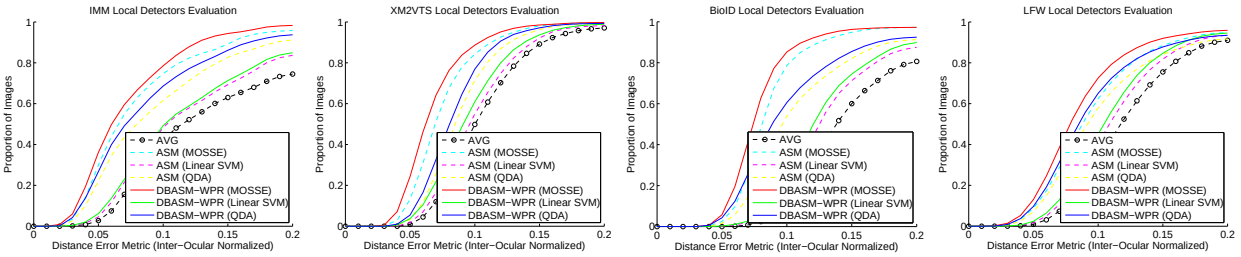
$$\mathcal{D}_i^{\text{quadratic}}(\mathbf{I}(\mathbf{y}_i)) = \mathbf{I}(\mathbf{y}_i)^T \mathbf{Q}_i \mathbf{I}(\mathbf{y}_i) + \mathbf{L}_i^T \mathbf{I}(\mathbf{y}_i) + b_i \tag{45}$$

with $\mathbf{Q}_i$ and $\mathbf{L}_i$ being the quadratic and linear terms, respectively. Finally, the MOSSE filter correlation (which is still a linear detector) gives

$$\mathcal{D}_i^{\text{MOSSE}}(\mathbf{I}(\mathbf{y}_i)) = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{I}(\mathbf{y}_i)\} \odot \mathbf{H}_i^*\} \tag{46}$$

where $\mathbf{H}_i^*$ is the MOSSE filter from Eq. 43.

Both linear and quadratic classifiers (linear-SVM [52] and Quadratic Discriminant Analysis) were trained using images from the IMM [36] dataset with 144 negative patch examples (for each landmark and image) being misaligned in translation by 12 pixels in x and y direction. 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. The desired output $\mathbf{G}$ (Eq. 43) is set to be a 2D Gaussian function centered at the landmark with 3 pixels of standard deviation.



(a) IMM database [36] (b) XM2VTS database [38] (c) BioID database [37] (d) LFW database [40]

Fig. 4: Fitting performance curves comparing different detectors (linear, quadratic and MOSSE filters) in the IMM, XM2VTS, BioID and LFW database, respectively. The AVG represents the average location provided by the initial estimate [51].

Later, in section 4.3, the performance of the fusion of detections is also evaluated. The additional detector is still a MOSSE filter built with the same settings but using the magnitude of image gradients ($\|\nabla \mathbf{I}_j\|$).

The global optimization method that best evaluates the detectors performance is the approach that relies the most in the output of the detector, i.e., the Active Shape Models (ASM) [1]. The results are presented in the form of fitting performance curves, which were also adopted by [53][9][20][13][17]. These curves show the percentage of faces that achieved convergence with a given error amount. Following common practice [9][12][26], the error metric is given by the mean error per landmark as fraction of the inter-ocular distance (measured between the center of the eyes), d_{eyes}, as

$$e_m(\mathbf{s}) = \frac{1}{v\ d_{\text{eyes}}} \sum_{i=1}^{v} \|\mathbf{s}_i - \mathbf{s}_i^{\text{gt}}\| \qquad (47)$$

where $\mathbf{s}_i^{\text{gt}}$ is the location of i^{th} landmark in the ground truth annotation (and v the number of landmarks).

The figure 4 shows the fitting performance curves that compare the three kinds of detectors using the ASM [1] optimization and our second order global BCLM technique using a Weighted Peak Response strategy (BCLM-WPR) without modelling the prior distribution. In fact, this method appears in the evaluation charts as DBASM-WPR (Discriminative Bayesian Active Shape Models) mainly because it is the original referenced name [41]. This approach relies in the standard PDM based prior (section 3.3.1) and uses the second order LDS inference technique (section 3.5.1). The initial estimate is also included in the evaluation. The presented AVG 'curve' represents the initial 2D mesh location provided by the Adaboost [51] face detector.

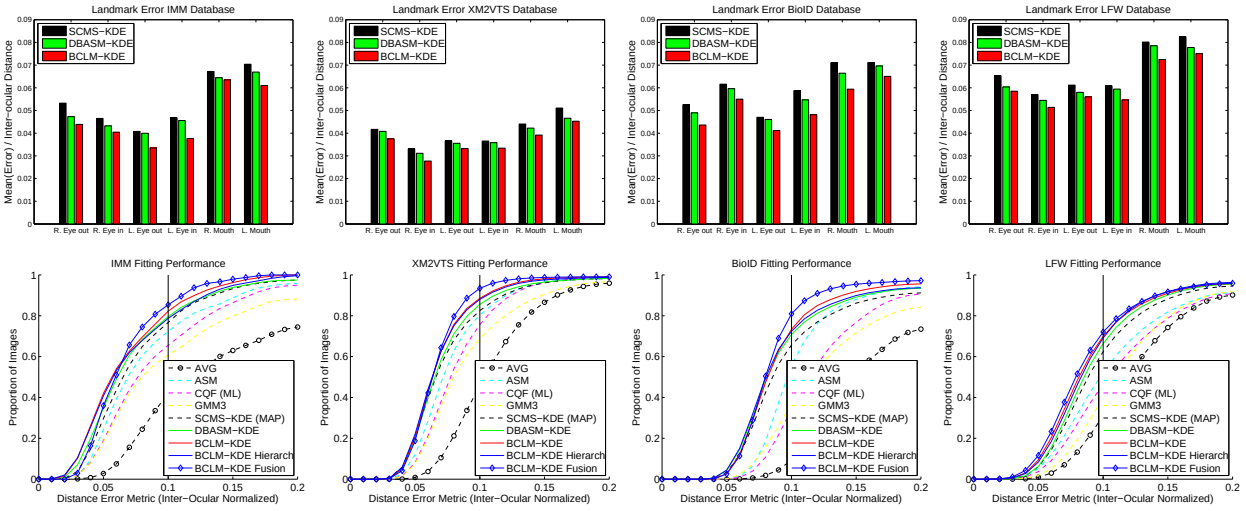
The results provide several conclusions: 1) the MOSSE filter always outperforms the others, specially when using simpler optimization methods; 2) the second order optimization improves the results even using simple detectors; 3) maximum performance can be achieved by using the MOSSE detector combined with the proposed optimization. The results show that the use of MOSSE filters is an interesting solution that works well in practice and is particularly suited to the detection of facial parts. However it is important to stress that it is not crucial for the performance of the Bayesian formulation. Our global optimization still improves performance when using standard linear-SVM detectors.

4.3 Evaluating Global Optimization Strategies

Performing a fair comparison requires that all the evaluated optimization strategies use the same local detector (same likelihood source), all methods are regularized by the same linear shape model (PDM) and they always start from the same initial estimate. In short, the BCLM optimization strategy was evaluated against similar CLM alignment solutions. The proposed BCLM and BCLM-Hierarchical methods (section 3.7) are compared against the ASM [1], CQF [13], BCQF [21], GMM [16] using 3 Gaussians (GMM3), SCMS (ML) [17] and finally the SCMS (MAP) [18]. The influence of modelling the prior distribution is included in the evaluation when comparing DBASM [41] with BCLM. Like in section 4.2, DBASM matches the proposed BCLM without the prior distribution modeling. The DBASM method uses the prior defined in section 3.3.1 whereas BCLM relies in the prior distribution defined in section 3.3.2. The BCQF is a maximum a posteriori version of CQF, likewise SCMS (ML) and SCMS (MAP) represent a maximum likelihood and maximum a posteriori versions of SCMS, respectively.

Once more, note that the BCLM optimization can be used with different local strategies to approximate the response maps (section 3.2.1), represented by the suffixes -WPR, -GR or -KDE. In fact, it is worth recall that the ASM [1], CQF [13] and SCMS [17] use as local optimizations the WPR, GR and KDE strategies, respectively. In all the local KDE methods, a bandwidth schedule of $\sigma_h^2 = (15, 10, 5, 2)$ is used (which applies to SCMS, DBASM and BCLM).

Both the shape model ($v = 58$ landmarks) and the MOSSE filters (proven to be the better detector) have been built with training images from the IMM [36] dataset. However, the results on this dataset use



(a) IMM database [36] (b) XM2VTS database [38] (c) BioID database [37] (d) LFW database [40]

Reference $e_m = 0.1$ (vertical line)	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)		LFW (13233 images)	
ASM [1]	72.3		80.3		55.5		52.2	
DBASM-WPR (○)	75.6	(+3.3)	83.4	(+3.1)	65.9	(+10.4)	67.9	(+15.7)
BCLM-WPR (⋆)	**78.5**	(+6.2)	**85.8**	(+5.5)	**70.2**	(+14.7)	70.4	(+18.2)
CQF (ML) [13]	65.6		75.7		34.3		46.7	
BCQF (MAP) [21]	67.8	(+2.2)	76.4	(+0.7)	36.2	(+1.9)	47.1	(+0.4)
GMM3 [16]	60.8	(-4.8)	68.8	(-6.9)	37.0	(+2.7)	40.0	(-6.7)
DBASM-GR (○)	68.4	(+2.8)	78.2	(+2.5)	39.4	(+5.1)	48.1	(+1.4)
BCLM-GR (⋆)	**69.2**	(+3.6)	**79.5**	(+3.8)	**40.1**	(+5.8)	48.6	(+1.9)
SCMS-KDE (ML) [17]	75.1		81.4		62.9		59.1	
SCMS-KDE (MAP) [18]	76.2	(+1.1)	82.5	(+1.1)	65.7	(+2.8)	62.4	(+3.3)
DBASM-KDE (○)	79.9	(+4.8)	85.6	(+4.2)	70.8	(+7.9)	66.4	(+7.3)
DBASM-KDE-Hierarchical (○)	80.3	(+5.2)	84.3	(+2.9)	70.2	(+7.3)	65.5	(+6.4)
BCLM-KDE (⋆)	**82.4**	(+7.3)	**88.6**	(+7.2)	**73.4**	(+10.5)	**69.9**	(+10.8)
BCLM-KDE-Hierarchical (⋆)	82.1	(+7.0)	88.1	(+6.7)	72.4	(+9.5)	69.2	(+10.1)
BCLM-KDE Fusion ($M = 2$) (⋆)	**85.4**	(+10.3)	**93.4**	(+12.0)	**80.9**	(+18.0)	**72.1**	(+13.0)

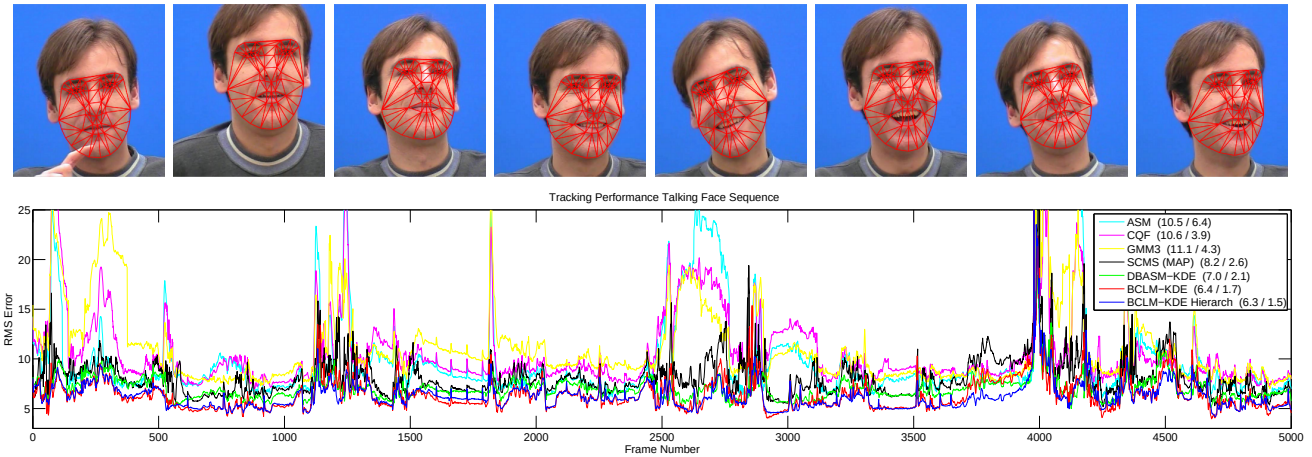
(⋆) → our method, (○) → our method w/o prior modeling (section 3.5.1)

Fig. 5: The bar charts display the (normalized) average location error of the most salient facial features in each dataset. The fitting performance curves are shown below. The table holds quantitative values taken by setting a fixed error amount ($e_m = 0.1$, i.e. the 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).

training images collected at our institution, which was done mainly due to incompatibility of the annotation format. In all cases, the nonrigid parameters started from zero, the similarity parameters were initialized by a face detection [51] (whose location appears as AVG in the evaluation charts, like before) and the model was fitted until convergence up to a maximum of 20 iterations.

Figure 5 shows a number of results: 1) Fitting performance curves, using the normalized inter-ocular error metric (Eq. 47), for the IMM, XM2VTS, BioID and LFW datasets, respectively; 2) A table with quantitative values taken by sampling the fitting curves using a fixed error metric amount ($e_m = 0.1$, shown as a vertical line in the graphics) and finally 3) a set of bar charts displaying the (inter-ocular) normalized average errors on the six most salient facial features (eyes and mouth corners) with the SCMS (MAP), DBASM and BCLM methods.

The results show that the CQF performs better than the 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 2). The BCQF is slightly better than CQF due to its improved parameter update (MAP update vs first order forwards additive). The same can be said between SCMS (ML) and SCMS (MAP). The MAP update penalizes large deformations of the shape model, being a proper regularization, whereas ML just makes unconstrained updates. The SCMS methods improve 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



RMS	ASM[1]	DBASM-WPR(○)	BCLM-WPR(⋆)
mean	10.47	7.78	7.18
std	6.42	2.19	2.09

RMS	CQF[13]	BCQF[21]	GMM3[16]	DBASM-GR(○)	BCLM-GR(⋆)
mean	10.64	9.91	11.15	9.17	9.06
std	3.92	2.76	4.33	2.10	2.07

RMS	SCMS[17]	SCMS(MAP)[18]	DBASM-KDE(○)	DBASM-KDE-H(○)	BCLM-KDE(⋆)	BCLM-KDE-H(⋆)	BCLM-KDE Fusion(⋆)
mean	9.45	8.16	7.03	7.16	6.39	6.26	5.89
std	3.58	2.57	2.09	2.16	1.68	1.50	1.45

(⋆) → our method, (○) → our method w/o prior modeling (section 3.5.1)

Fig. 6: Evaluation of the tracking performance of several fitting algorithms on the FGNET Talking Face [39] sequence. The values on legend box are the mean and standard deviation RMS errors, respectively. According, the table in the bottom shows a full comparative view of the same mean and standard deviation RMS errors but for all the evaluated algorithms. The top images show some BCLM-KDE fitting examples in the tested sequence. Best viewed in color.

Bayesian global optimization (BCLM) outperforms all previous methods. Explicitly modeling the prior distribution and using the covariance of the latent variables (that represent the confidence in the current parameters estimate) offers a significative increase in fitting performance. The effect of using an enhanced prior distribution is shown when DBASM versions are compared with BCLM. For instance, in the LFW dataset, which is the most representative of real world conditions, there is an improvement of 3.5% more converged images (while using a KDE local strategy). The results also show that the hierarchical annealing version of BCLM-KDE (BCLM-KDE-Hierarchical) has a comparable performance, but at the cost of more iterations.

Additionally, the Bayesian fusion of $M = 2$ detectors was also evaluated using the method that previously achieved the best performance (BCLM-KDE). The results of this approach (BCLM-KDE Fusion) show that including multiple sets of patch alignment classifiers further increase the fitting accuracy. In fact, this approach achieves the overall best results.

4.4 Tracking Performance

The tracking performance was evaluated in the FGNET Talking Face video sequence. As usual in this kind of experiments, each frame uses as initial estimate the previously estimated shape and pose parameters (the first frame starts with the mean shape, $\mathbf{b} = \mathbf{0}$, and the pose is initialized by a face detector [51]). Figure 6 shows the Root Mean Squared (RMS) error between the convergence of each algorithm and the ground truth annotation across the 5000 frames of the sequence. Once again, due to the incompatibility of landmark annotation, the error was only measured between the corresponded points. The same figure also shows, in the bottom, a comprehensive table with the RMS error mean and standard deviation on the full sequence for all the evaluated algorithms. The results show that the relative performance between all global optimization approaches is similar to the performance observed in previous experiments, where the BCLM technique confirms the best overall performance with the lowest RMS mean and standard deviation values.

Finally, the later figure 7 shows some qualitative evaluation results performed in the challenging Labeled Faces in the Wild (LFW) database [40] taken using the BCLM-KDE technique. Additional videos showing the performance of the proposed algorithms can be seen through the following link (videos).

5 Conclusions

This paper presents a novel and efficient Constrained Local Model (CLM) fitting solution to align facial parts in unseen images. A novel Bayesian formulation, aimed to solve the CLM global alignment problem in a *maximum a posteriori* (MAP) sense, is proposed. Two main technical insights are introduced. The first consists of inferring the posterior distribution of the global warp using a second order estimate of latent variables, accounting for the covariance of the shape

Fig. 7: Face alignment examples in the Labeled Faces the Wild dataset [40] taken using the BCLM-KDE fitting algorithm.

and pose parameters (by means of a Linear Dynamical System). The second main improvement relies in modelling the dynamic transitions of the PDM parameters, encoded by the prior distribution, using recursive Bayesian estimation techniques.

An extensive and thorough performance evaluation was conducted using several standard datasets (IMM, XM2VTS, BioID, Labeled Faces in the Wild and FGNET Talking Face sequence) where both local landmark detectors and global optimization strategies are compared. The performance evaluation, starts by demonstrating that the MOSSE correlation filters offer a superior landmark detection performance. Afterwards, the global optimization comparisons show that the proposed Bayesian approaches outperform other state-of-the-art CLM fitting solutions.

ACKNOWLEDGMENTS

This work was supported by the Portuguese Science Foundation (FCT) under the projects with grants: PTDC/EIA-CCO/108791/2008 and PTDC/EEA-CRO/122812/2010. P. Martins, J. Henriques and R. Caseiro also acknowledge the FCT through grants SFRH/BPD/90200/2012, SFRH/BD/75459/2010 and SFRH/BD/74152/2010, respectively.

REFERENCES

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[3] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 1, pp. 1080–1093, November 2005.

[4] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 135–164, November 2004.

[5] M. H. Nguyen and F. De la Torre, "Metric learning for image alignment," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 69–84, May 2010.

[6] P. Martins, R. Caseiro, and J. Batista, "Generative face alignment through 2.5d active appearance models," *Computer Vision and Image Understanding*, vol. 117, no. 3, pp. 250–268, March 2013.

[7] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Computer Graphics (ACM) SIGGRAPH*, 1999, pp. 187–194.

[8] Y. Zhou, L. Gu, and H.-J. Zhang, "Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[9] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *British Machine Vision Conference*, 2007.

[10] J. Saragih and R. Goecke, "A nonlinear discriminative approach to aam fitting," in *IEEE International Conference on Computer Vision*, 2007.

[11] P. Tresadern, H. Bhaskar, S. Adeshina, C. Taylor, and T. F. Cootes, "Combining local and global shape models for deformable object matching," in *British Machine Vision Conference*, 2009.

[12] D. Cristinacce and T. F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[13] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[14] P. Martins, R. Caseiro, and J. Batista, "Non-parametric bayesian constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[15] D. Bolme, J. Beveridge, B. Draper, and Y. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[16] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *European Conference on Computer Vision*, 2008.

[17] J. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *IEEE International Conference on Computer Vision*, 2009.

[18] ——, "Deformable model fitting by regularized landmark mean-shifts," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2010.

[19] Z. Xue, S. Z. Li, and E. K. Teoh, "Bayesian shape model for facial feature extraction and recognition," *Pattern Recognition*, vol. 36, no. 12, pp. 2819–2833, December 2003.

[20] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *British Machine Vision Conference*, 2006.

[21] U. Paquet, "Convexity and bayesian constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[22] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars,"

in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[23] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Computer Vision and Pattern Recognition*, 2012.
[24] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *IEEE Computer Vision and Pattern Recognition*, 2012.
[25] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *IEEE Computer Vision and Pattern Recognition*, 2014.
[26] T. F. Cootes, M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *European Conference on Computer Vision*, 2012.
[27] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[28] M. F. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *IEEE Computer Vision and Pattern Recognition*, 2010.
[29] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Computer Vision and Pattern Recognition*, 2013.
[30] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[31] X. Cheng, C. Fookes, S. Sridharan, J. Saragih, and S. Lucey, "Deformable face ensemble alignment with robust grouped-l1 anchors," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.
[32] B. M. Smith and L. Zhang, "Joint face alignment with non-parametric shape models," in *European Conference on Computer Vision*, 2012.
[33] X. Liu, "Discriminative face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1941–1954, November 2009.
[34] G. Fanelli, M. Dantone, and L. V. Gool, "Real time 3d face alignment with random forests-based active appearance models," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.
[35] G. Tzimiropoulos, J. A. i Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," in *Asian Conference on Computer Vision*, 2012.
[36] M. Nordstrom, M. Larsen, J. Sierakowski, and M. Stegmann, "The IMM face database - an annotated dataset of 240 face images," Technical University of Denmark, DTU, Tech. Rep., May 2004.
[37] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *International Conference on Audio and Video-based Biometric Person Authentication*, 2001.
[38] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *International Conference on Audio and Video-based Biometric Person Authentication*, 1999.
[39] FGNet, "Talking face video," 2004.
[40] G. Huang, M. Ramesh, T. Berg, and E.L.-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.
[41] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Discriminative bayesian active shape models," in *European Conference on Computer Vision*, 2012.
[42] ——, "Let the shape speak - discriminative face alignment using conjugate priors," in *British Machine Vision Conference*, 2012.
[43] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for computer vision," Imaging Science and Biomedical Engineering, University of Manchester, Tech. Rep., 2004.
[44] D. Comaniciu and P. Meer, "Mean Shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.
[45] C. Shen, M. Brooks, and A. Hengel, "Fast global kernel density mode seeking: Applications to localization and tracking," *IEEE Transactions On Image Processing*, vol. 16, no. 5, pp. 1457–1469, May 2007.
[46] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, 2004.
[47] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[48] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting with a mixture of local experts," in *IEEE International Conference on Computer Vision*, 2009.
[49] V. Rapp, K. Bailly, T. Senechal, and L. Prevost, "Multi-kernel appearance model," *Image and Vision Computing*, vol. 31, no. 8, pp. 542–554, 2013.
[50] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Likelihood-enhanced bayesian constrained local models," in *IEEE International Conference on Image Processing*, 2014.
[51] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, July 2002.
[52] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
[53] D. Cristinacce and T. F. Cootes, "Facial feature detection using adaboost with shape constraints," in *British Machine Vision Conference*, 2003.

Pedro Martins received both his M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Coimbra, Portugal in 2008 and 2012, respectively. Currently, he is a Postdoctoral researcher at the Institute of Systems and Robotics (ISR) in the University of Coimbra, Portugal. His main research include non-rigid image alignment, face tracking and facial expression recognition.

João F. Henriques received his M.Sc. degree in Electrical Engineering from the University of Coimbra, Portugal in 2009. He is currently a Ph.D. student at the Institute of Systems and Robotics, University of Coimbra. His current research interests include Fourier analysis and machine learning algorithms in general, with computer vision applications in detection and tracking.

Rui Caseiro received the B.Sc. degree in electrical engineering (specialization in automation) from the University of Coimbra, Portugal. Since 2007, he has been involved in several research projects, which include the European project Perception on Purpose and the National project 'Brisa-ITraffic'. He is currently a researcher with the Institute of Systems and Robotics and the Department of Electrical and Computer Engineering, University of Coimbra. His current research interests include the interplay of differential geometry with computer vision and pattern recognition.

Jorge Batista received the M.Sc. and Ph.D. degree in Electrical Engineering from the University of Coimbra in 1992 and 1999, respectively. He joins the Department of Electrical Engineering and Computers, University of Coimbra, Coimbra, Portugal, in 1987 as a research assistant where he is currently an Associate Professor. Jorge Batista is a founding member of the Institute of Systems and Robotics (ISR) in Coimbra, where he is a Senior Researcher and member of the Computer Vision and Robot Perception group. His research interest focus on a wide range of computer vision and pattern analysis related issues, including real-time vision, video surveillance, video analysis, non-rigid modeling and facial analysis. More recently his research activity also focus on the interplay of Differential Geometry in computer vision and pattern recognition problems. He has been involved on several national and European research projects, several of them as PI at UC.