Non-Parametric Bayesian Constrained Local Models

Pedro Martins, Rui Caseiro, Jorge Batista
Institute of Systems and Robotics, University of Coimbra, Portugal
{pedromartins,ruicaseiro,batista}@isr.uc.pt

Abstract

This work presents a novel non-parametric Bayesian formulation for aligning faces in unseen images. Popular approaches, such as the Constrained Local Models (CLM) or the Active Shape Models (ASM), perform facial alignment through a local search, combining an ensemble of detectors with a global optimization strategy that constraints the facial feature points to be within the subspace spanned by a Point Distribution Model (PDM). The global optimization can be posed as a Bayesian inference problem, looking to maximize the posterior distribution of the PDM parameters in a maximum a posteriori (MAP) sense. Previous approaches rely exclusively on Gaussian inference techniques, i.e. both the likelihood (detectors responses) and the prior (PDM) are Gaussians, resulting in a posterior which is also Gaussian, whereas in this work the posterior distribution is modeled as being non-parametric by a Kernel Density Estimator (KDE). We show that this posterior distribution can be efficiently inferred using Sequential Monte Carlo methods, in particular using a Regularized Particle Filter (RPF). The technique is evaluated in detail on several standard datasets (IMM, BioID, XM2VTS, LFW and FGNET Talking Face) and compared against state-of-the-art CLM methods. We demonstrate that inferring the PDM parameters non-parametrically significantly increase the face alignment performance.

1. Introduction

Facial alignment is a fundamental problem of computer vision (e.g. tracking, recognition, security, video compression, etc) which has been actively studied in the community with several degrees of success. Such alignment, also known as facial registration, is a key stage that has a huge impact on the robustness and quality of the later processes/applications. A widely used approach consists on seeking the parameters of a Point Distribution Model (PDM) that best represents the face in a target image. Traditionally, the proposed deformable face fitting methods can be divided in two major groups: generative (holistic) and discriminative (patch-based) approaches. In the generative paradigm, all the image pixels that describe the face are used to encode its appearance, typically using an eigen-based texture representation. The Active Appearance Models (AAM) [6, 22, 19] are probably the most popular generative method, achieving an impressive registration quality. However, this representation generalizes poorly beyond unseen data, when target individuals are not included in the training dataset. In recent years, there has been a growing interest on discriminative-based methods, such as the Constrained Local Models (CLM) [8, 9, 10, 32], as it circumvents several of the drawbacks of generative methods by improving the generic face representation. In this paradigm, both appearance and shape are combined by compelling a set of local feature detectors to lie within the subspace spanned by the PDM. In general, all instantiations of CLM are composed by a two phase fitting strategy. The first phase generates a response map for each PDM landmark (a likelihood map) using the local detectors. The second phase consists in a global optimization strategy that estimates the PDM parameters that jointly maximizes all the response maps at once. Most optimization strategies aim to approximate the responses maps by simple parametric forms (Weighted Peak Responses [8], Gaussians Responses [32, 26], Mixture of Gaussians [15]). However, due to the landmark's small support region and imperfect detectors (designed to be fast), some detection ambiguities exist. The Subspace Constrained Mean-Shift (SCMS) [27, 28] aims to deal with these ambiguities by using a non-parametric representation of the responses maps employing a Kernel Density Estimator (KDE). The mean-shift algorithm [5] is used to maximize over the KDE, and afterwards all the landmark updates are constrained to lie into the PDM subspace.

Recently a new paradigm emerged to solve the global optimization [26, 20, 21]. This new strategy suggests to formulate the global alignment as a Bayesian inference problem. The patch responses (likelihood) are embedded into a Bayesian framework, where the posterior distribution of

This work was supported by the Portuguese Science Foundation (FCT) under the project Differential Geometry for Computer Vision and Pattern Recognition (PTDC/EEA-CRO/122812/2010) and through grants SFRH/BPD/90200/2012 (Pedro Martins) and SFRH/BD74152/2010 (Rui Caseiro).

the global warp is inferred in a maximum a posteriori sense (MAP) [26, 20, 21]. The Bayesian CLM (BCLM) [26] makes basic inference of the PDM parameters using Gaussian assumptions of both likelihood and prior, leading to a posterior distribution that is also Gaussian. Afterwards, Discriminative Bayesian Active Shape Models (DBASM) [20] were proposed, where the alignment problem was formulated in terms of a Linear Dynamical System (LDS) using 2$^{\text{nd}}$ order updates of the parameters. The DBASM effectively accounts for the uncertainty on previous estimates as it models the covariance of the parameters. Although the increase in performance of DBASM, Gaussian inference techniques were still used, which in some scenarios can degrade alignment estimates (e.g. multimodal likelihood).

This work extends the CLM formulation generalizing the posterior distribution of the PDM parameters to be non-parametric, in particular, by a continuous KDE. Note that our approach is rather different than SCMS that just uses a non-parametric representation of the response maps (likelihood term). In SCMS, the landmark's mean-shift updates are just observations for the global optimization, which is a regularized projection onto the shape subspace. Here, the full non-parametric distribution of the response maps is accounted for, with the PDM parameters of the overall global alignment being inferred using Sequential Monte Carlo (SMC) methods [12, 11], in particular, with a Regularized Particle Filter (RPF) [24]. Additionally, KDE bandwidth selection is provided.

Recently, some remarkable face alignment techniques have been proposed, using: non-parametric shape models [1], part-based tree structure models [33] (providing face detection, pose estimation and feature localization), regression based shape updates [4], among others. We remark that these methods rely on non-parametric shape models and therefore should no be compared with ours. Nevertheless, this paper aims to extend the widely used CLM methodology, by non-parametric inference techniques, while maintaining its linear shape regularization model.

1.1. Main Contributions

1. This work presents a novel non-parametric Bayesian global optimization strategy that infers both the PDM and the pose parameters, in a *maximum a posteriori* (MAP) sense. Previous strategies make Gaussian assumptions of the posterior distribution, either making constant [26] or second order [20] predictions of the parameters. Here, we generalize the posterior distribution to be non-parametric, being approximated by a continuous KDE. We show that the posterior distribution of the global warp can be efficiently inferred using a Regularized Particle Filter (RPF) [24].

2. Previous PDM alignment methods evaluate the likelihood terms assuming conditional independence between landmarks [8, 32, 26, 27, 20, 21]. Typically, the likelihood parameters (landmark observations) are found by individual/independent parametric or non-parametric representations of the landmark's response maps. The approach presented in this paper does not make conditional independence assumptions and does not require a so-called local landmark optimization.

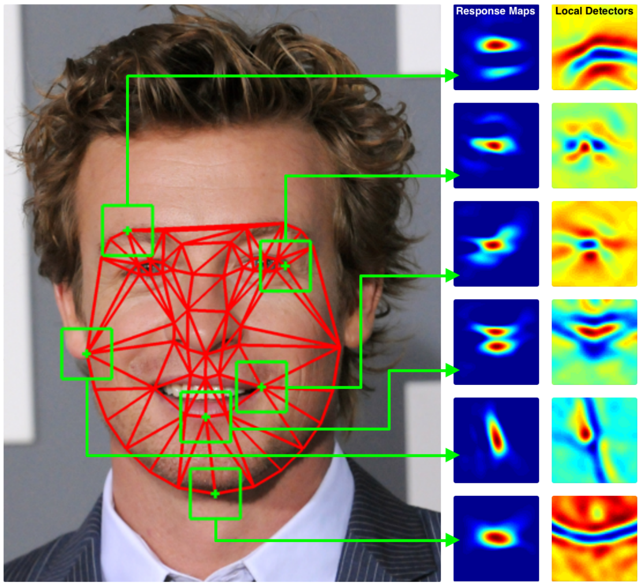


Figure 1. In Constrained Local Models (CLM) an ensemble of local feature detectors are constrained to lie in the span of a Point Distribution Model (PDM). The proposed global optimization strategy infers the PDM parameters, in a MAP sense, using a non-parametric posterior distribution. The left image shows the search regions for some highlighted landmarks, followed by a column with the detectors responses maps and their local detectors (MOSSE filters [3]), respectively.

3. Extensive evaluations were performed on a range of standard datasets (IMM [25], BioID [18], XM2VTS [23], FGNET Talking Face [13]) and the challenging LFW [16]) against state-of-the-art methods, while using the same local landmark detectors. We show that aligning the PDM using a non-parametric Bayesian approach offers a significative increase in performance.

The remaining of the paper is organized as follows: section 2 briefly explains the basics in ASM/CLM design. Section 3 revisits the existing methods of Bayesian global alignment and our non-parametric global optimization is presented in section 4. Sections 5 and 6 present the experimental evaluation and the conclusions, respectively.

2. Background

2.1. Linear Shape Model

The shape $\mathbf{s}$ of a Point Distribution Model (PDM) [7] with v landmarks is represented by a vector with the 2D vertex locations of a mesh $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$. In essence the PDM describes a shape by the following linear parametric model

$$\mathbf{s} = \mathcal{S}\left(\mathbf{s}_0 + \Phi\mathbf{b}, \mathbf{q}\right) \tag{1}$$

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%), $\mathbf{b}$ is a vector of shape parameters and $\mathcal{S}(., \mathbf{q})$ represents a similarity transformation function of the $\mathbf{q} = [s,\ \theta,\ t_x,\ t_y]^T$ pose parameters (s, θ, t_x, t_y are the scale, rotation and translations w.r.t. the base mesh $\mathbf{s}_0$, respectively). Refer to [7] for additional details in PDMs.

2.2. Local Detectors

The appearance model of an ASM/CLM consists of an ensemble of v local detectors [20, 27] (see figure 1). The correlation of the j^{th} landmark detector, evaluated at the pixel location $\mathbf{x}_j = (x_j, y_j)$, is given by

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

where $\mathbf{h}_j$ is a linear detector and $\mathbf{I}(\mathbf{x}_j)$ is a surrounding $L \times L$ support region (image patch, denoted by $\mathbf{\Omega}_{\mathbf{x}_j}$). Note that the landmark detectors are usually designed to operate at a given scale. The 2D 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_j to be a binary variable that denotes correct landmark alignment, the probability of pixel $\mathbf{z}_j \in \mathbf{\Omega}_{\mathbf{x}_j}$ being aligned is given by

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

where β_1 and β_0 are the regression coefficient and intercept, respectively. Note that a proper probability is used, always non-negative and $p(a_j = 1|\mathbf{I}(\mathbf{z}_j)) + p(a_j = -1|\mathbf{I}(\mathbf{z}_j)) = 1$.

3. Existing Global Optimization Strategies

In a Bayesian setting [20, 26], the optimal shape parameters $\mathbf{b}^*$ are given by the Bayes' theorem, where we seek to maximize the following posterior probability

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

with $\mathbf{y}$ being a $2v$ vector that represents the observed shape (measurement), $p(\mathbf{y}|\mathbf{b})$ is the likelihood term (that comes from the response maps) and $p(\mathbf{b})$ is the prior term that defines the knowledge of the model (the PDM). Conditional independence between landmarks is usually assumed, sampling each landmark independently, hence the overall likelihood becomes the individual contribution for each landmark, $p(\mathbf{y}|\mathbf{b}) \approx \prod_{j=1}^{v} p(\mathbf{y}_j|\mathbf{b})$.

3.1. Likelihood Term

In general, previous approaches define the likelihood term by the following Gaussian form

$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))^T \Sigma_{\mathbf{y}}^{-1} (\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))\right) \tag{5}$$

where $\Sigma_{\mathbf{y}}$ is the uncertainty of the spacial localization of the landmarks (being a $2v \times 2v$ block diagonal covariance matrix due to the conditional independence between landmarks assumed). Note that the shape measurement is done w.r.t. the base mesh $\mathbf{s}_0$. In general, the existing fitting approaches differ from each other in the way that the shape measurement $\mathbf{y}$ and its uncertainty $\Sigma_{\mathbf{y}}$ are obtained from the response maps. These methods can be considered as *local optimization strategies* and the most used are:

Active Shape Models (ASM): The first and most simple solution is to set each candidate to the localization where the response map has its maximum score [8]. The uncertainty is set to be inverse proportional to the peak value.

Convex Quadratic Fitting (CQF): The authors in [32] extend the ASM by approximating the response maps by a full Gaussian distribution. Generically, this means that the 'shape' of the response maps carries more useful information than just the amount of the detector score. This problem reduces to fitting a 2D Gaussian to weighted data.

Subspace Constrained Mean-Shifts (SCMS): In SCMS [27] the response maps were approximated by a non-parametric representation using a Kernel Density Estimator (KDE) [29] (isotropic Gaussian kernels with a given bandwidth). The mean-shift algorithm [5], with a decreasing annealing bandwidth schedule, was used to maximize over the KDE. In the original formulation [27] each shape observation consists of individual mean-shift landmark updates and the uncertainty was given by $\Sigma_{\mathbf{y}} = \mathbf{I}_{2v}$, this means that all landmarks will have the same weight and therefore contribute equally to the solution. Later in [28], a robust norm (Geman-McClure) was used to select the most reliable landmarks. Finally, the authors in [20] model the uncertainty of each mean-shift update by a full 2D Gaussian distribution.

3.2. Prior Term

By definition [30], the shape parameters $\mathbf{b}$, follow a multivariate Gaussian distribution $\mathbf{b} \propto \mathcal{N}(\mathbf{b}|\mathbf{0}, \Lambda)$, with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where λ_i denotes the PCA eigenvalue of the i^{th} mode of deformation. The prior term is then defined as

$$p(\mathbf{b}) \propto \mathcal{N}(\mathbf{b}|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \tag{6}$$

where $\mu_{\mathbf{b}} = \mathbf{0}$ and $\Sigma_{\mathbf{b}} = \Lambda$. The pose parameters (similarity) are modeled using a non-informative (uniform) prior.

3.3. Global MAP Solution

When the likelihood and the prior terms are both Gaussian distributions, the Bayes' theorem for Gaussian variables [2, 26, 20] states that the posterior is also a Gaussian distribution. Accordingly, the posterior is given by

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

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\Phi^T\Sigma_{\mathbf{y}}^{-1}\mathbf{y} + \Sigma_{\mathbf{b}}^{-1}\mu_{\mathbf{b}})$ and $\boldsymbol{\Sigma} = (\Sigma_{\mathbf{b}}^{-1} + \Phi^T\Sigma_{\mathbf{y}}^{-1}\Phi)^{-1}$. These equations are iteratively reused, where subscript k represents the iteration number, along with the response maps evaluated at the new updated locations, until convergence. Note that, the prior term is kept unchanged.

3.4. Inference by a Linear Dynamic System (LDS)

Faces are nonrigid structures that are described by continuous dynamic transitions, i.e. faces deform continuously in time. This constraint was exploited in DBASM [20], where global alignment was formulated in terms of a Linear Dynamic System (LDS). The LDS recursively computes a Gaussian posterior probability using incoming (also Gaussian) measurements and a linear model process. The state and measurement equations can be written as

$$\begin{aligned} \mathbf{b}_k &= \mathbf{I}_n\mathbf{b}_{k-1} + q & (8)\\ \mathbf{y} - \mathbf{s}_0 &= \Phi\mathbf{b}_k + r & (9) \end{aligned}$$

where is assumed that previous shape estimated parameters $\mathbf{b}_{k-1}$ are connected to the current parameters $\mathbf{b}_k$ by an identity relation $\mathbf{I}_n$ with noise. $q \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{b}})$ is the additive dynamic noise, $(\mathbf{y} - \mathbf{s}_0)$ is the observed shape deviation from the base mesh (related to the shape parameters by the linear relation Φ in eq.1) and r is the additive measurement noise following $r \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{y}})$. The LDS inference accounts with an adaptive prior, where the posterior distribution follow

$$p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k^{\mathbf{F}}, \boldsymbol{\Sigma}_k^{\mathbf{F}}) \tag{10}$$

with the mean $\boldsymbol{\mu}_k^{\mathbf{F}}$ and covariance $\boldsymbol{\Sigma}_k^{\mathbf{F}}$ given by the well-known Kalman Filter equations.

4. Non-Parametric Global Optimization

In the general case, the likelihood term $p(\mathbf{y}_k|\mathbf{b}_k)$ instead of being Gaussian, as in eq.5, it can be considered to be multimodal (or non-parametric), hence the posterior distribution $p(\mathbf{b}_k|\mathbf{y}_k)$ can not be analytically computed since no closed-form solution exists (opposed to the approaches reviewed in sections 3.3 and 3.4). A popular solution is to use Sequential Monte Carlo (SMC) methods [12, 11], which are also known as Particle Filters. These methods effectively allow to perform (approximate) inference of the posterior distribution when the likelihood (and prior) are arbitrary distributions. The most basic form of particle filtering is called Sequential Importance Sampling (SIS) where the main idea is to approximate the posterior distribution with a set of N weighted samples $\{w_k^{(i)}, \mathbf{b}_k^{(i)}\}_{i=1}^N$ (or particles). These particles are drawn from a proposal distribution [11] and recursively updated to obtain an approximation to the posterior distribution of the form $p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \approx \sum_{i=1}^N w_k^{(i)}\delta(\mathbf{b}_k - \mathbf{b}_k^{(i)})$ where $\delta(.)$ is the Dirac delta function centered at the particle $\mathbf{b}_k^{(i)}$. Each particle $\mathbf{b}_k^{(i)}$ represents a possible shape (according to eq.1) and its weight $w_k^{(i)}$ represents its fitting quality. The number of particles N is typically chosen as a trade-off between computational effort and estimation accuracy. In practice, the SIS filter leads to a *degeneracy problem* where only just a few particles will have a significant weight and all other particles will have very small weights. This degeneracy problem is typically dealt with resampling strategies.

The Sampling Importance Resampling (SIR) [14] also known as Bootstrap or Condensation [17] filter is a variant of SIS where the proposal distribution is taken form the state transition $p(\mathbf{b}_k|\mathbf{b}_{k-1})$ and resampling is applied at every iteration, in which the overall PDM alignment reduces to the following update equations

$$w_k^{(i)} \propto p(\mathbf{y}_k|\mathbf{b}_k^{(i)}) = \rho\left(\prod_{j=1}^{v} p(a_j = 1|\mathcal{D}_j, \mathbf{I}(\mathbf{y}_j)); \sigma\right) \tag{11}$$

$$\mathbf{b}_k^{(i)} \sim p(\mathbf{b}_k|\mathbf{b}_{k-1}^{(i)}) \propto \mathcal{N}(\mathbf{b}_k|\mathbf{b}_{k-1}, \Sigma_{\mathbf{b}}) \tag{12}$$

where the weight of particle $w_k^{(i)}$ is given by a robust measure of its alignment. The alignment metric is defined as the combined product of all landmarks $\prod_{j=1}^{v} p(a_j = 1|\mathcal{D}_j, \mathbf{I}(\mathbf{y}_j))$ (for the sake of computational stability is preferable to express alignment in terms of log-likelihoods, or $-\sum_{j=1}^{v}\log(p(a_j = 1|\mathcal{D}_j, \mathbf{I}(\mathbf{y}_j)))$), being $\rho(.;\sigma)$ a robust error norm and σ the scale parameters. Although several norms can be used (e.g. the Tukey's biweight, the Huber, or the Geman-McClure function), we simply use a nonlinear function that discards a given percentage of the worst scored landmarks (the scale parameter σ is the threshold e.g. 5%). In essence, a set of N possible/likely shapes are drawn from the eq.12, following the same dynamic model from section 3.4 and being weighted according to eq.11. It is worth saying that better results were found by exploring the search space by one shape parameter at a time.

The degeneracy problem is efficiently addressed using the SIR filter (resampling every iteration), although a new problem arises: *the sample impoverishment*, i.e. particles with large weights are likely to be drawn multiples times during resampling, whereas particles with small weights are not likely to be drawn at all (a lack of diversity problem).

Modified particle filtering algorithms have been suggested to handle the sample impoverish effect. A potential solution is to use the Regularized Particle Filter (RPF)

[24, 11]. In general terms, the RPF consists of a modified SIR particle filter in which the resampling process is performed upon a density estimation. The RPF resamples from a continuous approximation of the probability density $p(\mathbf{b}_k|\mathbf{y}_k,\ldots,\mathbf{y}_0)$, which is obtained by using the Kernel Density Estimator method [29]

$$p(\mathbf{b}_k|\mathbf{y}_k,\ldots,\mathbf{y}_0) \approx \sum_{i=1}^{N} w_k^{(i)} K_h(\mathbf{b}_k - \mathbf{b}_k^{(i)}) \qquad (13)$$

where K_h is the kernel density centered at $\mathbf{b}_k^{(i)}$ and h is the kernel bandwidth. The kernel density is a symmetric probability density function defined on $\mathcal{R}^n$, that satisfies $\int K_h(\mathbf{b}_k)d\mathbf{b}_k = 1$, $\int \mathbf{b}_k K_h(\mathbf{b}_k)d\mathbf{b}_k = 0$ and $\int ||\mathbf{b}_k||^2 K_h(\mathbf{b}_k)d\mathbf{b}_k < \infty$. A popular choice is the Gaussian kernel. Moreover, the kernel bandwidth can be optimally chosen, in the Mean Integrated Square Error (MISE) sense, by using a whitening transformation. Whitening consists of applying a linear transformation to achieve unit covariance of the data. In particular, the particles $\mathbf{b}_k^{(i)}$ are changed to $\mathbf{A}^{-1}\mathbf{b}_k^{(i)}$ where $\mathbf{S} = \mathbf{A}\mathbf{A}^T$ is the ensemble covariance ($\mathbf{A}$ could be recovered by a square root factorization, e.g. Cholesky factorization). The kernel density reduces to the following rescaled regularization kernel $\frac{\det(\mathbf{A})^{-1}}{h^n} K\left(\frac{\mathbf{A}^{-1}\mathbf{b}_k}{h}\right)$ (n is the dimension of the shape parameters), where the optimal bandwidth (Gaussian density with unit covariance [29]) is given by

$$h_{\text{opt}} = \left(\frac{4}{2N(n+2)}\right)^{\frac{1}{n+4}}. \qquad (14)$$

Finally, we update the model's parameters using the expectation of the posterior distribution $p(\mathbf{b}_k|\mathbf{y}_k,\ldots,\mathbf{y}_0)$ which is given by the average of the RPF resampled particles, as

$$\hat{\mathbf{b}}_k = \int_{-\infty}^{+\infty} \mathbf{b}_k \; p(\mathbf{b}_k|\mathbf{y}_k,\ldots,\mathbf{y}_0) \; d\mathbf{b}_k = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{b}}_k^{(i)}. \quad (15)$$

Pose and shape parameters are treated in two different optimizations. In this context is preferable to optimize the pose and shape parameters independently, dealing with lower dimensional problems each time (nevertheless it can be done all at once using a lot more particles). The algorithm 1 summarizes the overall alignment.

```
1  Precompute: PDM ($\mathbf{s}_0$, $\Phi$) and landmark detectors $\mathcal{D}_j$
2  Get an initial estimate of the shape/pose parameters ($\mathbf{b}_0, \mathbf{q}_0$)
3  repeat
4      Warp image $\mathbf{I}$ (into $\mathbf{s}_0$) using current pose parameters $\mathbf{q}$
5      for Landmark $j = 1$ to $v$ do
6          Eval response map: $p(a_j = 1|\mathcal{D}_j, \mathbf{I}(\mathbf{z}_j))$; $\mathbf{z}_j \in \mathbf{\Omega}_{\mathbf{x}_j}$
7      end
8      for Particle $i = 1$ to $N$ do
9          $\mathbf{b}_k^{(i)} = \mathbf{b}_{k-1}^{(i)} + d$;   $d \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{b}})$
10         Generate shape for the $i^{th}$ Particle: $\mathbf{s} = \mathbf{s}_0 + \Phi\mathbf{b}_k^{(i)}$
11         $w_k^{(i)} = \rho\left(-\sum_{j=1}^{v} \log(\, p(a_j = 1|\mathcal{D}_j, \mathbf{I}(\mathbf{s}_j))\,);\ \sigma\right)$
12     end
13     Normalize weights:   $\bar{w}_k^{(i)} = w_k^{(i)}(\sum_i^N w_k^{(i)})^{-1}$
14     $[\widetilde{\mathbf{b}}_k^{(i)}] \leftarrow$ Regularized Particle Filter resample $[\bar{w}_k, \mathbf{b}_k^{(i)}]$
15     Update model's parameters:   $\hat{\mathbf{b}}_k = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{b}}_k^{(i)}$
16 until $||\hat{\boldsymbol{b}}_k - \hat{\boldsymbol{b}}_{k-1}|| \leq \varepsilon$ or maximum number of iterations reached ;
```

Algorithm 1: Non-parametric Bayesian Constrained Local Models (npBCLM) algorithm. Note: optimizing the pose parameters require applying similar steps 7 to 15. The current MatLab implementation takes around 1 second per image.

5. Evaluation Results

The experimental evaluation was conducted in several standard databases with publicly available ground truth, namely: **(1)** The IMM [25] database that consists on 240 annotated images of 40 different individuals presenting different head pose, illumination, and facial expression (58 landmarks). **(2)** The BioID [18] dataset contains 1521 images, each showing a near frontal view of a face of 23 subjects (20 landmarks). **(3)** The XM2VTS [23] database has 2360 images frontal faces of 295 subjects (68 landmarks). **(4)** The Labeled Faces in the Wild (LFW) [16] database (12 landmarks) that contains images taken under variability in pose, lighting, facial expression, occlusions, etc. **(5)** Finally, tracking performance is also evaluated in the FGNet Talking Face [13] sequence that holds 5000 frames of video of an individual engaged in a conversation (68 landmarks).

As in [1], the overall face alignment challenge was evaluated in the different datasets, by creating 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 2 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.

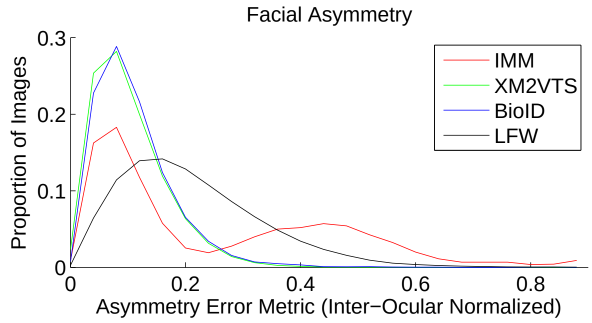
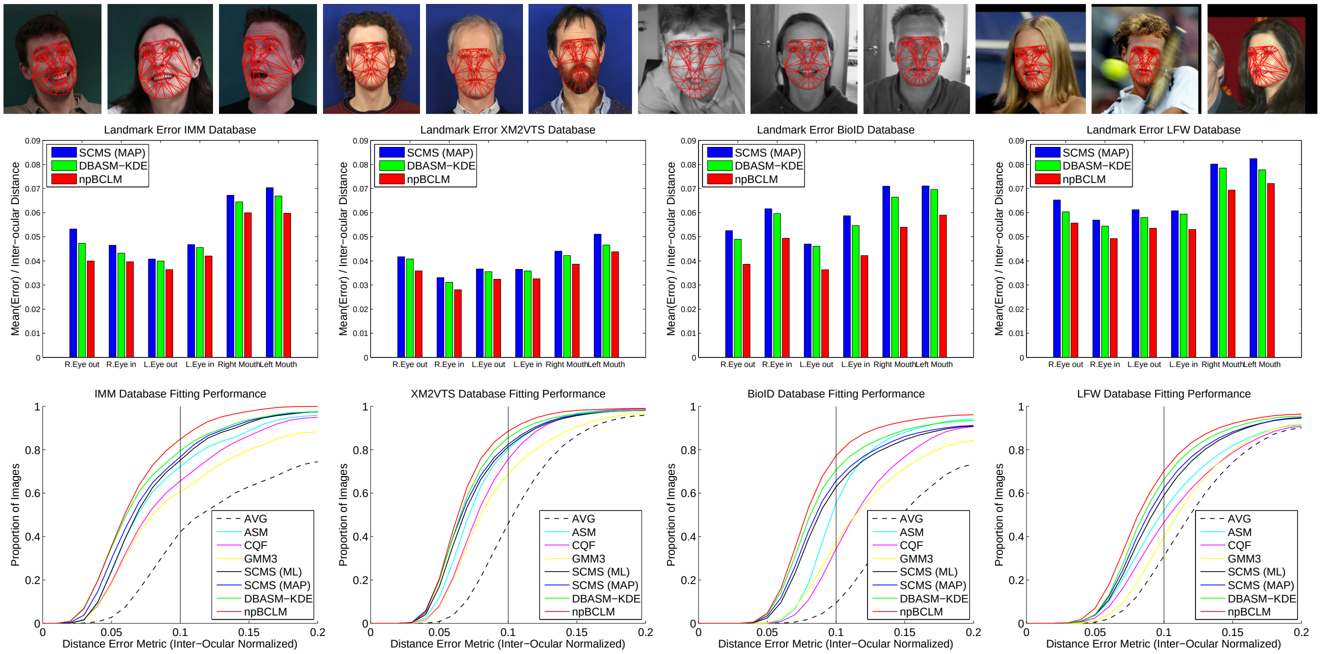


Figure 2. Distribution of face asymmetry in the evaluated datasets.



Reference $e_m = 0.1$ (vertical line)	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)		LFW (13233 images)	
ASM [8]	72.3		80.3		55.5		52.2	
CQF [32]	65.6		75.7		34.3		46.6	
BCLM [26]	67.1		76.4		35.8		48.2	
GMM3 [15]	60.8		68.8		37.0		39.8	
SCMS (ML) [27]	75.1	(0)	81.4	(0)	62.9	(0)	59.0	(0)
SCMS (MAP) [28]	76.7	(+1.6)	82.5	(+1.2)	65.7	(+2.8)	62.3	(+3.3)
DBASM-KDE [20]	79.5	(+4.4)	85.6	(+4.2)	70.8	(+7.9)	66.4	(+7.4)
npBCLM (our method)	**84.9**	(+9.7)	**88.6**	(+7.2)	**77.3**	(+14.3)	**70.2**	(+11.2)

Figure 3. 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) error than the reference.

5.1. Fitting Performance

In this section we aimed to make a fair comparison, making sure that all the evaluated optimization strategies use the same local detector (i.e. the same likelihood source) and are regularized by the same linear shape model. Therefore the evaluation was made against similar CLM solutions. All the experiments use the Minimum Output Sum of Squared Error (MOSSE) [3] filters as local landmark detectors, which has proven to perform better that most popular detectors [20], in particular when compared with linear classifiers built from aligned (positive) and misaligned (negative) examples [32, 27]. Both the shape model ($v = 58$ landmarks) and MOSSE filters have been built with training images from the IMM [25] dataset (however the results in this dataset use training images collected at our institution). The desired MOSSE correlation output (see [3]) was set to be a 2D Gaussian centered at the each landmark with 3 pixels of standard deviation. Each filter $\mathbf{h}_j$ has the size of 51×51 and it was used to scan a local region of 25×25 (i.e. size of the response maps in eq.3 - see figure 1). As described in section 4, the number of particles N is typically chosen as a trade-off between computational effort and estimation accuracy. A total of 2000 particles (100 per parameter) were used to estimate the shape parameters and 400 for the pose parameters. Notice that we can speed up the entire alignment procedure by rejecting particles whose shape falls outside of the local landmark search regions. In the experiments the robust norm $\rho(.;\sigma)$ discards the 7% worst scored landmarks (4 in a total of 58).

Our non-parametric Bayesian CLM approach, referred as npBCLM, was evaluated against standard and state-of-the-art global alignment solutions, in particular, the ASM [8], CQF [32], BCLM [26], GMM [15] using 3 Gaussians (GMM3), SCMS (ML) [27], SCMS (MAP) [28] and DBASM-KDE [20]. The BCLM 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. DBASM-KDE is the technique reviewed in section 3.4 where the KDE suffix means that the local optimization is based on a KDE representation of the response maps. A bandwidth schedule of $(15, 10, 5, 2)$ is used for the local KDE methods (which ap-

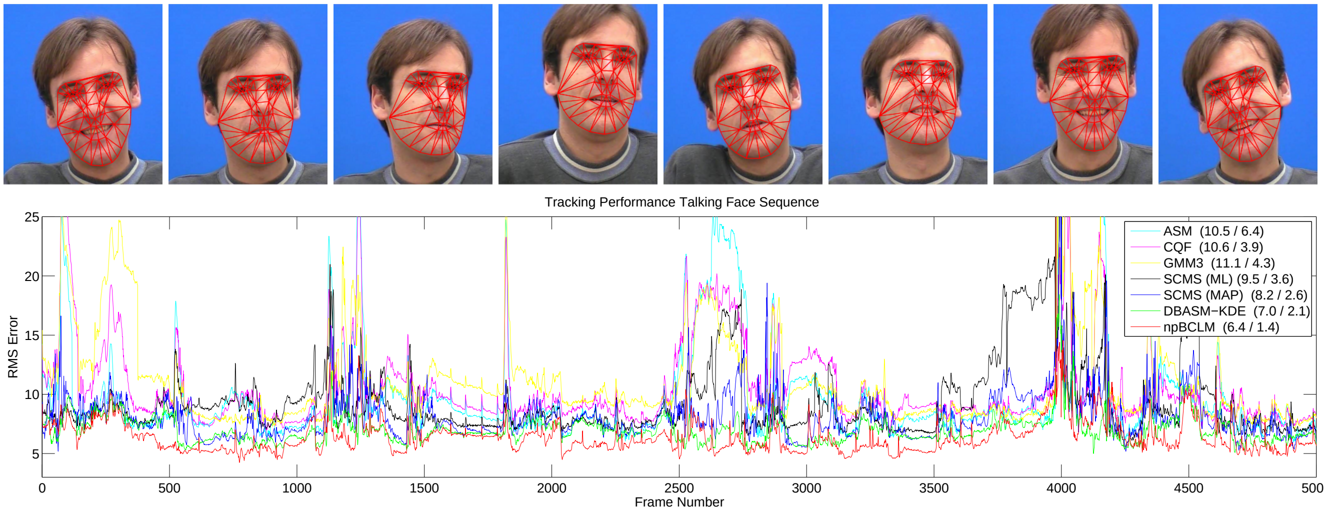


Figure 4. Tracking performance evaluation of several fitting algorithms in the FGNET Talking Face [13] sequence. The values on legend box are the mean and standard deviation RMS errors, respectively. The images on top show some fitting examples using our approach.

plies to SCMS and DBASM). Notice that our approach does not require tuning a kernel bandwidth (eq.14). In all cases, the initial shape parameters $\mathbf{b}_0$ start from zero, the pose parameters were initialized by a face detector [31] (whose location appears as 'AVG' method in the evaluation charts) and the model was fitted until convergence up to a maximum of 20 iterations.

The Figure 3 shows the fitting performance curves for all the evaluated methods on four different datasets. These curves, that were widely adopted in [9, 10, 32, 28, 20], are cumulative distribution functions that show the percentage of faces that achieved a given error amount (shown at the horizontal axis). Following common practice [9, 10], the error metric is given by the mean error per landmark as fraction of the inter-ocular distance, d_{eyes}, as

$$e_m(\mathbf{s}) = \frac{1}{v\, d_{\text{eyes}}} \sum_{i=1}^{v} \|\mathbf{s}_i - \mathbf{s}_i^{\text{gt}}\| \tag{16}$$

where $\mathbf{s}_i^{\text{gt}}$ is the location of i^{th} landmark in the shape ground truth annotation. Note that, the available annotations are different between datasets (and between our PDM model), hence the error metric was only measured over the corresponding landmarks. The table presented in the same figure shows quantitative values taken from sampling the curves setting a fixed error metric amount ($e_m = 0.1$, shown as a vertical line in the graphics). Figure 3 also includes bar charts with the (inter-ocular normalized) average errors on the six most salient facial features (eyes and mouth corners).

The results show that the MAP based approaches perform better than theirs maximum likelihood (ML) counterparts (BCLM vs CQF and SCMS-MAP vs SCMS-ML) as the MAP update penalizes large deformations of the shape model (it is a proper regularization) whereas ML just makes unconstrained updates. The SCMS approaches, as expected, achieve a high accuracy granted by the mean-shift algorithm. The excellent performance of the regular ASM, is mostly justified by the good performance of the local detectors (MOSSE filters). The DBASM-KDE improves on the results of the SCMS techniques, mainly because it fully accounts for the uncertainty in the responses maps and it uses an enhanced parameter update. Our proposed nonparametric global optimization (npBASM) outperforms all previous methods. The expectation of a KDE posterior representation does in fact provide a more accurate PDM update, that accounts with the multimodal distribution of the response maps together with a robust alignment metric.

5.2. Tracking Performance

The figure 4 shows the tracking performance evaluation in the FGNET Talking Face video sequence [13]. In this evaluation, a Root Mean Square (RMS) error metric was used. The relative performance between the global optimization approaches is similar to the previous experiments, where the npBCLM technique yields the best performance (with the lowest RMS error mean and standard deviation). Finally, the figure 5 shows examples of the qualitative evaluation of our npBCLM approach in the LFW [16] dataset.

6. Conclusions

This work presents a novel Bayesian global optimization strategy that infers both the Point Distribution Model (PDM) and the pose parameters, in a MAP sense, using a non-parametric posterior distribution (Kernel Density Estimator). The overall inference is done by a Regularized Particle Filter. Extensive evaluations were performed on several standard datasets against state-of-the-art CLM methods demonstrating a significant increase in performance.

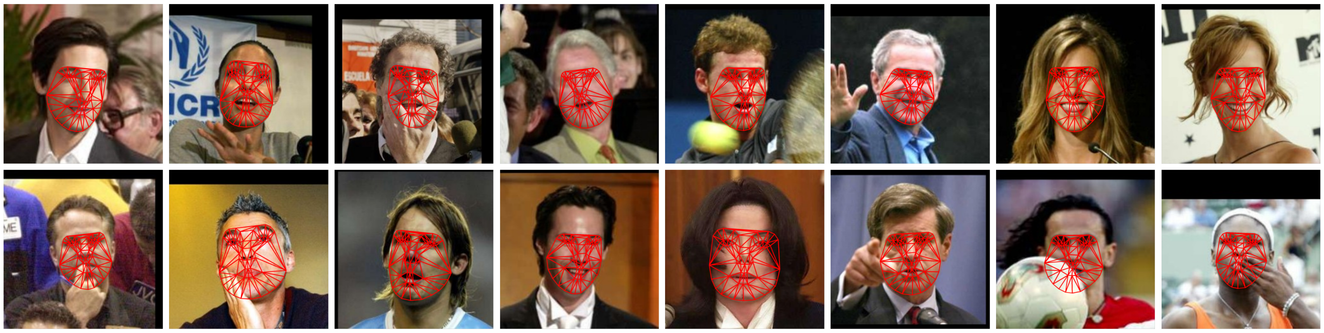


Figure 5. Qualitative npBCLM fitting results taken from the Labeled Faces in the Wild (LFW) dataset [16].

References

[1] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE CVPR*, 2011. 2, 5

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[3] D. Bolme, J. Beveridge, B. Draper, and Y. Lui. Visual object tracking using adaptive correlation filters. In *IEEE CVPR*, 2010. 2, 6

[4] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *IEEE CVPR*, 2012. 2

[5] D. Comaniciu and P. Meer. Mean Shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 24(5):603–619, May 2002. 1, 3

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, June 2001. 1

[7] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, Imaging Science and Biomedical Engineering, Univer. of Manchester, 2004. 2, 3

[8] T. Cootes, C. Taylor, D. Cooper, and J.Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995. 1, 2, 3, 6

[9] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007. 1, 7

[10] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 1, 7

[11] A. Doucet, N. de Feitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001. 2, 4, 5

[12] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Stat Comput*, 10(3):197–208, 2000. 2, 4

[13] FGNet. Talking face video, 2004. 2, 5, 7

[14] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Radar and Signal Processing*, pages 107–113, 1993. 4

[15] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *ECCV*, 2008. 1, 6

[16] G. 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, 2007. 2, 5, 7, 8

[17] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, 1996. 4

[18] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *AVBPA*, 2001. 2, 5

[19] P. Martins, R. Caseiro, and J. Batista. Generative face alignment through 2.5d active appearance models. *CVIU*, 117(3):250–268, 2013. 1

[20] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista. Discriminative bayesian active shape models. In *ECCV*, 2012. 1, 2, 3, 4, 6, 7

[21] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista. Let the shape speak - discriminative face alignment using conjugate priors. In *BMVC*, 2012. 1, 2

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

[23] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *AVBPA*, 1999. 2, 5

[24] C. Musso and N. Oudjane. Regularisation schemes for branching particle systems as a numerical solving method of the nonlinear filtering problem. In *ISSC*, 1998. 2, 5

[25] 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, May 2004. 2, 5, 6

[26] U. Paquet. Convexity and bayesian constrained local models. In *CVPR*, 2009. 1, 2, 3, 4, 6

[27] J. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *IEEE ICCV*, 2009. 1, 2, 3, 6

[28] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shifts. *IJCV*, 91(2):200–215, 2010. 1, 3, 6, 7

[29] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986. 3, 5

[30] M. E. Tipping and C. Bishop. Probabilistic principal component analysis. *JRSS*, 21(3):611–622, 1999. 3

[31] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, July 2002. 7

[32] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE CVPR*, 2008. 1, 2, 3, 6, 7

[33] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE CVPR*, 2012. 2