Int J Comput Vis (2011) 91: 200–215
DOI 10.1007/s11263-010-0380-4

Deformable Model Fitting by Regularized Landmark Mean-Shift

Jason M. Saragih · Simon Lucey · Jeffrey F. Cohn

Received: 12 May 2009 / Accepted: 10 September 2010 / Published online: 25 September 2010

Abstract Deformable model fitting has been actively pursued in the computer vision community for over a decade. As a result, numerous approaches have been proposed with varying degrees of success. A class of approaches that has shown substantial promise is one that makes independent predictions regarding locations of the model's landmarks, which are combined by enforcing a prior over their joint motion. A common theme in innovations to this approach is the replacement of the distribution of probable landmark locations, obtained from each local detector, with simpler parametric forms. In this work, a principled optimization strategy is proposed where nonparametric representations of these likelihoods are maximized within a hierarchy of smoothed estimates. The resulting update equations are reminiscent of mean-shift over the landmarks but with regularization imposed through a global prior over their joint motion. Extensions to handle partial occlusions and reduce computational complexity are also presented. Through numerical experiments, this approach is shown to outperform some common existing methods on the task of generic face fitting.

J.M. Saragih (✉)
ICT Center, CSIRO, Cnr Vimiera and Pembroke Rds, Sydney, NSW 2122, Australia
e-mail: jason.saragih@csiro.au

S. Lucey
ICT Center, CSIRO, 1 Technology Court Pullenvale, Brisbane, QLD 4069, Australia
e-mail: simon.lucey@csiro.au

J.F. Cohn
Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: jeffcohn@cs.cmu.edu

Keywords Deformable · Registration · Mean-shift

1 Introduction

Deformable model fitting is the problem of registering a parametrized shape model to an image such that its landmarks correspond to consistent locations on the object of interest. It is a difficult problem as it involves an optimization in high dimensions, where appearance can vary greatly between instances of the object due to lighting conditions, image noise, resolution and intrinsic sources of variability. Many approaches have been proposed for this problem with varying degrees of success. Of these, one of the most promising is that which models an object using local spatially-coherent image observations (i.e. image patches) centered around landmarks of interest within the object (Cootes and Taylor 1992; Cristinacce and Cootes 2004, 2006, 2007; Wang et al. 2008a). For computation and generalization purposes, these image patches are assumed to be conditionally independent of one another, an assumption that has shown superior performance in comparison to holistic approaches in recent literature (Liu 2007; Matthews and Baker 2004; Nguyen and De la Torre Frade 2008; Zhou and Comaniciu 2007). Local image patch detectors are typically learned, from labeled training images, for each landmark in the object. Due to their small local support and large appearance variation in training, however, these local detectors are plagued by the problem of ambiguity. This ambiguity can be observed in the non-parametric distribution of landmark locations (i.e., the response map) obtained from each landmark detector. The central dilemma addressed in this work is how to synergetically employ these non-parametric measures of the likely locations for each landmark, while limiting the effects of their ambiguity, when fitting the deformable model.

Our key contribution towards solving this dilemma lies in the realization that a number of popular optimization strategies are all, in some way, simplifying the non-parametric distribution of landmark locations obtained from each local detector parametrically. The motivation for this simplification is to ensure that the approximate objective function: (i) exhibits properties that make optimization efficient and numerically stable, and (ii) still approximately preserve the true certainty/uncertainty associated with each local detector. The question then remains: *how should one simplify these local distributions in order to satisfy (i) and (ii)*? In this work we propose a novel answer to this question that, unlike methods hitherto, does *not* require a parametric simplification, but still ensures an optimization that is efficient and numerically stable. Our non-parametric approach is reminiscent of the well known and understood mean-shift (Fukunaga and Hostetler 1975) mode seeking algorithm. The approach differs, however, from the traditional mean-shift algorithm as it is being applied over all landmarks simultaneously and also imposes a global prior over their joint motion. The resulting fitting algorithm is simple and efficient as well as affords significant improvements in convergence rate and accuracy over existing approaches.

We begin in Sect. 2 with a detailed overview of the problem of deformable model fitting with conditionally independent landmark detections. Many approaches are unified under a consistent formulation in which the observed behavior of the various approaches can be better understood. We present our approach in Sect. 3, which leverages on the formulation detailed in the previous section. Extensions to handle partial occlusions and reduce computational complexity of the proposed approach are also presented in this section. Empirical experiments, comparing the proposed approach against existing methods are presented Sect. 4. We conclude in Sect. 5 with a discussion and mention of future work.

2 Background

2.1 Problem Formulation

Most deformable model fitting methods employ a linear approximation to how the shape of a non-rigid object deforms, coined the point distribution model (PDM) by Cootes and Taylor (1992). It models non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

$$\mathbf{x}_i = s\mathbf{R}(\bar{\mathbf{x}}_i + \mathbf{\Phi}_i\mathbf{q}) + \mathbf{t}, \tag{1}$$

where $\mathbf{x}_i$ denotes the 2D-location of the PDM's i^{th} landmark and $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$ denotes the PDM parameters, which consist of a global scaling s, a rotation $\mathbf{R}$, a translation $\mathbf{t}$ and a set of non-rigid parameters $\mathbf{q}$. Here, $\bar{\mathbf{x}}_i$ denotes the mean location of the i^{th} PDM landmark in the reference frame (i.e. $\bar{\mathbf{x}}_i = [\bar{x}_i; \bar{y}_i]$ for a 2D model) and $\mathbf{\Phi}_i$ denotes the submatrix of the basis of variations, $\mathbf{\Phi}$, pertaining to that landmark. Such a model is both simple and efficient, and has been shown to adequately model the deformations of objects such as the human face (Cootes and Taylor 1992) and organs in medical image analysis (Zhou et al. 2005).

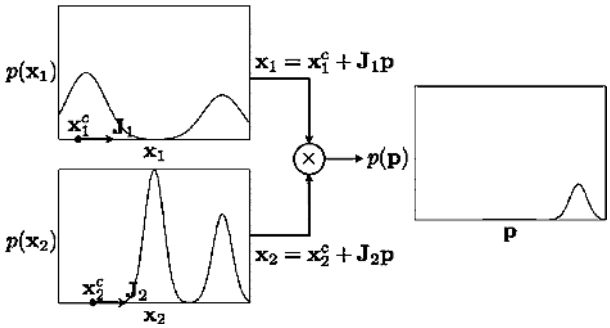

Fig. 1 An illustration of detection ambiguity reduction facilitated by a constraint on joint motion. The landmarks $\{\mathbf{x}_1, \mathbf{x}_2\}$ are constrained by the linear form: $\mathbf{x}_i = \mathbf{x}_i^c + \mathbf{J}_i\mathbf{p}$, which jointly parameterizes their locations with $\mathbf{p}$. Although the independent landmark likelihoods $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2)$ are multimodal, under the joint motion constraint the ambiguity is removed in: $p(\mathbf{p}) = p(\mathbf{x}_1)p(\mathbf{x}_2)$

In recent years, an approach that utilizes an independent set of local detectors for all PDM landmarks (see Cootes and Taylor 1992; Cristinacce and Cootes 2007, 2004, 2006; Wang et al. 2008a; Zhou et al. 2005, for example) has attracted some interest as it circumvents many of the drawbacks of holistic approaches, such as modeling complexity and sensitivity to lighting changes. The effects of ambiguous landmark detections, a result directly related to the limited support region assumed for detection, are reduced by virtue of the shape model that constrains the joint motion of these landmarks (see Fig. 1). Although we are primarily interested in approaches that utilize a statistical shape model, such as that in (1), the utility of such a framework has been demonstrated using more generic constraints in problems such as optical flow (Bruhn et al. 2005) and stereo matching (Sun et al. 2003), where constraints take the form of a smoothing process over the motion domain of landmarks. In this work, we will refer to these methods collectively as constrained local models (CLM).[1]

2.1.1 Fitting Objective

CLM fitting is generally posed as the search for the PDM parameters, $\mathbf{p}$, that jointly minimizes the misalignment error

[1] This term should not be confused with the work in Cristinacce and Cootes (2006) which is a particular instance of CLM in our nomenclature.

over all landmarks, regularized appropriately:

$$\mathcal{Q}(\mathbf{p}) = \mathcal{R}(\mathbf{p}) + \sum_{i=1}^{n} \mathcal{D}_i(\mathbf{x}_i; \mathcal{I}), \tag{2}$$

where $\mathcal{R}$ penalizes complex deformations (i.e. the regularization term) and $\mathcal{D}_i$ denotes the measure of misalignment for the i^{th} landmark at $\mathbf{x}_i$ in the image $\mathcal{I}$ (i.e. the data term). The form of regularization is related to the assumed distribution of PDM parameters describing plausible object shapes, common examples of which include the Gaussian (Basso et al. 2003) and Gaussian mixture model (GMM) (Gu and Kanade 2008) estimates. Examples of the misalignment error functions include the Mahalanobis distance over local patch appearance (Cootes and Taylor 1992) and the boosted Harr-like feature based classifier (Cristinacce and Cootes 2006).

Although it is possible to utilize general purpose optimization strategies to minimize (2), this is rarely done in practice. With the exception of tracking-targeted approaches, where $\mathcal{D}_i$ is often chosen as the least squares difference between the template and the image (Zhou et al. 2005), most variants of CLM fitting employ a specialized fitting strategy. One reason for this is that the misalignment error functions typically exhibit significant noise in the spatial domain of $\mathbf{x}_i$. As such, local deterministic optimization strategies, such as the Newton method, are often unstable. Stochastic optimization strategies, such as the simplex based method used in Cristinacce and Cootes (2004), are more stable since they do not make use of gradient information, which renders them somewhat insensitive to measurement noise. However, convergence may be slow when using these optimizers, especially for a complex PDM with a large number of parameters.

Since a landmark's misalignment error depends only on its spatial coordinates, an independent exhaustive local search for the location of each landmark can be performed efficiently (i.e. at all integer pixel locations around the estimated landmark locations). Therefore, most CLM variants implement a two step fitting strategy, where an exhaustive local search is first performed to obtain a *response map* for each landmark. Optimization is then performed over these response maps, which admit more sophisticated strategies compared to generic optimization methods that make no use of domain specific knowledge. An illustration of this two step procedure is presented in Fig. 2. It should be noted that this is made possible by the restricted search domains for $\{\mathbf{x}_i\}_{i=1}^{n}$, a condition specific to CLM's formulation. A detailed discussion of such strategies is presented in Sect. 2.2.

2.1.2 A Probabilistic Interpretation

The CLM objective in (2) can be interpreted as maximizing the likelihood of the model parameters such that all of its landmarks are aligned with their corresponding locations on the object in an image. The specific form of the objective implicitly assumes conditional independence between detections for each landmark, the probabilistic interpretation of which takes the form:

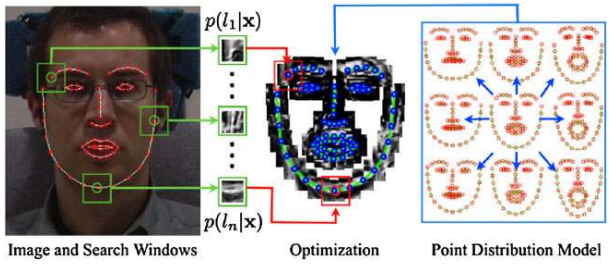

Fig. 2 Illustration of CLM fitting and its two components: (i) an exhaustive local search for feature locations to get the response maps $\{p(l_i = 1|\mathbf{x}, \mathcal{I})\}_{i=1}^{n}$, and (ii) an optimization strategy to maximize the responses of the PDM constrained landmarks

$$p(\mathbf{p}|\{l_i = 1\}_{i=1}^{n}, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^{n} p(l_i = 1|\mathbf{x}_i, \mathcal{I}), \tag{3}$$

where $l_i \in \{1, -1\}$ is a discrete random variable denoting wether the i^{th} landmark is aligned or misaligned. With this formulation, the regularization and misalignment error functions in (2) take the following forms:

$$\mathcal{R}(\mathbf{p}) = -\ln\{p(\mathbf{p})\} \tag{4}$$

$$\mathcal{D}_i(\mathbf{x}_i; \mathcal{I}) = -\ln\{p(l_i = 1|\mathbf{x}_i, \mathcal{I})\}. \tag{5}$$

To clarify exposition in the following sections, let us explicate the specific forms of the prior and likelihood in (3) utilized in this work. We model the likelihood of alignment at a particular landmark location, $\mathbf{x}$, as follows:

$$p(l_i = 1|\mathbf{x}, \mathcal{I}) = \frac{1}{1 + \exp\{l_i \mathcal{C}_i(\mathbf{x}; \mathcal{I})\}}, \tag{6}$$

where $\mathcal{C}_i$ denotes a classifier that discriminates aligned from misaligned locations. Notice that this likelihood is a proper probability mass function since it is non-negative everywhere, and:

$$p(l_i = 1|\mathbf{x}, \mathcal{I}) + p(l_i = -1|\mathbf{x}, \mathcal{I}) = 1. \tag{7}$$

For the classifier $\mathcal{C}_i$ we use the logistic regressor (Wang et al. 2008a):

$$\mathcal{C}_i(\mathbf{x}; \mathcal{I}) = \mathbf{w}_i^T \mathcal{P}(\mathcal{W}(\mathbf{x}; \mathcal{I})) + b_i, \tag{8}$$

where $\{\mathbf{w}_i, b_i\}$ respectively denote the gain and bias, and $\mathcal{P}(\mathbf{c})$ normalizes $\mathbf{c}$ to zero mean and unit variance. Here, $\mathcal{W}(\mathbf{x}; \mathcal{I})$ is an image patch:

$$\mathcal{W}(\mathbf{x}; \mathcal{I}) = [\mathcal{I}(\mathbf{z}_1); \ldots; \mathcal{I}(\mathbf{z}_P)]; \quad \{\mathbf{z}_i\}_{i=1}^{P} \in \mathbf{\Omega}_{\mathbf{x}}, \tag{9}$$

where $\mathbf{\Omega_x}$ denotes the set of integer pixel locations within a bounding box centered at $\mathbf{x}$. An advantage of using this classifier is that the response map can be computed using efficient convolution operations.

When assuming a non-informative (uniform) prior over the PDM parameters, the formulation in (3) leads to a Maximum Likelihood (ML) estimate, otherwise it leads to a Maximum *a-posterior* (MAP) estimate. When using a linear shape model attained by applying PCA to a set of registered shapes, the nonrigid shape parameters are often assumed to exhibit a Gaussian distribution, leading to the following prior:

$$p(\mathbf{p}) \propto \mathcal{N}(\mathbf{q}; \mathbf{0}, \mathbf{\Lambda}); \quad \mathbf{\Lambda} = \text{diag}\{[\lambda_1; \ldots; \lambda_m]\}, \tag{10}$$

where λ_i denotes the eigenvalue of the i^{th} mode of nonrigid deformation. Finally, a non-informative prior is commonly placed on the rigid transformation that places the model in the image frame, which assumes all rigid transformations are equally likely.

2.2 Existing Fitting Strategies

There are two sources of difficulty in optimizing (3): (i) how to avoid local optima whilst affording an efficient evaluation, and (ii) how to handle outlying detections. In the following sections, we show that existing optimization strategies entail replacing the true response maps with simpler parametric forms and performing optimization over these instead of the original response maps. The relative performance of these methods can be explained by the specific choice made for the parametric form in their approximations. As a general rule, the complexity of the approximated response maps dictates the computational cost of optimization and its sensitivity towards local minima as well as how faithful a representation it is of the true objective.

2.2.1 Isotropic Gaussian Estimate

The simplest optimization strategy for CLM fitting is that used in the Active Shape Model (ASM), first proposed by Cootes and Taylor (1992). The method entails first finding the location within each response map for which the maximum was attained: $\boldsymbol{\mu} = [\boldsymbol{\mu}_1; \ldots; \boldsymbol{\mu}_n]$. The objective of the optimization procedure is then to minimize the weighted least squares difference between the PDM and the coordinates of the peak responses, regularized appropriately:

$$\mathcal{Q}_{\text{ISO}}(\mathbf{p}) = \|\mathbf{q}\|^2_{\mathbf{\Lambda}^{-1}} + \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \boldsymbol{\mu}_i\|^2, \tag{11}$$

where the weights $\{w_i\}_{i=1}^n$ reflect the confidence over peak response coordinates and are typically set to some function of the responses at $\{\boldsymbol{\mu}_i\}_{i=1}^n$, making it more resistant towards such things as partial occlusion, where occluded landmarks will be more weakly weighted.

Equation (11) is iteratively minimized by taking a first order Taylor expansion of the PDM's landmarks:

$$\mathbf{x}_i \approx \mathbf{x}_i^c + \mathbf{J}_i \Delta\mathbf{p}, \tag{12}$$

and solving for the parameter update:

$$\Delta\mathbf{p} = -\mathbf{H}_{\text{ISO}}^{-1}\left(\tilde{\mathbf{\Lambda}}^{-1}\mathbf{p} + \sum_{i=1}^{n} w_i \mathbf{J}_i^T (\mathbf{x}_i^c - \boldsymbol{\mu}_i)\right), \tag{13}$$

which is then applied additively to the current parameters: $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$. Here, $\tilde{\mathbf{\Lambda}} = \text{diag}\{[\mathbf{0}; \lambda_1; \ldots; \lambda_m]\}$, $\mathbf{J} = [\mathbf{J}_1; \ldots; \mathbf{J}_n]$ is the PDM's Jacobian, $\mathbf{x}^c = [\mathbf{x}_1^c; \ldots; \mathbf{x}_n^c]$ is the current shape estimate, and:

$$\mathbf{H}_{\text{ISO}} = \tilde{\mathbf{\Lambda}}^{-1} + \sum_{i=1}^{n} w_i \mathbf{J}_i^T \mathbf{J}_i \tag{14}$$

is the Gauss-Newton Hessian.

From a probabilistic perspective, the ASM's optimization procedure is equivalent to approximating the response maps with isotropic Gaussian estimators[2]:

$$p(l_i = 1|\mathbf{x}_i, \mathcal{I}) \propto p(\mathbf{x}_i|l_i = 1, \mathcal{I}) \approx \mathcal{N}\big(\mathbf{x}_i; \boldsymbol{\mu}_i, \sigma_i^2\mathbf{I}\big), \tag{15}$$

where $w_i = \sigma_i^{-2}$. It is easily verified that substituting this approximation into (3) and taking its negative log results in the objective in (11).

2.2.2 Anisotropic Gaussian Estimate

Although the approximation described above is simple and efficient, in some cases it may be a poor estimate of the true response map. Firstly, the landmark detectors, such as the logistic regressor in (8), are usually imperfect in the sense that the maximum of the response may not always coincide with the correct landmark location. Secondly, as the features used in detection consist of small image patches they often contain limited structure, leading to detection ambiguities. A common example of this is the aperture problem, where detection confidence across the edge is better than along it (see example response maps for the nose bridge and chin in Fig. 3).

To account for these problems, a number of authors have proposed incorporating directional uncertainty into the response map estimate (see Nickels and Hutchinson 2002;

[2]The proportionality of $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$ and $p(\mathbf{x}_i|l_i = 1, \mathcal{I})$ stems from the assumption that $p(\mathbf{x}_i|\mathcal{I})$ is non-informative which is a direct result of assuming that all rigid transformations, which place the shape in the image frame, are equally likely.

Zhou et al. 2005; Wang et al. 2008a, for example). Similar to the approximation in (15), here the response maps are approximated by a full covariance Gaussian distribution:

$$p(l_i = 1|\mathbf{x}_i, \mathcal{I}) \approx \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{16}$$

The main difference between the various methods that utilize such an approximation is in how the mean and covariance are estimated. In Nickels and Hutchinson (2002), $\boldsymbol{\mu}_i$ is chosen as the maximum in the true response map, with the covariance set to the ML solution:

$$\boldsymbol{\Sigma}_i = \sum_{\mathbf{x}\in\boldsymbol{\Psi}_i} \frac{p(l_i = 1|\mathbf{x}, \mathcal{I})}{\sum_{\mathbf{y}\in\boldsymbol{\Psi}_i} p(l_i = 1|\mathbf{y}, \mathcal{I})} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T, \tag{17}$$

where $\boldsymbol{\Psi}_i$ is a 2D-rectangular grid over which the exhaustive local search is performed (i.e. the search window). In Wang et al. (2008a), a convex quadratic function was fit to the negative log of the response map, from which the mean and covariance of the approximating density can be inferred. In Zhou et al. (2005), where the summed-squared-difference is used as a measure of landmark fit, Laplace's approximation (Gelman et al. 1995) was used in conjunction with the small motion approximation to arrive at the scaled Grammian as the covariance estimate, with the mean defined as the ML optical flow solution and the scaling defined as the variance of appearance error at this solution.

Regardless of the strategy used in computing the anisotropic Gaussian estimate of the response map, by substituting this approximation into the objective in (3), the optimization problem can be written as the minimization of:

$$\mathcal{Q}_{\text{ANI}}(\mathbf{p}) = \|\mathbf{q}\|^2_{\boldsymbol{\Lambda}^{-1}} + \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}_i\|^2_{\boldsymbol{\Sigma}_i^{-1}}, \tag{18}$$

which can be solved for iteratively. The Gauss-Newton update for this objective takes the form:

$$\Delta\mathbf{p} = -\mathbf{H}^{-1}_{\text{ANI}} \left(\tilde{\boldsymbol{\Lambda}}^{-1}\mathbf{p} + \sum_{i=1}^{n} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_i^c - \boldsymbol{\mu}_i) \right), \tag{19}$$

where:

$$\mathbf{H}_{\text{ANI}} = \tilde{\boldsymbol{\Lambda}}^{-1} + \sum_{i=1}^{n} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i. \tag{20}$$

2.2.3 A Gaussian Mixture Model Estimate

Although the anisotropic Gaussian approximation of the response maps may overcome some of the drawbacks of its isotropic counterpart, its process of estimation can be poor in some cases. In particular, when the response map is strongly multimodal, such an approximation smoothes over the various modes (see the example response map for the eye corner in Fig. 3), limiting the fidelity of the resulting fit.

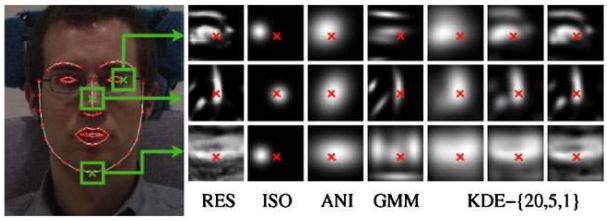

Fig. 3 Response maps, RES, and their approximations used in various methods, for the *outer left eye corner*, the *nose bridge* and *chin*. *Crosses* on the response maps denote the true landmark locations. The GMM approximation has five cluster centers. The KDE approximations are shown for $\rho \in \{20, 5, 1\}$

To account for this, in Gu and Kanade (2008) a Gaussian mixture model (GMM) was used to approximate the response maps:

$$p(l_i = 1|\mathbf{x}_i, \mathcal{I}) \approx \sum_{k=1}^{K_i} \pi_{ik} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \tag{21}$$

where K_i denotes the number of modes and $\{\pi_{ik}\}_{k=1}^{K_i}$ are the mixing coefficients for the GMM of the i^{th} landmark. Treating the mode membership for each landmark, $\{z_i\}_{i=1}^n$, as hidden variables, the maximum likelihood solution can be found using the expectation-maximization (EM) algorithm, which maximizes:

$$p(\mathbf{p}|\{l_i\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^{n} \sum_{k=1}^{K_i} p_i(z_i = k, l_i|\mathbf{x}_i, \mathcal{I}) \tag{22}$$

for $\{l_i = 1\}_{i=1}^n$.

The E-step of the EM algorithm involves computing the *posterior* distribution over the latent variables $\{z_i\}_{i=1}^n$:

$$p(z_i = k|l_i, \mathbf{x}_i, \mathcal{I}) = \frac{p(z_i = k) p(l_i|z_i = k, \mathbf{x}_i, \mathcal{I})}{\sum_{j=1}^{K_i} p(z_i = j) p(l_i|z_i = j, \mathbf{x}_i, \mathcal{I})}, \tag{23}$$

where $\{l_i = 1\}_{i=1}^n$, $p(z_i = k) = \pi_{ik}$ and:

$$p(l_i = 1|z_i = k, \mathbf{x}_i, \mathcal{I}) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}). \tag{24}$$

Letting $q(\mathbf{z}) = \prod_{i=1}^n p_i(z_i|l_i = 1, \mathbf{x}_i, \mathcal{I})$, the M-step of the EM algorithm involves minimizing the expectation of the negative log of the complete data:

$$\begin{aligned} \mathcal{Q}_{\text{GMM}}(\mathbf{p}) &= E_{q(\mathbf{z})} \left[-\ln \left\{ p(\mathbf{p}) \prod_{i=1}^{n} p(l_i = 1, z_i|\mathbf{x}_i, \mathcal{I}) \right\} \right] \\ &\propto \|\mathbf{q}\|^2_{\boldsymbol{\Lambda}^{-1}} + \sum_{i=1}^{n} \sum_{k=1}^{K_i} w_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_{ik}\|^2_{\boldsymbol{\Sigma}_{ik}^{-1}}, \end{aligned} \tag{25}$$

where $w_{ik} = p(z_i = k|l_i = 1, \mathbf{x}_i, \mathcal{I})$. This objective can be minimized iteratively using the Gauss-Newton optimization procedure, for which the updates take the form:

$$\Delta\mathbf{p} = -\mathbf{H}_{\text{GMM}}^{-1}\left(\tilde{\mathbf{\Lambda}}^{-1}\mathbf{p} + \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\Delta\mathbf{x}_{ik}\right), \quad (26)$$

where $\Delta\mathbf{x}_{ik} = \boldsymbol{\mu}_{ik} - \mathbf{x}_i^c$, and:

$$\mathbf{H}_{\text{GMM}} = \tilde{\mathbf{\Lambda}}^{-1} + \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\mathbf{J}_i. \quad (27)$$

Note that the Gauss-Newton optimization procedure does not guarantee convergence to the global minimum for non-linear least squares functions (for which this is one, due to the bilinear relationship between the pose and non-rigid parameters in the shape model). As such, this strategy is in fact the generalized EM algorithm (Dempster et al. 1977).

Although the GMM is a better approximation of the response map compared to a Gaussian approximation, it exhibits two major drawbacks. Firstly, the process of estimating the GMM parameters from the response maps is a non-linear optimization in itself. It is only locally convergent and requires the number of modes to be chosen *a-priori*. As GMM fitting is required for each PDM landmark, it constitutes a large computation overhead. Although some approximations can be made, they are generally suboptimal. For example, in Gu and Kanade (2008), the modes are chosen as the K_i-largest responses in the map. The covariances are parametrized isotropically, with their variance heuristically set as the scaled distance to the closest mode in the previous iteration of the fitting algorithm. Such an approximation allows an efficient estimate of the GMM parameters without the need for a costly EM procedure but with a poorer approximation of the true response map. The second drawback of the GMM response map approximation is that the approximated objective in (22) is almost always multimodal. As such, fitting with the GMM simplification is prone to terminating in local optima.

3 Regularized Landmark Mean-Shift

Due to the truncation used in PCA, the shape model can not perfectly reconstruct the true landmark locations in an image. The error in this estimate, which is assumed to originate from observation noise, is often modeled as homoscedastic isotropic Gaussian:

$$\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{\epsilon}_i; \mathbf{0}, \rho\mathbf{I}). \quad (28)$$

Here, ρ denotes the variance of the noise on landmark locations, which can be inferred from the training set as follows (Moghaddam and Pentland 1997):

$$\rho = \frac{1}{N-m}\sum_{i=m+1}^{N}\lambda_i, \quad (29)$$

which is simply the arithmetic average of the eigenvalues in the subspace orthogonal to $\mathbf{\Phi}$.

Let us assume that there exists a set of candidate locations for each landmark of the model that we denote $\{\mathbf{\Psi}_i\}_{i=1}^n$. This may be the case, for example, when the search is constrained to a local rectangular region, where $\mathbf{\Psi}_i$ denotes all integer pixel locations within this region. Treating the locations of the true landmarks as hidden variables, we marginalize them out of the likelihood that the landmarks are aligned:

$$p(l_i = 1|\mathbf{x}_i, \mathcal{I}) = \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} p(l_i = 1|\mathbf{y}_i, \mathcal{I})p(\mathbf{y}_i|\mathbf{x}_i), \quad (30)$$

where, from (28), we have:

$$p(\mathbf{y}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{y}_i; \mathbf{x}_i, \rho\mathbf{I}). \quad (31)$$

In (30), $p(l_i = 1|\mathbf{y}_i, \mathcal{I})$ denotes the likelihood that the i^{th} landmark is aligned at location $\mathbf{y}_i$ in image $\mathcal{I}$. The main difference between the formulation here and existing fitting strategies discussed in Sect. 2.2 is that the response maps are only evaluated at fixed locations defined through $\mathbf{y}_i \in \mathbf{\Psi}_i$, whereas optimization is performed over $\mathbf{p}$, which effects only $\{\mathbf{x}_i\}_{i=1}^n$. As such, rather than approximating the response map with a particular parametric form, here, a non-parametric estimate of the response map is made in the form of a homoscedastic isotropic Gaussian kernel density estimate (KDE) (Silverman 1986)[3]:

$$p(l_i = 1|\mathbf{x}_i, \mathcal{I}) = \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \pi_{\mathbf{y}_i}\mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho\mathbf{I}), \quad (32)$$

where $\pi_{\mathbf{y}_i} = p(l_i = 1|\mathbf{y}_i, \mathcal{I})$.

The quality of this nonparametric estimate of the response map depends largely on the choice of candidate landmark location sets $\{\mathbf{\Psi}_i\}_{i=1}^n$. If the candidates are sampled too sparsely, they may not adequately cover the space of variations and ρ, which is learned from training data as in (29), will be underestimated. However, since the space of $\mathbf{y}_i$ is the 2D image plane, it is computationally tractable to compute the likelihood of a dense set of candidates locally around the current PDM estimate through an exhaustive local search over all integer pixel locations.

Substituting (32) into (3), we get:

$$p(\mathbf{p}|\{l_i = 1\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p})\prod_{i=1}^{n}\sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \pi_{\mathbf{y}_i}\mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho\mathbf{I}). \quad (33)$$

[3] Since optimization is over the model parameters $\mathbf{p}$, which only effect $\{\mathbf{x}_i\}_{i=1}^n$, and not its fixed candidates $\{\mathbf{y}_i\}_{i=1}^n$, we have substituted the mean and variable of the Gaussian distribution from that in (31).

It is interesting to note that a graphical interpretation of CLM fitting using belief propagation (Yedidia et al. 2002) results in exactly in the same form for the marginal likelihood of the PDM parameters. An elaboration of such a perspective can be found in Appendix.

Equation (33) can be maximized using the EM algorithm, as in the case of the GMM approximated approach described in Sect. 2.2. Treating the true landmark locations $\{\mathbf{y}_i\}_{i=1}^n$ as hidden variables, in the E-step the *posterior* over the candidates are evaluated:

$$w_{\mathbf{y}_i} = p(\mathbf{y}_i|l_i = 1, \mathbf{x}_i, \mathcal{I}) = \frac{\pi_{\mathbf{y}_i}\mathcal{N}(\mathbf{x}_i; \mathbf{y}_i, \rho\mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} \pi_{\mathbf{z}_i}\mathcal{N}(\mathbf{x}_i; \mathbf{z}_i, \rho\mathbf{I})}. \quad (34)$$

Then, the M-step involves minimizing:

$$Q_{\mathrm{KDE}}(\mathbf{p}) = E_{q(\mathbf{y})}\left[-\ln\left\{p(\mathbf{q})\prod_{i=1}^{n} p(l_i = 1, \mathbf{y}_i|\mathbf{x}_i, \mathcal{I})\right\}\right]$$

$$\propto \|\mathbf{q}\|^2_{\mathbf{\Lambda}^{-1}} + \sum_{i=1}^{n}\sum_{\mathbf{y}_i \in \Psi_i} \frac{w_{\mathbf{y}_i}}{\rho}\|\mathbf{x}_i - \mathbf{y}_i\|^2, \quad (35)$$

where $q(\mathbf{y}) = \prod_{i=1}^n p(\mathbf{y}_i|l_i = 1, \mathbf{x}_i, \mathcal{I})$. Using the relationship: $\sum_{\mathbf{y}_i \in \Psi_i} w_{\mathbf{y}_i} = 1$, the solution for the linearized shape model can be written:

$$\Delta\mathbf{p} = -(\rho\tilde{\mathbf{\Lambda}}^{-1} + \mathbf{J}^T\mathbf{J})^{-1}(\rho\tilde{\mathbf{\Lambda}}^{-1}\mathbf{p} - \mathbf{J}^T\mathbf{v}), \quad (36)$$

where $\mathbf{v} = [\mathbf{v}_1; \ldots; \mathbf{v}_n]$ is the concatenation of the mean shift vectors from each landmark:

$$\mathbf{v}_i = \left(\sum_{\mathbf{y}_i \in \Psi_i} \frac{\pi_{\mathbf{y}_i}\mathcal{N}(\mathbf{x}_i^c; \mathbf{y}_i, \rho\mathbf{I})}{\sum_{\mathbf{z}_i \in \Psi_i} \pi_{\mathbf{z}_i}\mathcal{N}(\mathbf{x}_i^c; \mathbf{z}_i, \rho\mathbf{I})}\mathbf{y}_i\right) - \mathbf{x}_i^c. \quad (37)$$

Notice that in the case that a ML solution is desired[4] (i.e. $p(\mathbf{p})$ is non-informative), the solution for the parameter update $\Delta\mathbf{p}$ is simply the non-orthogonal projection of the mean shift vectors onto the subspace spanned by the PDM's Jacobian. In any case, the solution in (36) suggests a simple and efficient implementation, consisting of an alternation between computing the mean shift vectors and their regularization by the shape model. The complete fitting procedure, which we will refer to as regularized landmark mean-shift (RLMS), is outlined in Algorithm 1.

Fashing and Tomasi (2005) previously showed that mean-shift is a bound optimization. This was later extended in Carreira-Perpinan (2007) by showing that for Gaussian kernels, mean-shift is equivalent to employing the EM algorithm as an optimization strategy. With the derivation of RLMS presented here, we show that such an interpretation can be generalized further to problems with conditionally independent likelihoods. As such, the desirable properties of the EM algorithm, namely provably convergent and improving, are adopted by the RLMS optimization strategy.

[4]The derivation in this section is an extension of the ML formulation we originally proposed in Saragih et al. (2009).

Algorithm 1 Regularized landmark mean-shift

Require: $\mathcal{I}$ and $\mathbf{p}$
1: Compute responses {(6)}
2: **while** not_converged($\mathbf{p}$) **do**
3: Linearize shape model {(12)}
4: Compute mean-shift vectors {(37)}
5: Compute PDM parameter update {(36)}
6: Update parameters: $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$
7: **end while**
8: **return** $\mathbf{p}$

3.1 Avoiding Local Minima

The response map approximations discussed in Sect. 2.2 can be though of as a form of smoothing. This explains the relative performance of the various methods. Gaussian approximations smooth the most but approximate the true response map the poorest, whereas smoothing effected by GMMs are not as aggressive but exhibits a degree of sensitivity towards local optima. One might consider using the Gaussian and GMM approximations in tandem, where a Gaussian approximation is used to get within the convergence basin of a GMM approximation. However, such an approach is inelegant and affords no guarantee that the mode of the Gaussian approximation lies within the convergence basin of the GMM's.

With the KDE approximation in RLMS a more elegant approach can be devised, whereby the complexity of the response map estimate is directly controlled by the variance of the Gaussian kernel (see Fig. 3). The guiding principle here is similar to that of optimizing on a Gaussian pyramid. It can be shown that when using Gaussian kernels, there exists a $\rho < \infty$ such that the KDE is unimodal, regardless of the distribution of samples (Carreira-Perpinan and Williams 2003). As ρ is reduced, modes divide and smoothness of the objective's terrain decreases. However, it is likely that the optimum of the objective at a larger ρ is closest to the desired mode of the objective with a smaller ρ, promoting its convergence to the correct mode. As such, the policy under which ρ is reduced acts to guide optimization towards the global optimum of the true objective.

It should be noted that in formulating RLMS, the Gaussian kernel is in fact a particular incarnation of the likelihood of selecting a landmark candidate given the PDM's landmark estimate: $p(\mathbf{y}_i|\mathbf{x}_i)$. As such, given sufficient granularity in the local search (i.e. the choice of $\{\mathbf{\Psi}_i\}_{i=1}^n$), then the variance of the kernel, ρ, that best represents the like-

lihood is given by (29). However, since PCA retains the majority of total variance of the shape, typically in the order of 95–98%, ρ will generally be quite small, which results in a highly multimodal objective terrain. When initialization is far from the global maximum, optimization over this terrain is susceptible to terminating in local maxima. The variance tightening policy described above essentially replaces the optimal objective terrain with a smoothed estimate of itself, for which local maxima are reduced, but the global optimum is perturbed, the magnitude of which is directly related to the choice of ρ.

Drawing parallels with existing methods, as $\rho \to \infty$ the RLMS update approaches the solution of a homoscedastic Gaussian approximated objective function. As ρ is reduced, the KDE approximation resembles a GMM approximation, where the approximation for smaller ρ settings is similar to a GMM approximation with more modes.

3.2 Handling Partial Occlusions

One of the main limitations of RLMS as well as the other existing strategies described in Sect. 2.2 is that the likelihood that a landmark is aligned, $p(l_i = 1|\mathbf{x}_i, \mathcal{I})$, encodes no information regarding the effects of occlusion. Typically, it is trained on a set of image patches cropped from aligned (and misaligned in the case of a discriminative classifier) locations in occlusion free images. Modeling the alignment likelihood for occluded landmarks is not tractable in most realistic applications since one can not adequately cover the space of occluded appearance (i.e. the occluding object could be anything). However, ignoring such cases during training can have detrimental effects when the object of interest in the image is partially occluded.

Observations that do not adhere to the assumed model are generally referred to as outliers. In CLM fitting, outliers stem from non-Gaussian image noise, unseen appearance and occlusions. An example of non-Gaussian image noise often observed in medical image analysis is the signal drop-out effect (Zhou et al. 2005). Although non-Gaussian over the whole image, these outliers can be handled quite robustly by existing fitting strategies since their main effect is in increasing the spatial spread of the likelihood, which is equivalent to increasing the uncertainty in its estimation. This in turn results in a smaller contribution to the global objective, limiting deterioration effected by such noise. Outliers stemming from unseen appearance is not so problematic in a CLM fitting framework since only the appearance of a patch is considered when learning the alignment likelihood. These patches are typically quite small (i.e. in the order of (11×11)-pixels). As such, for many problems there exists an adequate amount of data to train a model that generalizes well over most instances of the object. Variations due to lighting changes can be partially accounted for by a power normalization (Wang et al. 2008a). This is in stark contrast to holistic based approaches (see Liu 2007; Matthews and Baker 2004; Nguyen and De la Torre Frade 2008; Zhou and Comaniciu 2007, for example), where correlations between all pixels within the object are considered.

Although existing approaches are somewhat robust towards the two aforementioned outlier types, the same can not be said for occlusions. The reason for this is that the patch appearance of an occluding object may be similar to that of the object of interest, which is a direct result of the conditional independence assumed for the aligned landmark likelihood. For example, the patch appearance of a landmark on the periphery of the face simply looks like an edge, which is easily confused with any occluding object with a strong edge.

The parameter update in the RLMS optimization strategy in (36) is essentially a regularized projection of the mean-shift vector for each landmark onto the space of plausible shape variations. As discussed previously, due to the misleading landmark likelihood in the presence of occlusion, the mean-shift vectors for occluded landmarks may be erroneous. As such, a least squares projection to regularize the solution is no longer suitable. A simple approach to handle such cases, therefore, is to use an M-estimator for this projection. Formally, this entails substituting the Q-function in (35) with:

$$Q_{\mathrm{KDE}}(\mathbf{p}) \propto \|\mathbf{q}\|^2_{\mathbf{\Lambda}^{-1}} + \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \omega_{\mathbf{y}_i} \varrho(\|\mathbf{x}_i - \mathbf{y}_i\|^2; \boldsymbol{\theta}), \quad (38)$$

where ϱ is an M-estimator, for example the Geman-McClure function, which has been used extensively in optical flow estimation (see Black and Anandan 1993; Blake et al. 1994, for example):

$$\varrho(r^2; \boldsymbol{\theta}) = \frac{r^2}{r^2 + \alpha^2}; \quad \boldsymbol{\theta} = \{\alpha\}. \quad (39)$$

Equation (38) can then be solved for using iteratively re-weighted least squares as follows:

$$\Delta\mathbf{p} = -\mathbf{H}^{-1}_{\mathrm{KDE}} \left[\tilde{\mathbf{\Lambda}}^{-1}\mathbf{p} + \sum_{i=1}^{n} \mathbf{J}_i^T \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \omega_{\mathbf{y}_i} \varrho'(\mathbf{x}_i^c - \mathbf{y}_i) \right], \quad (40)$$

where ϱ' denotes the derivative of the M-estimator evaluated at $\|\mathbf{x}_i^c - \mathbf{y}_i\|^2$, and the Hessian takes the form:

$$\mathbf{H}_{\mathrm{KDE}} = \tilde{\mathbf{\Lambda}}^{-1} + \sum_{i=1}^{n} \left(\sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \omega_{\mathbf{y}_i} \varrho' \right) \mathbf{J}_i^T \mathbf{J}_i. \quad (41)$$

Since the optimization strategy for RLMS is essentially the generalized EM algorithm, in order to preserve the prop-

erties of EM optimization, namely convergent and provably improving, the *posterior* weights $\omega_{\mathbf{y}_i}$ must be adjusted to reflect the new parameterization of $p(\mathbf{y}_i|\mathbf{x}_i)$ in accordance with the choice of ϱ (i.e. $\omega_{\mathbf{y}_i} \neq w_{\mathbf{y}_i}$ from (34)). For a particular ϱ, the landmark candidate likelihood takes the form:

$$p(\mathbf{y}_i|\mathbf{x}_i) = \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \exp\{-\varrho(\|\mathbf{x}_i - \mathbf{y}_i\|^2; \boldsymbol{\theta})\}, \tag{42}$$

where $\mathcal{Z}$ is a partition (normalizing) function that enforces (42) to be a PDF:

$$\mathcal{Z}(\boldsymbol{\theta}) = \int_{\mathbf{r}_i} p(\mathbf{r}_i|\mathbf{x}_i)\, d\mathbf{r}_i < \infty; \quad \mathbf{r}_i \in \boldsymbol{\Omega}, \tag{43}$$

with $\boldsymbol{\Omega}$ denoting the bounded spatial domain of the image. When the same hyperparameters $\boldsymbol{\theta}$ are chosen for the M-estimator of all candidates of all landmarks (i.e. a homoscedastic KDE), then the partition function need not be evaluated explicitly as it factors out of the equations as a scaling constant. The *posterior* over the candidates can now be written:

$$\omega_{\mathbf{y}_i} = \frac{\pi_{\mathbf{y}_i} \exp\{-\varrho(\|\mathbf{x}_i - \mathbf{y}_i\|^2; \boldsymbol{\theta})\}}{\sum_{\mathbf{z}_i \in \boldsymbol{\Psi}_i} \pi_{\mathbf{z}_i} \exp\{-\varrho(\|\mathbf{x}_i - \mathbf{z}_i\|^2; \boldsymbol{\theta})\}}. \tag{44}$$

With the use of a robust error function to reduce the effects of outlying landmark candidates, the resulting algorithm is equivalent to RLMS with a non-Gaussian kernel, where the type of kernel depends on the choice of ϱ. One of the complications introduced by such a choice, however, is the selection of the hyperparameters $\boldsymbol{\theta}$. For the case of a Gaussian kernel: $\boldsymbol{\theta} = \{\rho\}$, the optimal setting of which is given in (29). For more general ϱ, a closed form solution for the optimal $\boldsymbol{\theta}$ does not exist. Nonetheless, for certain ϱ reasonable estimates can be made without resorting to a Monte-Carlo strategy. For example, for the Geman-McClure robust function in (39) the inlier region is given by: $|r| \leq \frac{\alpha}{\sqrt{3}}$. Due to the assumed parametric shape model, the inlier standard deviation is given by $\sqrt{\rho}$. As such, following (Roberts et al. 2007) we can set: $\alpha = \gamma\sqrt{3\rho}$, where γ is the multiple of the inlier standard deviation, which is typically chosen as $\gamma \in [1, 3]$.

Finally, it should be noted that the robust formulation of RLMS described here defines outlying landmark candidates as those that are inconsistent with the shape model. Although such an assumption is reasonable for many cases, it is certainly possible that an outlying candidate is consistent with the shape model but does not represent the true landmark location. This occurrence becomes increasingly likely as the number of directions of variability of the shape model increases. Although the regularization induced by the prior, $p(\mathbf{p})$, partially addresses this problem, it is not a complete solution and how best to handle such cases remains an open question.

3.3 Practical Considerations

3.3.1 Similarity Normalized Search

Since the exhaustive local search used in CLM fitting is performed over the spatial dimensions only, such an approach poorly accounts for significant variations in scale and in-plane rotation. However, landmark detectors that are trained on images exhibiting such variations will lack in specificity, limiting the fidelity of the fitting procedure.

Generative holistic deformable model fitting algorithms, for example (Edwards et al. 1998; Matthews and Baker 2004), typically measure model fit in a predefined reference shape, often chosen according to the mean shape over the training data. We use the same principle here, whereby the image is transformed to the reference frame using the PDM's current estimate of scale and rotation. The fitting procedure outlined in Algorithm 1, is then performed on this *similarity normalized image*. The PDM parameters describing the shape in the image frame can then be found by composing the converged shape with the inverse of the similarity transform used to normalize the image. As the fitting procedure converges, the estimate of the similarity transform approaches that of the true pose of the object, improving the reliability of the landmark detections.

3.3.2 1D vs. 2D Search Regions

One of the remaining open questions with the CLM formulation is the selection of the search region for each landmark $\{\boldsymbol{\Psi}_i\}_{i=1}^n$. Some CLM variants, for example (Cootes and Taylor 1992; Gu and Kanade 2008), perform the exhaustive local search along a profile that is typically (manually) chosen to be perpendicular to the direction of largest edgedness of landmark appearance. Others search within a rectangular bounding box around the current landmark estimate (Cristinacce and Cootes 2007; Wang et al. 2008a).

The motivation for a profile search is two-fold. First, it is much cheaper to compute, with only d-detector evaluations required, as opposed to d^2 for square search regions. Secondly, it leverages on the limited structure of patches used for detection. As discussed in Sect. 2.2, landmarks located on edges tend to have poor discriminative capacity along the edge. As such, little motion information is gained by evaluating the detector at those locations.

Although common choices for landmarks of many deformable objects are indeed placed on edges, this is not strictly required. For example, landmarks on the human face almost always include eye and lip corners as they facilitate better consistency in manual annotations. For these landmarks, the local structure is sufficient to distinguish it from

image patches at neighboring locations. Therefore, restricting the search to a profile may bias the estimated objective function. This is made worse by the heuristic choice often made for the uncertainty in the direction perpendicular to the search profile (i.e. typically set to be equivalent to the uncertainty along the profile, Cootes and Taylor 1992; Gu and Kanade 2008). There is also some difficulty in determining the best profile direction for these landmarks. Although the optimal profile direction can be learned from the data through a cross-validation strategy, most methods simply define the direction as some function of the respective locations of the neighboring landmarks. Finally, in order to ensure robust fitting, the profile directions should be chosen such that they adequately cover the space of directions (i.e. a model with horizontal search profiles only can not move vertically).

Rectangular search regions do not suffer the aforementioned limitations of profile searches. Their only drawback is that they involve more detector evaluations, which may lead to inefficient fitting when complex detectors are utilized. In this work we use rectangular search regions since the linear classifier used for detection can be evaluated using efficient convolution operations. It is certainly possible to combine profile and rectangular search regions, where the choice of search region is specialized to the local structure exhibited by each landmark, however we do not pursue such an approach in this work.

3.3.3 Precomputed Grid for Efficiency

In the KDE representation of the response maps, the kernel centers are placed at the grid nodes defined by the search window. From the perspective of GMM fitting, these kernels represent candidates for the true landmark locations. Although no optimization is required for determining the number of modes, their centers and mixing coefficients, the number of candidates used here is much larger than what would typically be used in a general GMM estimate (i.e. GMM based representations typically use $K_i < 10$, whereas the search window size typically has > 100 nodes). As such, the computation of the *posterior* in (34) will be more costly. However, if the variance ρ is known *a-priori* (see Sect. 3.1), then some approximations can be made to significantly reduce computational complexity.

The main overhead when computing the mean-shift update is in evaluating the kernel between the current landmark estimate and every grid node in the response map. Since the grid locations are fixed and ρ is assumed to be known, one might choose to precompute the kernel for various settings of $\mathbf{x}_i$. In particular, a simple choice would be to precompute these values along a grid sampled at or above the resolution of the response map grid $\mathbf{\Psi}_i$. During fitting one simply finds the location in this grid closest to the current estimate of a PDM landmark and estimate the kernel evaluations by assuming the landmark is actually placed at that node (see Fig. 4). This only involves a table lookup and can be performed efficiently. The higher the granularity of the grid the better the approximation will be, at the cost of greater storage requirements but without a significant increase in computational complexity. An interpolation process over the lookup table might further improve estimation accuracy here, however we found that given sufficient granularity of the grid, such an addition offers little overall benefit. Finally, although this approximation ruins the strictly improving properties of EM, we empirically show in Sect. 4 that accurate fitting can still be achieved with this approximation. In our implementation, we found that such an approximation reduced the average fitting time by one half.

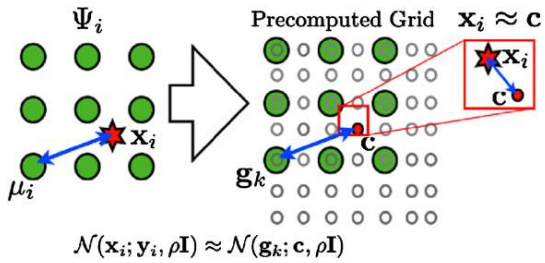

Fig. 4 Illustration of the use of a precomputed grid for efficient mean-shift. Kernel evaluations are precomputed between $\mathbf{c}$ and all other nodes in the grid. To approximate the true kernel evaluation, $\mathbf{x}_i$ is assumed to coincide with $\mathbf{c}$ and the likelihood of any response map grid location can be attained by a table lookup

4 Experiments

4.1 Empirical Comparison of Fitting Approaches

4.1.1 Still Images

The various CLM optimizations strategies discussed above were compared on the problem of generic frontal face fitting on two publicly available databases: (i) the CMU Pose, Illumination and Expression Database (MultiPie) (Gross et al. 2008), and (ii) the XM2VTS database (Messer et al. 1999). MultiPie is annotated with a 68-point markup used as ground truth landmarks. We used 762 frontal face images of 339 subjects. XM2VTS consists of 2360 frontal face images of 295 subjects for which ground truth annotations are publicly available but different from the 68-point markup we have for MultiPie. XM2VTS contains neutral expression only whereas MultiPie contains significant expression variations. The average interocular distance for the MultiPIE and XM2VTS databases are 80 and 100 pixels respectively. Four-fold cross validation experiments were performed on both MultiPie and XM2VTS, separately, where the images

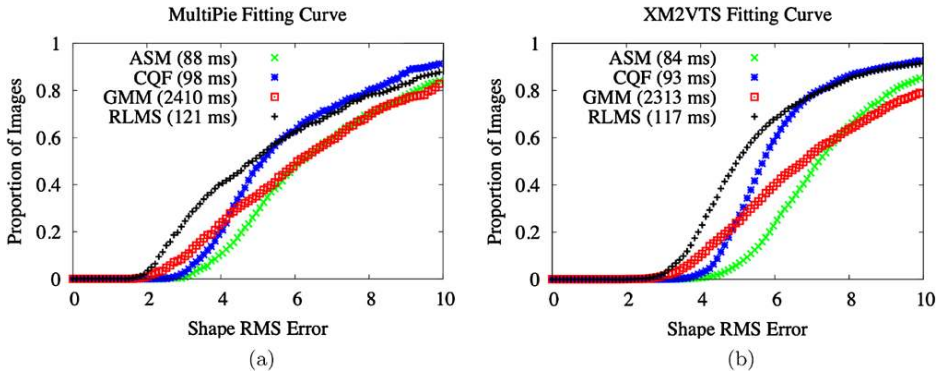

Fig. 5 Fitting curves for the ASM, CQF, GMM and RLMS optimization strategies on the MultiPie and XM2VTS databases

were partitioned into four sets of non-overlapping subject identities. In each trial, three partitions were used for training and the remainder for testing.

On these databases we compared four types of optimization strategies: (i) ASM (Cootes and Taylor 1992) described in Sect. 2.2.1, (ii) convex quadratic fitting (CQF) (Wang et al. 2008a) described in Sect. 2.2.2, (iii) GMM (Gu and Kanade 2008) described in Sect. 2.2.3, and (iv) RLMS derived in Sect. 3. For GMM, we empirically set $K_i = 5$ and used the EM algorithm to estimate the parameters of the mixture models describing the response maps. For RLMS, we used the efficient approximation described in Sect. 3.3.3 with a grid spacing of 0.1-pixels and a variance tightening policy of $\rho = \{20, 10, 5, 1\}$. In all cases the linear logistic regressor defined in (8) was used as landmark detectors. The local experts were (11×11)-pixels in size and the exhaustive local search was performed over a (15×15)-pixel window. As such, the only difference between the various methods compared here is the way in which the response maps are approximated along with their specialized optimization strategies. To better illustrate the performance difference between these methods, a ML formulation was used, where a non-informative prior is assumed over the PDM parameters. In all cases, the scale and location of the model was initialized by an off-the-shelf face detector, the rotation and non-rigid parameters in (1) set to zero (i.e. the mean shape), and the model fit until the optimization converged as measured by: $\sum_{i=1}^{n} \|\Delta \mathbf{x}_i\|^2 \leq 0.01$, where $\Delta \mathbf{x}_i$ denotes the change in the ith landmarks position between iterations.

Results of these experiments can be found in Fig. 5. The graphs (fitting curves) show the proportion of images at which various levels of maximum perturbation was exhibited, measured as the root-mean-squared (RMS) error between the ground truth landmarks and the resulting fit in the image frame. The average fitting times for the various methods on a 2.5 GHz Intel Core 2 Duo processor are shown in the legend.

The results show a consistent trend in the relative performance of the four methods. Firstly, CQF has the capacity to significantly outperform ASM. As discussed in Sect. 2.2.2 this is due to CQF's ability to account for directional uncertainty in the response maps as well as being more robust towards outlying responses. However, CQF has a tendency to over-smooth the response maps, leading to limited convergence accuracy. GMM shows an improvement in accuracy over CQF as shown by the larger number of samples that converged to smaller shape RMS errors. However, it exhibits sensitivity towards local optima due to its multimodal objective. This can be seen by its poorer performance than CQF for reconstructions errors above 4.2-pixels RMS in MultiPie and 5-pixels RMS in XM2VTS. In contrast, RLMS attains even better accuracies than GMM but still retains a degree of robustness against local optima, where its performance over grossly misplaced initializations is comparable to CQF. Finally, despite the significant improvement in performance, RLMS exhibits only a modest increase in computational complexity compared to ASM and CQF. This is in contrast to GMM that requires much longer fitting times, mainly due to the complexity of fitting a mixture model to the response maps.

4.1.2 Tracking in a Sequence

Evaluating the performance of fitting algorithms on images outside of a particular database is more meaningful as it gives a better indication on how well a method generalizes. However, this is rarely conducted as it requires the tedious process of annotating new images with the PDM configuration of the training set. Here, we utilize the freely available FGNet talking face sequence.[5] Quantitative analysis on this sequence is possible since ground truth annotations are available in the same format as that in XM2VTS. The same model used in the still image experiments was used here, except that it was trained on all images in XM2VTS. All four fitting methods were evaluated on this sequence using a ML formulation with the same criterion for convergence as described previously. We initialize the model using a face

[5] http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html.

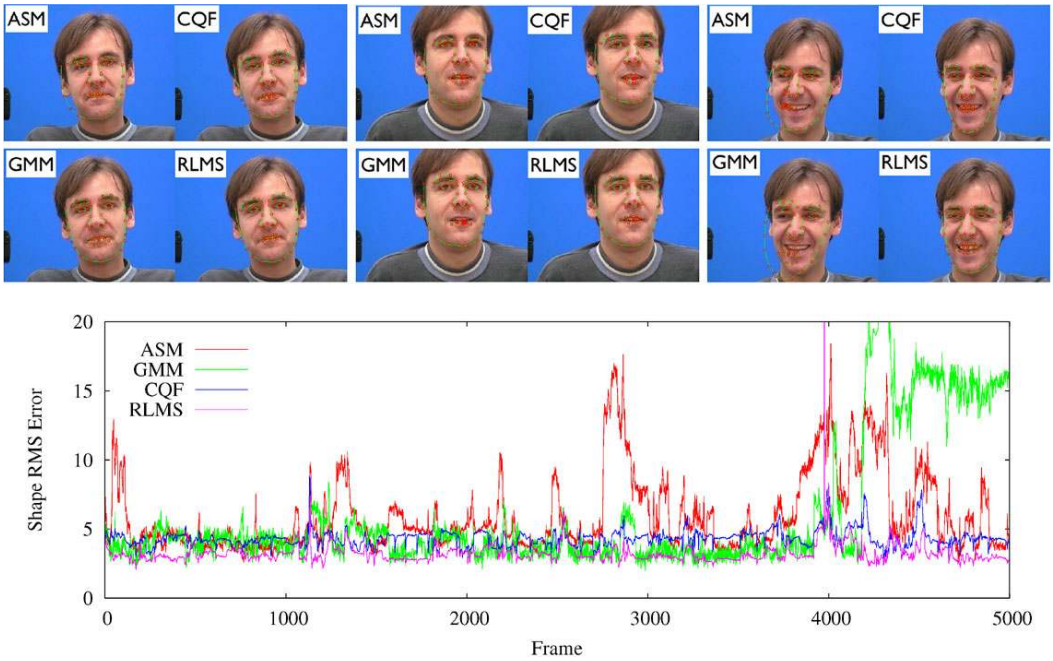

Fig. 6 *Top row:* Tracking results on the FGNet Talking Face database for frames {0, 1230, 4200}. *Clockwise from top left* are fitting results for ASM, CQF, RLMS and GMM. *Bottom:* Plot of shape RMS error from ground truth annotations throughout the sequence

detector in the first frame and fit consecutive frames using the PDM's configuration in the previous frame as an initial estimate. Although an improvement in performance may be attained by performing some kind of temporal smoothing, using a Kalman filter for example, this was not performed here as we are investigating the effects of the response map estimates and their subsequent optimization strategy on fitting performance.

In Fig. 6, the shape RMS error for each frame is plotted for the four optimization strategies being compared. The relative performance of the various strategies is similar to that observed in the still image experiments. CQF exhibits excellent stability throughout the sequence, but is limited in its fidelity. GMM outperforms CQF in many frames, but is not consistent. Furthermore, as with ASM, it appears particularly unstable on this sequence, loosing track at around frame 4200, and fails to recover until the end of the sequence. The unstable nature of GMM fitting here is due to its estimation of the mixture model describing the response maps, which is a nonlinear optimization procedure that is initialization dependent. As we initialize its K_i-centers based on the K_i-best responses, which can be noisy, very different mixture model estimates can be made for similar response maps (i.e. from neighboring frames). A similar observation can be made for ASM, explaining its instability. In contrast, RLMS exhibits the stability of CQF but with consistently better accuracy.

4.2 Qualitative Analysis and Occlusion Handling

A quantitative analysis of the performance of fitting algorithms on real occluded images is difficult as it requires the collection and annotation of occluded images. Typically evaluation is performed by synthesizing occlusions (Gross et al. 2004; Saragih 2008). However, the performance of a fitting method is often dependent on the choice made for the synthesized occlusions (Saragih 2008). Furthermore, it is difficult to model the effects of an occluding object on non-occluded parts of the image, such as that of shadowing. As such, in this work we analyze the effects of occlusion on RLMS qualitatively, by observing its behavior in a tracking sequences with real occlusions.

In Figs. 7 and 8 frames from two sequences with gross occlusions are shown, where tracking was performed using four variants of RLMS, namely the ML and MAP variants with Gaussian and Geman-McClure kernels. The model was trained on the entire MultiPie database with a 3D shape model learned by applying nonrigid structure from motion on the available annotations (Torresani et al. 2008). The types of occlusions in these sequences are quite challenging as they occlude the eye and mouth regions in Figs. 7 and 8, respectively, which generally give strong responses compared to other landmarks. Notice also the change in lighting over non-occluded landmarks effected by the occluding object in Fig. 7.

On these sequences, both ML and MAP variants of CLM fitting with a Gaussian kernel fail to track the subject, where

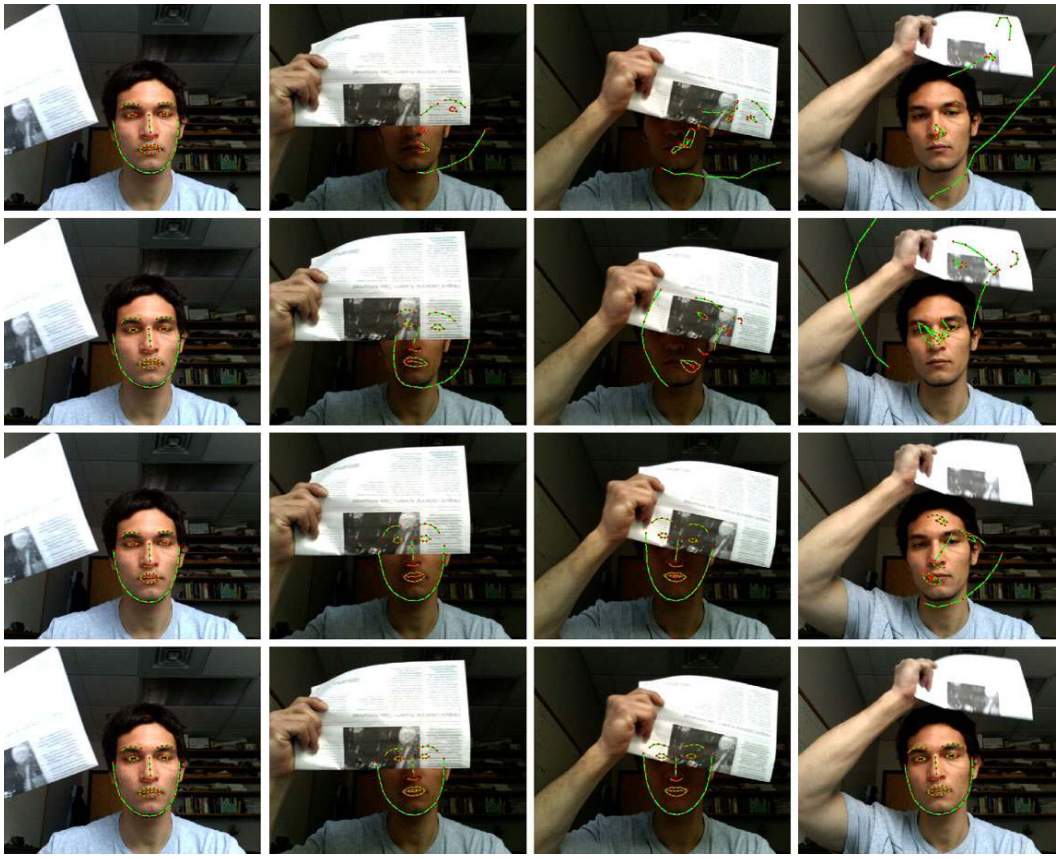

Fig. 7 Example of tracking with gross occlusion. *Top to bottom rows:* ML fitting with a Gaussian kernel, MAP fitting with a Gaussian kernel, ML fitting with a Geman-McClure kernel, and MAP fitting with a Geman-McClure kernel. *Left to right columns:* Frames 0, 30, 40 and 50

significant perturbations are immediately observed as the object starts occluding some landmarks. ML fitting with the Geman-McClure kernel exhibits better robustness, where the effects of occlusion are not as pronounced. However, the accumulation of errors lead to a loss of tracking towards the end of the sequence. As discussed in Sect. 3.2, the use of a robust kernel enforces occlusion invariance only through the consistency of the landmarks' joint motion with the shape model. As such, erroneous landmark candidates that are consistent with the shape model are not accounted for. Although the MAP formulation utilizes the same assumption, the use of a prior restricts the variability of the shape model, which reduces sensitivity of the fitting procedure towards such candidates. The tracking results for MAP fitting with the Geman-McClure kernel supports this analysis, where it successfully tracks until the end of the sequences.

Finally, a qualitative analysis of RLMS was performed on the Faces in the Wild database (Huang et al. 2007). It contains images taken under varying lighting, resolution, image noise and partial occlusion. As before, the model was initialized using a face detector and fit until convergence was attained. Some fitting results are shown in Fig. 9. Results suggest that RLMS exhibits a degree of robustness towards variations typically encountered in real images.

5 Conclusion

The optimization strategy for local experts-based deformable model fitting was investigated in this work. Various existing methods were posed within a consistent probabilistic framework where they were shown to make different parametric approximations to the true likelihood maps of landmark locations. From this perspective, the fitting behavior of a number of popular approaches were explained and validated through numerical experiments. To address the difficulties inherent in these approaches, and leveraging on insights gained from the aforementioned formulation, a new approximation of the likelihood maps was proposed that uses a nonparametric representation. Further innovations that reduce online computational complexity, avoid local optima and encourage robustness against partial occlusions were also proposed. The resulting fitting algorithm for this formulation was shown to be simple and efficient

Fig. 8 Example of tracking with gross occlusion. *Top to bottom rows:* ML fitting with a Gaussian kernel, MAP fitting with a Gaussian kernel, ML fitting with a Geman-McClure kernel, and MAP fitting with a Geman-McClure kernel. *Left to right columns:* Frames 0, 30, 40 and 50

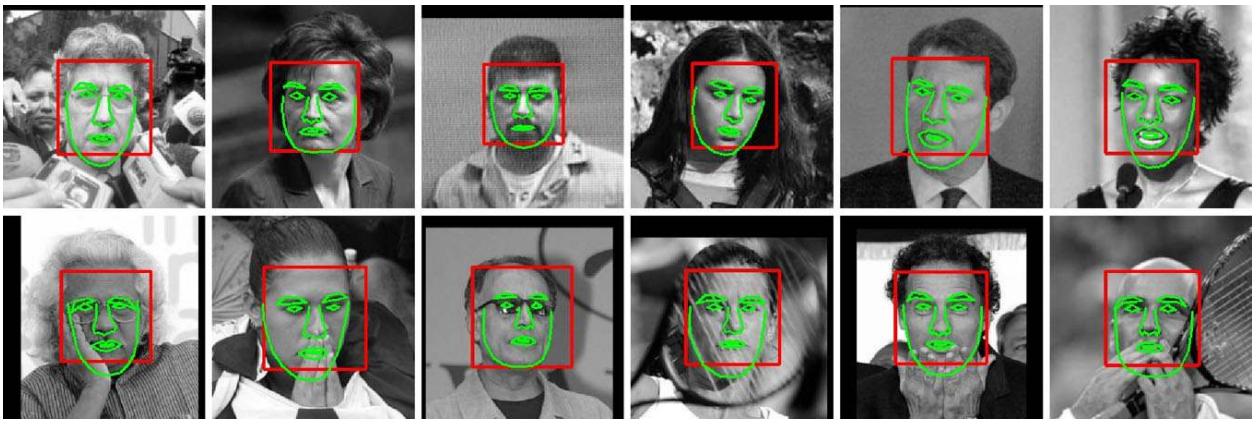

Fig. 9 Example fitting results on the Faces in the Wild database (*red rectangles* denote detected face regions used for initialization). *Top*: Un-occluded images. *Bottom*: Partially occluded

as well as exhibit superior performance over some existing approaches on the task of generic face fitting.

The approach proposed in this work is a framework for deformable model fitting rather than a complete system in itself. It is a generic optimization strategy in the sense that a number of its components can be specialized to a particular application. As such, future work will involve investigations into the use of different feature detectors (see Avidan 2004; Cootes and Taylor 1992; Cristinacce and Cootes 2006, for example), more sophisticated shape models (see Gu and

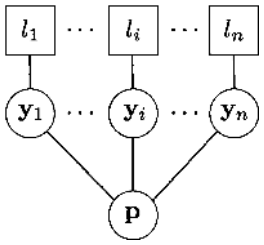

Fig. 10 Undirected graphical model of CLM fitting. *Squares* denote the observations and *circles* denote hidden variables

Kanade 2008; Romdhani et al. 1999, for example), the application of temporal smoothness constraints (see Wang et al. 2008b; Zhou et al. 2005, for example), and different kernel types, all of which can be integrated seamlessly into the proposed framework.

Appendix

Many problems in computer vision lend themselves naturally to representations using graphical models. Examples of this include stereo matching (Sun et al. 2003) and optical flow estimation (Felzenszwalb and Huttenlocher 2004) to name a few. In this section we motivate the formulation used in RLMS described in Sect. 3 from a graphical model perspective.

An undirected graph $\mathcal{G}$ is defined by a set of nodes $\mathcal{V}$ and a corresponding set of edges $\mathcal{E}$. For the problem of CLM fitting, let us define: $\mathcal{V} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{p}\}$, where, as in Sect. 3, $\mathbf{p}$ denotes the PDM parameters and $\mathbf{y}_i$ denotes a random variable describing the true location of the i^{th} landmark location in the image. The neighborhood of a node $a \in \mathcal{V}$ is defined as: $\mathbf{\Gamma}(a) = \{b | (a, b) \in \mathcal{E}\}$, where, for the problem of CLM fitting we have: $\mathbf{\Gamma}(\mathbf{y}_i) = \{\mathbf{p}\}$ and $\mathbf{\Gamma}(\mathbf{p}) = \{\mathbf{y}_i\}_{i=1}^n$. Letting the alignment labels $\{l_i\}_{i=1}^n$ denote the observed variables,[6] the graph describing this system takes the particularly simple acyclic form illustrated in Fig. 10.

Since the CLM graph is acyclic, the conditional distribution of all random variables can be directly calculated by a local message-passing algorithm known as belief propagation (BP) (Yedidia et al. 2002). At iteration t of the BP algorithm, each node a calculates a message $m_{a\to b}^t$ to be sent to each of its neighbors $b \in \mathbf{\Gamma}(a)$. For the CLM graph, only two types of messages need to be sent:

$$m_{\mathbf{y}_i\to\mathbf{p}}^t \propto \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} \phi(\mathbf{y}_i, \mathbf{p}) \varphi(\mathbf{y}_i, l_i) \tag{45}$$

$$m_{\mathbf{p}\to\mathbf{y}_i}^t \propto \int \phi(\mathbf{y}_i, \mathbf{p}) p(\mathbf{p}) \prod_{j=1, j\neq i}^{n} m_{\mathbf{y}_i\to\mathbf{p}}^{t-1} d\mathbf{p}, \tag{46}$$

[6]Typically the observed variables in graphical models of computer vision problems relate to image pixels. We depart from this convention here, in order to remain consistent with the discriminative interpretation of landmark alignment in Sect. 2.1, where the observation of the image is implicit in its formulation.

where $\phi(\mathbf{y}_i, \mathbf{p})$, $\varphi(\mathbf{y}_i, l_i)$ and $p(\mathbf{p})$ are the various *potentials* given by (31), (6) and (10), respectively.

Although the BP algorithm over the CLM graph converges within two iterations, evaluations of $m_{\mathbf{p}\to\mathbf{y}_i}^t$ are not analytically tractable (i.e. integration over a GMM). However, since we are interested only in the conditional marginal distribution of the PDM parameters:

$$p(\mathbf{p} | \{l_i = 1\}_{i=1}^n, \mathcal{I}) \propto p(\mathbf{p}) \prod_{i=1}^{n} m_{\mathbf{y}_i\to\mathbf{p}}^t, \tag{47}$$

the messages $\{m_{\mathbf{p}\to\mathbf{y}_i}^t\}_{i=1}^n$ need not be computed since they do not contribute to the messages $\{m_{\mathbf{y}_i\to\mathbf{p}}^t\}_{i=1}^n$ in (45). This is a direct result of assumed graph structure, where landmark detections are conditionally independent. Therefore, we arrive at the desired conditional marginal distribution by computing only the messages from the landmark candidate nodes to the PDM parameter node, yielding the RLMS objective in (33).

References

Avidan, S. (2004). Support vector tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *26*, 1064–1072.

Basso, C., Vetter, T., & Blanz, V. (2003). Regularized 3D morphable models. In *IEEE international workshop on higher-level knowledge in 3D modeling and motion analysis (HLK'03)* (p. 3).

Black, M., & Anandan, P. (1993). The robust estimation of multiple motions: affine and piecewise-smooth flow fields. Tech. rep., Xerox PARC.

Blake, A., Isard, M., & Reynard, D. (1994). Learning to track curves in motion. In *IEEE conference on decision theory and control* (pp. 3788–3793).

Bruhn, A., Weickert, J., & Schnörr, C. (2005). Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. *International Journal of Computer Vision*, *61*(3), 211–231.

Carreira-Perpinan, M. (2007). Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *29*(5), 767–776.

Carreira-Perpinan, M., & Williams, C. (2003). On the number of modes of a Gaussian mixture. *Lecture Notes in Computer Science*, *2695*, 625–640.

Cootes, T., & Taylor, C. (1992). Active shape models—'smart snakes'. In *British machine vision conference (BMVC'92)* (pp. 266–275).

Cristinacce, D., & Cootes, T. (2004). A comparison of shape constrained facial feature detectors. In *IEEE international conference on automatic face and gesture recognition (FG'04)* (pp. 375–380).

Cristinacce, D., & Cootes, T. (2006). Feature detection and tracking with constrained local models. In *British machine vision conference (BMVC'06)* (pp. 929–938).

Cristinacce, D., & Cootes, T. (2007). Boosted active shape models. In *British machine vision conference (BMVC'07)* (vol. 2, pp. 880–889).

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, *39*(1), 1–38.

Edwards, G., Taylor, C., & Cootes, T. (1998). Interpreting face images using active appearance models. In *IEEE international conference on automatic face and gesture recognition (FG'98)* (pp. 300–305).

Fashing, M., & Tomasi, C. (2005). Mean shift as a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *27*(3), 471–474.

Felzenszwalb, P., & Huttenlocher, D. (2004). Efficient belief propagation for early vision. In *IEEE conference on computer vision and pattern recognition (CVPR'04)* (vol. 1, pp. 261–268).

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*, 32–40.

Gelman, A., Carlin, J., Stern, H., & Rubinx, D. (1995). *Bayesian data analysis*. London/Boca Raton: Chapman & Hall/CRC Press.

Gross, R., Matthews, I., & Baker, S. (2004). Constructing and fitting active appearance models with occlusion. In *Proceedings of the IEEE workshop on face processing in video* (p. 72).

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2008). Multi-pie. In *IEEE international conference on automatic face and gesture recognition (FG'08)* (pp. 1–8).

Gu, L., & Kanade, T. (2008). A generative shape regularization model for robust face alignment. In *European conference on computer vision (ECCV'08)* (pp. 413–426).

Huang, G., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Tech. rep. 07-49, University of Massachusetts, Amherst.

Liu, X. (2007). Generic face alignment using boosted appearance model. In *IEEE conference on computer vision and pattern recognition (CVPR'07)* (pp. 1–8).

Matthews, I Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*, 135–164.

Messer, K., Matas, J., Kittler, J., Lüttin, J., & Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. In *International conference of audio- and video-based biometric person authentication (AVBPA'99)* (pp. 72–77).

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *19*(7), 696–710.

Nguyen, M., & De la Torre Frade, F. (2008). Local minima free parameterized appearance models. In *IEEE conference on computer vision and pattern recognition (CVPR'08)* (pp. 1–8).

Nickels, K., & Hutchinson, S. (2002). Estimating uncertainty in SSD-based feature tracking. *Image and Vision Computing*, *20*, 47–58.

Roberts, M., Cootes, T., & Adams, J. (2007). Robust active appearance models with iteratively rescaled kernels. In *British machine vision conference (BMVC'07)* (vol. 1, pp. 302–311).

Romdhani, S., Gong, S., & Psarrou, A. (1999). A multi-view nonlinear active shape model using kernel PCA. In *British machine vision conference (BMVC'99)* (pp. 438–492).

Saragih, J. (2008). The generative learning and discriminative fitting of linear deformable models. PhD thesis, The Australian National University, Australia.

Saragih, J., Lucey, S., & Cohn, J. (2009). Face alignment through subspace constrained mean-shifts. In *IEEE international conference on computer vision (ICCV'09)* (pp. 1034–1041).

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London/Boca Raton: Chapman & Hall/CRC Press.

Sun, J., Zheng, N., & Shum, H. (2003). Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *25*(7), 787–800.

Torresani, L., Hertzmann, A., & Bregler, C. (2008). Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *30*(5), 878–892.

Wang, Y., Lucey, S., & Cohn, J. (2008a). Enforcing convexity for improved alignment with constrained local models. In *IEEE conference on computer vision and pattern recognition (CVPR'08)* (pp. 1–8).

Wang, Y., Lucey, S., Cohn, J., & Saragih, J. (2008b). Non-rigid face tracking with local appearance consistency constraint. In *IEEE international conference on automatic face and gesture recognition (FG'08)*.

Yedidia, J., Freeman, W., & Weiss, Y. (2002). Constructing free energy approximations and generalized belief propagation algorithms. Tech. rep., Mitsubishi Electric Research Laboratories (MERL).

Zhou, S., & Comaniciu, D. (2007). Shape regression machine. In *Information processing in medical imaging (IPMI'07)* (pp. 13–25).

Zhou, X., Comaniciu, D., & Gupta, A. (2005). An information fusion framework for robust shape tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *27*(1), 115–129.