# Face Alignment through Subspace Constrained Mean-Shifts

Jason M. Saragih, Simon Lucey, Jeffrey F. Cohn
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jsaragih,slucey,jeffcohn}@cs.cmu.edu

## 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. This simplification substitutes the true objective with a smoothed version of itself, reducing sensitivity to local minima and outlying detections. In this work, a principled optimization strategy is proposed where a nonparametric representation of the landmark distributions is maximized within a hierarchy of smoothed estimates. The resulting update equations are reminiscent of mean-shift but with a subspace constraint placed on the shape's variability. This approach is shown to outperform other existing methods on the task of generic face fitting.*

## 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 with varying degrees of success. Of these, one of the most promising is one that uses a patch-based representation and assumes image observations made for each landmark are conditionally independent [2, 3, 4, 5, 16]. This leads to better generalization with limited data compared to holistic representations [10, 11, 14, 15], since it needs only account for local correlations between pixel values. However, it suffers from detection ambiguities as a direct result of its local representation. As such, care should be taken in combining detection results from the various local detectors in order to steer optimization towards the desired solution.

Our key contribution in this paper lies in the realization that a number of popular optimization strategies are all, in some way, simplifying the distribution of landmark locations obtained from each local detector using a parametric representation. The motivation of 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)*? We address this by using a nonparametric representation that leads to an optimization in the form of subspace constrained mean-shifts.

## 2. Background

### 2.1. Constrained Local Models

Most fitting methods employ a linear approximation to how the shape of a non-rigid object deforms, coined the point distribution model (PDM) [2]. 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 $i^{\text{th}}$ landmark and $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$ denotes the parameters of the PDM, which consist of a global scaling $s$, a rotation $\mathbf{R}$, a translation $\mathbf{t}$ and a set of non-rigid parameters $\mathbf{q}$.

In recent years, an approach to that utilizes an ensemble of local detectors (see [2, 3, 4, 5, 16]) has attracted some interest as it circumvents many of the drawbacks of holistic approaches, such as modeling complexity and sensitivity to lighting changes. In this work, we will refer to these methods collectively as constrained local models (CLM)[1].

[1]This term should not be confused with the work in [5] which is a particular instance of CLM in our nomenclature.

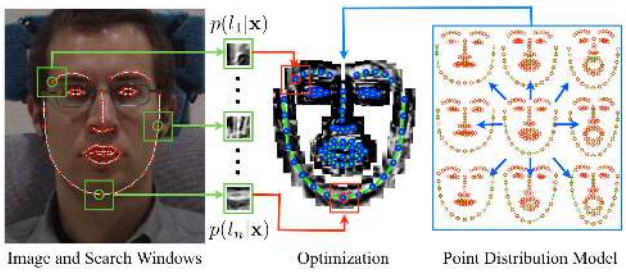



Figure 1. Illustration of CLM fitting and its two components: (i) an exhaustive local search for feature locations to get the response maps $\{p(l_i = \text{aligned}|I, \mathbf{x})\}_{i=1}^n$, and (ii) an optimization strategy to maximize the responses of the PDM constrained landmarks.

All instantiations of CLMs can be considered to be pursuing the same two goals: (i) perform an exhaustive local search for each PDM landmark around their current estimate using some kind of feature detector, and (ii) optimize the PDM parameters such that the detection responses over all of its landmarks are jointly maximized. Figure 1 illustrates the components of CLM fitting.

**Exhaustive Local Search:** In the first step of CLM fitting, a likelihood map is generated for each landmark position by applying local detectors to constrained regions around the current estimate. A number of feature detectors have been proposed for this purpose. One of the simplest, proposed in [16], is the linear logistic regressor which gives the following response map for the $i^{\text{th}}$ landmark[2]:

$$p(l_i = \text{aligned} \mid I, \mathbf{x}) = \frac{1}{1 + \exp\{\alpha\, C_i(I; \mathbf{x}) + \beta\}}, \quad (2)$$

where $l_i$ is a discrete random variable denoting whether the $i^{\text{th}}$ landmark is correctly aligned or not, $I$ is the image, $\mathbf{x}$ is a 2D location in the image, and $C_i$ is a linear classifier:

$$C_i(I; \mathbf{x}) = \mathbf{w}_i^T \left[\, I(\mathbf{y}_1)\ ;\ \ldots\ ;\ I(\mathbf{y}_m) \,\right] + b_i, \quad (3)$$

with $\{\mathbf{y}_i\}_{i=1}^m \in \mathbf{\Omega}_{\mathbf{x}}$ (i.e. an image patch). An advantage of using this classifier is that the map can be computed using efficient convolution operations. Other feature detectors have also been used to great effect, such as the Gaussian likelihood [2] and the Haar-based boosted classifier [3].

**Optimization:** Once the response maps for each landmark have been found, by assuming conditional independence, optimization proceeds by maximizing:

$$p(\{l_i = \text{aligned}\}_{i=1}^n \mid \mathbf{p}) = \prod_{i=1}^n p(l_i = \text{aligned} \mid \mathbf{x}_i) \quad (4)$$

with respect to the PDM parameters $\mathbf{p}$, where $\mathbf{x}_i$ is parameterized as in Equation (1) and dependence on the image $I$ is dropped for succinctness. It should be noted that some forms of CLMs pose Equation (4) as minimizing the summation of local energy responses (see §2.2).

The main difficulty in this optimization is how to avoid local optima whilst affording an efficient evaluation. Treating Equation (4) as a generic optimization problem, one may be tempted to utilize general purpose optimization strategies here. However, as the responses are typically noisy, these optimization strategies have a tendency to be unstable. The simplex based method used in [4] has been shown to perform reasonably for this task since it is a gradient-free based generic optimizer, which renders it somewhat insensitive to measurement noise. However, convergence may be slow when using this method, especially for a complex PDM with a large number of parameters.

## 2.2. Optimization Strategies

In this section, a review of current methods for CLM optimization is presented. These methods entail replacing the true response maps, $\{p(l_i|\mathbf{x})\}_{i=1}^n$, with simpler parametric forms and performing optimization over these instead of the original response maps. As these parametric density estimates are a kind of smoothed version of the original responses, sensitivity to local minima is generally reduced.

**Active Shape Models:** The simplest optimization strategy for CLM fitting is that used in the Active Shape Model (ASM) [2]. The method entails first finding the location within each response map for which the maximum was attained: $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1; \ldots; \boldsymbol{\mu}_n\right]$. 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:

$$Q(\mathbf{p}) = \sum_{i=1}^n w_i \|\mathbf{x}_i - \boldsymbol{\mu}_i\|^2, \quad (5)$$

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

$$\mathbf{x}_i \approx \mathbf{x}_i^c + \mathbf{J}_i \Delta\mathbf{p}, \quad (6)$$

[2]Not all CLM instances require a probabilistic output from the local detectors. Some, for example [2, 5], only require a similarity measure or a match score. However, these matching scores can be interpreted as the result of applying a monotonic function to the generating probability. For example, the Mahalanobis distance used in [2] is the negative log of the Gaussian likelihood. In the interest of clarity and succinctness, discussions in this work assume that responses are probabilities.

and solving for the parameter update:

$$\Delta \mathbf{p} = \left( \sum_{i=1}^{n} w_i \mathbf{J}_i^T \mathbf{J}_i \right)^{-1} \sum_{i=1}^{n} w_i \mathbf{J}_i^T \left( \boldsymbol{\mu}_i - \mathbf{x}_i^c \right), \quad (7)$$

which is then applied additively to the current parameters: $\mathbf{p} \leftarrow \mathbf{p} + \Delta \mathbf{p}$. Here, $\mathbf{J} = [\mathbf{J}_1; \ldots; \mathbf{J}_n]$ is the Jacobian and $\mathbf{x}^c = \left[\mathbf{x}_1^c; \ldots; \mathbf{x}_n^c\right]$ is the current shape estimate.

From the probabilistic perspective introduced in §2.1, the ASM's optimization procedure is equivalent to approximating the response maps with an isotropic Gaussian estimator:

$$p(l_i = \text{aligned} \mid \mathbf{x}) \approx \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I}), \quad (8)$$

where $w_i = \sigma_i^{-2}$. With this approximation, taking the negative log of the likelihood in Equation (4) results in the objective in Equation (5).

**Convex Quadratic Fitting:** 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 linear classifier described in §2.1, 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. The simplest 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 Figure 2).

To account for these problems, a method coined convex quadratic fitting (CQF) has been proposed recently [16]. The method fits a convex quadratic function to the negative log of the response map. This is equivalent to approximating the response map with a full covariance Gaussian:

$$p(l_i = \text{aligned} \mid \mathbf{x}) \approx \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (9)$$

The mean and covariance are maximum likelihood estimates given the response map:

$$\boldsymbol{\Sigma}_i = \sum_{\mathbf{x} \in \boldsymbol{\Psi}_{\mathbf{x}_i^c}} \alpha_{\mathbf{x}}^i \left( \mathbf{x} - \boldsymbol{\mu}_i \right) \left( \mathbf{x} - \boldsymbol{\mu}_i \right)^T ; \; \boldsymbol{\mu}_i = \sum_{\mathbf{x} \in \boldsymbol{\Psi}_{\mathbf{x}_i^c}} \alpha_{\mathbf{x}}^i \, \mathbf{x}, \quad (10)$$

where $\boldsymbol{\Psi}_{\mathbf{x}_i^c}$ is a 2D-rectangular grid centered at the current landmark estimate $\mathbf{x}_i^c$ (i.e. the search window), and:

$$\alpha_{\mathbf{x}}^i = \frac{p(l_i = \text{aligned} \mid \mathbf{x})}{\sum_{\mathbf{y} \in \boldsymbol{\Psi}_{\mathbf{x}_i^c}} p(l_i = \text{aligned} \mid \mathbf{y})}. \quad (11)$$

With this approximation, the objective can be written as the minimization of:

$$Q(\Delta \mathbf{p}) = \sum_{i=1}^{n} \| \mathbf{x}_i^c + \mathbf{J}_i \Delta \mathbf{p} - \boldsymbol{\mu}_i \|^2_{\boldsymbol{\Sigma}_i^{-1}}, \quad (12)$$

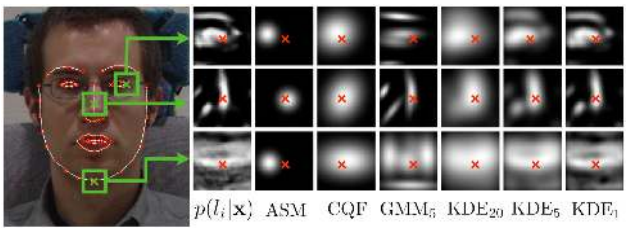



Figure 2. Response maps, $p(l_i = \text{aligned}|\mathbf{x})$, and their approximations used in various methods, for the outer left eye corner, the nose bridge and chin. Red crosses on the response maps denote the true landmark locations. The GMM approximation has five cluster centers. The KDE approximations are shown for $\sigma^2 \in \{20, 5, 1\}$.

the solution of which is given by:

$$\Delta \mathbf{p} = \left( \sum_{i=1}^{n} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{J}_i^T \boldsymbol{\Sigma}_i^{-1} \left( \boldsymbol{\mu}_i - \mathbf{x}_i^c \right). \quad (13)$$

**A Gaussian Mixture Model Estimate:** Although the response map approximation in CQF may overcome some of the drawbacks of ASM, 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 Figure 2).

To account for this, in [8] a Gaussian mixture model (GMM) was used to approximate the response maps:

$$p(l_i = \text{aligned} \mid \mathbf{x}) \approx \sum_{k=1}^{K} \pi_{ik} \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \quad (14)$$

where $K$ denotes the number of modes and $\{\pi_{ik}\}_{k=1}^K$ are the mixing coefficients for the GMM of the $i^{\text{th}}$ PDM 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:

$$p(\{l_i\}_{i=1}^n | \mathbf{p}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_i(z_i = k, l_i | \mathbf{x}_i). \quad (15)$$

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

$$p(z_i = k \mid l_i, \mathbf{x}_i) = \frac{p(z_i = k) \, p(l_i | z_i = k, \mathbf{x}_i)}{\sum_{j=1}^{K} p(z_i = j) \, p(l_i | z_i = j, \mathbf{x}_i)}, \quad (16)$$

where $p(z_i = k) = \pi_{ik}$ and:

$$p(l_i = \text{aligned} \mid z_i = k, \mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i \; ; \; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}). \quad (17)$$

In the M-step, the expectation of the negative log of the complete data is minimized:

$$Q(\mathbf{p}) = E_{q(\mathbf{z})}\left[-\log\left\{\prod_{i=1}^{n} p(l_i = \text{aligned}, z_i|\mathbf{x}_i)\right\}\right], \tag{18}$$

where $q(\mathbf{z}) = \prod_{i=1}^{n} p_i(z_i|l_i = \text{aligned}, \mathbf{x}_i)$. Linearizing the shape model as in Equation (6), this $Q$-function takes the form:

$$Q(\Delta\mathbf{p}) \propto \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\|\mathbf{J}_i\Delta\mathbf{p} - \mathbf{y}_{ik}\|^2_{\mathbf{\Sigma}_{ik}^{-1}} + \text{const}, \tag{19}$$

where $w_{ik} = p_i(z_i = k|l_i = \text{aligned}, \mathbf{x}_i)$ and $\mathbf{y}_{ik} = \boldsymbol{\mu}_{ik} - \mathbf{x}_i^c$, the solution of which is given by:

$$\Delta\mathbf{p} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\mathbf{J}_i\right)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\mathbf{y}_{ik}. \tag{20}$$

Although the GMM is a better approximation of the response map compared to the Gaussian approximation in CQF, it exhibits two major drawbacks. Firstly, the process of estimating the GMM parameters from the response maps is a nonlinear 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 [8], the modes are chosen as the $K$-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 CLM fitting algorithm. Such an approximation allows an efficient estimate of the GMM parameters without the need for a costly EM procedure at the cost of a poorer approximation of the true response map.

The second drawback of the GMM response map approximation is that the approximated objective in Equation (15) is multimodal. As such, CLM fitting with the GMM simplification is prone to terminating in local optima. Although good results were reported in [8], in that work the PDM was parameterized using a mixture model as opposed to the more typical Gaussian parameterization, which places a stronger prior on the way the shape can vary.

## 3. Subspace Constrained Mean-Shifts

Rather than approximating the response maps for each PDM landmark using parametric models, we consider here the use of a nonparametric representation. In particular, we propose the use of a homoscedastic kernel density estimate (KDE) with an isotropic Gaussian kernel:

$$p(l_i = \text{aligned}|\mathbf{x}) \approx \sum_{\boldsymbol{\mu}_i \in \mathbf{\Psi}_{\mathbf{x}_i^c}} \alpha^i_{\boldsymbol{\mu}_i}\,\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \sigma^2\mathbf{I}), \tag{21}$$

where $\alpha^i_{\boldsymbol{\mu}_i}$ is the normalized true detector response defined in Equation (11). With this representation the kernel centers are fixed as defined through $\mathbf{\Psi}_{\mathbf{x}_i^c}$ (i.e. the grid nodes of the search window). The mixing weights, $\alpha^i_{\boldsymbol{\mu}_i}$, can be obtained directly from the true response map. Since the response is an estimate of the probability that a particular location in the image is the aligned landmark location, such a choice for the mixing coefficients is reasonable. Compared to parametric representations, KDE has the advantage that no nonlinear optimization is required to learn the parameters of its representation. The only remaining free parameter is the variance of the Gaussian kernel, $\sigma^2$, which regulates the smoothness of the approximation. Since one of the main problems with a GMM based representation is the computational complexity and suboptimal nature of fitting a mixture model to the response maps, if $\sigma^2$ is set *a-priori*, then optimizing over the KDE can be expected to be more stable and efficient.

Maximizing the objective in Equation (4) with a KDE representations is nontrivial as the objective is nonlinear and typically multimodal. However, in the case where no shape prior is placed on the way the PDM's landmarks can vary, the problem reverts to independent maximizations of the KDE for each landmark location separately. This is because the landmark detections are assumed to be independent, conditioned on the PDM's parameterization. A common approach for maximization over a KDE is to use the well known mean-shift algorithm [1]. It consists of fixed point iterations of the form:

$$\mathbf{x}_i^{(\tau+1)} \leftarrow \sum_{\boldsymbol{\mu}_i \in \mathbf{\Psi}_{\mathbf{x}_i^c}} \frac{\alpha^i_{\boldsymbol{\mu}_i}\mathcal{N}\left(\mathbf{x}_i^{(\tau)}; \boldsymbol{\mu}_i, \sigma^2\mathbf{I}\right)}{\sum_{\mathbf{y}\in\mathbf{\Psi}_{\mathbf{x}_i^c}} \alpha^i_{\mathbf{y}}\,\mathcal{N}\left(\mathbf{x}_i^{(\tau)}; \mathbf{y}, \sigma^2\mathbf{I}\right)}\boldsymbol{\mu}_i, \tag{22}$$

where $\tau$ denotes the time-step in the iterative process. This fixed point iteration scheme finds a mode of the KDE, where an improvement is guaranteed at each step by virtue of its interpretation as a lower bound maximization [6]. Compared to other optimization strategies, mean-shift is an attractive choice as it does not use a step size parameter or a line search. Equation (22) is simply applied iteratively until some convergence criterion is met.

To incorporate the shape model constraint into the optimization procedure, one might consider a two step strategy: (i) compute the mean-shift update for each landmark, and (ii) constrain the mean-shifted landmarks to adhere to the PDM's parameterization using a least-squares fit:

$$Q(\mathbf{p}) = \sum_{i=1}^{n}\left\|\mathbf{x}_i - \mathbf{x}_i^{(\tau+1)}\right\|^2. \tag{23}$$

This is reminiscent of the ASM optimization strategy, where the location of the response map's peak is replaced with the mean-shifted estimate. Although such a strategy is attractive in its simplicity, it is unclear how it relates to the global

**Algorithm 1** Subspace Constrained Mean-Shifts

**Require:** $I$ and $\mathbf{p}$.

**while** not_converged($\mathbf{p}$) **do**
  Compute responses {Eqn. (2)}
  Linearize shape model {Eqn. (6)}
  Precompute pseudo-inverse of Jacobian ($\mathbf{J}^\dagger$)
  Initialize parameter updates: $\Delta\mathbf{p} \leftarrow \mathbf{0}$
  **while** not_converged($\Delta\mathbf{p}$) **do**
    Compute mean-shifted landmarks {Eqn. (22)}
    Apply subspace constraint {Eqn. (24)}
  **end while**
  Update parameters: $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$
**end while**
**return** $\mathbf{p}$

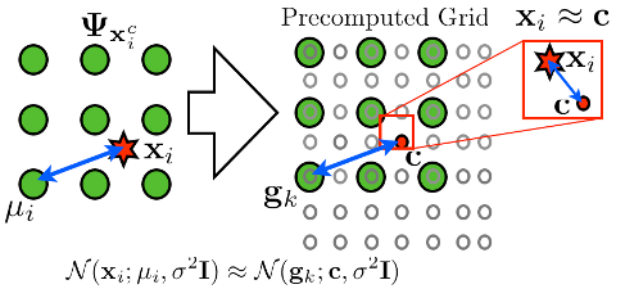



Figure 3. Illustration of a the use of a precomputed grid for efficient mean-shift. Kernel evaluations are precomputed between $\mathbf{c}$ and all other nodes in the grid. To approximate the true kernel evaluation, $\mathbf{x}_i$ is assumed to coincide with $\mathbf{c}$ and the likelihood of any response map grid location can be attained by a table lookup.

objective in Equation (4).

Given the form of the KDE representation in Equation (21), one can treat it simply as a GMM. As such, the discussions in §2.2 on GMMs are directly applicable here, replacing the number of candidates $K$ with the number of grid nodes in the search window $\mathbf{\Psi}_{\mathbf{x}_i^c}$, the mixture weights $\pi_{ik}$ with $\alpha^i_{\boldsymbol{\mu}_i}$, and the covariances $\mathbf{\Sigma}_{ik}$ with the scaled identity $\sigma^2\mathbf{I}$. When using the linearized shape model in Equation (6) and maximizing the global objective in Equation (4) using the EM algorithm, the solution for the so called $Q$-function of the M-step takes the form:

$$\Delta\mathbf{p} = \mathbf{J}^\dagger \left[\mathbf{x}_1^{(\tau+1)} - \mathbf{x}_1^c \; ; \; \ldots \; ; \; \mathbf{x}_n^{(\tau+1)} - \mathbf{x}_n^c\right], \qquad (24)$$

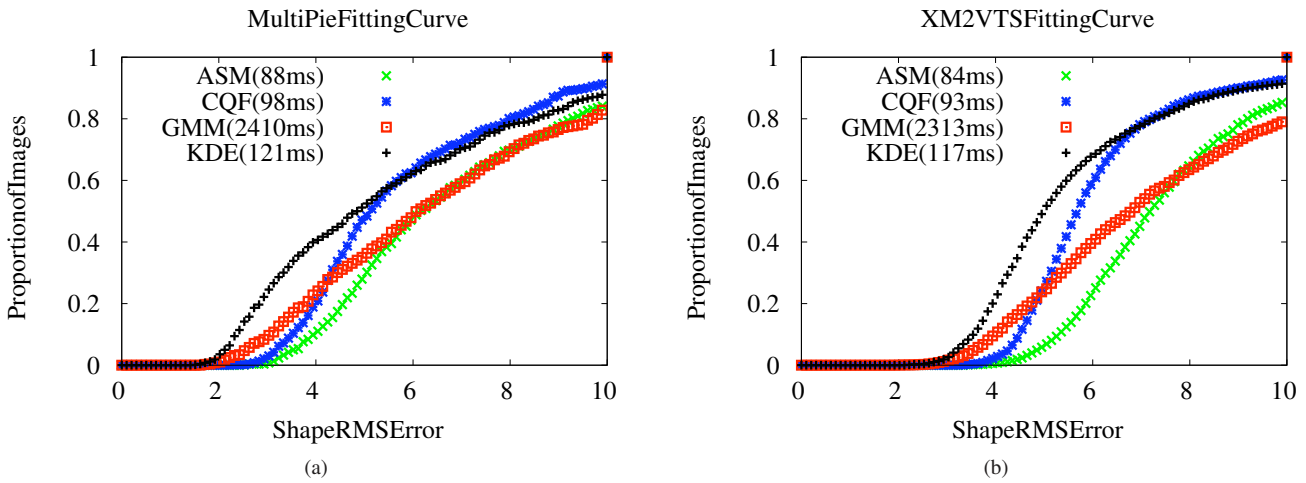
where $\mathbf{J}^\dagger$ denotes the pseudo-inverse of $\mathbf{J}$, and $\mathbf{x}_i^{(\tau+1)}$ is the mean shifted update for the $i^{\text{th}}$ landmark given in Equation (22). This is simply the Gauss Newton update for the least squares PDM constraint in Equation (23). As such, under a linearized shape model, the two step strategy for maximizing the objective in Equation (4) with a KDE representation shares the properties of a general EM optimization, namely: provably improving and convergent. The complete fitting procedure, which we will refer to as subspace constrained mean-shifts (SCMS), is outlined in Algorithm 1. In the following, two further innovations are proposed, which address difficulties regarding local optima and the computational expense of kernel evaluations.

**Kernel Width Relaxation:** The response map approximations discussed in §2.2 can be though of as a form of smoothing. This explains the relative performance of the various methods. The Gaussian approximations smooth the most but approximate the true response map the poorest, whereas smoothing effected by the GMM is not as aggressive but exhibits of a degree of sensitivity towards local optima. One might consider using the Gaussian and GMM approximations in tandem, where the Gaussian approximation is used to get within the convergence basin of the 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 SCMS 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 Figure 2). 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 $\sigma^2 < \infty$ such that the KDE is unimodal, regardless of the distribution of samples [13]. As $\sigma^2$ 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 $\sigma^2$ is closest to the desired mode of the objective with a smaller $\sigma^2$, promoting its convergence to the correct mode. As such, the policy under which $\sigma^2$ is reduced acts to guide optimization towards the global optimum of the true objective.

Drawing parallels with existing methods, as $\sigma^2 \to \infty$ the SCMS update approaches the solution of a homoscedastic Gaussian approximated objective function. As $\sigma^2$ is reduced, the KDE approximation resembles a GMM approximation, where the approximation for smaller $\sigma^2$ settings is similar to a GMM approximation with more modes.

**Precomputed Grid:** 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 $K < 10$, whereas the search window size typically has $> 100$ nodes). As such, the computation of the *posterior* in Equation (16) will be more costly. However, if the vari-




(a) (b)

Figure 4. Fitting Curves for the ASM, CQF, GMM and KDE optimization strategies on the MultiPie and XM2VTS databases.

ance $\sigma^2$ is known *a-priori*, then some approximations can be made to significantly reduce computational complexity.

The main overhead when computing the mean-shift update is in evaluating the Gaussian kernel between the current landmark estimate and every grid node in the response map. Since the grid locations are fixed and $\sigma^2$ 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 $\mathbf{\Psi}_{\mathbf{x}_i^c}$. 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 Figure 3). 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.

Although such an approximation ruins the strictly improving properties of EM, we empirically show in §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.

## 4. Experiments

**Database Specific Experiments:** We compared the various CLM optimizations strategies discussed above on the problem of generic frontal face fitting on two databases: (i) the CMU Pose, Illumination and Expression Database (MultiPie) [7], and (ii) the XM2VTS database [12]. The MultiPie database is annotated with a 68-point markup used as ground truth landmarks. We used 762 frontal face images of 339 subjects. The XM2VTS database 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. A 4-fold cross validation was performed on both MultiPie and XM2VTS, separately, where the images were partitioned into three 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 [2], (ii) CQF [16], (iii) GMM [8], and (iv) the KDE method proposed in §3. For GMM, we empirically set $K = 5$ and used the EM algorithm to estimate the parameters of the mixture model. For KDE, we used a variance relaxation policy of $\sigma^2 = \{20, 10, 5, 1\}$ and a grid spacing of 0.1-pixels in its efficient approximation. In all cases the linear logistic regressor described in §2.1 was used. The local experts were $(11 \times 11)$-pixels in size and the exhaustive local search was performed over a $(15 \times 15)$-pixel window. As such, the only difference between the various methods compared here is their optimization strategy. 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 Equation (1) set to zero (i.e. the mean shape), and the model fit until the optimization converged.

Results of these experiments can be found in Figure 4, where 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. The average fitting times for the various methods on a 2.5GHz 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 capac-

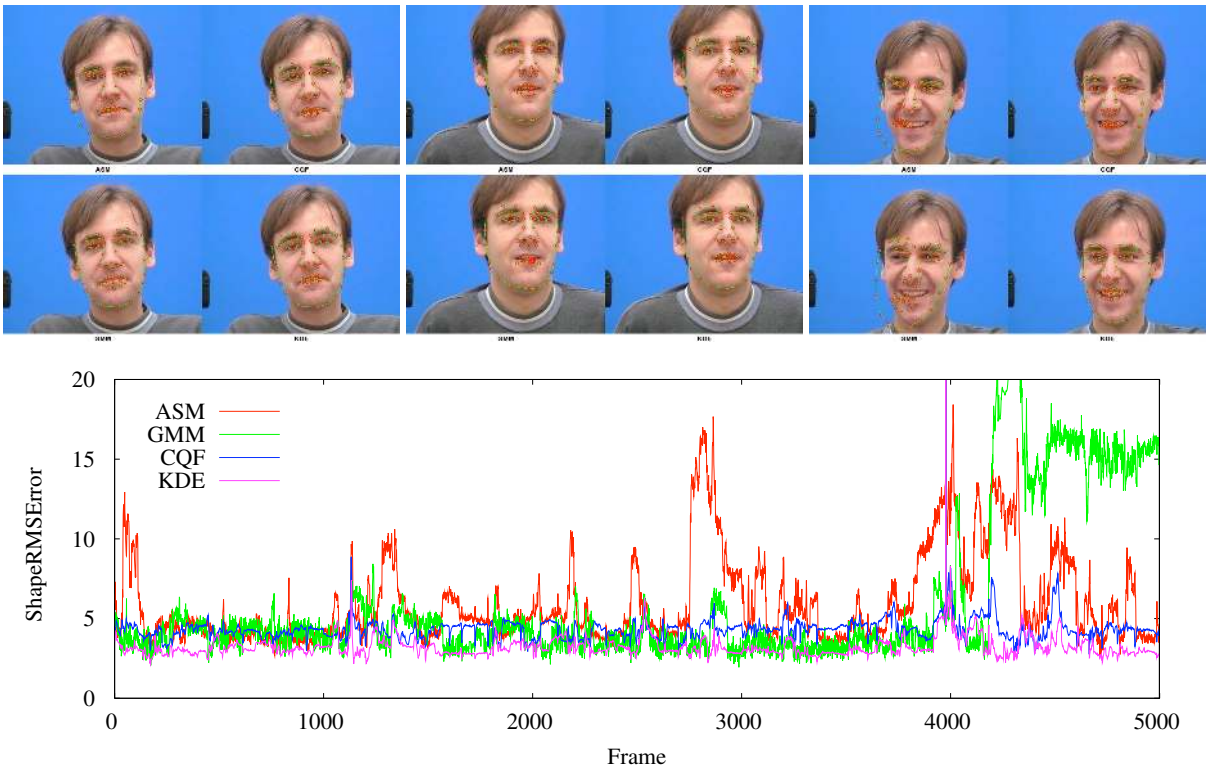



Figure 5. **Top row:** Tracking results on the FGNet Talking Face database for frames $\{0, 1230, 4200\}$. Clockwise from top left are fitting results for ASM, CQF, KDE and GMM. **Bottom:** Plot of shape RMS error from ground truth annotations throughout the sequence.

ity to significantly outperform ASM. As discussed in §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 has the tendency to terminate in 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, KDE is capable of attaining even better accuracies than GMM but still retains a degree of robustness towards local optima, where its performance over grossly misplaced initializations is comparable to CQF. Finally, despite the significant improvement in performance, KDE 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.

**Out-of-Database Experiments:** Testing the performance of fitting algorithms on images outside of a particular database is more meaningful as it gives a better indication on how well the 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[3]. Quantitative analysis on this sequence is possible since ground truth annotations are available in the same format as that in XM2VTS. We initialize the model using a face detector in the first frame and fit consecutive frames using the PDM's configuration in the previous frame as an initial estimate. The same model used in the database-specific experiments was used here, except that it was trained on all images in XM2VTS. In Figure 5, 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 in the database-specific experiments, with KDE yielding the best performance. ASM and GMM are particularly unstable on this sequence, with GMM loosing track at around frame 4200, and fails to recover until the end of the sequence.

Finally, we performed a qualitative analysis of KDE's performance on the Faces in the Wild database [9]. 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 using the XM2VTS

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

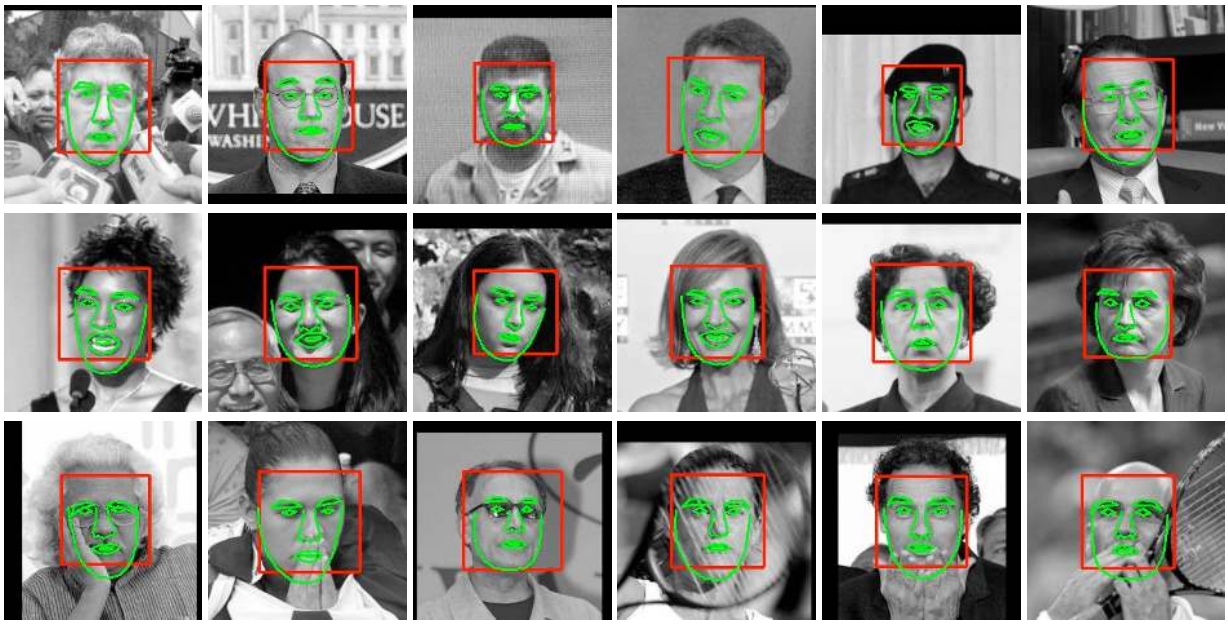


Figure 6. Example fitting results on the Faces in the Wild database using a model trained using the XM2VTS database. **Top row:** Male subjects. **Middle row:** female subjects. **Bottom row:** partially occluded faces.

trained model. Some fitting results are shown in Figure 6. Results suggest that KDE exhibits a degree of robustness towards variations typically encountered in real images.

## 5. Conclusion

The optimization strategy for 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. A new approximation was then proposed that uses a nonparametric representation. Two further innovations were proposed in order to reduce computational complexity and avoid local optima. The proposed method was shown to outperform three other optimization strategies on the task of generic face fitting. Future work will involve investigations into the effects of different local detectors types and shape priors on the optimization strategies.

## References


[1] Y. Cheng. Mean Shift, Mode Seeking, and Clustering. *PAMI*, 17(8):790–799, 1995.

[2] T. F. Cootes and C. J. Taylor. Active Shape Models - 'Smart Snakes'. In *BMVC*, pages 266–275, 1992.

[3] D. Cristinacce and T. Cootes. Boosted Active Shape Models. In *BMVC*, volume 2, pages 880–889, 2007.

[4] D. Cristinacce and T. F. Cootes. A Comparison of Shape Constrained Facial Feature Detectors. In *FG*, pages 375–380, 2004.

[5] D. Cristinacce and T. F. Cootes. Feature Detection and Tracking with Constrained Local Models. In *EMCV*, pages 929–938, 2004.

[6] M. Fashing and C. Tomasi. Mean Shift as a Bound Optimization. *PAMI*, 27(3), 2005.

[7] R. Gross, I. Matthews, S. Baker, and T. Kanade. The CMU Multiple Pose, Illumination and Expression (MultiPIE) Database. Technical report, Robotics Institute, Carnegie Mellon University, 2007.

[8] L. Gu and T. Kanade. A Generative Shape Regularization Model for Robust Face Alignment. In *ECCV'08*, 2008.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[10] X. Liu. Generic Face Alignment using Boosted Appearance Model. In *CVPR*, pages 1–8, 2007.

[11] I. Matthews and S. Baker. Active Appearance Models Revisited. *IJCV*, 60:135–164, 2004.

[12] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *AVBPA*, pages 72–77, 1999.

[13] M. A. C.-P. nán and C. K. I. Williams. On the Number of Modes of a Gaussian Mixture. *Lecture Notes in Computer Science*, 2695:625–640, 2003.

[14] M. H. Nguyen and F. De la Torre Frade. Local Minima Free Parameterized Appearance Models. In *CVPR*, 2008.

[15] J. Saragih and R. Goecke. A Nonlinear Discriminative Approach to AAM Fitting. In *ICCV*, 2007.

[16] Y. Wang, S. Lucey, and J. Cohn. Enforcing Convexity for Improved Alignment with Constrained Local Models. In *CVPR*, 2008.