Deformable Model Fitting with a Mixture of Local Experts

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

Local experts have been used to great effect for fitting deformable models to images. Typically, the best location in an image for the deformable model's landmarks are found through a locally exhaustive search using these experts. In order to achieve efficient fitting, these experts should afford an efficient evaluation, which often leads to forms with restricted discriminative capacity. In this work, a framework is proposed in which multiple simple experts can be utilized to increase the capacity of the detections overall. In particular, the use of a mixture of linear classifiers is proposed, the computational complexity of which scales linearly with the number of mixture components. The fitting objective is maximized using the expectation maximization (EM) algorithm, where approximations to the true objective are made in order to facilitate efficient and numerically stable fitting. The efficacy of the proposed approach is evaluated on the task of generic face fitting where performance improvement is observed over two existing methods.

1. Introduction

Deformable model fitting is the problem of finding the optimal configuration of a parameterized shape model that best describes the object of interest in an image. Objects that are typically modeled in this way include the human face [3, 19] and organs in medical image analysis [22, 24]. Numerous representations and fitting strategies have been proposed for these objects, most of which can be categorized based on their representations as being either holistic or patch-based.

Holistic representations, for example [3, 12, 21], model the appearance of all image pixels describing the object. The advantage of such a representation is that all available data is used simultaneously during fitting. As such, these methods have the capacity to attain highly accurate fitting. However, such a representation generalizes poorly when the object of interest exhibits large amounts of variability, such as in the case of the human face under variations in identity, expression, pose and lighting [8, 19]. This is due to the high dimensionality of the represented appearance and the typically limited amount of available training data. It has been shown in [1, 19] that a parts-based representation can improve the model's representation capacity as it accounts only for local correlations between pixels values.

Patch-based approaches, for example [4, 6, 25], model the appearance around each landmark of the parameterized shape model independently of all others. They exhibit good generalization with limited data and can offer a degree of robustness towards changes in lighting conditions. Unlike holistic based approaches, where fitting is generally posed either as a regression between the image and the parameter updates [20, 21, 28] or as the deterministic minimization of some kind of fitting criterion [12, 14, 17], patch-based deformable model fitting typically proceeds by exhaustively searching for the best landmark locations in the image that are then constrained to adhere to the shape model's parameterization. Care should to be taken here with regards to how the landmark locations are constrained to reflect confidence over their detections [10, 25]. Local experts used in the exhaustive search generally require an efficient evaluation. This places limits on the complexity, and hence capacity, of these experts. For complex visual objects, simple local experts may be unable to discriminate correct from incorrect locations, limiting the fidelity of such an approach.

In this work, we propose a principled way of combining responses for each landmark from an ensemble of simple local experts. This is achieved by posing the fitting problem probabilistically, where the likelihood of each landmark location is approximated using a mixture model. Each component of the mixture is of the result of an exhaustive search with a local expert. The parameters of the shape model describing the object in the image are found by maximizing the likelihood over all landmark location through the expectation maximization (EM) algorithm. We begin in §2 with a brief overview of relevant work, where the parameterization of the shape model and local experts are discussed. The fitting algorithm is then presented in §3. Results of empirical experiments investigating the efficacy of this approach are presented in §4. We conclude in §5 with a brief overview and mention of future work.

2. Background

Deformable model fitting is a heavily researched topic. In this section, we give a brief overview of advances in local experts based deformable model fitting, motivating the use of mixture of experts in §3, where the parameterization is discussed in §2.1 and the local experts in §2.2.

2.1. Constrained Local Models

The most typical parameterization of a deformable visual object is that of the point distribution model (PDM) [4]. It assumes a linear generative model on the non-rigid variations of the object [1]:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{\Phi}\,\mathbf{q}, \tag{1}$$

where $\mathbf{s}$ denotes the coordinates of the shape's landmarks in the pose normalized frame, $\bar{\mathbf{s}}$ denotes the mean shape, $\mathbf{\Phi}$ is a matrix whose columns consist of the directions of variability, and $\mathbf{q}$ are the parameters of the model. In this work we utilize a 3D linear shape model (i.e. $\mathbf{s} = [x_1; y_1; z_1; \ldots; x_n; y_n; z_n]$), where $\bar{\mathbf{s}}$ and $\mathbf{\Phi}$ are found by applying non-rigid structure from motion on a set of 2D shapes [23]. The shape of the visual object in the image frame is obtained by projecting $\mathbf{s}$ onto the image with a suitable scaling, rotation and translation:

$$\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] = \alpha\,(\mathbf{I} \otimes \mathbf{R})\,\mathbf{s} + \mathbf{1} \otimes [t_x; t_y]. \tag{2}$$

Here, α denotes the global scaling, $[t_x; t_y]$ denotes the global translation and $\mathbf{R}$ denotes a truncated rotation matrix (i.e. we use a weak perpective model). In discussions that follow, the parameter set describing the PDM is denoted $\mathbf{p} = \{\alpha, \mathbf{R}, t_x, t_y, \mathbf{q}\}$.

Patch-based approaches to deformable model fitting typically involve an exhaustive local search for the best location of each PDM landmark that are then constrained to adhere to the PDM's parameterization. This is typically achieved through a least squares fit:

$$Q(\mathbf{p}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}_i\|^2_{\mathbf{W}_i}, \tag{3}$$

where $\{\boldsymbol{\mu}_i\}_{i=1}^n$ denote the locations found by the exhaustive local search procedure and $\{\mathbf{W}_i\}_{i=1}^n$ are weighting matrices that represent the importance of matching to any particular landmark. In this work, we will collectively refer to all methods that utilize such a fitting strategy as constrained local models (CLM) [2].

2.2. Local Experts

The exhaustive local search for the best location of each PDM landmark is typically realized through the use of local experts. These local experts can be roughly divided into two types: generative and discriminative.

Generative local experts model the distribution of local patch appearance across the many instantiations of the visual object. The simplest patch expert assumes a Gaussian distribution on patch appearance. When the appearance distribution of the model is assumed to be full rank, the matching criterion used by the Gaussian local expert is the Mahalanobis distance [4]. However, such a criterion is computationally expensive to evaluate. In [4], both the patch and search region are constrained to lie along profiles, allowing an efficient evaluation. In general, where a rectangular patch representation and search window are utilized, a more efficient criterion is the distance-from-feature-space [16]. This involves a truncated basis of patch appearance, the dimensionality of which is usually much smaller than the dimensionality of the patch. However, this may still incur significant computational costs. In [6], a combined deterministic/stochastic optimization strategy is utilized, where a search for the optimal PDM parameters is alternated with solutions for the optimal parameters of the truncated patch appearance model. For a fixed shape, the optimal patch appearance parameters can be computed in closed form by virtue of the use of a linear generative appearance model. For a fixed appearance, the optimal landmark locations are found through a stochastic sampling of the spatial domain (i.e. the exhaustive local search).

Although generative approaches have an intuitive appeal, they are optimally constructed for synthesis rather than classification. To address this drawback, some authors propose using discriminative local experts, where both aligned and misaligned examples are considered during training [5, 25]. In the interest of facilitating an efficient evaluation, discriminative local experts are generally limited in their capacity. For example, in [25], a linear support vector machine (SVM) was used for this purpose. It allows an extremely efficient evaluation as it requires only the inner product of two patch-sized vectors in its evaluation. However, a linear classifier may be too restrictive to account for the large inter-class variability exhibited by many deformable objects. In [5], a boosted ensemble of weak classifiers was used that allowed an efficient evaluation through the utility of Haar-based features.

A third category of CLM fitting methods is one that uses a displacement expert for each PDM landmark [5, 26]. Such

[1] **Notation:** Function names are written in upper case, scalars in lower case, vectors in lowercase bold and matrices in uppercase bold, where $\mathbf{I}$ denotes the identity matrix. Greek letters denote either vectors or matrices depending on context. The $\otimes$ operator denotes the Kronecker (tiling) product. $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a Gaussian probability density function over the random variable $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$. The elliptical error norm is written: $\|\mathbf{x}\|^2_{\mathbf{W}} = \mathbf{x}^T\mathbf{W}\mathbf{x}$.

[2] This term should not be confused with the work in [6] which is a particular instance of CLM in our nomenclature.

an approach has the potential to be more efficient since a locally exhaustive search is avoided. However, since the displacement expert needs to determine both the direction and magnitude of misalignment, such an approach is more difficult to train and exhibits poorer generalization.

3. Fitting with a Mixture of Local Experts

As previously mentioned, local experts in the CLM fitting framework require an efficient evaluation as well as accurate estimates of the correct landmark locations. The effects of these two opposing criteria are most clearly observed when the object of interest exhibits large intra-class variability. If the model is too simple, it lacks the capacity to accurately distinguish aligned from missaligned locations (see experimental results in §4). If the model is too complex it requires a large computational overhead as well as exhibiting generalization difficulties.

In terms of computational efficiency, no expert is better than the linear classifier used in [25]. Further computational savings can be made through an intelligent feature selection scheme, such as that proposed in [18]. However, its domain of application is limited to cases where the data is linearly separable. In this section, we make the case for the use of a mixture of linear experts and demonstrate how they can be efficiently integrated into a CLM fitting framework. Although the discussion here centers on the use of linear classifiers, the proposed optimization strategy can be applied to any combination of local experts that afford a probabilistic interpretation.

3.1. Mixture of Linear Experts

In this work we are interested in a particular class of non-linear classifiers that take the form of an additive ensemble of simple classifiers:

$$E(\mathbf{d}) = \sum_{i=1}^{K} \pi_i \; E_i(\mathbf{d}), \tag{4}$$

where $\mathbf{d}$ denotes the observed data (i.e. the cropped image patch), E_i denotes the i^{th} simple classifier and $\{\pi_i\}_{i=1}^{K}$ are the mixing coefficients. Classifiers that exhibit such a form include the boosted set of weak learners and kernel SVM. In particular, we are interested in a probabilistic interpretation of classification, where E_i denotes the likelihood of $\mathbf{d}$ being positive data (i.e. observed from the aligned landmark location) given that the i^{th} simple expert is selected for classification, and π_i denotes the probability of selecting that expert (i.e. $\sum_{i=1}^{K} \pi_i = 1$).

Within a probabilistic CLM framework, the parametric form of a mixture of experts for the i^{th} PDM landmark takes the following form:

$$p(l_i|\mathbf{x}_i, I) = \sum_{k=1}^{K_i} p(z_i = k|\mathbf{x}, I) \; p(l_i|z_i = k, \mathbf{x}_i, I), \tag{5}$$

where $l_i \in \{\text{aligned}, \text{missaligned}\}$, I denotes the image, z_i denotes the mixture membership and K_i denotes the number of mixture components for the i^{th} landmark. Here, we model the likelihood of a landmark being correctly aligned, given the choice of its location and mixture component, using the exponential of the negative hinge loss (i.e. the data term of a linear SVM) [2]:

$$p(l_i = \text{aligned}|z_i = k, \mathbf{x}_i, I) = e^{-\max(1-\mathbf{w}_{ik}^T C(\mathbf{x}_i;I)+b_{ik},0)}, \tag{6}$$

where C crops the image I around $\mathbf{x}$:

$$C(\mathbf{x}; I) = [I(\mathbf{y}_1); \ldots; I(\mathbf{y}_m)] \; ; \; \{\mathbf{y}_i\}_{i=1}^{m} \in \mathbf{\Omega}_{\mathbf{x}}, \tag{7}$$

with $\mathbf{\Omega}_{\mathbf{x}}$ denoting a rectangular region centered around $\mathbf{x}$ in the image (i.e. the image patch). The mixture weights in Equation (5) capture the likelihood of a particular mixture component membership. For this, we use a data-driven model, a convenient choice of which is the multinomial log-linear model [11]:

$$\pi_{ik}(\mathbf{x}; I) = p(z_i = k|\mathbf{x}, I) = \frac{e^{\gamma \mathbf{v}_{ik}^T C(\mathbf{x};I)}}{\sum_{j=1}^{K_i} e^{\gamma \mathbf{v}_{ij}^T C(\mathbf{x};I)}}, \tag{8}$$

where C is as in Equation (7) and γ controls the "softness" of the assignment that can be learned from the data though cross validation. Making the mixture weights data-dependent allows more flexibility in modeling and the ability to cope with ambiguous data labels.

The advantage of a mixture of linear experts over more general classification schemes is twofold. First, the computational complexity of a mixture of linear experts grows only linearly with the number of mixture components. Second, the generalization capacity of the model is directly related to the number of mixtures used, akin to the number of weak learners used in a boosting framework [13]. Compared to kernel SVM, for example, mixture of linear SVMs has been shown to exhibit similar classification accuracy and generalization whilst affording a much reduced computational complexity [7].

3.2. Optimization through the EM Algorithm

With a probabilistic interpretation of the local expert responses, the problem of CLM fitting can be posed as finding the parameters of the PDM that maximize the likelihood that its landmarks are correctly aligned[3]:

$$p(\{l_i = \text{aligned}\}_{i=1}^n | \{\mathbf{x}_i\}_{i=1}^n) = \prod_{i=1}^n p(l_i = \text{aligned}|\mathbf{x}_i), \tag{9}$$

where landmark detections are assumed to be conditionally independent. Regardless of the type of classifier used to define $\{p(l_i|\mathbf{x}_i)\}_{i=1}^n$, Equation (9) is a nonlinear function of the PDM parameters $\mathbf{p}$, which define $\{\mathbf{x}_i\}_{i=1}^n$ through Equations (1) and (2). Although a solution can be obtained by using a general purpose optimization strategy, the particular form of the objective allows a more specialized treatment.

Treating the local expert membership $\mathbf{z} = \{z_i\}_{i=1}^n$ as latent variables, the objective in Equation (9) can be optimized using the EM algorithm. The E-step involves the computation of the *posterior* over the latent variables:

$$p(z_i = k|l_i, \mathbf{x}_i) = \frac{\pi_{ik}(\mathbf{x}_i)\; p(l_i|z_i = k, \mathbf{x}_i)}{\sum_{j=1}^{K_i} \pi_{ij}(\mathbf{x}_i)\; p(l_i|z_i = j, \mathbf{x}_i)}. \tag{10}$$

In the M-step, the expectation of the negative log of the complete data is minimized with respect to the parameters of the PDM:

$$Q(\mathbf{p}) = E_{q(\mathbf{z})}\left[-\log \left\{ \prod_{i=1}^n p(l_i = \text{aligned}, z_i|\mathbf{x}_i) \right\} \right], \tag{11}$$

where $q(\mathbf{z}) = \prod_{i=1}^n p(z_i|l_i = \text{aligned}, \mathbf{x}_i^c)$, with $\mathbf{x}_i^c$ denoting the current estimate of the i^{th} landmark. Maximizing this objective function with respect to the PDM parameters $\mathbf{p}$ is difficult as both $\pi_{ik}(\mathbf{x}_i)$ and $p(l_i = \text{aligned}|z_i, \mathbf{x}_i)$ are nonlinear in $\mathbf{x}_i$. In addition, the landmark locations are related nonlinearly to the PDM parameters, as defined through Equations (1) and (2). In the following, we describe some approximations that greatly simplify the computations involved in the EM algorithm.

Likelihood Approximation: Regardless of the type of classifier used in the mixture model, $p(l_i = \text{aligned}|z_i, \mathbf{x}_i)$ will generally be nonlinear as it involves the extraction of pixel intensities from the image, defined in Equation (7), which are generally related to the PDM parameters nonlinearly. Recently, in a method coined convex quadratic fitting (CQF) [25], the authors propose substituting the true responses for each landmark with a convex quadratic that best matches the responses locally. This greatly simplifies the optimization procedure, allowing a closed form solution to be attained. Furthermore, such an approximation has been shown to approximately preserve the expert's directional uncertainty (i.e. the aperture problem), which is prevalent in patch based detection due to the limited structure of the data. A probabilistic interpretation of CQF is that the expert responses denote the negative log of a generating probability density function (PDF), in which the quadratic approximation is equivalent to assuming the PDF is Gaussian in the spatial dimensions. Utilizing such an approximation here, the likelihood of the i^{th} PDM landmark being correctly aligned is given by:

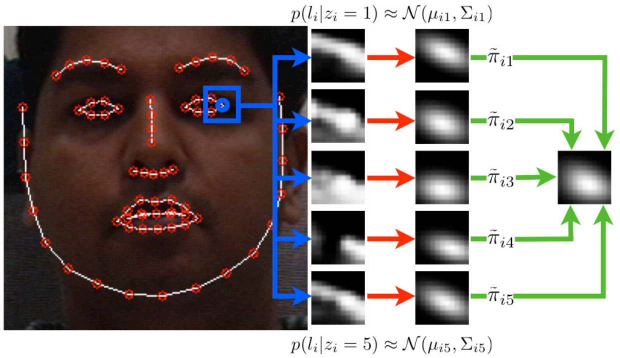
$p(l_i|z_i = 1) \approx \mathcal{N}(\mu_{i1}, \Sigma_{i1})$
$\tilde{\pi}_{i1}$
$\tilde{\pi}_{i2}$
$\tilde{\pi}_{i3}$
$\tilde{\pi}_{i4}$
$\tilde{\pi}_{i5}$
$p(l_i|z_i = 5) \approx \mathcal{N}(\mu_{i5}, \Sigma_{i5})$

Figure 1. Approximating the left eye corner landmark likelihood with a Gaussian mixture model. The true responses for each local expert, $\{p(l_i = \text{aligned}|z_i, \mathbf{x}_i)\}_{z_i=1}^{K_i}$, are first approximated by a Gaussian, then combined using the approximate mixing weights: $p(l_i = \text{aligned}|\mathbf{x}_i) \approx \sum_{k=1}^{K_i} \tilde{\pi}_{ik}\, \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$.

$$p(l_i = \text{aligned}|z_i = k, \mathbf{x}_i) \approx \mathcal{N}(\mathbf{x}_i\; ;\; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \tag{12}$$

where the mean and covariances are set to their maximum likelihood estimate over the search region:

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\mathbf{y}\in\boldsymbol{\Psi}_{\mathbf{x}_i}} p_{\mathbf{y}}} \sum_{\mathbf{y}\in\boldsymbol{\Psi}_{\mathbf{x}_i}} p_{\mathbf{y}}\, \mathbf{y} \tag{13}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{\mathbf{y}\in\boldsymbol{\Psi}_{\mathbf{x}_i}} p_{\mathbf{y}}} \sum_{\mathbf{y}\in\boldsymbol{\Psi}_{\mathbf{x}_i}} p_{\mathbf{y}} (\mathbf{y} - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T, \tag{14}$$

where $p_{\mathbf{y}} = p(l_i = \text{aligned}|z_i = k, \mathbf{y})$ and $\boldsymbol{\Psi}_{\mathbf{x}_i}$ denotes the rectangular search region centered at $\mathbf{x}_i$.

Rather than applying the Gaussian approximation to the response map of each local expert separately, it is also possible to apply the approximation directly to the combined response maps $\{p(l_i = \text{aligned}|\mathbf{x}_i)\}_{i=1}^n$. However, such an approximation may over-smooth the response maps, limiting the fidelity of the resulting fit. In contrast, the approximation made here loosely preserves the modalities of the true response map originating from the various local experts. Although it is possible to directly fit a Gaussian mixture model to the true response map, this process is computationally expensive and not provably optimal (i.e. dependent on initialization).

Mixing Weight Approximation: The data dependent mixing weights, $\{\pi_{ik}\}_{k=1}^{K_i}$, also inject nonlinearities into Equa-

[3] Throughout the rest of this paper, dependence on the image I will be dropped for succinctness, but implicitly assumed in all function.

tion (11). For this, we assume that both aligned and misaligned patches for a given image acquire the same local expert membership through the gating function defined in Equation (8). Such an assumption is reasonable since the experts primarily cluster the data based on patch appearance variations between instances of the visual object. A simple approximation, therefore, is the average membership likelihood:

$$\pi_{ik}(\mathbf{x}) \approx \tilde{\pi}_{ik} = \frac{\sum_{\mathbf{y}\in\mathbf{\Psi}_{\mathbf{x}_i}} \pi_{ik}(\mathbf{y})}{\sum_{j=1}^{K_i} \sum_{\mathbf{y}\in\mathbf{\Psi}_{\mathbf{x}_i}} \pi_{ij}(\mathbf{y})}; \forall \mathbf{x} \in \mathbf{\Psi}_{\mathbf{x}_i}. \quad (15)$$

This approximation, along with the Gaussian estimated landmark likelihood described previously, constitutes replacing $p(l_i = \text{aligned}|\mathbf{x}_i)$ in Equation (9) with a Gaussian mixture model (GMM) with fixed mixing weights. An illustration of this approximation is shown in Figure 1.

It should be noted that this approximation is only required when the prior over mixture membership is data driven, such as for the multinomial log-linear model in Equation (8). When the mixing weights are independent of the data, as in the case of typical additive classifier ensembles, then Equation (15) is no longer an approximation.

Shape Model Approximation: With the GMM approximation described above, the so called Q-function in Equation (11) simplifies to:

$$Q(\mathbf{p}) \propto \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}\|\mathbf{x}_i - \boldsymbol{\mu}_{ik}\|^2_{\mathbf{\Sigma}_{ik}^{-1}} + \text{const}, \quad (16)$$

where $w_{ik} = p(z_i = k|l_i = \text{aligned}, \mathbf{x}_i)$ given in Equation (10). Although this form is much simplified compared to that in Equation (11), it is still nonlinear due to the way shape is parameterized in Equations (1) and (2). For this, we apply a first order Taylor expansion of the shape model around the current estimate of its parameters:

$$\mathbf{x}_i = \mathbf{x}_i^c + \mathbf{J}_i \Delta\mathbf{p}, \quad (17)$$

where $\mathbf{x}_i^c$ are the 2D-coordinates of the i^{th} landmark of the PDM under its current parameter estimate, $\mathbf{J}_i$ is the Jacobian of that landmark's coordinates, and $\Delta\mathbf{p}$ is the sought parameter update, which is to be applied additively to the current estimate: $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$[4]. With this approximation, the solution for the M-step of the EM algorithm takes the form:

$$\Delta\mathbf{p} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\mathbf{J}_i\right)^{-1} \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}\mathbf{J}_i^T\mathbf{\Sigma}_{ik}^{-1}\mathbf{d}_{ik}, \quad (18)$$

where $\mathbf{d}_{ik} = \boldsymbol{\mu}_{ik} - \mathbf{x}_i^c$. The complete CLM fitting algorithm with a mixture of local experts is outlined in Agorithm 1.

Algorithm 1 CLM Fitting with a Mixture of Local Experts

```
Require: $I$ and $\mathbf{p}$.
while not_converged($\mathbf{p}$) do
    Compute $\{\{\pi_{ik}\}_{k=1}^{K_i}\}_{i=1}^{n}$ {Eqn. (8)}
    Compute $\{\{p(l_i|z_i = j, \mathbf{x}_i)\}_{j=1}^{K_i}\}_{i=1}^{n}$ {Eqn. (6)}
    Compute Gaussian approximation {Eqn.(12)}
    Compute mixture weight approximation {Eqn.(15)}
    Linearize Shape Model {Eqn. (17)}
    Initialize parameter updates: $\Delta\mathbf{p} \leftarrow \mathbf{0}$
    while not_converged($\Delta\mathbf{p}$) do
        E-step {Eqn. (10)}
        M-step {Eqn. (18)}
    end while
    Update parameters: $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$
end while
return $\mathbf{p}$
```

4. Experiments

The performance of CLM with a mixture of local experts (CLMix) was evaluated using the CMU Pose, Illumination and Expression Database (MultiPie) [9]. A collection of 2457 images of 339 subjects were hand labeled with 68-points that were used as ground truth. The collection contains significant variations in identity, facial expression and pose. The 3D shape model was learned using structure from motion [23], retaining 15 modes of non-rigid shape variation. The images were partitioned into four parts for use in a 4-fold cross validation procedure, where three parts were used for training and the remainder for testing in each of the four trials. During testing, the PDM was randomly perturbed from its optimal configuration in each training image and CLM fitting performed until convergence, as measured through the change in landmark locations between iterations. Fitting performance was measured as the root-mean-squared (RMS) distance between the converged shape and the ground truth.

CLMix was compared against two other methods, namely the active shape model (ASM) [4], which acts as a baseline for deformable model fitting, and CQF [25], which is equivalent to CLMix with one mixture component for each landmark. In all methods the linear SVM was used for the local experts, where the training data consisted of (11×11)-patches. Positive data was cropped from the image at the ground-truth coordinates and negative data at a distance $(5 \leq \delta \leq 20)$-pixels from it. The mixture of local experts were trained using the EM-based method described in [7], initialized using K-means on the positive data patch appearance. Models were trained for $K_i = \{1, \ldots, 5\}$, where $\{K_i = K_j; \forall\ i, j \in [1, \ldots, n]\}$. During fitting, the exhaustive local search for all methods was performed within a (11×11)-window. As such, all methods share the

[4] To apply the additive parameter update to the weak perspective model in Equation (2), we utilize the small angle rotation matrix approximation in the Taylor expansion. Details on this can be found in [15].

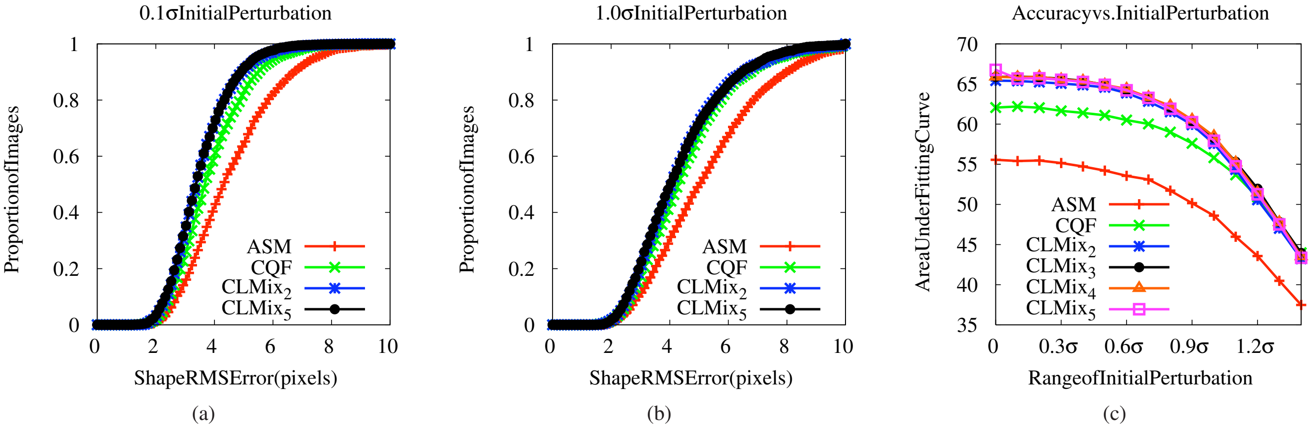

Figure 2. Fitting performance using ASM, CQF and CLMix. Subscripts for CLMix in the legend denote the number of mixture components. **(a):** Fitting curves with an initial perturbation range of 0.1σ. **(b):** Fitting curves with an initial perturbation range of 1.0σ. **(c):** Plot of area under the fitting curves against the range of initial perturbation.

same parameterization and local expert type, differing only in the number of experts used and their optimization strategy[5].

The combined results of these experiments are presented in Figure 2. The fitting curves in Figures 2(a) and 2(b) show the proportion of images at which various levels of maximum error was exhibited at convergence. Figure 2(c) plots the area under the fitting curves, which is a measure of overall convergence accuracy, against the range of initial perturbations as measured in fractions of σ, the standard deviation of variation exhibited in the database[6]. The fitting curves in Figures 2(a) and 2(b) are for initial perturbation ranges of 0.1σ and 1.0σ, respectively. Some examples of fitting results are shown in Figure 3

The results show a significant improvement in performance afforded by CLMix over CQF and ASM. This is most pronounced when the range of initial perturbation is small. When the model is grossly misplaced (i.e. an initial perturbation range greater than σ), CLMix performs similarly to CQF but much better than ASM. This suggests that CLMix is more sensitive towards local minima than CQF. However, CLM lacks the capacity to account for the large amount of variability exhibited by faces in the database due to its use of only a single linear classifier as its local expert. This can be seen by its limited accuracy even when optimally initialized (i.e. zero initial perturbation range).

The results also show that performance improvement is marginal when using more than two mixture components. This suggests that two mixture components are sufficient to distinguish aligned from misaligned landmarks in this database. This further motivates the use of a mixture of linear classifiers as opposed to more general nonlinear classifiers, since good performance can be attained with as little as two mixture components. Examining Figure 3, one notices that the ASM and CQF fail to accurately detect landmarks on the outline of the face. Since patches extracted from these landmarks will include background pixels, two disjoint cases exist: when the background is lighter than the face, and *vice-versa*. A linear classifier can not accurately distinguish such cases, but a mixture of two linear classifiers can. This is a partial explanation of the performance improvement afforded by CLMix. When the database exhibits other sources of variability, changes in lighting conditions for example, more than two mixture components may be required to accurately distinguish aligned from misaligned landmark locations. In practice, the number of mixture components used for each landmark should reflect the data for that landmark in order to maximize the utility of such a framework.

Finally, despite the significant improvement in performance, the computational complexity of CLMix scales only linearly with the number of mixture components. In the experiments presented here, the average fitting time for C/C++ implementations of ASM, CQF, CLMix$_2$, CLMix$_3$, CLMix$_4$ and CLMix$_5$ on a 2.5GHz Intel Core 2 Duo were 101ms, 113ms, 178ms, 223ms, 285ms and 353ms, respectively. In addition, further computational savings can be attained when parallel processing is involved. The computation of the response maps for each local expert of each landmark constitutes the bulk of the computational load in

[5] ASMs typically use a generative Gaussian expert and a profile search region. The particular instance of ASM being compared here was chosen to highlight the utility of mixture of experts and the proposed optimization strategy.

[6] The various parameters of a PDM typically exhibit differing scales, resulting in different variances across the database. We use σ here as a generic indicator of parameter variance. For example, a perturbation range of 0.1σ denotes the ranges within 0.1 of the variance in scale, translation rotation and nonrigid parameters within the database, independently.

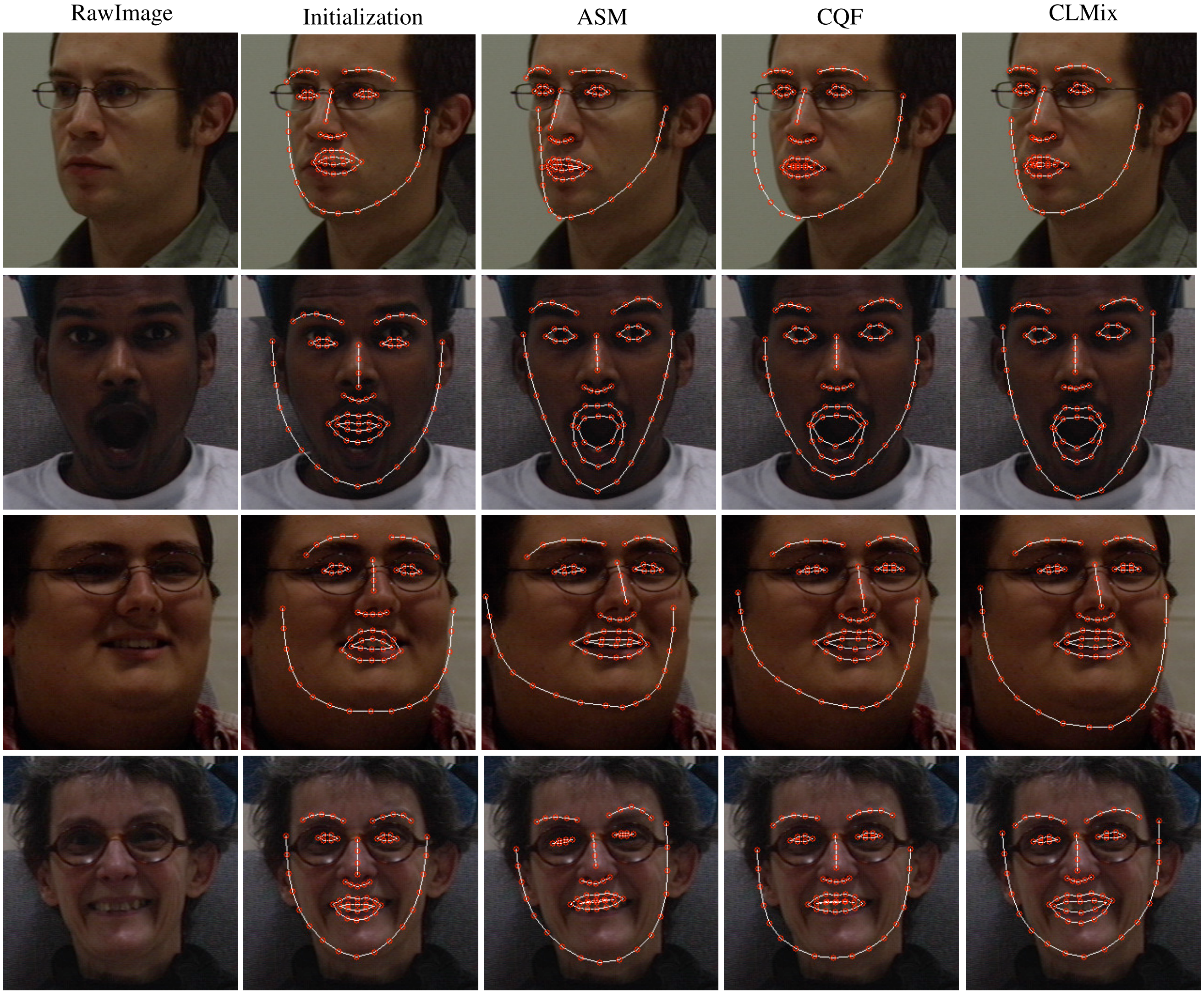

Figure 3. Examples of fitting results for ASM, CQF and CLMix$_2$.

CLMix. However, the computation of each response map is independent of all others, due to the conditional independence assumption made in Equation (9) regarding landmark detections. This facilitates a multithreaded implementation, computing all response maps in parallel rather than in series.

5. Conclusion

In this work, an approach for combining multiple local experts for deformable model fitting was presented. The approach makes use of a mixture of linear classifiers for each landmark of the model's shape. Optimization is performed using the EM algorithm, where some approximations to the true objective are made in order to achieve efficient fitting. Experiments were performed, comparing the proposed approach with two existing methods, where improvements in fitting fidelity was observed when using multiple local experts. Furthermore, the computational complexity of the approach was shown to scale only linearly with the number of mixture components used.

Further improvements to the proposed framework are also possible. In particular, the local experts can be trained specifically for use in the proposed approach. This would involve training local experts to generate distribution of landmark likelihoods that are better approximated by a Gaussian. For example, in [17, 27] an expert is learned that can generate approximately convex responses. Other improvements may include relaxing the simplifying assumptions made in §3.2, optimizing the true objective rather than an approximation thereof. This may be achieved by fitting on a hierarchy of smoothed estimates to avoid local minima and promote numerical stability.

References

[1] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D-faces. In *SIGGRAPH*, 1999.

[2] O. Chapelle. Training a Support Vector Machine in the Primal. *Neural Computation*, 9(5), 2007.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. In *ECCV*, pages 484–498, 1998.

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

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

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

[7] Z. Fu and A. Robles-Kelly. On Mixtures of Linear SVMs for Nonlinear Classification. *SSPR/SPR*, pages 489–499, 2008.

[8] R. Gross, I. Matthews, and S. Baker. Generic vs. Person Specific Active Appearance Models. *IVC*, 23:1080–1093, 2005.

[9] 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.

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

[11] M. I. Jordan and R. A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation 6*, page 181214, 1994.

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

[13] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting Algorithms as Gradient Descent. In *NIPS*, pages 512–518, 2000.

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

[15] I. Matthews, J. Xiao, and S. Baker. On the Dimensionality of Deformable Face Models. Technical report, Robotics Institute, 2006.

[16] B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. *PAMI*, 19(7):696–710, 1997.

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

[18] M. H. Nguyen, J. Perez, and F. De la Torre. Facial Feature Detection with Optimal Pixel Reduction SVMs. In *FG*, 2008.

[19] J. Peyras, A. Bartoli, H. Mercier, and P. Dalle. Segmented AAMs Improve Person-Independent Face Fitting. In *BMVC*, 2007.

[20] J. Saragih and R. Goecke. Iterative Error Bound Minimisation for AAM Alignment. In *ICPR*, volume 2, pages 1192–1195, 2006.

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

[22] K. Sjöstrand, M. B. Stegmann, and R. Larsen. Sparse Principal Component Analysis in Medical Shape Modeling. In *International Symposium on Medical Imaging*, number 6144, 2006.

[23] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid Structure-from-Motion: Estimating Shape and Motion with Hierarchical Priors. *PAMI*, 2008.

[24] B. v. Ginneken, M. B. Stegmann, and M. Loog. Segmentation of Anatomical Structures in Chest Radiographs using Supervised Methods: A Comparative Study on a Public Database. *Medical Image Analysis*, 10(1):19–40, 2006.

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

[26] O. Williams, A. Blake, and R. Cipolla. A Sparse Probabilistic Learning Algorithm for Real-time Tracking. In *ICCV*, volume 1, pages 353–360, 2003.

[27] H. Wu, X. Liu, and G. Doretto. Face Alignment via Boosted Ranking Models. In *CVPR*, 2008.

[28] S. Zhou and D. Comaniciu. Shape Regression Machine. In *IPMI*, 2007.