Convexity and Bayesian Constrained Local Models

Ulrich Paquet
Imense Ltd
Cambridge, UK
ulrich@imense.com

Abstract

The accurate localization of facial features plays a fundamental role in any face recognition pipeline. Constrained local models (CLM) provide an effective approach to localization by coupling ensembles of local patch detectors for non-rigid object alignment. A recent improvement has been made by using generic convex quadratic fitting (CQF), which elegantly addresses the CLM warp update by enforcing convexity of the patch response surfaces. In this paper, CQF is generalized to a Bayesian inference problem, in which it appears as a particular maximum likelihood solution. The Bayesian viewpoint holds many advantages: for example, the task of feature localization can explicitly build on previous face detection stages, and multiple sets of patch responses can be seamlessly incorporated. A second contribution of the paper is an analytic solution to finding convex approximations to patch response surfaces, which removes CQF's reliance on a numeric optimizer. Improvements in feature localization performance are illustrated on the Labeled Faces in the Wild and BioID data sets.

1. Introduction

The task of parsing and recognizing faces in an unconstrained environment often assumes that processing happens in a *detection–alignment–recognition* pipeline, which separates the tasks of detecting faces, aligning them by locating key fiducial points, and finally basing any recognition task on that alignment. The recent release of the Labeled Faces in the Wild (LFW) data set emphasizes this pipeline [7], highlighting the dependence of each stage on its predecessor.

The alignment stage is a problem of combining shape and texture information from a training set with texture information from a target image to locate the fiducial point locations on the target face. A Bayesian framework the alignment stage is proposed here, and similar to LFW the assumption is made that faces are detectable by a Viola-Jones (VJ) face detection algorithm [17]—an assumption that can be explicitly incorporated into a prior alignment distribution.

Constrained Local Models (CLMs) are ideally suited to facial feature alignment and general non-rigid object registration, as they merge shape and texture information by coupling (or constraining) an ensemble of local patch or feature detectors at a global shape level [2, 3]. This has proved to outperform Active Appearance Models (AAMs) [1] as it is more robust to occlusion and changes in appearance and no texture warps are required.

A prime example of a CLM was given by Wang *et al*. [19], which, through generic convex quadratic fitting (CQF), turns the global CLM warp update into a convex problem. By finding convex approximations to the local patch response surfaces of feature alignment classifiers, this circumvented the need for computationally expensive optimizers (except maybe for fitting the convex surfaces). A pleasing consequence was that a specific form of the Lucas-Kanade [12] gradient descent image alignment algorithm can be viewed as a generic CQF. It was also shown to be superior to exhaustive local search (ELS) [18], which constrains local patch response maxima to be consistent with the shape prior.

This paper argues that for further progress to be made in tasks similar to LFW, the generic CQF method of Wang *et al*. [19] can be generalized and improved even further:

1. The convex patch responses can be folded into a Bayesian inference problem, where the posterior distribution of the global warp needs to be inferred. In this Bayesian constrained local model (BCLM) formulation both generic CQF and ELS appear as maximum likelihood solutions, and more than one feature classifier for each feature can easily be included.

2. The Bayesian framework allows the alignment stage's shape prior to explicitly model our beliefs about the range and distribution of faces that will be received from a given face detector.

3. The generic CQF method finds convex approximations to patch response surfaces by solving a quadratically

constrained quadratic program, or by simplifying the problem to only include axis-aligned functions. A simple and effective analytic method for finding general convex approximations is proposed here.

The alignment of fiducial data points are illustrated on faces from the LFW and BioID [8] data sets. To be truly general, all patch classifiers were trained on a different set of random Internet images, where, similar to LFW, faces are detectable by a given VJ algorithm. The local patch classifiers used here are linear and fast, but therefore sacrifice a degree of accuracy. In cases where noisy classifiers imply that maximum likelihood estimates will not be sufficient, the use of a Bayesian approach becomes evident and improves alignment errors.

Numerous efforts, both similar and complimentary to the approach presented here, have been made for facial feature detection. Methods on which this work builds include Cristinacce *et al.*'s pairwise reinforcement of feature responses [4], Cristinacce and Cootes' CLMs [2, 3], and Wang *et al.*'s ELS and generic CQF approaches [18, 19]. A "shape-constrained" Markov random field is used by Liang *et al.* to model a face [10]. Another Bayesian generative model is Gu and Kanade's treatment of multiple candidate feature alignment positions as unobserved latent variables, through which the shape-and-pose posterior mode can be found with an expectation maximization algorithm [5]. Liu aligns images by iteratively maximizing the score of a classifier—a boosted appearance model—that distinguishes between correct and incorrect alignments, and updating a low-rank shape parameter [11]; this approach is extended with a boosted ranking model by Wu *et al.* [20].

The rest of the paper lays out BCLMs in section 2, and shows a simple analytic method for fitting local convex energy functions in section 3. Section 4 presents the patch classifiers that are used in this work, while a comparison of BCLMs against generic CQFs and AAMs is given in section 5.

2. Bayesian Constrained Local Model

The alignment stage combines shape and texture information to locate fiducial points on a face. In the BCLM presented here, shape information appears in a prior distribution, which models the range of faces that a given face detector can detect, while texture information is summarized in convex functions in a log-likelihood. An iterative algorithm is presented for feature alignment, and CQF and ELS are shown to be maximum likelihood solutions in this Bayesian set-up.

Let $\mathbf{x}$ be a vector indexing feature locations across an object, for example a face. If $\mathbf{x}_i = (x_i, y_i)$ denotes the centre of feature i, then $\mathbf{x} = (x_1, y_1, \ldots, x_I, y_I)$; with $D = 2I$ we have $\mathbf{x} \in \mathbb{R}^D$.

For a shape prior, we assume that a point distribution model transforms a lower-dimensional ($K \leq D$) latent variable $\mathbf{z} \in \mathbb{R}^K$, with prior $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$, to $\mathbf{x}$ with

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z} \,. \tag{1}$$

This view of the usual point distribution model is as a generative model where the only uncertainty is in $\mathbf{z}$, i.e. a noise-free formulation of Bayesian PCA [15].

In this instance $\mathbf{x}$ will be a set of feature locations relative to a window given by a VJ face spotter. The face detector window is scaled to a standard size of 110-by-110 pixels, implying an inter-ocular distance of roughly 50 pixels. This operation standardizes over global scale and translation, whereas procrustes analysis is normally applied. The model is determined from a set of marked-up faces containing true fiducial points $\{\mathbf{x}^{(n)}\}_{n=1}^N$. Although posterior densities over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ can plausibly be incorporated, a point mass estimate $\boldsymbol{\mu} = \frac{1}{N}\sum_n \mathbf{x}^{(n)}$ is used in this paper. An eigenvalue decomposition $\mathbf{U}\mathbf{D}\mathbf{U}^\top = \frac{1}{N}\sum_n (\mathbf{x}^{(n)} - \boldsymbol{\mu})(\mathbf{x}^{(n)} - \boldsymbol{\mu})^\top$ gives $\boldsymbol{\Lambda} = \mathbf{U}_K \mathbf{D}_K^{1/2}$, where $\mathbf{U}_K$ is a submatrix of the K eigenvectors is $\mathbf{U}$ corresponding to the largest eigenvalues in $\mathbf{D}$.

The texture model for aligning feature i is represented by a convex energy function centered at $\mathbf{c}_i$,

$$\mathcal{E}_i(\mathbf{x}_i) = \frac{1}{2}(\mathbf{x}_i - \mathbf{c}_i)^\top \mathbf{A}_i (\mathbf{x}_i - \mathbf{c}_i) \,, \tag{2}$$

with $\mathbf{A}_i$ being positive definite. For now our only assumption is that $\mathbf{c}_i$ and $\mathbf{A}_i$ are chosen such that $\mathcal{E}_i(\mathbf{x}_i)$ is small if pixel $\mathbf{x}_i$ lies close to the true location of fiducial point i, and large otherwise. This is the same setup as the generic CQF [19].

Section 3 presents an analytic approach for determining $\mathcal{E}_i(\mathbf{x}_i)$—or approximating $\mathbf{A}_i$ and $\mathbf{c}_i$—from outputs of feature i's patch alignment classifier. An alignment classifier uses a local *patch* of pixels around a certain point to determine the probability of it aligning to some fiducial point; this is discussed in section 4. Figure 1 illustrates examples of $\mathcal{E}_i(\mathbf{x}_i)$, along with the local patch classifier outputs and a final alignment.

2.1. An explicit Bayesian formulation

Using the linear relation in (1), each local convex energy function $\mathcal{E}_i(\mathbf{x}_i)$ can be treated as a negative log-likelihood for $\mathbf{z}$, *given* some knowledge of $\mathbf{c}_i$ and $\mathbf{A}_i$. This gives I different likelihood functions for the warp $\mathbf{z}$, for which we already have a prior.

Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$ correspond to the appropriate rows of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, and from (1) assume a deterministic relationship without additive noise: $\mathbf{x}_i = \boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i \mathbf{z}$. Let

$$\Delta\mathbf{m}_i = \mathbf{c}_i - \boldsymbol{\mu}_i \tag{3}$$

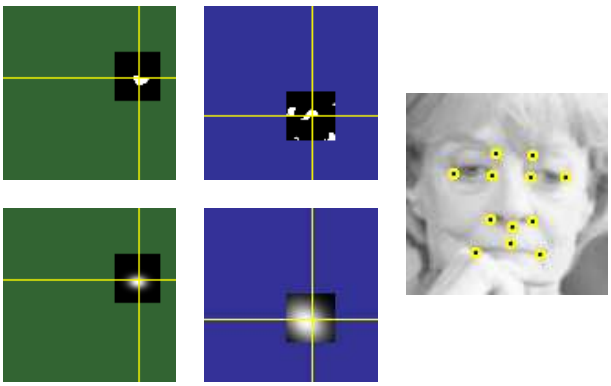

Figure 1. Alignment classifiers and convex energy functions: At the top patch classifier outputs (15, 19) are shown for the right eye and nose corners, for *each* pixel in a window $\mathcal{W}(\mathbf{x}_i^*; L)$ of width L pixels centered on some $\mathbf{x}_i^*$. Their convex approximations $\mathcal{E}_i(\mathbf{x}_i)$ in (2) are shown at the bottom. Uncertainty in patch classifications (see the middle illustration that matches the right nose corner) can be treated by iteratively centering $\mathcal{W}(\mathbf{x}_i^*; L)$ on a current alignment and reducing L in algorithm 1. The final alignment is shown on the right face, and located at the crossed lines in the other figures.

be the offset of the local energy function from the mean feature location. This observed quantity is dependent on $\mathbf{z}$ in a generative model: With

$$\mathcal{E}_i(\mathbf{x}_i) = \mathcal{E}_i(\boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i\mathbf{z}) \tag{4}$$

$$= \frac{1}{2}(\Delta\mathbf{m}_i - \boldsymbol{\Lambda}_i\mathbf{z})^\top \mathbf{A}_i(\Delta\mathbf{m}_i - \boldsymbol{\Lambda}_i\mathbf{z}) \tag{5}$$

being a negative log likelihood for $\mathbf{z}$, we have $p(\Delta\mathbf{m}_i|\mathbf{z}) = \frac{1}{Z}\exp(-\mathcal{E}_i(\mathbf{x}_i))$, and therefore the likelihood for each local alignment is

$$p(\Delta\mathbf{m}_i|\mathbf{z}) = \mathcal{N}(\Delta\mathbf{m}_i; \boldsymbol{\Lambda}_i\mathbf{z}, \mathbf{A}_i^{-1}) \,. \tag{6}$$

As $\Delta\mathbf{m}_i \in \mathbb{R}^2$, let $\Delta\mathbf{m} \in \mathbb{R}^{2I}$ be the vector concatenation of all I patch alignment offsets $\Delta\mathbf{x}_i$, and let $\mathbf{A} = \mathrm{diag}(\{\mathbf{A}_i\}) \in \mathbb{R}^{2I\times 2I}$ have submatrices $\mathbf{A}_i$ along its diagonal.

Bayes' theorem provides the posterior distribution of $\mathbf{z}$,

$$p(\mathbf{z}|\Delta\mathbf{m}) = \frac{p(\Delta\mathbf{m}|\mathbf{z})p(\mathbf{z})}{p(\Delta\mathbf{m})} = \frac{\prod_i p(\Delta\mathbf{m}_i|\mathbf{z})p(\mathbf{z})}{p(\Delta\mathbf{m})} \,, \tag{7}$$

which is Gaussian $\mathcal{N}(\mathbf{z}; \boldsymbol{\nu}, \mathbf{S})$ with covariance and mean

$$\mathbf{S} = (\boldsymbol{\Lambda}^\top\mathbf{A}\boldsymbol{\Lambda} + \mathbf{I})^{-1} \tag{8}$$

$$\boldsymbol{\nu} = \mathbf{S}\boldsymbol{\Lambda}^\top\mathbf{A}\Delta\mathbf{m} \,. \tag{9}$$

The mean is also the maximum a posteriori (MAP) estimate.

The global model for $\mathbf{x}$ constrains the posterior to be consistent with a low-rank representation of typical alignments. This constraint can be relaxed by increasing K to D, but with a range of appearance variations and occlusions the patch classifiers will not be exact, and a workable compromise has to be made. In this work the warp $\mathbf{z}$ has $K = 5$ degrees of freedom.

Algorithm 1 Bayesian Constrained Local Model

initialize (Preprocessed) face image $\mathcal{I}$ from VJ detector; patch experts $\{\mathbf{w}_i\}_{i=1}^I$ (section 4); $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$; initial window size L; minimum window size $L_{\min}$; initial warp $\boldsymbol{\nu} = \mathbf{0}$.
repeat
  for $i = 1$ to I **do**
    Find $\mathbf{x}_i^* \leftarrow \boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i\boldsymbol{\nu}$ and determine $\mathcal{W}(\mathbf{x}_i^*; L)$.
    Determine $p_i(\mathbf{x}_i)$ for each possible alignment centre $\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)$ using (19).
    Find $\mathbf{c}_i$ and $\mathbf{A}_i$ in (16).
  end for
  $\Delta\mathbf{m} \leftarrow \mathbf{c} - \boldsymbol{\mu}$ and $\mathbf{A} \leftarrow \mathrm{diag}(\{\mathbf{A}_i\})$
  $\boldsymbol{\nu} \leftarrow (\boldsymbol{\Lambda}^\top\mathbf{A}\boldsymbol{\Lambda} + \mathbf{I})^{-1}\boldsymbol{\Lambda}^\top\mathbf{A}\Delta\mathbf{m}$
  $L \leftarrow L - 2$
until $L < L_{\min}$
return $\mathbf{x}^* \leftarrow \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\nu}$

2.2. A practical algorithm

A practical algorithm has to address cases, as illustrated in figure 1, where the patch classifiers can give false or noisy responses. One solution is to iteratively use the posterior mode to shift and decrease the window size over which $\mathcal{E}(\mathbf{x}_i)$ is estimated, as false responses can skew $\mathcal{E}$. Equation (7) implies that "imprecise" energy functions are always correlated with (and aided by) more precise patch responses. Similar to an annealing schedule, the energy functions therefore become more sharply peaked as false responses are disregarded.

Let $\mathcal{W}(\mathbf{x}_i^*; L)$ be a square window of width L pixels centered on $\mathbf{x}_i^*$, so that all possible alignments around the current estimate $\mathbf{x}_i^*$ are considered when estimating the "observation" parameters $\mathbf{c}_i$ and $\mathbf{A}_i$ of $\mathcal{E}(\mathbf{x}_i)$. These parameters are determined through a range of I classifiers, each of which gives the probability $p_i(\mathbf{x}_i)$ that some $\mathbf{x}_i$ aligns with an unknown fiducial point i. The probability $p_i(\mathbf{x}_i)$ typically takes a local *patch* of pixels around $\mathbf{x}_i$ into account, and this process is explained in sections 3 and 4.

In algorithm 1 this method is referred to as a Bayesian constrained local model (BCLM).

2.3. Maximum likelihood solutions

The generic CQF and ELS methods of Wang *et al.* [18, 19] iteratively fit versions of the maximum likelihood (ML) estimate of $p(\Delta\mathbf{m}|\mathbf{z})$ in (7). As ML solutions are equivalent to MAP solutions with a non-informative prior,

this is useful when we don't have any prior belief about the (low-rank) distribution of features on a face. In practice this argument is less strong, as alignment methods often fail when initialized too far away from the "truth". Prior knowledge about typical faces that pass through the detection stage may also be available.

Generic CQF The energy function in (2) can be rearranged so that the likelihood function appears in terms of *warp updates* $\Delta\mathbf{z} = \mathbf{z} - \boldsymbol{\nu}^*$, with $\boldsymbol{\nu}^*$ coming from a preceding iteration in algorithm 1. Using $\Delta\mathbf{x}_i = \mathbf{x}_i - \mathbf{x}_i^*$ we have

$$\mathcal{E}_i(\mathbf{x}_i) = \frac{1}{2}\Big(\Delta\mathbf{x}_i - (\mathbf{c}_i - \mathbf{x}_i^*)\Big)^\top \mathbf{A}_i \Big(\Delta\mathbf{x}_i - (\mathbf{c}_i - \mathbf{x}_i^*)\Big) \,. \tag{10}$$

The relation $\Delta\mathbf{x}_i = \boldsymbol{\Lambda}_i\Delta\mathbf{z}$ specifies how much the fiducial points should be adjusted given a warp update $\Delta\mathbf{z}$, as $\mathbf{x}^* = \boldsymbol{\Lambda}\boldsymbol{\nu}^* + \boldsymbol{\mu}$.

If $\Delta\mathbf{m}_i' = \mathbf{c}_i - \mathbf{x}_i^*$ is defined as an observation of the offset of the local energy function from the current fiducial point estimate, and (10) is treated as a negative log likelihood, then $p(\Delta\mathbf{m}_i'|\Delta\mathbf{z}) = \mathcal{N}(\Delta\mathbf{m}_i' \,;\, \boldsymbol{\Lambda}_i\Delta\mathbf{z}, \mathbf{A}_i^{-1})$, and the ML solution to $\max_{\Delta\mathbf{z}} \prod_i p(\Delta\mathbf{m}_i'|\Delta\mathbf{z})$,

$$\widehat{\Delta\mathbf{z}} = (\boldsymbol{\Lambda}^\top\mathbf{A}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^\top\mathbf{A}\Delta\mathbf{m}' \,, \tag{11}$$

is equivalent to the warp update of equation (9) in [19].

ELS Each step in ELS searches locally, i.e. in a window $\mathcal{W}(\mathbf{x}_i^*; L)$, for the highest alignment response, defined as

$$\mathbf{c}_i = \underset{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)}{\arg\max}\; p(\mathbf{x}_i) \,. \tag{12}$$

By defining the certainty in the approximation to be e.g. proportional to $\mathbf{A}_i = \mathrm{diag}(p(\mathbf{c}_i), p(\mathbf{c}_i))$, and using it in (10), the ML solution in (11) is equivalent to the ELS *weighed least squares optimization* warp update.

As ELS is based on local maxima in the patch response surfaces, it is clear why CQF is superior when the patch classifiers are noisy, as it incorporates responses in a whole window.

2.4. Multiple sets of feature detectors

The Bayesian framework allows different patch alignment classifiers to be seamlessly incorporated into the model. Generalizing further, let $r = 1, \ldots, R$ index sets of patch alignment classifiers $\{\mathcal{M}_{ri}\}_{i=1}^I$, with each giving rise to a convex error function with parameters $\mathbf{c}_i^{(r)}$ and $\mathbf{A}_i^{(r)}$, where features $i = 1, \ldots, I$ are to be aligned.

With multiple observations $\Delta\mathbf{m}_i^{(r)} = \mathbf{c}_i^{(r)} - \boldsymbol{\mu}_i$, the posterior distribution of $\mathbf{z}$ is again Gaussian with covariance and mean

$$\mathbf{S} = \left[\boldsymbol{\Lambda}^\top\Big(\sum_r \mathbf{A}^{(r)}\Big)\boldsymbol{\Lambda} + \mathbf{I}\right]^{-1} \tag{13}$$

$$\boldsymbol{\nu} = \mathbf{S}\boldsymbol{\Lambda}^\top \sum_r \mathbf{A}^{(r)}\Delta\mathbf{m}^{(r)} \,. \tag{14}$$

This approach is motivated by the fact that classifiers—especially when designed to be fast—may give a few false responses. Section 5 illustrates how a combination of $R = 2$ sets of classifiers can improve on the alignments resulting from when either set is used individually.

3. Local convex energy functions

In this section a simple analytic method is presented for determining the convex energy functions. Wang *et al*. postulated using a quadratically constrained quadratic program that is costly to solve directly [19], but rather simplified the problem to a quadratic program by constraining the form of $\mathbf{A}_i$ to be axis-aligned. This restriction is not ideal when the feature responses are e.g. diagonal, and the $\mathcal{E}$ is required to model these possibilities.

Let $\mathbf{x}_i = (x_i, y_i)$ be the centre of a $P \times P$ patch of pixels. Define the patch on the (possibly preprocessed) image $\mathcal{I}$ as $\mathcal{I}(\mathbf{x}_i)$: it is the vector concatenation of the P^2 patch of pixels in $\mathcal{I}$, centered at $\mathbf{x}_i$. Finally define the binary variable a_i such that

$$p_i(\mathbf{x}_i) = p(a_i = 1 \,|\, \mathcal{I}(\mathbf{x}_i), \mathcal{M}_i) \tag{15}$$

is the probability that $\mathbf{x}_i$ aligns with (or is centered at) the i^{th} fiducial point, *given* its surrounding patch $\mathcal{I}(\mathbf{x}_i)$ and a patch classification model $\mathcal{M}_i$.

Parameters $\mathbf{c}_i$ and $\mathbf{A}_i$ in (2) can be found by minimizing

$$\underset{\mathbf{A}_i, \mathbf{c}_i}{\arg\min} \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)\, \mathcal{E}_i(\mathbf{x}_i) \,, \tag{16}$$

which equivalently fits a Gaussian density to weighted data in $\mathcal{W}(\mathbf{x}_i^*; L)$. The sufficient requirement for $\mathbf{A}_i$ being positive definite is that $p_i(\mathbf{x}_i) > 0$, so that any positive function which gives proportionally higher weight to good alignments can realistically be used. With $s = \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)$ the minimum is straight-forward:

$$\mathbf{c}_i = \frac{1}{s} \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)\, \mathbf{x}_i \tag{17}$$

$$\mathbf{A}_i^{-1} = \frac{1}{s} \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_i)(\mathbf{x}_i - \mathbf{c}_i)^\top \,. \tag{18}$$

Figure 1 illustrates $p_i(\mathbf{x}_i)$ and the resulting convex approximations.

4. The patch classifiers

The patch classifiers can be constructed around any probabilistic method. Fully probabilistic kernel-based classifiers [14] are ideal, but the evaluation of even a moderate amount of kernels to classify a P^2 patch for *each* pixel in $\mathcal{W}(\mathbf{x}_i^*; L)$ poses a time vs. accuracy trade-off. Linear logistic regression (similar to the linear support vector machine in [19]) is only based on a dot product, and used here because of its computational speed.

For each feature i a data set $\mathcal{D}_i = \{\mathcal{I}(\mathbf{x}_i^{(m)}), a_i^{(m)}\}_{m=1}^M$ was created, where $a_i^{(m)} = 1$ if the patch is aligned and -1 otherwise. The data set was purposefully built around faces from publicly available Internet images, that were detected by a VJ detector. This formed the basis of our prior assumption of faces entering the alignment stage of the pipeline, and parallels the LFW assumptions [7]. A total of $N = 1750$ faces were marked up with 33 feature points each (for use in an AAM in section 5), of which $I = 12$ were used for patch classifiers. For *each* feature i a total of 7000 aligned and 17500 misaligned patches of size $P = 19$ were used to give $M = 24500$ training examples. The aligned patches were sampled from the exact alignment, and from the alignment offset with a random 1-pixel shift. Half the misaligned patches were sampled from within a 3-19 pixel window around the exact alignment, while the second half were drawn randomly from the rest of the image. Each $\mathcal{I}(\mathbf{x}_i^{(m)})$ is normalized in a preprocessing step by subtracting the mean pixel value of all the patches in $\mathcal{D}_i$, and dividing by the variance of all the pixels in $\mathcal{D}_i$. The same preprocessing is done when aligning patch i on a new face. Finally $\mathcal{I}(\mathbf{x}_i)$ is concatenated with an additional offset value clamped at 1, so that $\mathcal{I}(\mathbf{x}_i) \in \mathbb{R}^{P^2+1}$.

Let the probability of a correct (or incorrect) alignment be

$$p(a_i \,|\, \mathcal{I}(\mathbf{x}_i), \mathbf{w}_i) = \sigma\left(a_i \mathbf{w}_i^\top \mathcal{I}(\mathbf{x}_i)\right) \ , \qquad (19)$$

where $\mathbf{w}_i \in \mathbb{R}^{P^2+1}$ defines a patch classifier, and $\sigma(z) = 1/(1+e^{-z})$. The parameters $\mathbf{w}_i$ are found by minimizing the error function (a negative log likelihood plus negative log prior) for each patch i,

$$\mathcal{L}(\mathbf{w}_i) = -\sum_m \log\left[(1-\epsilon)\,\sigma\left(a_i^{(m)} \mathbf{w}_i^\top \mathcal{I}(\mathbf{x}_i^{(m)})\right) + \epsilon\right] + \frac{\alpha}{2}\mathbf{w}_i^\top \mathbf{w}_i \ , \qquad (20)$$

with a conjugate gradient-based algorithm. A small value $\epsilon > 0$ aids in numeric stability when optimizing for $\mathbf{w}_i$, and can be interpreted as a probability of *label noise*: when a subject wears dark glasses, for example, we account for the possibility of the patch not being a good example of an alignment. The value of α is chosen through cross-validation.

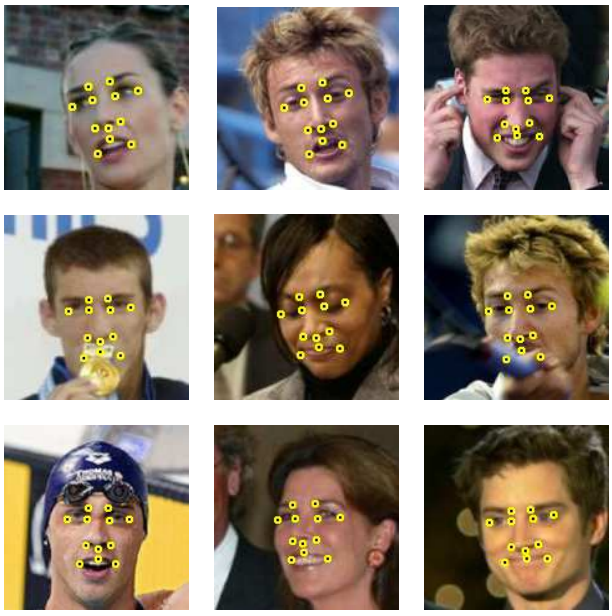

Figure 2. A selection of BCLM alignments from the LFW data set, using $p_{i,\mathrm{prod}}(\mathbf{x}_i)$ as patch classifier.

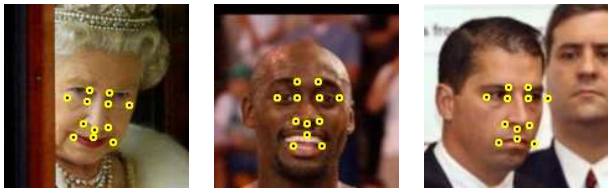

Figure 3. A few example errors from the LFW data set: a weak eyebrow and eyebrow-like eye response misaligns the eyes; a false left mouth corner detection; a lack of a strong right eye corner detection enforces a best guess. $p_{i,\mathrm{prod}}(\mathbf{x}_i)$ was used as patch classifier.

On data with a similar distribution to $\mathcal{D}_i$ the classifiers gave error rates of 8 to 13 percent over the various features. False positives, e.g. a "centre of upper lip" patch, which will give a positive response along the upper lip and not merely for one or two pixels, can be judiciously treated, as was proposed in section 2.2.

5. Results

The accuracy of algorithm 1's BCLM was tested on faces from the LFW [7] and BioID [8] data sets. For each test set 200 faces were marked up with 33 (for an AAM tested against here) true feature point locations, of which 12 were used for a generic CQF and BCLM. CQF was already shown to be superior to ELS [19], and is not compared against here. The training set comprised of 1750 similarly marked up faces from publicly available Internet images, and contains faces similar to the LFW data set.

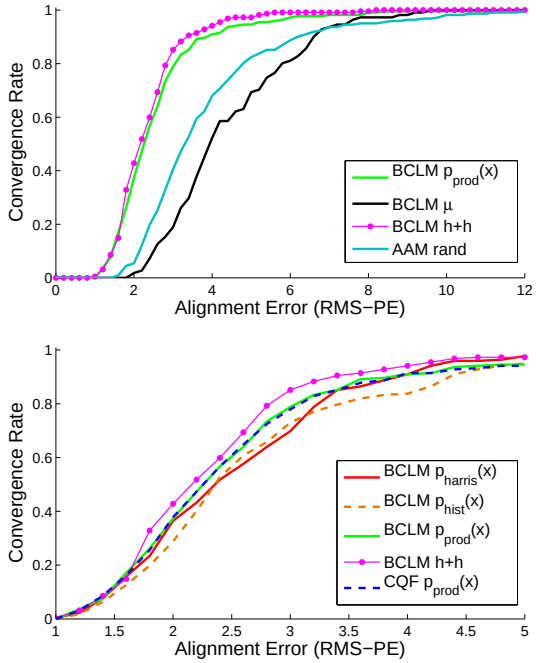

Figure 4. The alignment error for different methods on the LFW data set. The AAM shape parameter was randomly initialized according to the prior described in section 2. "h+h" refers to a combination of $p_{i,\mathrm{harris}}(\mathbf{x}_i)$ and $p_{i,\mathrm{hist}}(\mathbf{x}_i)$ in the posterior mode in (14), and improves an AAM, a CLM, and their separate use in BCLMs.

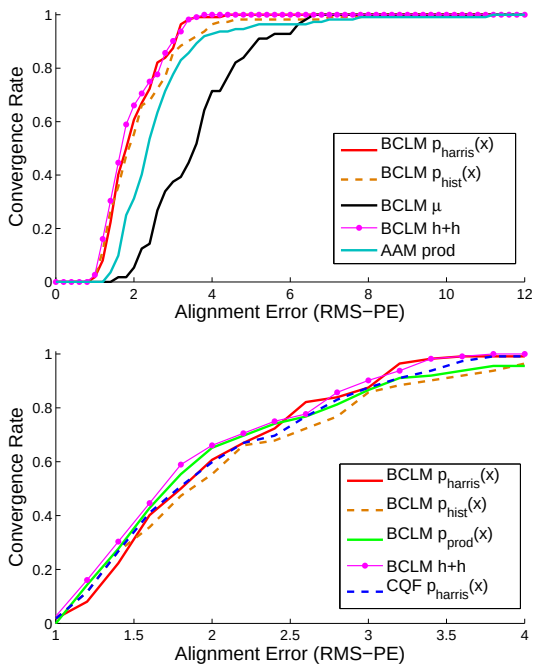

Figure 5. The alignment error for different methods on the BioID data set. The AAM shape parameter was randomly initialized using the BCLM feature point locations with $p_{i,\mathrm{prod}}(\mathbf{x}_i)$.

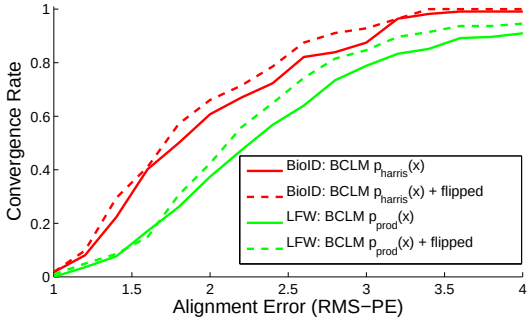

Figure 6. Improving results by taking the average of an alignment and an alignment of a horizontal mirror image of the same face.

Four different BCLM settings are explored. The face is preprocessed in two ways, so that $R = 2$ sets of linear classifiers are trained and used:

1. A histogram-normalized image gives $p_{i,\mathrm{hist}}(\mathbf{x}_i)$ for feature i's centre.

2. A Harris image, where the image is preprocessed with a Harris corner and edge detector [6], and the (log) gradient magnitude used for each pixel, gives $p_{i,\mathrm{harris}}(\mathbf{x}_i)$.

Thirdly, the classifier outputs are also multiplied with

$$p_{i,\mathrm{prod}}(\mathbf{x}_i) = p_{i,\mathrm{harris}}(\mathbf{x}_i) \times p_{i,\mathrm{hist}}(\mathbf{x}_i)\ , \qquad (21)$$

so that a strong positive response occurs when both classifiers are in agreement. Finally, the outputs from $p_{i,\mathrm{harris}}(\mathbf{x}_i)$ and $p_{i,\mathrm{hist}}(\mathbf{x}_i)$ are combined using (14). Figures 2 and 3 illustrate typical alignments, and example misalignments that can occur.

The AAM [1] was trained on the same set of (scale and translation normalized) faces, and used a 7-dimensional shape parameter. The appearance parameters were kept separate [13] as they are useful as a largely shape-independent feature vector in the recognition stage [9, 16]. For each test set the *best* set of patch classifiers from the BCLM results were employed in the generic CQF. The BCLM started with an initial window size of $L = 31$, which proved to be sufficiently large for a variety of features to be detected in the first iteration of the algorithm. The majority of computation time for alignment is consumed by classifier evaluations; using a single set of 12 patch classifiers a face is aligned in 0.2 to 0.3 seconds on a 2 GHz processor.

Test results are plotted in figures 4 and 5 in terms of alignment convergence curves, which show the percentage of faces that achieve a better average root mean squared point error (RMS-PE) than the given x-axis value. A combination of classifiers in a BCLM generally improves on their separate use and on a generic CQF, and shows a marked

improvement over an AAM. The mean in (1) is shown as a baseline.

The final result in figure 6 shows that instead of combining two sets of classifiers, a better fit can also be achieved by averaging the alignment with the alignment of a horizontal mirror image of the same face. This result can be expected: because each $\mathcal{D}_i$ was created from random patch samples from the training images, $p(a_i \,|\, \mathcal{I}(\mathbf{x}_i), \mathbf{w}_i)$ is not symmetric around horizontal flips of $\mathcal{I}(\mathbf{x}_i)$.

In practice, the tail of the alignment convergence curve is significant in a detection–alignment–recognition pipeline, as it represents the fraction of "badly aligned" faces that would typically be passed to the recognition stage. Considering results on the LFW dataset in figures 4 and 6, the BCLM is therefore most useful when faces appear in an unconstrained environment.

6. Conclusion

The generic CQF can be generalized to a Bayesian version, which allows multiple sets of patch alignment classifiers to be used. For even better alignment, the choice of K (where more degrees of freedom will allow for more accurate fits) needs to be traded against the quality and speed of the patch classifiers. As the BCLM formulation is probabilistic, other parameters and beliefs can be included in its Bayesian network to further improve accuracy in this task.

Acknowledgments

The author thanks Blaise Thomson, Chris Town, and David Sinclair for many fruitful discussions.

References

[1] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001.

[2] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *British Machine Vision Conference*, pages 929–938, 2006.

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

[4] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *British Machine Vision Conference*, pages 231–240, 2004.

[5] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *Proceedings of The 10th European Conference on Computer Vision*, 2008.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[7] 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.

[8] O. Jesorsky, K.J. Kirchberg, and R.W. Frischholz. Robust face detection using the Hausdorff distance. In J. Bigun and F. Smeraldi, editors, *Audio and Video based Person Authentication*, pages 90–95. Springer, 2001.

[9] A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[10] L. Liang, F. Wen, Y. Xu, X. Tang, and H. Shum. Accurate face alignment using shape constrained Markov network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1313–1319, 2006.

[11] X. Liu. Generic face alignment using boosted appearance model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[12] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[13] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[14] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2000, 2000.

[15] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1990.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[17] Paul Viola and Michael Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[18] Y. Wang, S. Lucey, and J. Cohn. Non-rigid object alignment with a mismatch template based on exhaustive local search. In *IEEE Workshop on Non-rigid Registration and Tracking through Learning*, 2007.

[19] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[20] H. Wu, X. Liu, and G. Doretto. Face alignment via boosted ranking model. In *Proceedings of the European Conference on Computer Vision*, volume 2, pages 72–85, 2008.