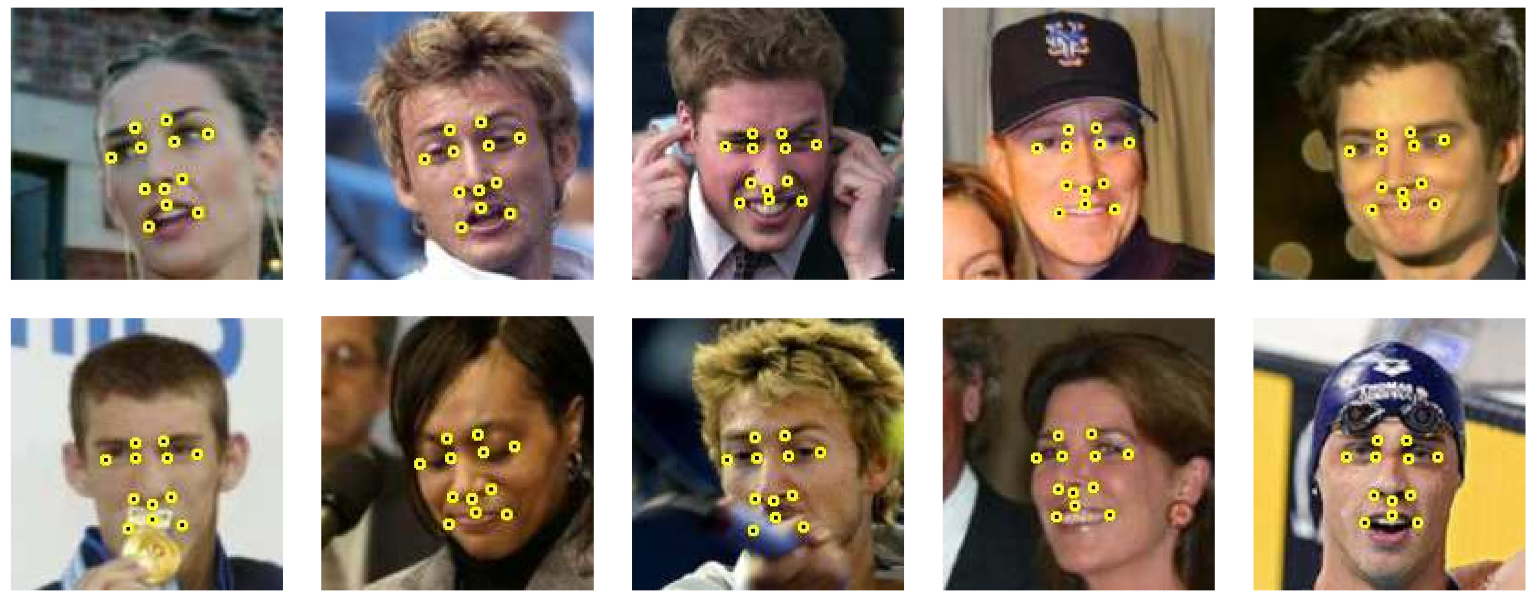
Convexity and Bayesian Constrained Local Models

Ulrich Paquet

Imense Ltd., Cambridge, UK

get the big picture



A selection of Bayesian Constrained Local Model (BCLM) alignments from the Labeled Faces in the Wild (LFW) data set.

What's it all about?

- Facial (nonrigid object) feature alignment.
- A Bayesian formulation of Constrained Local Models (CLMs): **likelihood + prior**.
- Various feature "patch classifiers" can be seamlessly incorporated into **likelihood** functions.
- In a *detection–alignment–recognition* face recognition pipeline, the alignment stage's **prior** can be explicitly based on the first stage face detector.

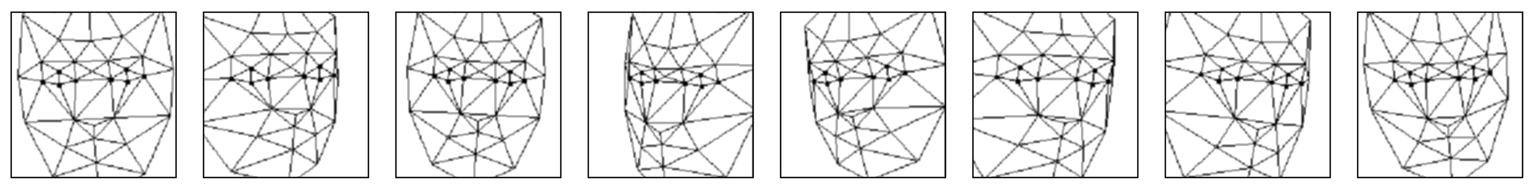
Notation. $\mathbf{x}$ indexes feature locations across an object (face). If $\mathbf{x}_i = (x_i, y_i)$ = centre of feature i, then $\mathbf{x} = (x_1, y_1, \ldots, x_I, y_I)$.

Point distribution model. A distribution on typical faces received from a detector, e.g. Viola–Jones (VJ).

- Lower dimensional $\mathbf{z} \in \mathbb{R}^K$ has *prior* $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ and is transformed to $\mathbf{x}$ with

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

- A generative model; noise-free Bayesian PCA.
- $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are estimated from marked-up VJ detected faces (posterior densities for them can be incorporated). Pipeline assumption.



Feature locations $\mathbf{x}$ generated from (1) and random draws from $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$.

Convex energy functions. Centered at $\mathbf{c}_i$, the **texture model** for aligning feature i is represented by ($\mathbf{A}_i$ pos. def.)

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

- Assumption: $\mathcal{E}_i(\mathbf{x}_i)$ is small if pixel $\mathbf{x}_i$ lies close to the true location of fiducial point i, and large otherwise.

An explicit Bayesian formulation

- **Offset** of the local energy function from the mean feature location: $\Delta\mathbf{m}_i = \mathbf{c}_i - \boldsymbol{\mu}_i$ = observed and dependent on $\mathbf{z}$ in a generative model.
- Negative log likelihood for $\mathbf{z}$ *given* observation $\Delta\mathbf{m}_i$ and some knowledge of $\mathbf{c}_i$ and $\mathbf{A}_i$,

$$\mathcal{E}_i(\mathbf{x}_i) = \mathcal{E}_i(\boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i\mathbf{z}) = \frac{1}{2}(\Delta\mathbf{m}_i - \boldsymbol{\Lambda}_i\mathbf{z})^\top \mathbf{A}_i (\Delta\mathbf{m}_i - \boldsymbol{\Lambda}_i\mathbf{z}) \ ,$$

gives a local alignment likelihood $p(\Delta\mathbf{m}_i|\mathbf{z}) = \frac{1}{Z}\exp(-\mathcal{E}_i(\mathbf{x}_i))$, or

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

Bayes' theorem. The the posterior distribution of $\mathbf{z}$ is Gaussian,

$$p(\mathbf{z}|\Delta\mathbf{m}) = \frac{p(\Delta\mathbf{m}|\mathbf{z})p(\mathbf{z})}{p(\Delta\mathbf{m})} = \frac{\prod_i p(\Delta\mathbf{m}_i|\mathbf{z})p(\mathbf{z})}{p(\Delta\mathbf{m})} = \mathcal{N}(\mathbf{z}; \boldsymbol{\nu}, \mathbf{S}) \tag{3}$$

with covariance $\mathbf{S} = (\boldsymbol{\Lambda}^\top \mathbf{A} \boldsymbol{\Lambda} + \mathbf{I})^{-1}$ and mean $\boldsymbol{\nu} = \mathbf{S}\boldsymbol{\Lambda}^\top \mathbf{A} \Delta\mathbf{m}$.

Multiple sets of feature detectors. Different patch alignment classifiers $r = 1, \ldots, R$ give different $\mathbf{c}_i^{(r)}$ and $\mathbf{A}_i^{(r)}$.

- Multiple observations $\Delta\mathbf{m}_i^{(r)} = \mathbf{c}_i^{(r)} - \boldsymbol{\mu}_i$ give a Gaussian posterior for $\mathbf{z}$ with covariance and mean

$$\mathbf{S} = \left[\boldsymbol{\Lambda}^\top \Big(\sum_r \mathbf{A}^{(r)}\Big)\boldsymbol{\Lambda} + \mathbf{I}\right]^{-1} \quad \text{and} \quad \boldsymbol{\nu} = \mathbf{S}\boldsymbol{\Lambda}^\top \sum_r \mathbf{A}^{(r)} \Delta\mathbf{m}^{(r)} \ .$$

Energy functions from patch classifiers

- Let $\mathbf{x}_i = (x_i, y_i)$ = centre of a $P \times P$ patch of pixels $\mathcal{I}(\mathbf{x}_i)$ in image $\mathcal{I}$.
- Define the binary variable $a_i \in \{-1, +1\}$ such that

$$p_i(\mathbf{x}_i) = p(a_i = 1 \,|\, \mathcal{I}(\mathbf{x}_i), \mathcal{M}_i) \tag{4}$$

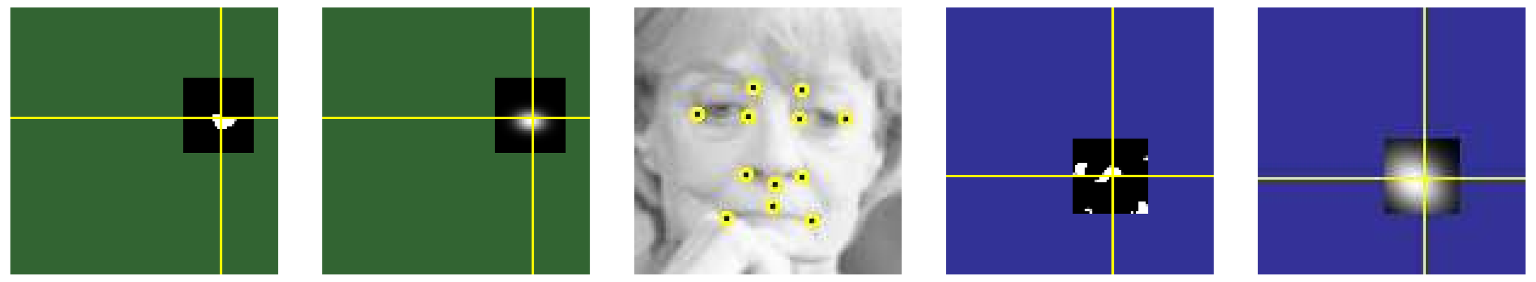
is the probability that $\mathbf{x}_i$ is centered at the i^{th} fiducial point, *given* its surrounding patch $\mathcal{I}(\mathbf{x}_i)$ and a patch classification model $\mathcal{M}_i$.

Local convex energy functions. Parameters $\mathbf{c}_i$ and $\mathbf{A}_i$ in (2) can be found analytically by minimizing

$$\operatorname*{arg\,min}_{\mathbf{A}_i, \mathbf{c}_i} \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)\, \mathcal{E}_i(\mathbf{x}_i) \ , \tag{5}$$

which equivalently fits a Gaussian density to weighted data in $\mathcal{W}(\mathbf{x}_i^*; L)$. With $s = \sum_{\mathbf{x}_i \in \mathcal{W}(\mathbf{x}_i^*; L)} p_i(\mathbf{x}_i)$ the minimum is straight-forward:

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



Alignment classifiers outputs $p_i(\mathbf{x}_i)$ and convex energy function approximations $\mathcal{E}_i(\mathbf{x}_i)$ for the right eye and nose corners, for each pixel $\mathbf{x}_i$ in a window $\mathcal{W}(\mathbf{x}_i^*; L)$ of width L pixels centered on some $\mathbf{x}_i^*$.

Logistic regression. For speed (as no kernel function evaluations are required)

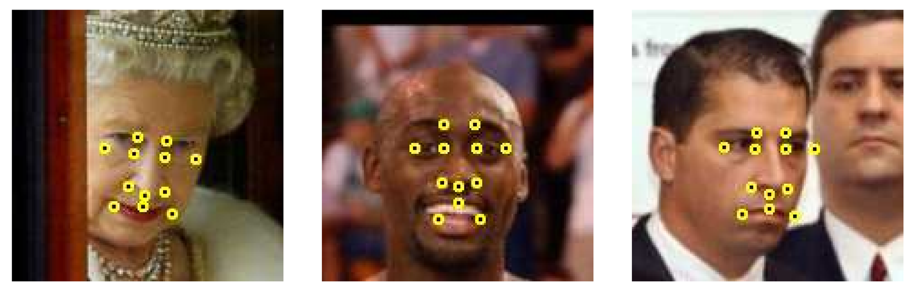
$$p(a_i \,|\, \mathcal{I}(\mathbf{x}_i), \mathbf{w}_i) = \sigma(a_i \mathbf{w}_i^\top \mathcal{I}(\mathbf{x}_i)) \tag{6}$$

is used. Hence $\mathbf{w}_i$ defines a patch classifier, and $\sigma(z) = 1/(1 + e^{-z})$. Training data sets were built around faces from publicly available Internet images, that were detected by a VJ detector (mirroring the LFW assumption).

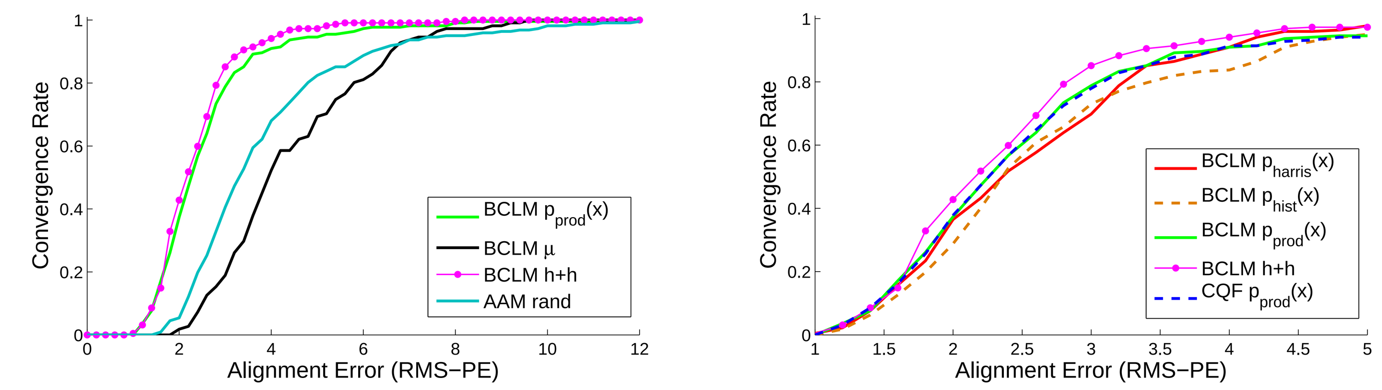
Results and illustrations

Bayesian Constrained Local Model algorithm.

- **initialize:** (preprocessed) face image $\mathcal{I}$ from detector **;** patch experts $\{\mathbf{w}_i\}_{i=1}^I$ **;** $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ **;** initial window size L **;** minimum window size $L_{\min}$ **;** initial warp $\boldsymbol{\nu} = \mathbf{0}$
- **repeat until** $L < L_{\min}$ **:**
  - **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 (6) **;** find $\mathbf{c}_i$ and $\mathbf{A}_i$ in (5)
  - $\Delta\mathbf{m} \leftarrow \mathbf{c} - \boldsymbol{\mu}$ and $\mathbf{A} \leftarrow \mathrm{diag}(\{\mathbf{A}_i\})$ **;** $\boldsymbol{\nu} \leftarrow (\boldsymbol{\Lambda}^\top \mathbf{A} \boldsymbol{\Lambda} + \mathbf{I})^{-1} \boldsymbol{\Lambda}^\top \mathbf{A} \Delta\mathbf{m}$ **;** $L \leftarrow L - 2$
- **return:** $\mathbf{x}^* \leftarrow \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\nu}$



A few example errors from the LFW data set.



The alignment error for different methods on the LFW data set, including an Active Appearance Model (AAM) and generic Convex Quadratic Fit (CQF).