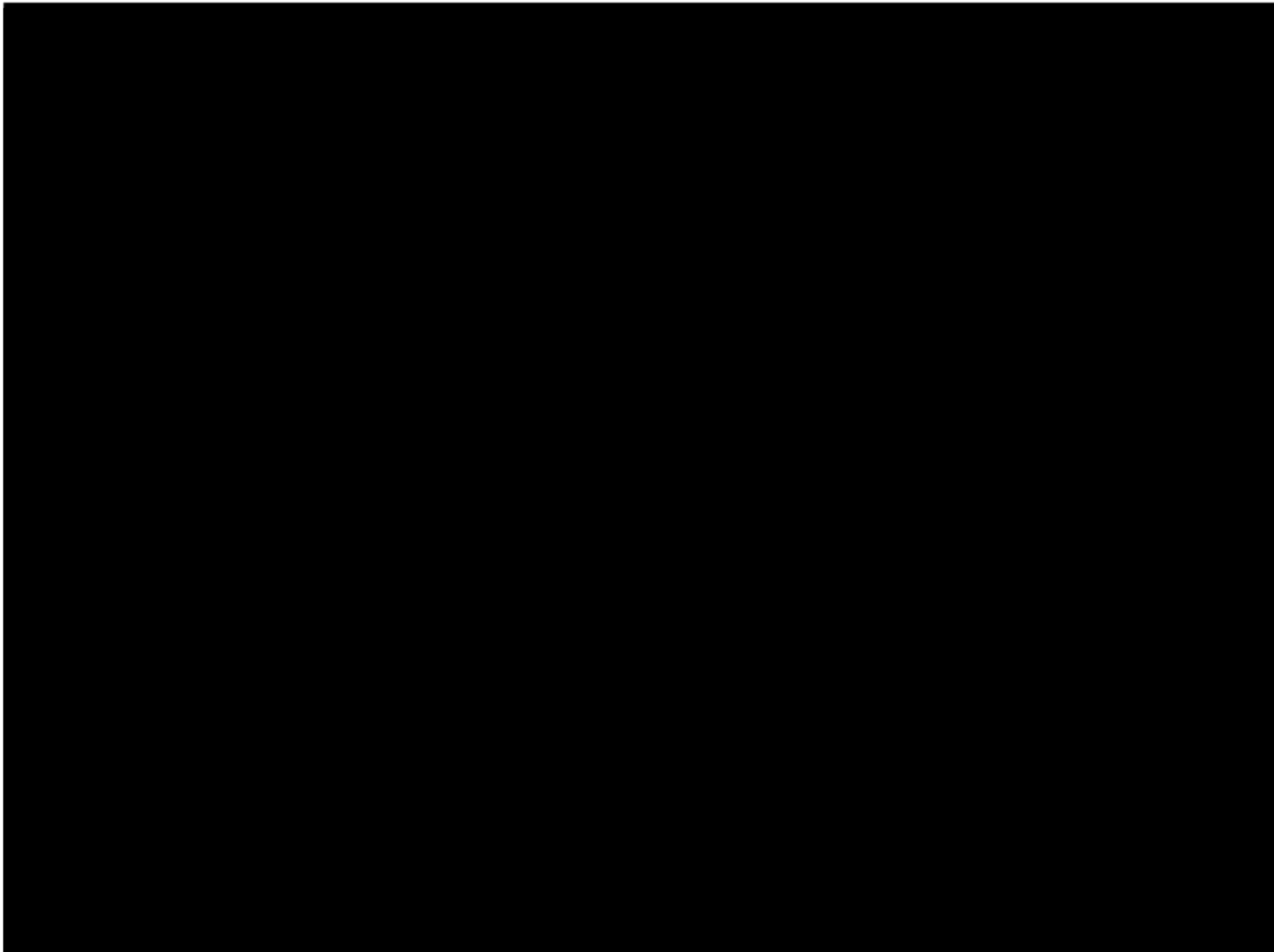
Let the Shape Speak - Discriminative Face Alignment using Conjugate Priors

Pedro Martins, Rui Caseiro, João F. Henriques, Jorge Batista
http://www.isr.uc.pt/~pedromartins

Institute of Systems and Robotics
University of Coimbra
Portugal

British Machine Vision Conference 2012

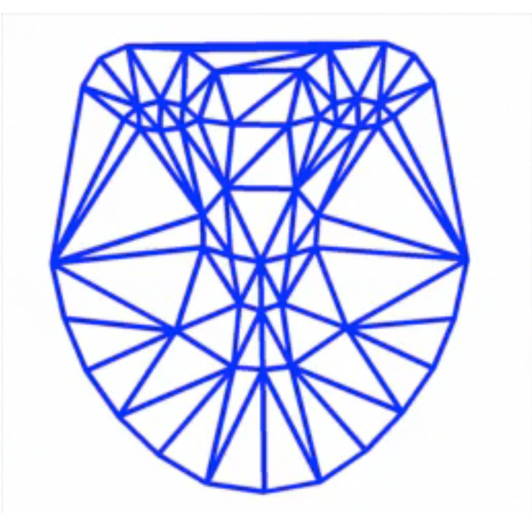
Demo Video

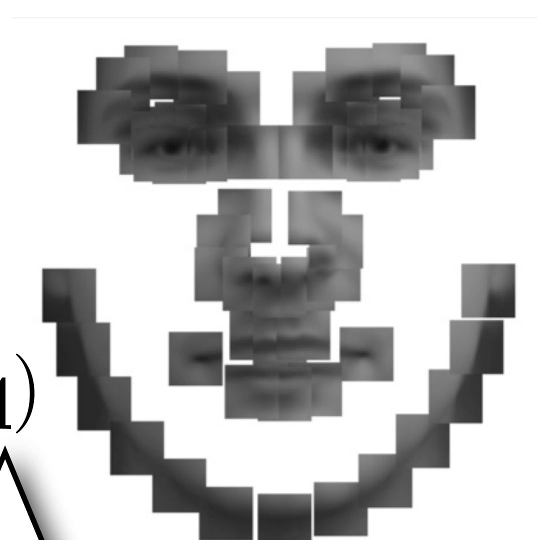


Introduction

- **Goal:** Face alignment in unseen images.
- Closely related to Constrained Local Models (CLM) and Active Shape Models (ASM), where a set of local detectors is constrained to lie in the subspace spanned by a Point Distribution Model (PDM).
    - Two step fitting approach:
        - **(1)** Local search using the local detectors (response maps for each landmark)
        - **(2)** Global optimization strategy that finds the PDM parameters that jointly maximize all the detections.
- **Proposed Work:** New Bayesian global optimization strategy where the prior distribution encodes the transition of the PDM parameters.
- The prior distribution is modeled using recursive Bayesian estimation. The mean and covariance are assumed to be unknown and treated as random variables.

Related Work - Parametric Image Alignment





Point Distribution Model

$$\mathbf{s} = \mathcal{S}\left(\mathbf{s}_0 + \Phi\mathbf{b}; \mathbf{q}\right)$$

Shape Parameters **Pose Parameters**

- **Generative / Holistic methods**
  - Active Appearance Models (AAM)
    T.F.Cootes, G.J.Edwards, C.J.Taylor - ECCV 98
  - 3D Morphable Models (3DMM)
    V.Blanz, T.Vetter - SIGGRAPH 99
  - Real Time Combined 2D+3D Active Appearance Models
    J.Xing, S.Baker, I.Matthews, T.Kanade - CVPR 2004
- **Discriminative / Patch-Based**
  - Active Shape Models (ASM)
    T.F.Cootes, G.J.Edwards, C.J.Taylor - CVIU 95
  - Constrained Local Model (CLM)
    D.Cristinance, T.F.Cootes - BMVC 2006
  - Convex Quadratic Fitting (CQF)
    Y.Wang, S.Lucey, J.Cohn - CVPR 2008
  - Bayesian Constrained Local Model (BCLM)
    U.Paquet - CVPR 2009
  - Subspace Constrained Mean-Shifts (SCMS)
    J.Saragih, S.Lucey, J.Cohn - ICCV 2009

The Alignment Goal

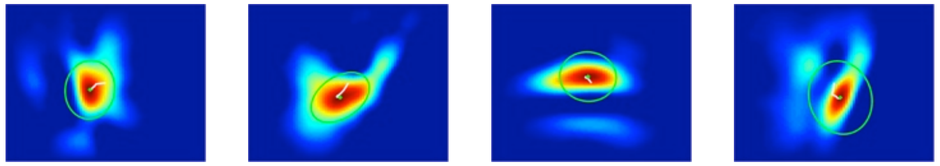
- Given a shape observation vector **(y)**, find the optimal set of shape (and pose) parameters **(b)** that maximize the posterior probability

$$\mathbf{b}^* = \arg\max_{\mathbf{b}} p(\mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b})$$

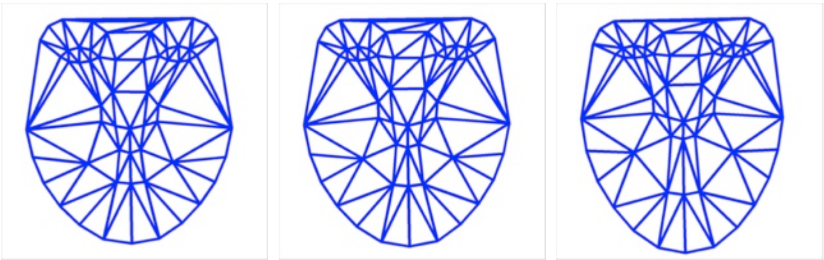
- Assuming:
  - Conditional independence between landmarks
  - Close to a solution

$$p(\mathbf{b}|\mathbf{y}) \propto \left(\prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b})\right) p(\mathbf{b}|\mathbf{b}^*_{k-1})$$

Likelihood from the local detectors



Prior on how parameters change

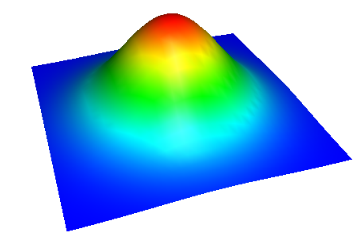


The Likelihood Term

- Convex energy function:

Observed shape **(y)**

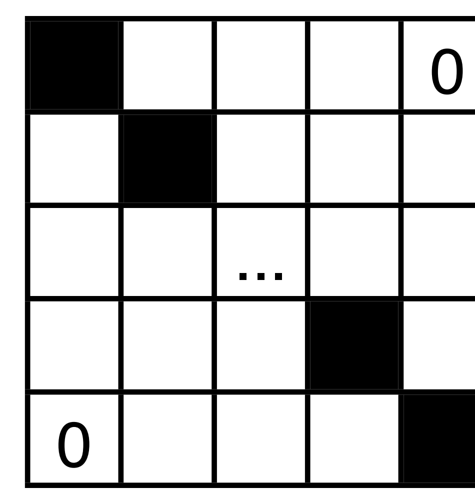
$$p(\mathbf{y}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}(\underbrace{\mathbf{y} - (\mathbf{s}_0}_{\Delta\mathbf{y}} + \Phi\mathbf{b}))^T \Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b}))\right)$$



Uncertainty covariance

Difference between the **observed** and the **mean shape**

$$\Sigma_{\mathbf{y}}$$

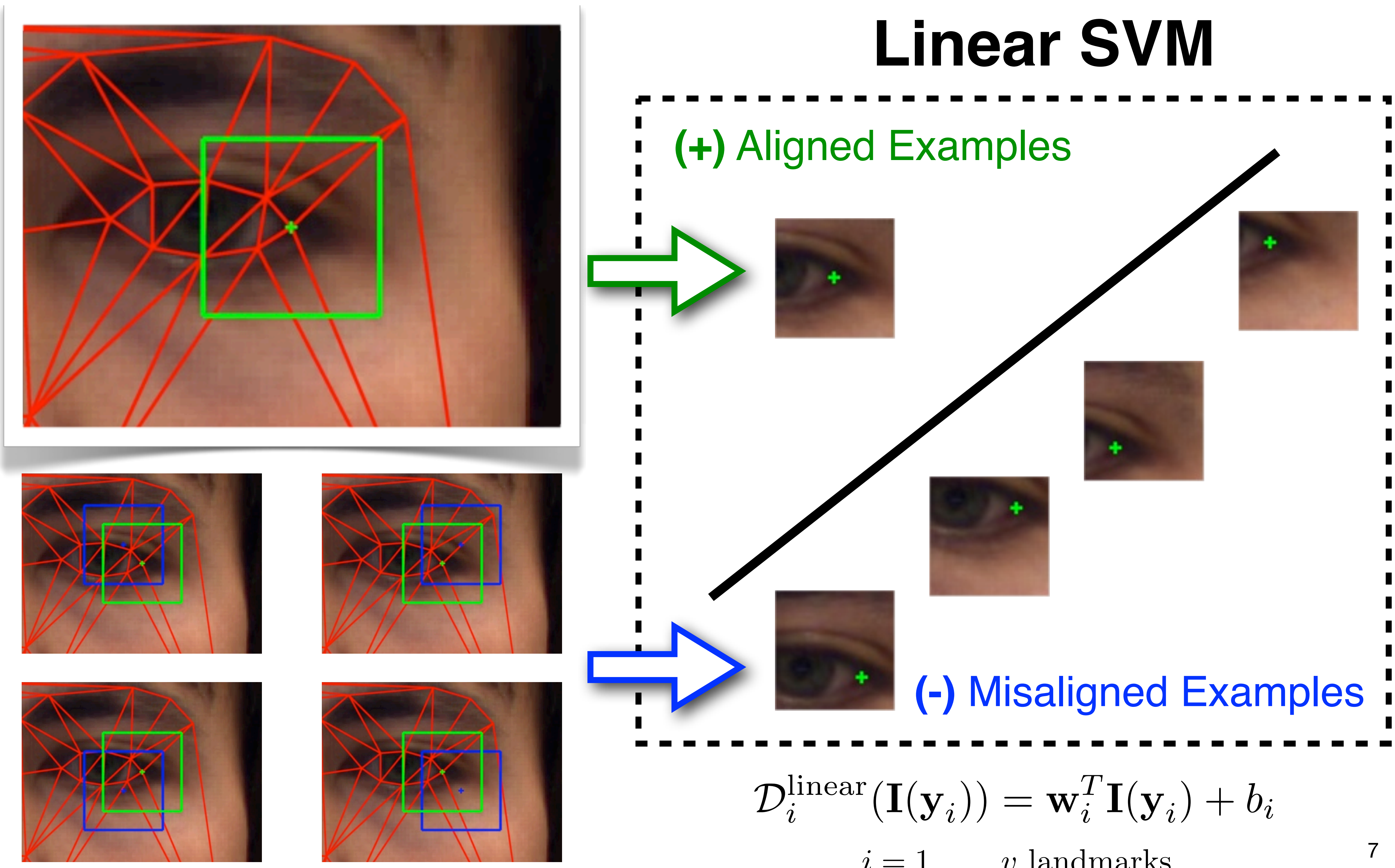


$2v \times 2v$ Block diagonal

The likelihood follow a Gaussian distribution

$$p(\mathbf{y}|\mathbf{b}) \propto \mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_{\mathbf{y}})$$

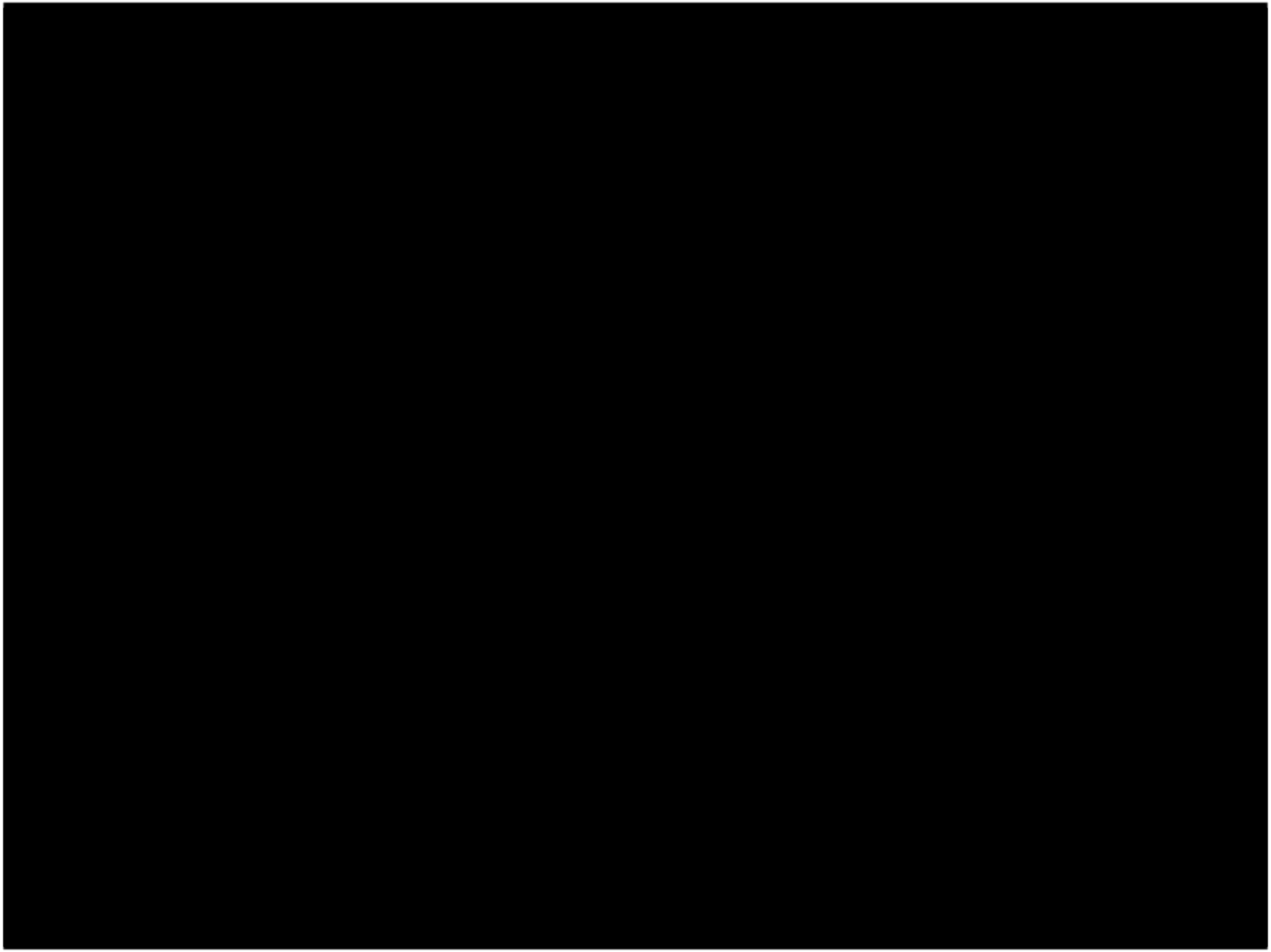
Local Landmark Detectors



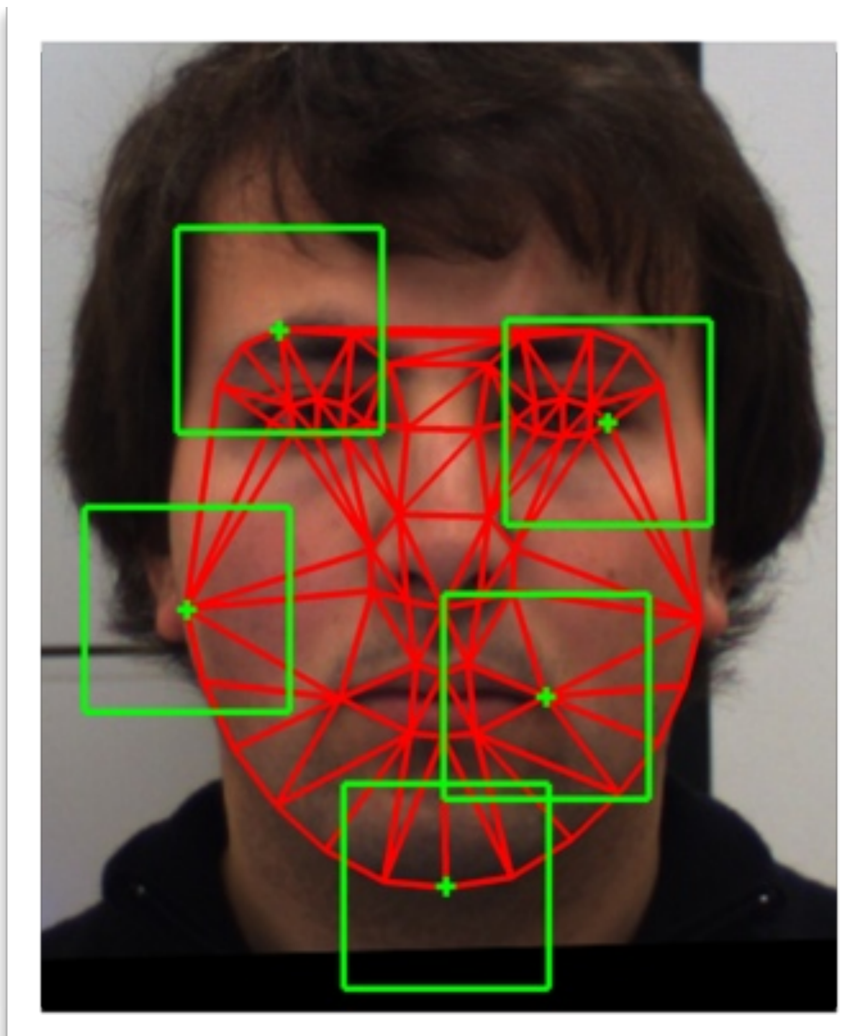
$$\mathcal{D}_i^{\text{linear}}(\mathbf{I}(\mathbf{y}_i)) = \mathbf{w}_i^T \mathbf{I}(\mathbf{y}_i) + b_i$$

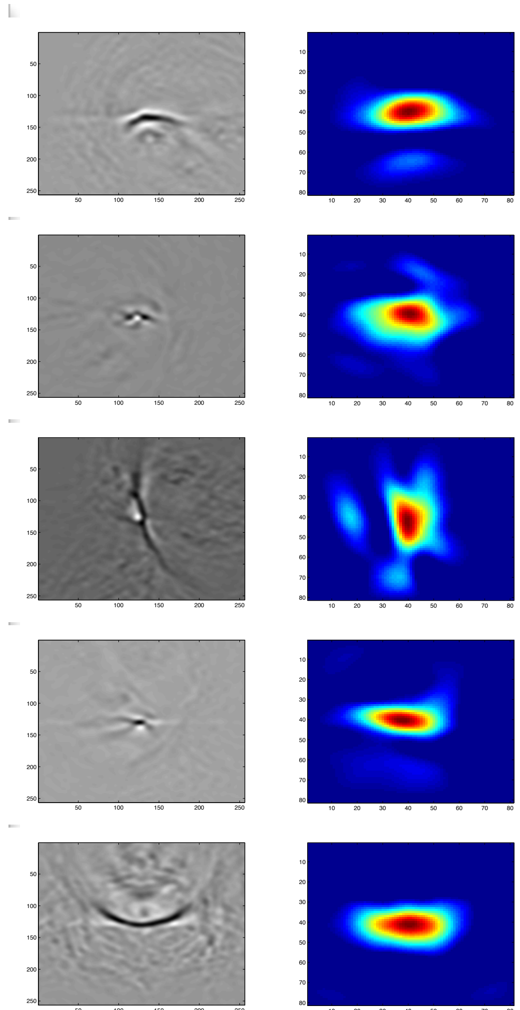
$i = 1, \ldots, v$ landmarks

Local Detectors



Local Landmark Detectors - MOSSE Filters

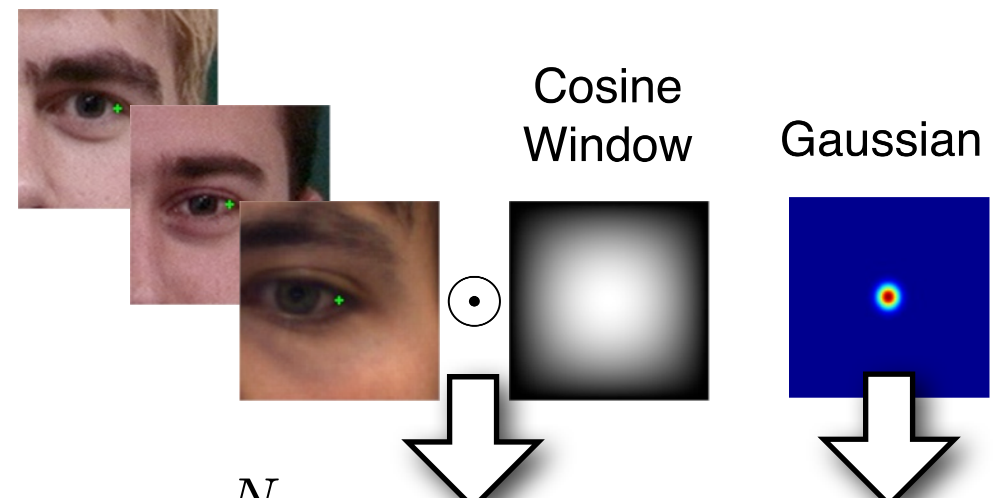




$\mathcal{F}^{-1}\{\mathbf{H}_i^*\}$ $\mathcal{D}_i^{\mathrm{MOSSE}}(\mathbf{I}(\mathbf{y}_i))$

- Correlation in Fourier Domain

$$\mathbf{G} = \mathcal{F}\{\mathbf{I}\} \odot \mathbf{H}^*$$



$$\min_{\mathbf{H}^*} \sum_{j=1}^{N} \left(\mathcal{F}\{\mathbf{I}_j\} \odot \mathbf{H}^* - \mathbf{G}_j\right)^2$$

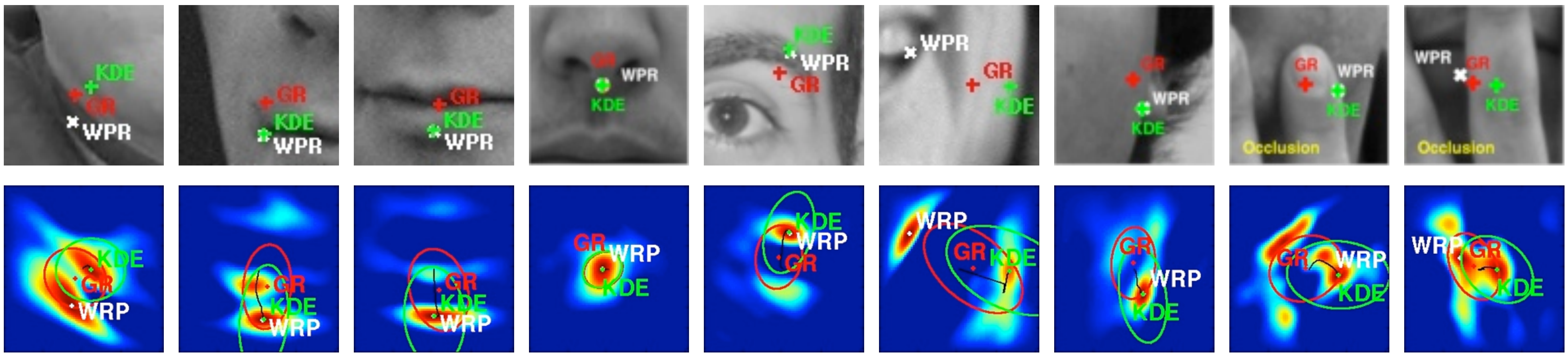
MOSSE Filter

$$\mathbf{H}^* = \frac{\sum_{j=1}^{N} \mathbf{G}_j \odot \mathcal{F}\{\mathbf{I}_j\}^*}{\sum_{j=1}^{N} \mathcal{F}\{\mathbf{I}_j\} \odot \mathcal{F}\{\mathbf{I}_j\}^*}$$

$$\mathcal{D}_i^{\mathrm{MOSSE}}(\mathbf{I}(\mathbf{y}_i)) = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{I}(\mathbf{y}_i)\} \odot \mathbf{H}_i^*\}$$

- Visual object tracking using adaptive correlation filters
D.Bolme, J.Beveridge, B.Draper, Y.Lui, CVPR 2010

Local Optimization Strategies

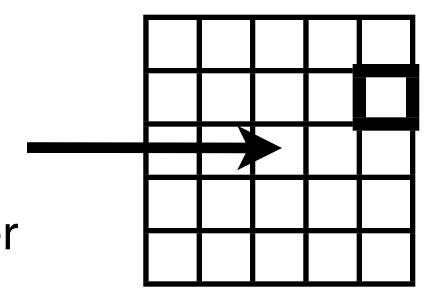


$p_i(\mathbf{z}_i)$

Prob. $\mathbf{z}_i$ is aligned

$\mathbf{y}_i^c$

Center



$\mathbf{z}_i = (x_i, y_i)$

Pixel candidate to ith landmark location

$\boldsymbol{\Omega}_{\mathbf{y}_i^c}$ Local search grid

Patches under occlusion

Weighted Peak Response (WPR)

$$\mathbf{y}_i^{\mathrm{WPR}} = \max_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} (p_i(\mathbf{z}_i))$$

$$\Sigma_{\mathbf{y}_i}^{\mathrm{WPR}} = diag(p_i(\mathbf{y}_i^{\mathrm{WPR}})^{-1})$$

ASM

Gaussian Response (GR)

$$\mathbf{y}_i^{\mathrm{GR}} = \frac{1}{d} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)\mathbf{z}_i \qquad d = \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)$$

$$\Sigma_{\mathbf{y}_i}^{\mathrm{GR}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{GR}})^T$$

CQF, BCLM

Kernel Density Estimator (KDE)

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}$$

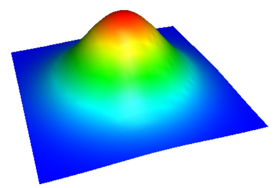
$$\Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \boldsymbol{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T$$

SCMS

KDE Demo Video

MAP Global Alignment

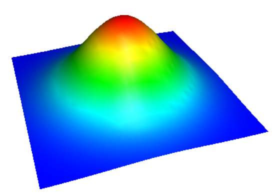
Likelihood



$$p(\mathbf{y}|\mathbf{b}) \propto \mathcal{N}(\Delta\mathbf{y}|\Phi\mathbf{b}, \Sigma_{\mathbf{y}})$$

Mean that is function of **b**
Covariance independent of **b**

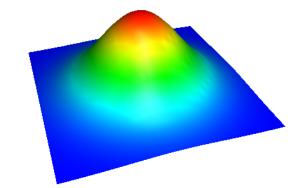
Prior



$$p(\mathbf{b}_k|\mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$$

Bayes's Theorem for
Gaussian variables

Posterior



$$p(\mathbf{b}_k|\mathbf{y}) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = (\Sigma_{\mathbf{b}}^{-1} + \Phi^T \Sigma_{\mathbf{y}}^{-1} \Phi)^{-1}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}(\Phi^T \Sigma_{\mathbf{y}}^{-1} \mathbf{y} + \Sigma_{\mathbf{b}}^{-1} \mu_{\mathbf{b}})$$

Model the Covariance of **b**
(2nd Order Estimate)
+
Bayesian Fusion of
Multiple Detectors

$$p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\Sigma}_k = \left((\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} + \Phi^T \sum_{m=1}^{M} \left(\Sigma_{\mathbf{y}_{(m)}}^{-1} \right) \Phi \right)^{-1}$$

$$\boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k \left(\Phi^T \sum_{m=1}^{M} \left(\Sigma_{\mathbf{y}_{(m)}}^{-1} \Delta\mathbf{y}_{(m)} \right) + (\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} \mu_{\mathbf{b}_k} \right)$$

$\Delta\mathbf{y}_{(m)}, \Sigma_{\mathbf{y}_{(m)}}$
Multiple Likelihood
Observations

The Prior Term

- Mean and Covariance $(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ are assumed to be **unknown** and modeled as **random variables**.

Unknown

$$p(\mathbf{b}_k | \mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k | \boxed{\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}})$$

Observable vector **b**

Bayes Theorem: $$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}} | \mathbf{b}) \propto p(\mathbf{b} | \mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) \; p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$$

Joint Posterior
Normal Inverse-Wishart

Joint Prior
Normal Inverse-Wishart

Parameters ⇩

Conjugate Prior for a Gaussian with unknown mean and covariance is a Normal Inverse-Wishart distribution

Degrees of freedom $\upsilon_k = \upsilon_{k-1} + m$

Number of measurements $\kappa_k = \kappa_{k-1} + m$

Mean $\theta_k = \frac{\kappa_{k-1}}{\kappa_{k-1} + m} \theta_{k-1} + \frac{m}{\kappa_{k-1} + m} \overline{\mathbf{b}}$

m - Number of samples to update the model
$\overline{\mathbf{b}}$ - Mean of all samples

Scale matrix $\Lambda_k = \Lambda_{k-1} + \frac{\kappa_{k-1} m}{\kappa_{k-1} + m} (\overline{\mathbf{b}} - \theta_{k-1})(\overline{\mathbf{b}} - \theta_{k-1})^T$

The Prior Term

- Using the expectation of marginal posterior distributions as the model parameters update.

Expectation

Joint Posterior
Normal Inverse-Wishart

$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ ⇨

Marginalizing
w.r.t $\Sigma_{\mathbf{b}}$

$p(\mu_{\mathbf{b}}|\mathbf{b}) \propto$ Multivariate
Student t

$\mu_{\mathbf{b}_k} = E(\mu_{\mathbf{b}}|\mathbf{b}) = \theta_k$

Marginalizing
w.r.t $\mu_{\mathbf{b}}$

$p(\Sigma_{\mathbf{b}}|\mathbf{b}) \propto$ Inv-Wishart

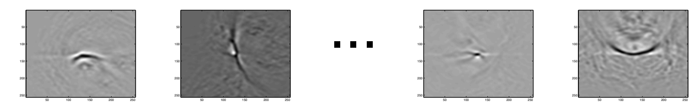
$\Sigma_{\mathbf{b}_k} = E(\Sigma_{\mathbf{b}}|\mathbf{b}) = (\upsilon_k - n - 1)^{-1} \Lambda_k$

The Prior distribution is **continuously kept up to date**

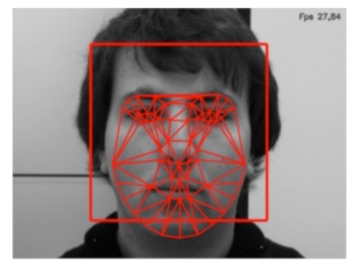
The Algorithm

Precompute: PDM: $\mathbf{s}_0, \Phi, \Psi, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$

MOSSE Filters: $\mathbf{H}_i^* \quad {}_{i=1,\ldots,v}$

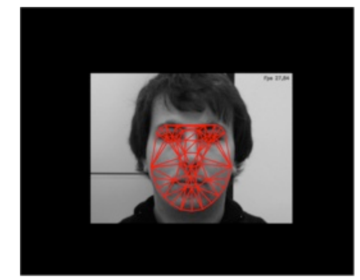


Initial estimate $(\mathbf{b}_0, \Sigma_0), (\mathbf{q}_0, \Sigma_0^q)$

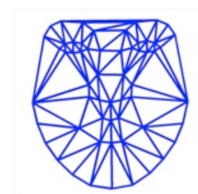


```
for k=1:1:MaxIterations
```

Warp Image to the base mesh, using the current pose parameters



Generate current shape $\mathbf{s} = \mathcal{S}\left(\mathbf{s}_0 + \Phi\mathbf{b}_k; \mathbf{q}_k\right)$

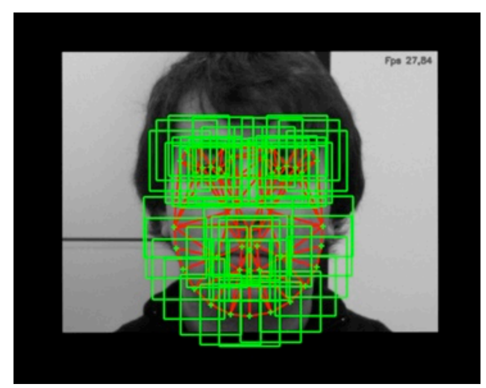


```
for i=1:1:Landmarks
```

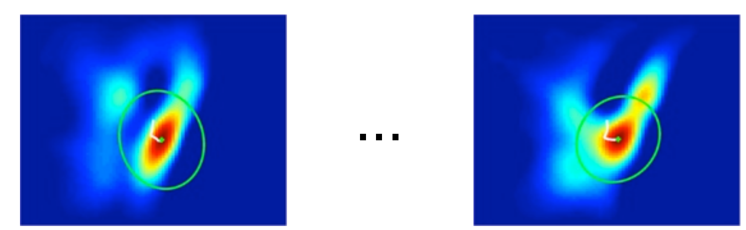
Evaluate detectors response

Find the likelihood parameters $\mathbf{y}_i, \Sigma_{\mathbf{y}_i}$

```
end
```



WPR, **GR** or **KDE** Local Strategy



Estimate the shape/pose parameters:

Update the parameters of Normal Inv-Wishart distribution $\rightarrow \upsilon_k, \kappa_k, \theta_k, \mathit{\Lambda}_k$

Expectation of the prior shape parameters $\longrightarrow \mu_{\mathbf{b}_k} = \theta_k, \quad \Sigma_{\mathbf{b}_k} = (\upsilon_k - n - 1)^{-1}\mathit{\Lambda}_k$

Evaluate the **global** shape parameters and the covariance $\rightarrow \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

```
end
```

Hierarchical Search (BASM-KDE-H)

- When response maps are approximated by KDE representations.

Mean-Shift Landmark Update

$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \mathbf{\Omega}_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}$$

Bandwidth schedule

$$\sigma_h^2 = (15, 10, 5, 2)$$

Standard Search

for k=1:1:MaxIterations

for i=1:1:v (LandMarks)

Evaluate de Detectors Response

Mean-Shift Landmark Update $\boxed{\sigma_h^2 = (15, 10, 5, 2)}$

end

Global Optimization

end

Hierarchical Search

for k=1:1:MaxIterations

for $\boxed{\sigma_h^2 = (15, 10, 5, 2)}$

for i=1:1:v (LandMarks)

Evaluate de Detectors Response

Mean-Shift Landmark Update $\sigma_{h_j}^2$

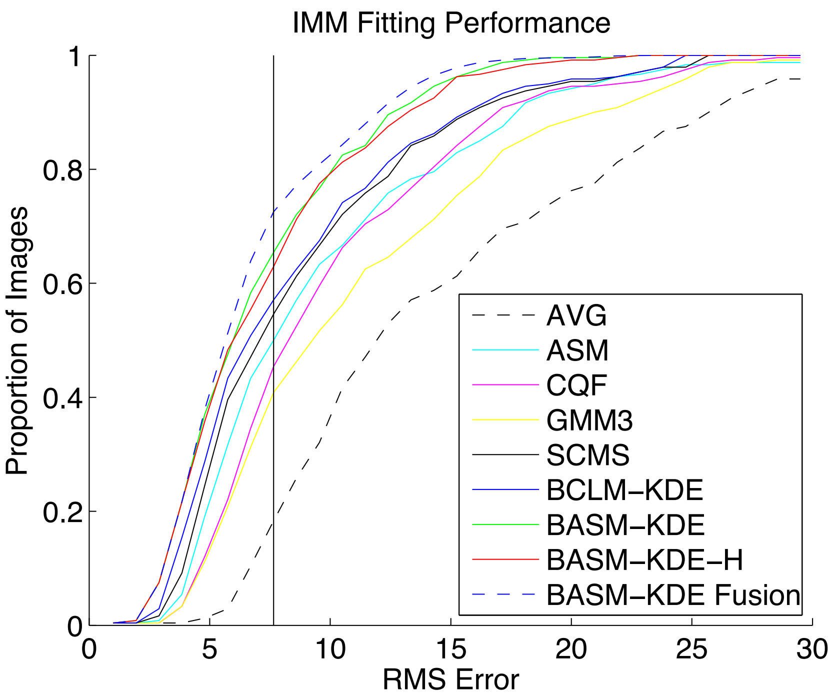
end

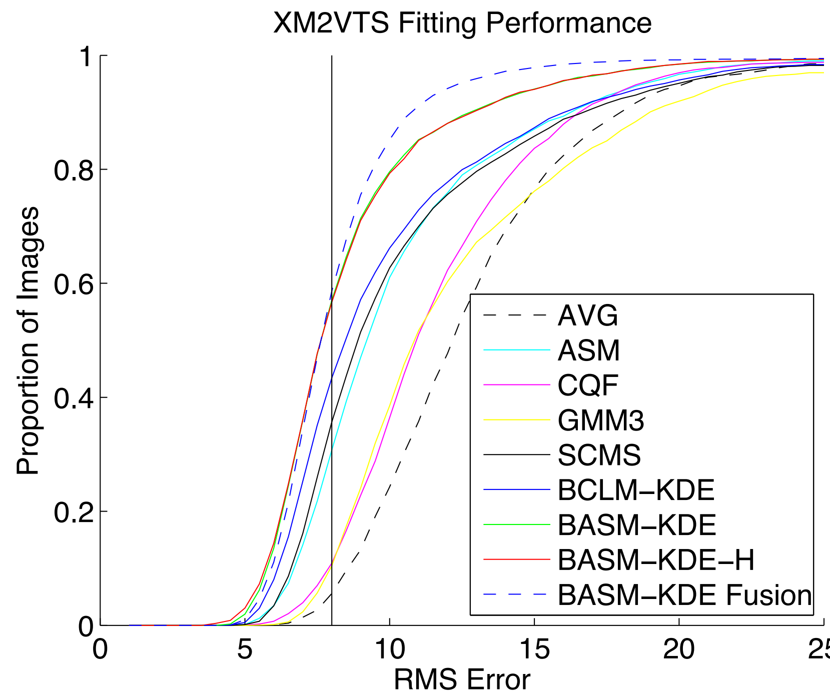
Global Optimization

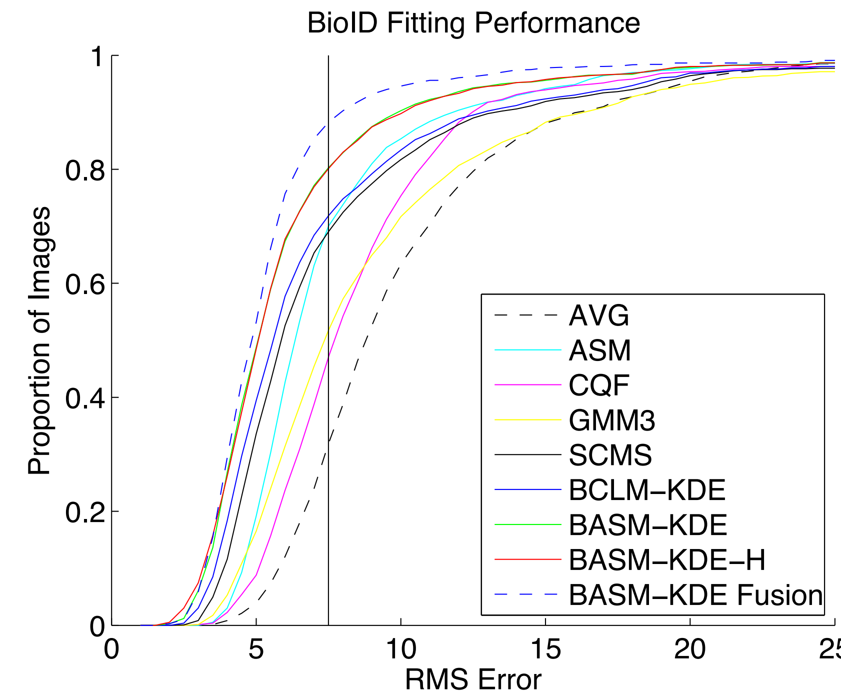
end

end

Evaluation Results



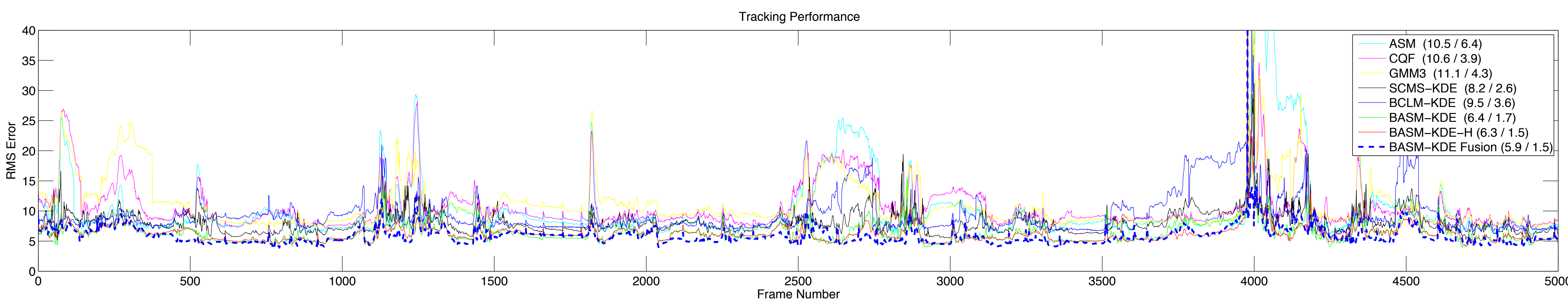




Reference 7.5 RMS	IMM (240 images)		XM2VTS (2360 images)		BioID (1521 images)	
ASM	50.0		30.7		70.0	
BASM-WPR (our method)	**58.4**	(+8.4)	**47.4**	(+16.7)	**77.1**	(+7.1)
CQF	45.4		10.9		47.0	
GMM3	40.8	(-4.6)	10.4	(-0.5)	51.7	(+4.7)
BCLM-GR	48.3	(+2.9)	15.9	(+5.0)	54.2	(+7.2)
BASM-GR (our method)	**51.8**	(+6.4)	**19.7**	(+8.8)	**63.5**	(+16.5)
SCMS-KDE	54.6		35.7		69.0	
BCLM-KDE	57.1	(+2.5)	43.4	(+7.7)	71.9	(+2.9)
BASM-KDE (our method)	**65.4**	(+10.8)	**57.0**	(+21.3)	**80.3**	(+11.3)
BASM-KDE-H (our method)	64.0	(+9.4)	56.6	(+20.9)	79.9	(+10.9)
BASM-KDE Fusion of 2 Detectors	**72.5**	(+17.9)	**58.7**	(+23.0)	**88.2**	(+19.2)

Tracking Performance

- FGNET Talking Face Video Sequence



RMS Error	ASM	CQF	GMM3	SCMS-KDE	BCLM-KDE	BASM-KDE	BASM-KDE-H	BASM-KDE Fusion
Mean	10.5	10.6	11.1	8.2	9.5	6.4	6.3	5.9
Standard Deviation	6.4	3.9	4.3	2.6	3.6	1.7	1.5	1.5

Tracking Performance

Labeled Faces in the Wild (LFW)

Conclusions

- Bayesian formulation for aligning faces in unseen images.
- New global optimization strategy infers both shape and pose parameters, in MAP sense, by explicitly modeling the prior distribution.
- The prior distribution is continuously kept up to date.
- Recursive Bayesian estimation is used to treat the mean and covariance as random variables.
- Extensive evaluations were performed on several standard datasets (IMM, XM2VTS, BioID, LFW and FGNET Talking Face) against state-of-the-art methods while using the same local detectors.

Acknowledgements

- Work supported by the Portuguese Science Foundation (*Fundação para a Ciência e Tecnologia* - FCT) under the project grant PTDC/EIA-CCO/108791/2008.
- Pedro Martins, Rui Caseiro and João F. Henriques also acknowledge the FCT through the grants SFRH/BD/45178/2008, SFRH/BD74152/2010 and SFRH/BD/75459/2010, respectively.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Qualitative Evaluation - Labeled Faces in the Wild