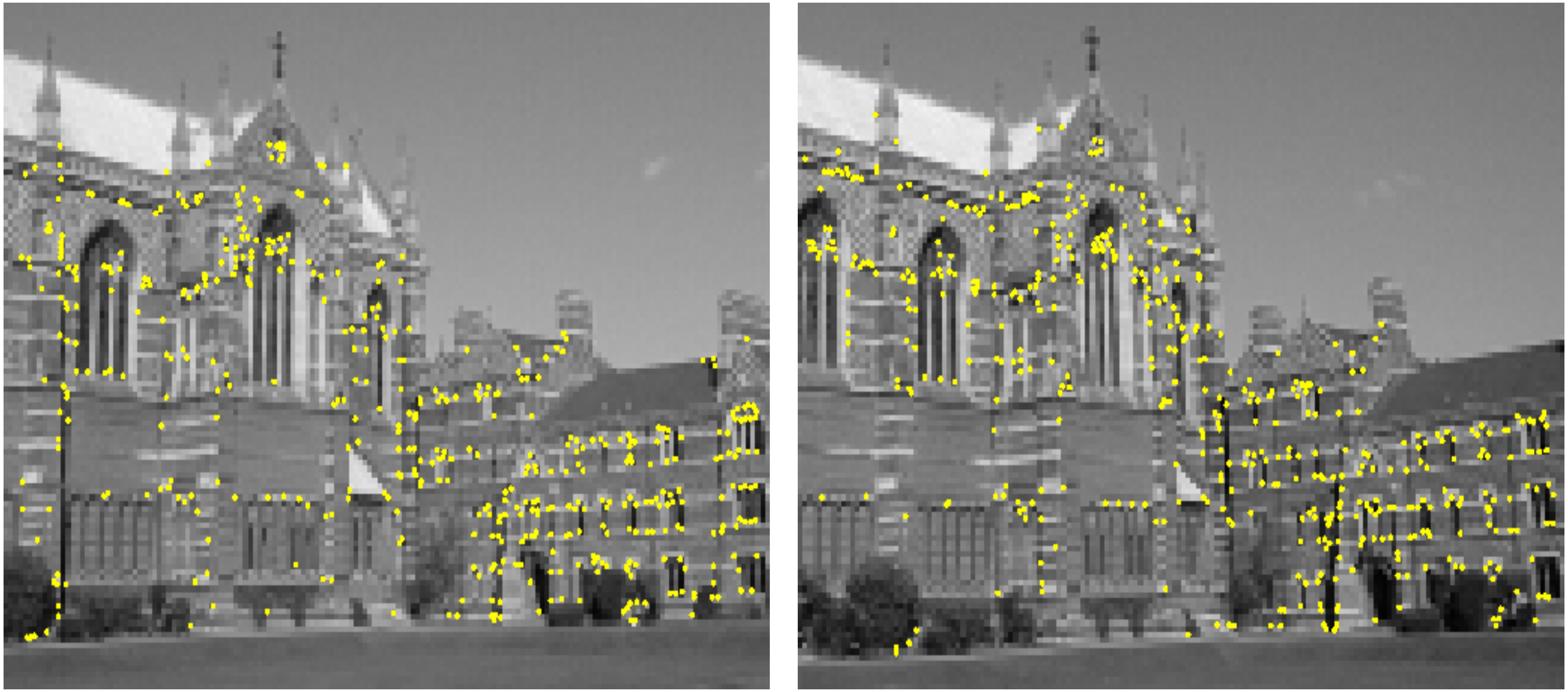
Features & A Smidgen of Learning

Frank Dellaert

Slides Adapted from Cordelia Schmid and David Lowe's Short course at CVPR

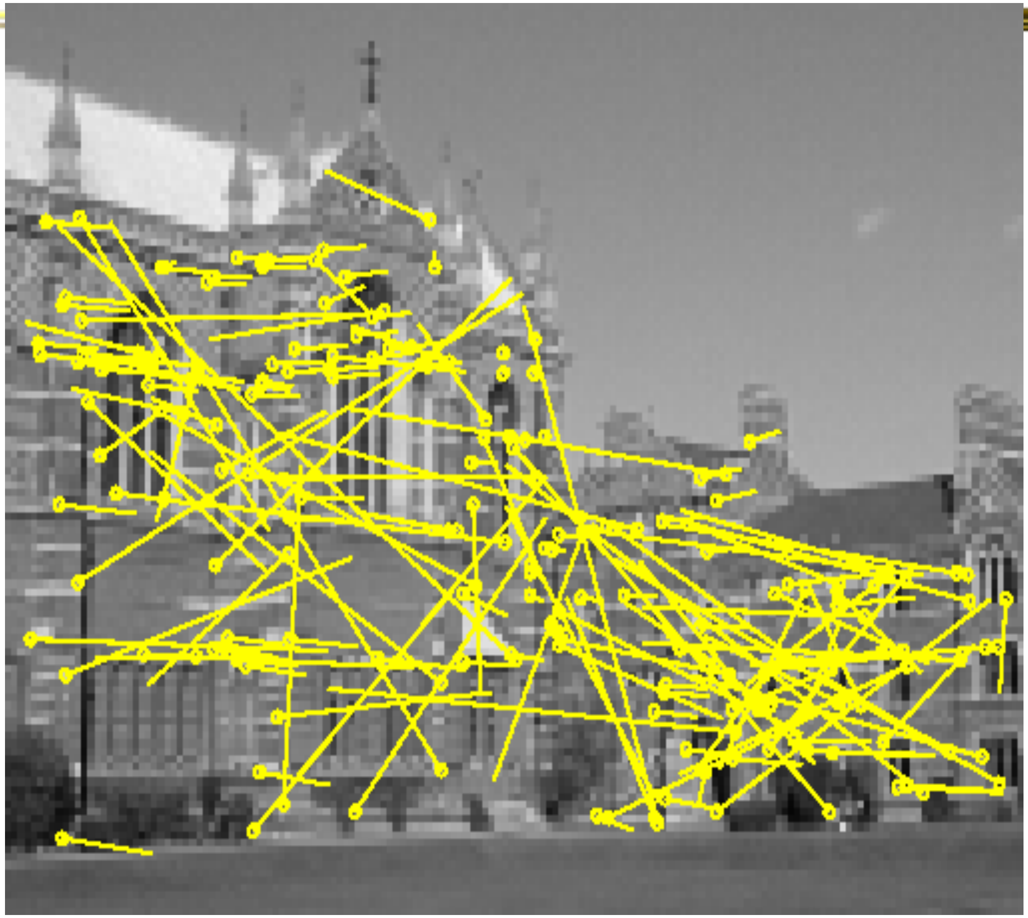
(used with permission)

Harris detector



Interest points extracted with Harris (~ 500 points)

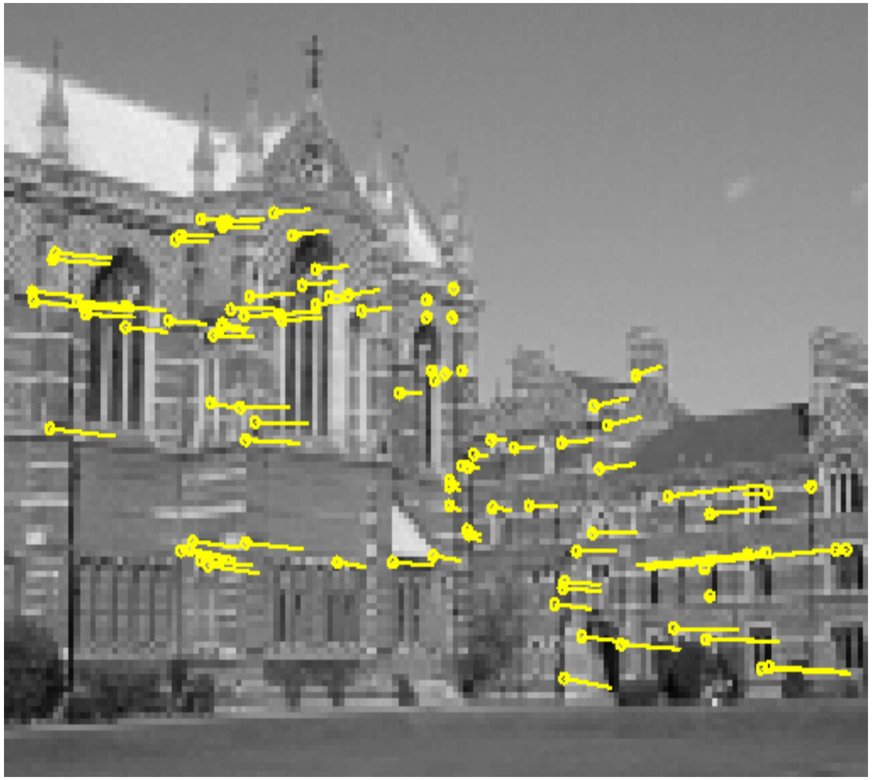
Cross-correlation matching



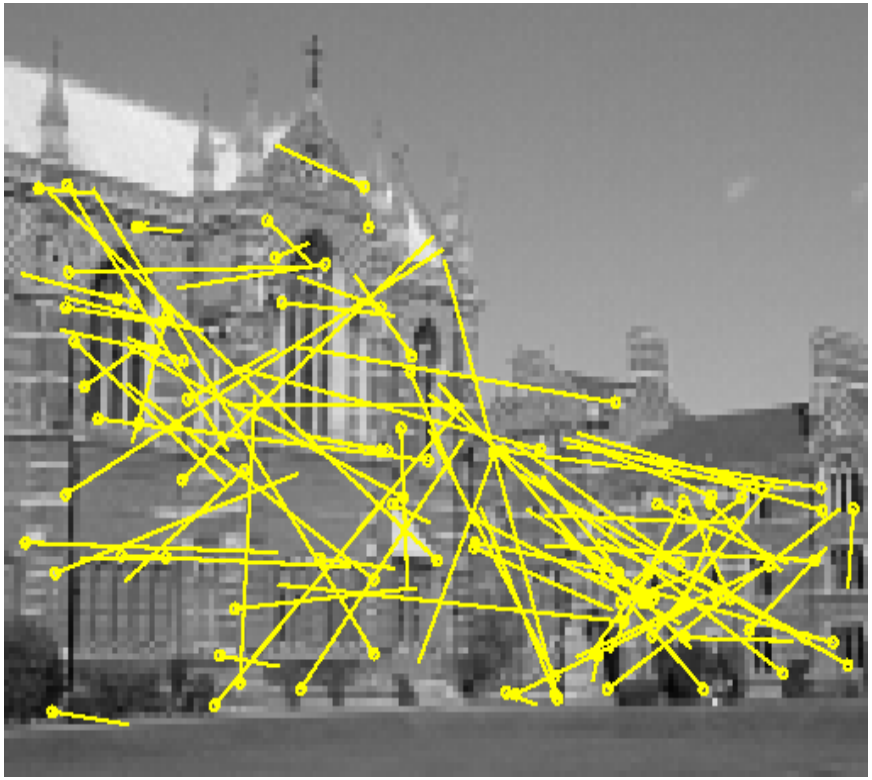
Initial matches (188 pairs)

RANSAC (again!)

Robust estimation of the fundamental matrix

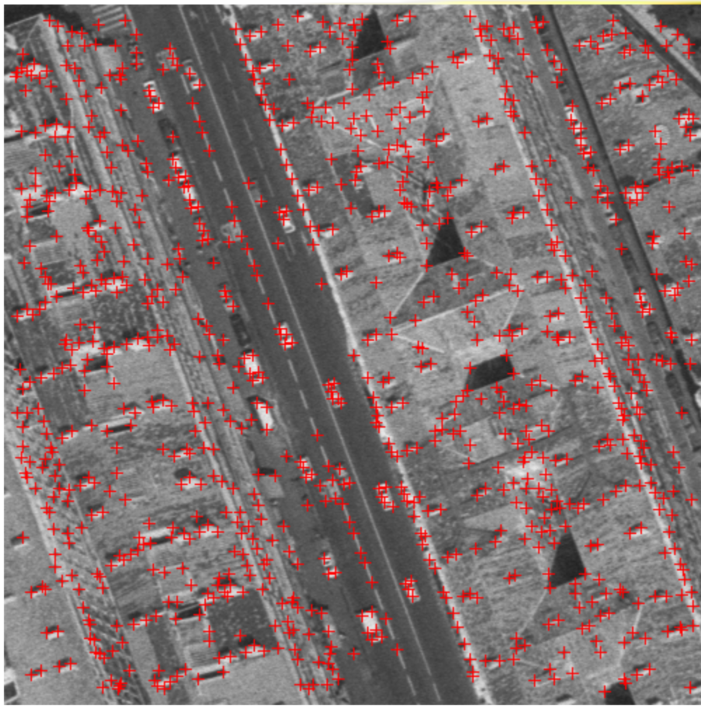


99 inliers



89
outliers

Interest points



Geometric features

➡ repeatable under transformations

2D characteristics of the signal

➡ high informational content

Comparison of different detectors [Schmid98] ➡ Harris detector

Harris detector

Based on the idea of auto-correlation

Important difference in all directions => interest point

Harris detector

Auto-correlation function for a point (x,y) and a shift $(\Delta x,\Delta y)$

$$f(x,y)=\sum_{(x_k,y_k)\in W}\left(I(x_k,y_k)-I(x_k+\Delta x,y_k+\Delta y)\right)^2$$

Discrete shifts can be avoided with the auto-correlation matrix

with $I(x_k+\Delta x,y_k+\Delta y)=I(x_k,y_k)+\begin{pmatrix}I_x(x_k,y_k) & I_y(x_k,y_k)\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix}$

$$f(x,y)=\sum_{(x_k,y_k)\in W}\left(\begin{pmatrix}I_x(x_k,y_k) & I_y(x_k,y_k)\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix}\right)^2$$

Harris detector

$$= \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \begin{bmatrix} \sum\limits_{(x_k, y_k) \in W} (I_x(x_k, y_k))^2 & \sum\limits_{(x_k, y_k) \in W} I_x(x_k, y_k) I_y(x_k, y_k) \\ \sum\limits_{(x_k, y_k) \in W} I_x(x_k, y_k) I_y(x_k, y_k) & \sum\limits_{(x_k, y_k) \in W} (I_y(x_k, y_k))^2 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

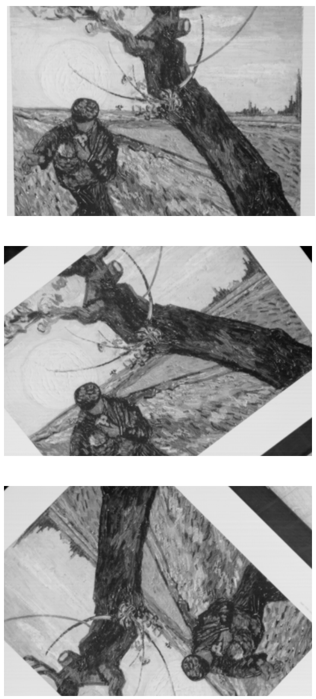
Auto-correlation matrix

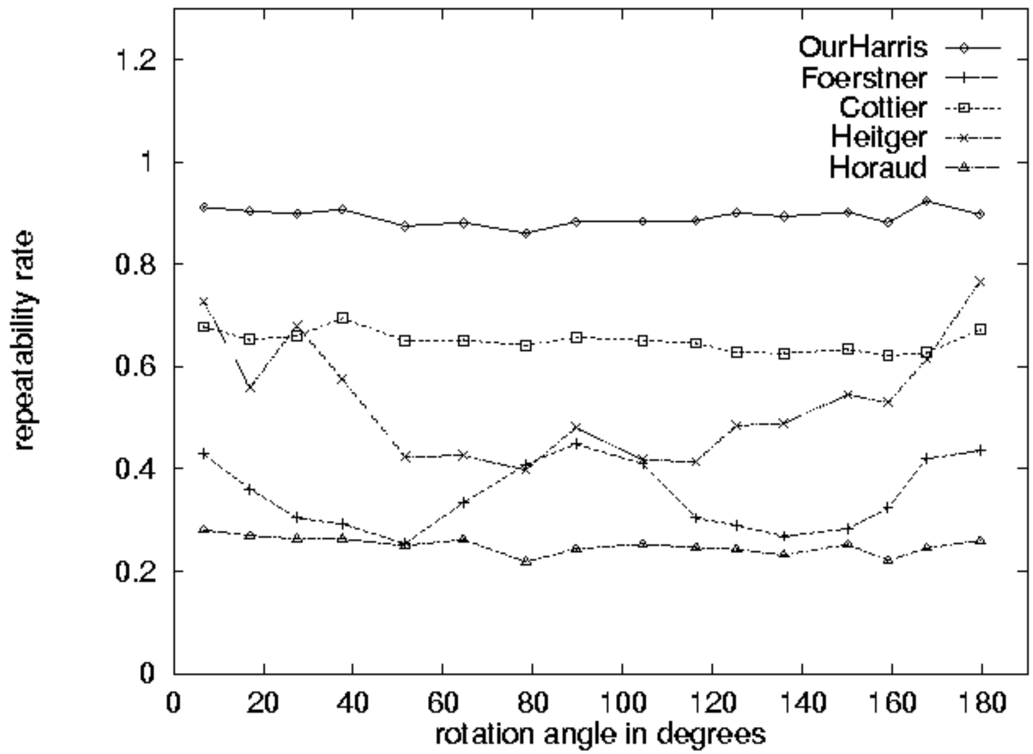
Harris detection

- Auto-correlation matrix
  - captures the structure of the local neighborhood
  - measure based on eigenvalues of this matrix
    - 2 strong eigenvalues => interest point
    - 1 strong eigenvalue => contour
    - 0 eigenvalue => uniform region
- Interest point detection
  - threshold on the eigenvalues
  - local maximum for localization

Comparison of different detectors

repeatability - image rotation

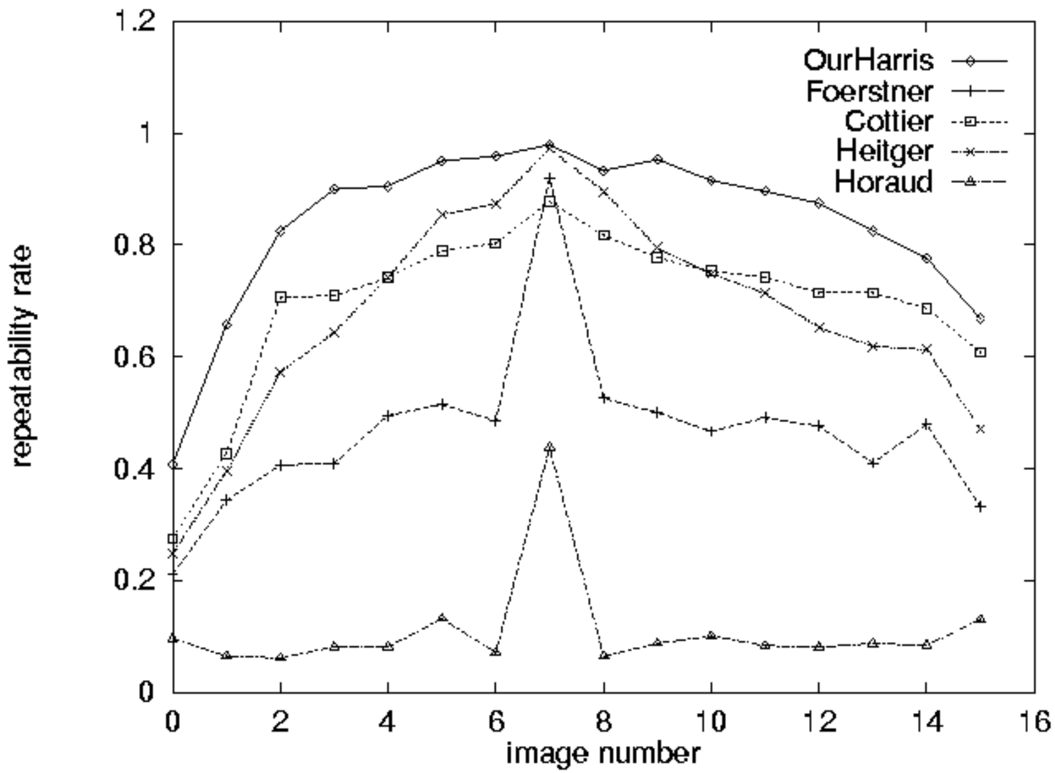




[Comparing and Evaluating Interest Points, Schmid, Mohr & Bauckhage, ICCV 98]

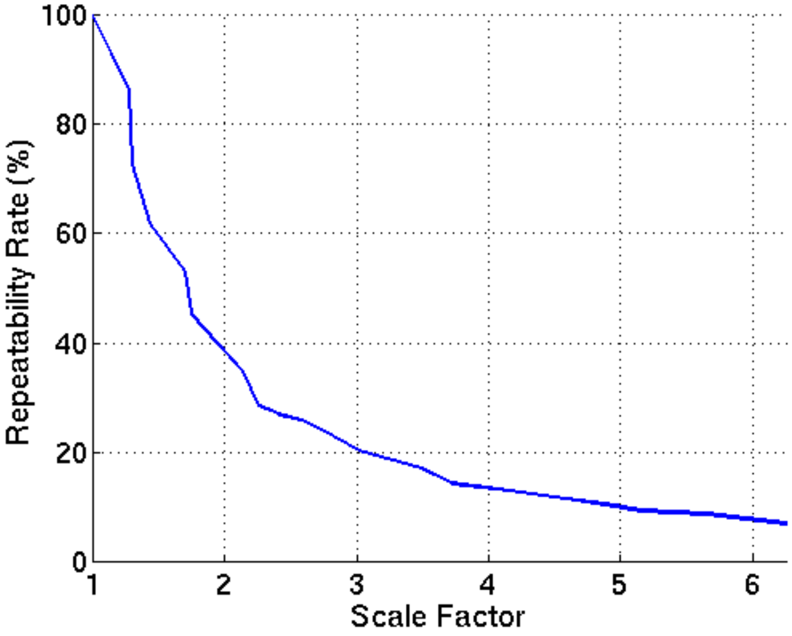
Comparison of different detectors

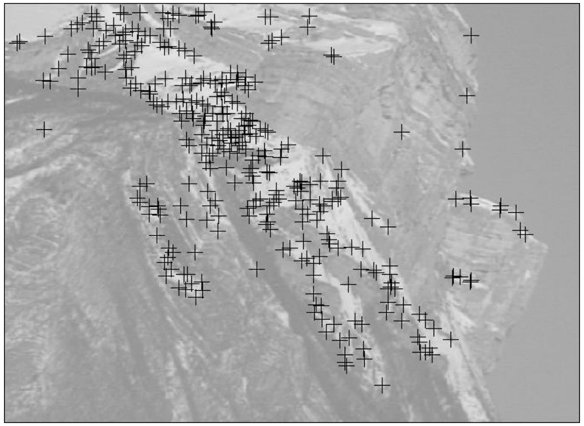
repeatability – perspective transformation

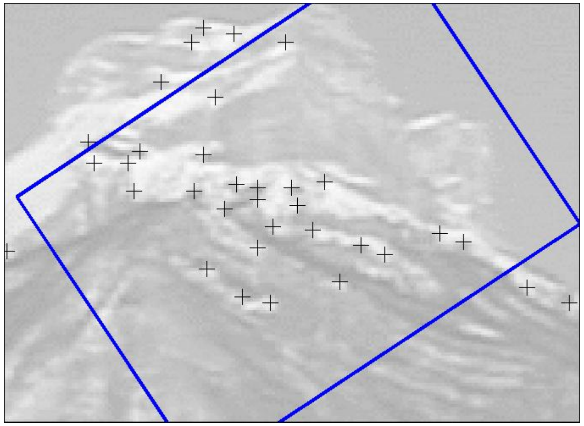


[Comparing and Evaluating Interest Points, Schmid, Mohr & Bauckhage, ICCV 98]

Harris detector + scale changes



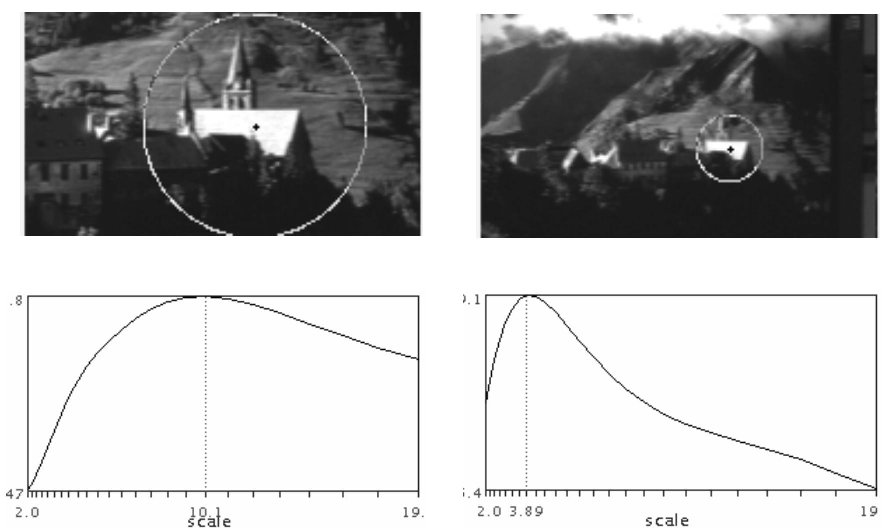




Scale invariant Harris points

- Multi-scale extraction of Harris interest points

- Selection of points at characteristic scale in scale space

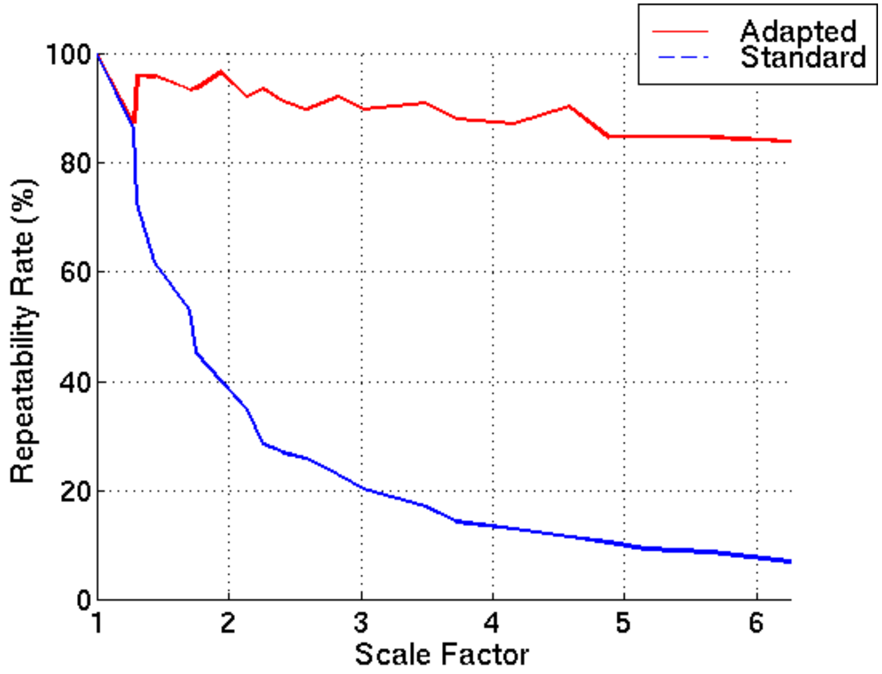


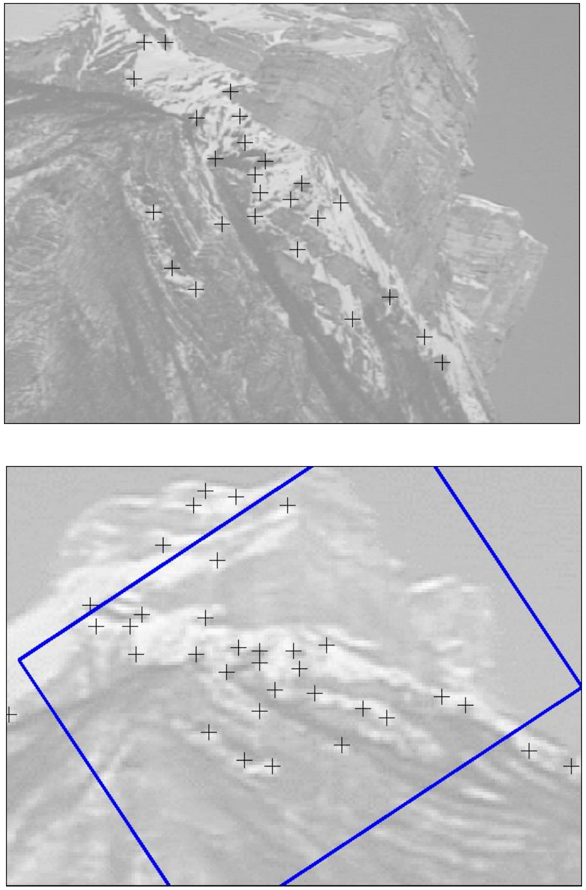
Laplacian

Chacteristic scale :

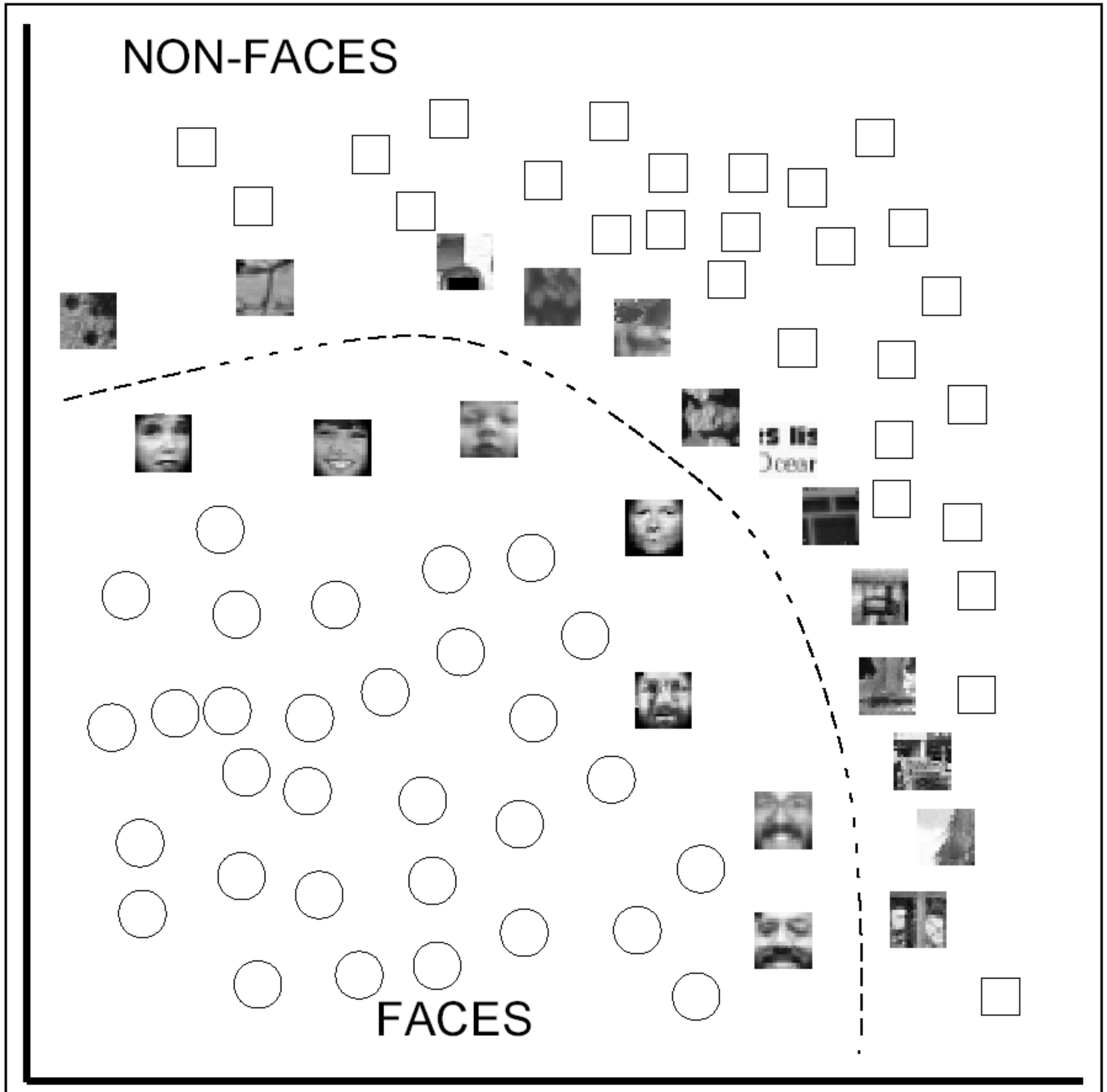
- maximum in scale space
- scale invariant

Harris detector – adaptation to scale





Decision Boundaries



In Powerpoint !
(via Andrew Blake)

In Powerpoint !
(via Andrew Blake)

Weak learners

- Accurate learners = hard
- “so-so” learners = easy

- Example:
  - if “buy” occurs in email, classify as SPAM

- Weak learner = “rule of thumb”

More Weak Learners

- Perceptron
- Decision stumps
- Haar wavelets cfr. Papageorgiou et al, and later Viola+Jones:

AdaBoost

- Given a set of weak classifiers

$$h_j(\mathbf{x}) \in \{+1, -1\}$$

None much better than random

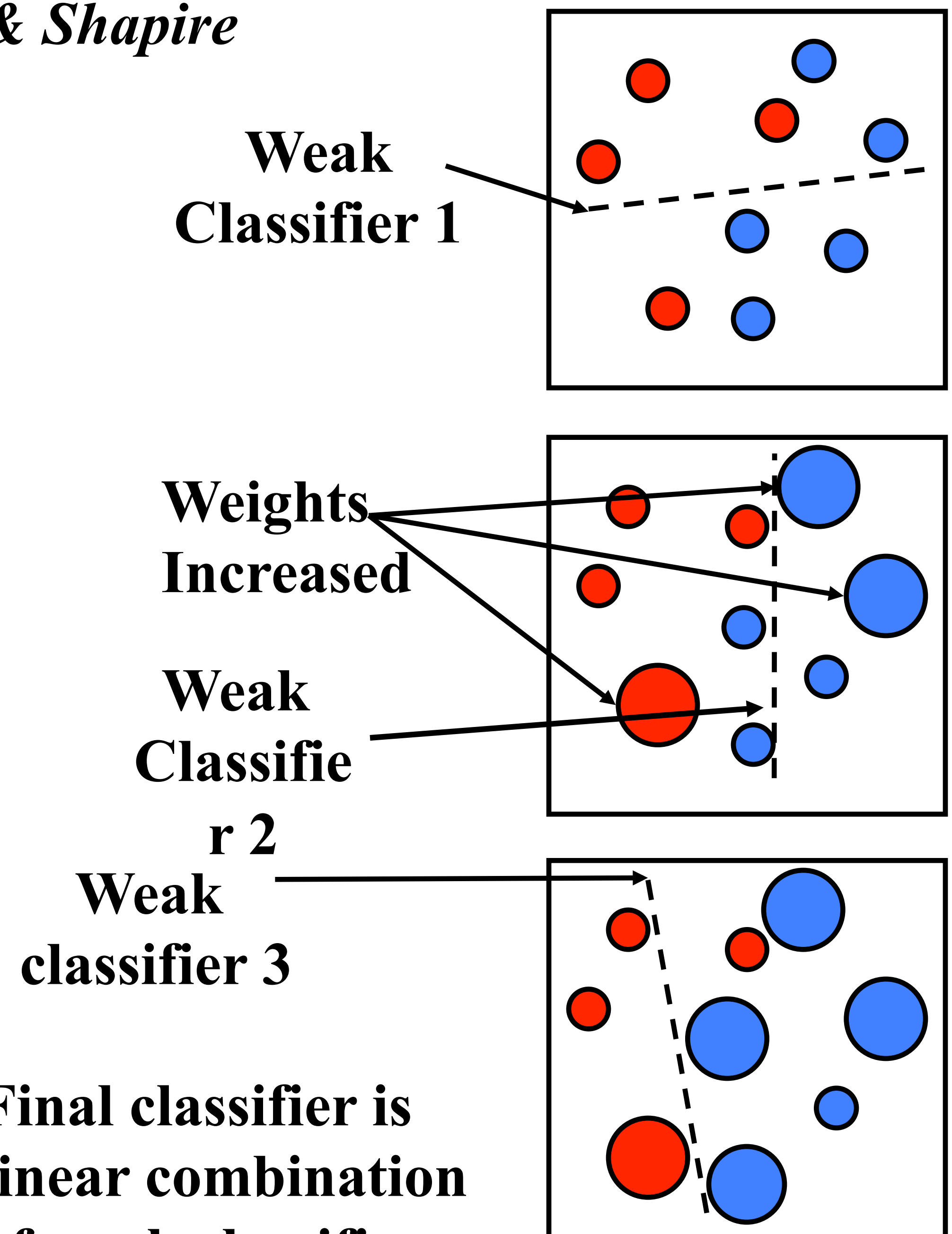
- Iteratively combine classifiers
  - Form a linear combination

$$C(x) = \theta\left(\sum_t h_t(x) + b\right)$$

  - Training error converges to 0 quickly
  - Test error is related to training margin

Freund & Shapire

AdaBoost



Final classifier is linear combination of weak classifiers

Adaboost Algorithm

For $t = 1, \ldots, T$:

Freund & Shapire

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \{-1, +1\}$ with error

$$\epsilon_t = \Pr_{i \sim D_t}\left[h_t(x_i) \neq y_i\right].$$

- Choose $\boxed{\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)}$
- Update:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} \\ &= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \end{aligned}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

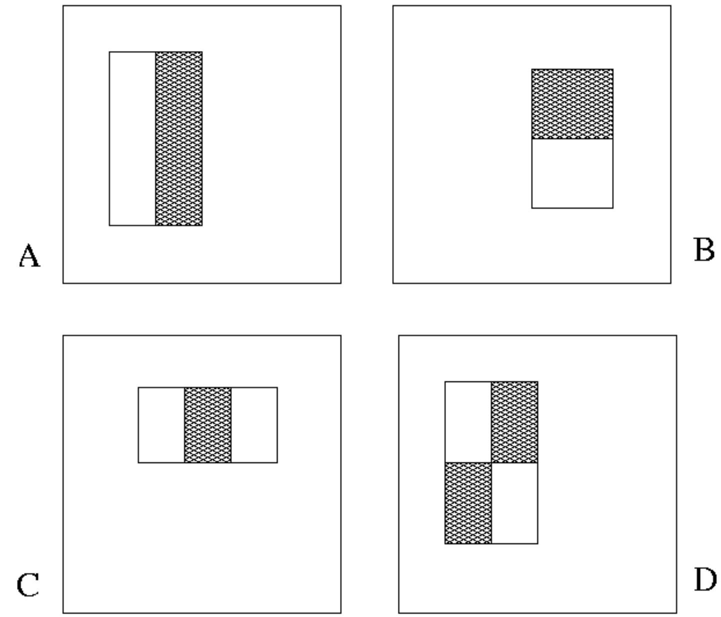
$$\boxed{H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)}$$

AdaBoost: Super Efficient Feature Selector

- Features = Weak Classifiers
- Each round selects the optimal feature given:
  - Previous selected features
  - Exponential Loss

Boosted Face Detection: Image Features

“Rectangle filters”



Similar to Haar wavelets
Papageorgiou, et al.

$$h_t(x_i) = \begin{cases} \alpha_t & \text{if } f_t(x_i) > \theta_t \\ \beta_t & \text{otherwise} \end{cases}$$

$$C(x) = \theta\left(\sum_t h_t(x) + b\right)$$

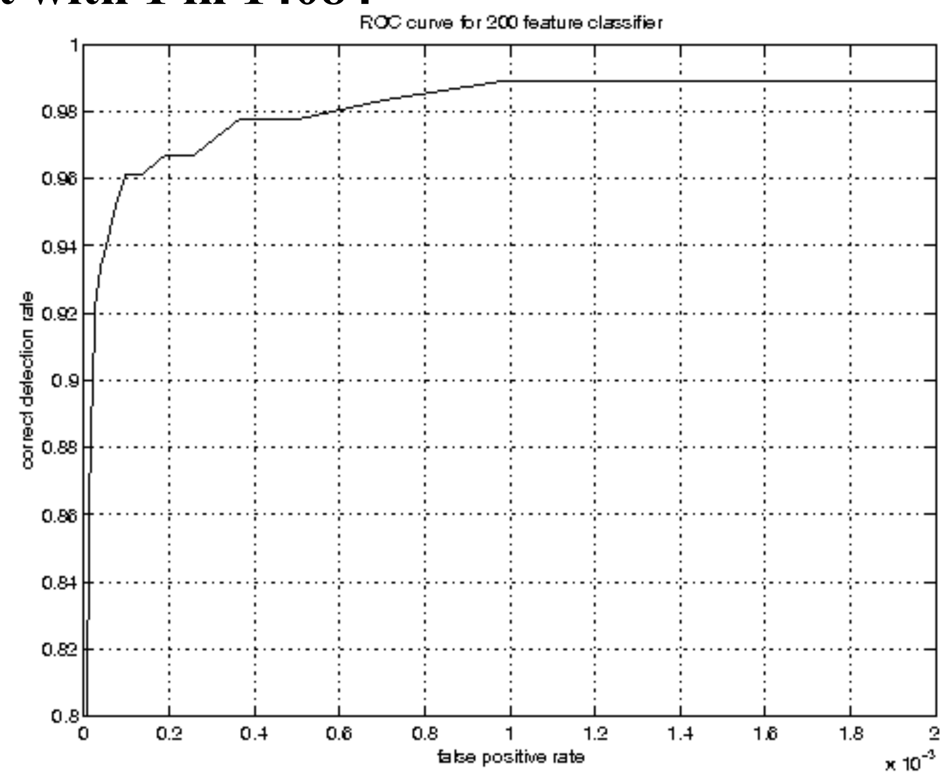
$60{,}000 \times 100 = 6{,}000{,}000$
Unique Binary Features

Example Classifier for Face Detection

A classifier with 200 rectangle features was learned using AdaBoost

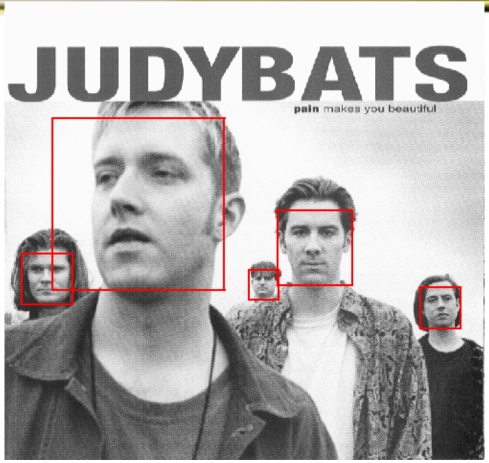
95% correct detection on test set with 1 in 14084 false positives.

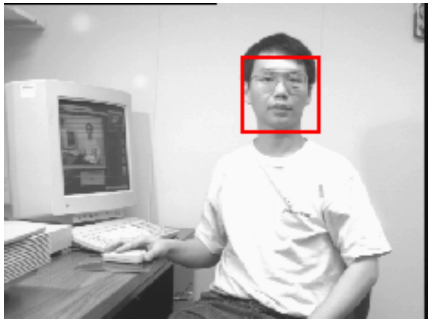
Not quite competitive...

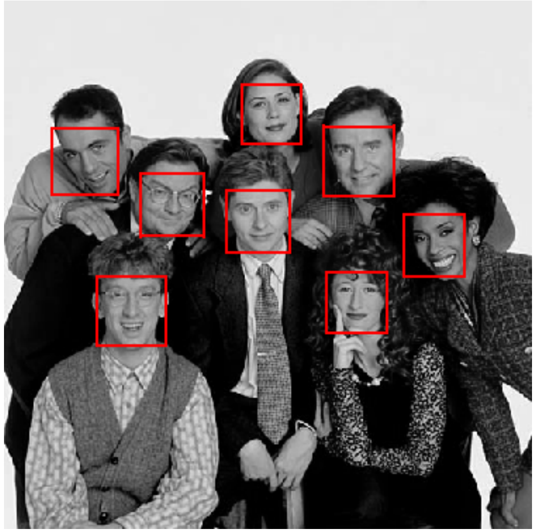


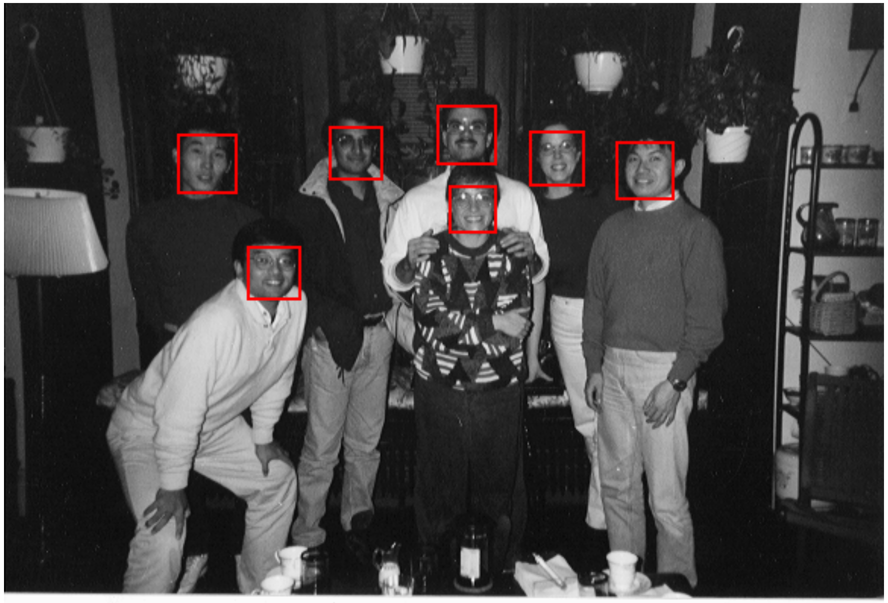
ROC curve for 200 feature classifier

Output of Face Detector on Test Images

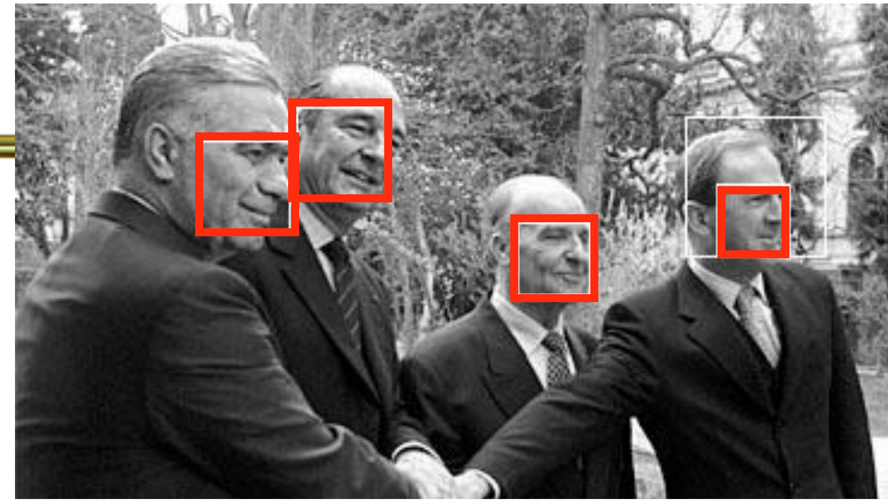








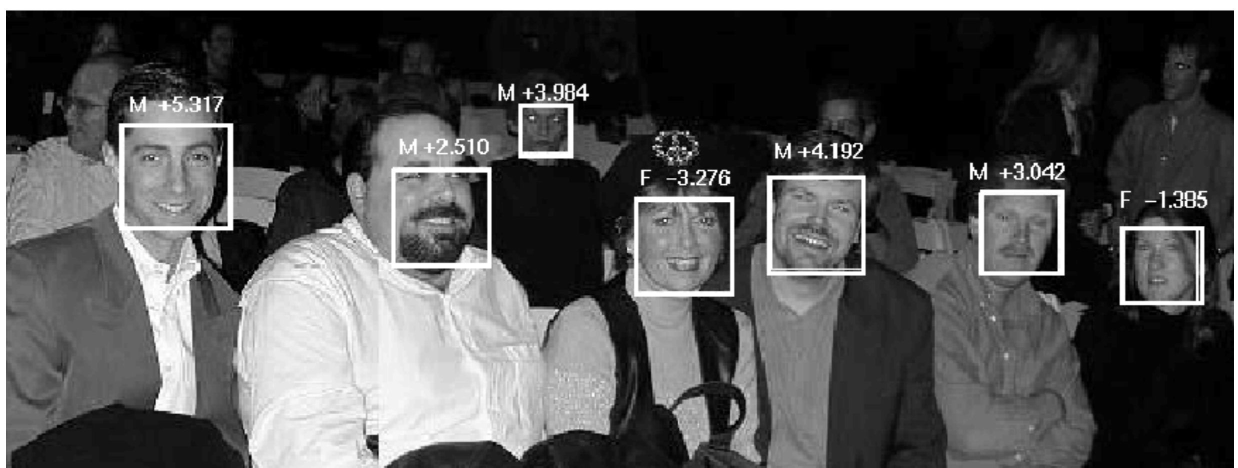
Solving other "Face" Tasks



Facial Feature Localization

Profile Detection

Demographic Analysis



Feature Localization Features