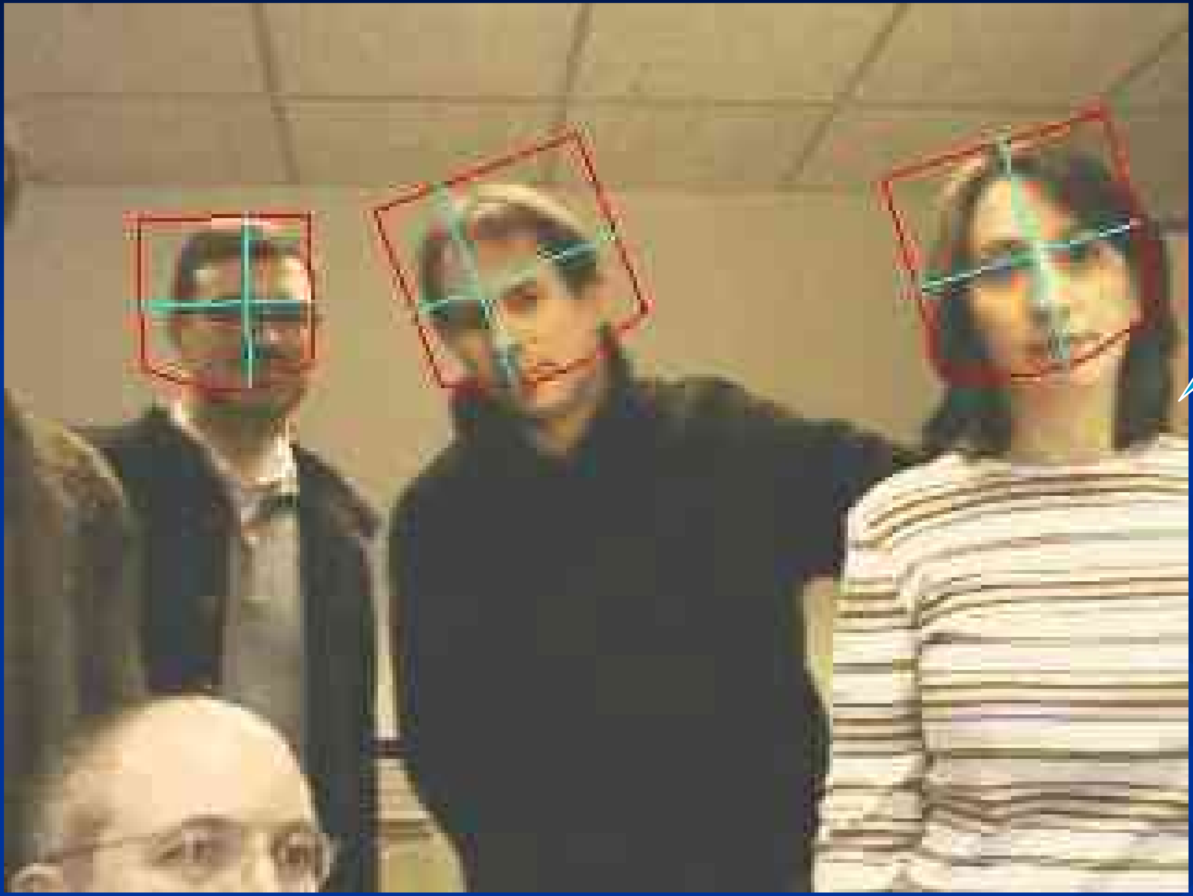
Synergistic Face Detection and Pose Estimation

M. Osadchy M.Miller Y. LeCun

Technion NEC Labs NYU

Our System



No tracking!

- Detects faces independently of their poses.
- Estimates head poses.

Our System

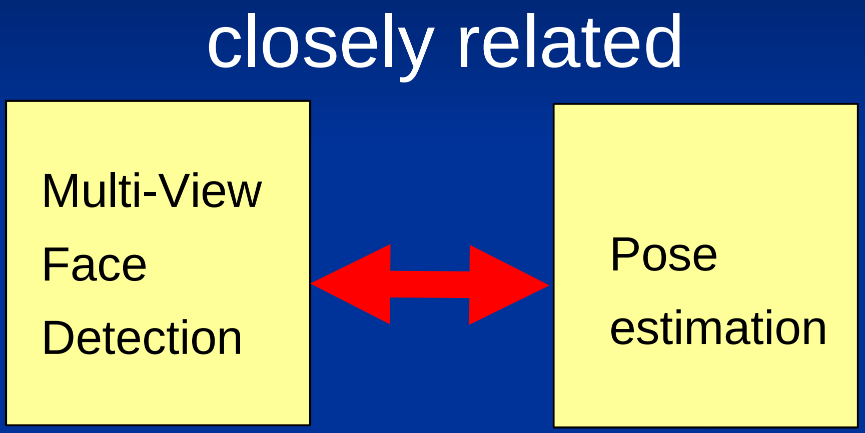
Robust to: yaw (from left to right profile), roll (-45, 45), and pitch (-60, 60).

Single Detector is applied to all poses.

Pose estimation: Within 15° error about 90% of poses are estimated correctly.

Near real-time: 5 frames per second on standard hardware.

Synergy


closely related

Multi-View Face Detection

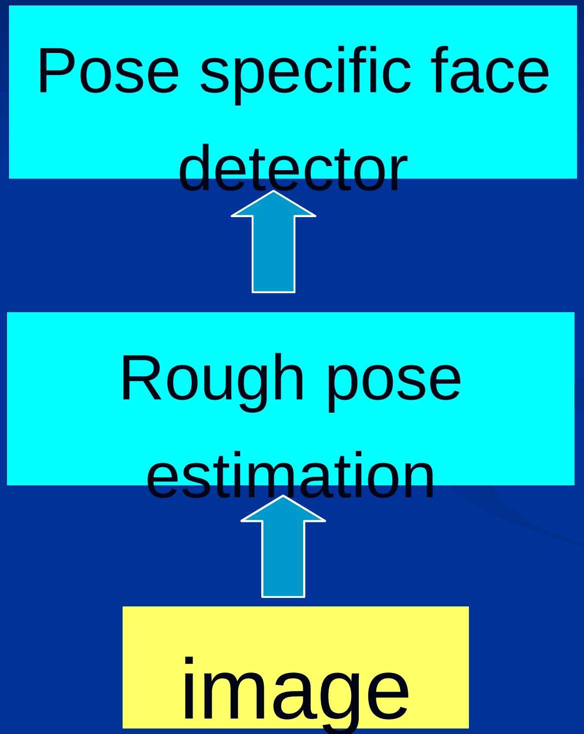
Pose estimation

Common Problems

- Inner class variation (skin color, hair style, etc.)
- Lighting Variations
- Scale Variations
- Facial Expressions
- …

Train together → Better generalization

Integrating Face Detection and Pose Estimation: Previous Methods



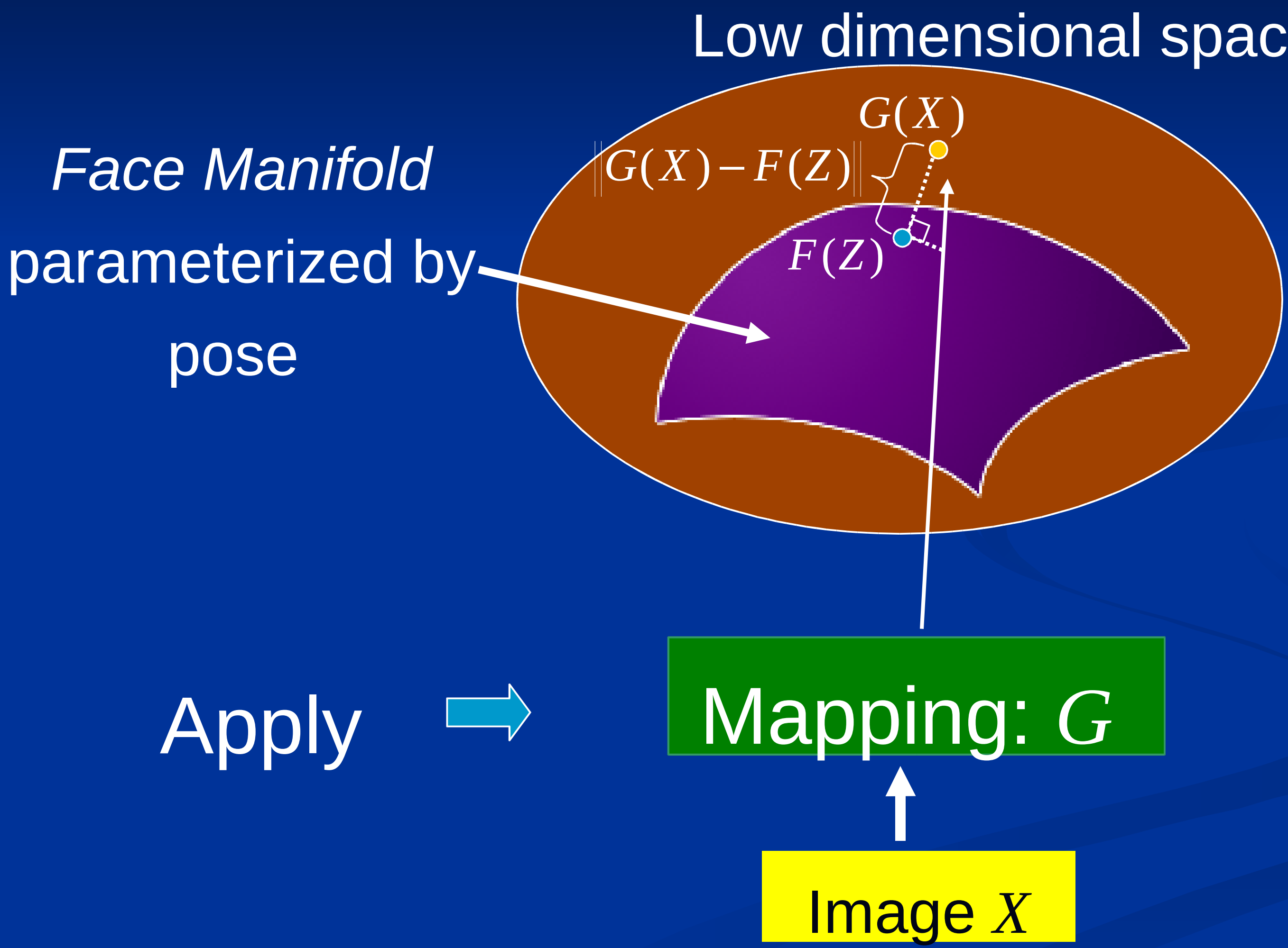
Pose specific face detector

Rough pose estimation

image

Unmanageable in real problems

Integrating Face Detection and Pose Estimation: Our Approach



Low dimensional space

$G(X)$

$\|G(X)-F(Z)\|$

$F(Z)$

Face Manifold parameterized by pose

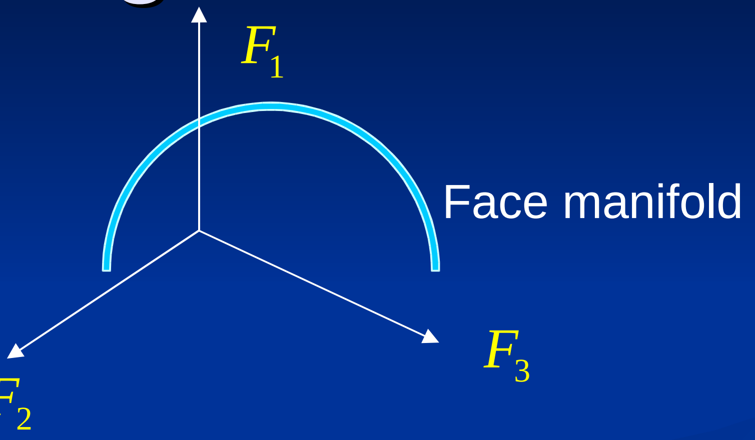
Apply

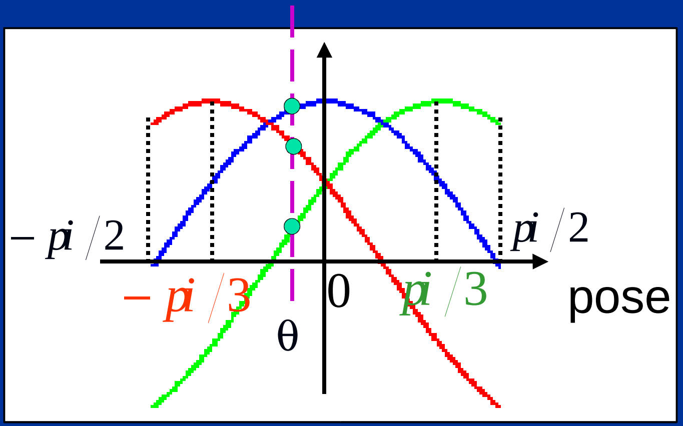
Mapping: G

Image X

Parameterization of the Face Manifold – Single Parameter

Yaw: $Z = \theta = [-pi/2,\ pi/2]$





$$F_i(\theta) = \cos(\theta - \alpha_i) \qquad \alpha = [-pi/3, 0, pi/3]$$

$$i = 1,2,3$$

$$\bar{\theta} = \arctan \frac{\sum_{i=1}^{3} G_i \cos \alpha_i}{\sum_{i=1}^{3} G_i \sin \alpha_i}$$

Parameterization of the Face Manifold – Two Parameters

Yaw and roll $Z = (\theta, \varphi)$:

$$\left.\begin{aligned}\theta &= [-pi/2,\ pi/2]\\ \varphi &= [-pi/4,\ pi/4]\end{aligned}\right\} \cong$$ a portion of the surface of a sphere

$$F_{ij}(\theta,\varphi) = \cos(\theta - \alpha_i)\cos(\varphi - \beta_j); \quad i, j = 1,2,3 \quad \alpha, \beta = [-pi/3, 0, pi/3]$$

$$\bar{\theta} = 0.5(\operatorname{atan2}(cs + sc, cc - ss) + \operatorname{atan2}(sc - cs, cc + ss))$$

$$\bar{\varphi} = 0.5(\operatorname{atan2}(cs + sc, cc - ss) - \operatorname{atan2}(sc - cs, cc + ss))$$

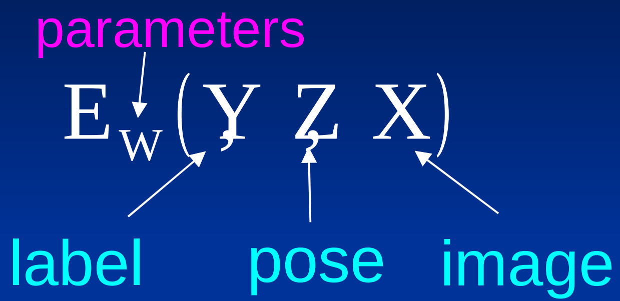
where

$$cc = \sum_{ij} G_{ij}(X)\cos(\alpha_i)\cos(\beta_j) \qquad cs = \sum_{ij} G_{ij}(X)\cos(\alpha_i)\sin(\beta_j)$$

$$ss = \sum_{ij} G_{ij}(X)\sin(\alpha_i)\sin(\beta_j) \qquad sc = \sum_{ij} G_{ij}(X)\cos(\alpha_i)\cos(\beta_j)$$

Minimum Energy Machine

- Energy function:



parameters

$E_W(Y, Z, X)$

label

pose

image

$\{Z\} = [-90, 90] \times [-45, 45]$

$$Y = \begin{cases} 1 & \text{face} \\ 0 & \text{non face} \end{cases}$$

- $E_W(Y, Z, X)$ measures compatibility between X,Z,Y.

- If X is a face with pose Z then we want:

$$E_W(1, Z, X) < E_W(0, Z, X), \quad \forall Z' \nleq$$

$$E_W(1, Z, X) < E_W(1, Z, X), \quad \forall Z' \nleq Z$$

Operating the Machine

- Clamp X to the observed value (the image)
- Find Z and Y such that:

$$\left(\bar{Y},\bar{Z}\right) = \underset{Y\in\{Y\},Z\in\{Z\}}{\arg\min}\ E_W\left(Y, Z, X\right)$$

- Complete energy:

$$E_W\left(Y, Z, X\right) = Y\cdot\left\|G_W(X) - F(Z)\right\| + \xcancel{\left(1-Y\right)}\cdot T$$

X is a face → Y=1

Operating the Machine

- Clamp X to the observed value (the image)
- Find Z and Y such that:

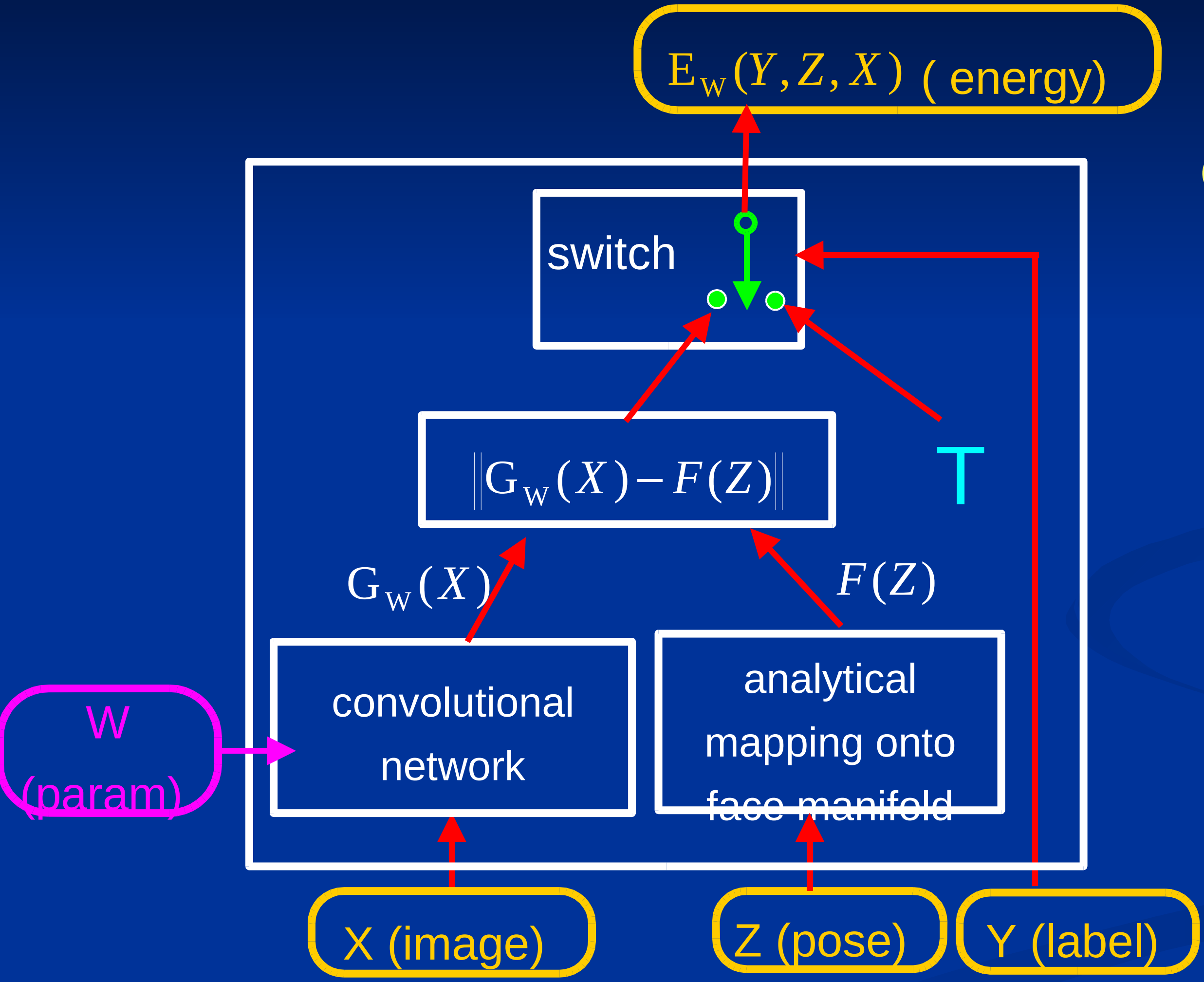
$$\left(\overline{\mathrm{Y}}, \overline{\mathrm{Z}}\right) = \arg\min_{\mathrm{Y}\in\{\mathrm{Y}\}, \mathrm{Z}\in\{\mathrm{Z}\}} \mathrm{E}_{\mathrm{W}}\left(\mathrm{Y}, \mathrm{Z}, \mathrm{X}\right)$$

- Complete energy:

$$\mathrm{E}_{\mathrm{W}}\left(\mathrm{Y}, \mathrm{Z}, \mathrm{X}\right) = \mathrm{Y}\cdot \left\|\cancel{G_W(X) - F(Z)}\right\| + \left(1-\mathrm{Y}\right)\cdot \mathrm{T}$$

X is not a face → Y=0

Architecture



Operating the machine:

$$\overline{Z} = \arg\min_{Z\in\{Z\}} \|G_W(X) - F(Z)\|$$

$$\overline{Y} = \begin{cases} 1 & \|G_W(X) - F(Z)\| < T \\ 0 & \text{otherwise} \end{cases}$$

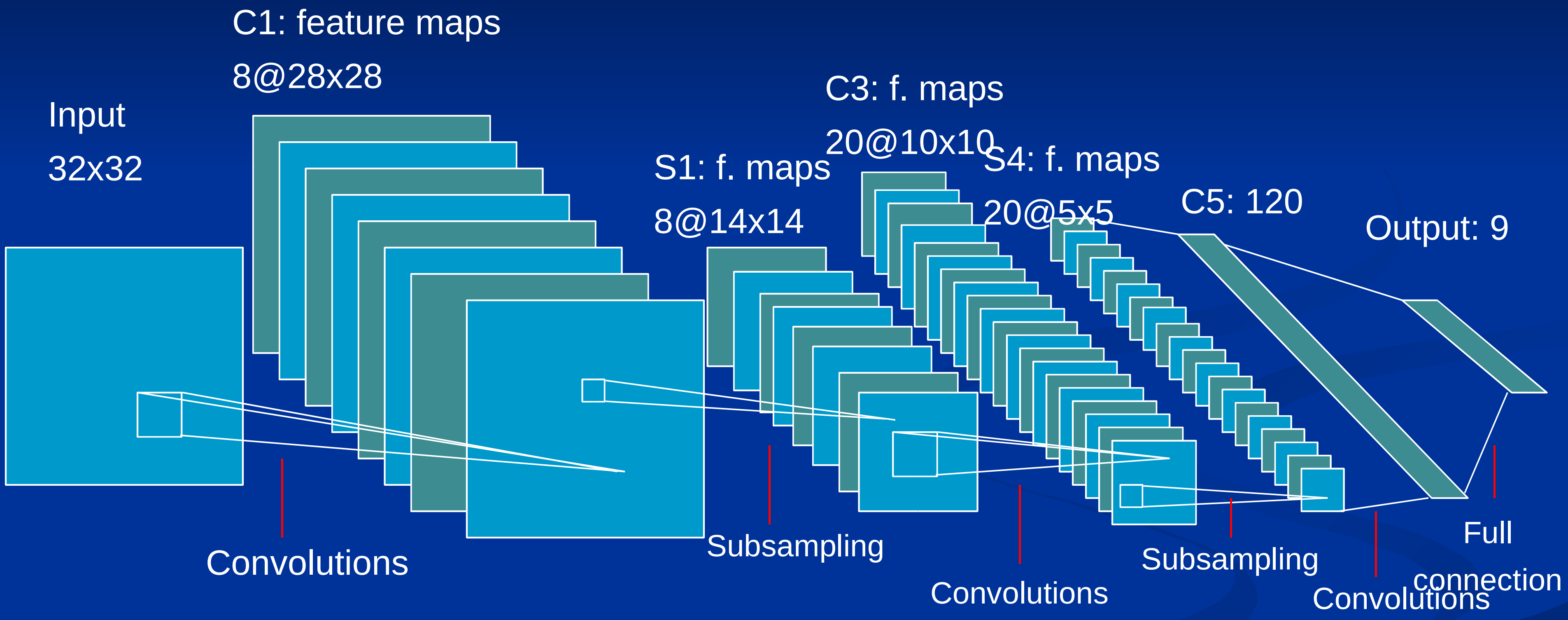
Convolutional Network

- “end-to-end” trainable systems from low-level features to high-level representations.
- Easily learn the type of shift-invariant features, relevant to object recognition.
- Can be replicated over large images much more efficiently than traditional classifiers.

↓

Considerable advantage for
real-time systems!

Similar to LeNet5, with more maps:



Training with Discriminative Loss Function

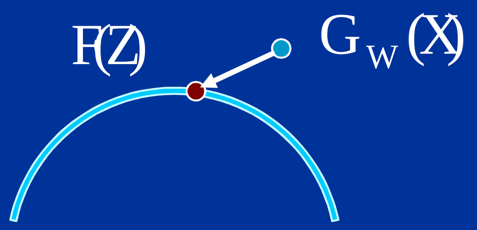
Minimize: $$L(W)=\frac{1}{|S_1|}\sum_{i\in S_1}\overbrace{L_1\left(W,Z^i,X^i\right)}^{\text{loss for face sample with known pose}}+\frac{1}{|S_0|}\sum_{i\in S_0}\overbrace{L_0\left(W,X^i\right)}^{\text{loss for non-face sample}}$$

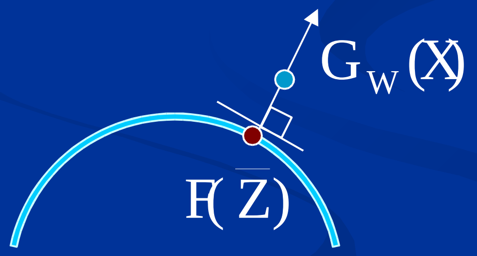
training faces

training non-faces

$$L_1(W,1,Z,X)=E_W(1,Z,X)^2$$

$$L_0(W,0,X)=K\exp\left[-E(1,\bar{Z},X)\right]$$





We showed that this loss function causes the machine to exhibit proper behavior:

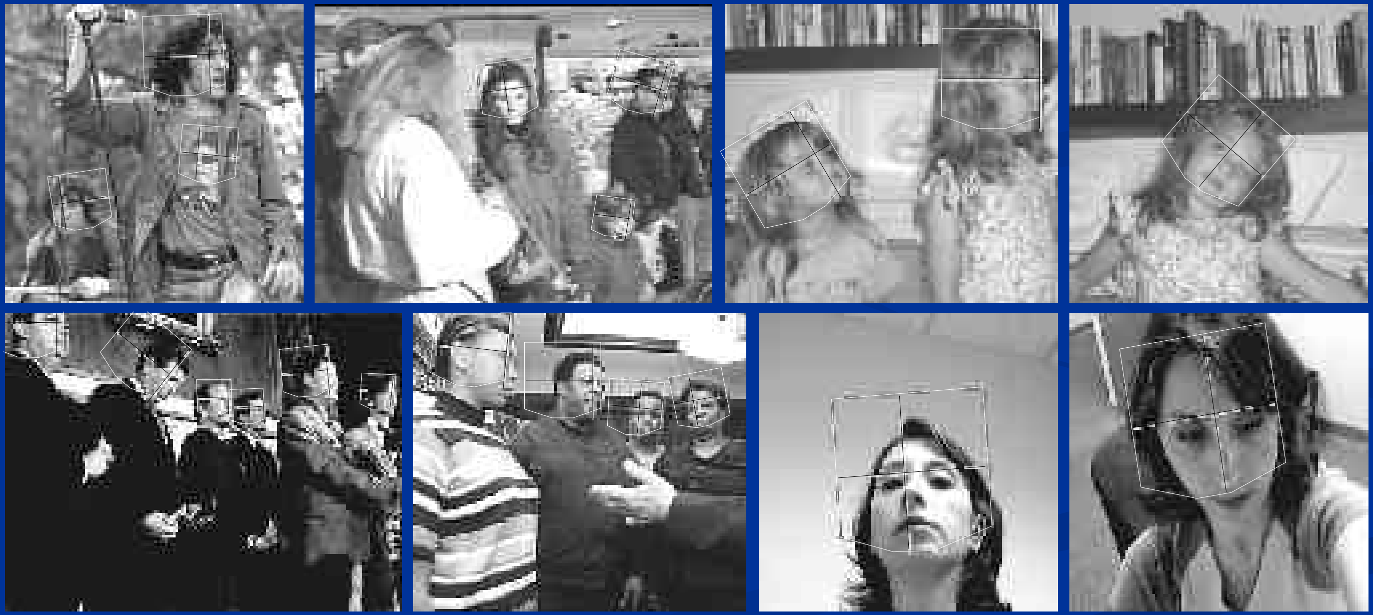
$$\mathrm{E}\left(\mathrm{Y}^{\text{desired}},\ldots\right)<\mathrm{E}\left(\mathrm{Y}^{\text{undesired}},\ldots\right)+\text{margin}$$

Running the Machine

- Works on grey-level images.
- Applied at range of scales stepping by a factor of $\sqrt{2}$.
- The network is replicated over the image at each scale, stepping by 4 pixels in x and y.
- Overlapping detections are replaced by the strongest.

Results

- Our system is robust to yaw ± 90, in-plane rotation , and pitch ± 45 ± 60



Training

- 52,850, 32x32 grey-level images of faces (NEC Labs hand annotated set) with uniform distribution of poses.
- Initial negative set: 52,850 random non-face natural images.
- Second phase: half of the initial negative set was replaced by false positives of the initial version of the detector.
- Each training image was used 5 times with random variation in scale, in-plane rotation, brightness and contrast.
- 9 passes on the data: 26 hours on 2Ghz Pentium 4.
- The system converged to an EER of 5% on training set and 6% on test set of 90,000 images.

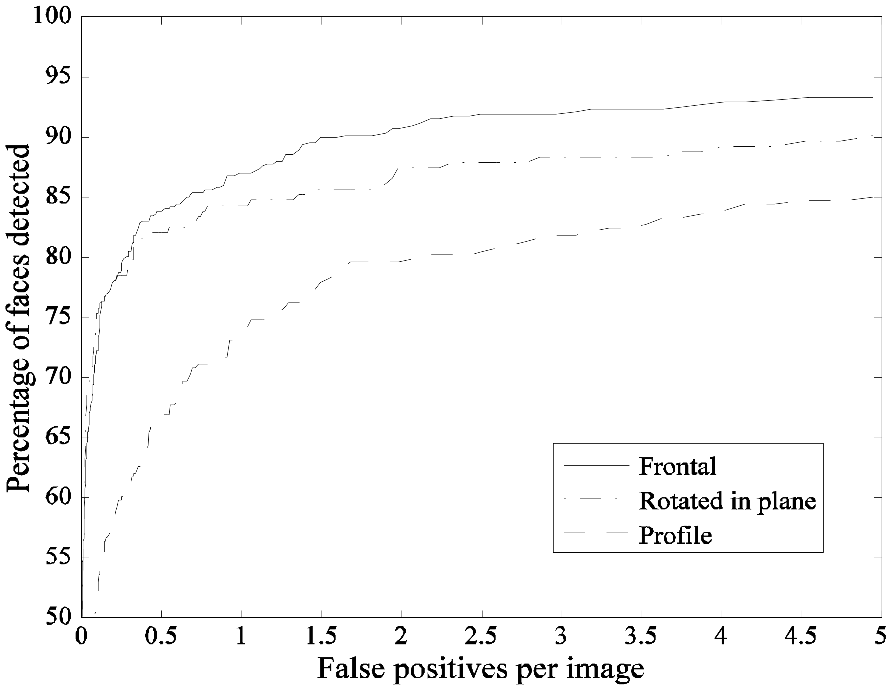
Test on Standard Data Sets

- No standard set tests all poses, that our system is designed to detect.
- 3 standard sets focusing on particular pose variation: tilted, profile, and frontal.

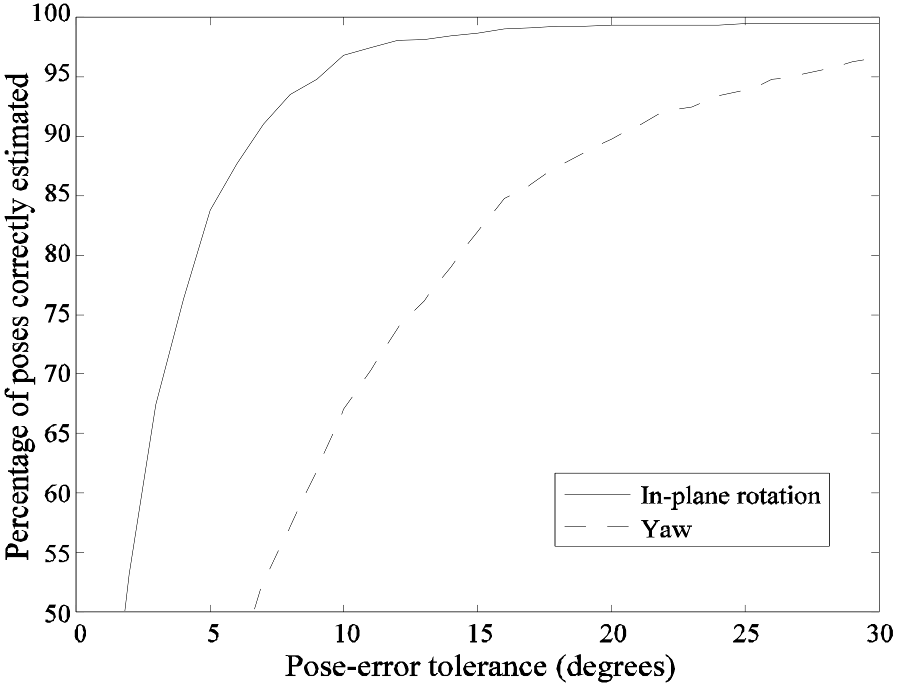
	Data Set->	TILTED		PROFILE		MIT+CMU	
	False positives per image->	4.42	26.9	0.47	3.36	0.5	1.28
Real time	**Our Detector**	**90%**	**97%**	**67%**	**83%**	**83%**	**88%**
	Jones & Viola (tilted)	**90%**	**95%**	x		x	
	Jones & Viola (profile)	x		**70%**	**83%**	x	
	Rowley *et al*	89%	96%			x	
	Schneiderman & Kanade			86%	93%	x	

Standard Sets

Detection



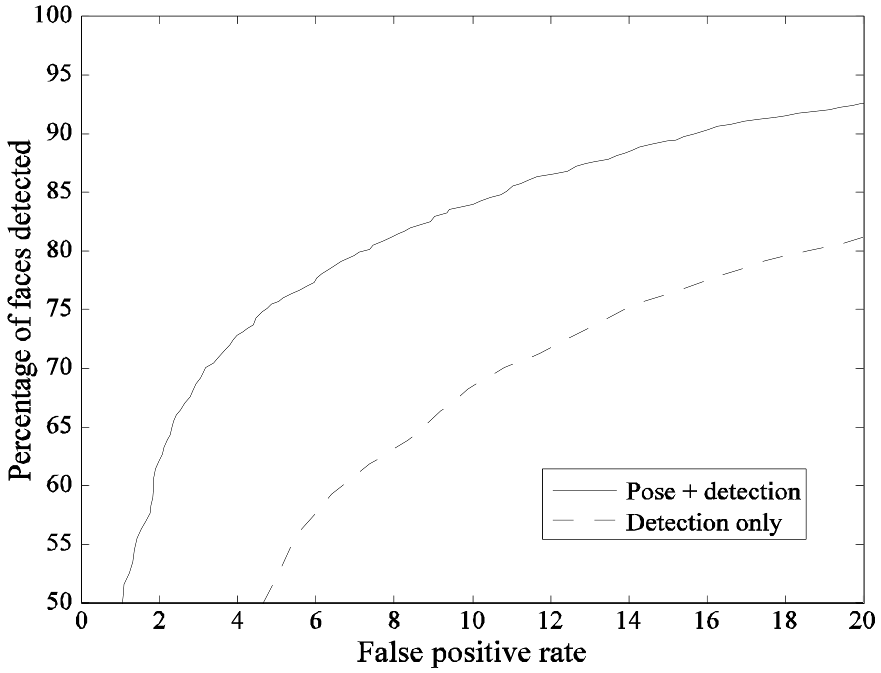
Pose Estimation of the detected faces



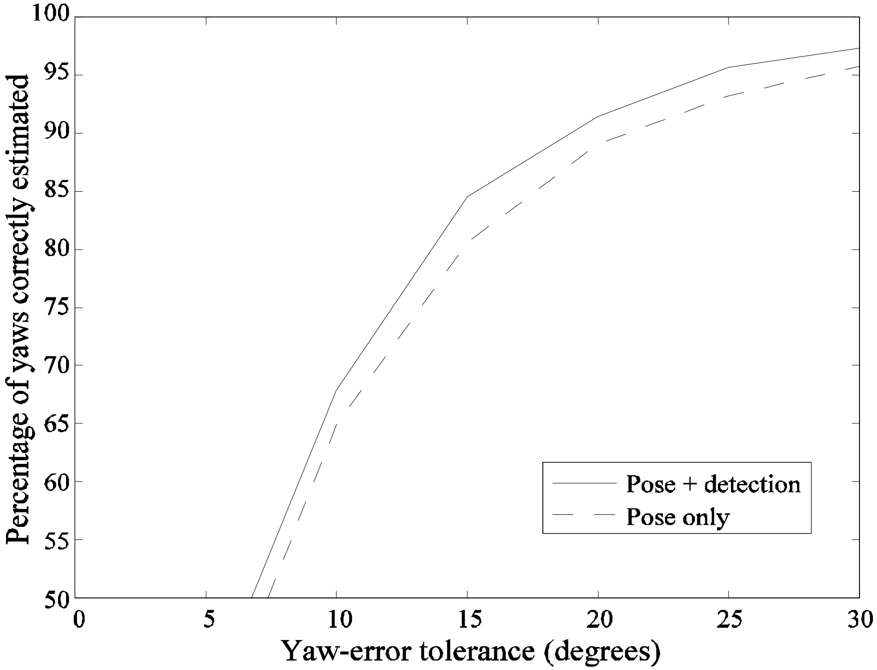
Note: typical pose estimation systems input centered faces; when we hand localize this faces we get: 89% of yaw and 100% of in-plane rotations within 15 degrees.

Synergy Test

Detection



Pose Estimation


Percentage of yaws correctly estimated
Yaw-error tolerance (degrees)
Pose + detection
Pose only