Synergistic Face Detection and Pose Estimation with Energy-Based Models

M. Osadchy (Technion) M.Miller (NEC Labs) Y. LeCun (NYU)

We developed a method for

Simultaneous face detection and pose estimation.
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.

Integrating face detection and pose estimation

Synergy

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

Previous Methods

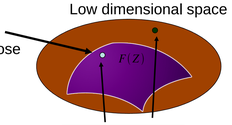
Pose specific face detector
⇧
Rough pose estimation
⇧
image

Unmanageable in real problems

Our Approach

Training:

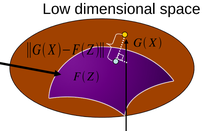
Face Manifold parameterized by pose — Low dimensional space



Train ⇨ Mapping: G

Applying:

Face Manifold parameterized by pose — Low dimensional space



Apply ⇨ Mapping: G ⇨ Image X

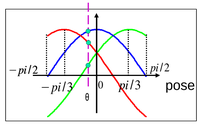
Parameterization of the face space

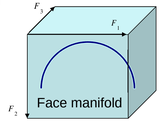
Single pose parameter

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

$F_i(\theta)=\cos(\theta-\alpha_i)$ $i=1,2,3$

$\alpha=[-pi/3, 0, pi/3]$





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

More Pose Parameters

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

$\theta=[-pi/2, pi/2]$
$\phi=[-pi/4, pi/4]$
} ∈ a portion of the surface of a sphere

$F_{ij}(\theta,\phi)=\cos(\theta-\alpha_i)\cos(\phi-\beta_j)$; $i,j=1,2,3$

$\alpha,\beta=[-pi/3, 0, pi/3]$

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

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

Learning Machine

Minimum Energy Machine

Energy function: $E_W(Y,Z,X)$ (parameters: label, pose, image)

measures compatibility between X,Z,Y.

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

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

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

$E_W(1,Z,X)<E_W(1,\bar{Z},X)$, $\forall\bar{Z}$;
$E_W(1,Z,X)<E_W(0,Z,X)$, $\forall\bar{Z}\neq Z$

Operating the Machine

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

$$(\hat{Y},\hat{Z})=\underset{Y\in[Y],Z\in[Z]}{\operatorname{argmin}} E_W(Y,Z,X)$$

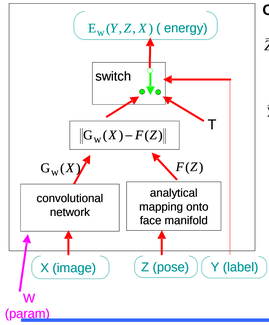
Complete energy:

$$E_W(Y,Z,X)=Y\cdot\|G_W(X)-F(Z)\|+(1-Y)\cdot T$$

X is a face, Y=1 ⇨ $E_W(1,Z,X)=\|G_W(X)-F(Z)\|$

X is not a face, Y=0 ⇨ $E_W(0,Z,X)=T$

Architecture

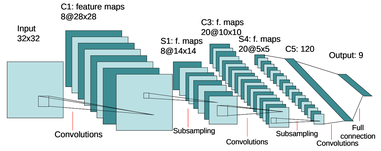


Operating the machine:

$$\hat{Z}=\underset{Z\in[Z]}{\operatorname{argmin}}\|G_W(X)-F(Z)\|$$

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

Convolutional Network:



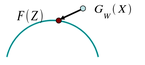
"end-to-end" trainable systems from low-level features to high-level representations.
Easily learn the type of shift-invariant features, relevant to object recognition.
Can be replicated over large images much more efficiently than traditional classifiers.

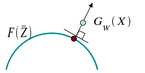
Discriminative Loss Function

Minimize: $L(W)=\frac{1}{|S_1|}\sum_{i\in S_1} L_1(W,Z^i,X^i)+\frac{1}{|S_0|}\sum_{i\in S_0} L_0(W,X^i)$

loss for face sample with known pose — loss for non-face sample
training faces — training non-faces

$L_1(W,1,Z,X)=E_W(1,Z,X)^2$ $L_0(W,0,X)=K\exp[-E(1,\bar{Z},X)]$





We showed that this loss function causes the machine to exhibit proper behavior: $E(Y^{\text{desired}},\ldots)<E(Y^{\text{undesired}},\ldots)+\text{margin}$

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

Training

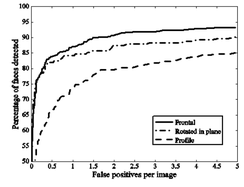
"52,850, 32x32 grey-level images of faces (NEC Labs hand annotated set) with uniform distribution of poses.
Initial negative set: 52,850 random non-face natural images.
Second phase: half of the initial negative set was replaced by false positives of the initial version of the detector.
Each training 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

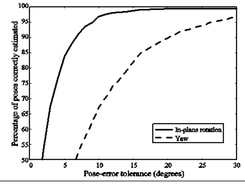
No standard set tests all poses, that our system is designed to detect.
3 standard sets focusing on particular pose variation: tilted, profile, and frontal.

Detection:

	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	



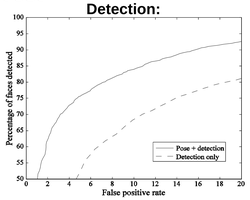
Pose Estimation:



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: