Development in Object Detection

Junyuan Lin

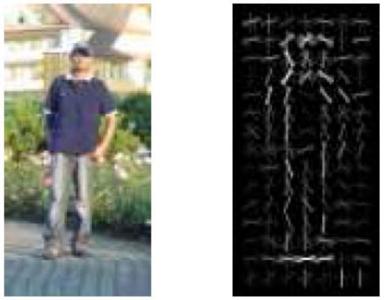
May 4th

Line of Research

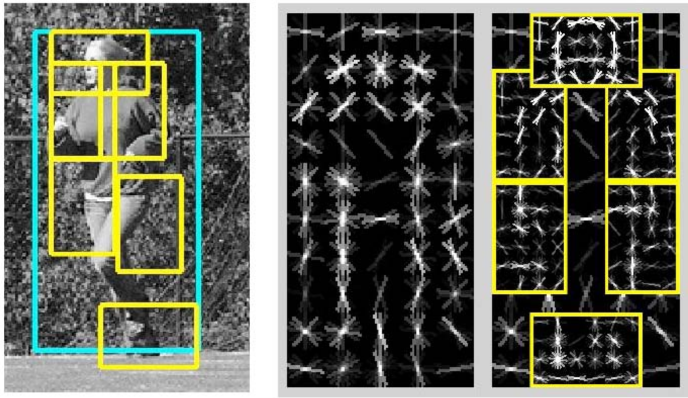
[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection, CVPR 2005.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model, CVPR 2008.

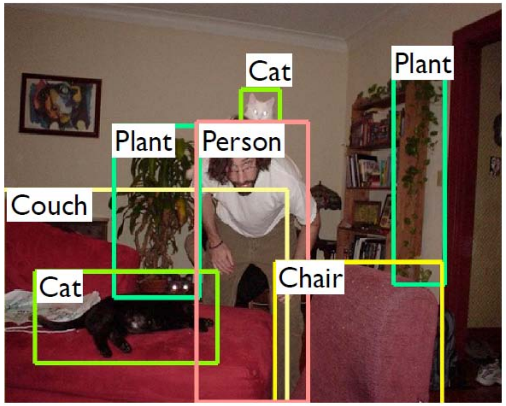
[3] C. Desai, D. Ramanan, C. Fowlkes. Discriminative Models for Multi-Class Object Layout, ICCV 2009.



HOG Feature template



Deformable Part Model



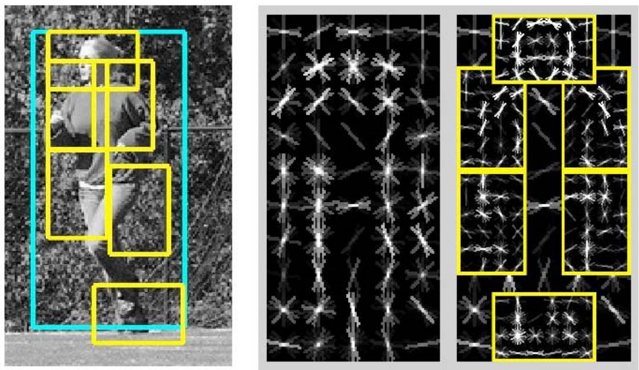
Multi-class Object Relationship

Max-margin Formulation

HOG Feature template

Linear SVM

$$f(x) = w^T \Phi(x)$$



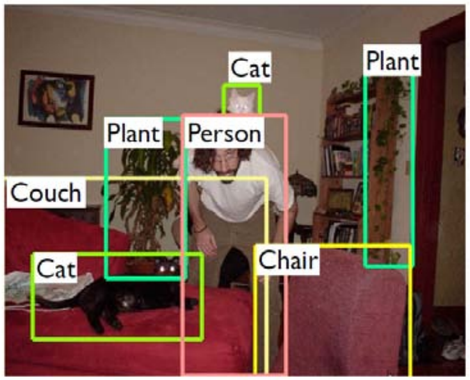
Deformable Part Model

Latent SVM

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$$

β are model parameters

z are latent values



Multi-class Object Relationship

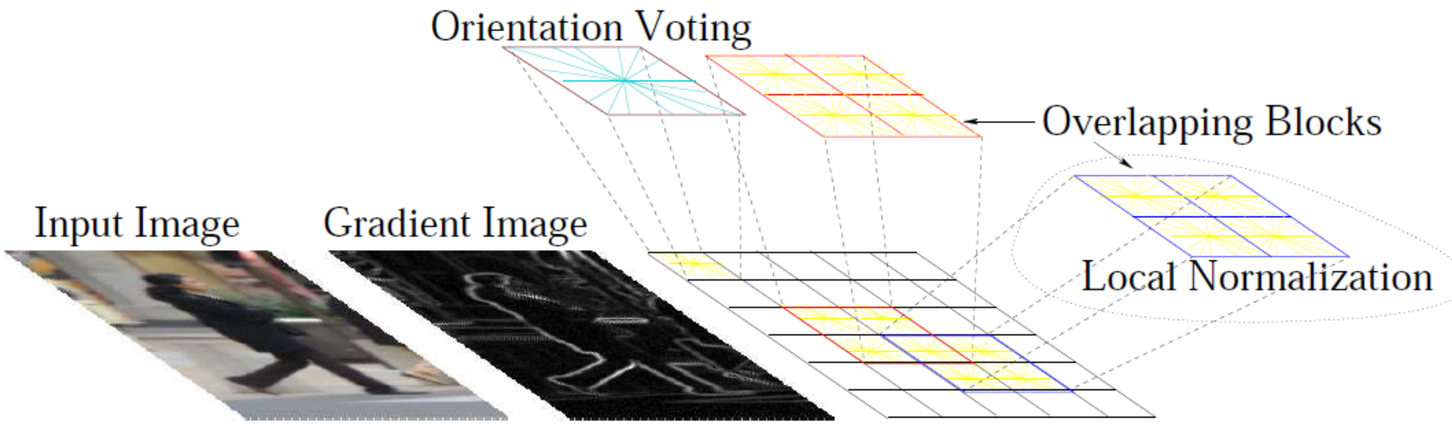
Structural SVM

$$f(x) = argmax_{y \in Y} w^T \Phi(x, y)$$

Y structured output

$$Y = \{y_i : i = 1 \ldots M\}$$

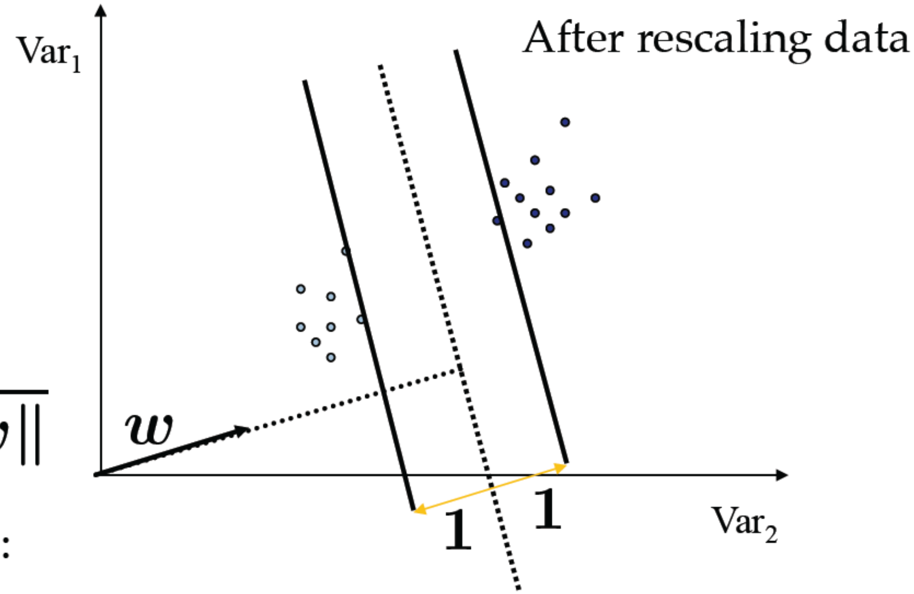
Dalal & Triggs Detector



- Concatenate HOG Features in a search window into feature vector $\Phi(x)$
- Formulate the object detection as a binary linear classifier problem
  $f(x) = w^T \Phi(x)$
- Sliding window scanning over the HOG feature pyramid
- Output location with highest score f(x)
- Possible post process (Non-maxima suppression)

Linear SVM classifier

$f(x) = w^T \Phi(x)$



Problem:

$$\max_{w} \frac{2}{||w||}$$

subject to:

$$w.x + b \geq 1, \ \forall x \text{ of class } 1$$

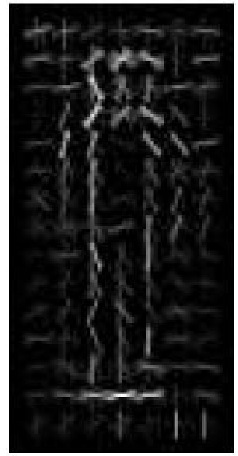
$$w.x + b \leq -1, \ \forall x \text{ of class -1}$$

minimize $\quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$

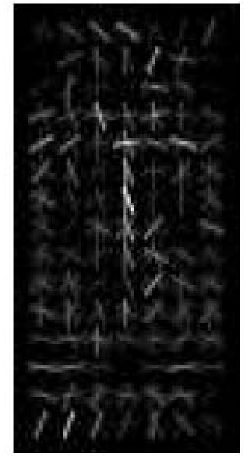
subject to $\quad y_i(\langle w, x_i \rangle + b) - 1 + \xi_i \geq 0$ and $\xi_i \geq 0$

Learned weighted Filter

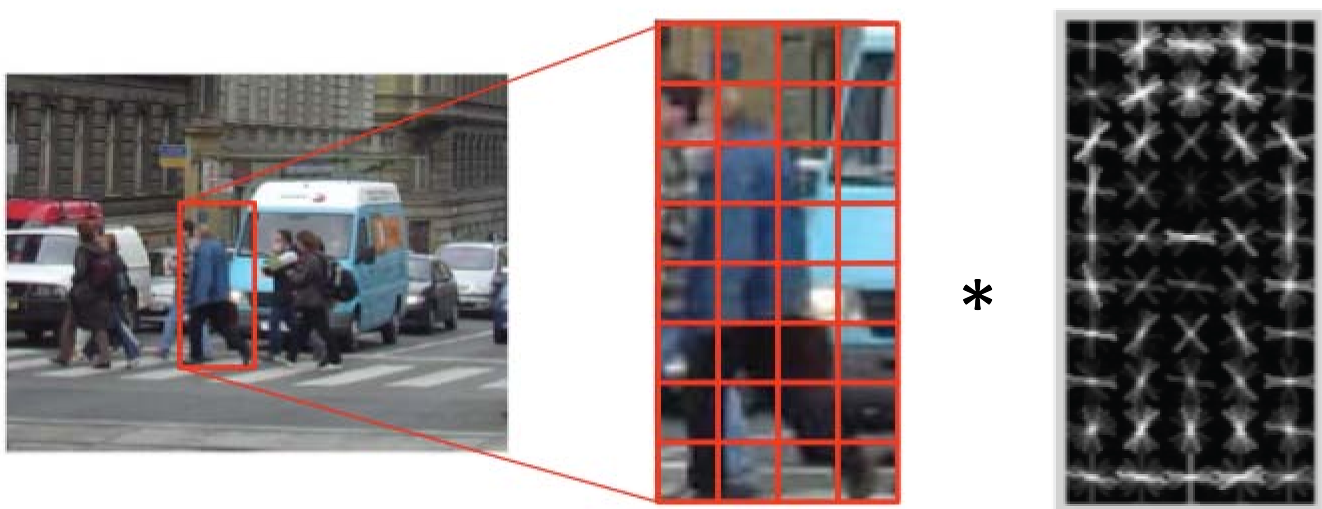
Training sample



Positive weights

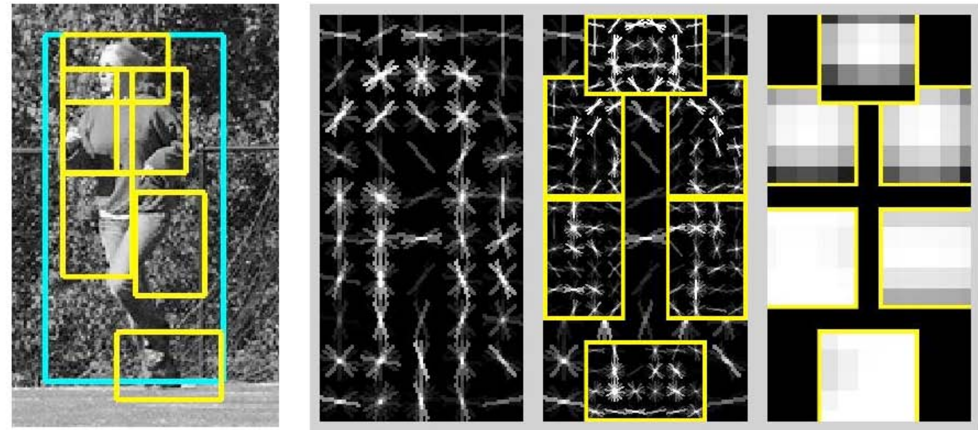


Negative weights



Deformable Part Model

Multiscale model captures features at two-resolution



detection root filter part filters deformation models

$z = (p_0, \ldots, p_n)$

p_0 : location of root

$p_1, \ldots, p_n$: location of parts relative to p_0

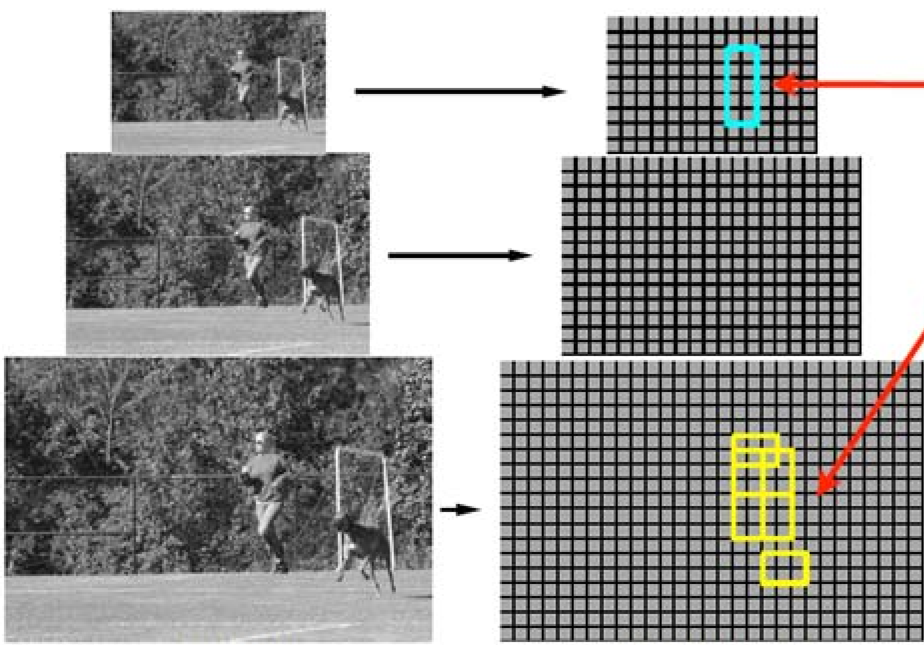


Image pyramid HOG feature pyramid

Score is sum of filter scores minus deformation costs

Slide by P. Felzenszwalb

Score of a hypothesis

"data term"

"spatial prior"

$$\text{score}(p_0, \ldots, p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2)$$

filters

deformation parameters

displacements

relative to anchor position

$$\text{score}(z) = \beta \cdot \Psi(H, z)$$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$$

Latent variable $z = (p_0, \ldots, p_n)$

Slide by P. Felzenszwalb

Latent SVM Training

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$$ Learn β

$$L_D(\beta) = \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i))$$

Hinge loss

Not convex in general !

Semi-convexity:

- $f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$ is convex in β
- $\max(0, 1 - y_i f_\beta(x_i))$ is convex for negative examples $(y_i = -1)$
- $\max(0, 1 - y_i f_\beta(x_i))$ is concave for positive examples $(y_i = +1)$

For positive examples make $L_D(\beta)$ convex by
fixing the latent variable z_p for each positive training example.

Latent SVM Training cont'd

$$L_D(\beta) = \min_{Z_p} L_D(\beta, Z_p).$$

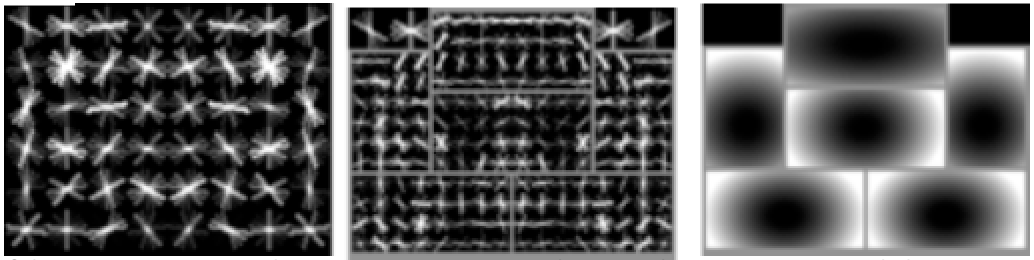
Coordinate descent optimization approach:

Initialize β and iterate:

- Fix β pick best z_p for each positive example.

$$z_i = \text{argmax}_{z \in Z(x_i)} \beta \cdot \Phi(x_i, z).$$

- Fix z_p optimize $L_D(\beta, Z_p)$ over β by quadratic programming or gradient descent



Model training procedure:

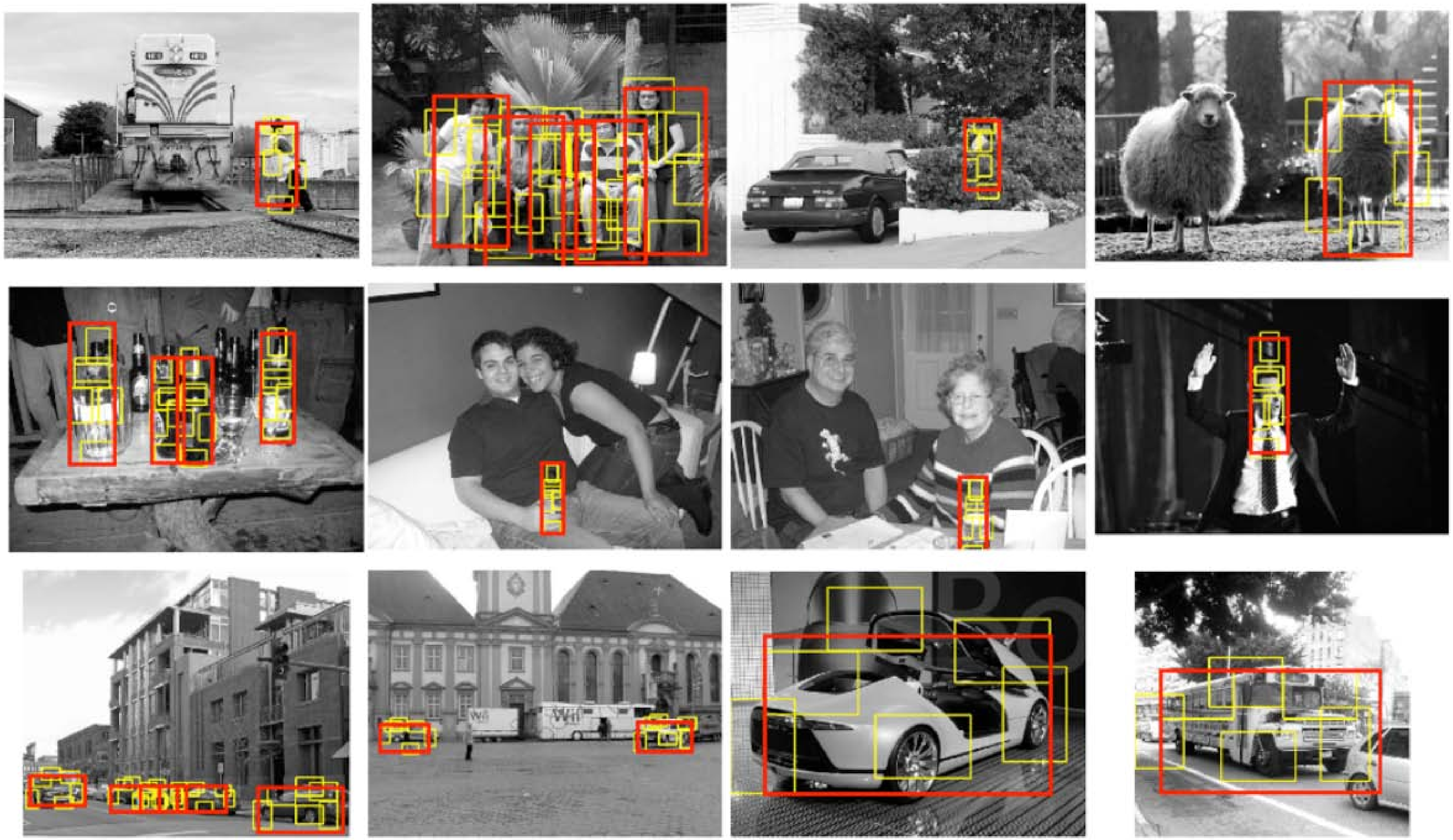
- Initialization:

  -*Root Filter :* train an initial root filter F_0 using linear SVM without latent variable.

  -*Part Filter :* Initialize six parts from root filter F_0 with the same shape, place at anchor positions with most positive energy in F_0 . The size of each part filter a is determined to take 80% of the area of F_0.

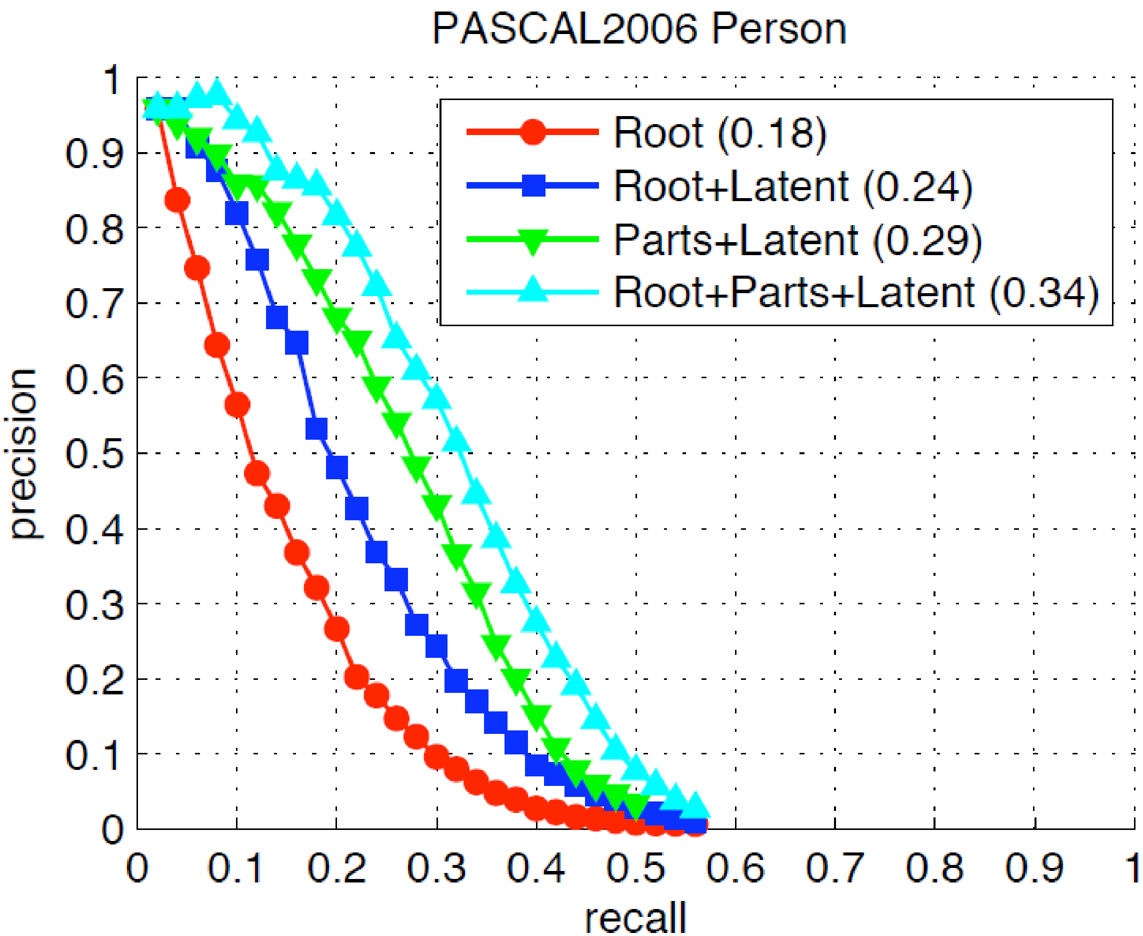
- Train the latent SVM model

Example Results



Slide by P. Felzenszwalb

Quantitative Result



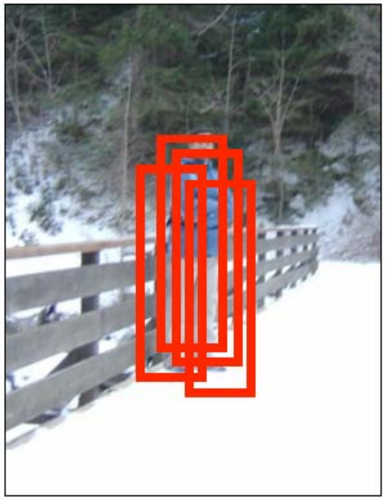
Multi-class Object Layout

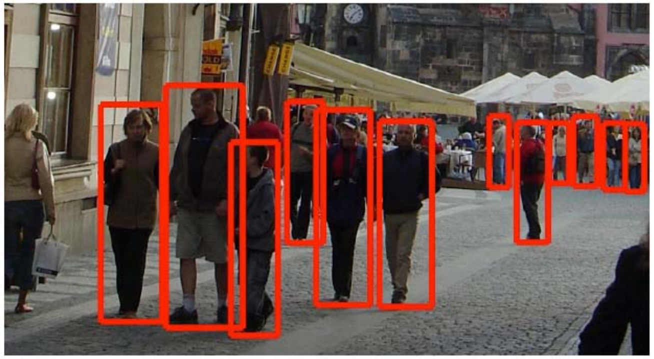
Pascal Dataset: Multiple objects of multi-class in a single image.

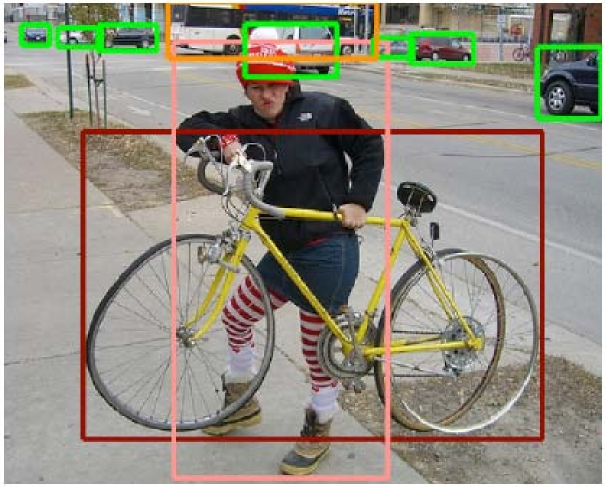
Problems with traditional binary classification + sliding window approach:

- Intra-class: Require ad-hoc non-maxima suppression as post processing.
- Inter-class: multiple class models are searched independently over images.

  Heuristically forcing mutual exclusion.





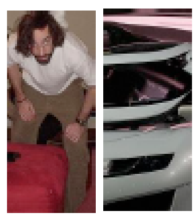


Need a principal way to model spatial statistics between objects

Formulation

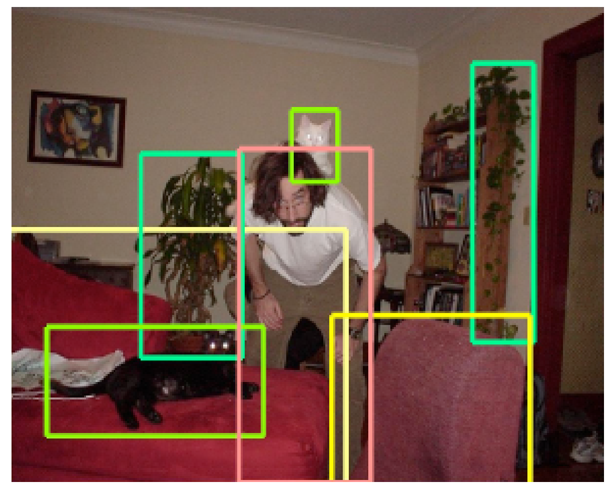
Formulate as a structured prediction problem.

Classification



x = image window
$y \in \{0,1\}$

Structured, sparse label



X = entire image
$Y = [\ldots 4 \ldots 3 \ldots 2 . 7 . . 1 . .]$

X: a collection of overlapping windows at various scales covering the entire image
Y: the entire label vector for the set of all windows in an image.
spatial relationship between prediction y_i

Global scoring function

$$S_w(X,Y) = \underbrace{\boxed{\sum_i w_{y_i}^T x_i}}_{\text{sum of per-window classifier scores}} + \underbrace{\boxed{\sum_{i,j} w_{y_i,y_j}^T d_{ij}}}_{\text{sum of pairwise window-label interactions}}$$

$X = \{x_i\}$ $Y = \{y_i\}$

x_i = feature vector extracted from i^{th} window (e.g. HOG)

y_i = class label (0..K)

w_{yi} = template for class y_i

w_{car}

w_{bus}

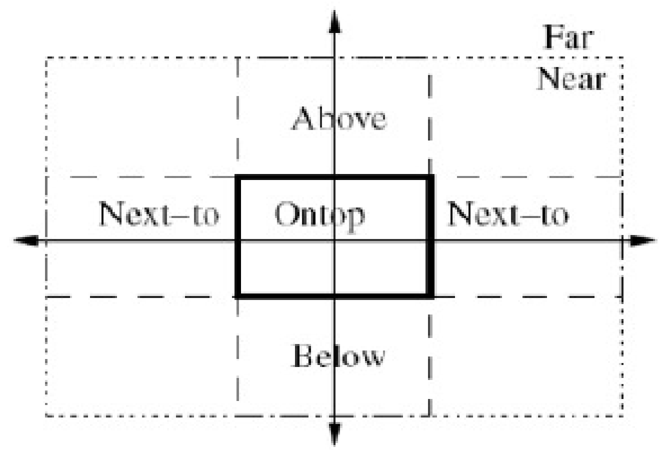
$\vdots$

d_{ij} = spatial context descriptor for window i and j

w_{yiyj} = spatial interaction model for class y_i & y_j

Output score of local detectors

The parameter learning and inference can be formulated as a structural SVM framework.



Spatial context descriptor d_{ij}

Structural SVM

$$\min_{\vec{w}_1,\ldots,\vec{w}_n,\vec{\xi}} \quad \sum_{i=1}^{k} \vec{w}_i^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

Multi-class SVM

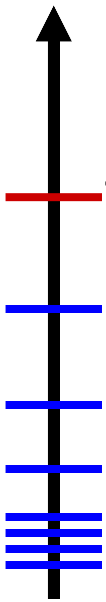
$$s.t. \quad \forall j \neq y_1 : \vec{w}_{y_1}^T \vec{x}_1 \geq \vec{w}_j^T \vec{x}_1 + 1 - \xi_1$$

$$Y = \{y_i : i = 1 \ldots M\}$$

$$\ldots$$

$$\forall j \neq y_n : \vec{w}_{y_n}^T \vec{x}_n \geq \vec{w}_j^T \vec{x}_n + 1 - \xi_n$$

$\vec{w}^T \Phi(x_n, y_n)$



(x_n, y_n)

- Exponential Parameters
- Exponential Constraints

Joint feature map $\Phi(x, y)$ represent Y using features extract from X,Y

Learn weights $\vec{w}$ so that $\vec{w}^T \Phi(x, y)$ is max for correct *y*

$$\min_{\vec{w}} \quad \frac{1}{2}\vec{w}^T \vec{w}$$

$$s.t. \quad \forall y \in Y \backslash y_1 : \vec{w}^T \Phi(x_1, y_1) \geq \vec{w}^T \Phi(x_1, y) + 1$$

$$\ldots$$

$$\forall y \in Y \backslash y_n : \vec{w}^T \Phi(x_n, y_n) \geq \vec{w}^T \Phi(x_n, y) + 1$$

Structural SVM cont'd

n-Slack Formulation: (margin rescaling)

$$\min_{\vec{w},\vec{\xi}} \quad \frac{1}{2}\vec{w}^T\vec{w} + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\begin{aligned} s.t. \quad & \forall y' \in Y : \vec{w}^T\Phi(x_1, y_1) - \vec{w}^T\Phi(x_1, y') \geq \Delta(y_1, y) - \xi_1 \\ & \ldots \\ & \forall y' \in Y : \vec{w}^T\Phi(x_n, y_n) - \vec{w}^T\Phi(x_n, y') \geq \Delta(y_n, y) - \xi_n \end{aligned}$$

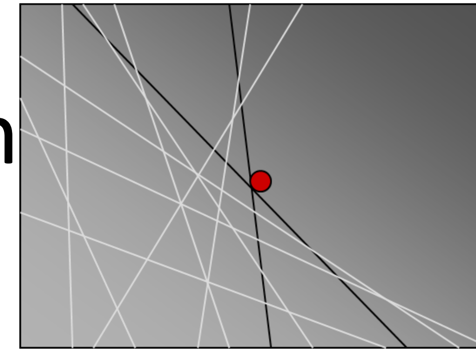
1-Slack Formulation:

$$\min_{\vec{w},\xi} \quad \frac{1}{2}\vec{w}^T\vec{w} + C\xi$$

$$s.t. \quad \forall y'_1 \ldots y'_n \in Y : \frac{1}{n}\sum_{i=1}^{n}\left[\vec{w}^T\Phi(x_i, y_i) - \vec{w}^T\Phi(x_i, y'_i)\right] \geq \frac{1}{n}\sum_{i=1}^{n}\left[\Delta(y_i, y'_i)\right] - \xi$$

Slide by T. Joachims

Cutting-Plane Algorithm



- **Input:** $(x_1, y_1), \ldots, (x_n, y_n), C, \epsilon$
- $S \leftarrow \emptyset, \vec{w} \leftarrow 0, \xi \leftarrow 0$
- **REPEAT**
  - FOR $i = 1, \ldots, n$
    - Compute $y_i' = argmax_{y \in Y}\{\Delta(y_i, y) + \vec{w}^T \Phi(x_i, y)\}$
  - ENDFOR
  - IF $\sum_{i=1}^{n} \left[\Delta(y_i, y_i') - \vec{w}^T[\Phi(x_i, y_i) - \Phi(x_i, y_i')]\right] > \xi + \epsilon$
    - $S \leftarrow S \cup \{\vec{w}^T \frac{1}{n} \sum_{i=1}^{n} [\Phi(x_i, y_i) - \Phi(x_i, y_i')] \geq \frac{1}{n} \sum_{i=1}^{n} \Delta(y_i, y_i') - \xi\}$
    - $[\vec{w}, \xi] \leftarrow$ optimize StructSVM over S
  - ENDIF
- **UNTIL** S **has not changed during iteration**

Find most violated constraint

Violated by more than ε ?

Add constraint to working set

[Jo06] [JoFinYu08]

Theoretical Bound

Theorem: Given any $\varepsilon > 0$, the number of constraints added to working set S is bounded by

$$\left\lceil \log_2\left(\frac{\Delta}{4R^2C}\right)\right\rceil + \left\lceil \frac{16R^2C}{\varepsilon} \right\rceil$$

where $0 \leq \Delta(y_i,y) \leq \Delta$

$2||\Phi(x,y)|| \leq R$

so that $(\vec{w}, \xi + \varepsilon)$ is a feasible solution.

- Number of constraints is independent of training examples.
- Linear time training algorithm.

Summary of Structural SVM

- Training: (cutting plane algorithm)

$$\min_{\vec{w},\xi} \quad \frac{1}{2}\vec{w}^T\vec{w} + C\xi$$

$$s.t. \quad \forall y'_1 \dots y'_n \in Y : \frac{1}{n}\sum_{i=1}^{n}\left[\vec{w}^T\Phi(x_i,y_i) - \vec{w}^T\Phi(x_i,y'_i)\right] \geq \frac{1}{n}\sum_{i=1}^{n}\left[\Delta(y_i,y'_i)\right] - \xi$$

- Prediction:

$$\hat{y} = argmax_{y\in Y}\{\vec{w}^T\Phi(x,y)\}$$

- Application Specific Design of Model
  - Loss function: $\Delta(y_i, y)$
  - Representation: joint feature map $\Phi(x, y)$
  - Algorithm to compute: $\hat{y} = argmax_{y\in Y}\{\vec{w}^T\Phi(x,y)\}$

$$\hat{y} = argmax_{y\in Y}\{\Delta(y_i,y) + \vec{w}^T\Phi(x_i,y)\}$$

Component Formulation

$$S_w(X,Y) = \sum_i w_{y_i}^T x_i + \sum_{i,j} w_{y_i,y_j}^T d_{ij}$$

$$S_w(X,Y) = w^T \Psi(X,Y)$$

joint feature map

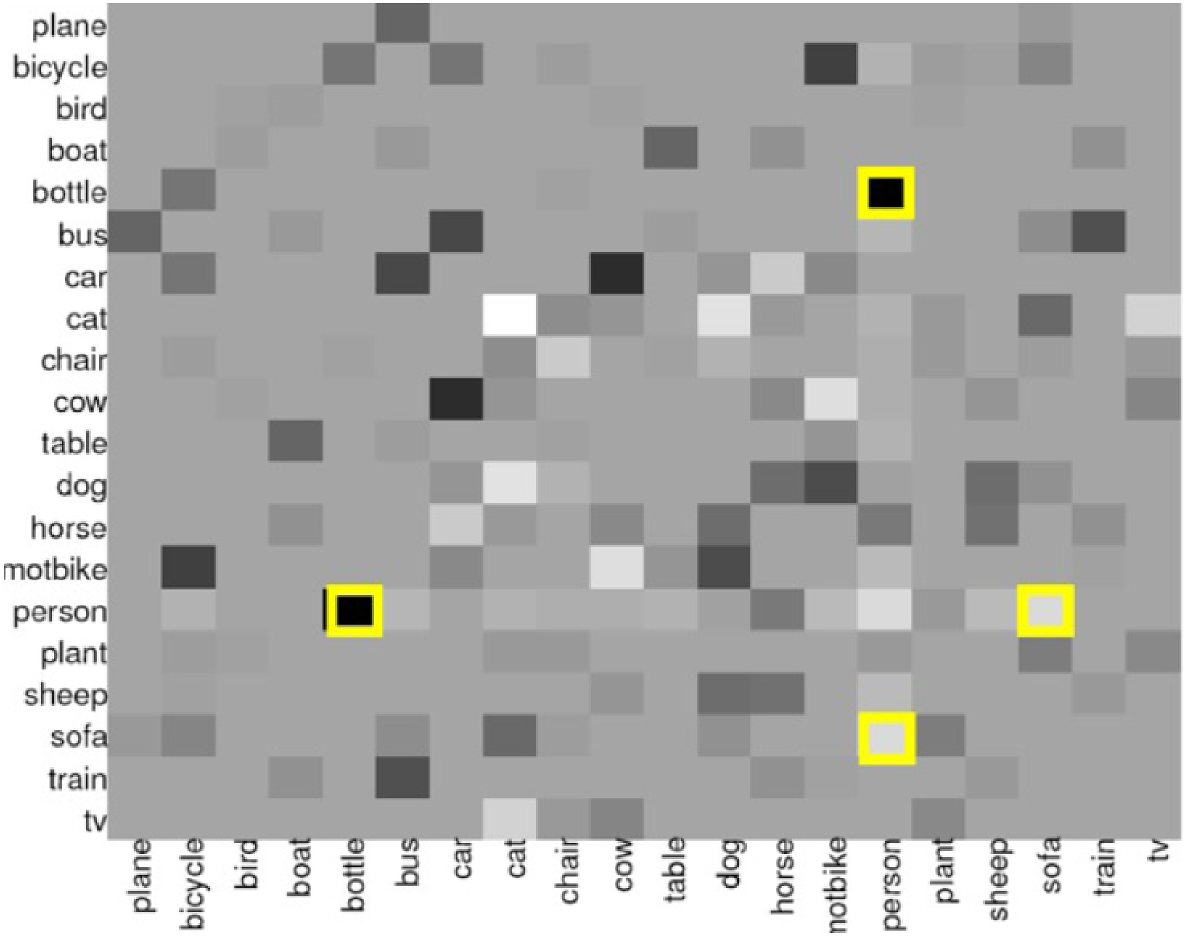
$\Delta(y_i, y)$ 0-1 loss function

$$\hat{y} = argmax_{y \in Y} S(X,Y)$$

Greedy forward search algorithm

(1) Initialize all labels to bg
Initialize per-window scores with local template
(2) Select highest scoring un-instanced window
(3) Instance it and add pairwise contribution to remaining windows
(4) Stop when remaining windows score < 0

Overlap feature in pairwise potential

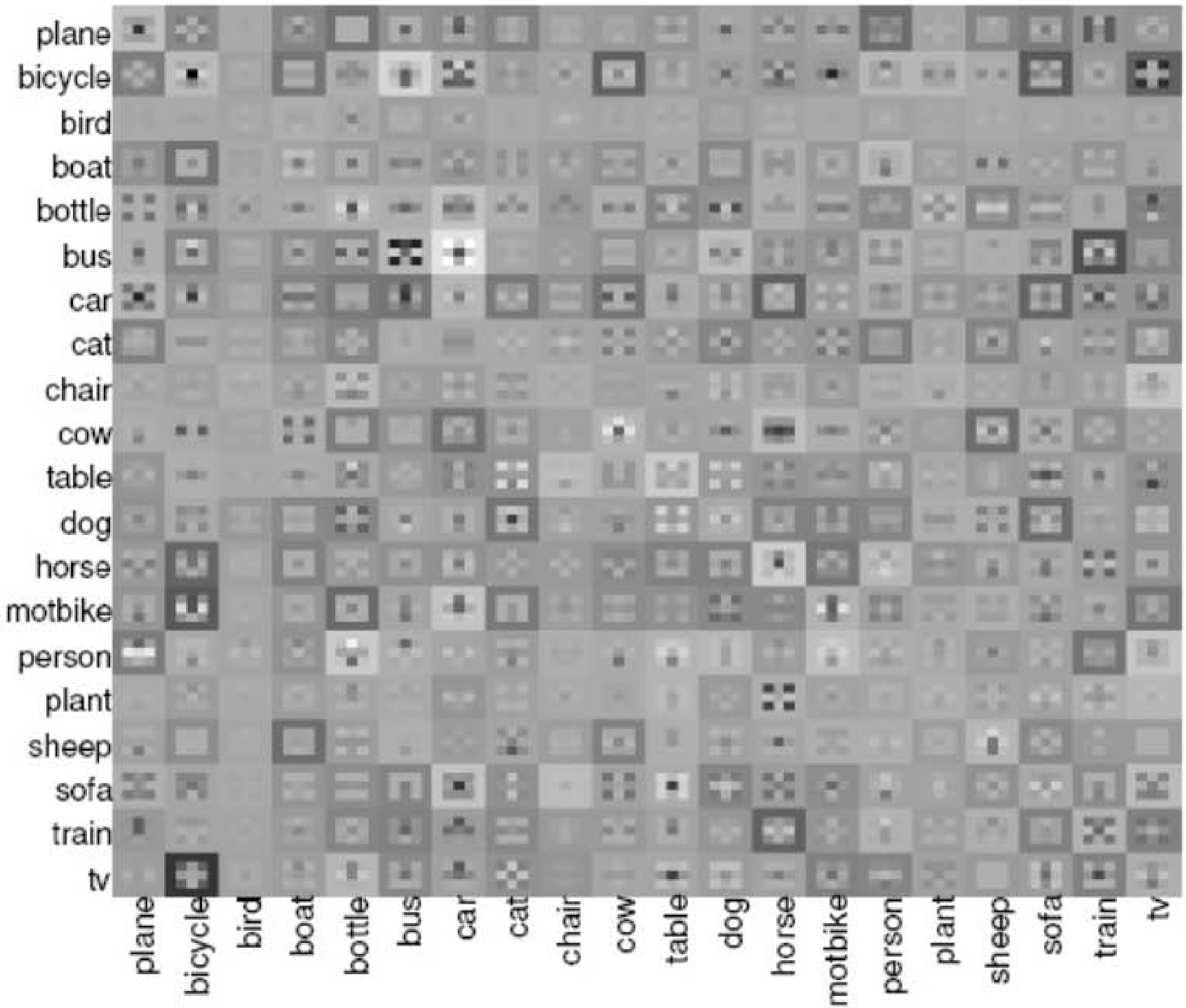


Mutual exclusion can be subtle

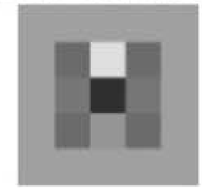
Parameters are trained with knowledge of local detectors

Slide by D. Ramanan

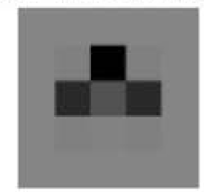
Remaining pairwise potentials



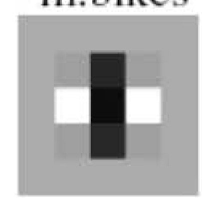
bottles wrt tables



cars wrt trains



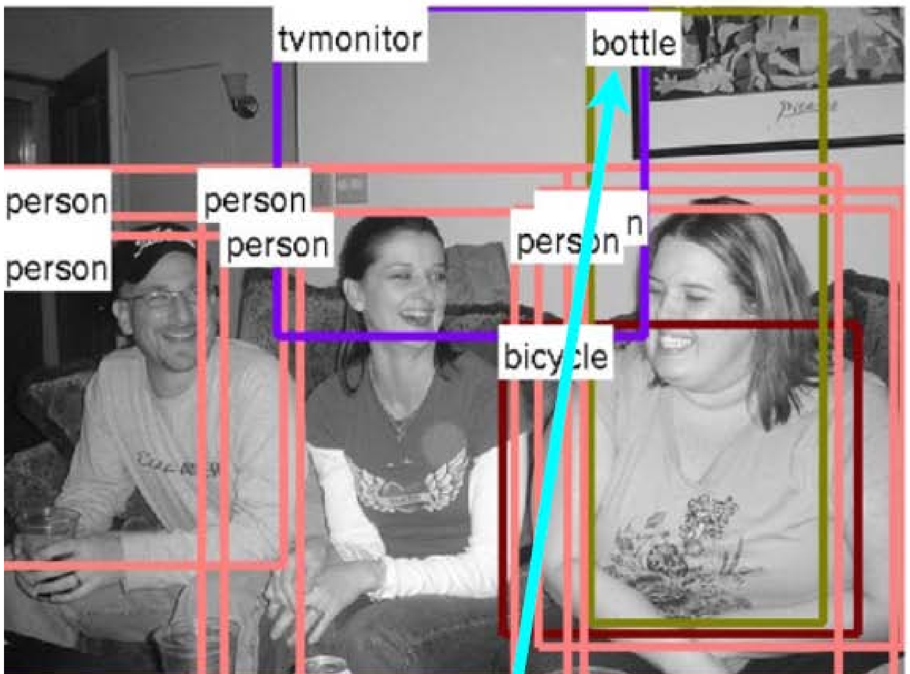
m.bikes wrt m.bikes



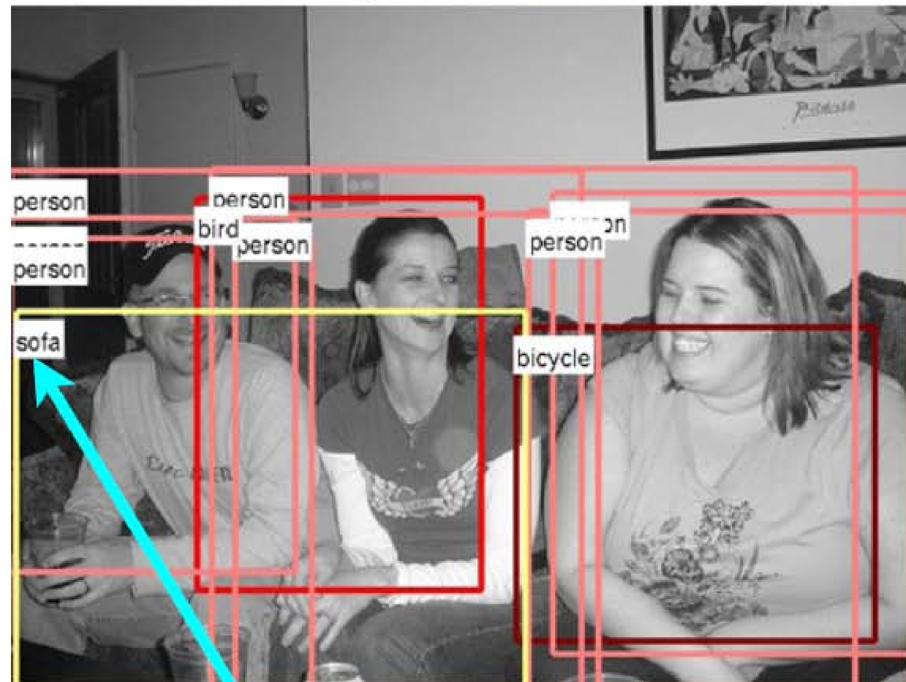
Slide by D. Ramanan

Results

Top 10 detections for baseline



Our top 10 detections



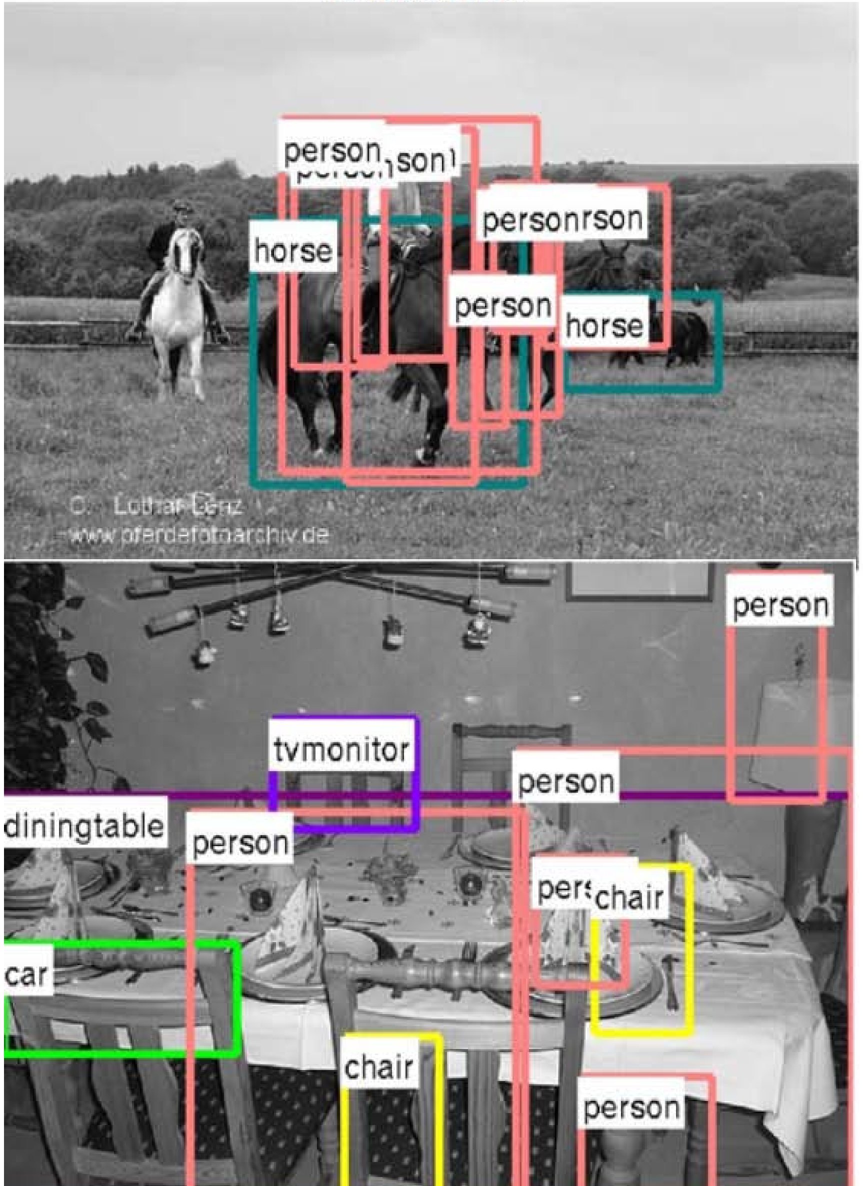
Inhibit
overlapping people & bottles
because local detectors confuse them

Favor
overlapping people & sofas
because people sit on sofas

Slide by D. Ramanan

Results

Baseline



Our model

Slide by D. Ramanan