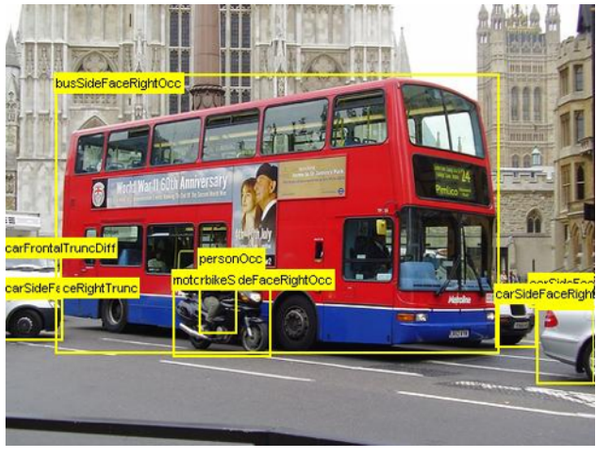
From Rigid Templates to Grammars: Object Detection with Structured Models

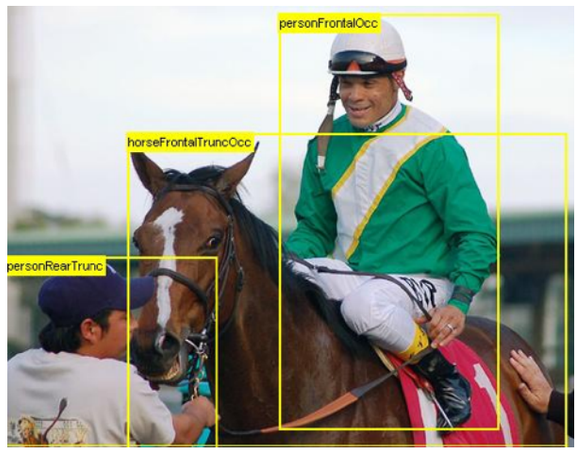
Ross B. Girshick

Dissertation defense

April 20, 2012

The question





What objects are where?

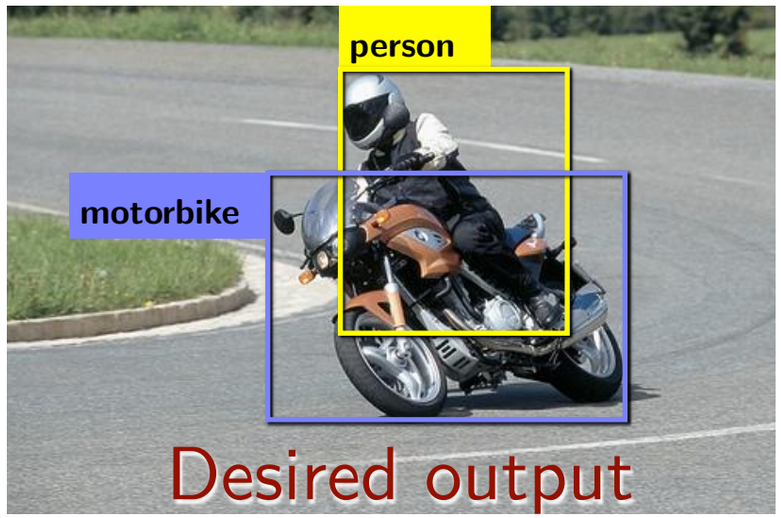
Why it matters

- Intellectual curiosity
  - How do we extract this information from the signal?
- Applications
  - Semantic image and video search
  - Human-computer interaction (*e.g.*, Kinect)
  - Automotive safety
  - Camera focus-by-detection
  - Surveillance
  - Semantic image and video editing
  - Assistive technologies
  - Medical imaging
  - ...

Proxy task: PASCAL VOC Challenge

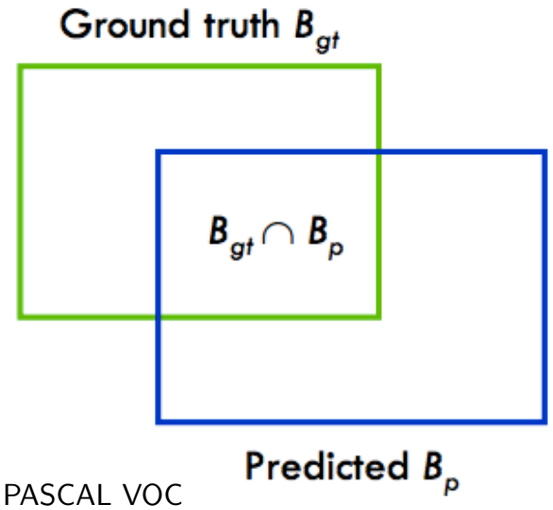
- Localize & name (*detect*) 20 basic-level object categories
  - Airplane, bicycle, bus, cat, car, dog, person, sheep, sofa, monitor, *etc.*


Input



Desired output

- 11k training images with 500 to 8000 instances / category
- Evaluation: bounding-box overlap; average precision (AP)



$$\text{overlap}(B_{\text{gt}}, B_{\text{p}}) = \frac{|B_{\text{gt}} \cap B_{\text{p}}|}{|B_{\text{gt}} \cup B_{\text{p}}|}$$

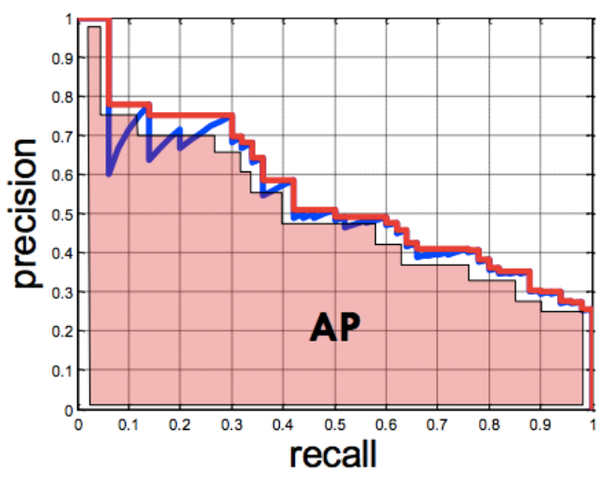


Image credits: PASCAL VOC

Challenges

- Deformation
- Viewpoint
- Subcategory
- Variable structure
- Occlusion
- Background clutter
- Photometric

Challenges

- Deformation

Image credit: http://i173.photobucket.com/albums/w78/yahoozy/MultipleExposures2.jpg

Challenges

- Viewpoint

Image credits: PASCAL VOC

Challenges

- Subcategory — “airplane” images

Image credits: PASCAL VOC

Challenges

- Variable structure

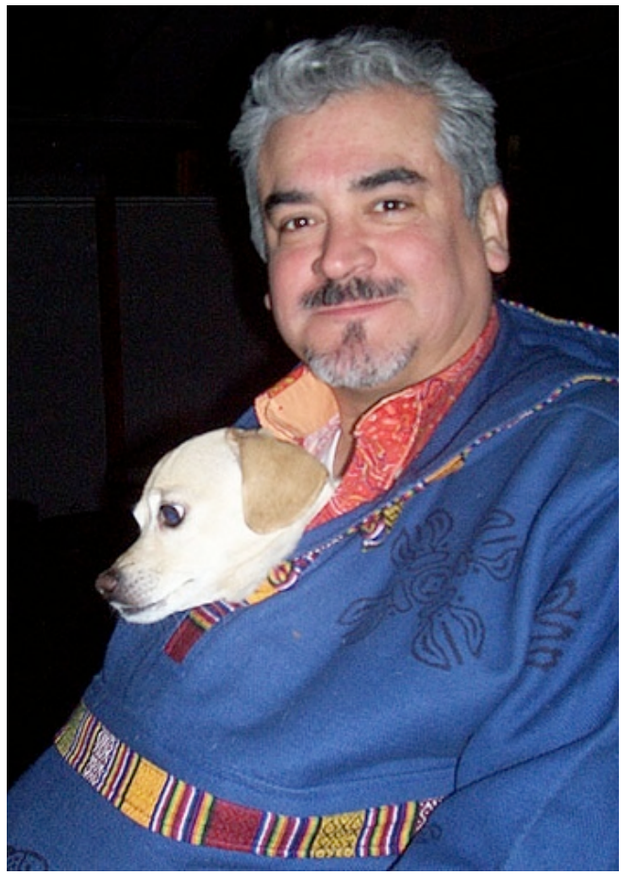


Image credits: PASCAL VOC

PASCAL VOC Challenges 2007-2011

- 2007 Challenge
  - Winner: Deformable part models & Latent SVM [FMR'08]
  - 21% mAP
  - Baseline for dissertation

↑

Prior work

This work

- Winners of 2008 & 2009 Challenges

↓

- Fast forward to the 2011 Challenge
  - Our system (**voc-release4**): 34% mAP
  - Top system (NLPR): 41% mAP
  - NLPR method: **voc-release4** + LBP image features + richer spatial model (GMM) + more context rescoring
  - Second (MIT-UCLA) and third place (Oxford) also based on **voc-release4**

Contributions — By area

- Object representation*
  - Mixture models (in PAMI'10); Latent orientation; **Person grammar model**
- Efficient detection algorithms*
  - Cascaded detection for DPM (oral at CVPR'10)
- Learning*
  - **Weak-label structural SVM** (spotlight at NIPS'11)
- Detection post-processing
  - Bounding box prediction & context rescoring
- Image representation
  - Enhanced HOG features; features for boundary truncation & small objects
- Software
  - voc-release{2,3,4} – currently the "go to" object detection system

Object representation

Model lineage – Dalal & Triggs

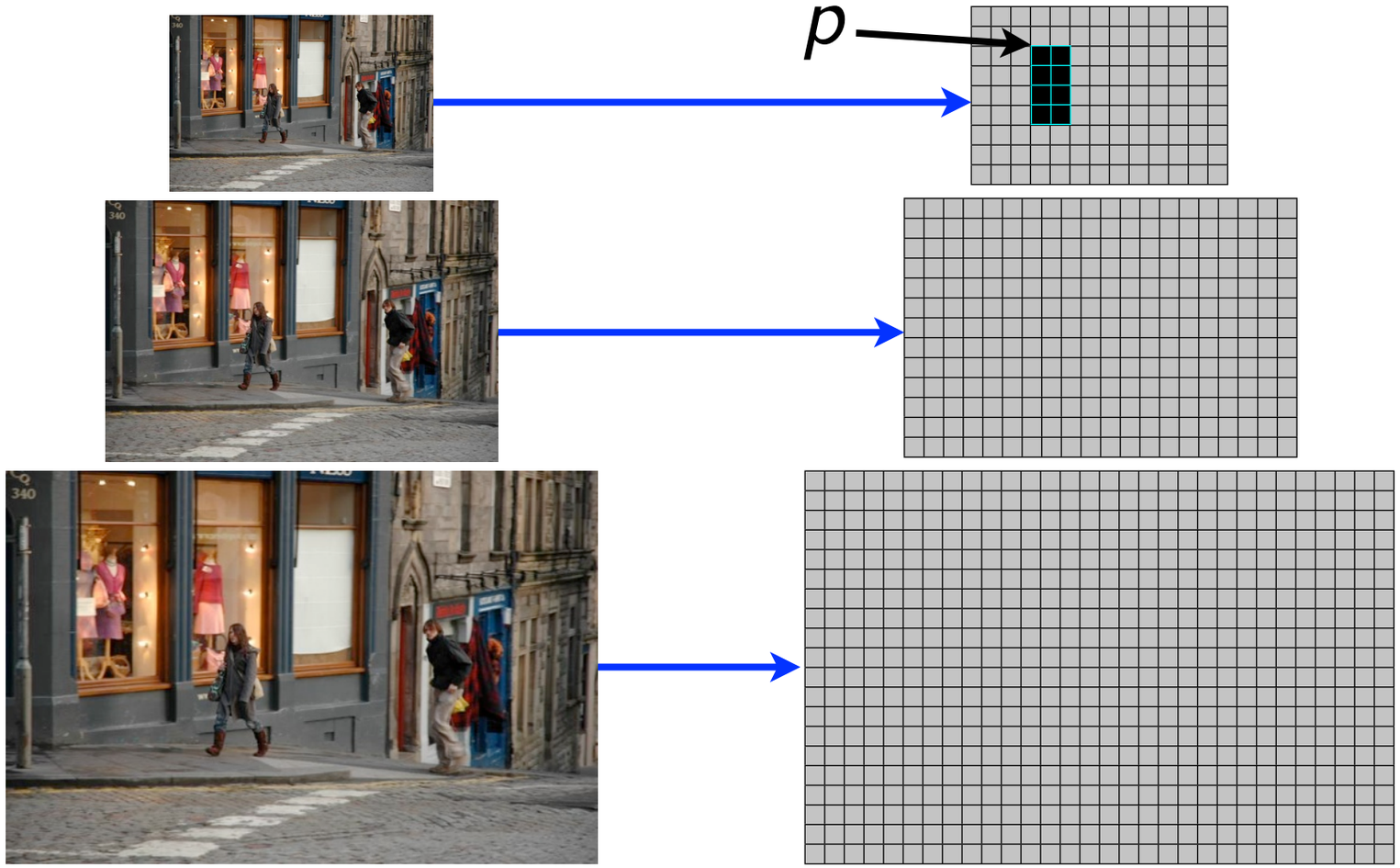


Image pyramid

HOG feature pyramid

$$\text{score}(I, p) = \mathbf{w} \cdot \psi(I, p)$$

[Dalal and Triggs '05]

"Root filter"

- Histogram of Oriented Gradients (HOG) features
- Scanning window detector (linear filter)
- $\mathbf{w}$ learned by SVM

Model lineage – Latent SVM DPM

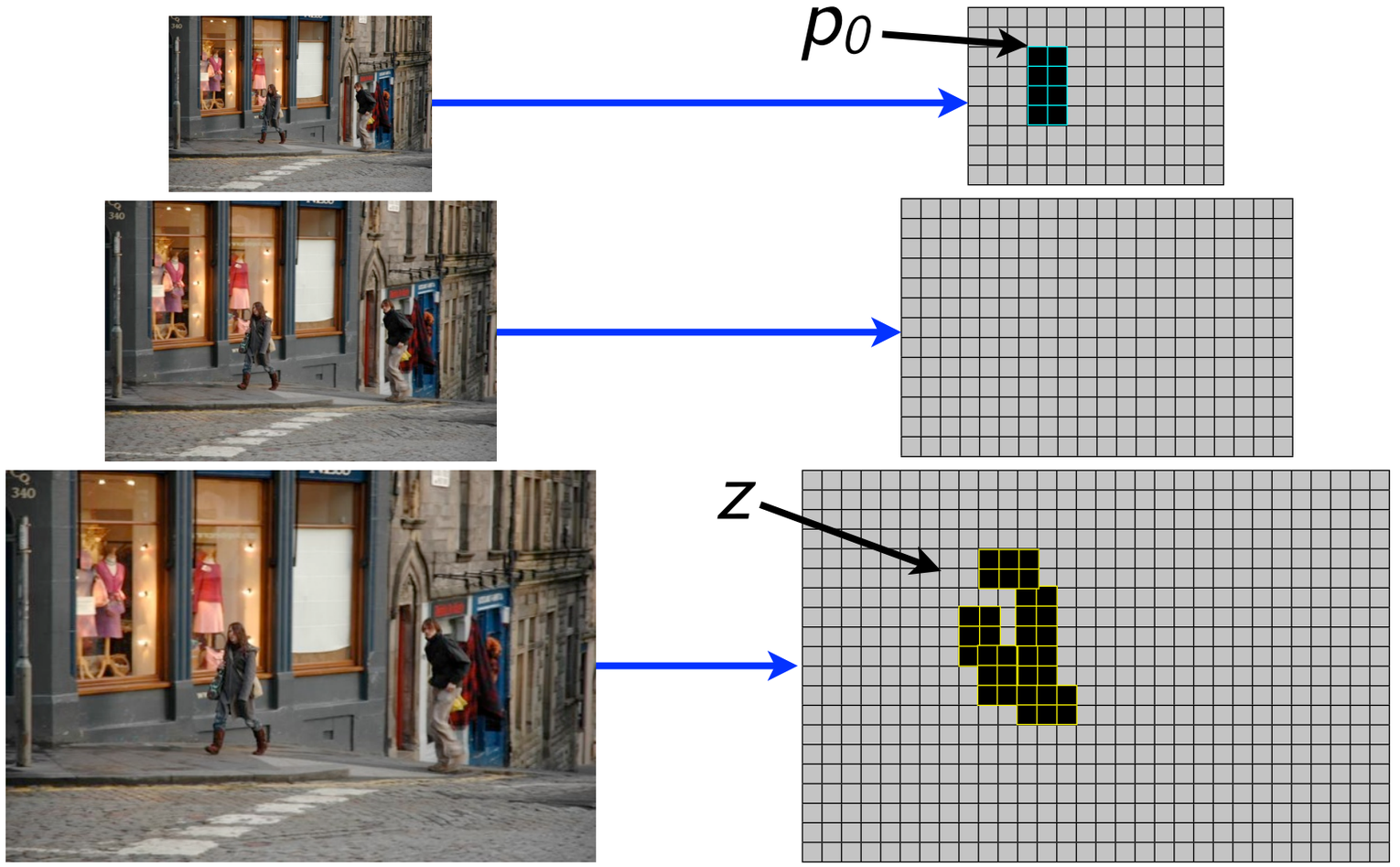
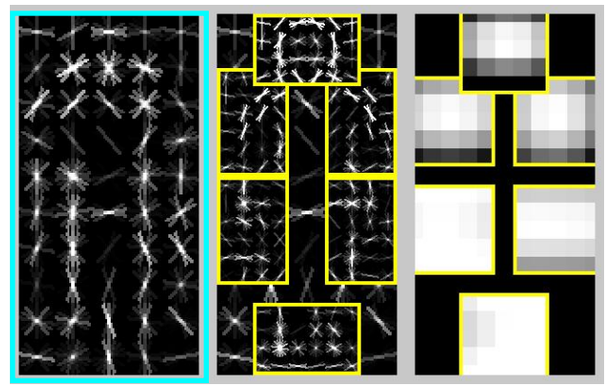


Image pyramid

HOG feature pyramid

$$\text{score}(I, p_0) = \max_{z \in \mathcal{Z}(I)} \mathbf{w} \cdot \psi(I, (p_0, z))$$

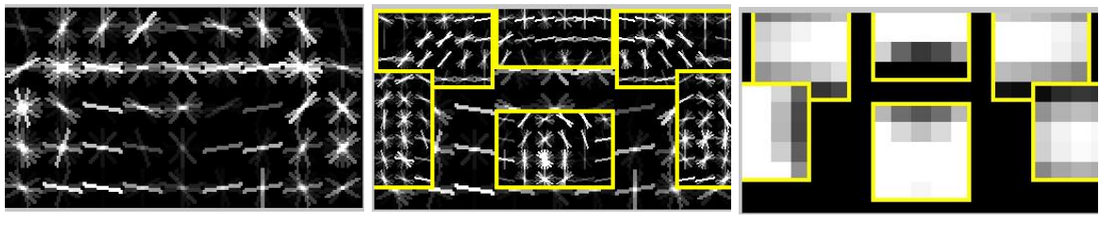
$$z = (p_1, \ldots, p_n)$$



[FMR'08]

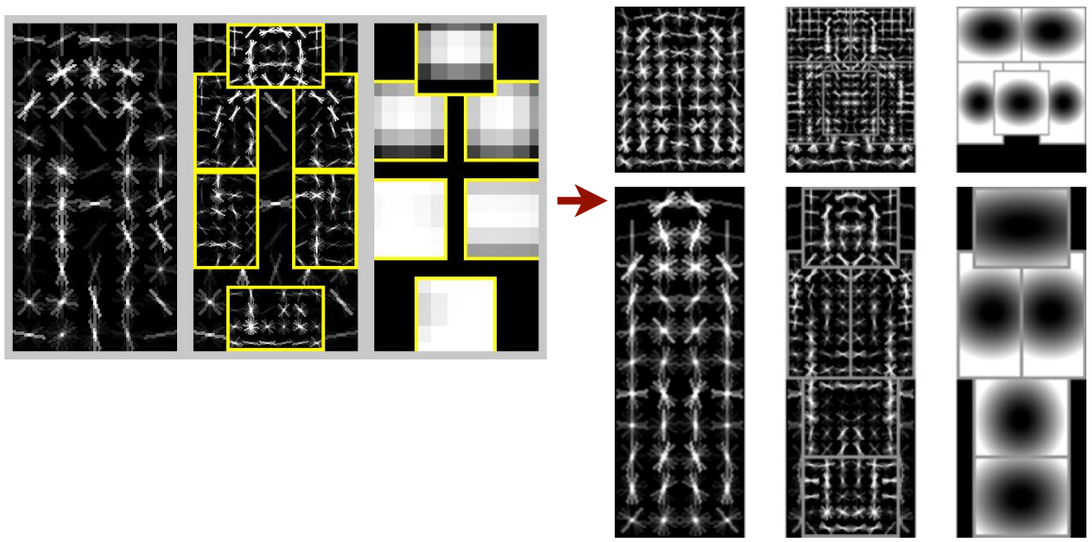
- Dalal & Triggs + Parts in a deformable configuration z
- Scanning window detection: max over z at each p_0
- $\mathbf{w}$ learned by latent SVM

Superposition of views

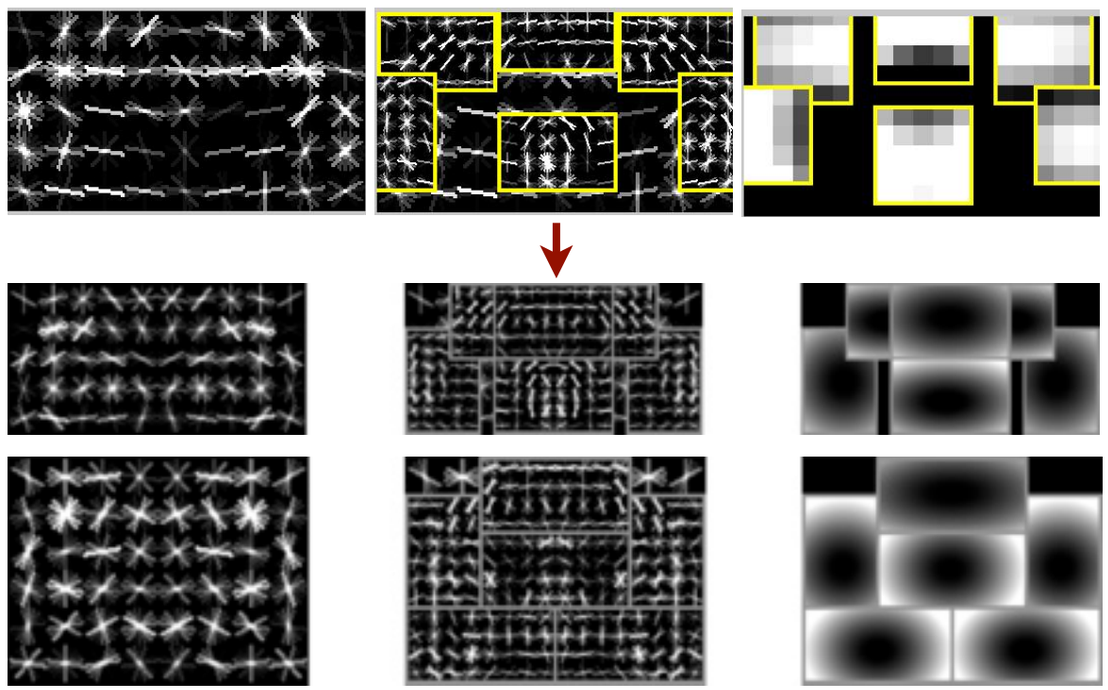


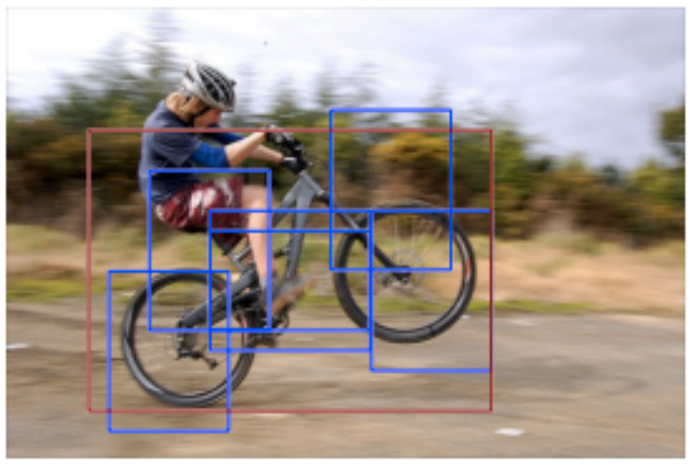
Mixture of DPMs

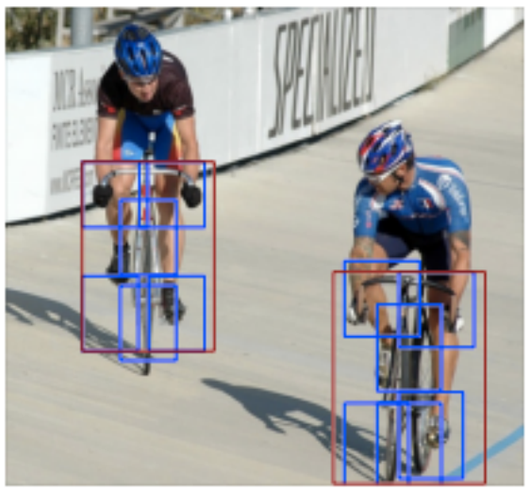
Person



Car



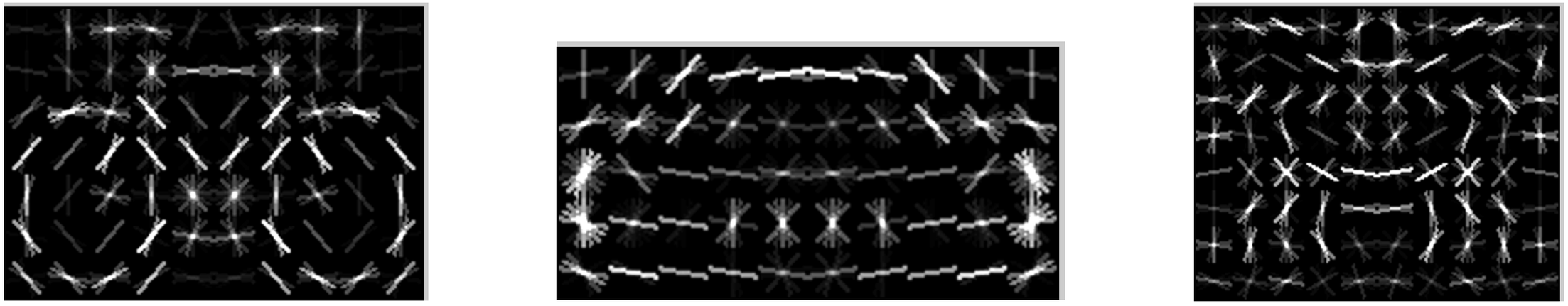




- Training (component labels are hidden)
  - Cluster training examples by bounding-box aspect ratio
  - Initialize root filters for each component (cluster) independently
  - Merge components into mixture model and train with latent SVM

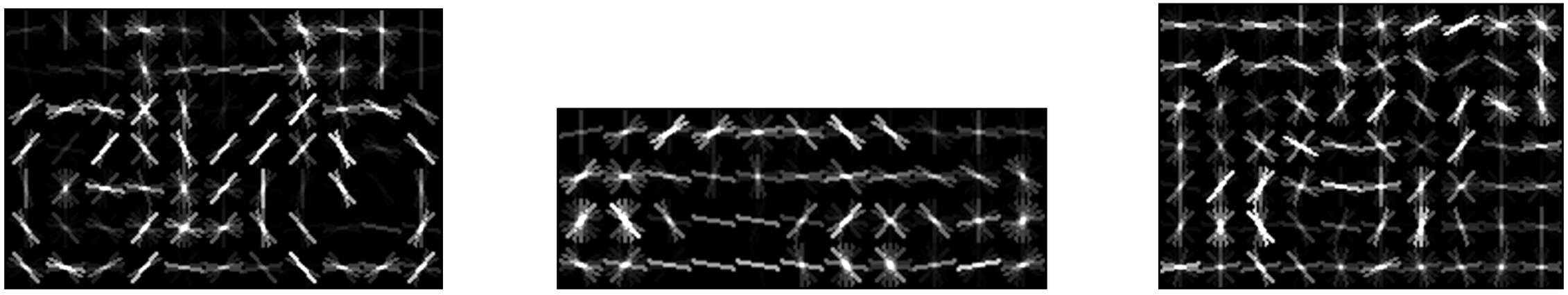
Mixtures with latent orientation

("pushmi-pullyu"
instead of horse)



Learning *without* latent orientation

(right-facing horse)

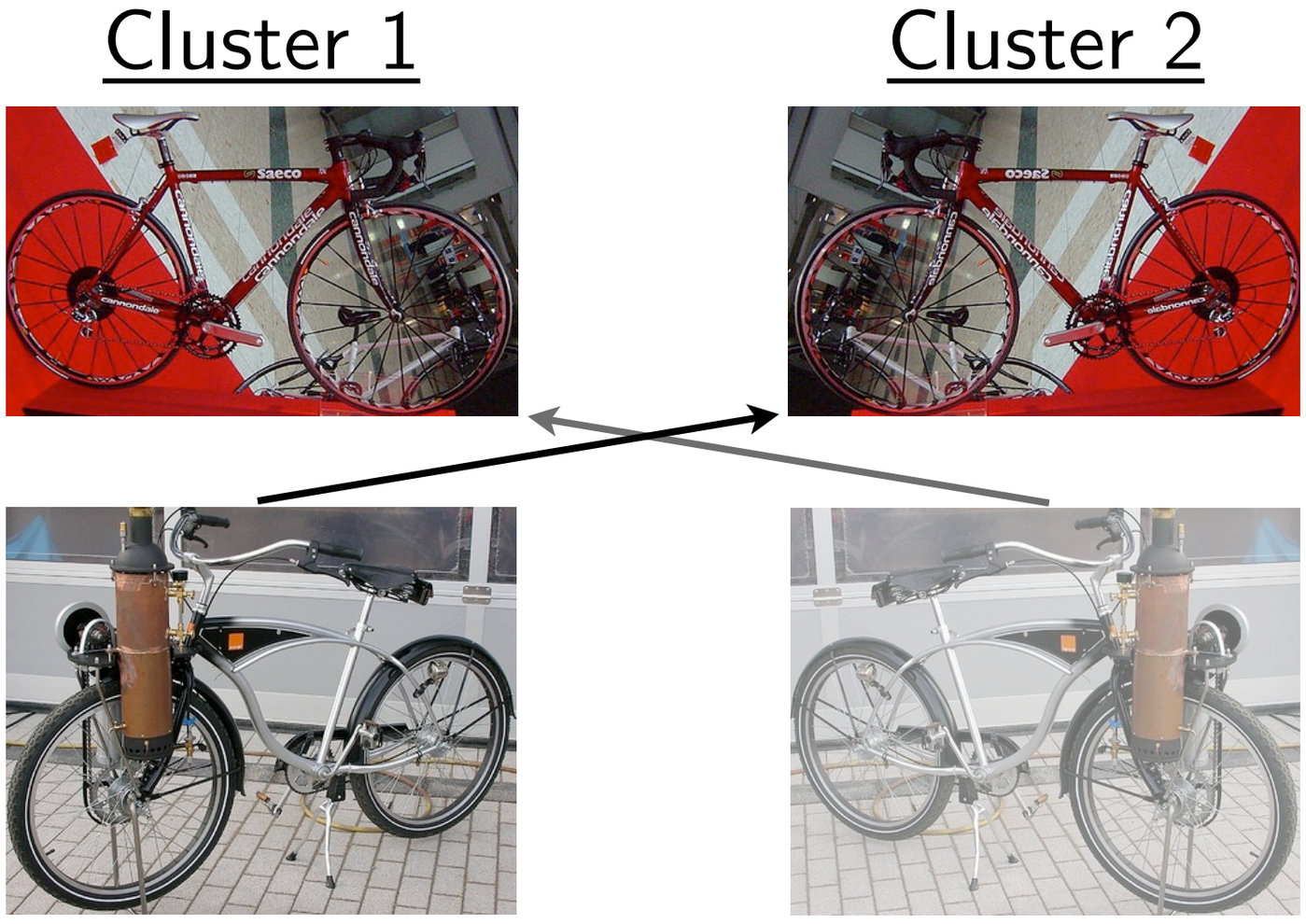


Learning *with* latent orientation

[GFM voc-release4]

Unsupervised orientation clustering

- Online clustering with a hard constraint



Assign i^{th} example to nearest cluster

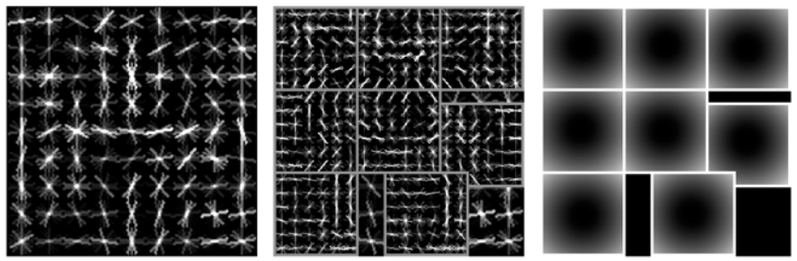
Flipped example *must* go to the other cluster

Latent orientation improves performance

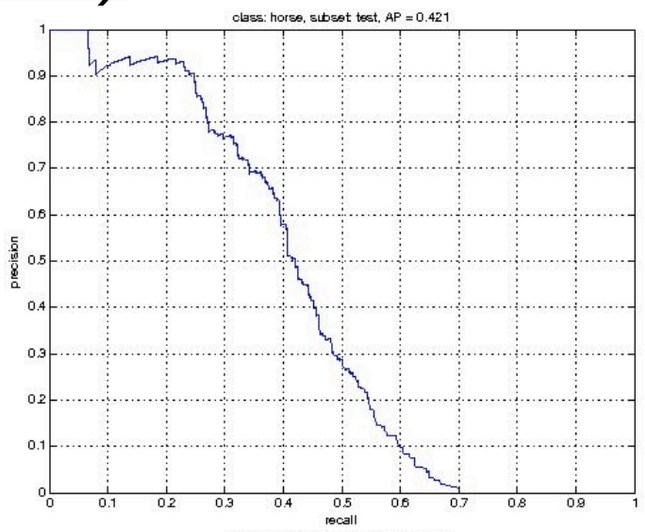
Horse model type

AP (2007)

Single component

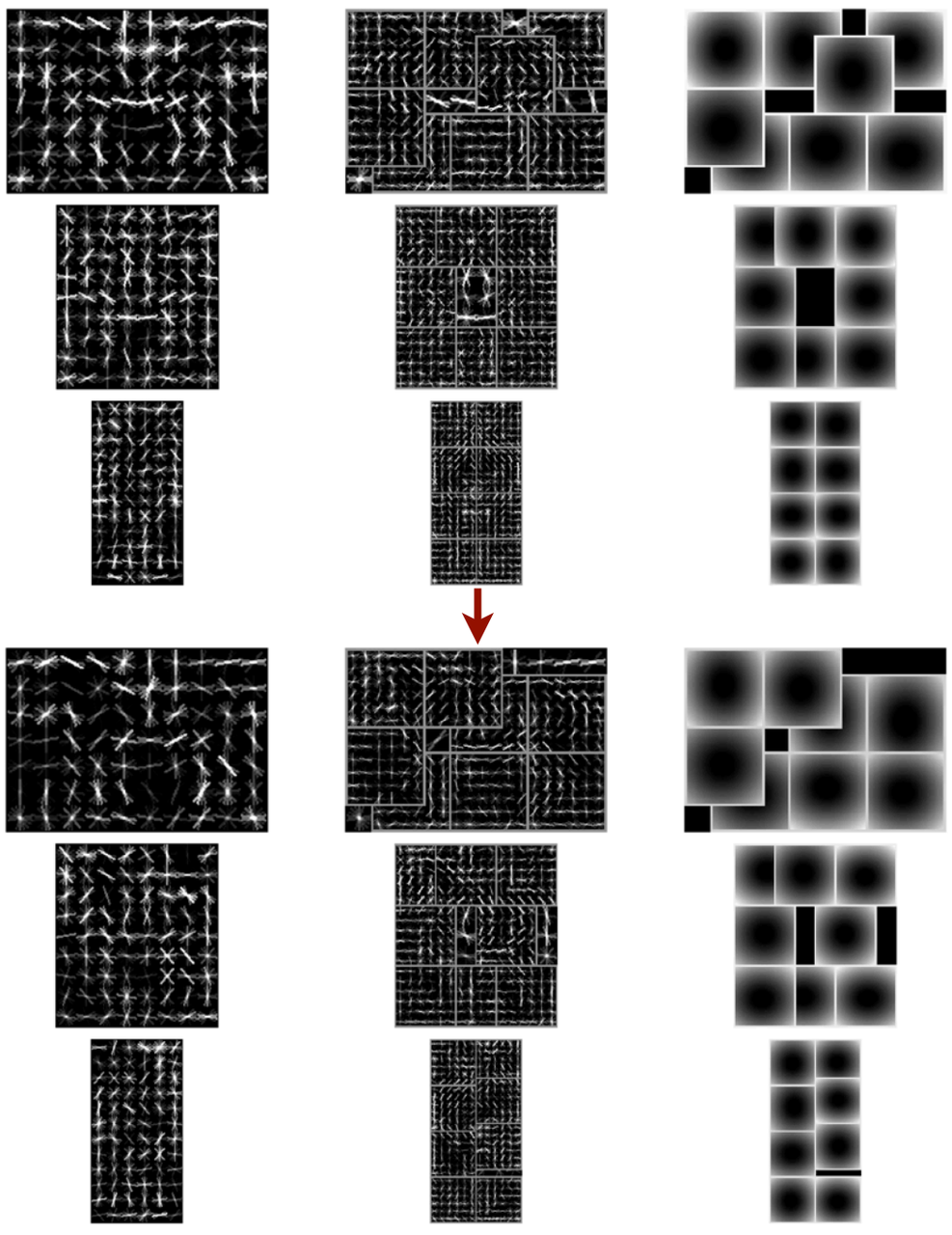


42.1

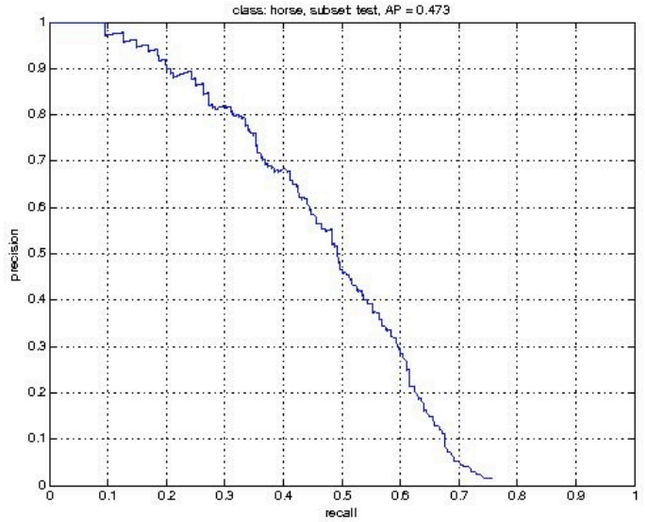


Mixture model

3 components



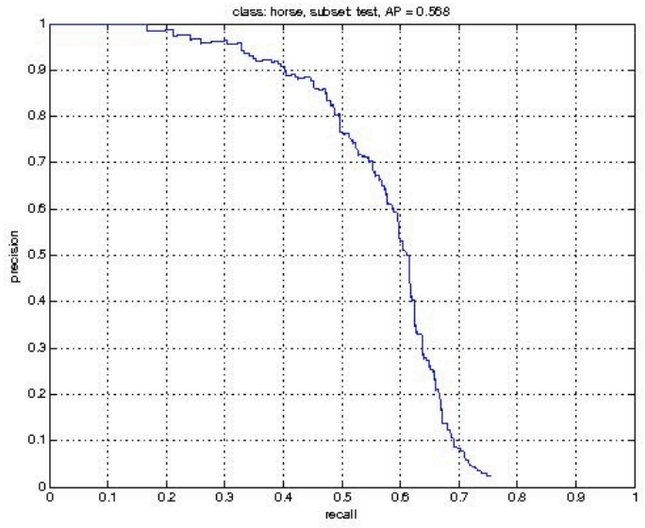
47.3



Latent orientation

2x3 components

56.8



Results – Mixture models and latent orientation

- Mixture models boost mAP by 3.7 points
- Latent orientation boost mAP by 2.6 AP points

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
LSVM CVPR [34]	18.0	41.1	9.2	9.8	24.9	34.9	39.6	11.0	15.5	16.5	11.0	6.2	30.1	33.7	26.7	14.0	14.1	15.6	20.6	33.6	22.3
voc-release4.01: EHOG + TRUNC + BBOX + OPTIM + MAX-REG																					
	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
star model	21.5	49.7	9.3	9.3	27.3	39.8	53.3	11.4	13.5	17.8	28.6	3.7	42.1	38.3	33.1	13.4	15.1	20.4	28.7	42.7	26.0
MIX	33.2	58.0	2.3	15.7	26.5	49.3	55.5	13.9	18.1	19.1	24.6	12.6	47.3	42.6	40.5	14.4	17.0	27.4	37.9	39.0	29.7
⋯ + ORIENT	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
⋯ + CNTXT	31.2	61.5	11.9	17.4	27.0	49.1	59.6	23.1	23.0	26.3	24.9	12.9	60.1	51.0	43.2	13.4	18.8	36.2	49.1	43.0	34.1

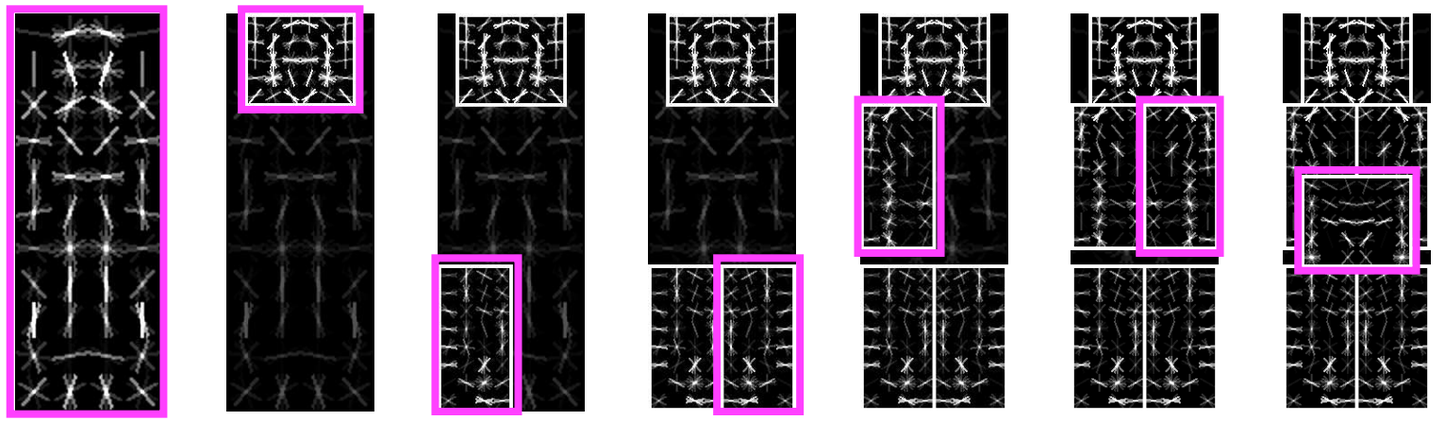
AP scores using the PASCAL 2007 evaluation

- 12 AP point improvement (>50% relative) over the baseline

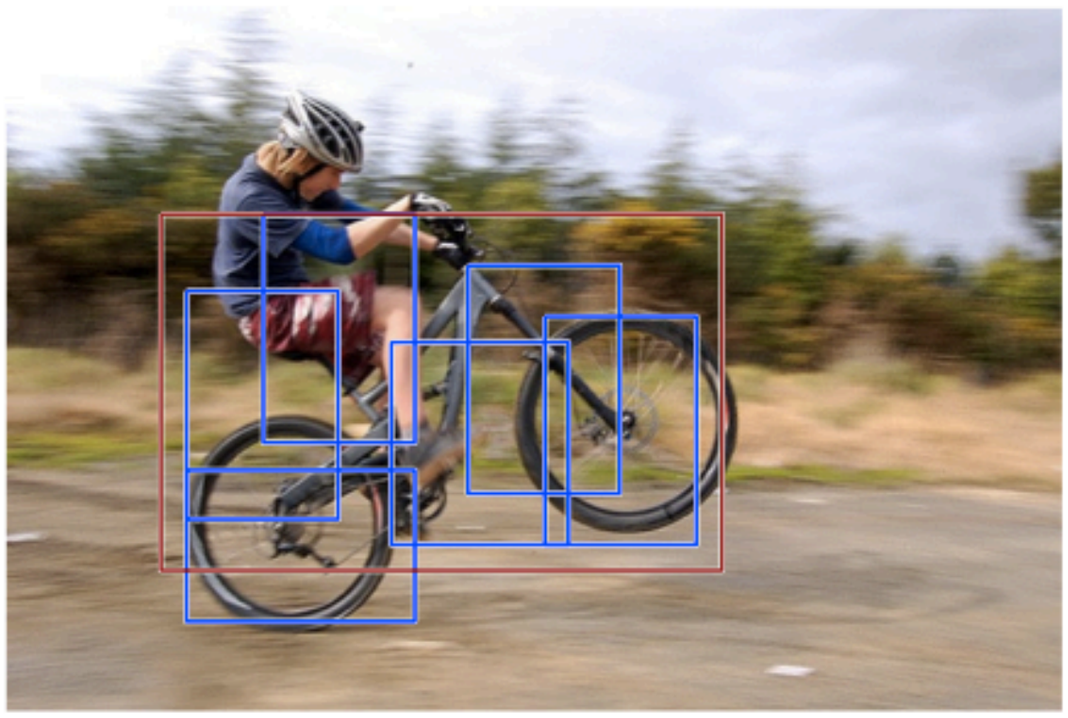
Efficient detection

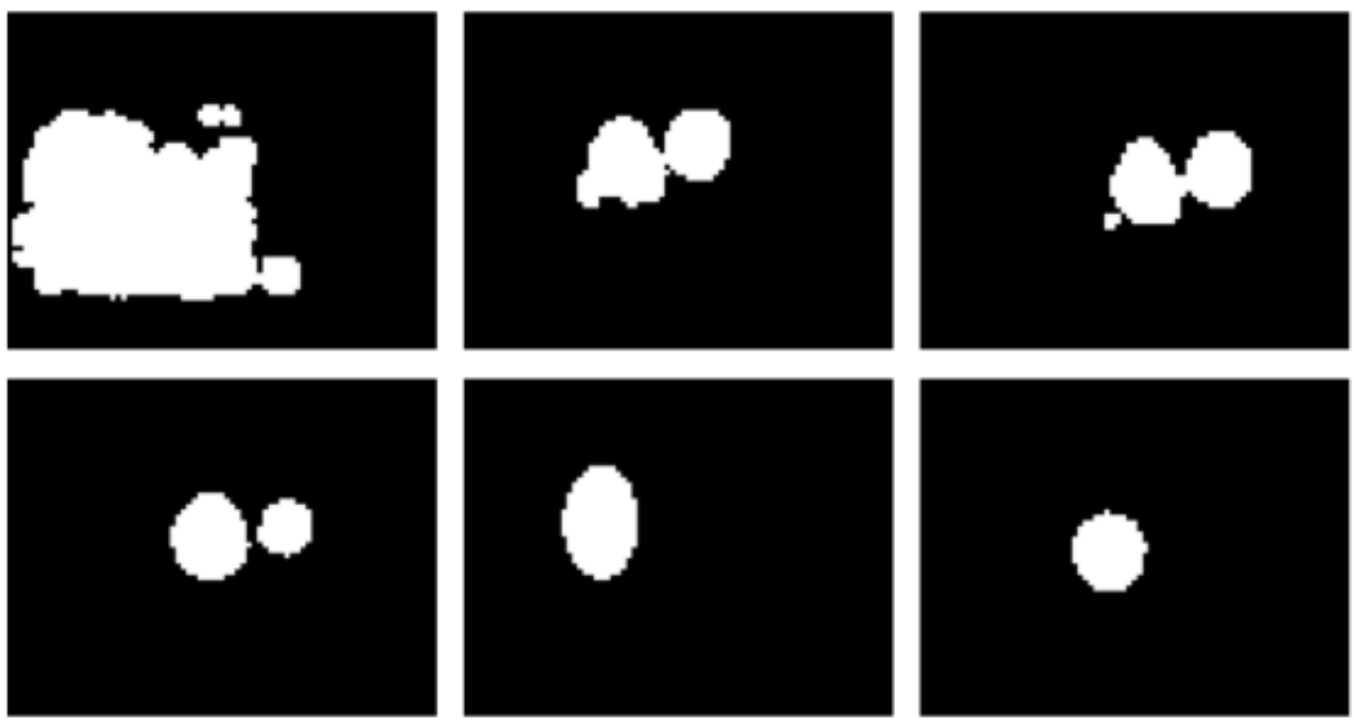
Cascaded detection for DPM

- Add in parts one-by-one and prune partial scores



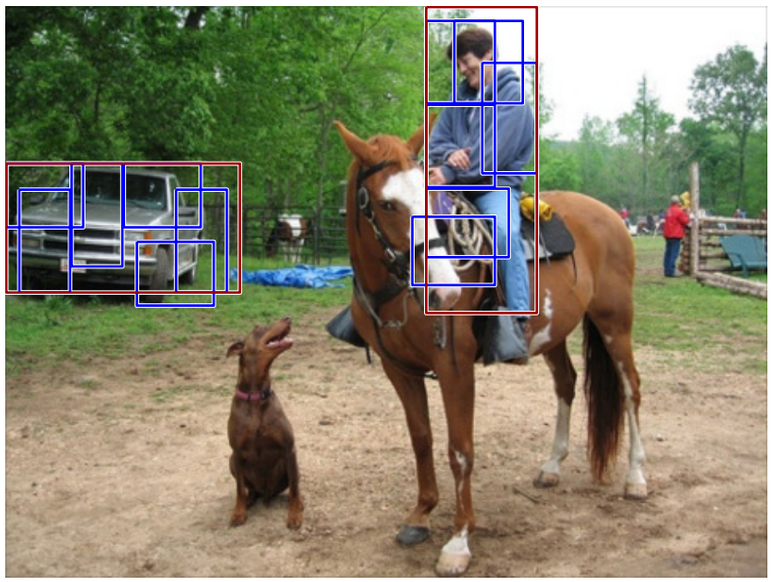
- Sparse dynamic programming tables (reuse computation!)

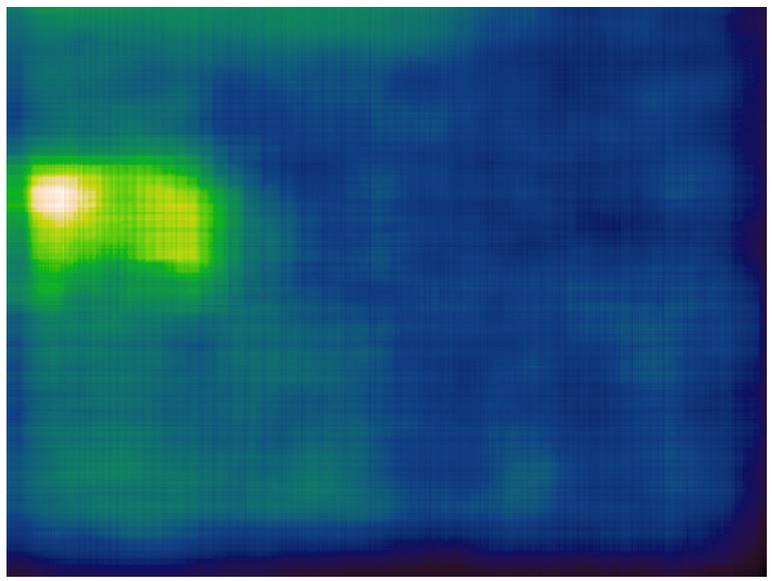


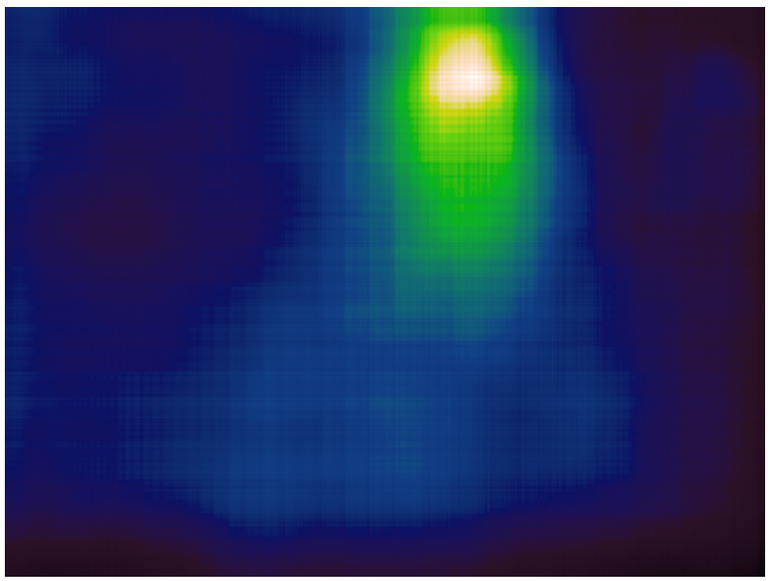


Threshold selection & PCA filters

- Data-driven threshold selection
  - Based on statistics of partial scores on training data
  - Provably safe ("probably approximately admissible" thresholds)
  - Empirically effective
- 2-stage cascade with simplified appearance models
  - Use PCA of HOG features (or model filters)
  - Stage 1: place low-dimensional filters; Stage 2: place original filters

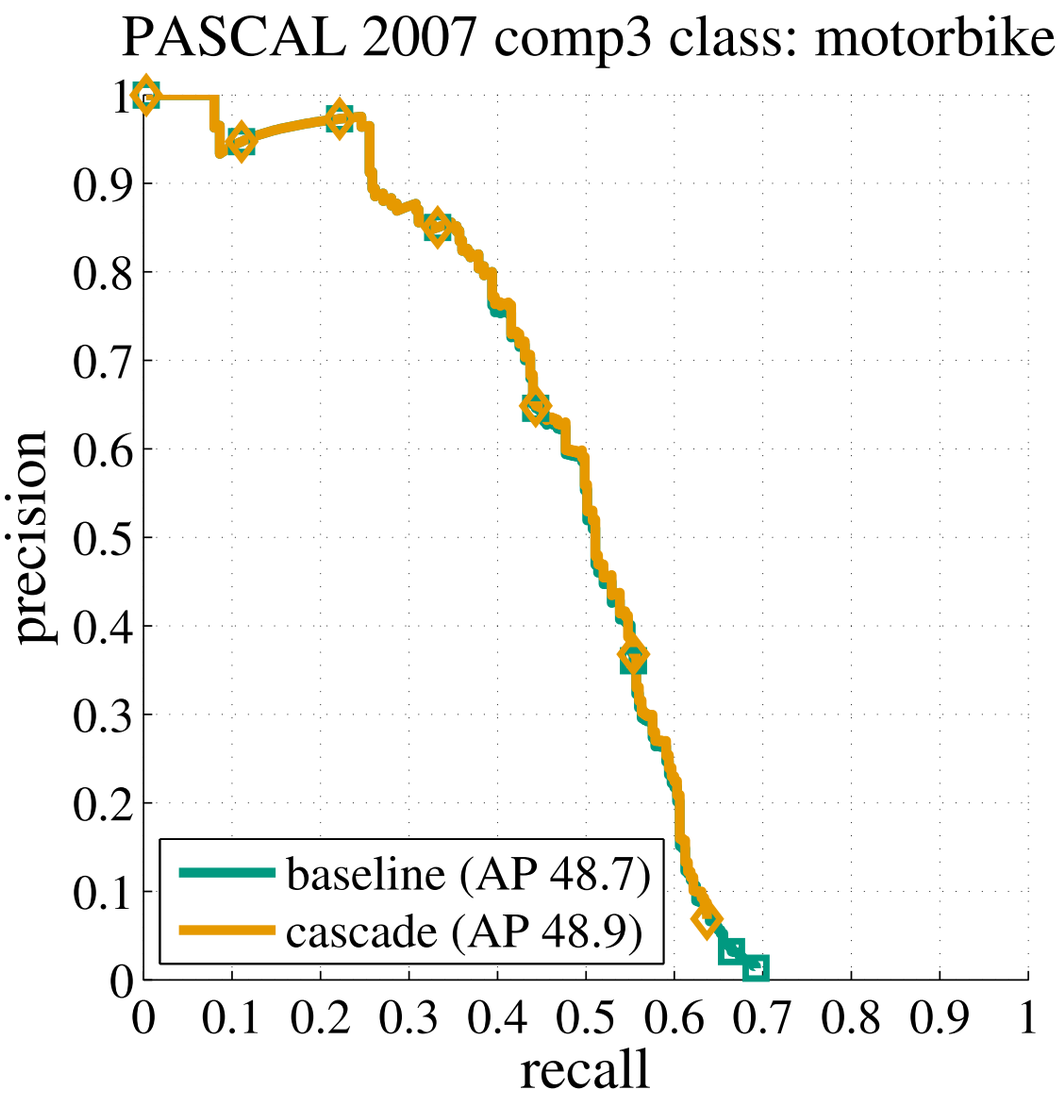






Results — 15x speedup (no loss in mAP)

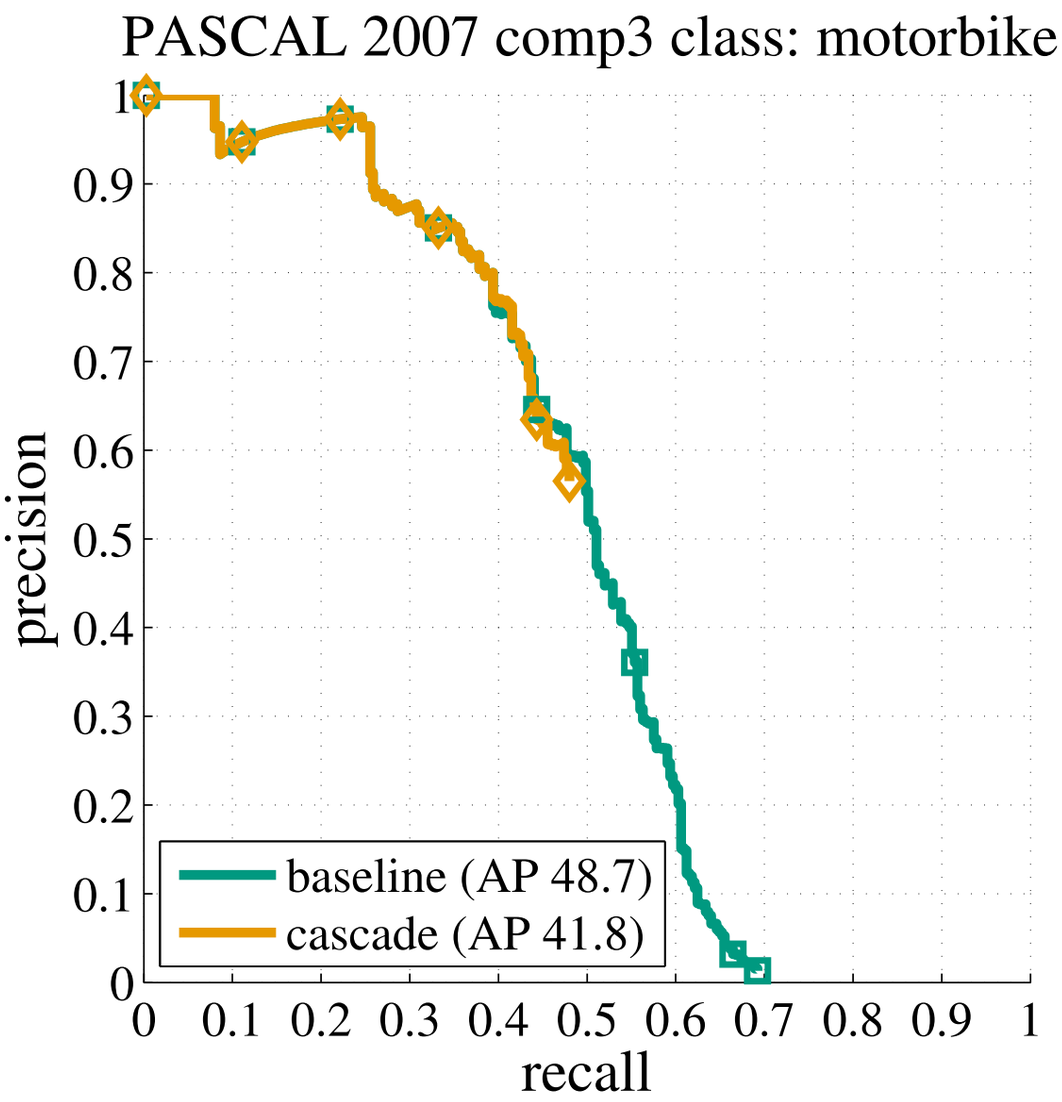
High recall



23.2x faster

(618ms per/image)

Lower recall ⇒ faster



31.6x faster

(454ms per/image)

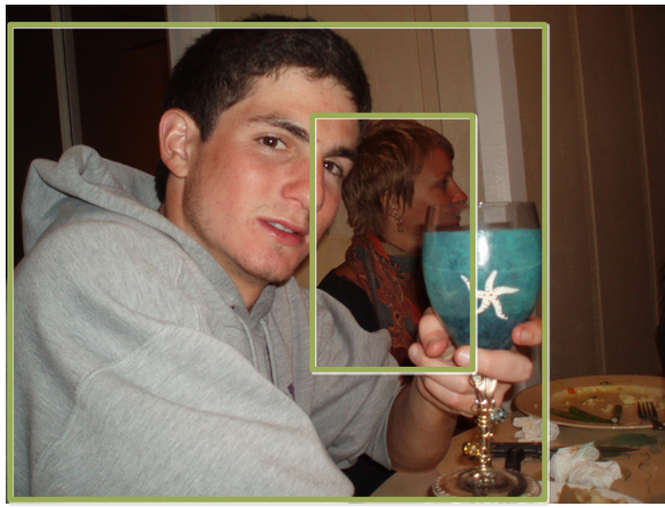
Towards richer grammar models

People are complicated

Helmet,
occluded left side

Ski cap, no face,
truncated

Pirate hat, dresses,
long hair

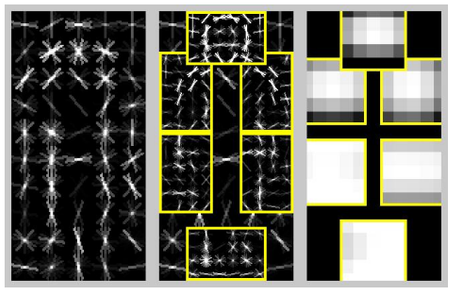


Truncation, holding glass,
heavy occlusion

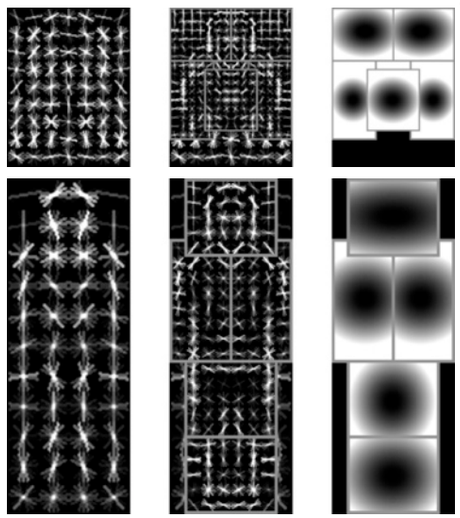
Objects from visually rich categories
have diverse *structural* variation

Compositional models

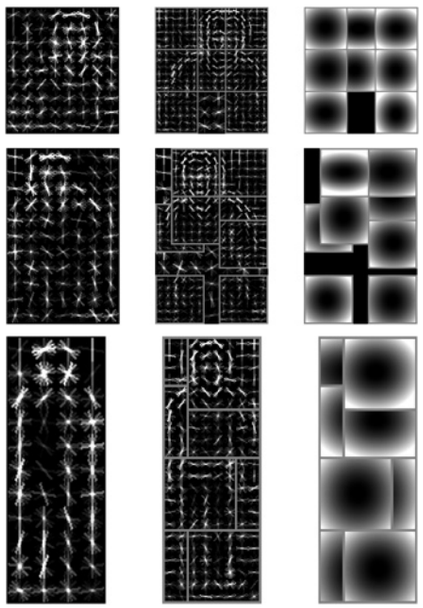
[DT'05]
AP 0.12



[FMR'08]
AP 0.27



[FGMR'10]
AP 0.36



[GFM voc-release4]
AP 0.42

More mixture components?

There are too many combinations!

Instead...

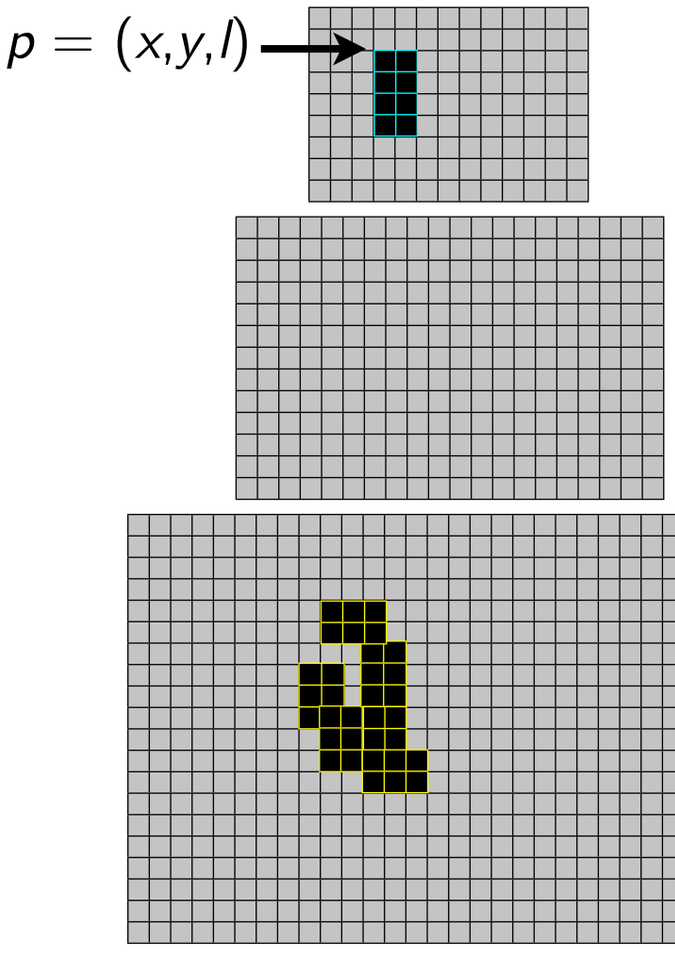
... compositional models defined by grammars

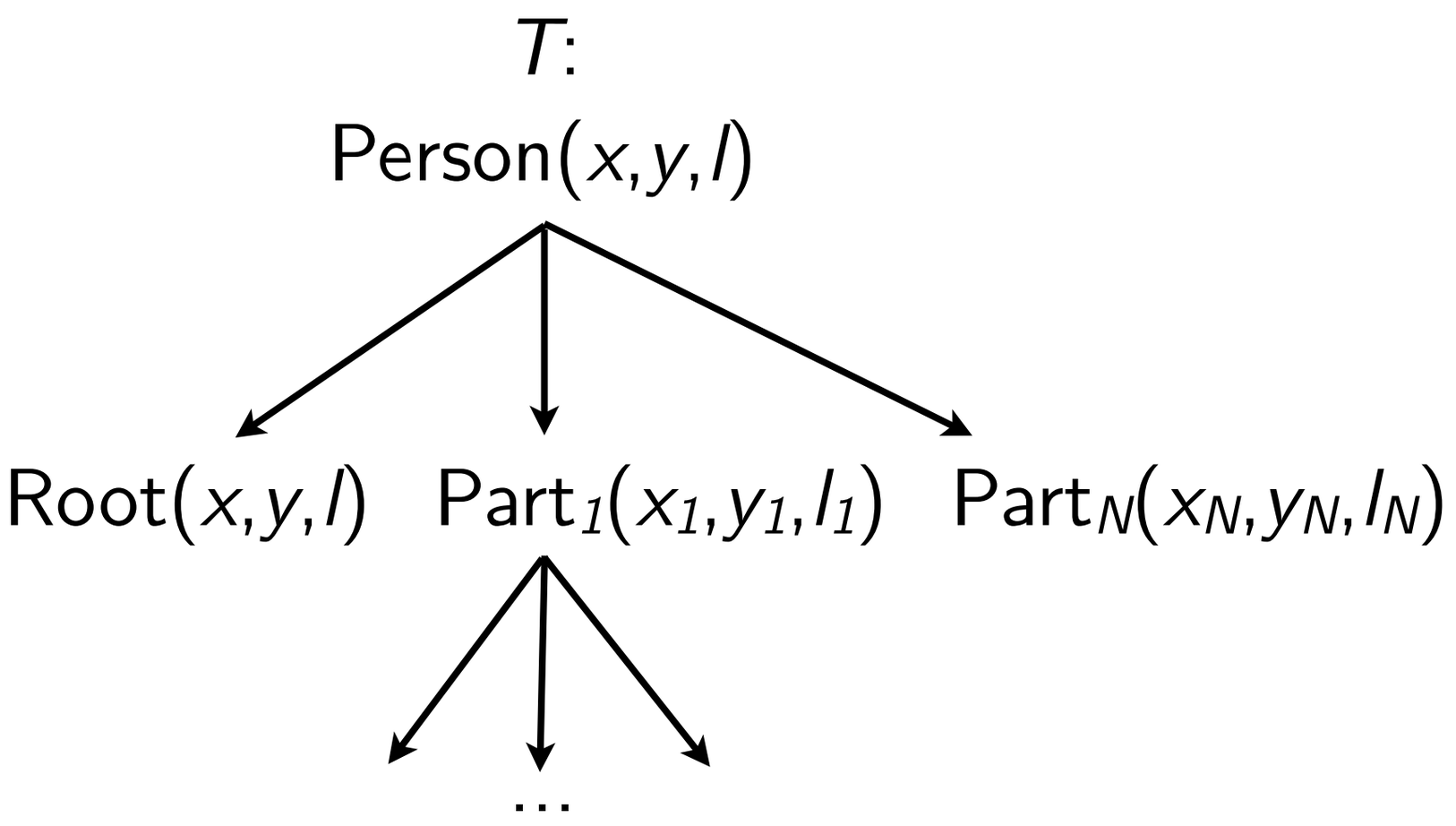
Object detection grammars

- A modeling language for building object detectors [FM'10]
  - Terminals (model image data appearance)
  - Nonterminals (objects, parts, ...)
  - Weighted production rules (define composition, variable structure)
- Composition
  - Objects are recursively composed of other objects (parts)
- Variable structure
  - Expanding different rules produces different structures
- Person $\rightarrow$ Head, Torso, Arms, Legs
- Head $\rightarrow$ Eye, Eye, Mouth
- Mouth $\rightarrow$ Smile OR Mouth $\rightarrow$ Frown

Object detection grammars

- Object hypothesis = derivation tree T





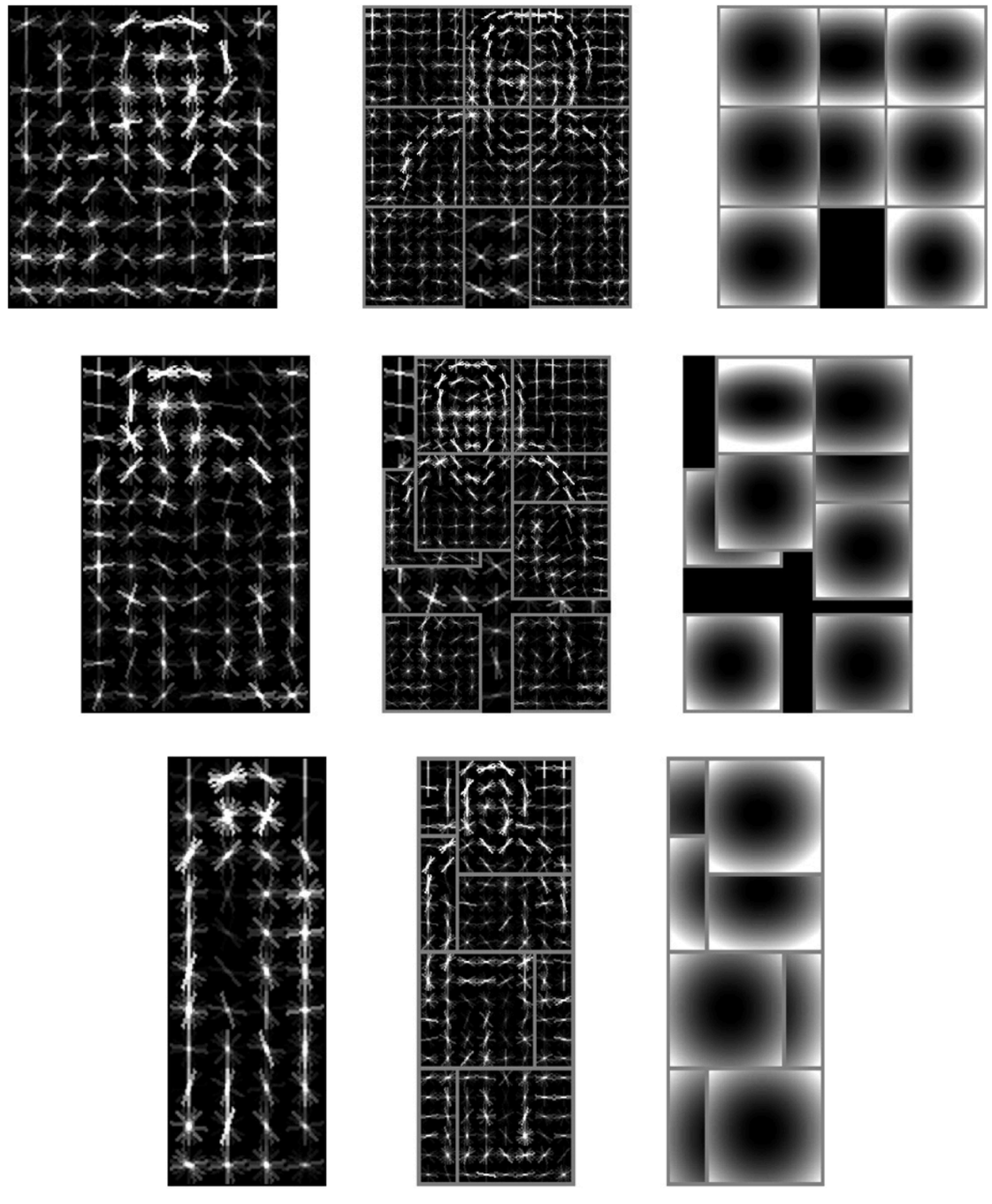
- Linear score function

$$\text{score}(I, T) = \mathbf{w} \cdot \boldsymbol{\psi}(I, T)$$

Detection with DP

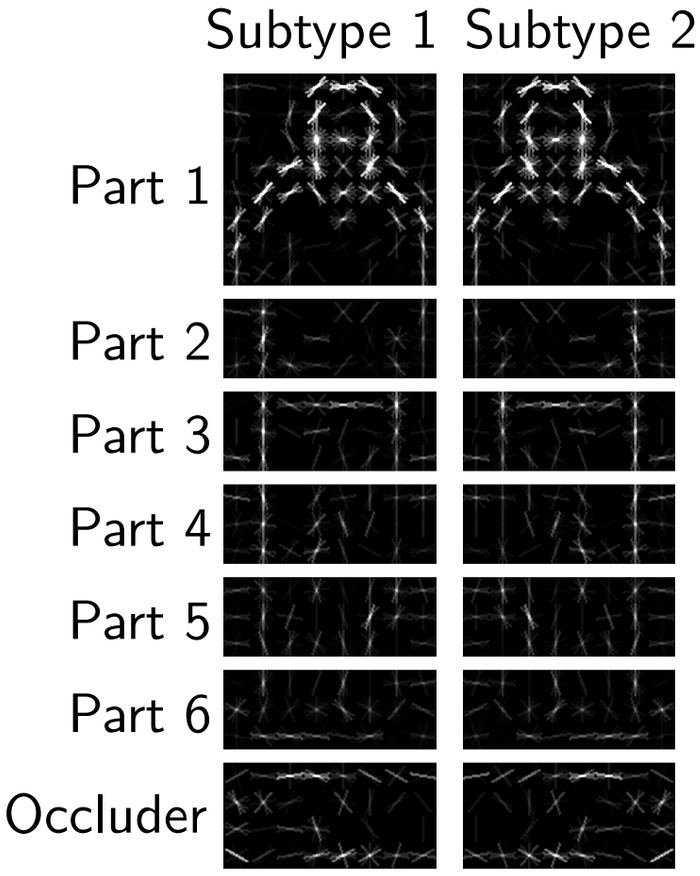
$$T^*(p) = \underset{T \in \mathcal{T}_p}{\text{argmax}}\, \mathbf{w} \cdot \boldsymbol{\psi}(I, T)$$

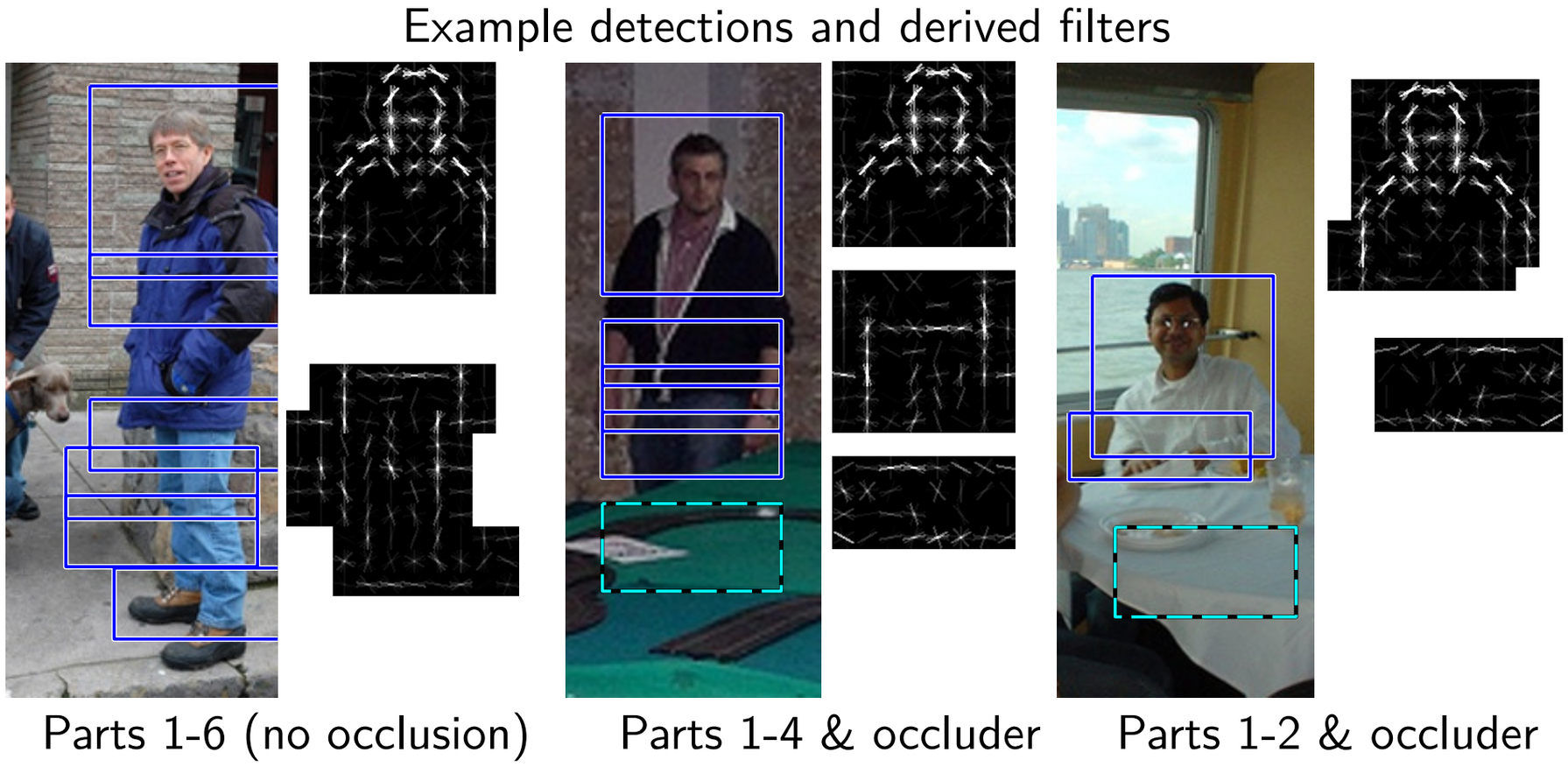
Build on what works



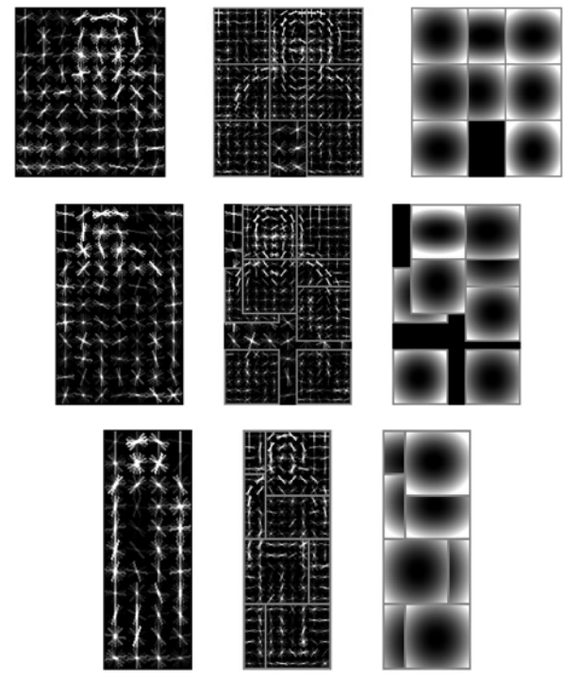
Can we build a better person detector?

Case study: a person detection grammar

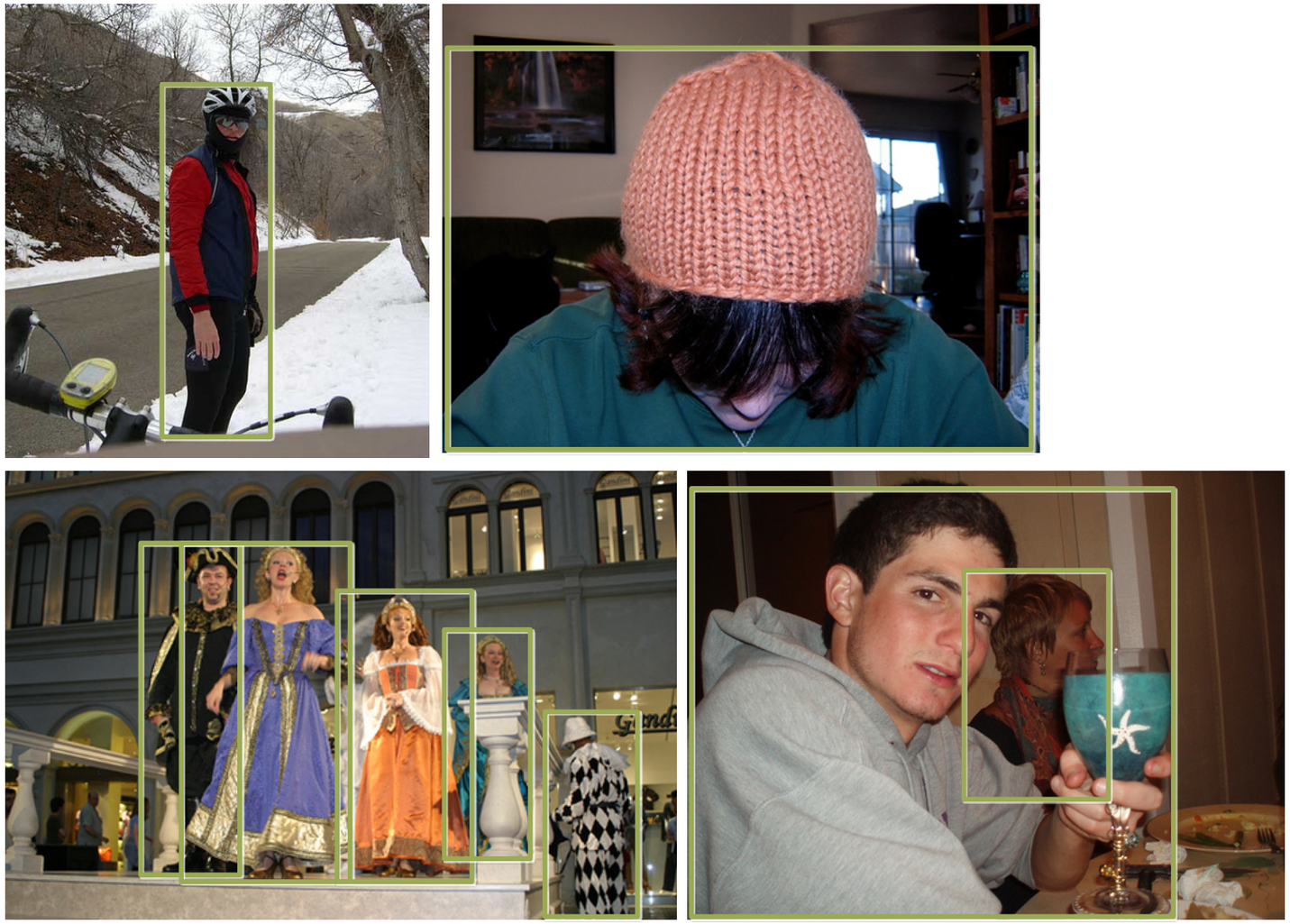




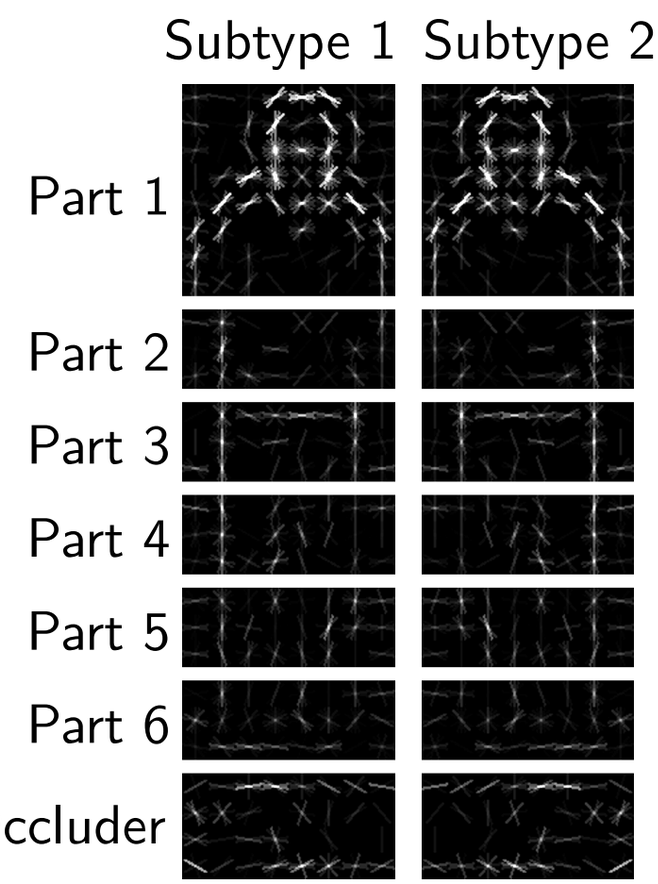
- Fine-grained occlusion
- Sharing across all derivations
- Model of the stuff that causes occlusion
- Part subtypes and multiple resolutions
- Parts have subparts (not pictured)



Training models

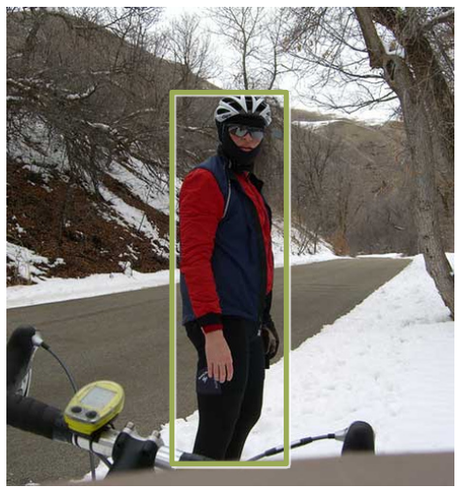
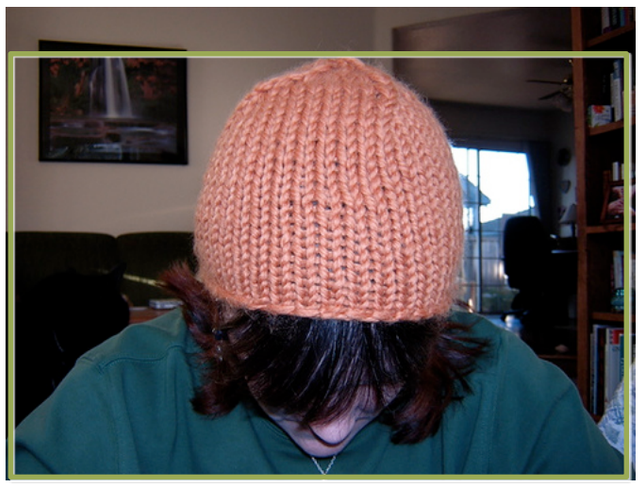


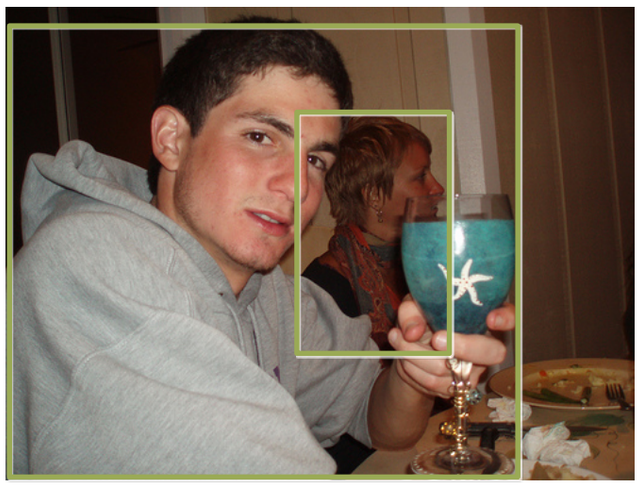
training



- PASCAL data: bounding-box labels
- No derivation trees given! (weakly-supervised learning)
- Learn the parameters $\mathbf{w}$

Defining examples



- Each bounding box is a foreground example
- All locations in background images are background examples
- From these examples, learn the prediction rule

$$\text{Predicted output} \longrightarrow f_{\mathbf{w}}(x) = \underset{s \in \mathcal{S}(x)}{\operatorname{argmax}} \mathbf{w} \cdot \psi(x, s)$$

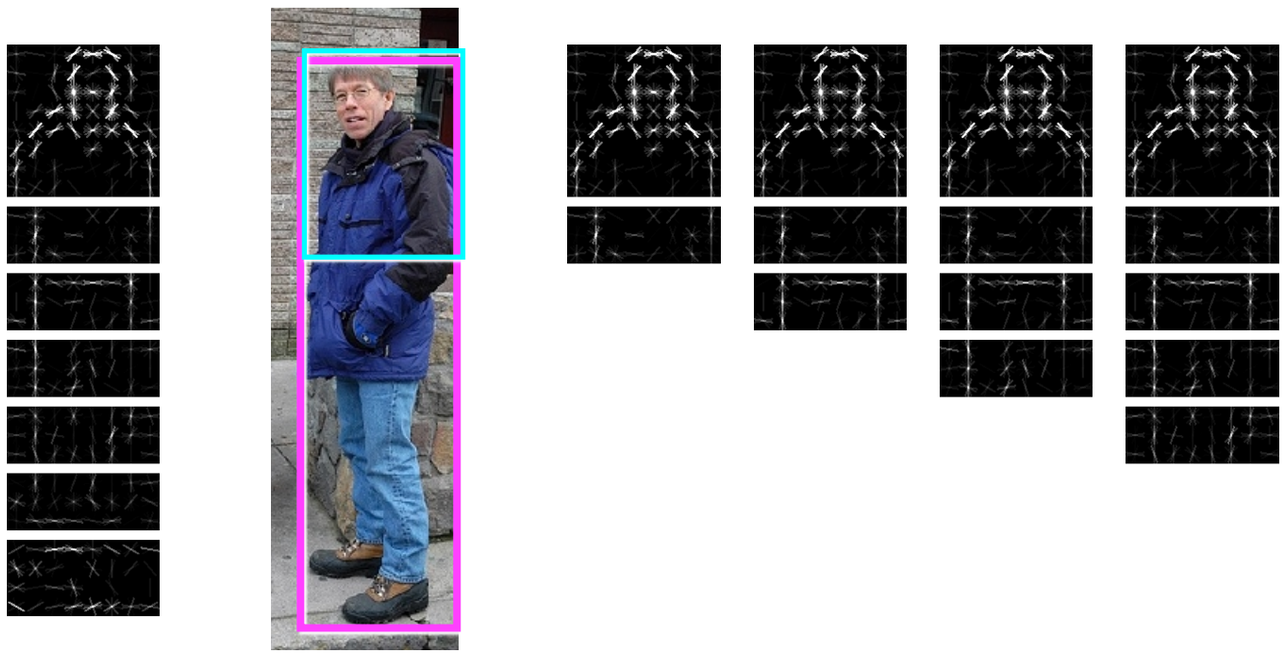
Input example

Possible outputs (derivation trees)

Feature map

Parameter learning

- Richer models, richer problems



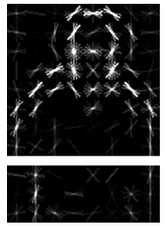
One good output... and many bad ones!

- Which learning framework should we use?

Classification training

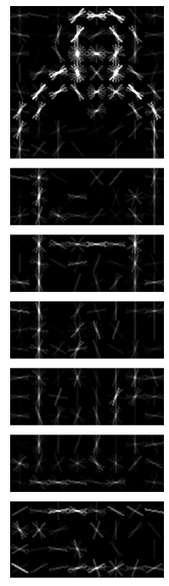
$$E_{\text{lsvm}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\left[0, 1 - y_i \max_{s \in \mathcal{S}(x_i)} \mathbf{w} \cdot \psi(x_i, s)\right]$$

Training:



LSVM objective:

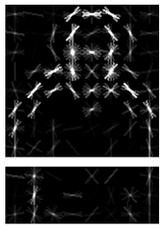
"score +1 here"



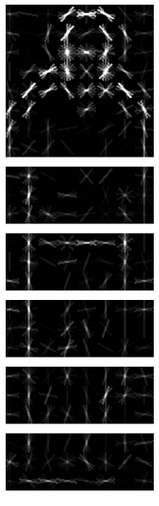
LSVM objective:

"score +1 here"

Testing:



vs.



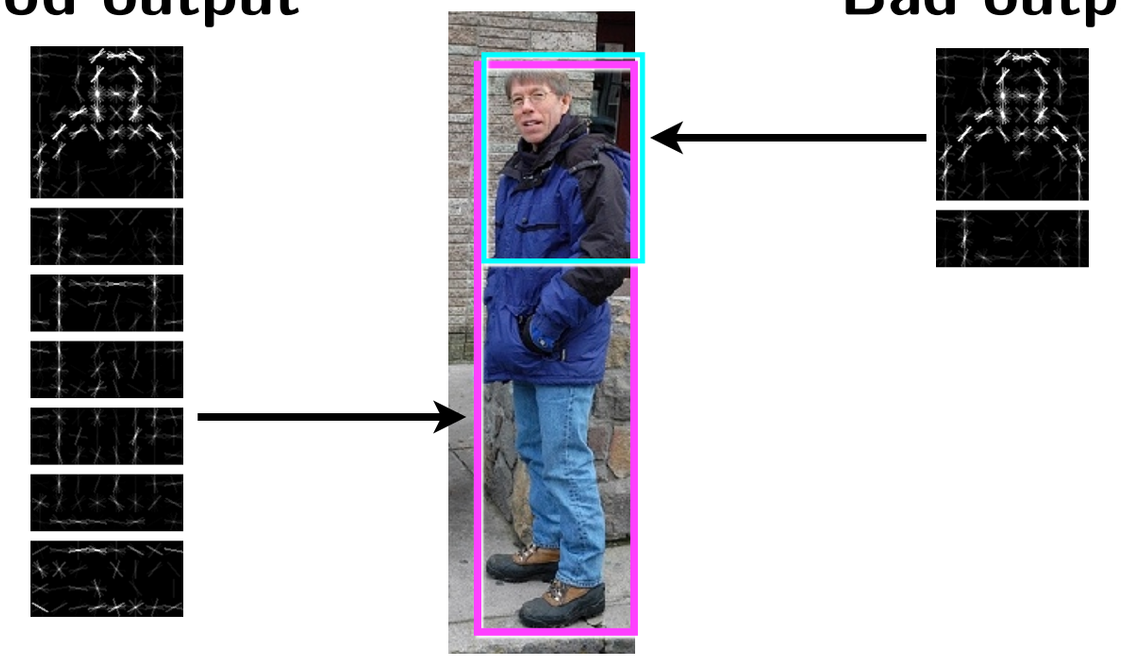
Who wins?

Both derivations were trained to score +1.

Structured output training

Good output

Bad output

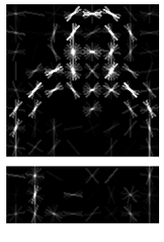


Training:

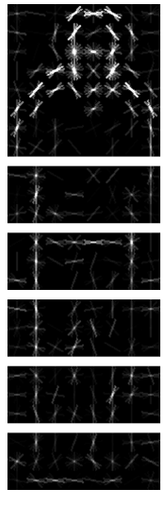
"outscore *all* other outputs by a margin"

"score lower by a margin"

Testing:



vs.



A "good" output should win.

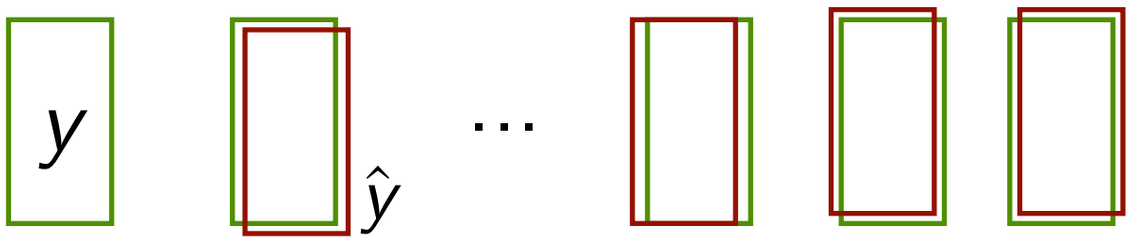
Latent structural SVM

[Yu and Joachims]

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L_{\text{lssvm}}(\mathbf{w}, x_i, y_i)$$

$$L_{\text{lssvm}}(\mathbf{w}, x, y) = \max_{(\hat{y},\hat{z})\in\mathcal{Y}\times\mathcal{Z}} [\mathbf{w}\cdot\boldsymbol{\psi}(x,\hat{y},\hat{z}) + L_{\text{margin}}(y,\hat{y})] - \max_{\hat{z}\in\mathcal{Z}} \mathbf{w}\cdot\boldsymbol{\psi}(x,y,\hat{z})$$

- Objective and task loss (L_{margin}) might be inconsistent

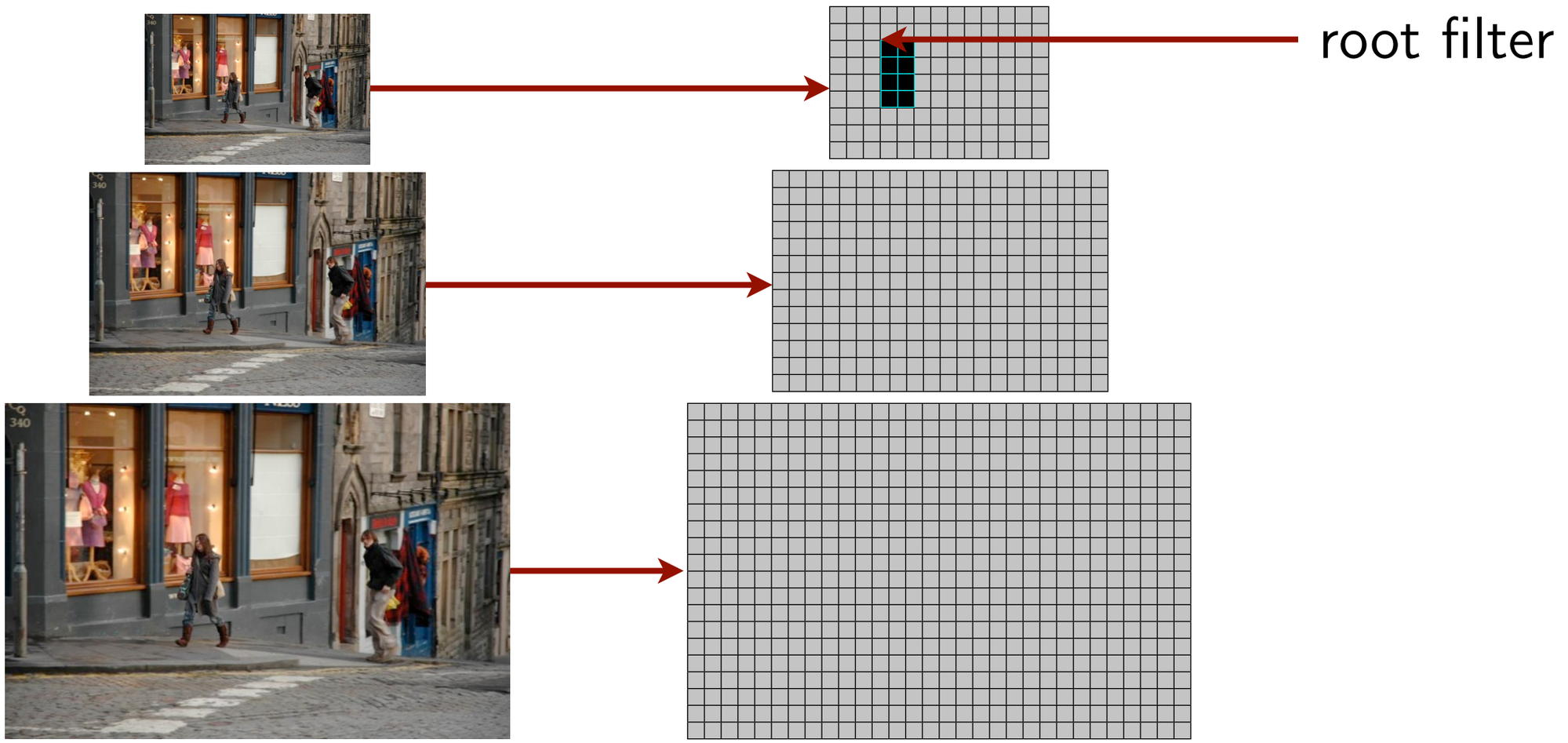


$$L_{\text{margin}}(x,y) = 0 \text{ and } L_{\text{margin}}(x,\hat{y}) = 0$$

  - Many outputs with zero loss — LSSVM "requires" the training label

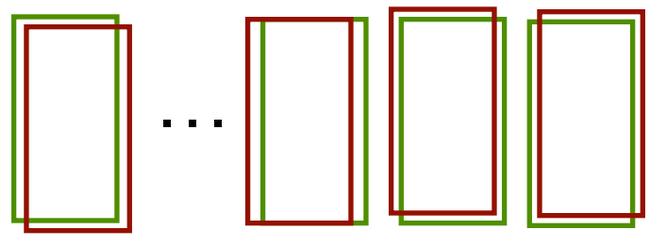
LSSVM requires label space = output space

- A simple example where label space != output space
  - Label space is ***all pixel-accurate*** bounding boxes
  - Outputs are bounding boxes ***on a low-res. grid at some scales***
  - Does not naturally fit the LSSVM framework



Structured learning desiderata

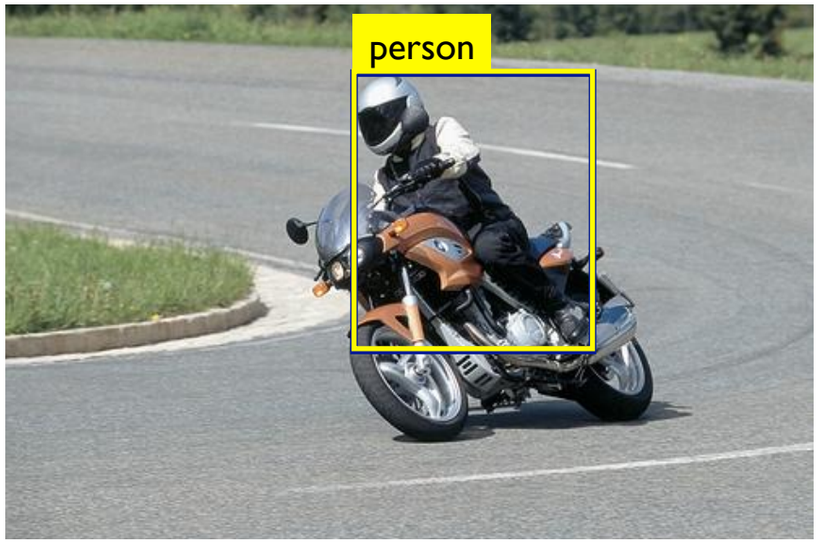
- Model can make any low-loss prediction
  - Many outputs might be compatible with one label
  - The model is free to choose between them



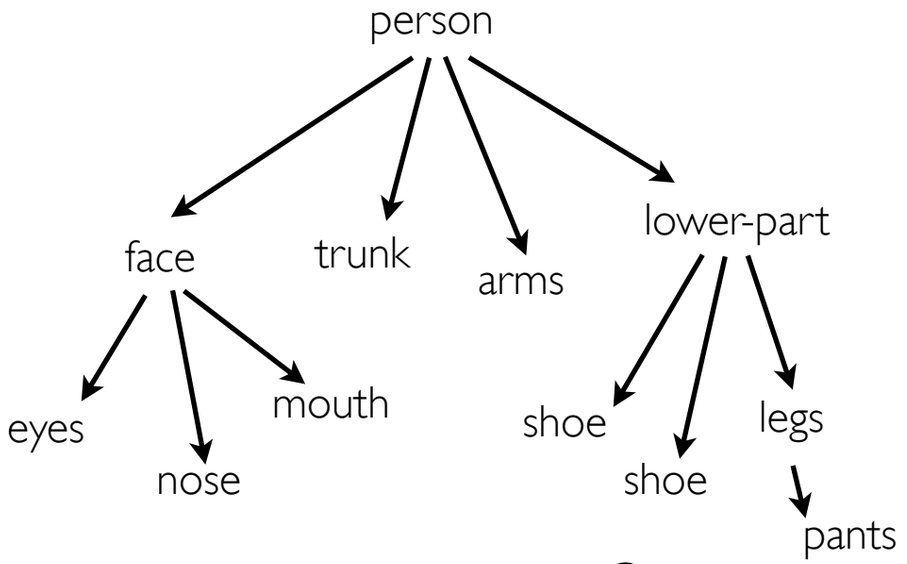
- Label space and output space can be different
  - *E.g.*, bounding boxes labels and derivation tree outputs
- Generalize frameworks that work well
  - Structural SVM
  - Latent structural SVM
  - Latent SVM

Label space != output space

- Allowing different label spaces and output spaces



label $\in \mathcal{Y}$



output $\in \mathcal{S}$

- Connect the spaces with loss functions of the form

$$L : \mathcal{Y} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$$

Weak-label structural SVM

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L(\mathbf{w}, x_i, y_i)$$

$$L(\mathbf{w}, x, y) = \max_{s \in \mathcal{S}(x)} \left[\mathbf{w} \cdot \psi(x, s) + L_{\text{margin}}(y, s)\right] - \max_{s \in \mathcal{S}(x)} \left[\mathbf{w} \cdot \psi(x, s) - L_{\text{output}}(y, s)\right]$$

- Allows different label spaces and output spaces
- Not "required" to predict the training label
  - Many outputs may be compatible with a label — labels are "weak"
  - The model can pick any output with low L_{output}
- Generalizes many frameworks
  - SSVM, LSSVM, LSVM, structural ramp loss

Person grammar results

- Person grammar results

	Grammar	+bbox	+context	UoC-TTI [26]	+bbox	+context	Poselets [26]
AP	47.5	47.6	**49.5**	44.4	45.2	47.5	48.5

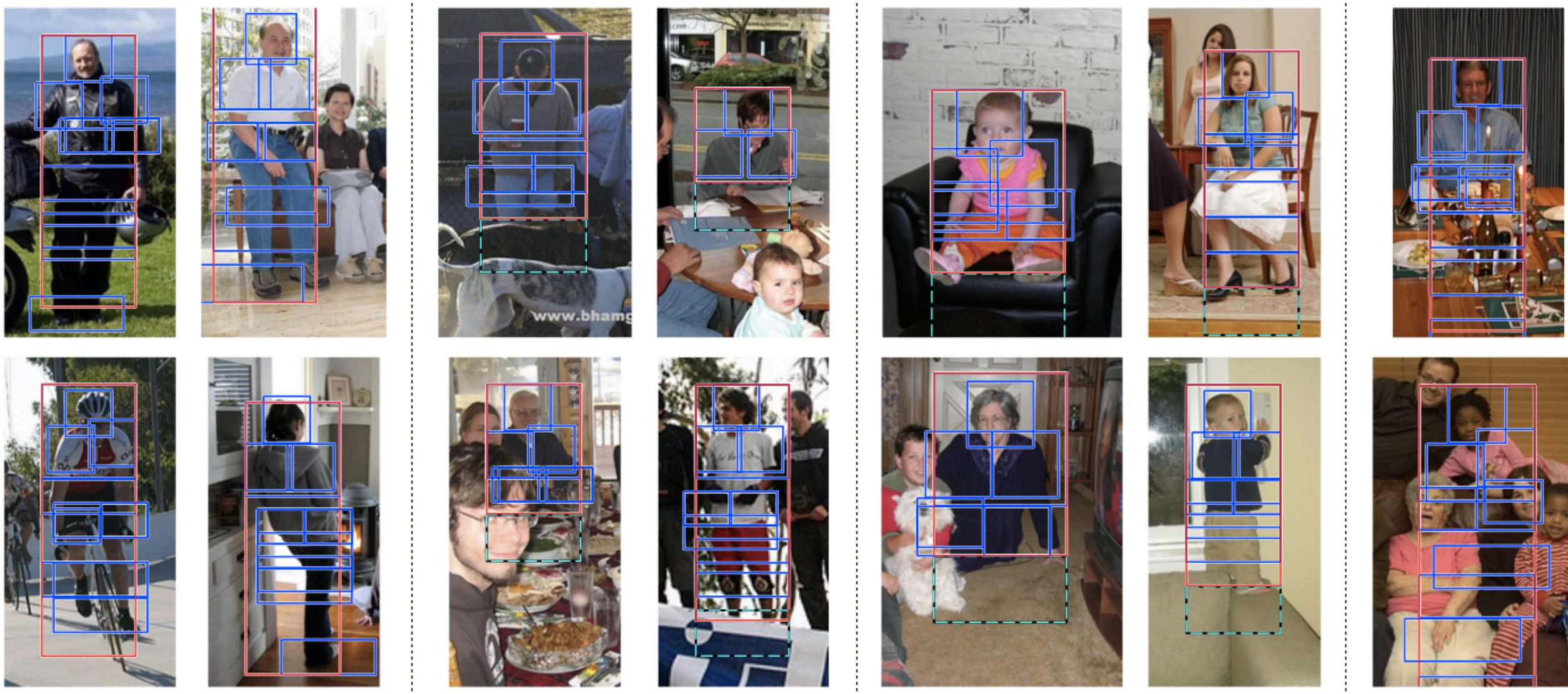
AP scores on PASCAL 2010

- WL-SSVM *vs.* LSVM

Person grammar objective	mAP	AP scores on splits 1	2	3	4	5
LSVM	46.0	47.5	46.3	45.3	45.3	45.6
WL-SSVM	48.0	49.0	48.4	48.4	48.0	46.5

AP scores on 5 PASCAL 2011 train+val splits

Example detections



(a) Full visibility (b) Occlusion boundaries (c) Early termination (d) Mistakes

Summary of contributions

- Richer models + post-processing + features + learning
  - $>$ 50% relative improvement in the state-of-the-art
- Cascaded detection for DPM
  - 15x speedup of detection with no loss in performance
- Person detection grammar and WL-SSVM
  - Highest-performing person detector
  - More general & natural learning framework for many problems
- Improved image features
- Detection post-processing
- Software
  - voc-release5 will be available soon!

Open directions

- Grammar structure learning
  - Perhaps from more detailed annotations
- Score compatibility and linear grammars
  - Nonlinearities to normalize score ranges — neural grammars?
- Rethink our low-level features
  - HOG features are too coarse to model fine detail
  - We are likely saturating the performance of HOG features
- Optimizing nonconvex objectives with latent variables
  - How can we free ourselves from careful (often model specific) initialization?

Challenges

- Subcategory — “car” images

Image credits: PASCAL VOC

Organizing principles

- Gradually build richer models
  - Central research methodology
- Compositional models
  - Object Detection Grammars [FM'10]
  - Deformation, viewpoint, subcategory, composition – in a unified framework
- Efficient computation
  - Tree-structured models
  - Cascaded detection
- Train models from weakly-labeled data
  - New models, old annotations

Preliminaries — Object detection grammars

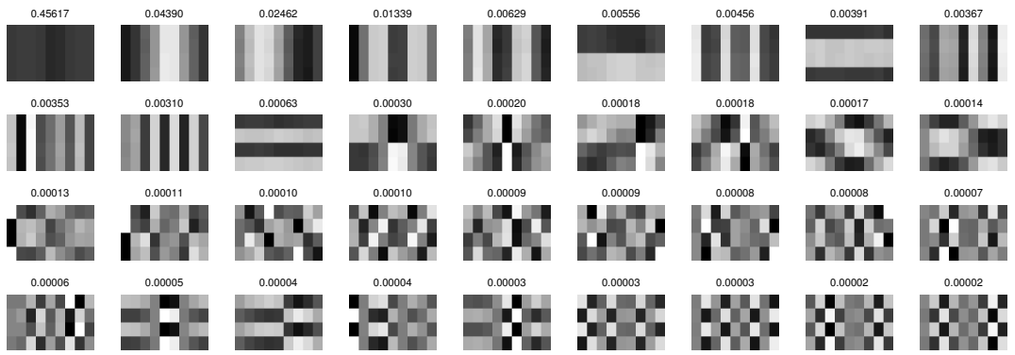
- Linear score function

$$\begin{aligned}
\text{score}(I, T) &= \sum_{(r,z)\in\text{int}(T)} \beta_r(z) + \sum_{A(\omega)\in\text{leaf}(T)} \text{score}(I, A, \omega) \\
&= \sum_{(r,z)\in\text{int}(T)} \mathbf{w}\cdot\boldsymbol{\phi}_r(z) + \sum_{A(\omega)\in\text{leaf}(T)} \mathbf{w}\cdot\boldsymbol{\phi}_A(I,\omega) \\
&= \mathbf{w}\cdot\left[\sum_{(r,z)\in\text{int}(T)} \boldsymbol{\phi}_r(z) + \sum_{A(\omega)\in\text{leaf}(T)} \boldsymbol{\phi}_A(I,\omega)\right] \\
&= \mathbf{w}\cdot\boldsymbol{\psi}(I, T).
\end{aligned}$$

- Detection = find high scoring derivations
  - Efficient dynamic programming algorithm

Image representation & features

- Enhanced HOG features
  - 36D $\rightarrow$ 31D with more information
  - Contrast sensitive and insensitive

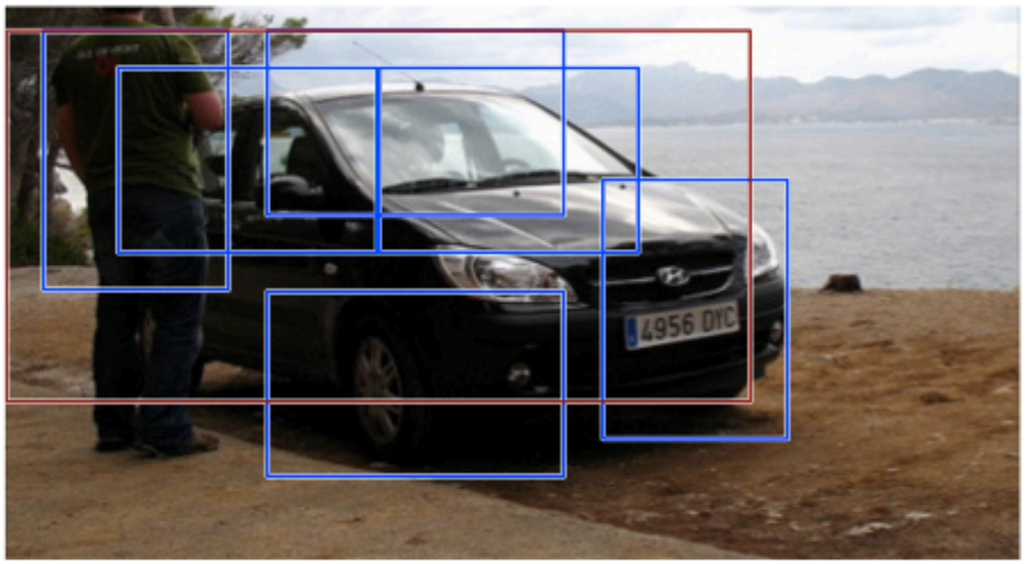


- Boundary truncation
  - Nonzero scores outside the image

- Small objects
  - Scale-dependent score bias

Detection post-processing

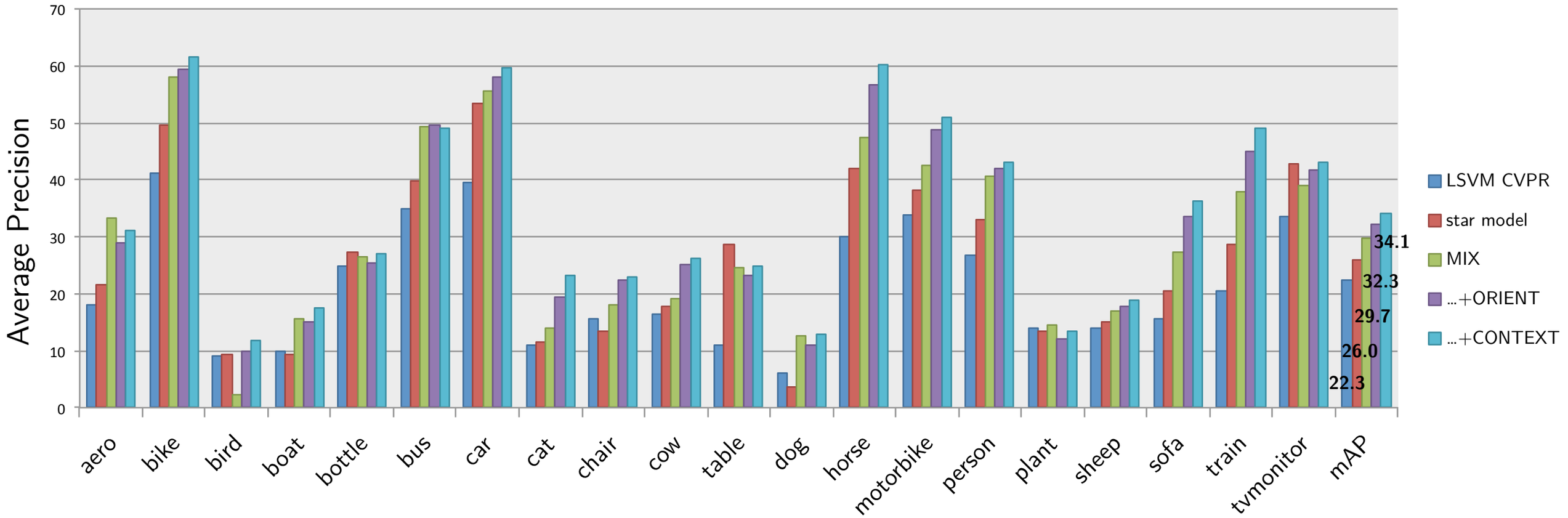
- Bounding box prediction



- Contextual information

Results – Mixture model and latent orientation

- Mixture models boost mAP 3.7 points
- Latent orientation boost mAP 2.6 AP points



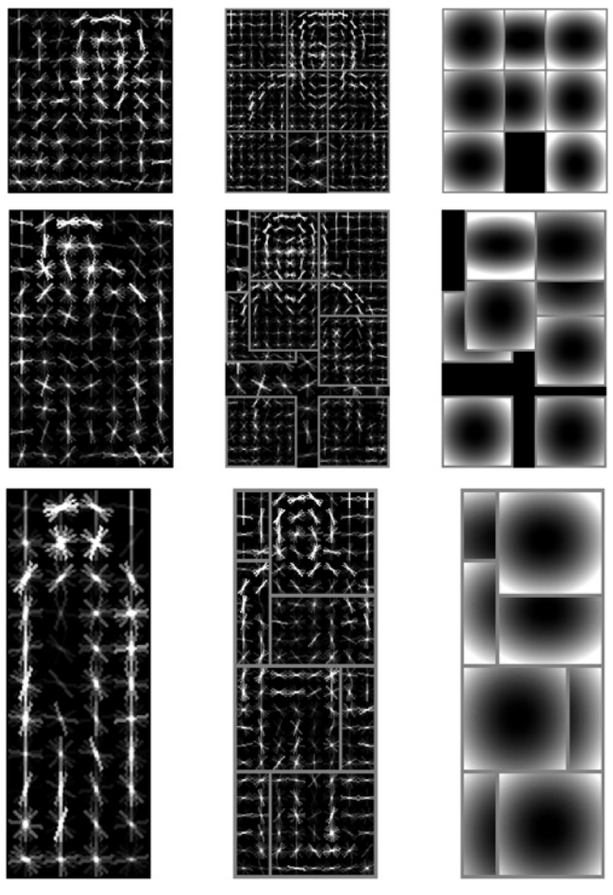
AP scores using the PASCAL 2007 evaluation

- 12 AP point improvement (>50% relative) over the baseline

Structure learning

What's the model class?

Number of components?



Number of parts?

Root filter sizes?

Anchor positions?

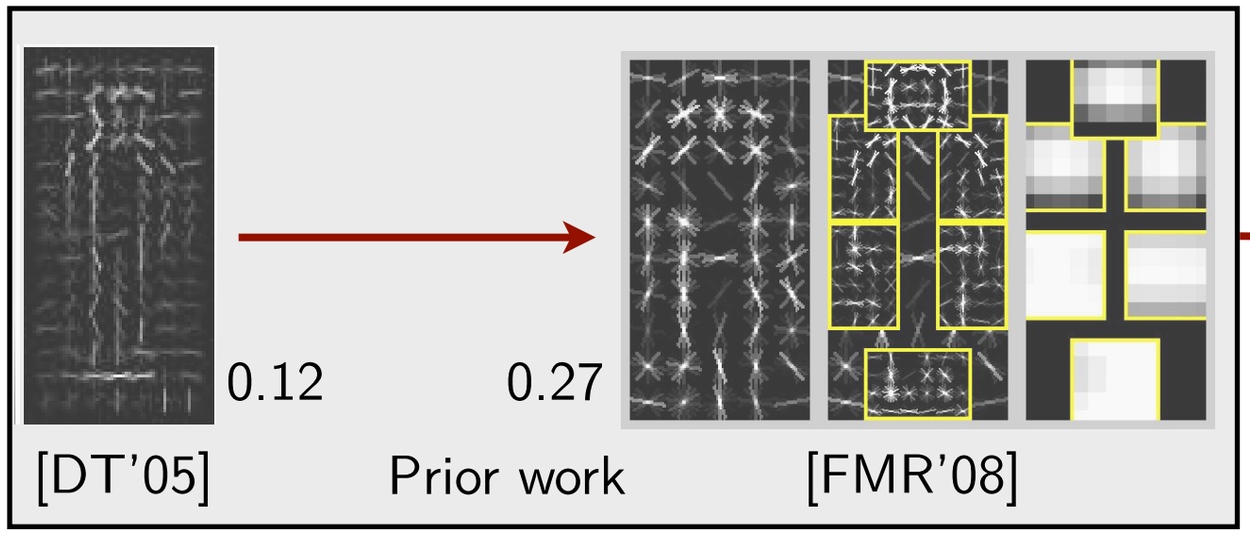
Root filter shapes?

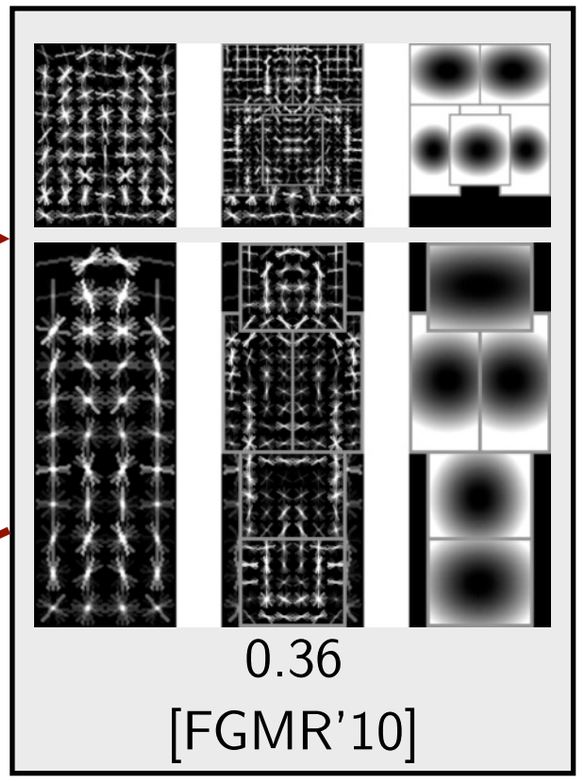
Part shapes and sizes?

What are the grammar productions?

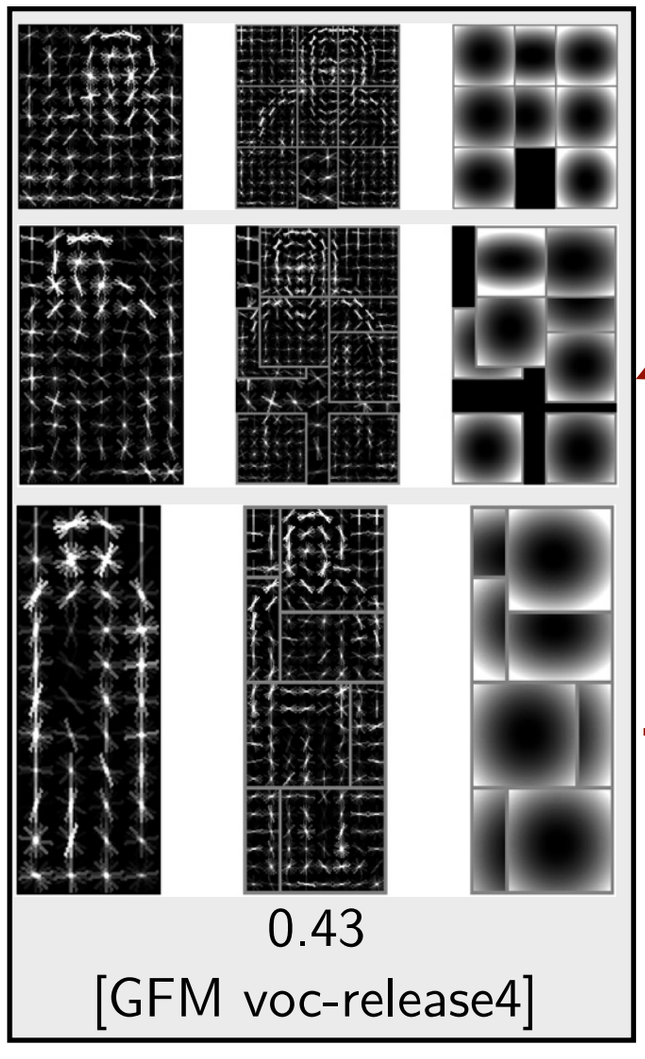
Heuristics, cross validation, insight (from humans)

Object representation — Summary

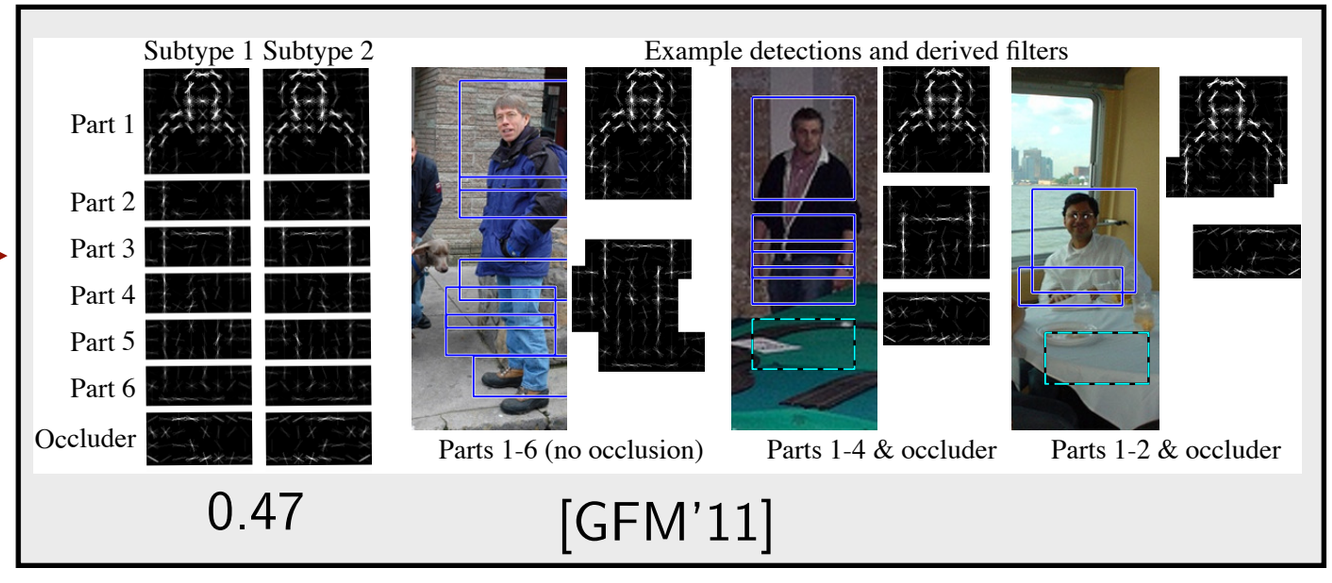




Mixture models



Latent orientation



Person detection grammar WL-SSVM

Structural SVM

[Tsochantaridis et al., Taskar et al.]

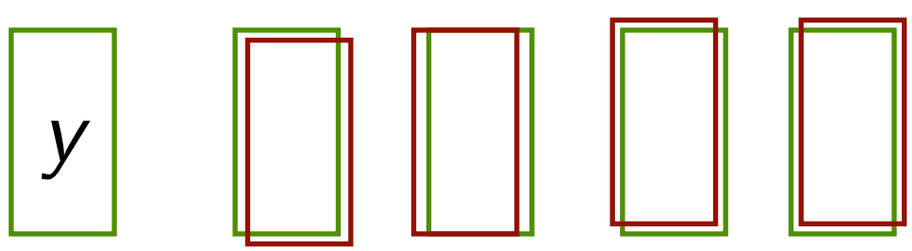
- No latent variables

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L_{\text{ssvm}}(\mathbf{w}, x_i, y_i)$$

$$L_{\text{ssvm}}(\mathbf{w}, x, y) = \max_{\hat{y}\in\mathcal{Y}} \left[\mathbf{w}\cdot\boldsymbol{\psi}(x,\hat{y}) + L_{\text{margin}}(y,\hat{y})\right] - \mathbf{w}\cdot\boldsymbol{\psi}(x,y)$$

- Objective and task loss (L_{margin}) might be inconsistent
  - Two outputs with zero loss — SSVM "requires" the training label

$$L_{\text{margin}}(\mathbf{w}, x, y) = 0 \text{ and } L_{\text{margin}}(\mathbf{w}, x, \hat{y}) = 0$$



Optimizing WL-SSVM

- Use the convex-concave procedure

$$E(\mathbf{w}) = E_{\text{convex}}(\mathbf{w}) + E_{\text{concave}}(\mathbf{w})$$

$$E_{\text{convex}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max_{s \in \mathcal{S}(x_i)} [\mathbf{w} \cdot \psi(x_i, s) + L_{\text{margin}}(y_i, s)]$$

$$E_{\text{concave}}(\mathbf{w}) = -C\sum_{i=1}^{n} \max_{s \in \mathcal{S}(x_i)} [\mathbf{w} \cdot \psi(x_i, s) - L_{\text{output}}(y_i, s)]$$

- Sequence of convex slave problems

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max_{s \in \mathcal{S}(x_i)} [\mathbf{w} \cdot \psi(x_i, s) + L_{\text{margin}}(y_i, s)] - \mathbf{w} \cdot \psi(x_i, s_i(\mathbf{w}_t))$$

Object detection grammars

- A modeling language for building object detectors [FM'10]
  - Terminals (model image data appearance)
  - Nonterminals (objects, parts, ...)
  - Weighted production rules (define compositions, subtypes)
- Composition

$$X \xrightarrow{\beta} \{\ Y_1, \ldots, Y_n\ \}$$

$$\text{Person} \xrightarrow{\beta} \{\ \text{Head}, \text{Arms}, \text{Torso}, \ldots\ \}$$

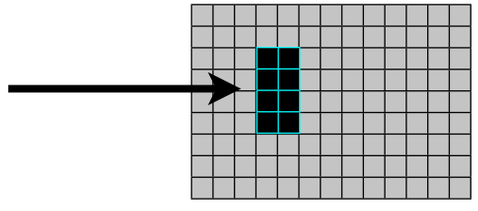
$$\text{Head} \xrightarrow{\beta'} \{\ \text{Eye}, \text{Eye}, \text{Mouth}\ \}$$

- Subtypes (choice yields variable structure)

$$X \xrightarrow{\beta_1} \{\ W_1, \ldots, W_n\ \}$$

$$X \xrightarrow{\beta_2} \{\ Y_1, \ldots, Y_m\ \}$$

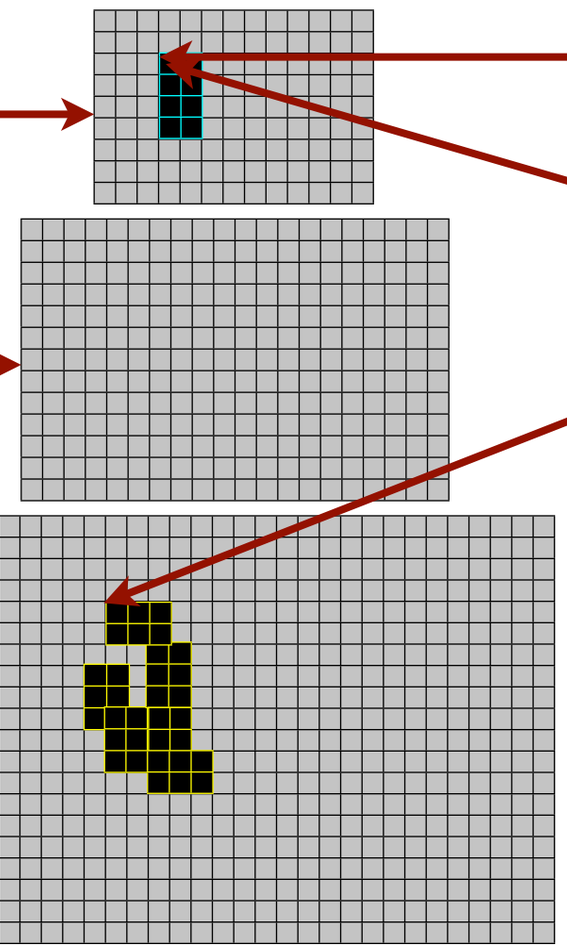
X is a "smiling face" or a "frowning face"

- Symbols are placed $Y(p)$ 

Object detection grammars

- Symbols are placed
- Terminals model appearance (HOG filters)

Image pyramid

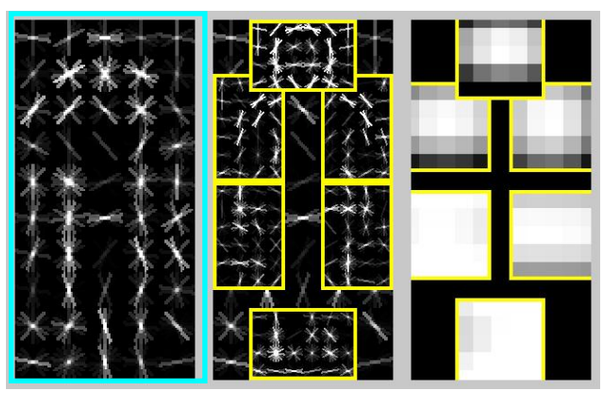


HOG feature pyramid

Person(x,y,l)

Root(x,y,l)

Part$_i(x_i,y_i,l_i)$



[FMR'08]

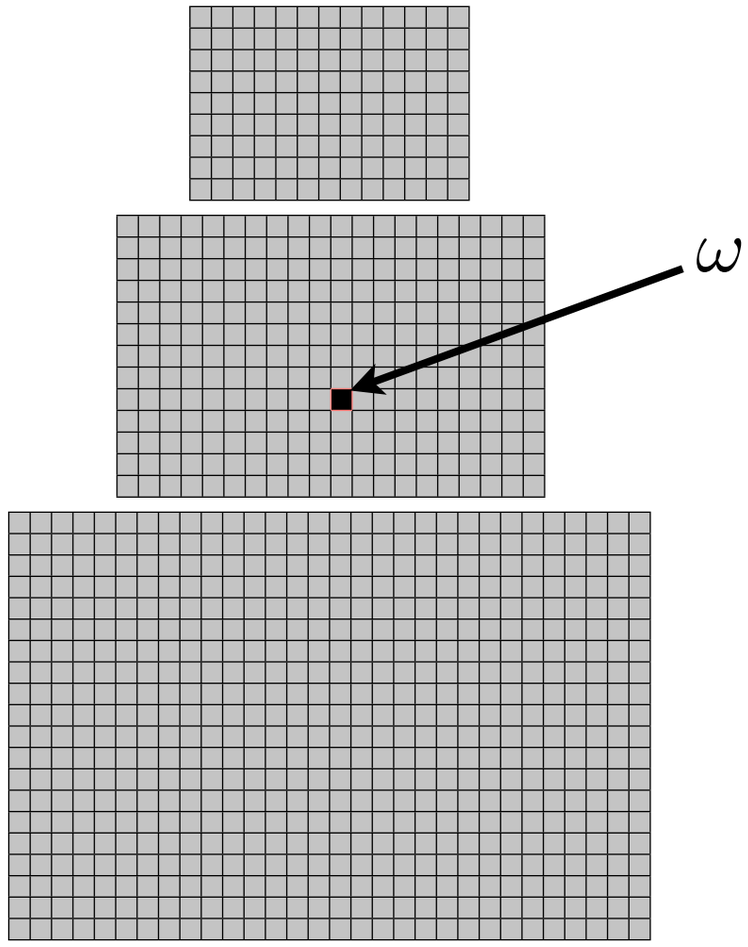
Definition of $S(x)$ – foreground examples

Example (x, y) where x specifies an image I and $y = B$ is a bounding box. Let $\mathcal{S}(x)$ be:

1. Derivations T with $\text{overlap}(\text{box}(T), B) \geq 0.1$ *and* $\text{overlap}(\text{box}(T), B') < 0.5$ for all B' in I such that $B' \neq B$.
2. The background output $\perp$.

Definition of $S(x)$ – background examples

Example (x, y) where x specifies an image I and position ω and $y = \perp$ is the background symbol.
Let $\mathcal{S}(x) = \mathcal{T}_\omega \cup \{\perp\}$.



HOG feature pyramid