Exemplars for Object Detection

Noah Snavely

CS7670: September 5, 2011

Announcements

- Office hours: Thursdays 1pm – 2:30pm
- Course schedule is now online

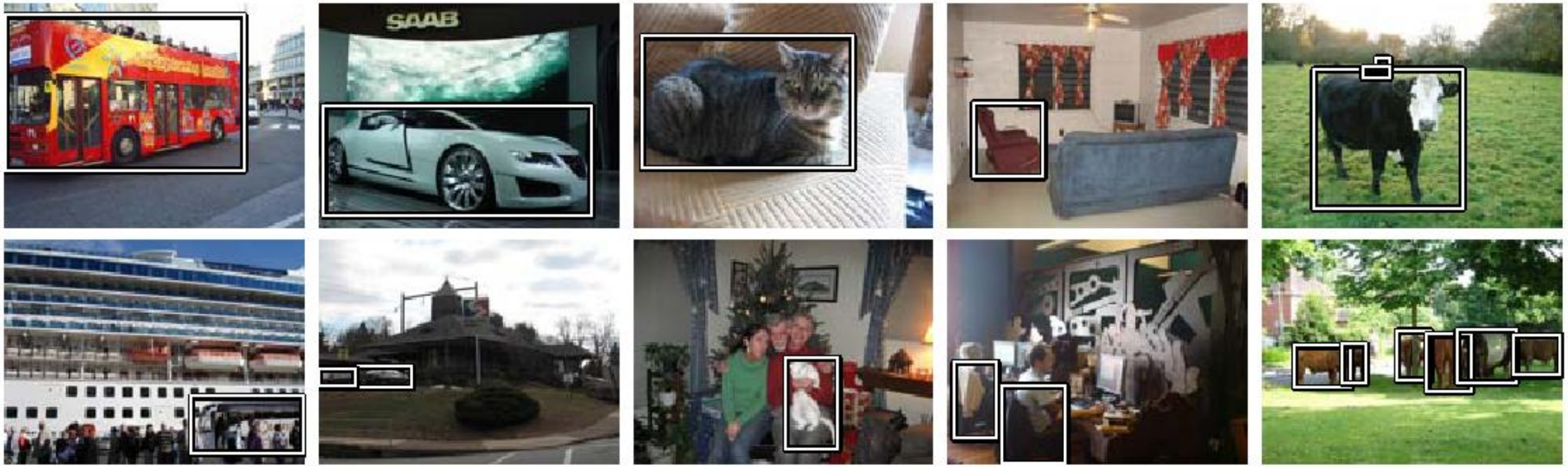
Object detection: where are we?

Credit: Flickr user neilalderney123

- Incredible progress in the last ten years
- Better features, better models, better learning methods, better datasets
- Combination of science and hacks

The 800-lb Gorilla of Vision Contests

- PASCAL VOC Challenge



- 20 categories
- Annual classification, detection, segmentation, … challenges

Object detection performance (2010)

Average Precision (AP %)

	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog
BONN_FGT_SEGM	52.7	33.7	13.2	11.0	14.2	43.1	31.9	35.6	5.7	25.4	14.4	20.6
BONN_SVR_SEGM	50.5	24.4	17.1	13.3	10.9	39.5	32.9	36.5	5.6	16.0	6.6	22.3
CMIC_SYNTHTRAIN	-	28.9	-	-	-	30.2	13.3	-	-	-	-	-
CMIC_VARPARTS	-	28.2	-	-	-	26.9	13.7	-	-	-	-	-
CMU_RANDPARTS	23.8	31.7	1.2	3.4	11.1	29.7	19.5	14.2	0.8	11.1	7.0	4.7
CMU_RANDPARTS_MAXSCORE	-	-	2.7	-	-	-	-	16.2	-	10.6	8.5	-
LJKINPG_HOG_LBP_LTP_PLS2ROOTS	32.7	29.7	0.8	1.1	19.8	39.4	27.5	8.6	4.5	8.1	6.3	11.0
MITUCLA_HIERARCHY	54.2	48.5	15.7	19.2	29.2	55.5	43.5	41.7	16.9	28.5	26.7	30.9
NLPR_HOGLBP_MC_LCEGCHLC	53.3	55.3	19.2	21.0	30.0	54.4	46.7	41.2	20.0	31.5	20.7	30.3
NUS_HOGLBP_CTX_CLS_RESCORE_V2	49.1	52.4	17.8	12.0	30.6	53.5	32.8	37.3	17.7	30.6	27.7	29.5
TIT_SIFT_GMM_MKL	10.5	1.6	1.2	0.9	0.1	2.8	1.6	6.7	0.1	2.0	0.4	3.0
TIT_SIFT_GMM_MKL2	20.0	14.5	3.8	1.2	0.5	17.6	8.1	28.5	0.1	2.9	3.1	17.5
UC3M_GENDISC	15.8	5.5	5.6	2.3	0.3	10.2	5.4	12.6	0.5	5.6	4.5	7.7
UCI_DPM_SP	46.1	52.6	13.8	15.5	28.3	53.2	44.5	26.6	17.6	-	16.1	20.4
UMNECUIUC_HOGLBP_DHOGBOW_SVM	40.4	34.7	2.7	8.4	26.0	43.1	33.8	17.2	11.2	14.3	14.4	14.9
UMNECUIUC_HOGLBP_LINSVM	37.9	33.7	2.7	6.5	25.3	37.5	33.1	15.5	10.9	12.3	12.5	13.7
UOCTTI_LSVM_MDPM	52.4	54.3	13.0	15.6	35.1	54.2	49.1	31.8	15.5	26.2	13.5	21.5
UVA_DETMONKEY	56.7	39.8	16.8	12.2	13.8	44.9	36.9	47.7	12.1	26.9	26.5	37.2
UVA_GROUPLOC	58.4	39.6	18.0	13.3	11.1	46.4	37.8	43.9	10.3	27.5	20.8	36.0

Object detection performance (2010)

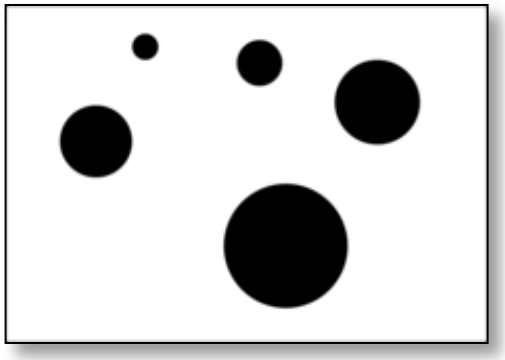
	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor
BONN_FGT_SEGM	38.1	41.7	25.0	5.8	26.3	18.1	37.6	28.1
BONN_SVR_SEGM	24.9	29.0	29.8	6.7	28.4	13.3	32.1	27.2
CMIC_SYNTHTRAIN	26.2	28.1	13.2	-	-	-	18.8	25.7
CMIC_VARPARTS	23.5	24.7	16.1	-	-	-	18.8	24.5
CMU_RANDPARTS	16.4	31.5	16.0	1.1	15.6	10.2	14.7	21.0
CMU_RANDPARTS_MAXSCORE	-	-	17.9	-	-	-	15.7	-
LJKINPG_HOG_LBP_LTP_PLS2ROOTS	22.9	34.1	24.6	3.1	24.0	2.0	23.5	27.0
MITUCLA_HIERARCHY	48.3	55.0	41.7	9.7	35.8	30.8	47.2	40.8
NLPR_HOGLBP_MC_LCEGCHLC	48.6	55.3	46.5	10.2	34.4	26.5	50.3	40.3
NUS_HOGLBP_CTX_CLS_RESCORE_V2	51.9	56.3	44.2	9.6	14.8	27.9	49.5	38.4
TIT_SIFT_GMM_MKL	2.0	4.4	2.0	0.3	1.1	1.2	2.1	1.9
TIT_SIFT_GMM_MKL2	7.2	18.8	3.3	0.8	2.9	6.3	7.6	1.1
UC3M_GENDISC	11.3	12.6	5.3	1.5	2.0	5.9	9.1	3.2
UCI_DPM_SP	45.5	51.2	43.5	11.6	30.9	20.3	47.6	-
UMNECUIUC_HOGLBP_DHOGBOW_SVM	31.8	37.3	30.0	6.4	25.2	11.6	30.0	35.7
UMNECUIUC_HOGLBP_LINSVM	29.6	34.5	33.8	7.2	22.9	9.9	28.9	34.1
UOCTTI_LSVM_MDPM	45.4	51.6	47.5	9.1	35.1	19.4	46.6	38.0
UVA_DETMONKEY	42.1	51.9	25.7	12.1	37.8	33.0	41.5	41.7
UVA_GROUPLOC	39.4	48.5	22.9	13.0	36.8	30.5	41.2	41.9

The 2011 server opened for submissions today!

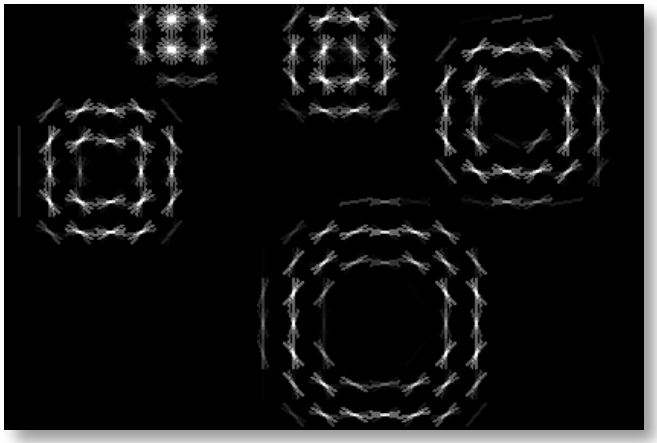
Machine learning for object detection

- What features do we use?
  - intensity, color, gradient information, …
- Which machine learning methods?
  - generative vs. discriminative
  - k-nearest neighbors, boosting, SVMs, …
- What hacks do we need to get things working?

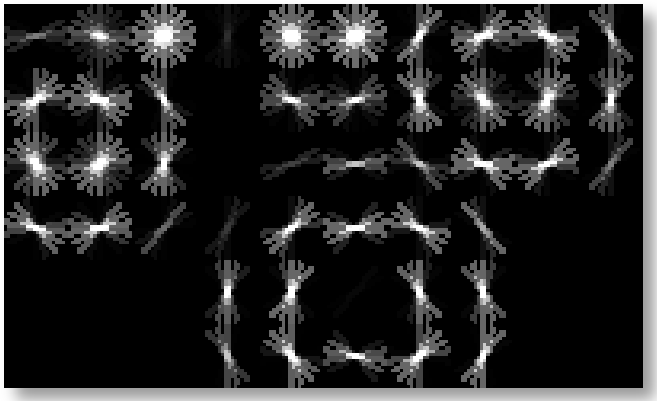
Histogram of Oriented Gradients (HoG)



HoGify



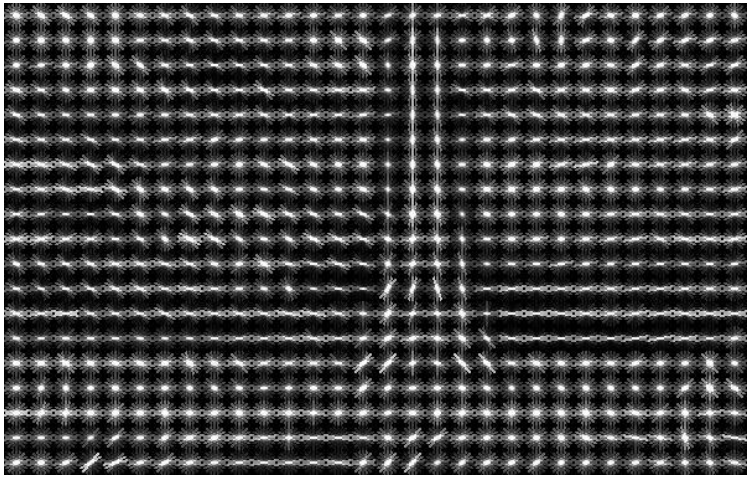
10x10 cells



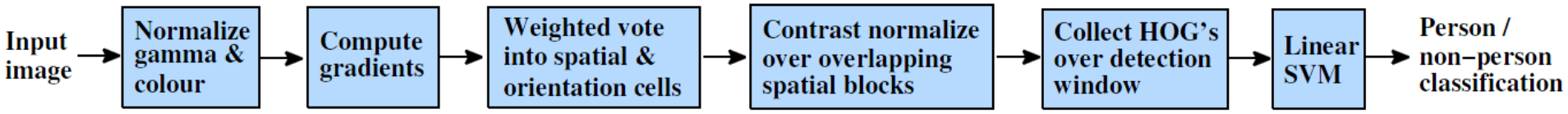
20x20 cells

[Dalal and Triggs, CVPR 2005]

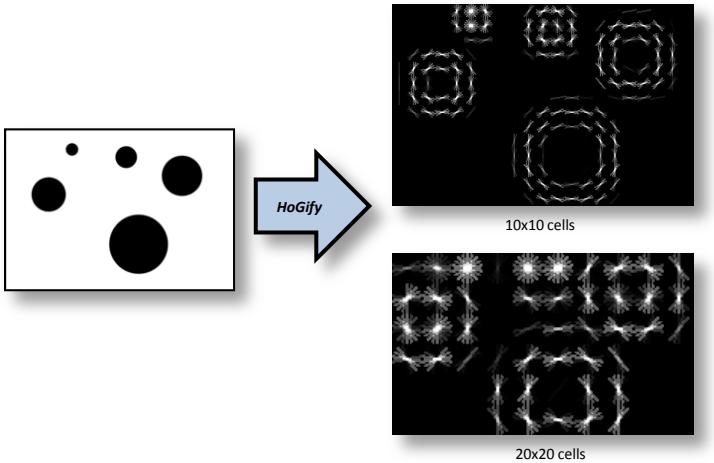
Histogram of Oriented Gradients (HoG)



Histogram of Oriented Gradients (HoG)



- Like SIFT (Scale Invariant Feature Transform), but…
  - Sampled on a dense, regular grid
  - Gradients are contrast normalized in overlapping blocks



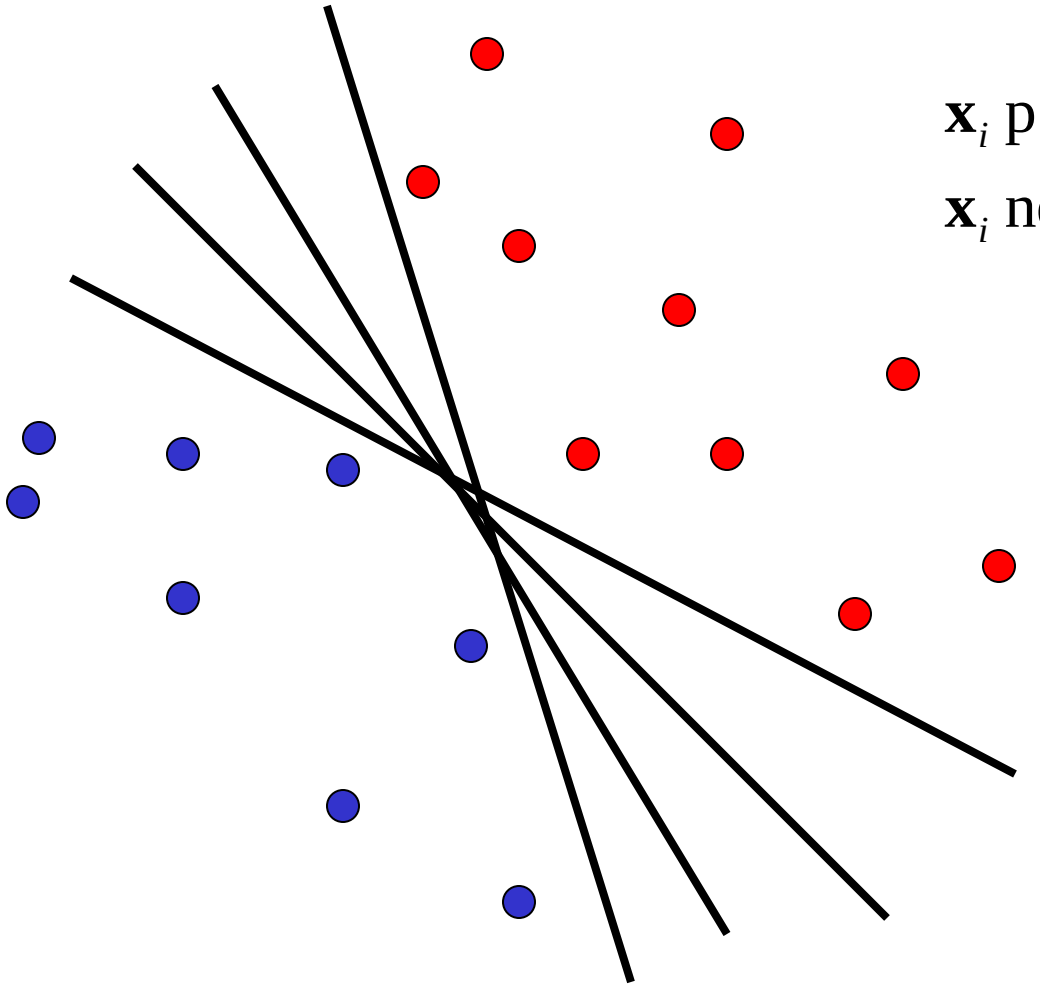
[Dalal and Triggs, CVPR 2005]

Histogram of Oriented Gradients (HoG)

- First used for application of person detection [Dalal and Triggs, CVPR 2005]
- Cited since in thousands of computer vision papers

Linear classifiers

- Find linear function to separate positive and negative examples



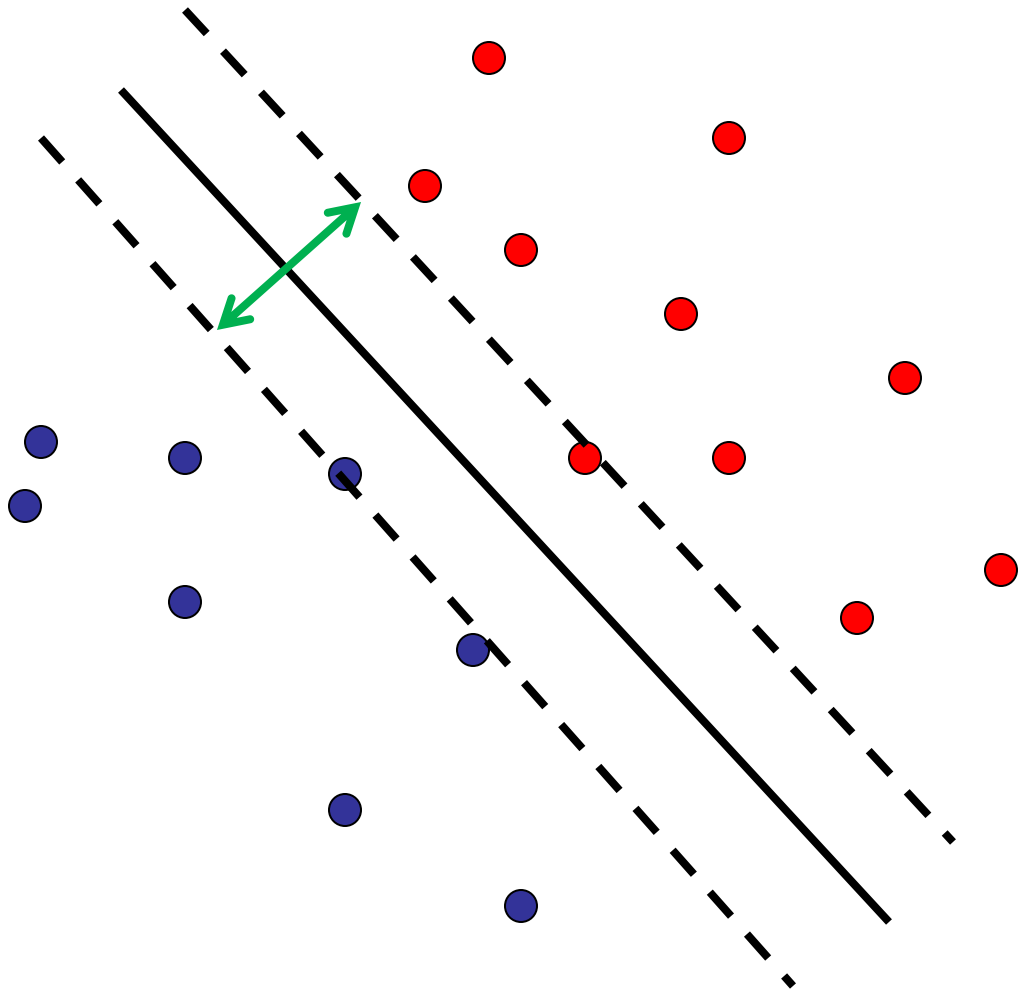
$\mathbf{x}_i$ positive: $\mathbf{x}_i \cdot \mathbf{w} + b \geq 0$

$\mathbf{x}_i$ negative : $\mathbf{x}_i \cdot \mathbf{w} + b < 0$

Which line is best?

[slide credit: Kristin Grauman]

Support Vector Machines (SVMs)

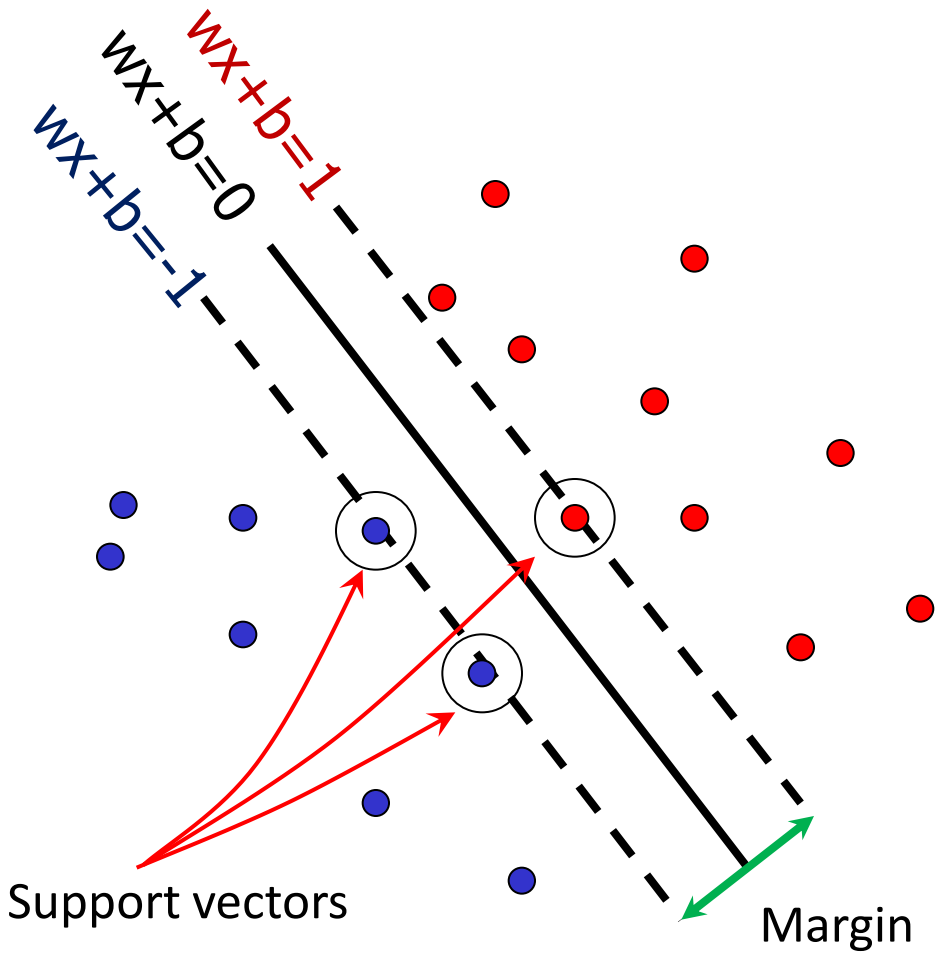


- Discriminative classifier based on *optimal separating line (for 2D case)*
- Maximize the *margin* between the positive and negative training examples

[slide credit: Kristin Grauman]

Support vector machines

- Want line that maximizes the margin.



$\mathbf{x}_i$ positive $(y_i = 1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \geq 1$

$\mathbf{x}_i$ negative $(y_i = -1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \leq -1$

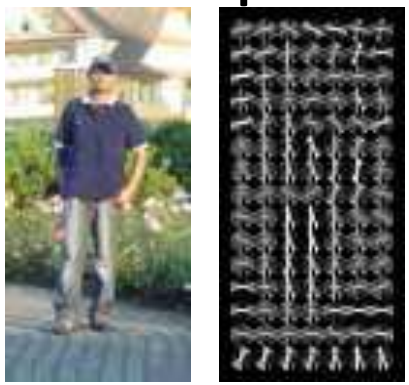
For support, vectors, $\mathbf{x}_i \cdot \mathbf{w} + b = \pm 1$

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

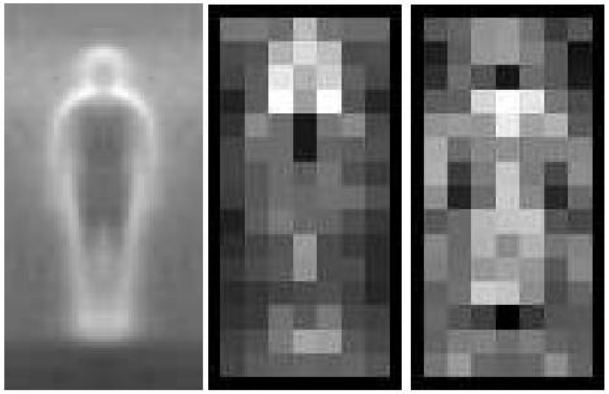
[slide credit: Kristin Grauman]

Person detection, ca. 2005

1. Represent each example with a single, fixed HoG template



2. Learn a single [linear] SVM as a detector



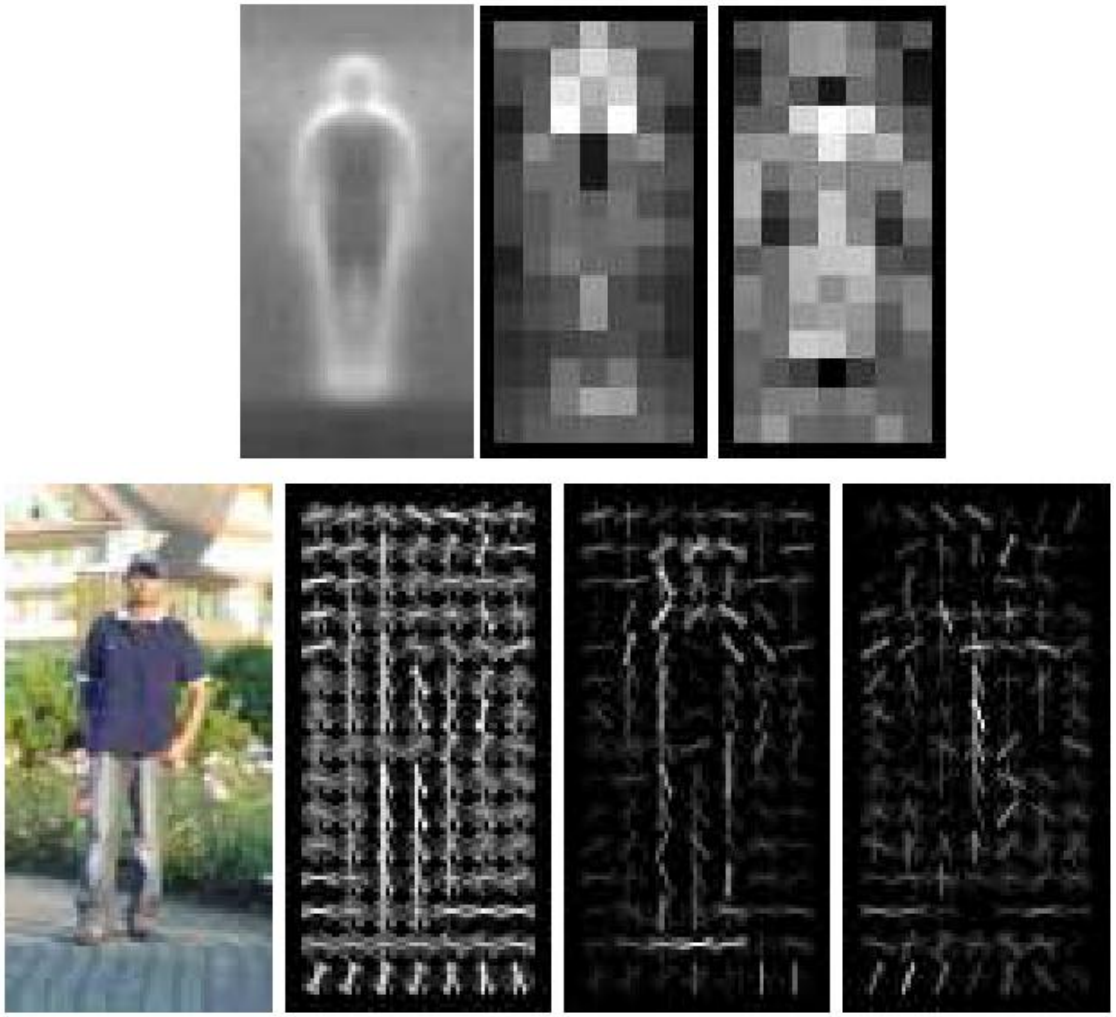
Code available: http://pascal.inrialpes.fr/soft/olt/

Positive and negative examples

+ thousands more…

+ millions more…

HoG templates for person detection



Person detection with HoG & linear SVM

[Dalal and Triggs, CVPR 2005]

Are we done?

Are we done?

- Single, rigid template usually not enough to represent a category
  - Many objects (e.g. humans) are articulated, or have parts that can vary in configuration

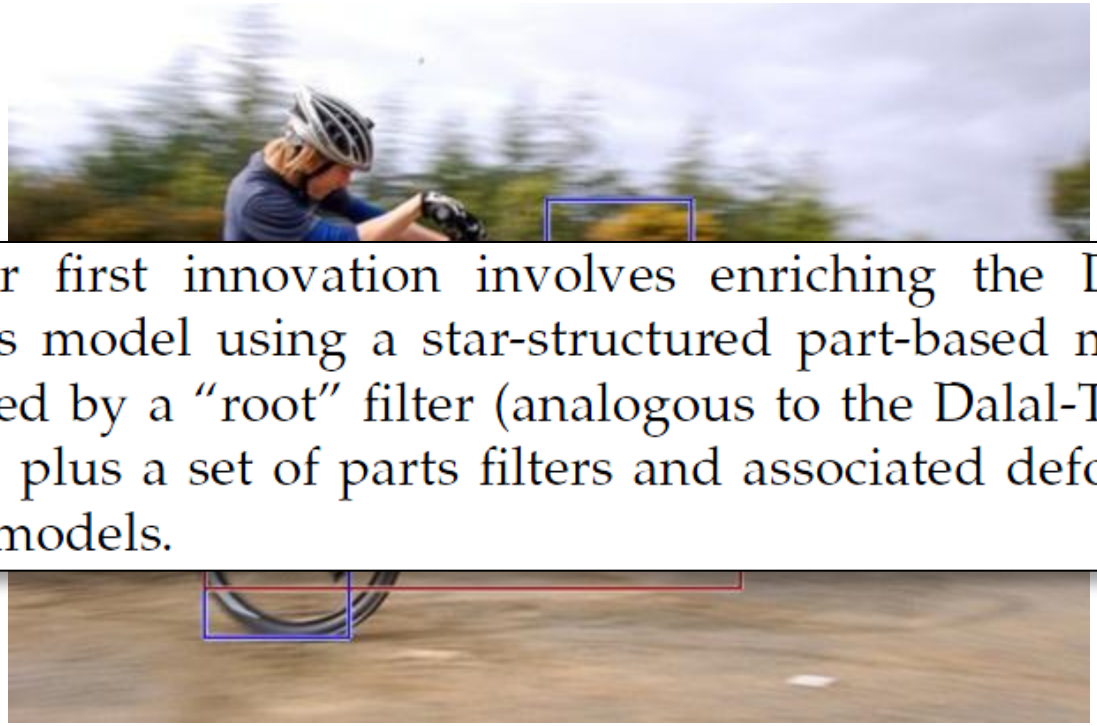
  - Many object categories look very different from different viewpoints, or from instance to instance

Difficulty of representing positive instances

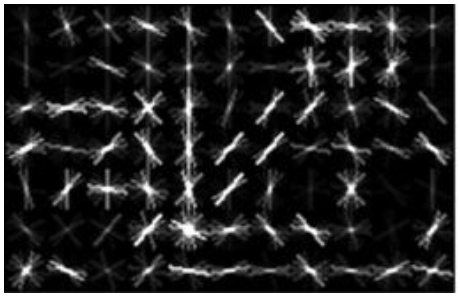
- Discriminative methods have proven very powerful
- But linear SVM on HoG templates not sufficient?

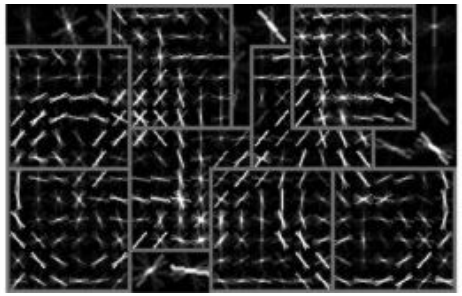
- Alternatives:
  - Parts-based models [Felzenszwalb et al. CVPR 2008]
  - Latent SVMs [Felzenszwalb et al. CVPR 2008]
  - Today's paper [Exemplar-SVMs, Malisiewicz, et al. ICCV 2011]

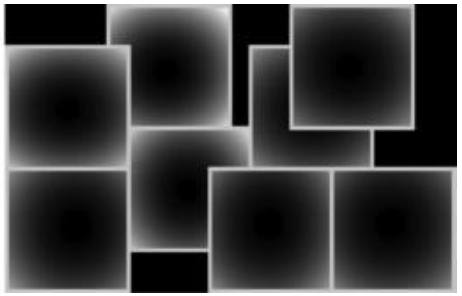
Parts-based models



Our first innovation involves enriching the Dalal-Triggs model using a star-structured part-based model defined by a "root" filter (analogous to the Dalal-Triggs filter) plus a set of parts filters and associated deformation models.





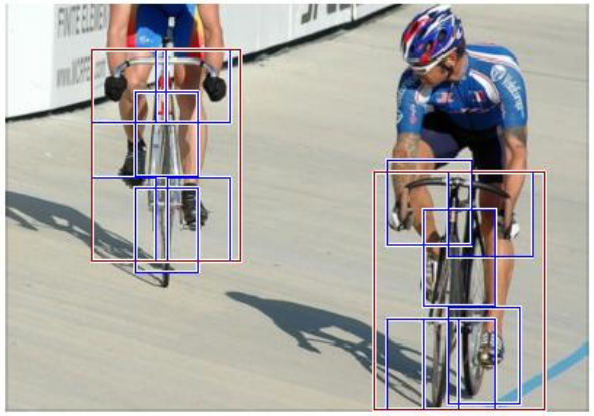


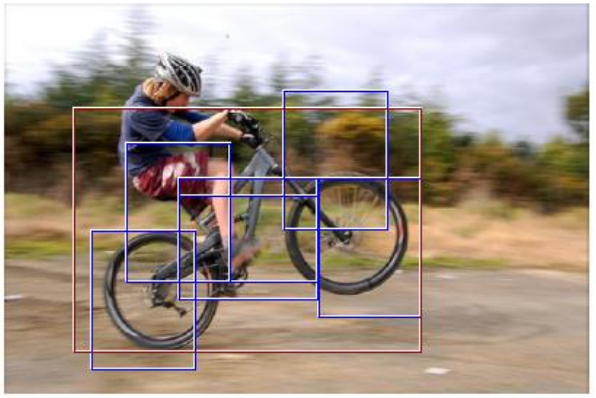
Felzenszwalb, et al., **Discriminatively Trained Deformable Part Models,**
http://people.cs.uchicago.edu/~pff/latent/

Latent SVMs

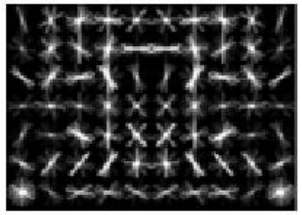
- Rather than training a single linear SVM separating positive examples…
- … cluster positive examples into "components" and train a classifier for each (using all negative examples)

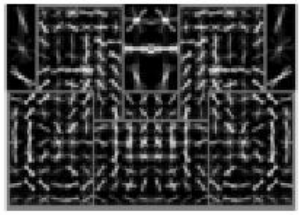
Two-component bicycle model

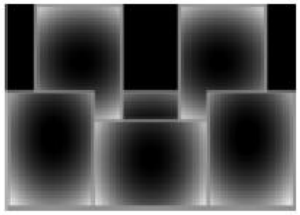




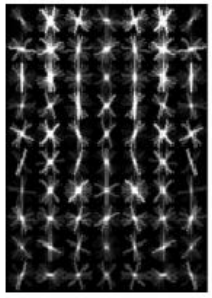
“side” component

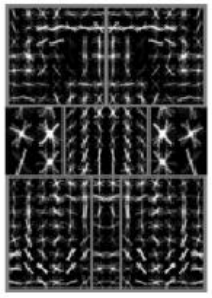


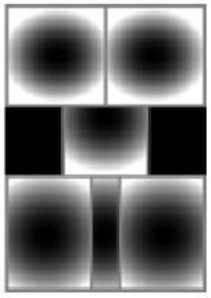




“frontal” component







Latent SVMs

Our second class of models represents an object category by a mixture of star models. The score of a mixture model at a particular position and scale is the maximum over components, of the score of that component model at the given location. In this case the latent information, z, specifies a component label and a configuration for that component.

- *Latent* because component labels are unknown in advance

Training of Latent SVMs

- Components are initialized by clustering positive instances by **bounding box aspect ratio**
- Linear SVM is learned for each component
- Each positive instance reassigned to the component that gives the max SVM response
- SVMs are retrained, and the process repeats

- Before training, training data is doubled through flipping

[Felzenszwalb et al., PAMI 2010]

Exemplar-SVMs for Object Detection

- Brings us to today…
- Why do discriminative techniques work so well?
  - When there are 100s of millions of training instances, kNN is infeasible
  - Parametric classifiers very good at generalizing from millions of negative examples
  - This paper's claim: parametric classifiers ***aren't*** the right way to represent positive examples

Ensemble of Exemplar-SVMs for Object Detection and Beyond. Malisiewicz, Gupta, Efros, ICCV 2011.

Representing positive examples

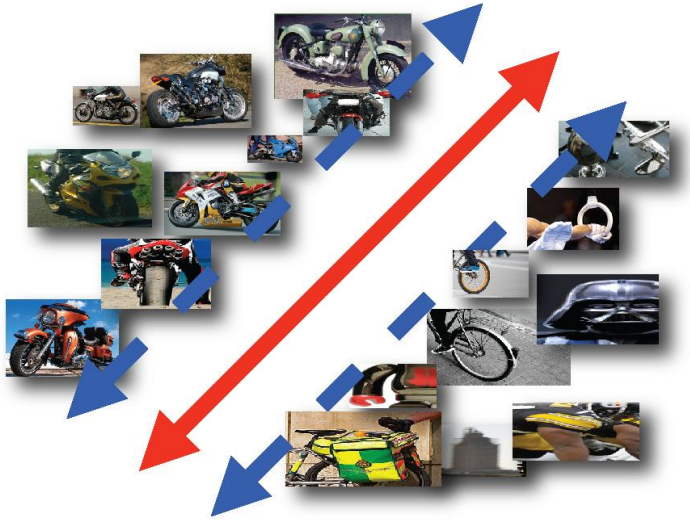
However, the parametric nature of these classifiers, while a blessing for handling negative data, becomes more problematic when representing the positives. Typically, all positive examples of a given object category are represented as a whole, implicitly assuming that they are all related to each other *visually*. Unfortunately, most standard *semantic* categories (e.g., "car", "chair", "train") do not form coherent *visual* categories [14], thus treating them parametrically results in weak and overly-generic detectors.

address this problem, a number of approaches have used semi-parametric mixture models, grouping the positives into clusters based on meta-data such as bounding box aspect ratio [9], object scale [15], object viewpoint [11], part labels [3], etc. But the low number of mixture components used in practice means that there is still considerable variation within each cluster. As a result, the alignment, or visual correspondence, between the learned model and a detected instance is too coarse to be usable for object association and label transfer. While part-based models [9] allow dif-

Exemplar-SVMs

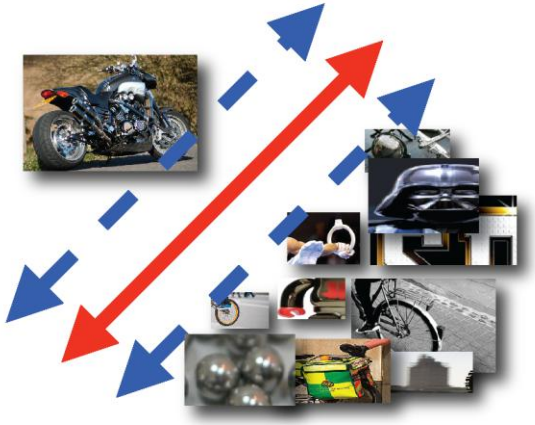
- This paper goes to the extreme, and learns a separate classifier for *every* positive example (and millions of negative examples)
- Each positive instance becomes an *exemplar* with an associated linear SVM; at test time each classifier is applied to a test image
- "Non-parametric when representing the positives, but parametric… when representing the negatives
- Allows for more accurate correspondence and information transfer

Category-SVM



vs.

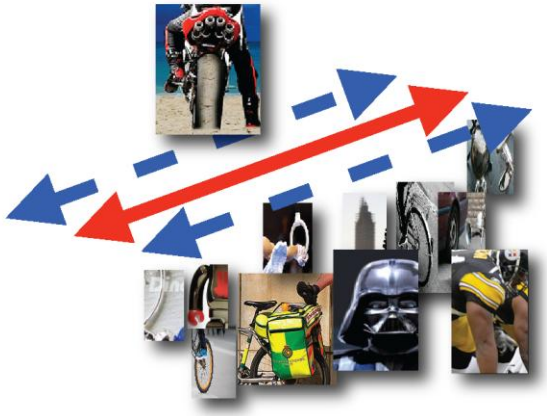
Exemplar-SVM 1



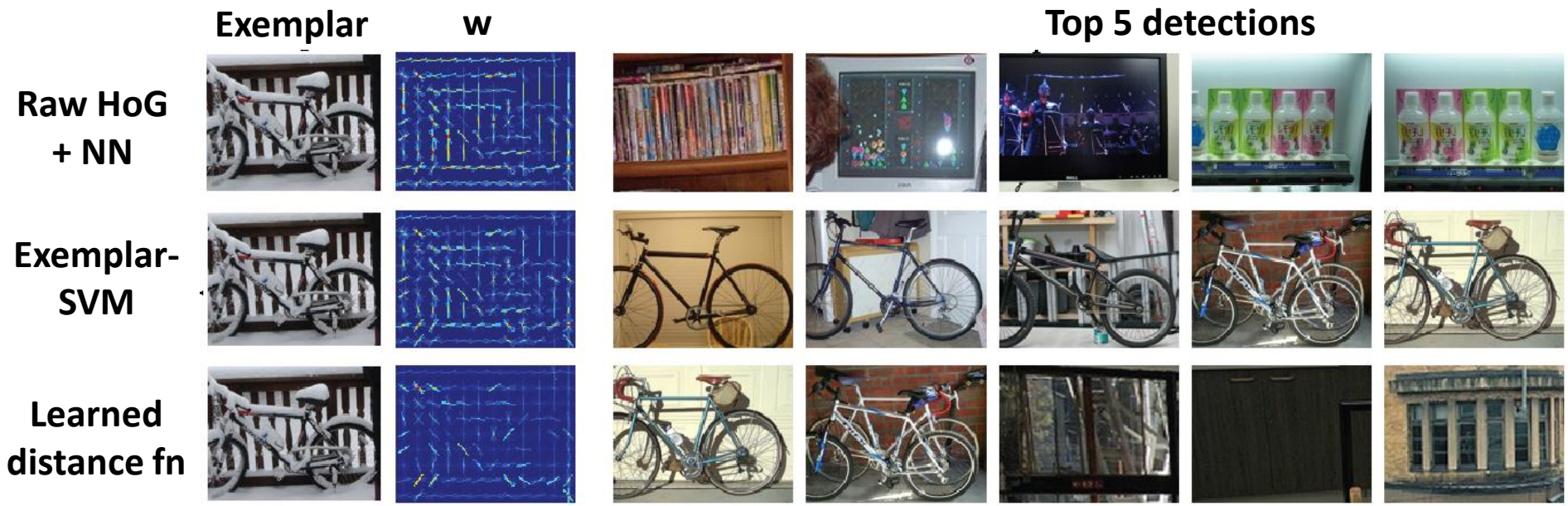
Exemplar-SVM 2

...

Exemplar-SVM N



Example

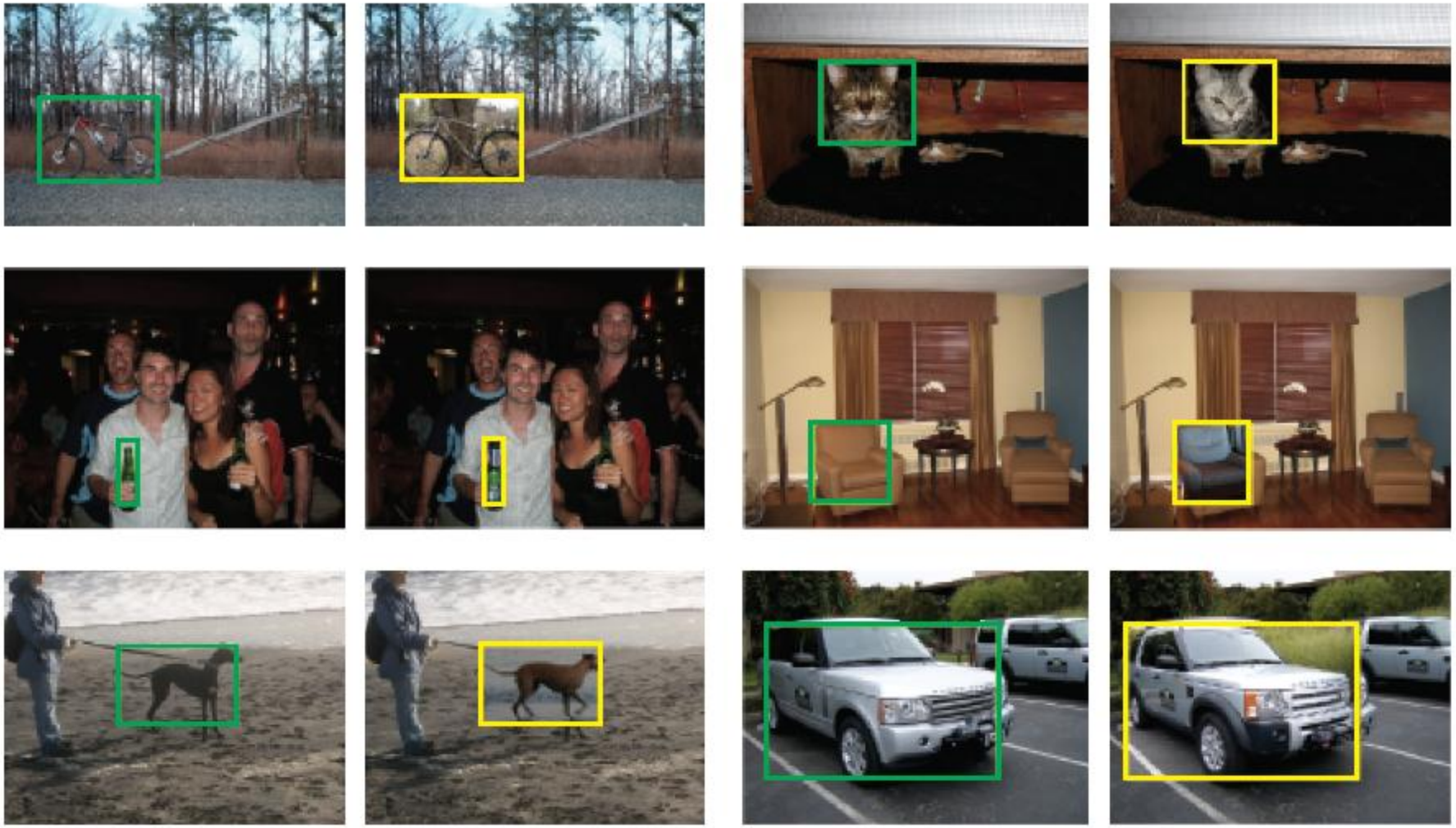


"learns what the exemplar is *not*"

Multiple instances of a category

- Each classifier fires on *similar* trains

Successful classifications



Failed classifications

Does it really work?

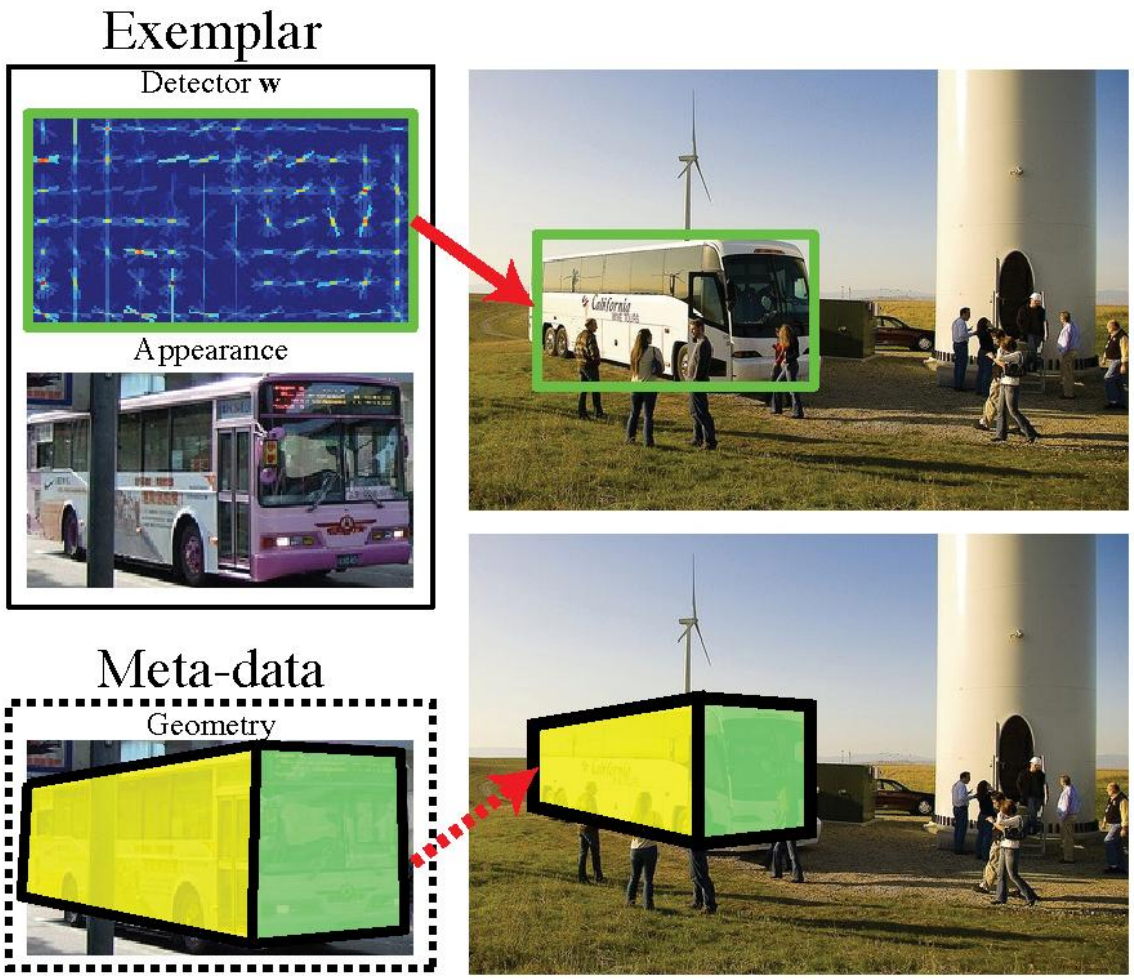
Approach	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog
NN	.006	.094	.000	.005	.000	.006	.010	.092	.001	.092	.001	.004
NN+Cal	.056	.293	.012	.034	.009	.207	.261	.017	.094	.111	.004	.033
DFUN+Cal	.162	.364	.008	.096	.097	.316	.366	.092	.098	.107	.002	.093
E-SVM+Cal	.204	.407	.093	.100	.103	.310	.401	.096	.104	.147	.023	.097
E-SVM+Co-occ	.208	.480	.077	.143	.131	.397	.411	.052	.116	.186	.111	.031
CZ [6]	.262	.409	–	–	–	.393	.432	–	–	–	–	–
DT [7]	.127	.253	.005	.015	.107	.205	.230	.005	.021	.128	.014	.004
LDPM [9]	.287	.510	.006	.145	.265	.397	.502	.163	.165	.166	.245	.050

...

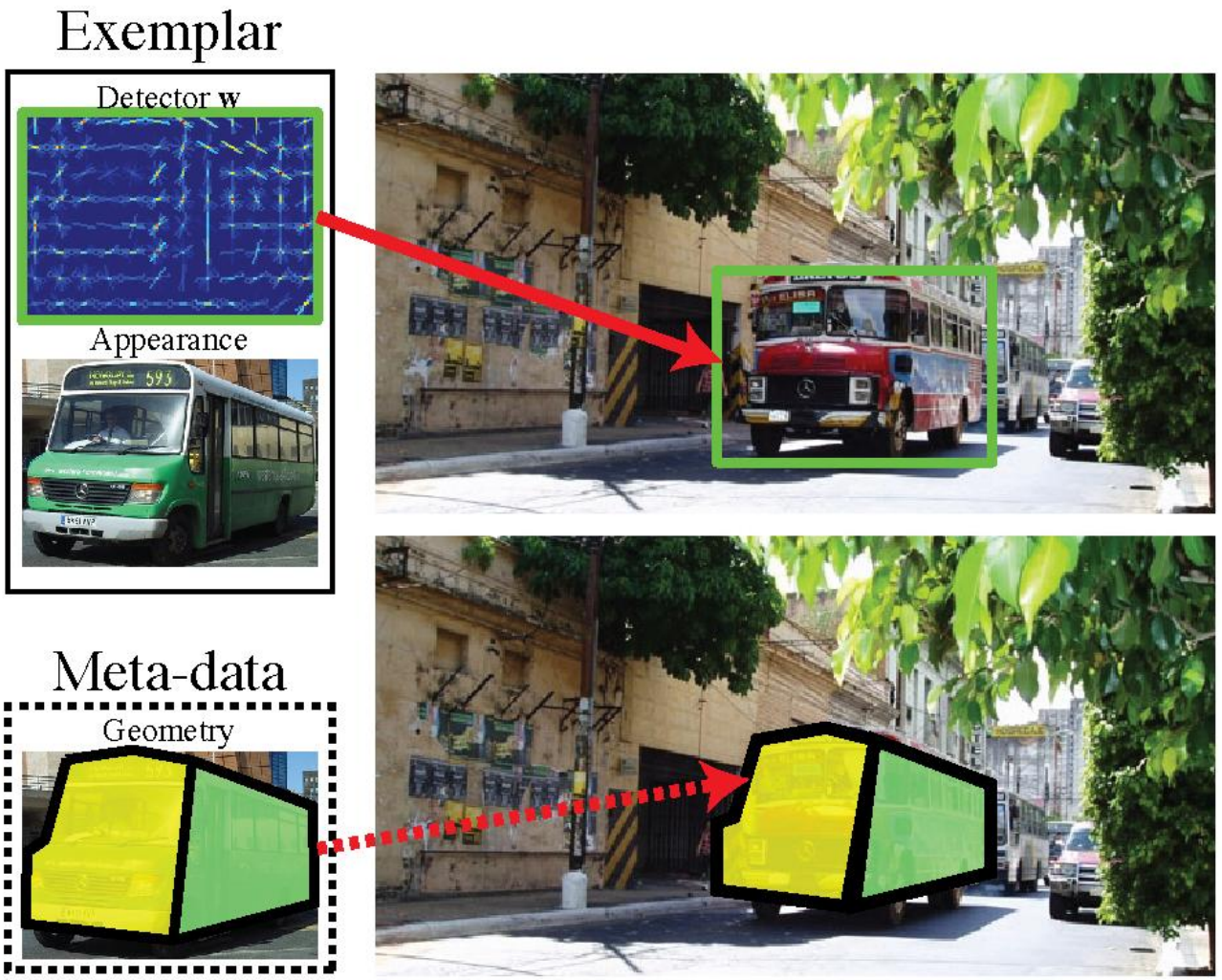
...

horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
.096	.094	.005	.018	.009	.008	.096	.144	.039
.243	.188	.114	.020	.129	.003	.183	.195	.110
.234	.223	.109	.037	.117	.016	.271	.293	.155
.384	.320	.192	.096	.167	.110	.291	.315	.198
.447	.394	.169	.112	.226	.170	.369	.300	.227
–	.375	–	–	–	–	.334	–	–
.122	.103	.101	.022	.056	.050	.120	.248	.097
.452	.383	.362	.090	.174	.228	.341	.384	.266

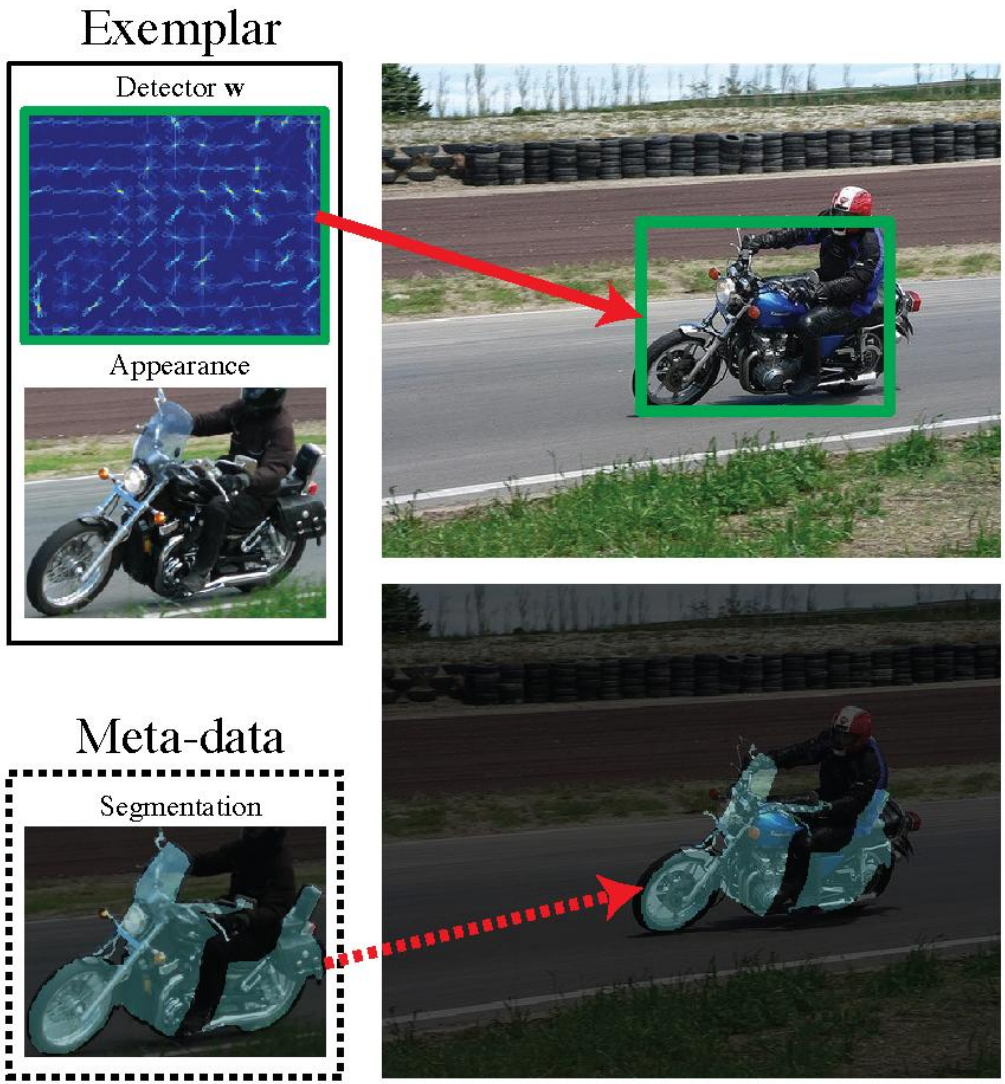
Geometry transfer



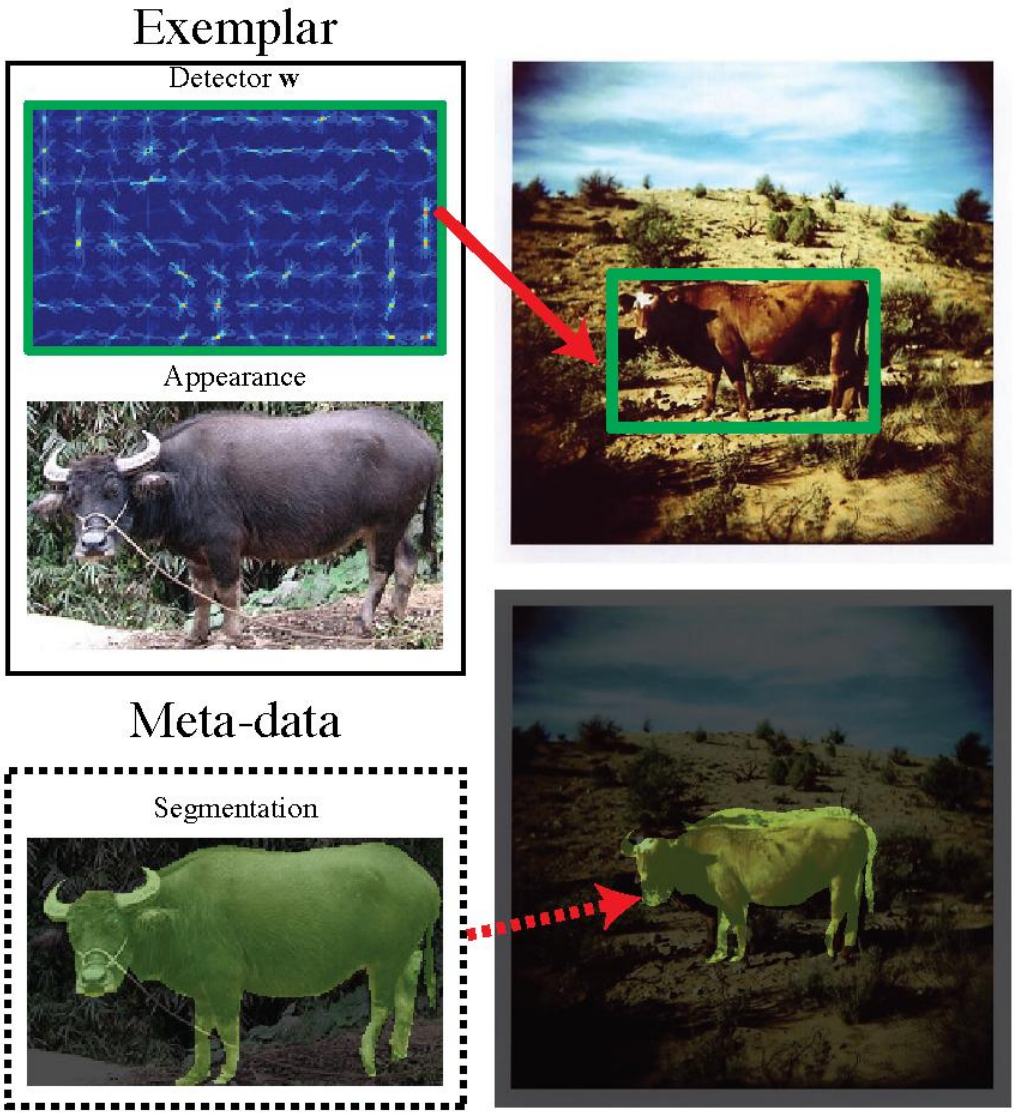
Geometry transfer



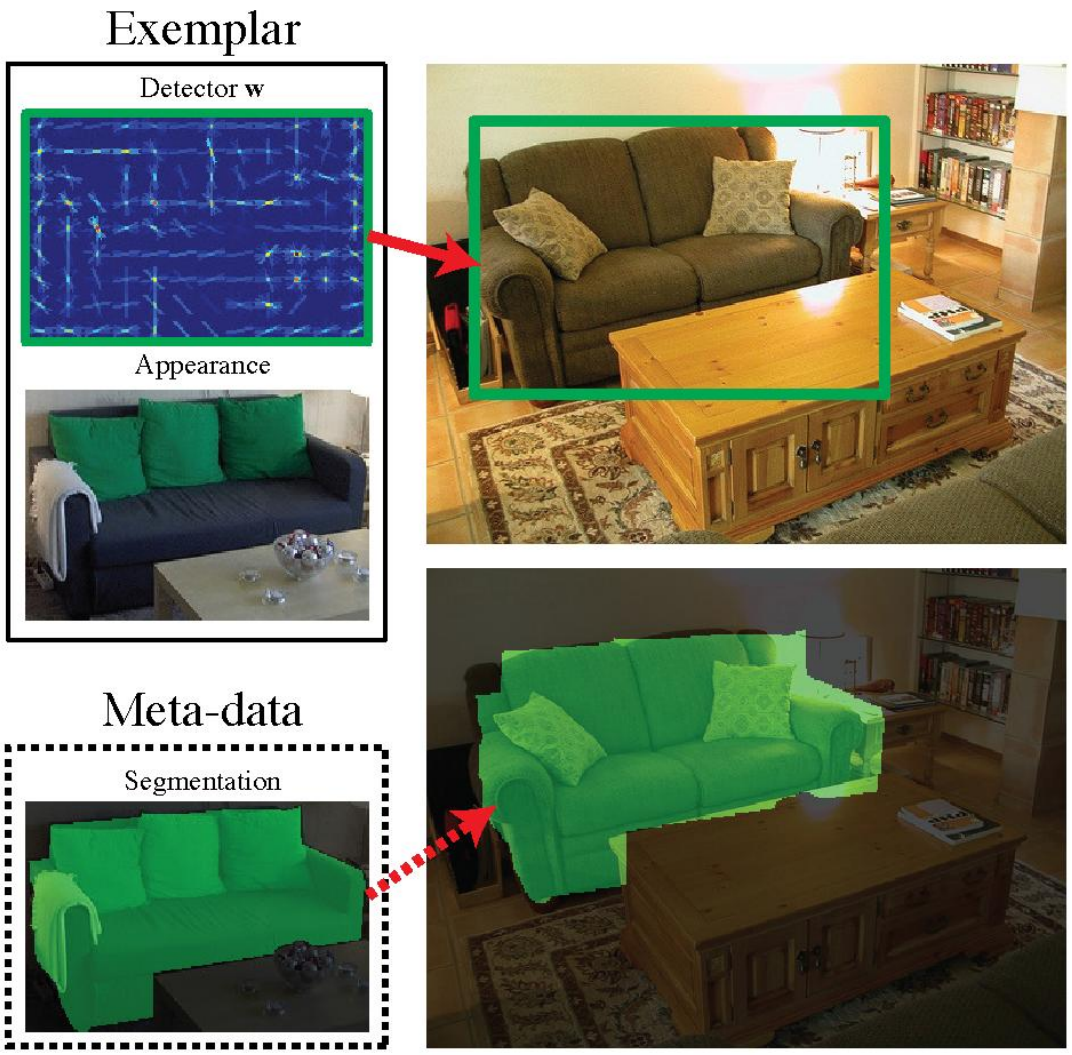
Segmentation transfer



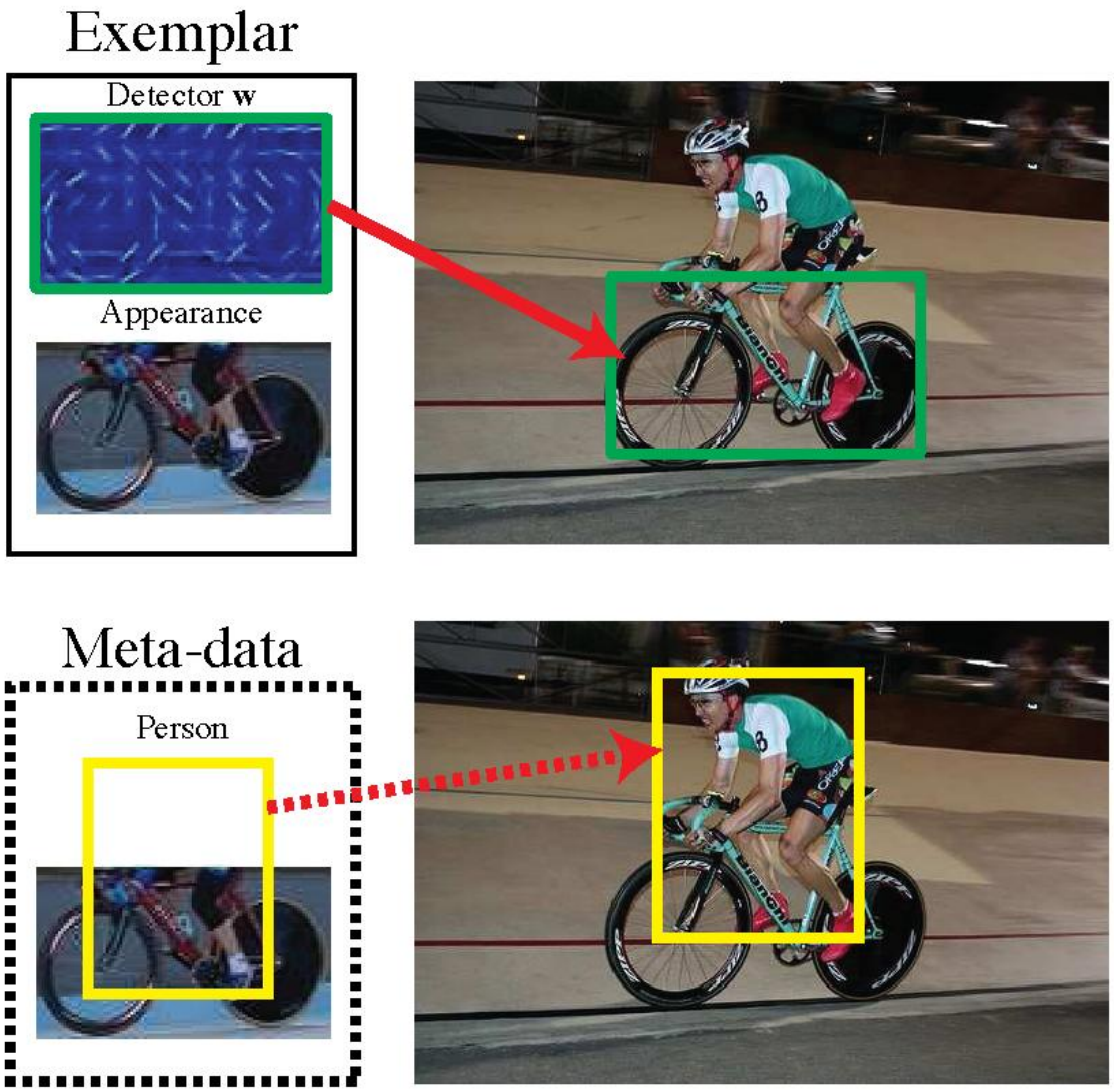
Segmentation transfer



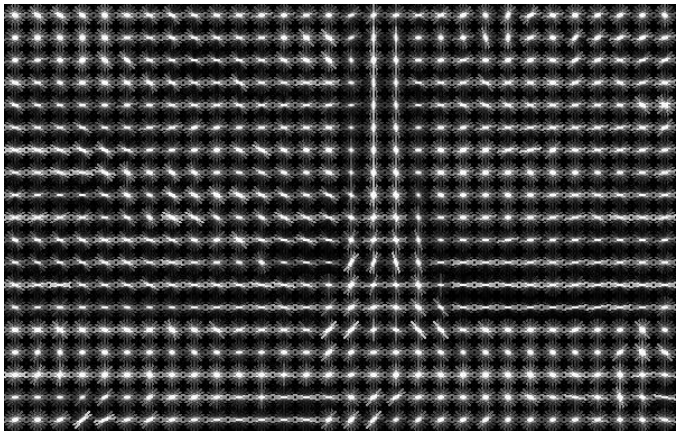
Segmentation transfer

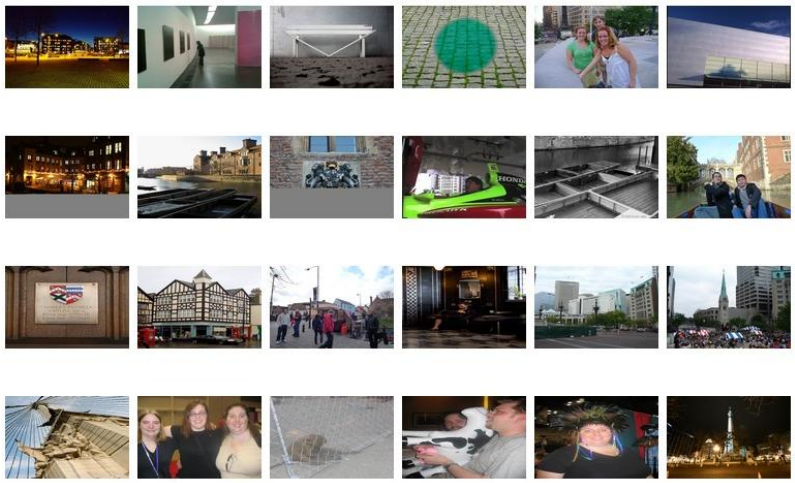


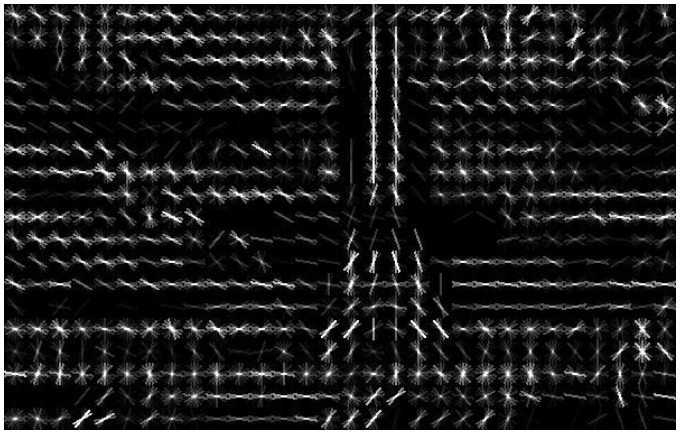
“Object priming” transfer



Histogram of Oriented Gradients (HoG)







Conclusions and open issues

- Interesting new idea for object detection
- … but does it really work? Seems to perform well on some categories, but not others
- Maybe this is too extreme -- some grouping of positives seems like a good idea
- How to come up with better ways of clustering?