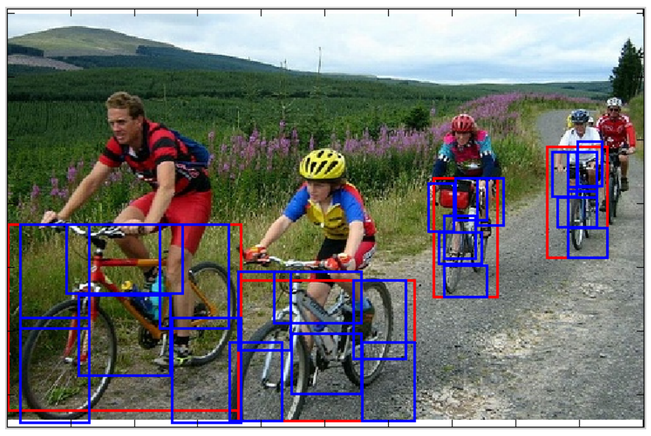
Object Detection with Heuristic Coarse-to-Fine Search

Ross B. Girshick and Pedro F. Felzenszwalb
Department of Computer Science, University of Chicago

What's the problem?

Object detection: localize all instances of a generic object class in a real-world image.

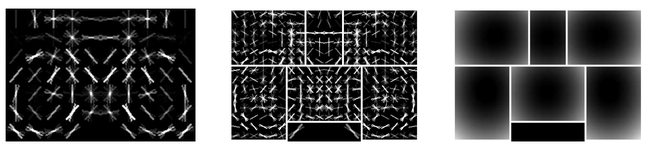


Example: Find all bicycles in this image.

Our research focuses on a general class of deformable part models that we call *visual grammars*. We introduce this class with two examples.

Example 1. Multiscale star models

This simple model is a collection of templates called *filters*.



(A) root filter (B) part filters (C) deformation model

The root filter captures the whole object at a coarse resolution. The part filters capture local regions, such as wheels, at a higher resolution. Each part filter is anchored at a position relative to the root, but may move according to the deformation model.

Detection algorithm for star models

Optimization problem:
Find all local maxima of the model's score function $S_M(L)$, above a threshold T, where $L = (l_1, \ldots, l_n)$ specifies an object hypothesis in a feature pyramid.

Can be solved efficiently with dynamic programming and generalized distance transforms.

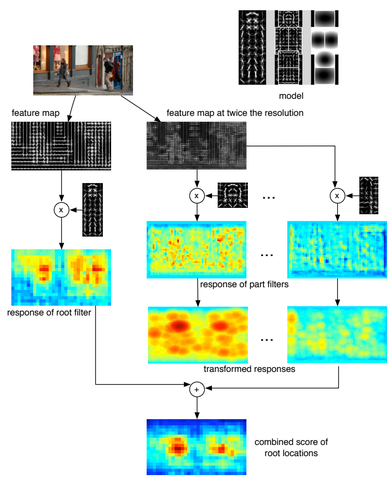


Figure: Computing $S_M(L)$ at one scale.

The algorithm is linear in the number of filters, but the constant factor is large, e.g., $\sim 250M$ floating point multiplications for a 640×480 image.

Example 2. Mixtures of multiscale star models

Mixture models can capture extreme intra-class variation.

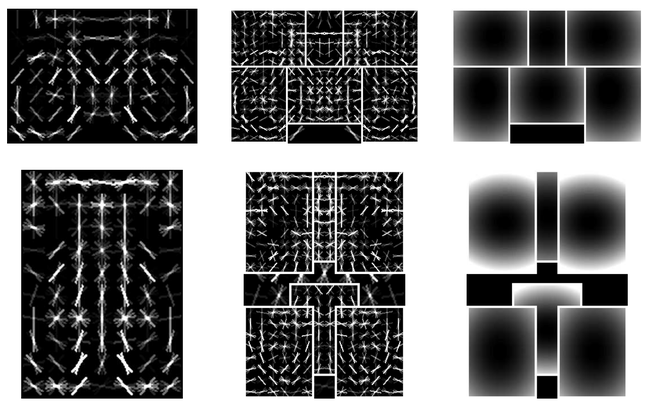


Figure: A mixture model for two bicycle poses.

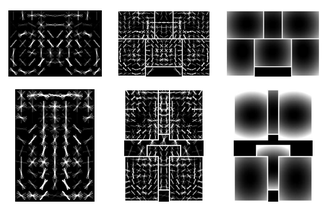
We can apply the detection algorithm for each component *independently*, and then take a max over components.

A mixture of multiscale star models is a special case of a much more general class of deformable object models: *visual grammar models*.

Visual grammar models

Visual grammars define rich deformable part models in terms of acyclic context-free grammars. For example:

Mixture model $\equiv$ Visual grammar



$$B \rightarrow R_1 \mid R_2$$
$$R_1 \rightarrow P_1 P_2 P_3 P_4 P_5 P_6$$
$$R_2 \rightarrow P_7 P_8 P_9 P_{10} P_{11} P_{12}$$

Grammar models generalize deformable object models in a number of important ways:

Example grammar

$$C \rightarrow SC_1 | \ldots | SC_k$$
$$SC_1 \rightarrow P_1 P_2 \ldots P_{n_1}$$
$$\vdots$$
$$SC_k \rightarrow P_1 P_5 P_7$$
$$P_1 \rightarrow MP_1 | MP_2$$
$$MP_1 \rightarrow SP_1 SP_2$$
$$MP_2 \rightarrow SP_1 SP_3$$
$$\vdots$$

- an object class (C) is defined by k subclasses (SC_i),
- shared parts, e.g., a wheel (P_1),
- parts modeled as mixtures (MP_i), *head* $\rightarrow$ *front head*|*side head*,
- parts composed of subparts (SP_i). *front head* $\rightarrow$ *eyes nose mouth*

The path to rich grammar models

It is surprisingly difficult to improve detection results by enriching models.

Three fundamental reasons:

1. Learning often requires judicious use of hidden or latent information (Latent SVM).
2. Good initialization is extremely important, but challenging.
3. The computational efficiency of inference must be maintained as model complexity increases.

Computational efficiency is a prerequisite for finding solutions to issues 1 & 2. **Our approach: coarse-to-fine detection + heuristic best-first search.**

Coarse-to-fine detection

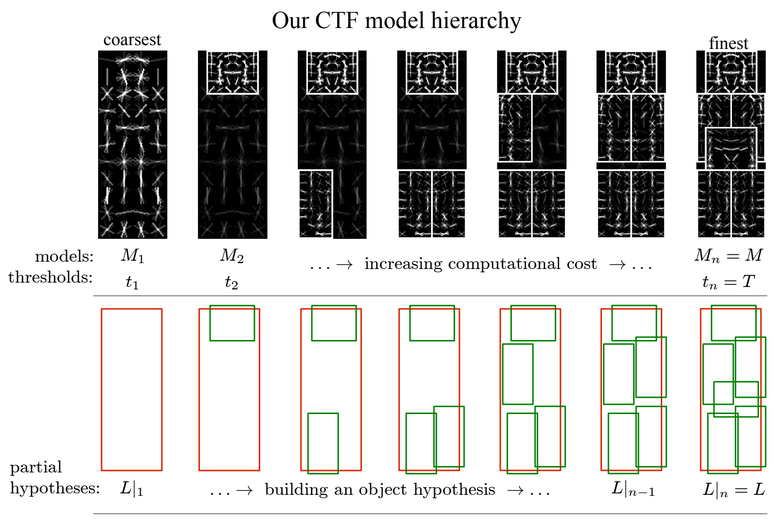
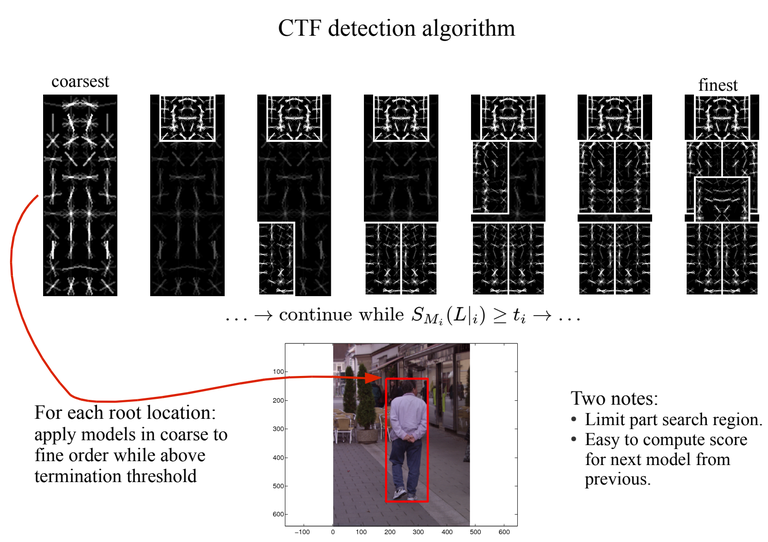


Figure: A fixed part order defines the CTF hierarchy for a star model.



Heuristic coarse-to-fine detection

Observation: when computing the local maxima of a function above a threshold, best-first search allows additional pruning by applying non-maximal suppression on the fly.

Admissible best-first heuristic: upper bound how much a partial object hypothesis can improve.

$$\hat{h}_i \geq S_M(L) - S_{M_i}(L|_i) \text{ for } i = 1, \ldots, n$$
$$\hat{h}_n = 0 \text{ and } S_M(L) \geq T$$

Algorithm:

1. Reorder coarse-to-fine search with a priority queue of partial object hypotheses.
2. Prioritize by partial score + heuristic function.
3. Do non-maximal suppression for each solution.

Probably approximately admissible heuristics

Problem: we cannot efficiently determine a heuristic function that gives good performance and is admissible.

Solution: use a *good* inadmissible heuristic function.

Let $h_i^*(L) = S_M(L) - S_{M_i}(L|_i)$ be a random variable, where L is from an unknown distribution D over $\{L \mid S_M(L) \geq T\}$.

Let $\mathcal{H}_i$ be a sample of h_i^*.

The rule $\hat{h}_i = \max(\mathcal{H}_i)$ is a *good* in the sense that we can provide a PAC-like bound on the error rate.

Let $err(\hat{h}_1, \ldots, \hat{h}_n) = P_{L \sim D}(\hat{h}_1 < h_1^*(L) \vee \ldots \vee \hat{h}_n < h_n^*(L))$.

Theorem: using the rule $\hat{h}_i = \max(\mathcal{H}_i)$, for fixed ϵ and δ, if $|\mathcal{H}_i| > \frac{n}{\epsilon} \ln \frac{n}{\delta}$, then $P(err(\hat{h}_1, \ldots, \hat{h}_n) > \epsilon) < \delta$.

The heuristic is *probably admissible* with *high probability*.

Empirical results

Experimental results for two-component mixture models using heuristic coarse-to-fine detection *with inadmissible heuristics and thresholds* during *testing and training* on 12 of the 20 PASCAL 2007 object classes.

PASCAL 2007 Testing Time

class	DP	HCTF	Speedup
aeroplane	5.70h	3.86h	1.48
bicycle	5.79h	2.37h	2.44
bottle	4.54h	2.28h	1.99
bus	5.75h	2.85h	2.02
car	4.37h	3.82h	1.15
cow	6.09h	3.40h	1.79
horse	6.00h	4.27h	1.41
motorbike	6.01h	2.21h	2.72
person	4.95h	4.45h	1.11
sheep	4.81h	2.85h	1.69
train	6.59h	2.54h	2.59
tvmonitor	9.63h	3.07h	3.13

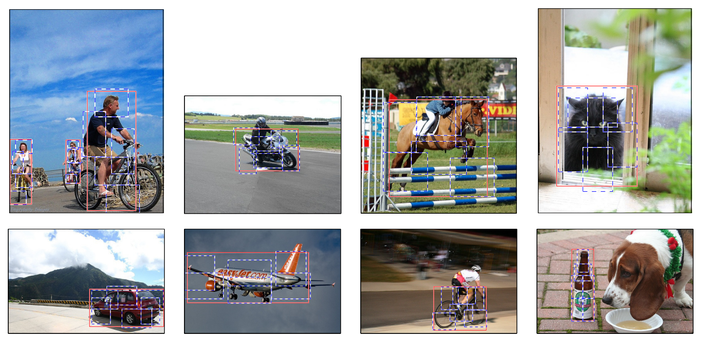
PASCAL 2007 Average Prec.* (comp3)

class	DP	HCTF	% change
aeroplane	0.281	0.285	1.41%
bicycle	0.558	0.548	-1.80%
bottle	0.269	0.261	-3.07%
bus	0.437	0.443	1.42%
car	0.465	0.464	-0.06%
cow	0.207	0.195	-5.93%
horse	0.438	0.432	-1.23%
motorbike	0.384	0.397	3.48%
person	0.332	0.336	1.31%
sheep	0.196	0.200	2.43%
train	0.340	0.371	9.02%
tvmonitor	0.384	0.370	-3.84%

(* Prior to any post-processing steps, not comparable to published results.)

Conclusion: 2-3x speedup for 6 classes while maintaining good AP scores.

Example detections



References

P. Felzenszwalb, D. McAllester, D. Ramanan.
A Discriminatively Trained, Multiscale, Deformable Part Model.
Proceedings of the IEEE CVPR, 2008.

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan.
Object Detection with Discriminatively Trained Part Based Models.
Under preparation, 2009.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman
The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results
http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html

http://people.cs.uchicago.edu/~rbg/ {rbg,pff}@cs.uchicago.edu