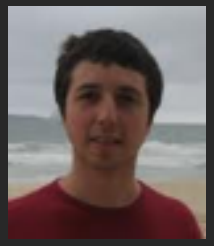
Part models: some thoughts on why they work and what next

Deva Ramanan

Collaborators

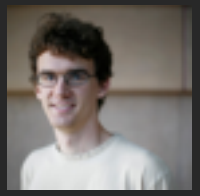
David McAllester



Pedro Felzenswzalb

Ross Girshick

Xiangxin Zhu



Carl Vondrick

Charless Fowlkes

Jitendra Malik

Bharath Hariharan

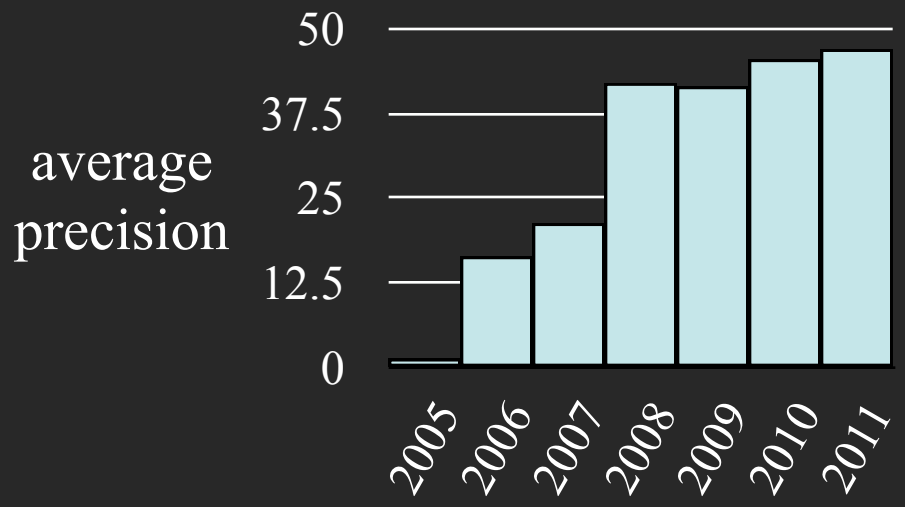
Outline

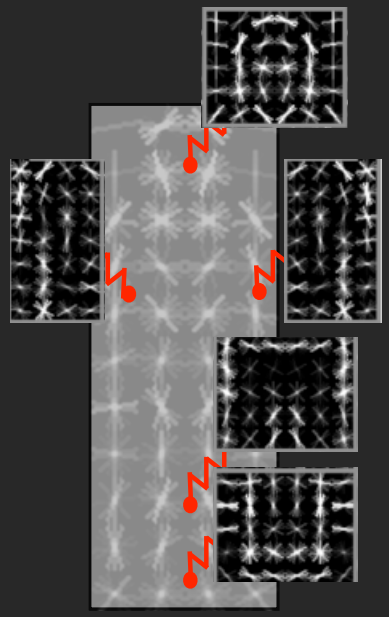
I. Why do part models work?

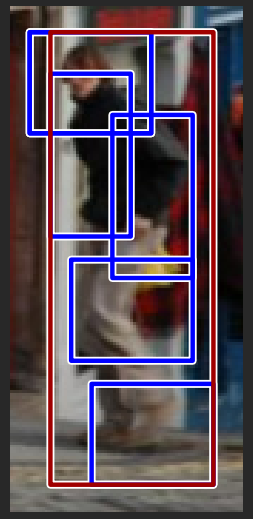
II. A retrospective on PASCAL

III. A wishlist for PASCAL 2.0

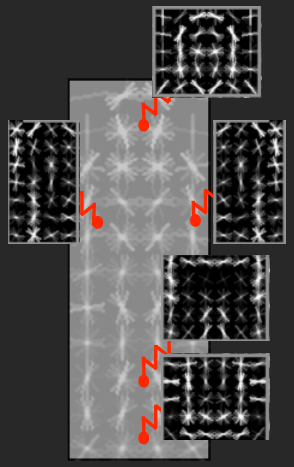
5 years of PASCAL people detection





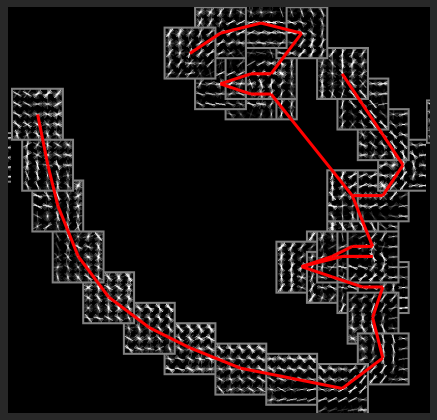


Different flavors of part models



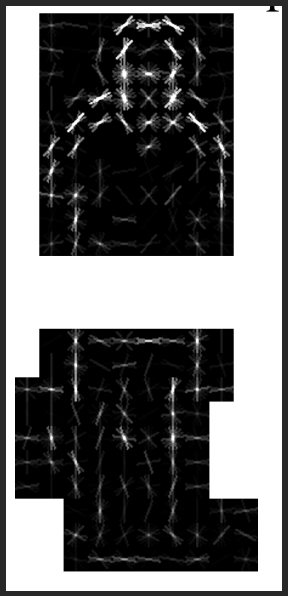
Star models

Girshick, Felzenszwalb, McAllester & Ramanan 10



Tree models

Yang & Ramanan 11
Zhu & Ramanan 12



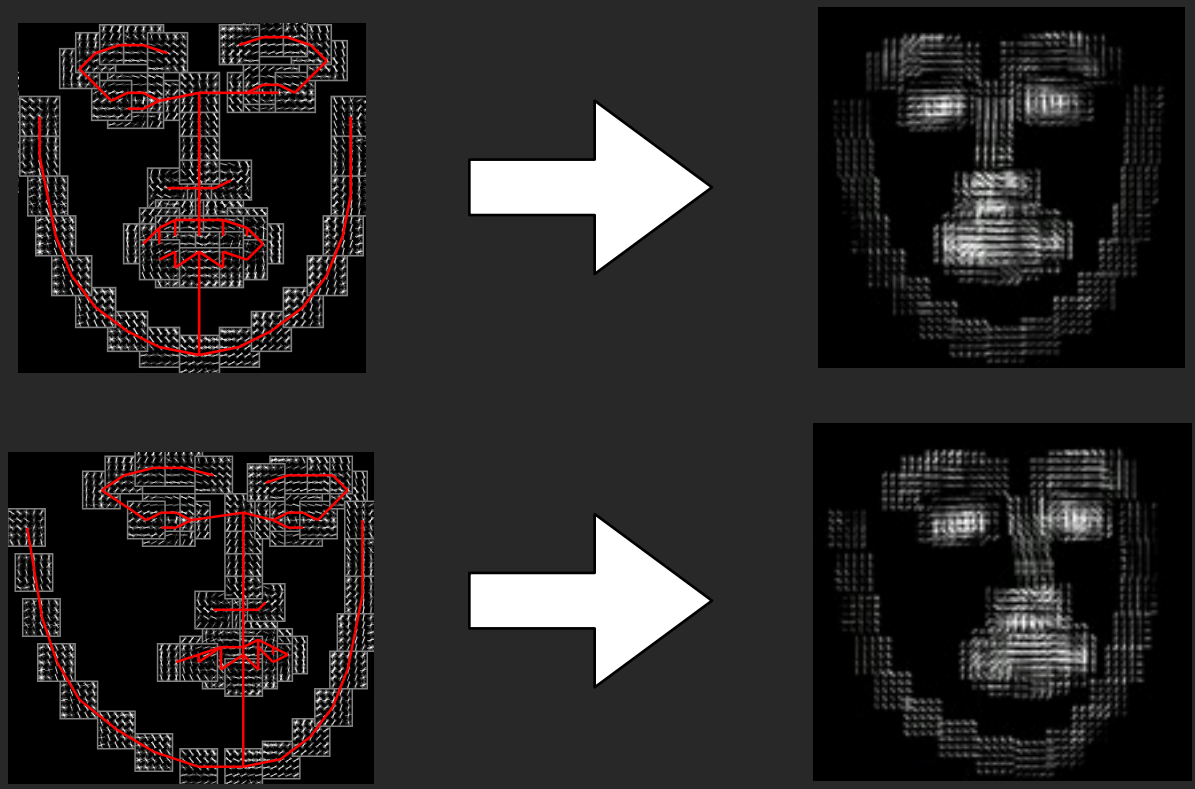
Grammar models

Girshick, Felzenszwalb, & McAllester 11

First, a look back at part models

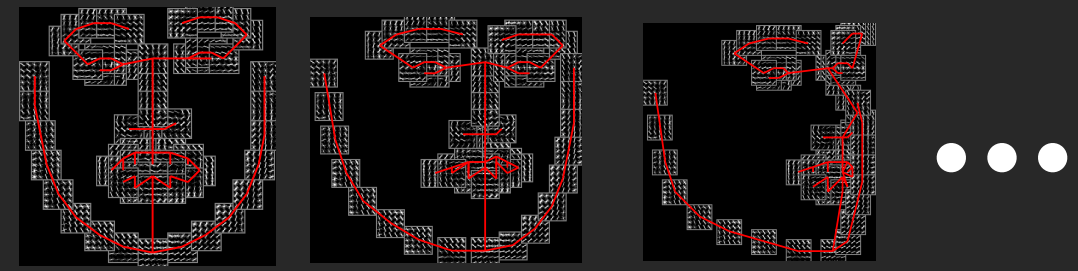
Why do part models “work”?

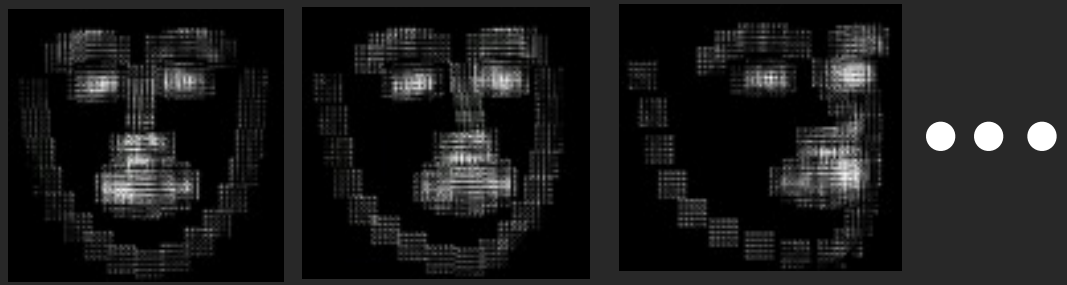
Each distinct placement of parts yields a unique global template



First, a look back at part models

Compositional models allow us to represent an exponentially-large family of global templates

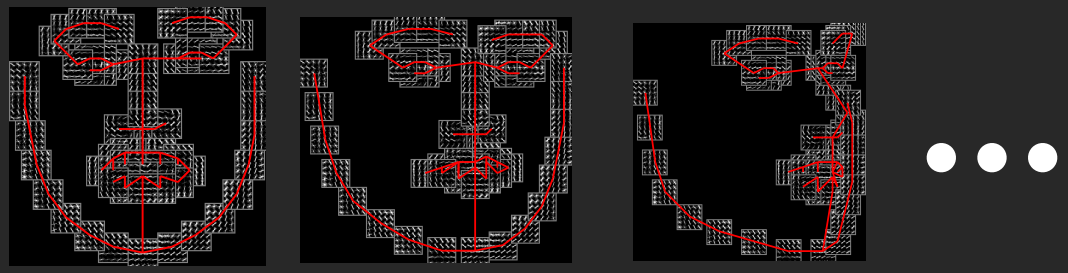




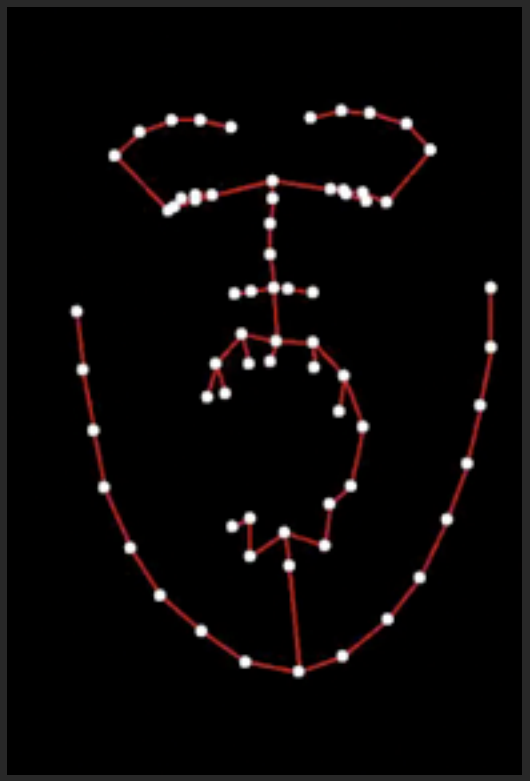
First, a look back at part models

Spatial model defines bias or “prior”

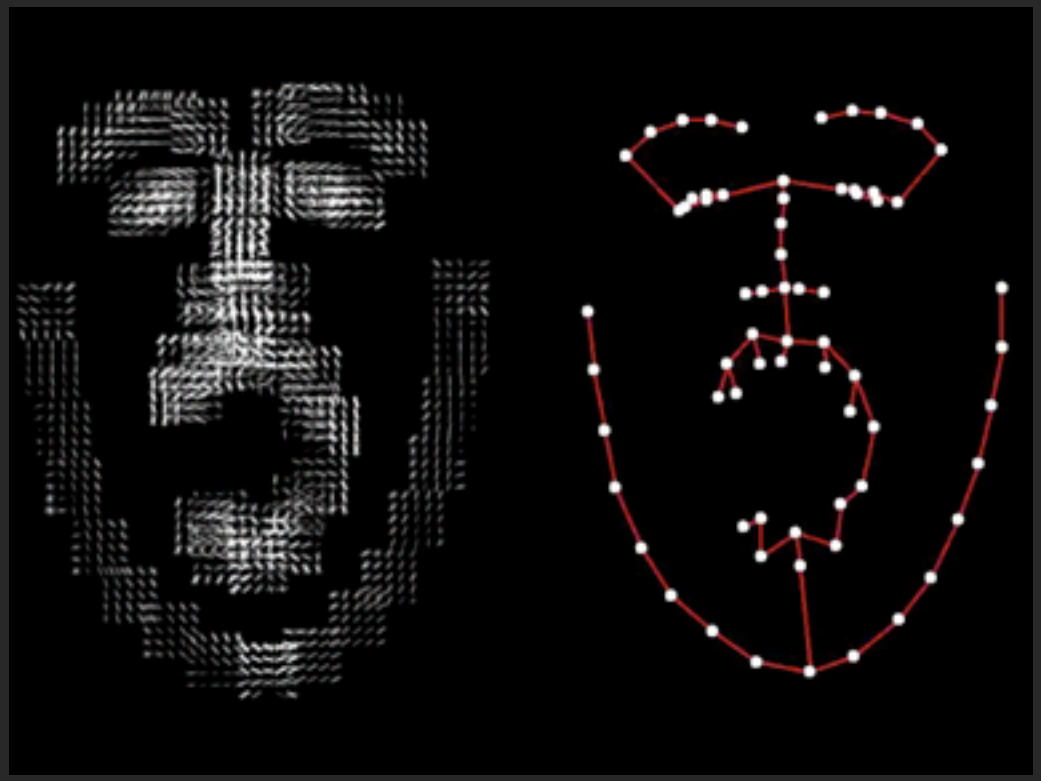
$$f(x) = \max_{z \in Z} w_z \cdot x + b_z$$



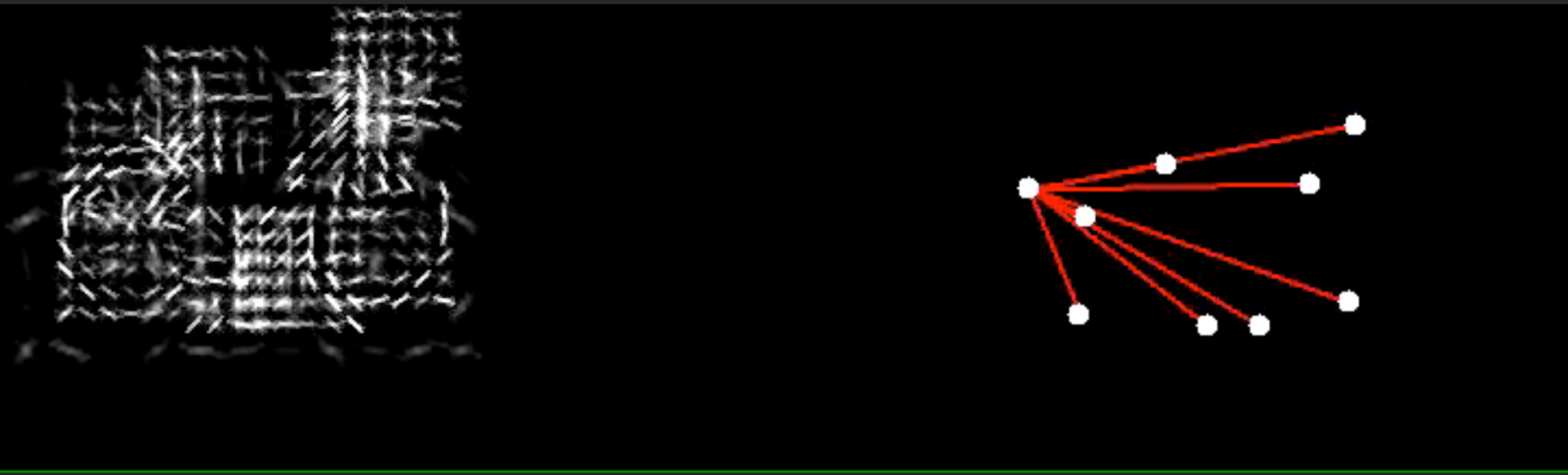
Some fun visualizations



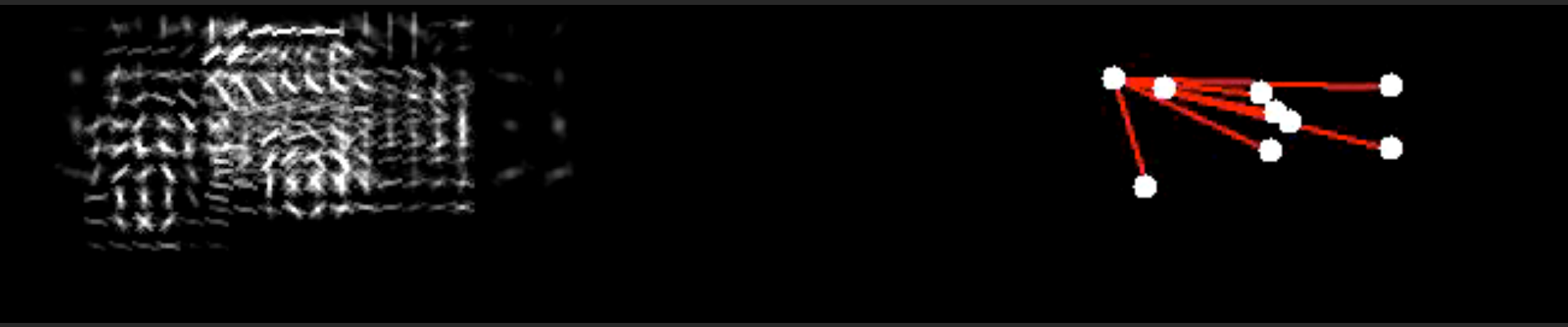
Some fun visualizations



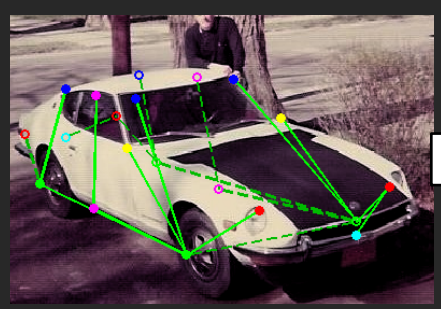
Star bike model (PASCAL 2007)

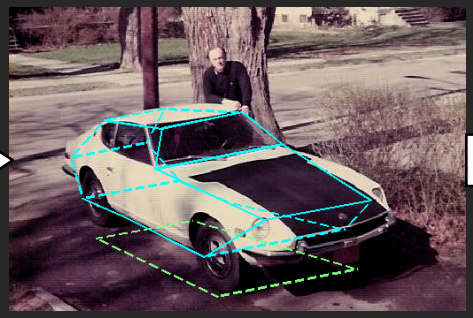


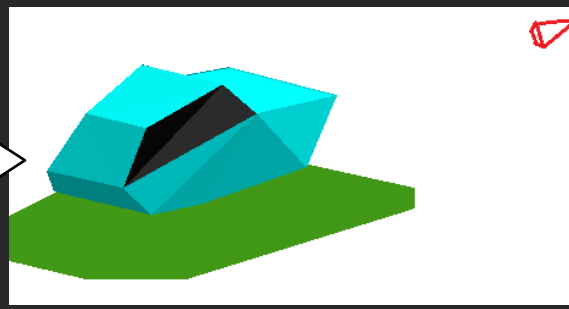
Star car model (PASCAL 2007)



Tree car model
(with local colored mixtures)

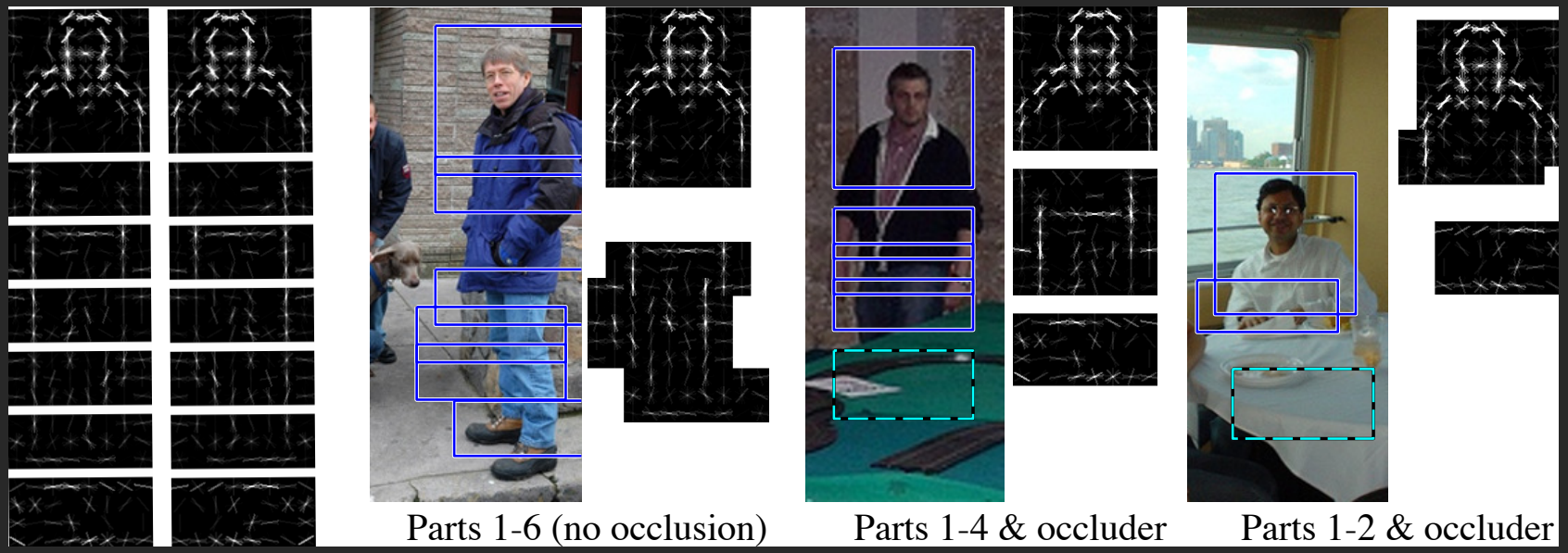






Hejrati & Ramanan
NIPS12

Variable-structure grammar models



Girshick, Felzenszwalb, & McAllester 11

DPMs as large-mixture models

$$f(x) = \max_{z \in Z} w_z \cdot x + b_z$$

- “Double-counting” manifests simply as too strong of a weight

- Suggests jointly learning parts is crucial

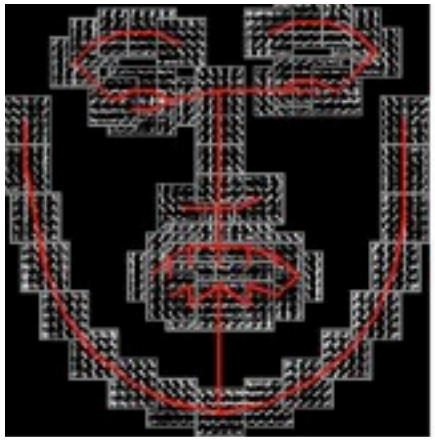
(verified empirically)

Case study

Qualitative results

Zhu & Ramanan CVPR 12

Face detection

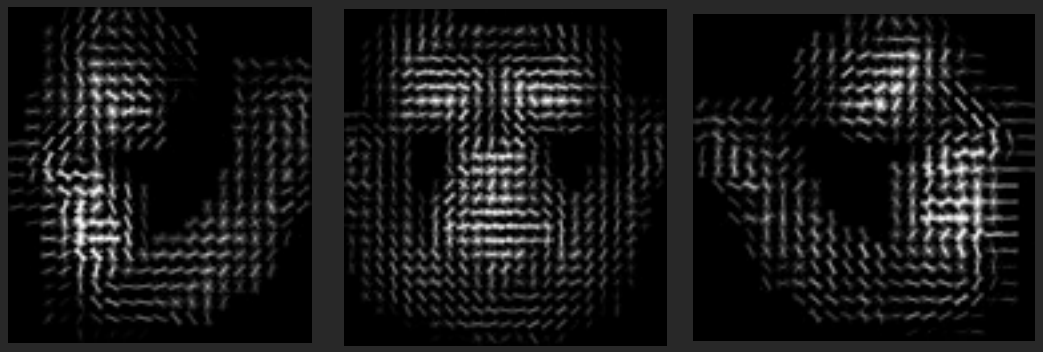


We train on hundreds of faces

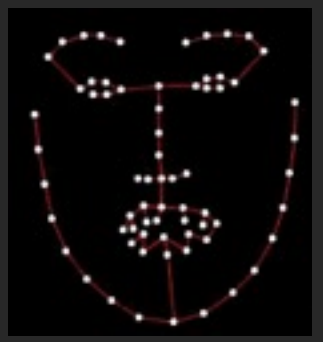
vs

Taigman & Wolf "Leveraging Billions of Faces..." 2011

DPMs vs explicit mixtures



Mixtures of rigid templates



Part model

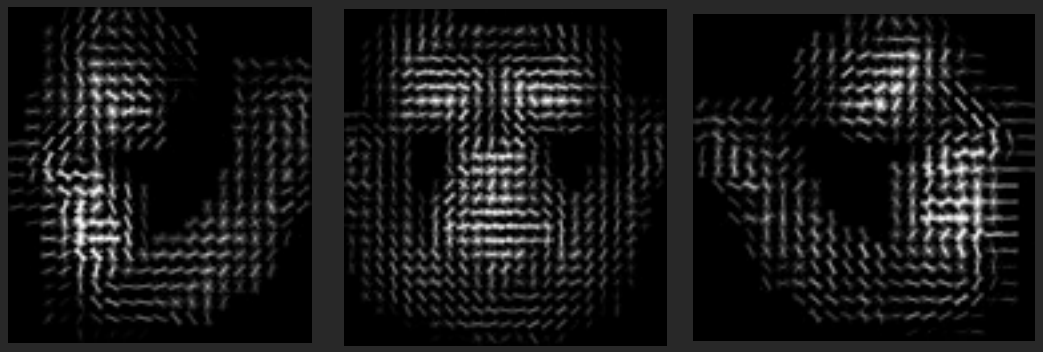
"Exemplar SVMs"

Malisiewicz et al ICCV 11

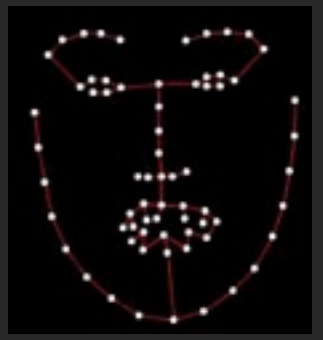
An analysis of part models

Zhu et al, BMVC 2012

Why do explicit mixtures not (appear to) approach DPM performance?



Mixtures of rigid templates

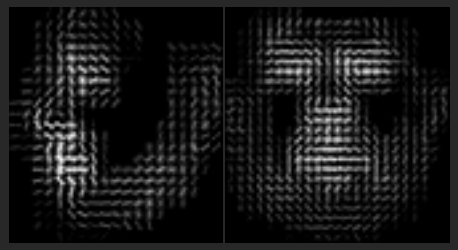


Part model

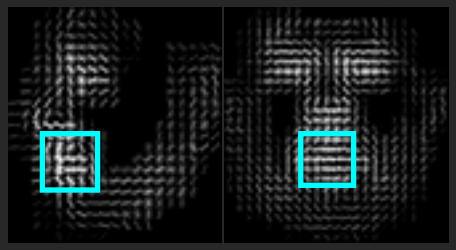
Compared to a mixture of exemplars (Malisiewicz et al), part models...

1) Share parameters across mixtures
2) “Synthesize” new rigid templates not seen during training
3) Efficiently search over mixtures using dynamic programming

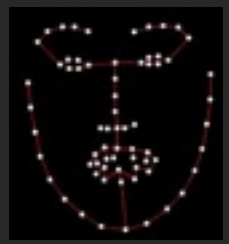
Part model vs. large collections of templates



Mixtures of rigid templates



Mixtures of rigid templates with tied parameters (given by parts)



Part model

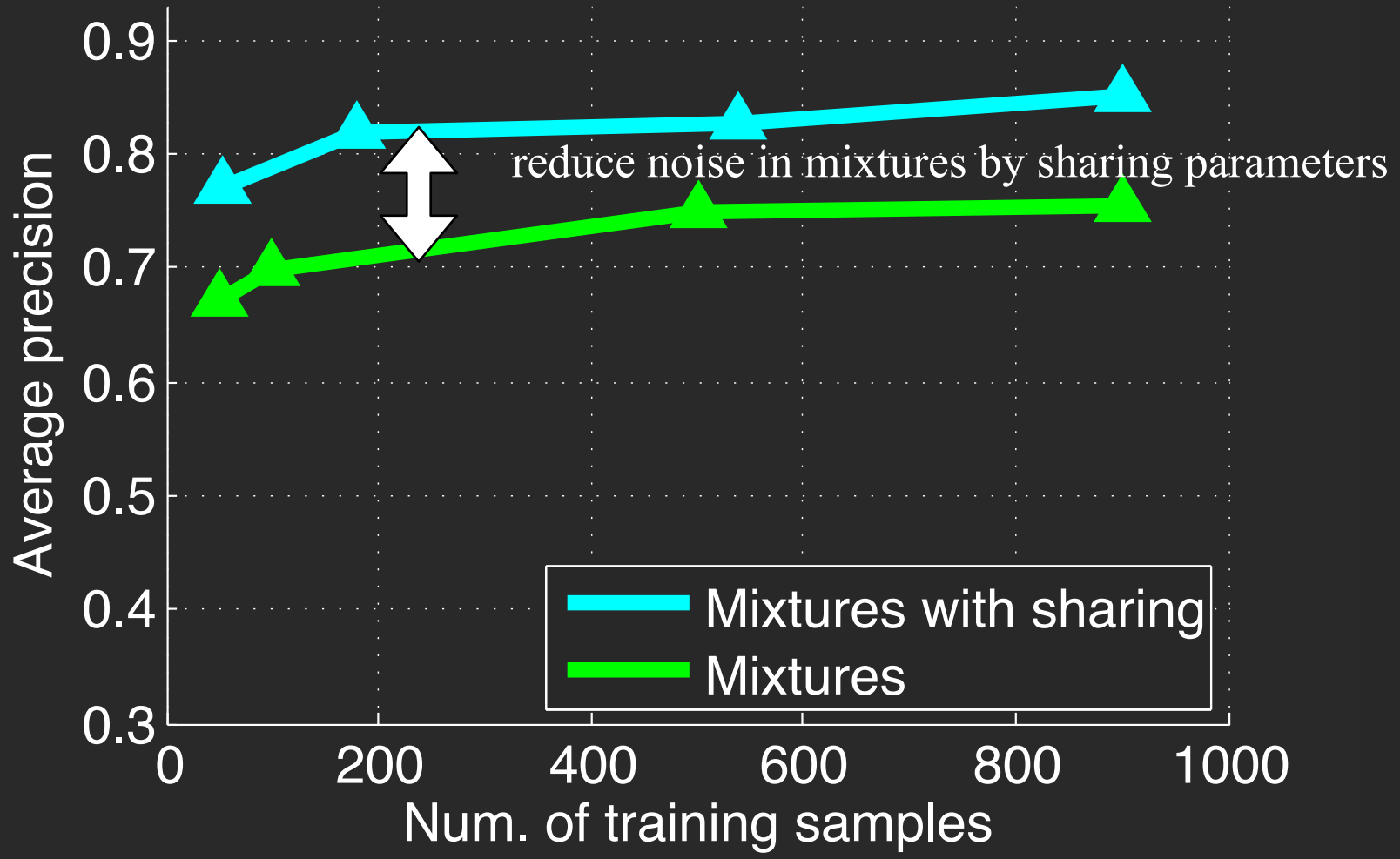
1) Share parameters across mixtures
2) "Synthesize" new rigid templates not seen during training

To examine (1) vs (2), lets define mixture of exemplars with sharing

An analysis of part models

Zhu et al, BMVC 2012

An analysis of part models

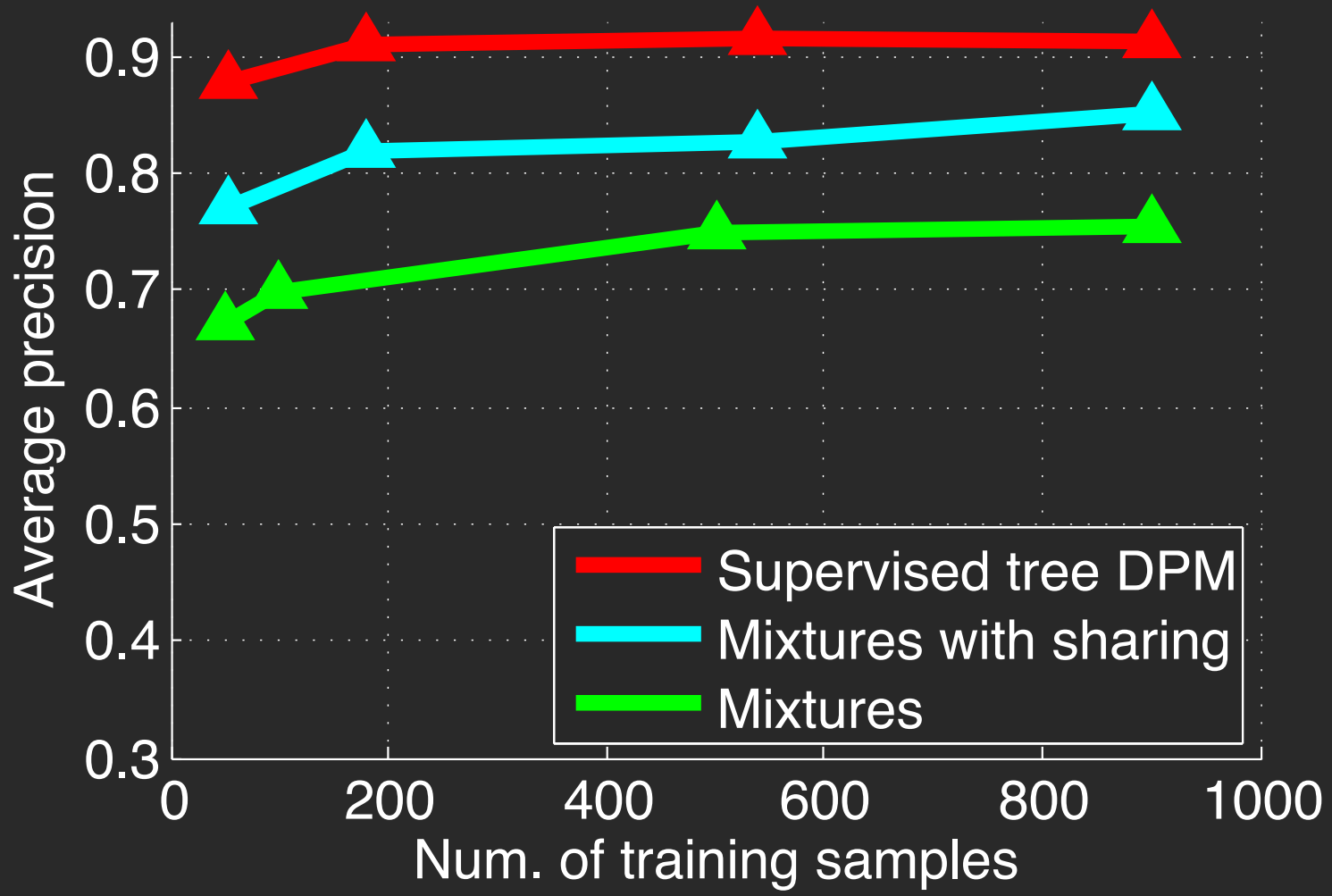


Zhu et al, BMVC 2012

An analysis of part models

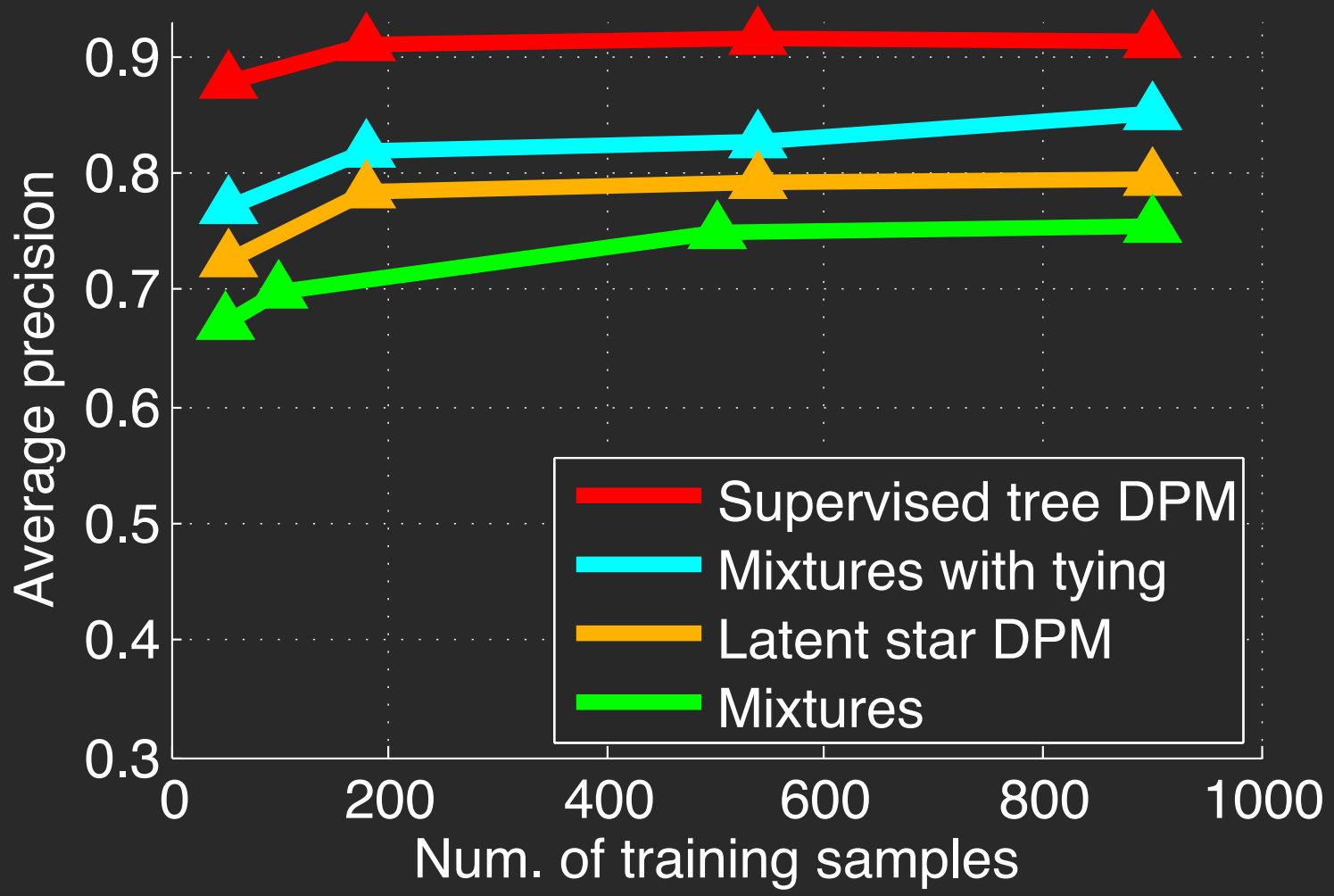
Hallucinating new templates is even more beneficial than sharing

An argument against "big-data"



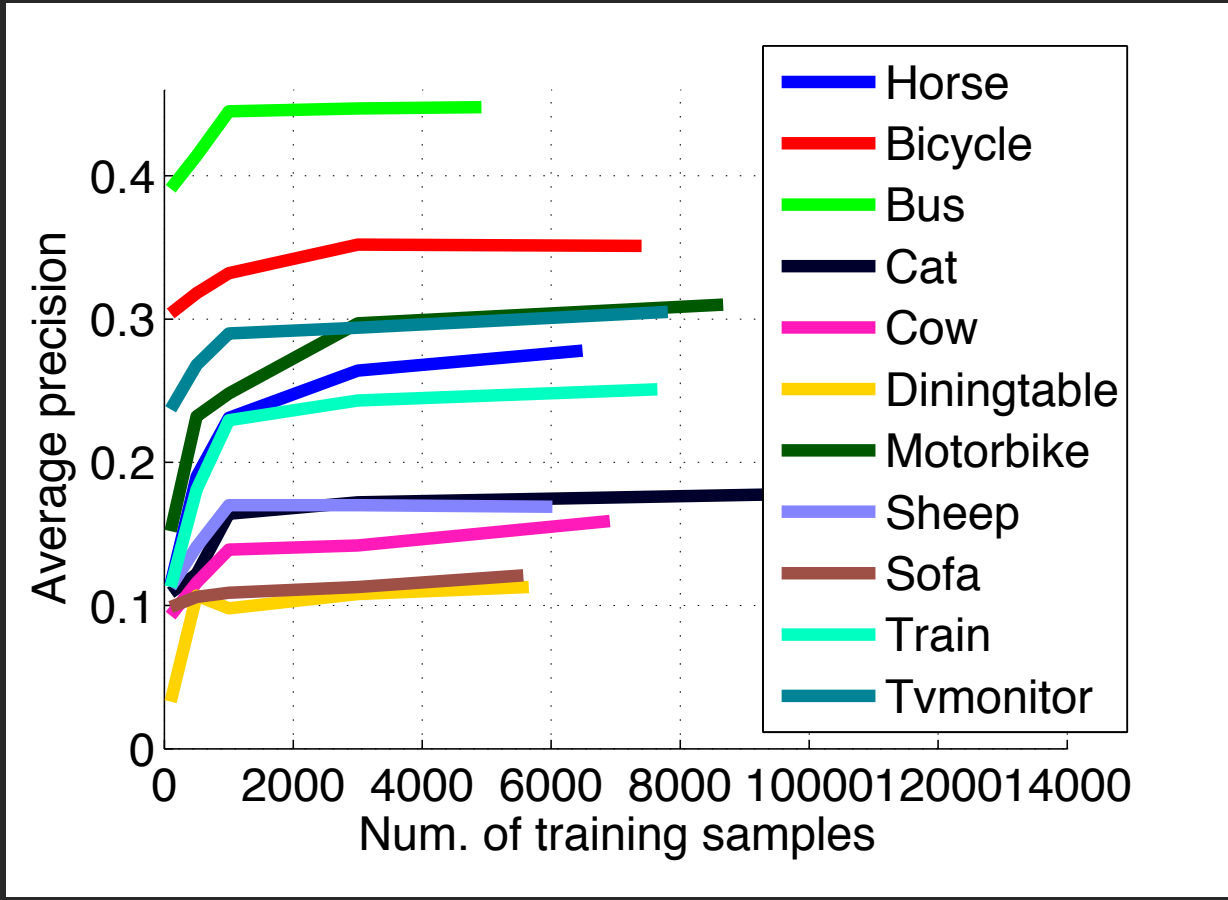
One can train a state-of-art face detector (Google Picassa & Facebook's face.com) with 100 faces!

An argument against "big-data"



Supervised tree structure is important

PASCAL 10X data

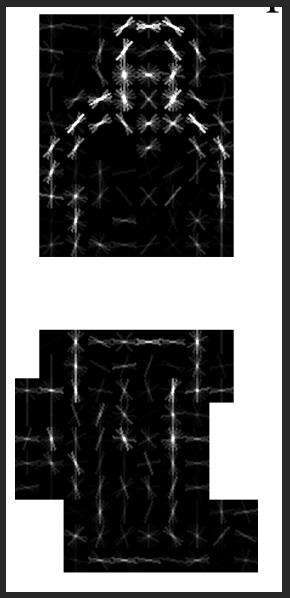


Zhu et al, BMVC 2012

(without parts)

Claim: representation more important than data

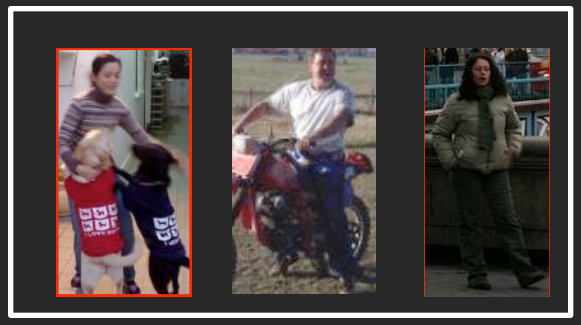
Trees with local mixtures



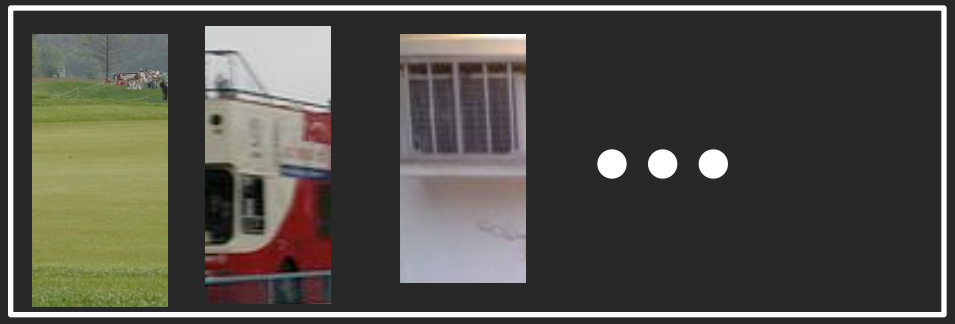
Grammars

But don’t we need to mine through lots of hard “negative” examples?

pos

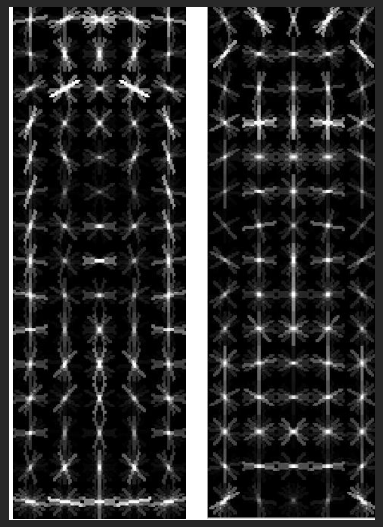


neg

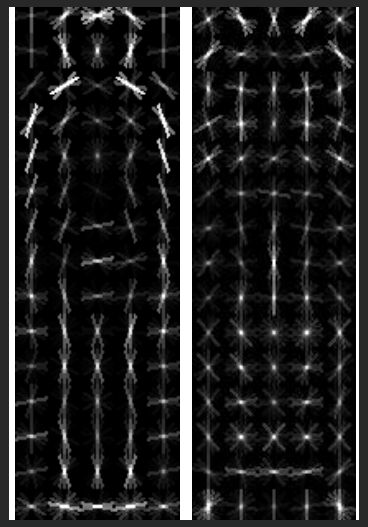


But don’t we need to mine through lots of hard “negative” examples?

Perhaps not...

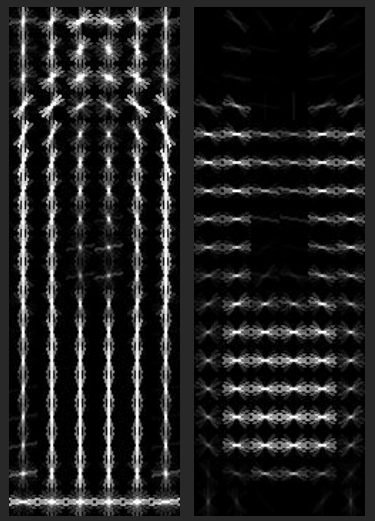


SVM



Gaussian model

$$w = \Sigma^{-1}(\mu_1 - \mu_0)$$



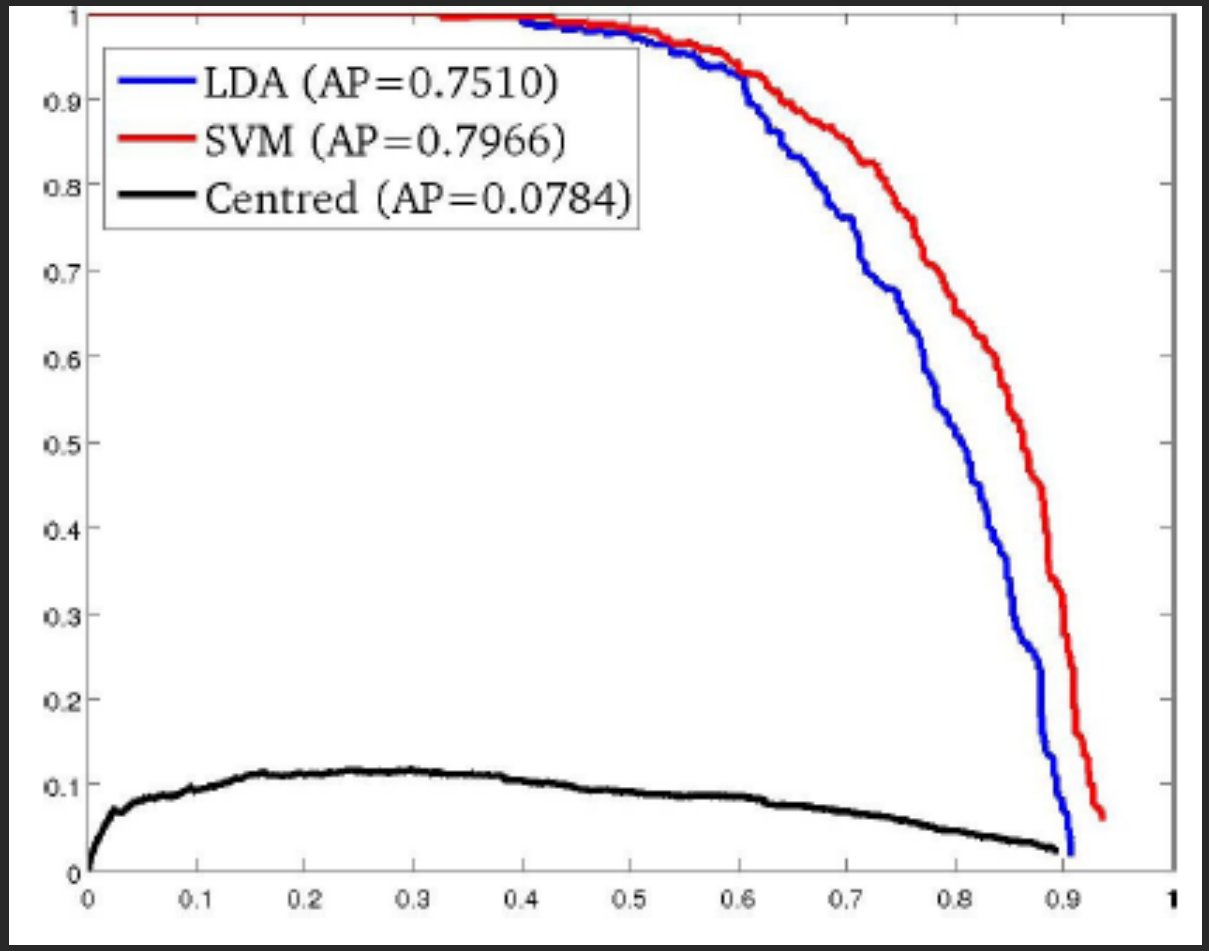
Centered model

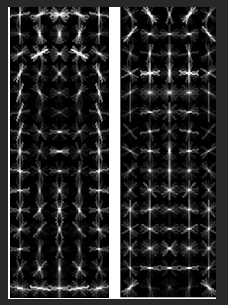
$$w = \mu_1 - \mu_0$$

Learn templates with simple statistical (de)correlation models

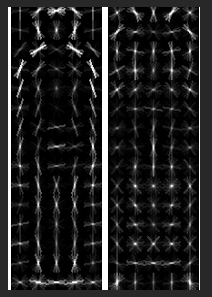
Hariharan, Malik, Ramanan ECCV 12

Linear discriminant (LDA) models

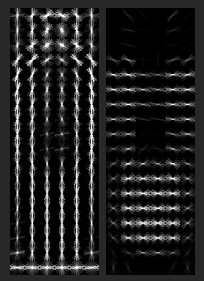




SVM



LDA

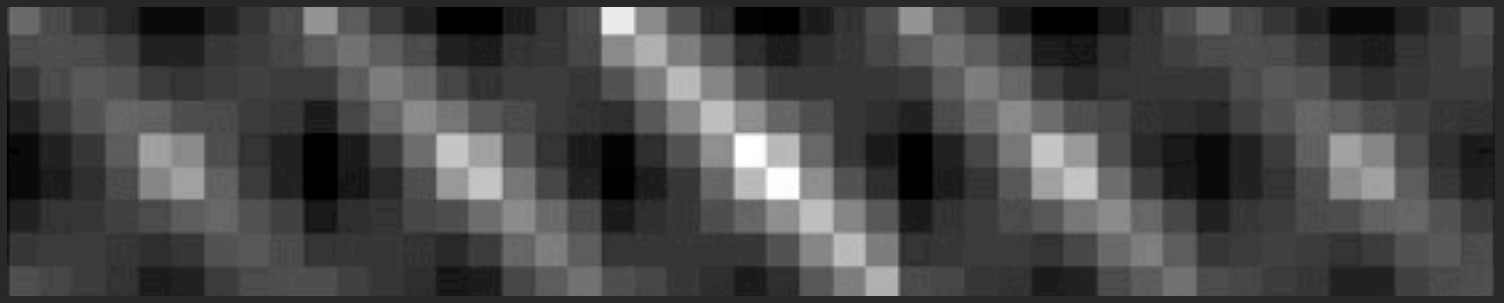


Cen

Properties of spatial covariance matrix

1) Stationairy: $cov(x_i,x_j) = cov(x_i - x_j)$

Can be efficiently encoded with a set of 36x36 matrices Sig_{i-j}

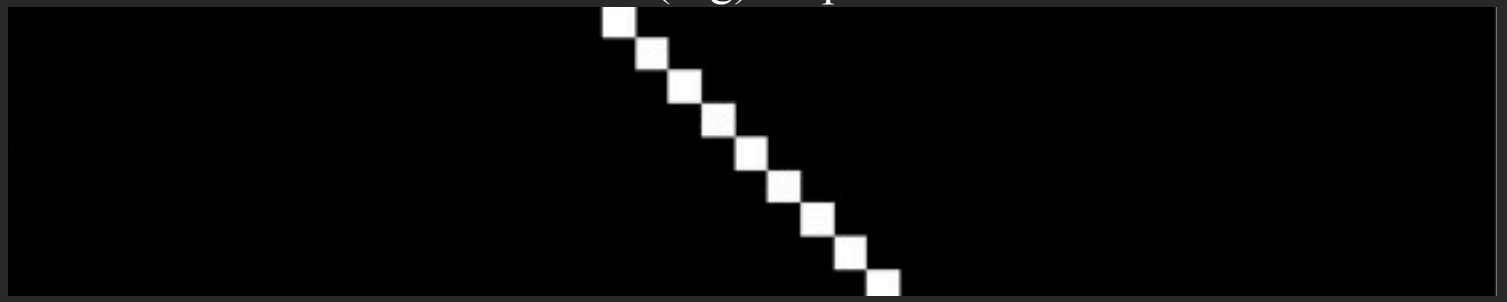


Sig_{-2} Sig_{-1} Sig_o Sig_1 Sig_2

Properties of spatial precision matrix

Inv(Sig) is sparse

Inv(Sig) > eps



Inv(Sig) < -eps

Inv(Sig) subtracts correlated gradients (at neighboring orientations and windows)

Outline

I. Why do part models work?

II. A retrospective on PASCAL

III. A wishlist for PASCAL 2.0

Outline

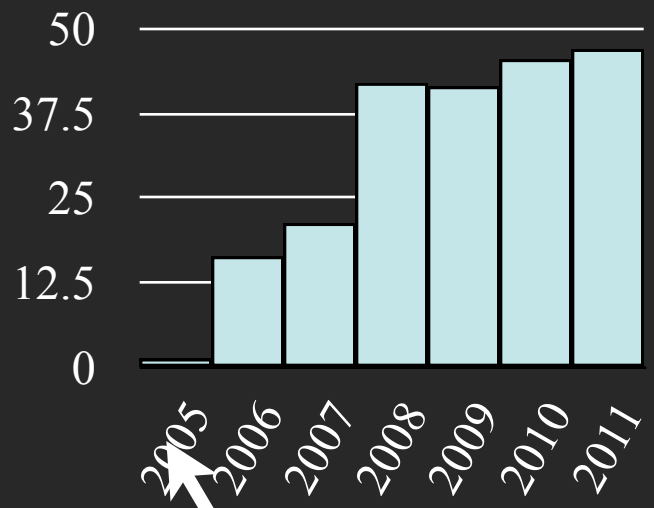
I. Why do part models work?

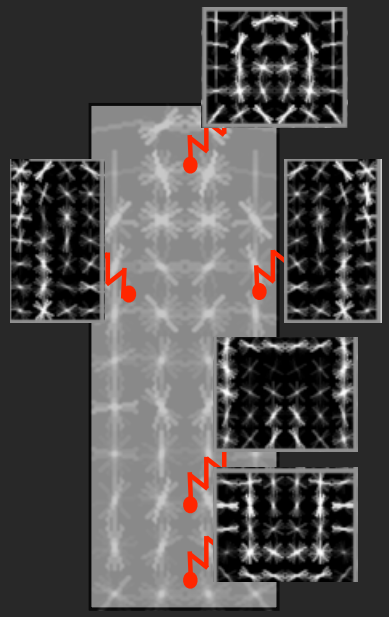
II. A retrospective on PASCAL

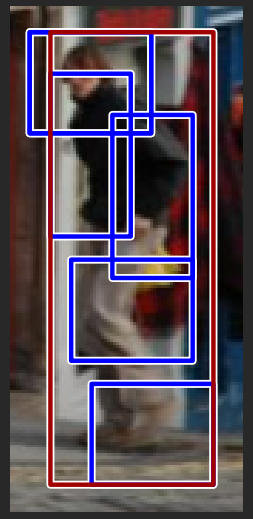
What worked, what didn't?

III. A wishlist for PASCAL 2.0

1. A gut check





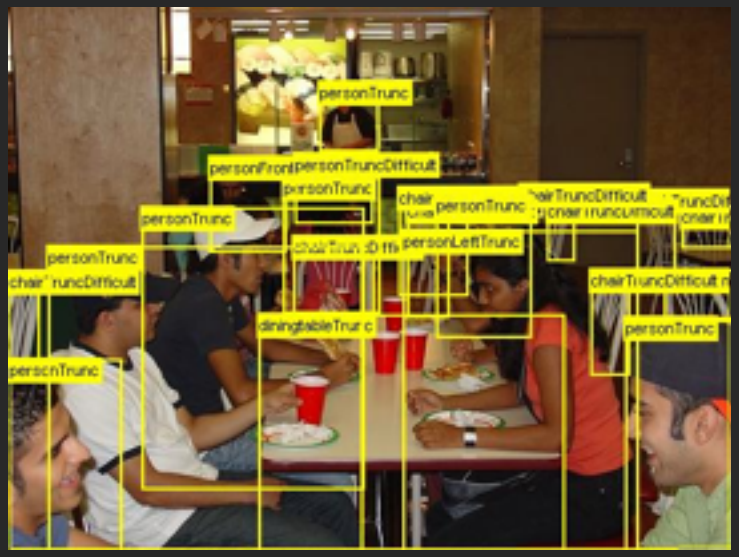


Close to 0 percent performance!

PASCAL made it okay to be “honest” about the state-of-affairs

Address reviewer complaint: “Why doesn’t your approach do better?”

The data is “golden”

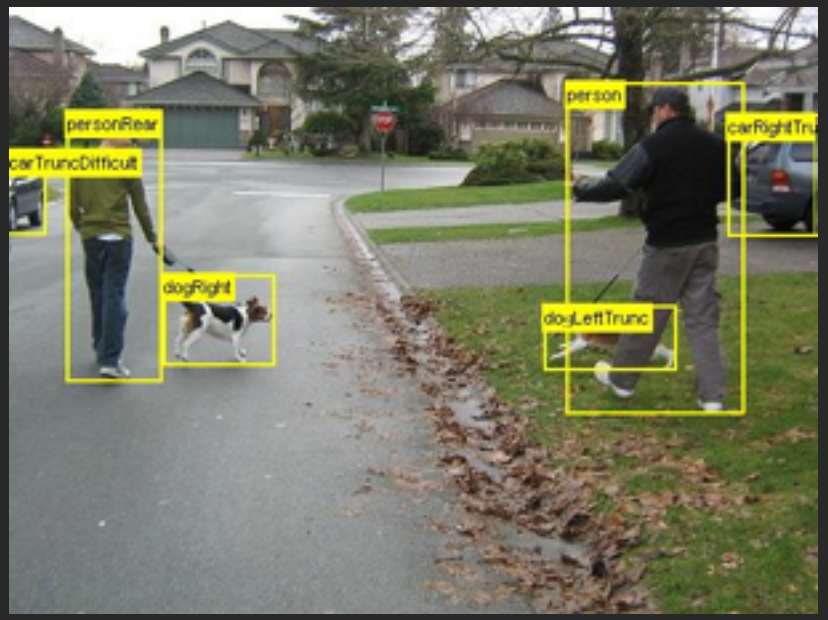


Our first attempt at PASCAL was a curve-based model

Lesson: rather than starting with a model, start with the data

2. Detection does not immediately follow from classification

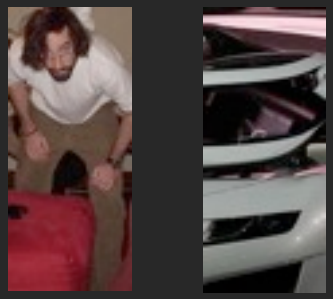
Image classification
Caltech 101/256



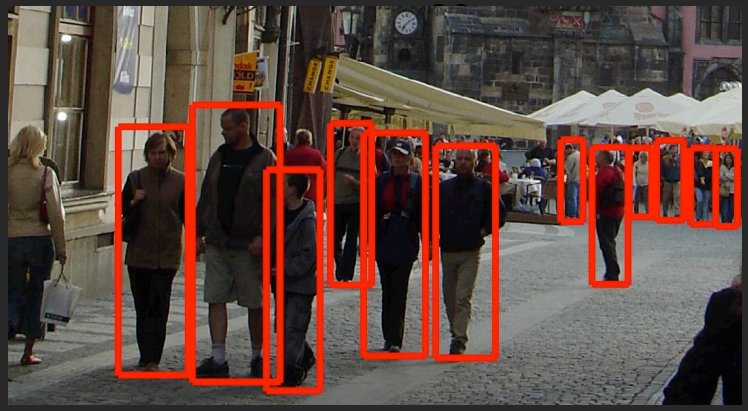
Object detection
Pascal

“One drives a car, not an engine”

Paraphrased from Hao Zhang



Pattern classification

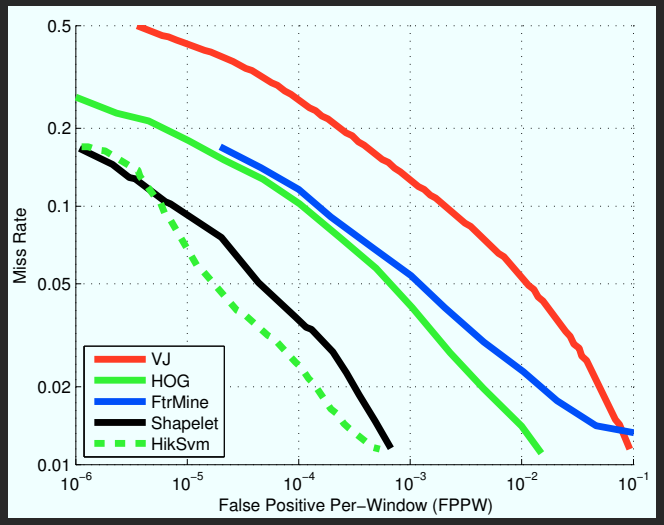


Object recognition in cluttered scenes

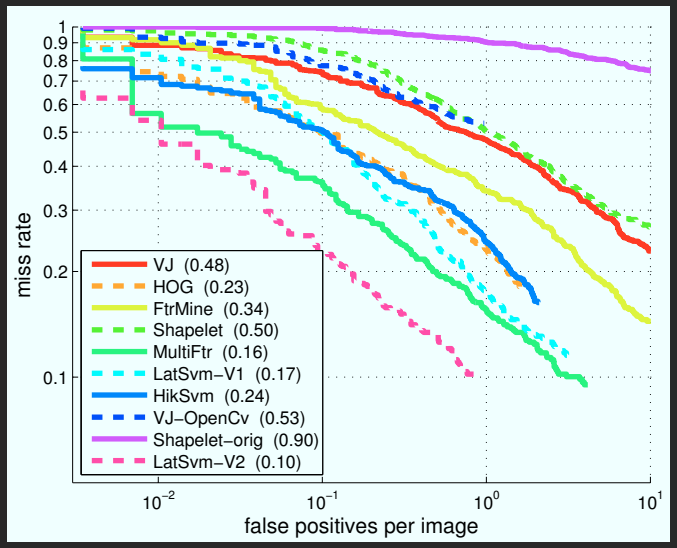
Nuisance issues

1) Positives are not aligned perfectly (search over coordinate frames - translations, euclidean, affine?)
2) High imbalanced class distributions (“infinite” set of negatives)
2) NMS for overlapping detections (smooth response functions?)

Benchmark evaluation - Dollar, Wojek, Schiele, Perona CVPR 09



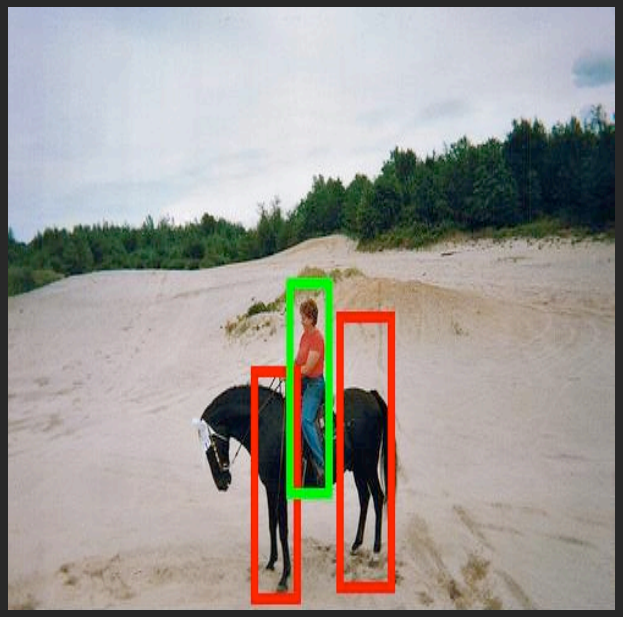
Classification results



Detection results

Over half of DalalTriggs++ papers are worse than DalalTriggs when used as detectors on real images

Large-scale learning



Our test set distribution is highly imbalanced; so should be the training set

(hundreds of positives, hundreds of millions of negatives)

SVMs are attractive because they generate sparse learning problems

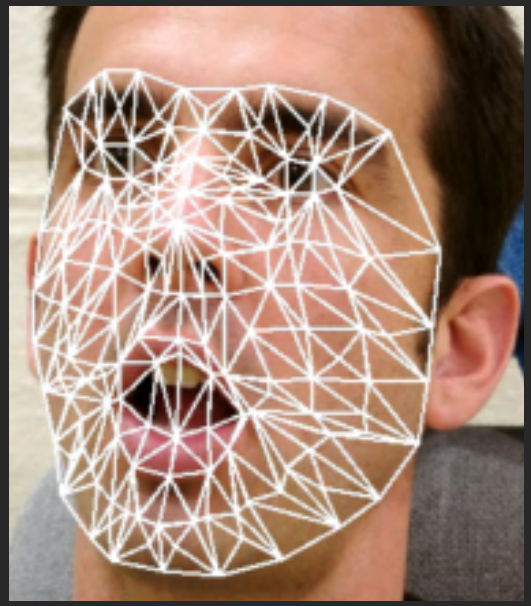
(One can solve problems that are too big to fit in memory)
(hard-negative mining different for SVMs vs Boosting....)

Generative models seem to deal better with imbalanced problems and noisy data
(success of LDA?)

Takeo Kanade's 3 most important vision problems?

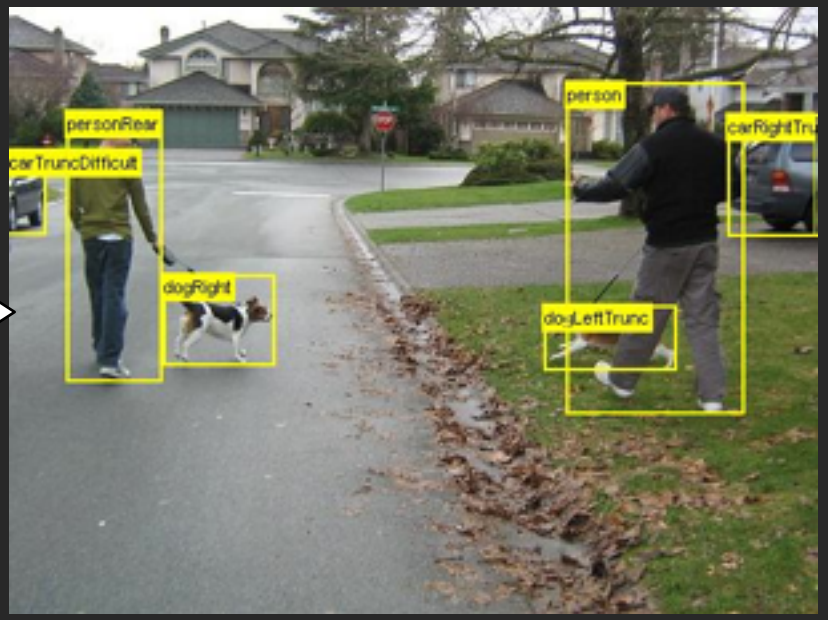
Takeo Kanade's 3 most important vision problems?

1. Alignment
2. Alignment
3. Alignment



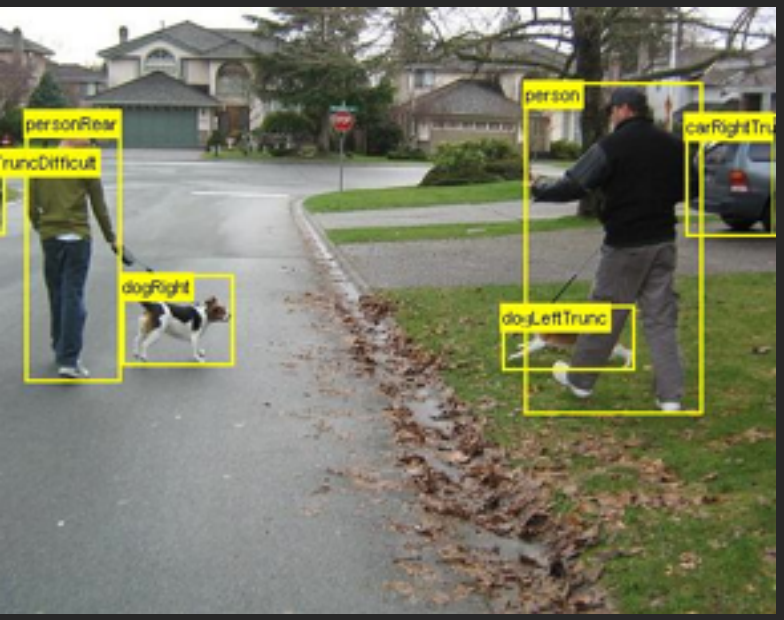
3. Focused community on understanding spatial layout

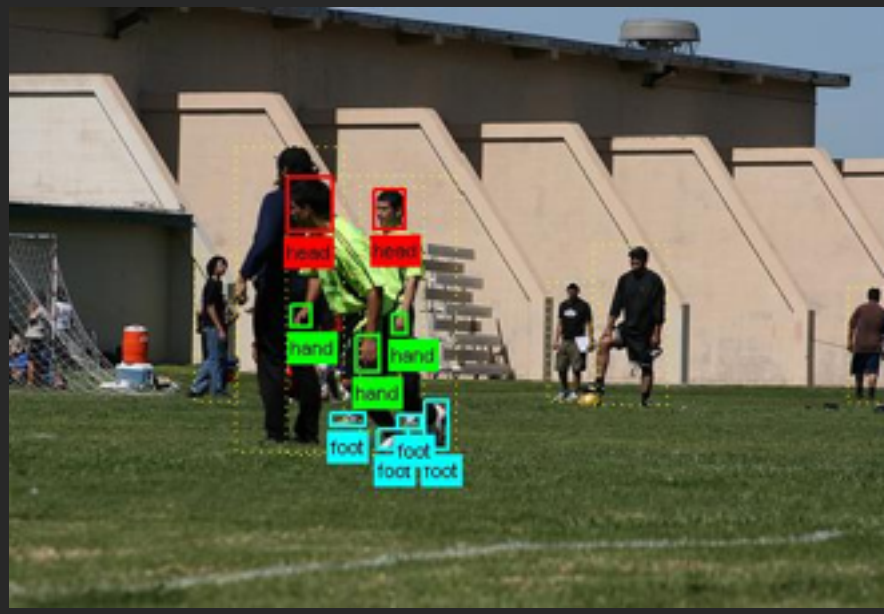
Caltech 101/256



PASCAL VOC

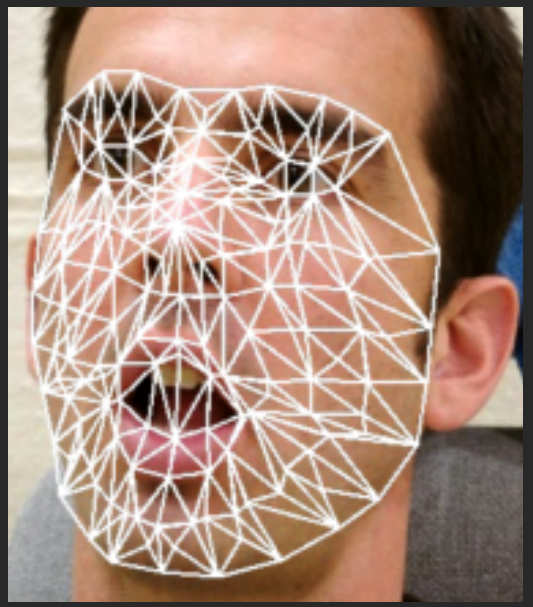
Person Layout





Difficult to score (50% overlap too strict?)
Difficult to attempt (heavy occlusion / truncation)

Claim: alignment is what separates vision from pattern classication

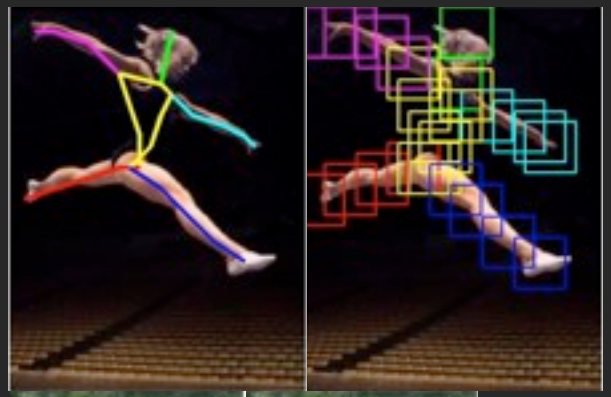


Can a deep belief net output the latter?
If so, then I think its reasoning about shape (which is good!)

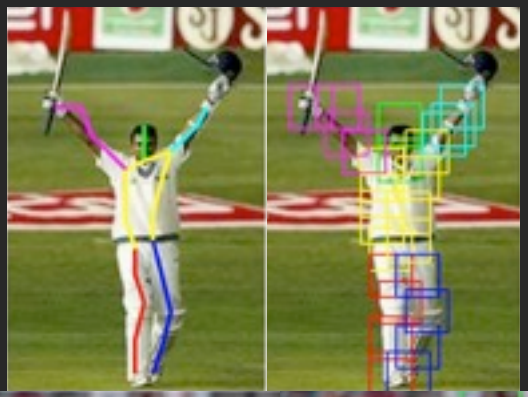
“Fine-grain” shape estimation

Shape gives us a way to define an extremely large set of categories with shared structure

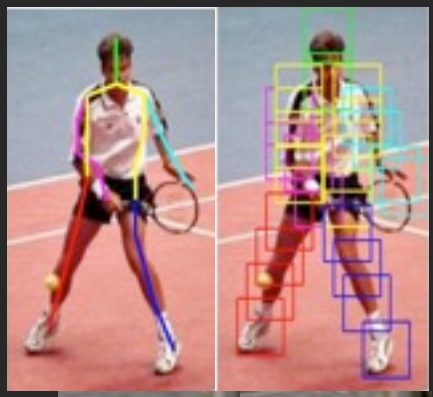
gymnastics

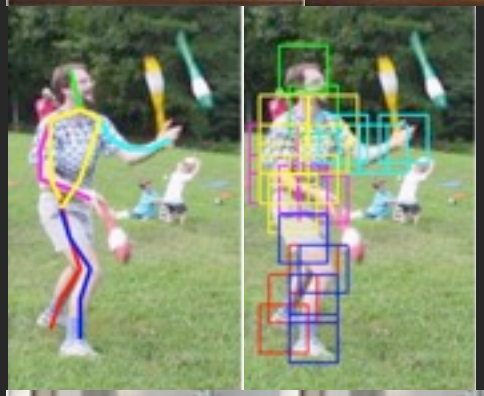


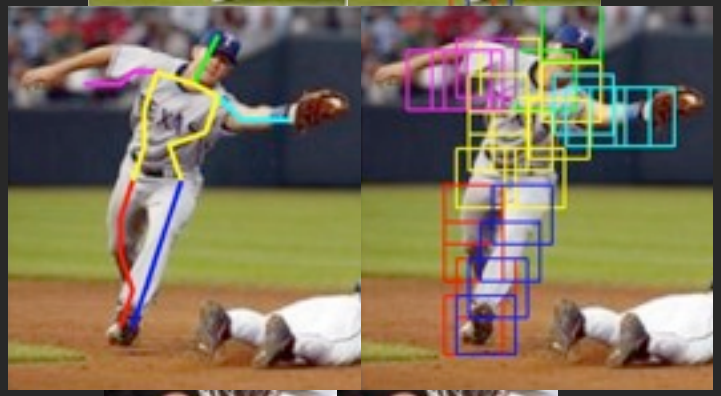
cricket

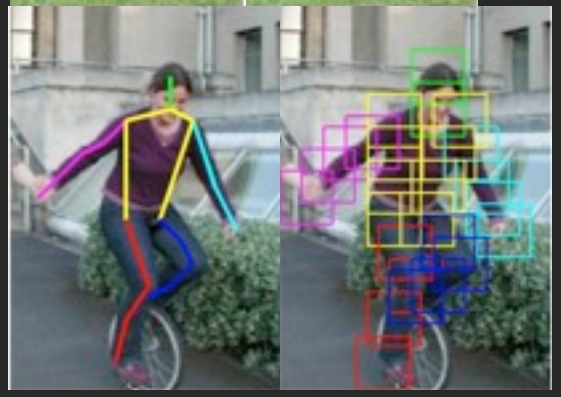


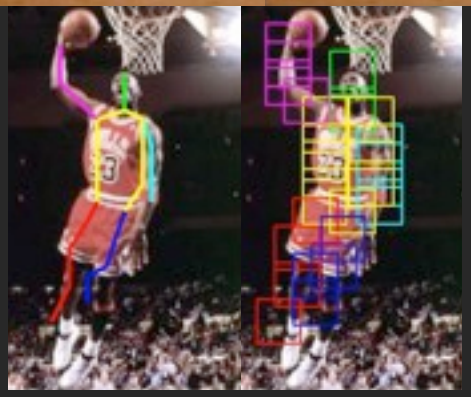
forehand



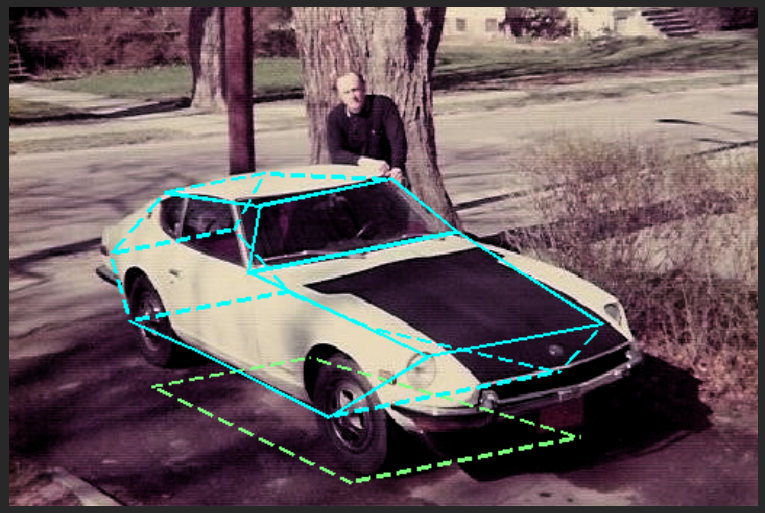


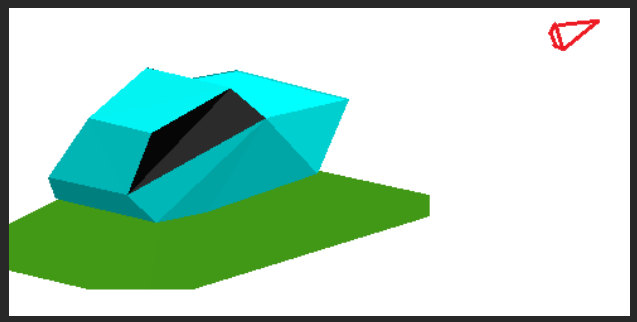


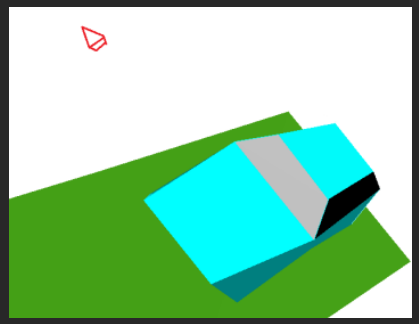




“Fine-grain viewpoint”
= 3D viewpoint estimation



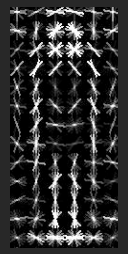




Recall the recognition landscape @ 2005

Structureless

No reasoning about shape
"Parts" had fallen out of favor

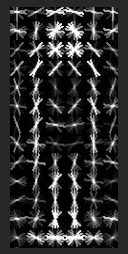


Rigid

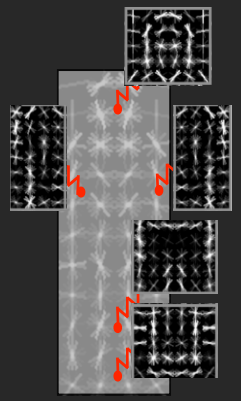
Recall the recognition landscape @ 2005

Structureless

No reasoning about shape
"Parts" had fallen out of favor



Rigid



PASCAL VOC was vital to putting parts,
localization, and "shape" back into the discussion

Common criticisms of PASCAL

1. It encouraged uniformity of thought

1. It encouraged uniformity of thought

I agree with this one

Soln? See Hoeim et al ECCV12

We already collect special purpose datasets to explore a particular phenomena (scale changes, extreme poses). Why not use a single annotated dataset?

2. It has stifled progress

“My method can’t beat the best numbers; I can’t publish”

2. It has stifled progress

"My method can't beat the best numbers; I can't publish"

I don't agree with this

a. It should be hard to do good research (9/10 ideas, at least for me, don't work)

b. The onus is us as researchers to both have a good idea and communicate it. Empirical results are one way to communicate. There are other creative ways; e.g., attributes (introduce problem and create a dataset).

c. See previous soln (more detailed benchmarking)

3. It encourages incremental research

3. It encourages incremental research

I don’t really agree

Are mixture models incremental
(we originally thought so) ?

Is feature engineering incremental
(HOG+LBP+...)?

Are multiscale models incremental?

The original DPM paper (CVPR08) was
dinged for being incremental

Outline

I. Why do part models work?

II. A retrospective on PASCAL

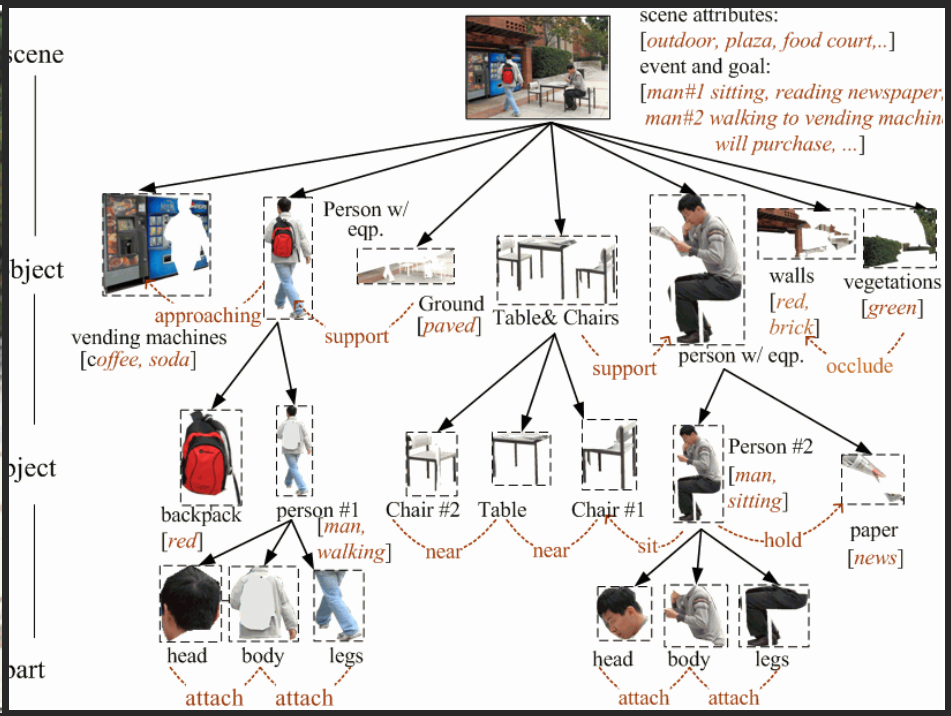
III. A wishlist for PASCAL 2.0

Standard wishlist

More categories, denser labels, etc...

“Hyper annotation”

“Lotus hill”-style annotation

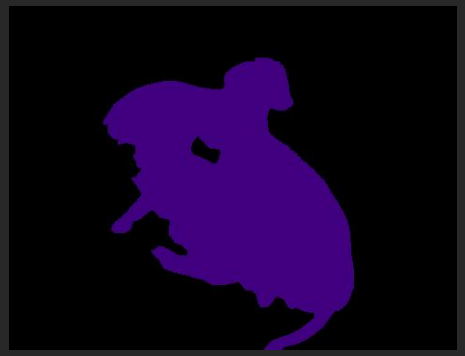


Let’s set our sights a bit more modest

Getting rid of bounding boxes

Getting rid of bounding boxes

Combine detection and segmentation



Semantic segmenation

Instance-level semantic segmentation

Getting rid of bounding boxes

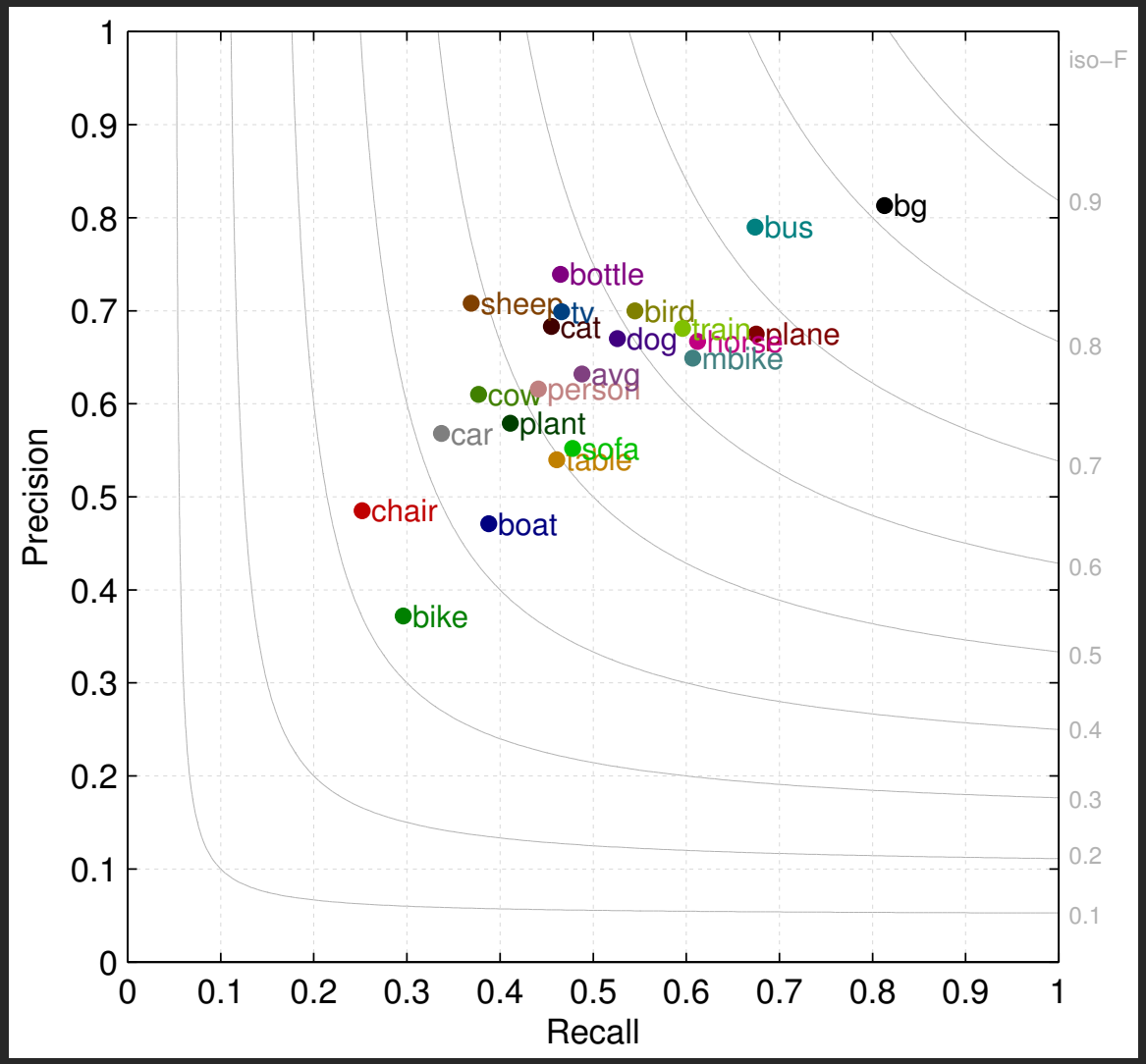
Instance-level semantic segmentation

-Define candidate segment to be a “good” match if intersection/union > .5
-Evaluation criteria is no longer bounding-box dependent (useful for articulation)
-If desired, one could require a “globally-consistent” interpretation of an image

-Assume dogs & cats are confused with one another. One can artificially increase recall of both detectors by returning overlapping detections

-Force detectors to pick an precision-recall operating point

Instance-level segmentation



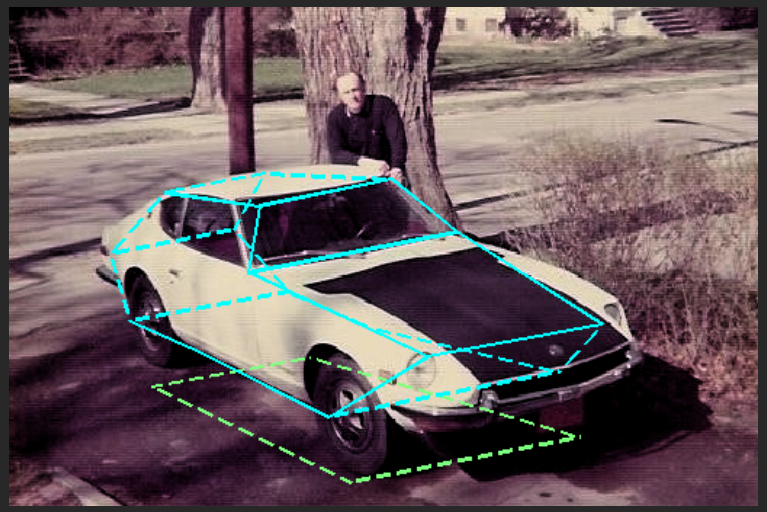
Yang et al , IJCV 11

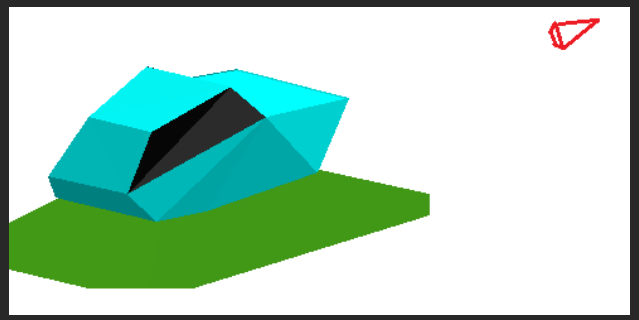
Action classification => Action recognition

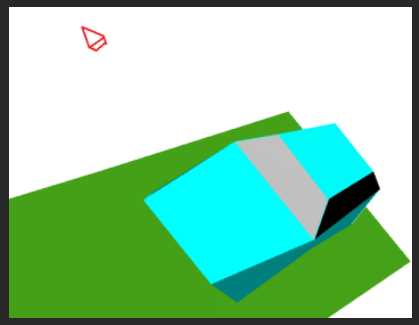
1. Localize person (+ interacting object)
2. Classify action of each detected instance
3. Estimate pose of person (+ interacting object)

“Fine-grain viewpoint” = 3D viewpoint estimation

(and 3D shape?)







A look back

I. Why do part models work? Extrapolation to unseen data

II. A retrospective on PASCAL It was great - thanks!

III. A wishlist for PASCAL 2.0 More annotations and diagnostics