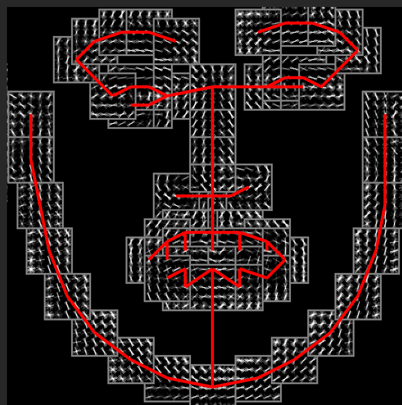
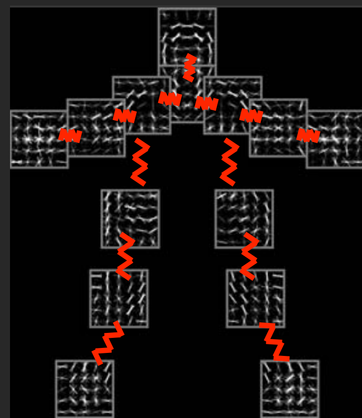
Extensions of part models

Deva Ramanan

UC Irvine





Outline

"Core" deformable part model system

(This morning)

"Extensions" of deformable part models

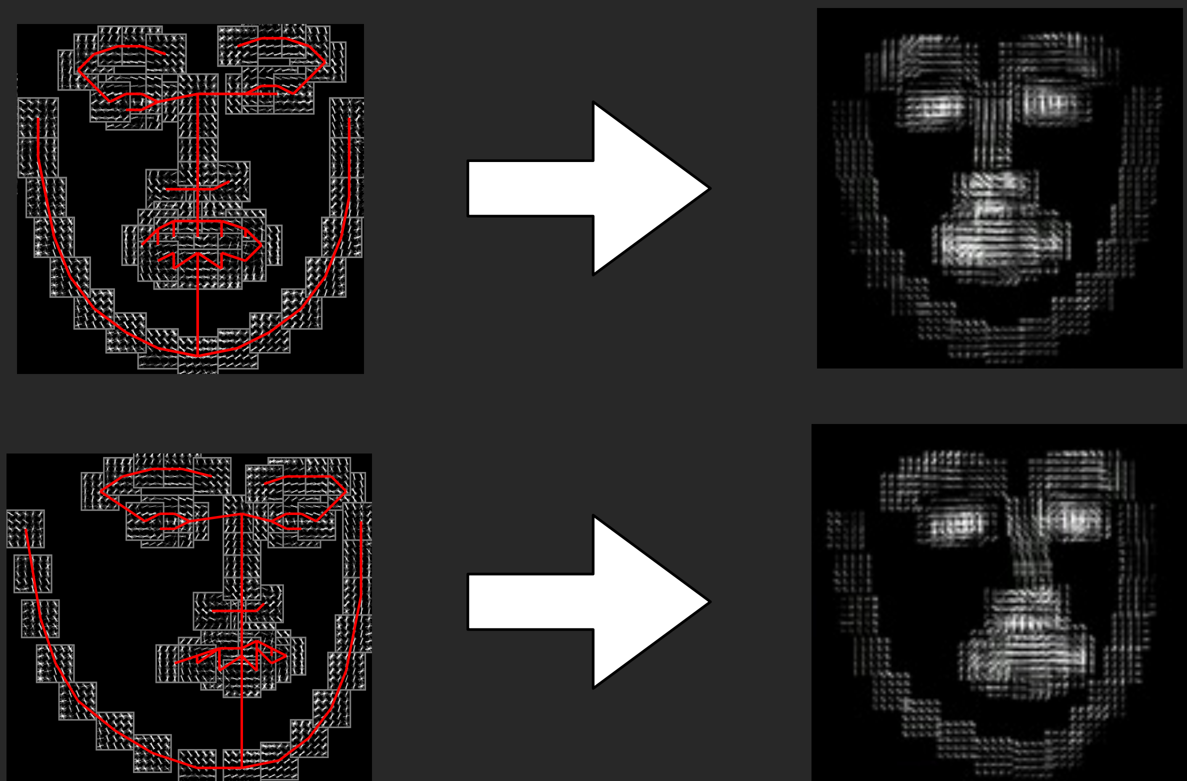
(This afternoon)

Parts as large mixture models

$$S(x,z) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{ij \in E} w_{ij} \cdot \psi(z_i, z_j)$$

Each distinct placement of parts yields a unique global template

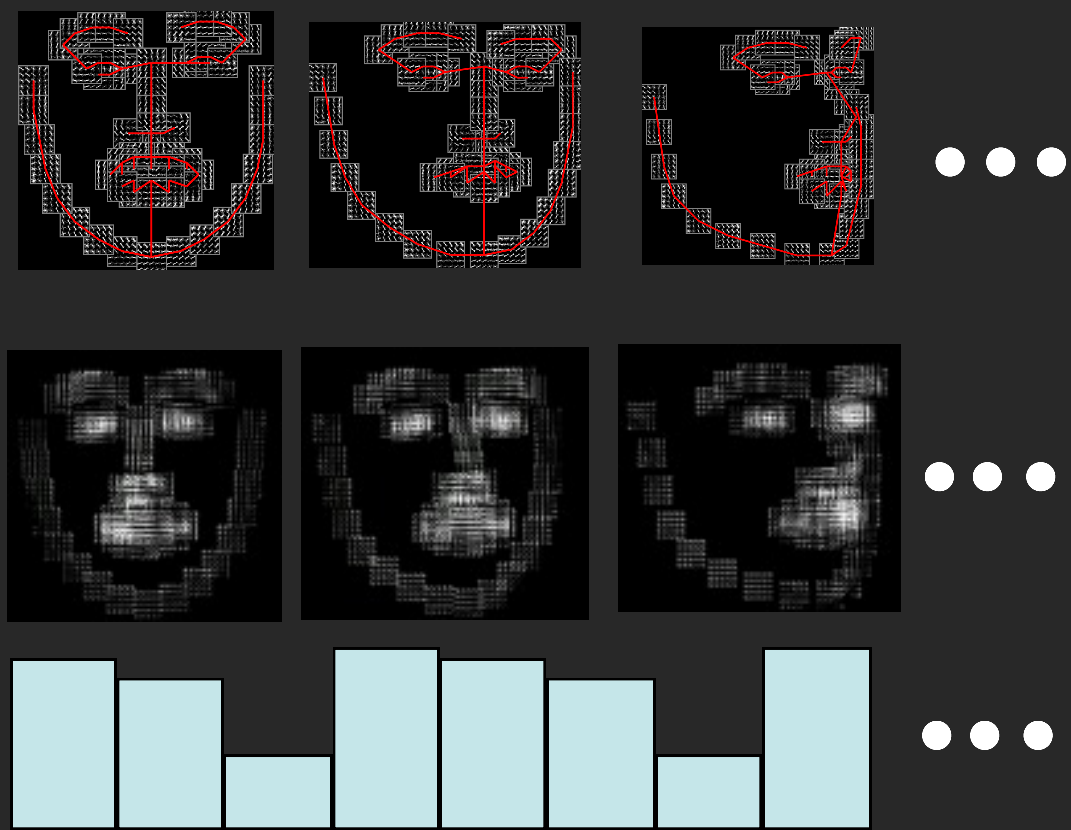
$$S(x,z) = w_z \cdot x + b_z$$



Parts as mixture models

Spatial model defines bias or "prior"

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

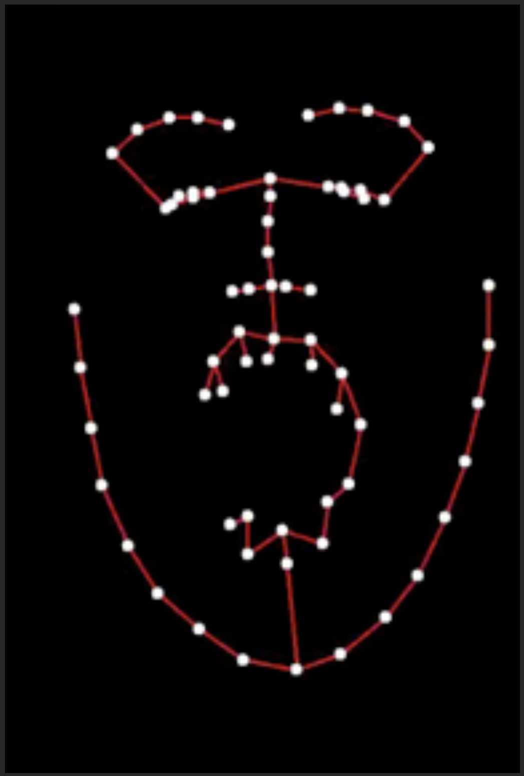


Parts as mixture models

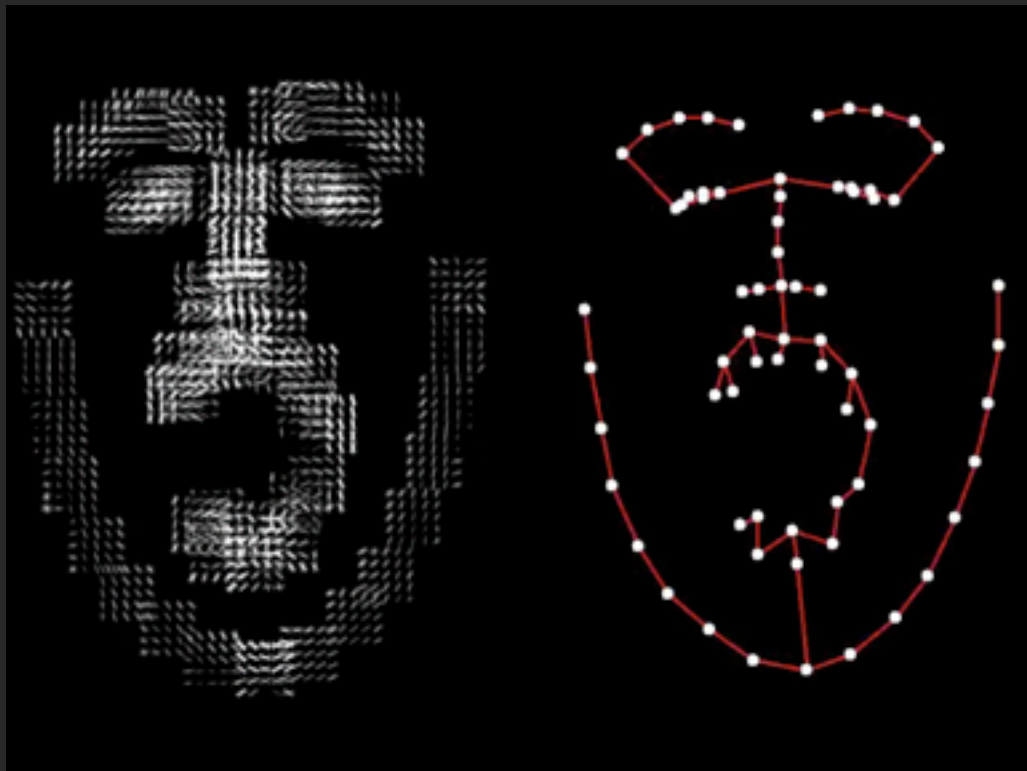
Part models allow us to represent an exponentially-large family of global templates

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

Deformation modes



Deformation modes



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

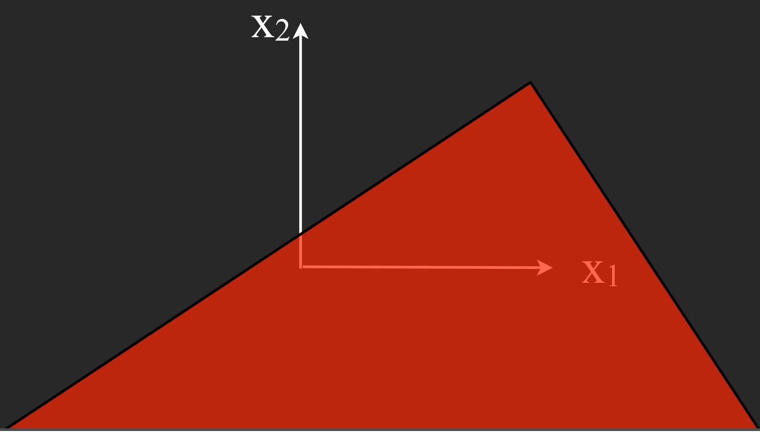
(more on that later)

Revisit latent (vs linear) classification

$$f_w(x) = w \cdot x$$

Score is linear in x

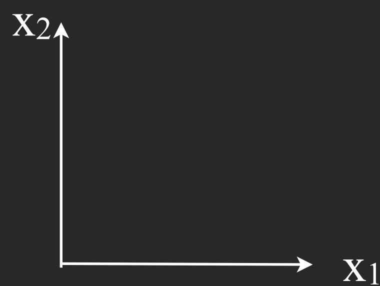
Positive set $\{x : f_w(x) > 0\}$ is half-space



$$f_w(x) = \max_z w_z \cdot x$$

Score is ?

Positive set is ?

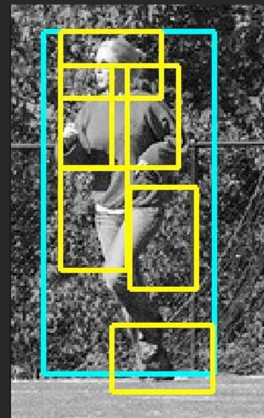


Revisit latent (vs linear) classification

$$f_w(x) = w \cdot x$$

Score $f_w(x)$ is linear in x

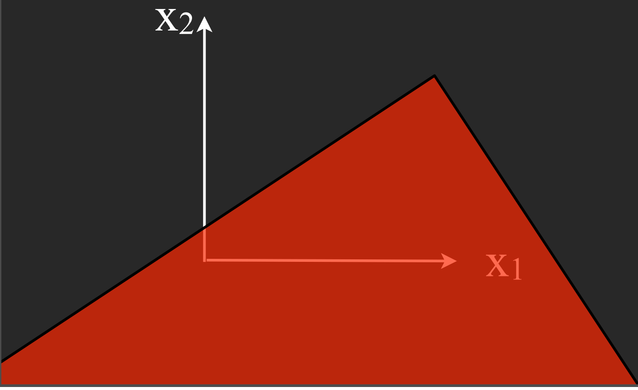
Positive set $\{x : f_w(x) > 0\}$
is half-space

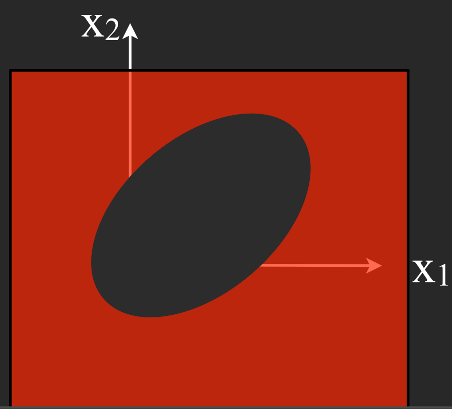


$$f_w(x) = \max_z w_z \cdot x$$

Score $f_w(x)$ is convex in x

Positive set $\{x : f_w(x) > 0\}$
is concave

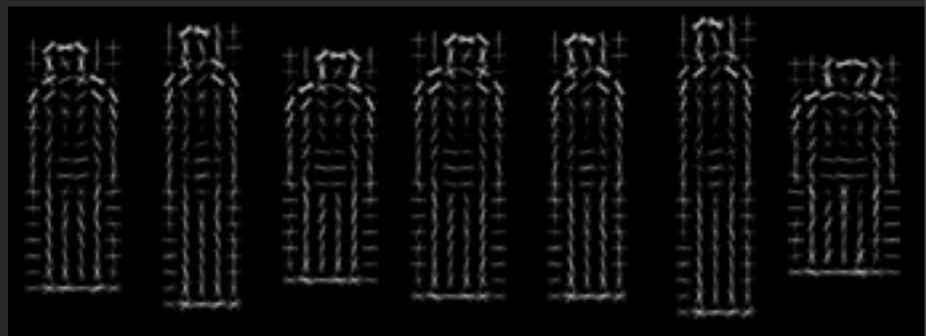




Crucial aspects

1) Efficient discriminative learning

2) Efficient inference



Efficient learning

Lots of large-scale solvers for quadratic programs (SVMS)

Two flavors

Batch: Require access to all training data

(guarantees on convergence)

Online: Require access to on-the-fly training data

(usually stochastic in practice)

In-between: Support-vectors fit in memory, but data doesn't

(Relatively unexplored!)

Online dual solvers

In practice, can get near-optimal models with a single pass through large datasets

A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with LaRank. In *ICML*, pages 89–96. ACM, 2007. 7

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *The Journal of Machine Learning Research*, 6:1579–1619, 2005. 8

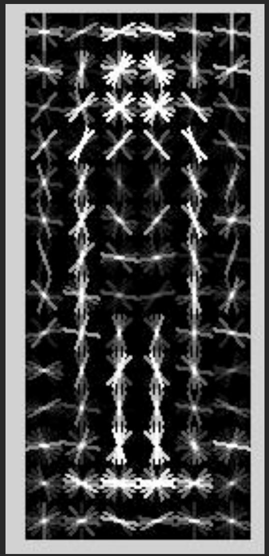
L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20:161–168, 2008. 1

http://www.csie.ntu.edu.tw/~cjlin/liblinear/

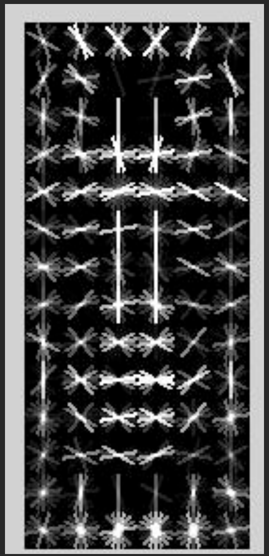
Recall: why are we bothering training large-scale classifiers?

$$w_{pos} \cdot x > w_{neg} \cdot x$$

Pedestrian template



$>$

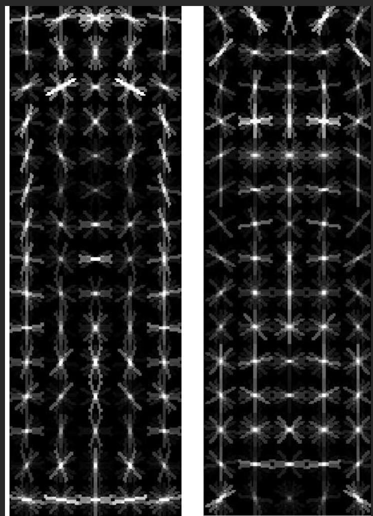


Pedestrian background template

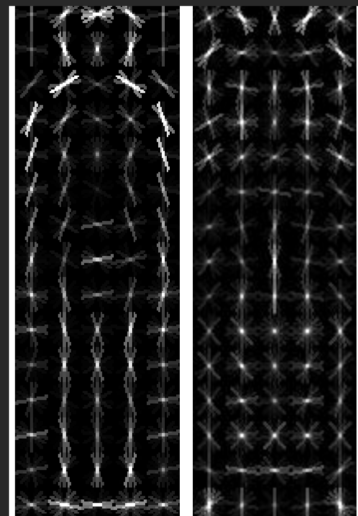
Right approach is to compete pedestrian, pillar, doorway... models

Background class is hard to model - easier to penalize particular vertical edges

Do we really need this machinery?

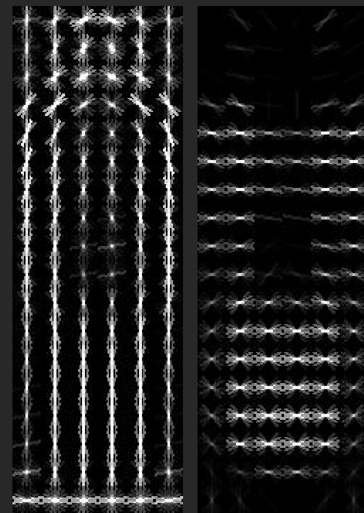


SVM



Gaussian model

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



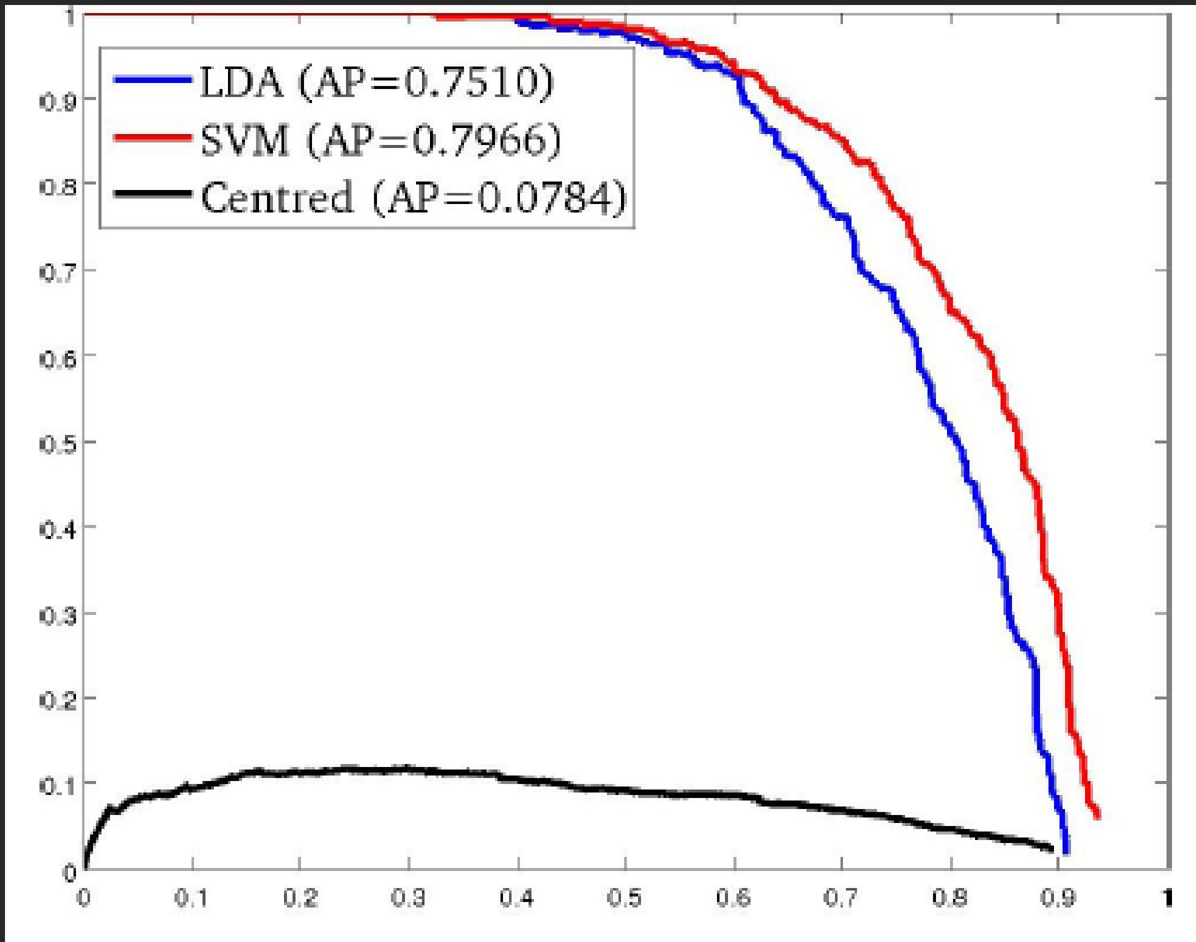
Centered model

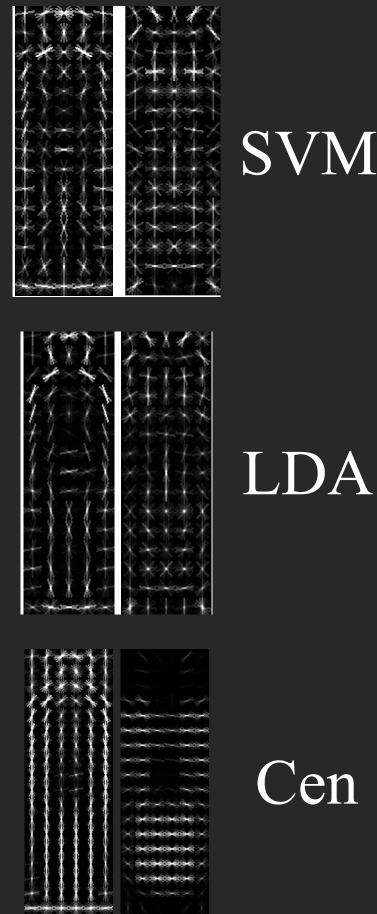
$$w = \mu_1 - \mu_0$$

Learn templates with simple statistical (de)correlation models

Hariharan, Malik, Ramanan ECCV 12

Linear discriminant (LDA) models

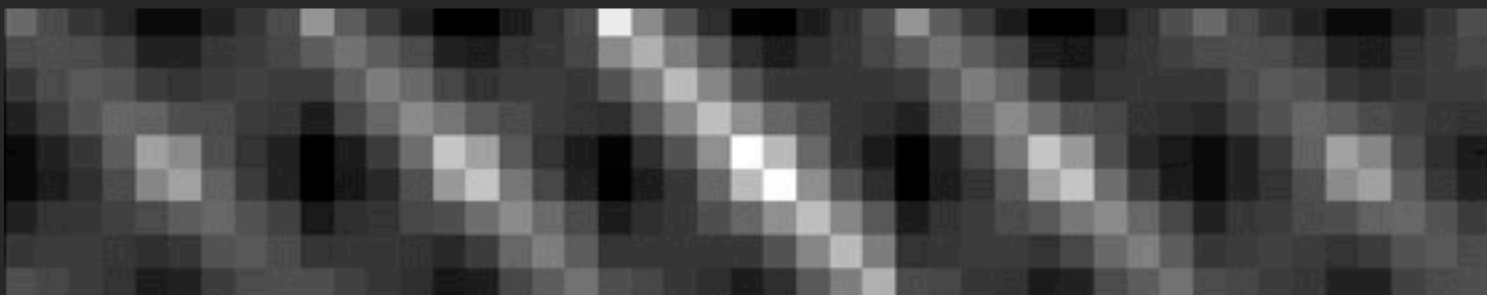




Properties of spatial covariance matrix

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

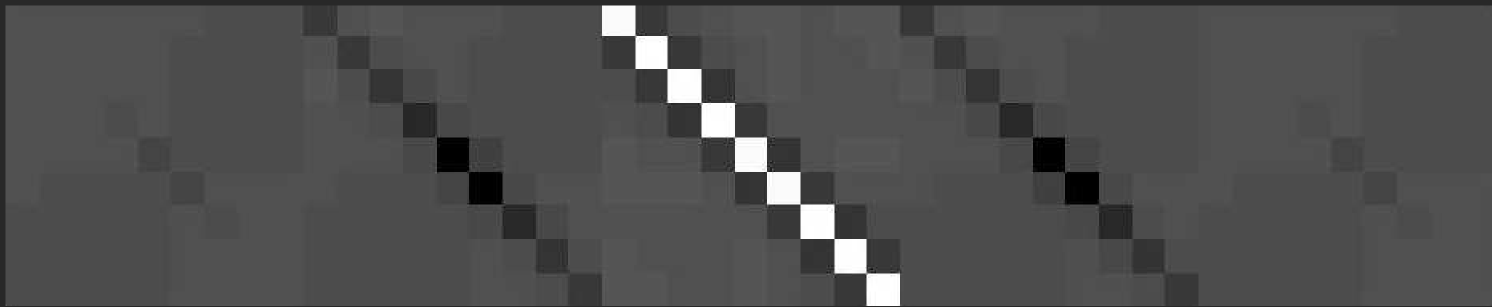
Can be efficiently encoded with a set of 36x36 matrices Sig_{i-j}



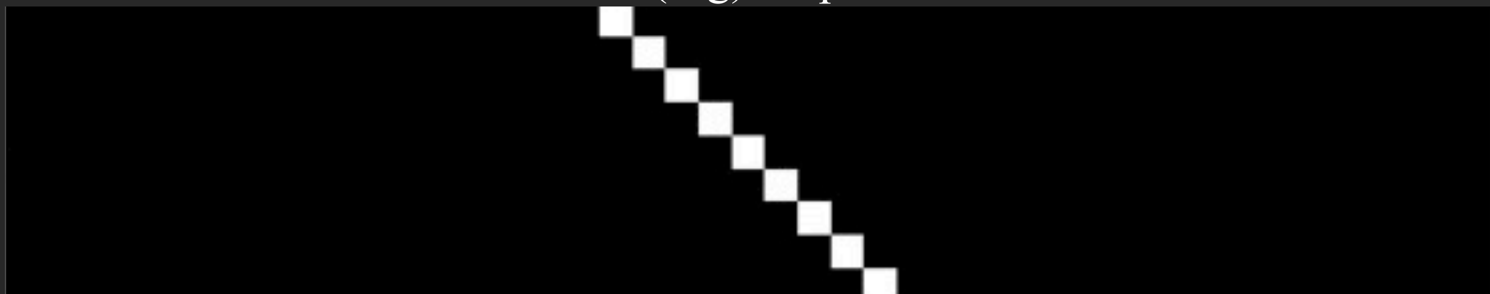
Sig_{-2} Sig_{-1} Sig_o Sig_1 Sig_2

Properties of spatial precision matrix

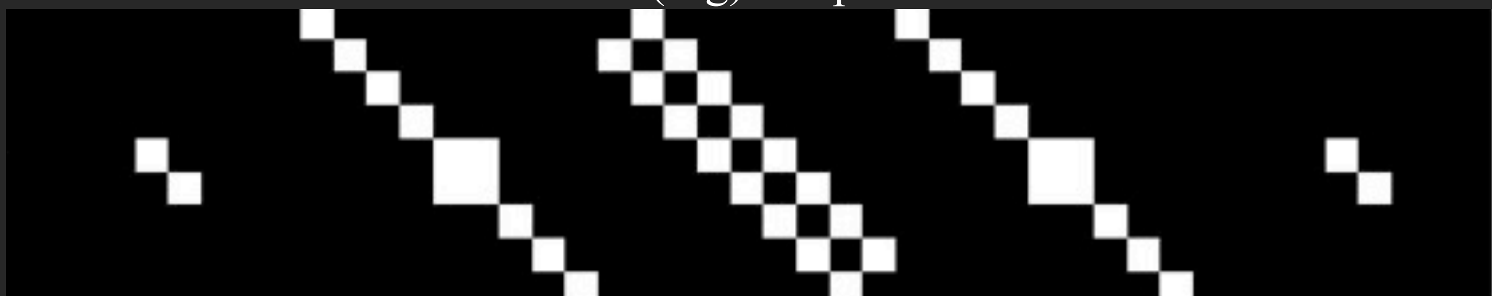
Inv(Sig) is sparse



Inv(Sig) > eps



Inv(Sig) < -eps



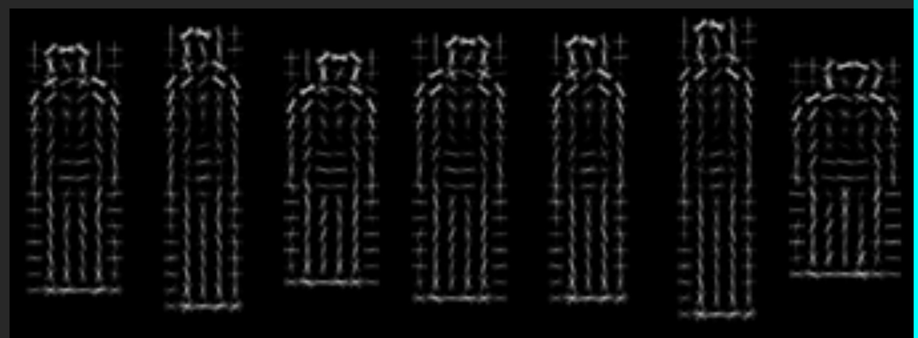
Inv(Sig) subtracts correlated gradients (at neighboring orientations and windows)

Crucial aspects

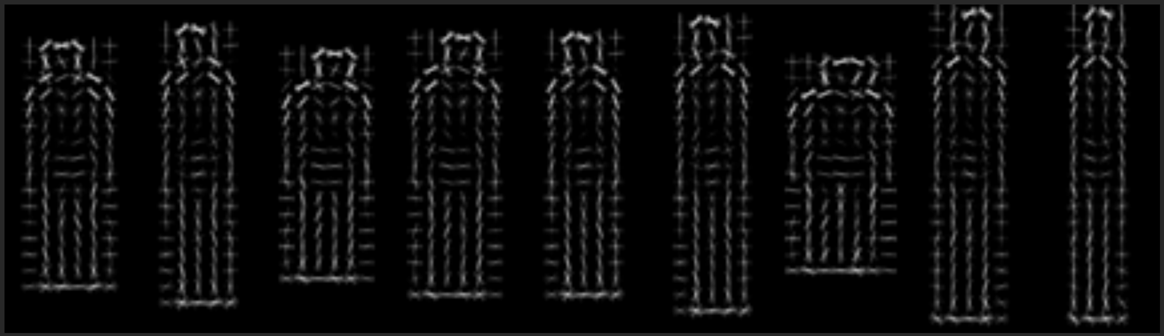
1) Efficient discriminative learning

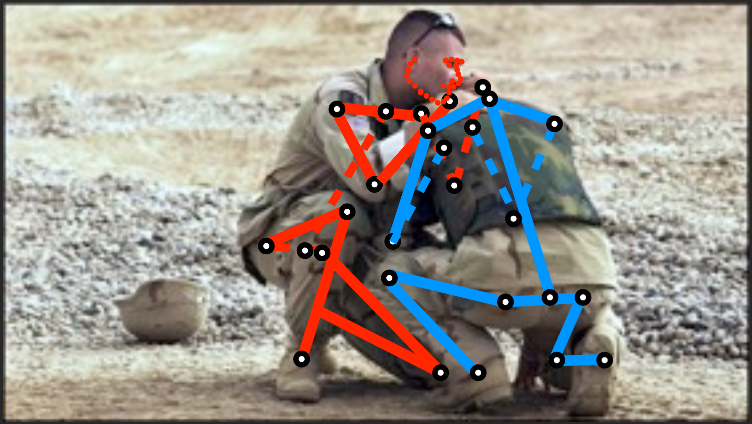
Bottom-line: parameter can be tuned with a single-pass over data

2) Efficient inference

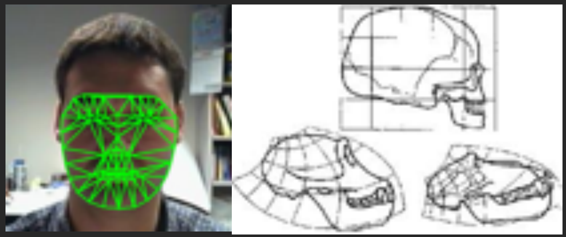


Are quasi-rigid templates enough?



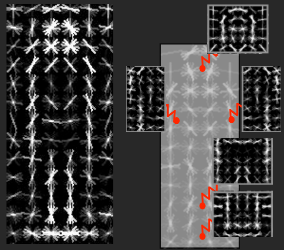


Spectrum of shape models



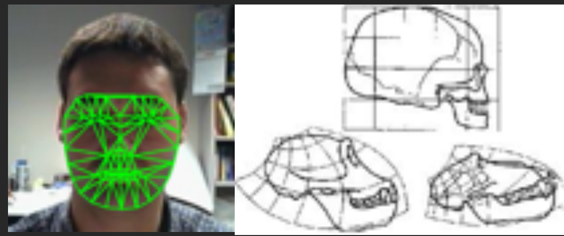
Elastic

Structureless



Quasi-rigid

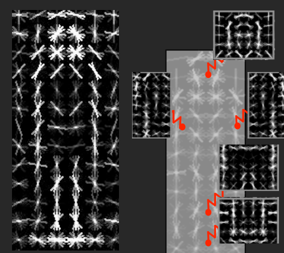
Spectrum of shape models



Elastic

right model, but hard
to compute with

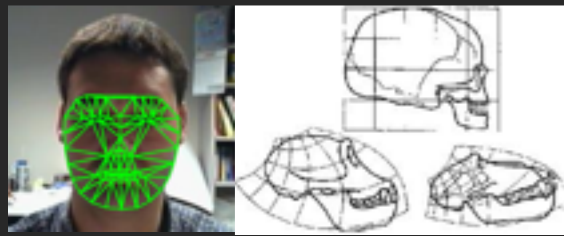
wrong models, but
simple to compute with



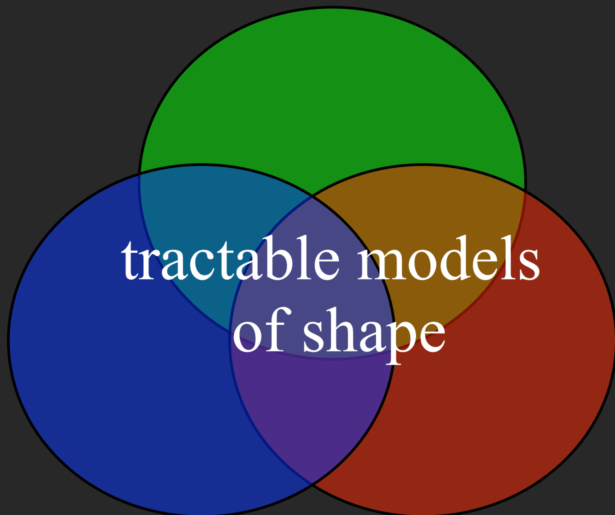
Structureless

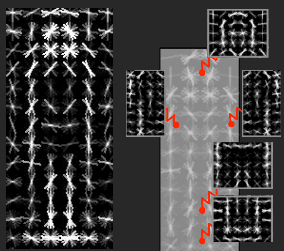
Quasi-rigid

Trifecta of shape



Flexible models

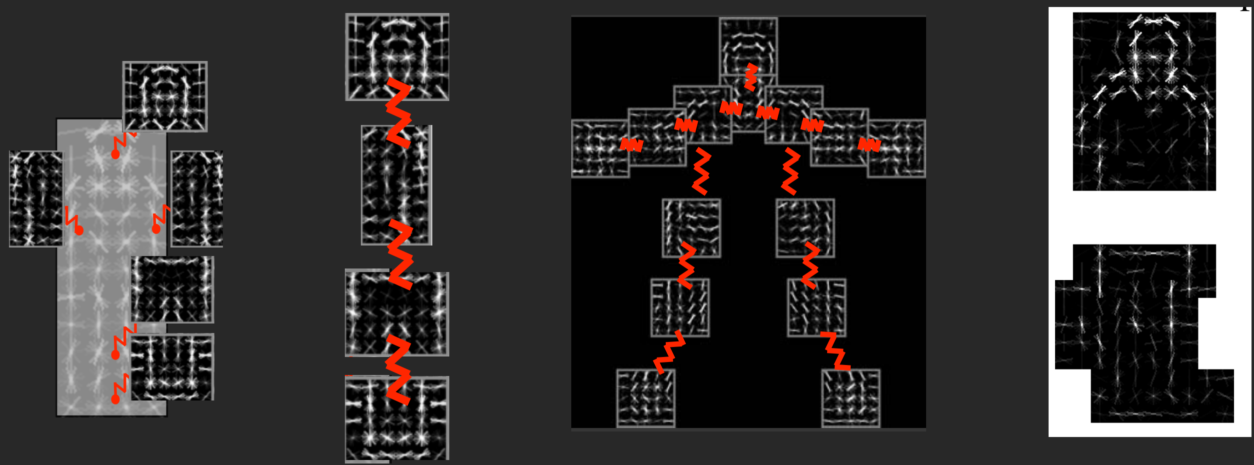




Quasi-rigid

Structureless

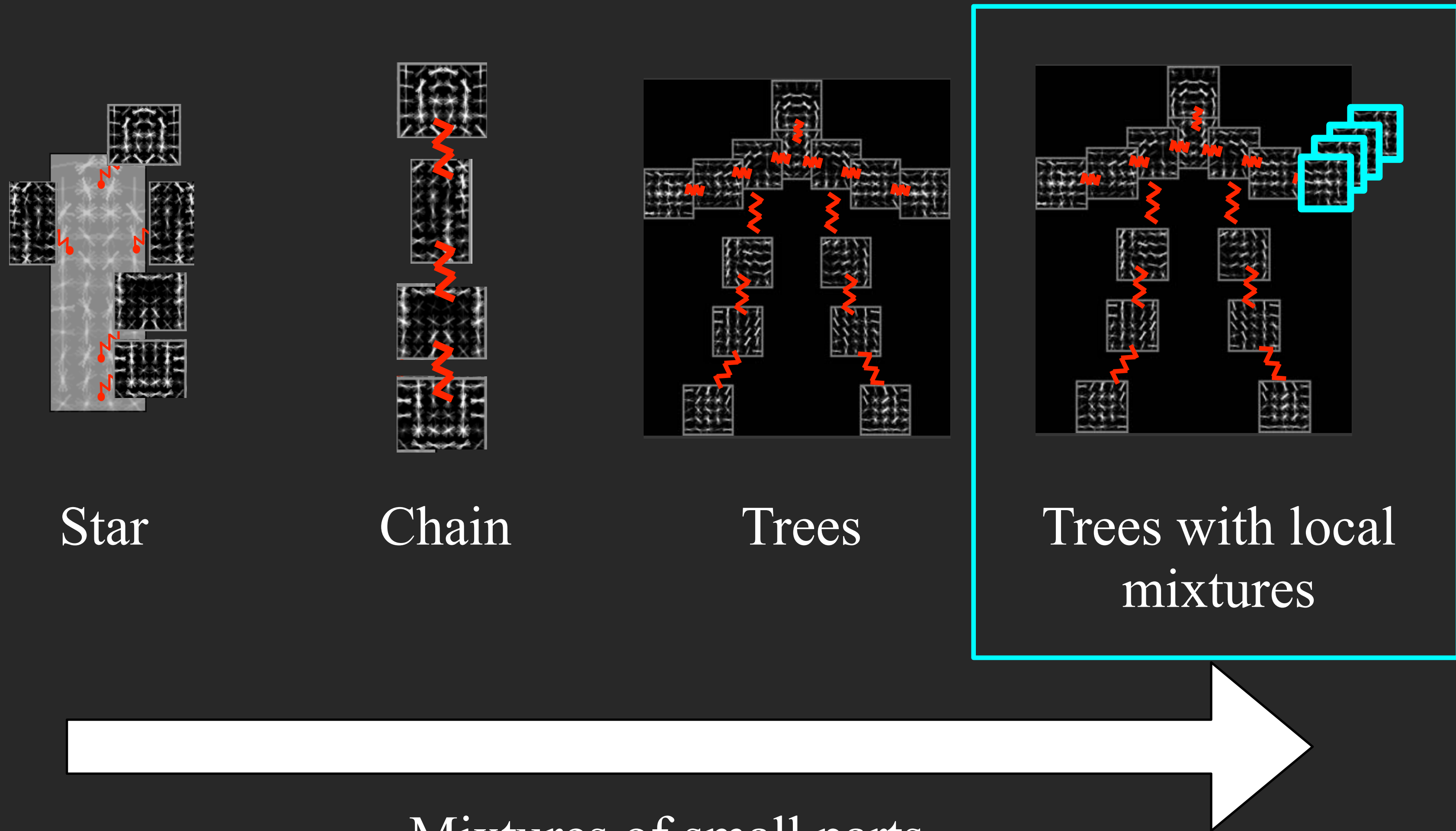
Tractable shape



Star Chain Tree Grammars

Increasingly flexible

Recent work: flexible representations



Star

Chain

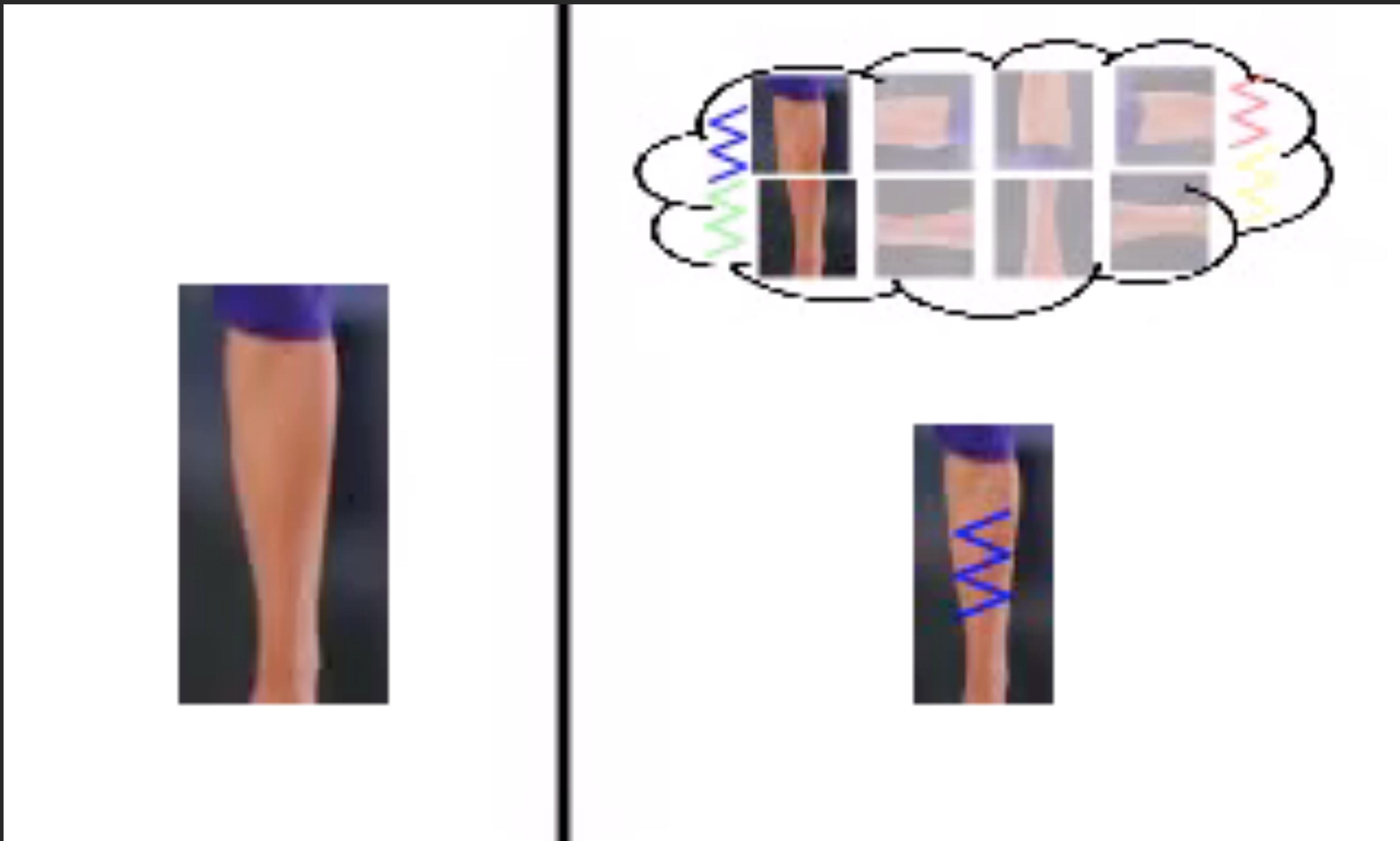
Trees

Trees with local mixtures

Mixtures of small parts

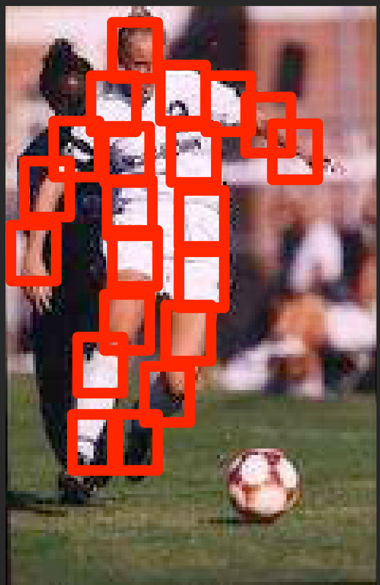
Make use of heavily supervised correspondences

One solution: local mixtures of small patches

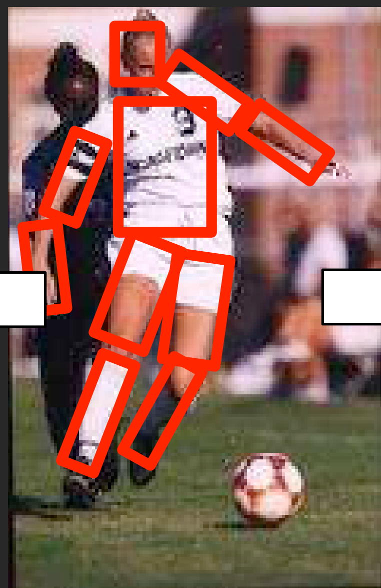


Any smooth spatial transformation is locally rigid

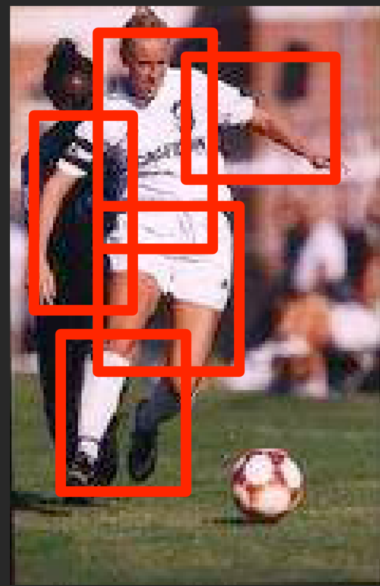
What are the right parts?



Smaller parts



Larger parts



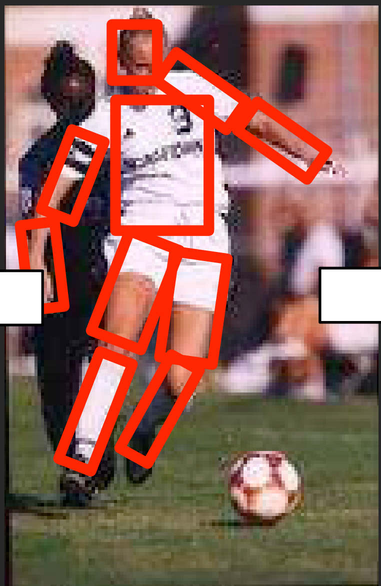
Patches

Skeleton

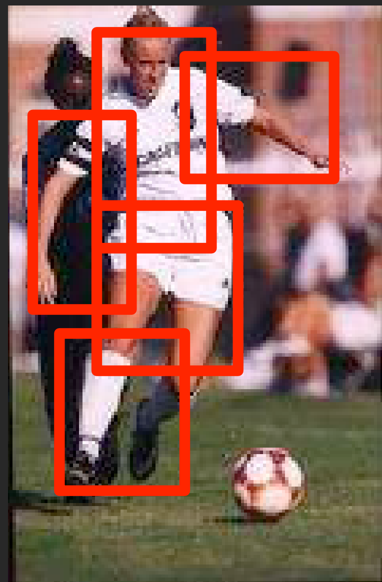
Poselets

What are the right parts?

er



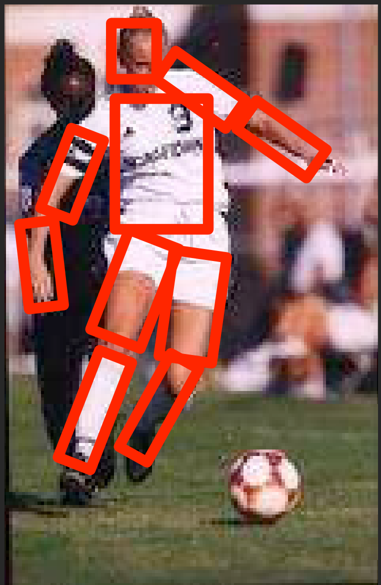
Larger
parts



Skeleton

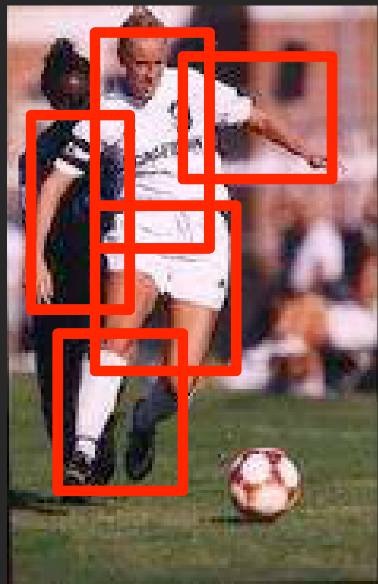
Poselets

Coarser representations



Skeleton

Ioffe & Forsyth
zenswalb & Huttenlocher
Johnson & Everingham
Andruikula et al.
Ferrari et al.



Poselets

Bourdev & Malik
Maji et al.
Yang & Mori
Wang & Yang

Exemplars

Malisiewicz et al
Mori & Malik
Shaknarovich & Darrell
Johnson & Everingham

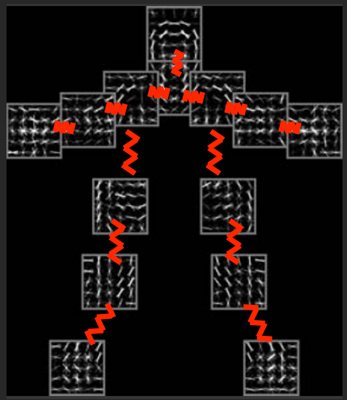
Visual Phrases

Sadeghi and Fahardi

The flaw behind "classic" parts

(Flawed) assumption: local appearance and global geometry are independent

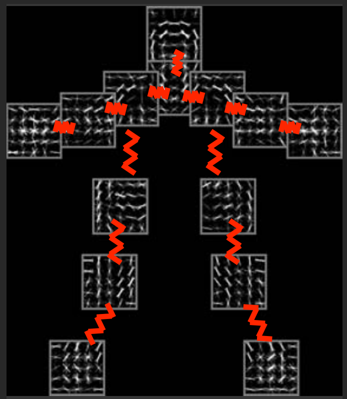
(e.g., head looks the same no matter the geometry of the rest of the body)



The flaw behind "classic" parts

(Flawed) assumption: local appearance and global geometry are independent

(e.g., head looks the same no matter the geometry of the rest of the body)



Fails for....

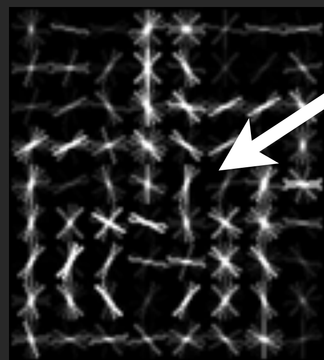
Occlusion

3D viewpoint

Articulation

Visual Phrases

Sadeghi and Fahardi, *CVPR* 11

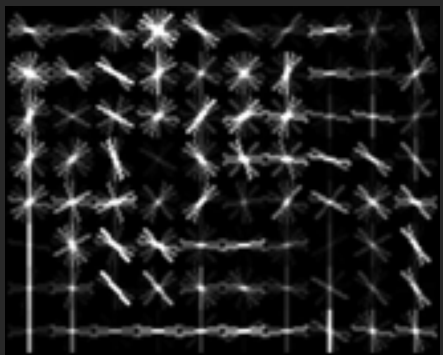


Occluded leg not present in template

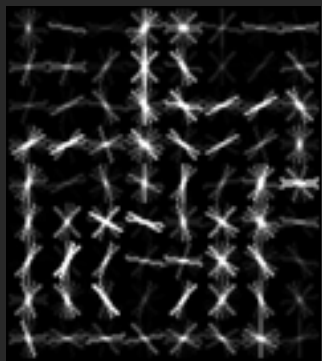
Person on horse

Visual Phrases

Sadeghi and Fahardi, *CVPR* 11



Person on jumping horse



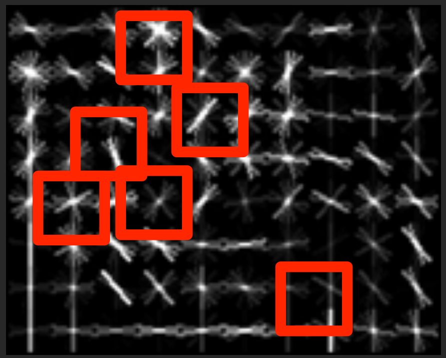
Person on horse

Person standing next to horse

Problem: one may need lots of large composite templates

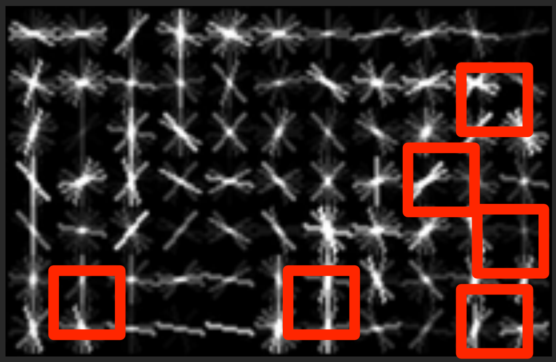
Visual Phrases

Sadeghi and Fahardi, *CVPR* 11



Person on jumping horse

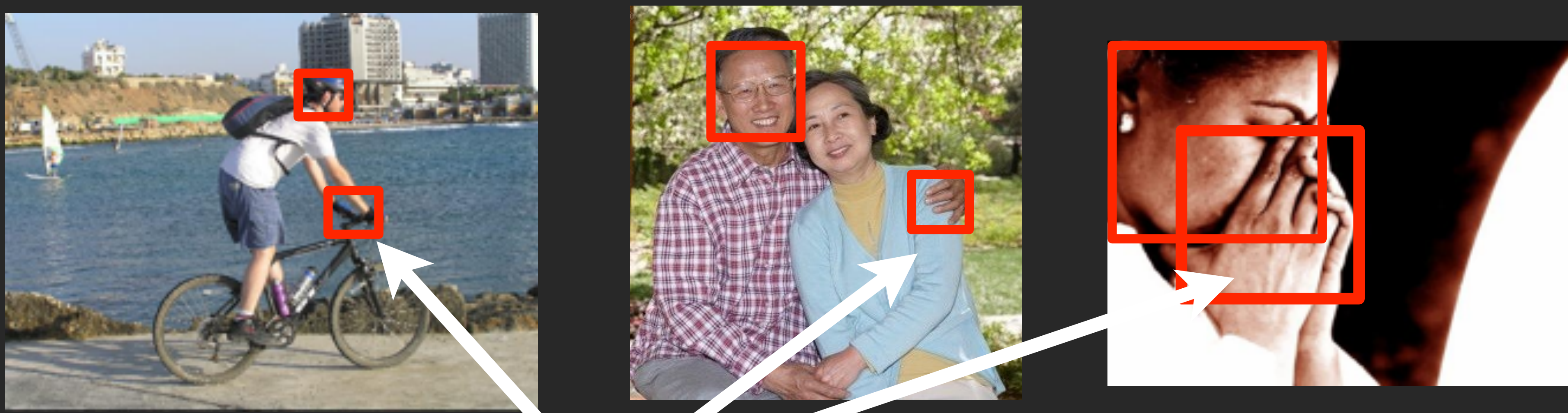
Person on horse



Person standing next to horse

Solution: cut up composites into patches that can be mixed and matched

Visual "phraselets"



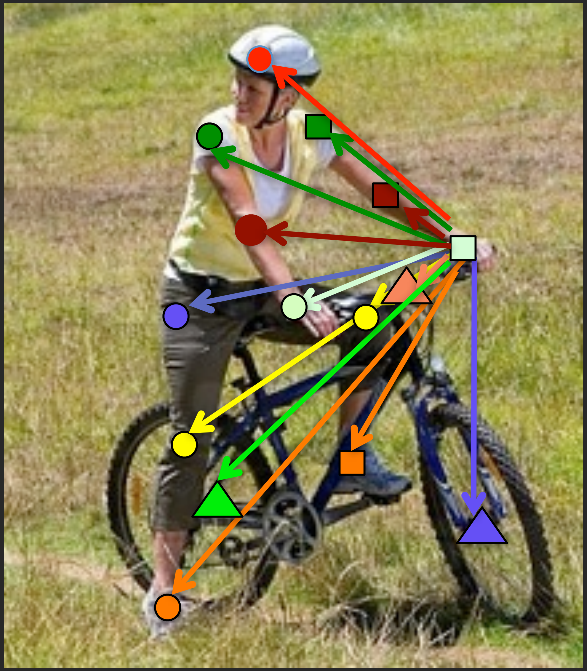
Hand looks different due to interactions with global geometry

We'll encode such visual differences as local part mixtures

Learning phraselets

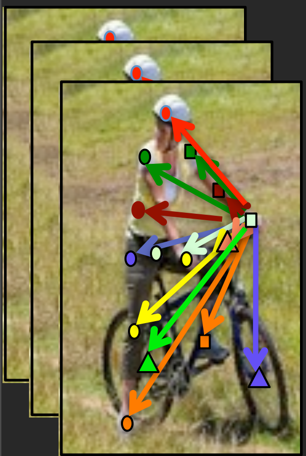
Define phraselets as commonly-occuring geometric configurations

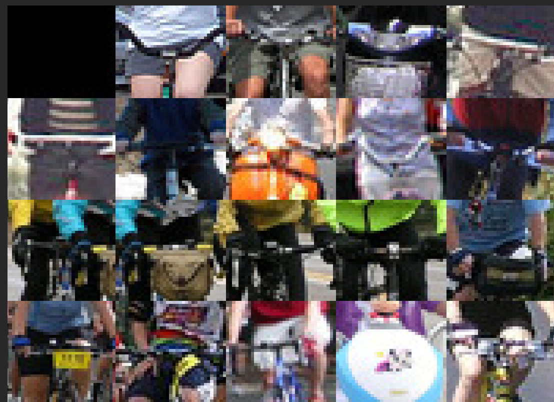
"Poselet-like" clusters



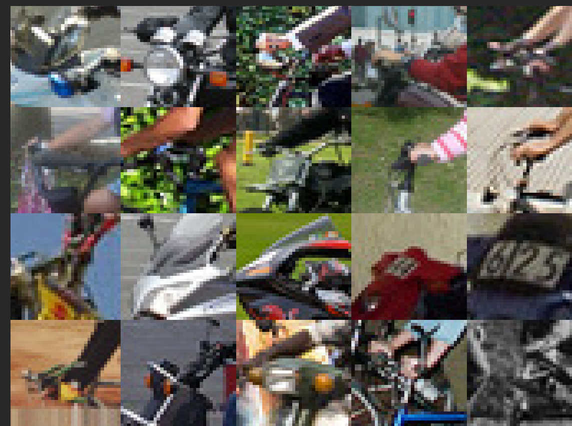
Given labelled training data, find clusters of keypoint configurations relative to each joint

Geometrically-defined hand clusters

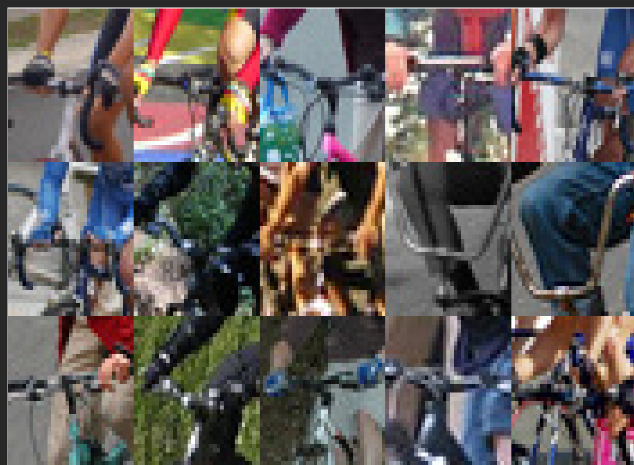




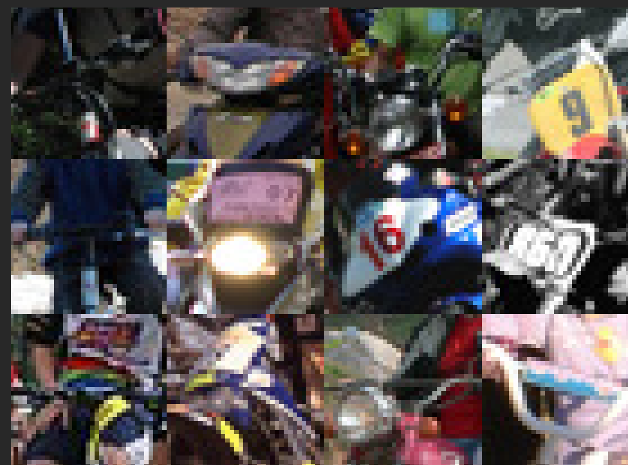
frontal



side-view



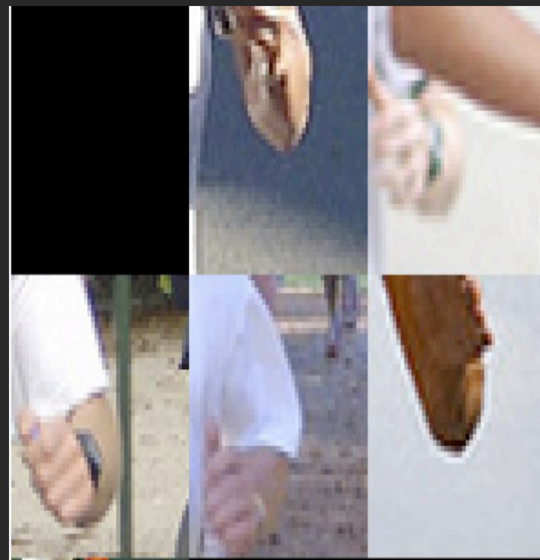
3/4 road bikes



3/4 motorbikes

Model occlusions with separate clusters

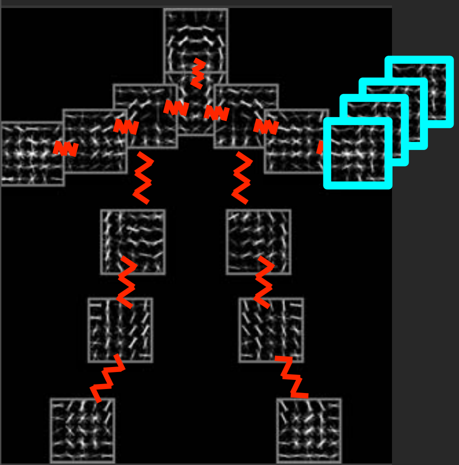
Visible left elbow

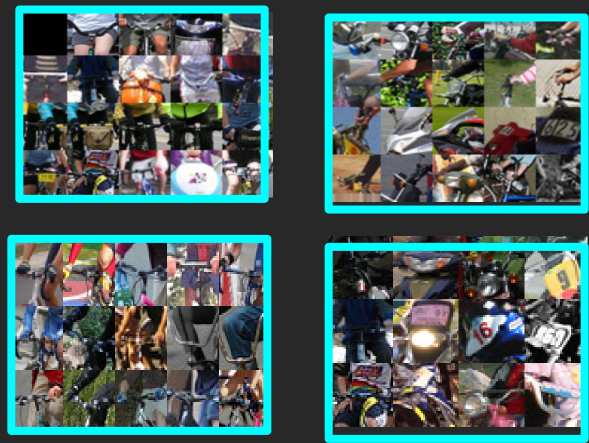


Occluded left elbow

Mixture label corresponds to orientation/ viewpoint and visible/occlusion state

Local mixtures of parts





$$p_i = (x_i, y_i)$$

$$t_i \in \{1, \ldots, T\}$$

$$S(x,p,t) = \sum_i w_i^{t_i} \cdot \phi(x,p_i) + \sum_{ij \in E} w_{ij}^{t_i,t_j} \cdot \psi(p_i,p_j) + S(t)$$

Each part has a position ‘p’ and mixture type ‘t’

Score local model with one of T templates

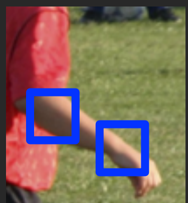
Score deformation with one of T^2 springs (interdependence of geometry + appearance)

Learn rigidity

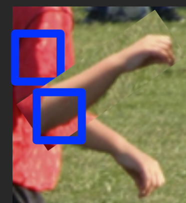
(when mixtures correspond to orientation)

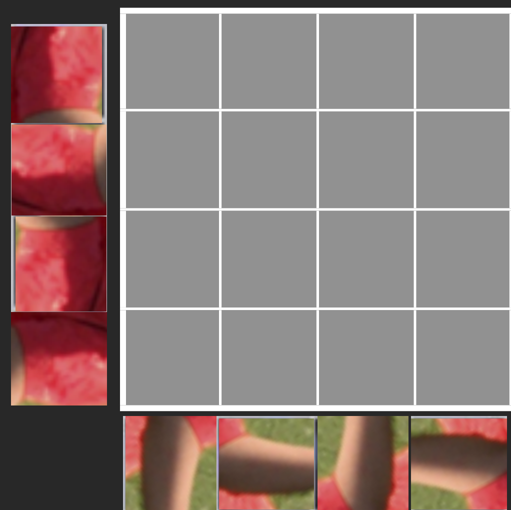
$$S(t) = \sum_{ij \in E} b_{ij}^{t_i, t_j}$$

Rigid relation



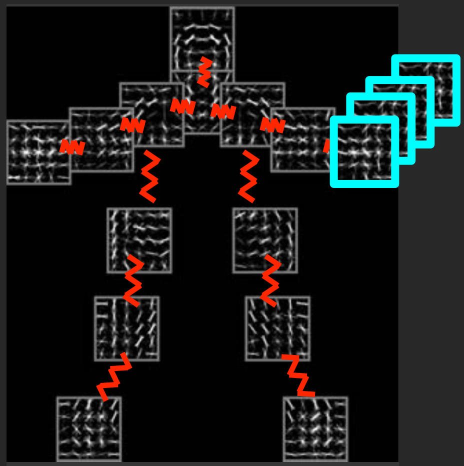
Flexible relation





Learn self-occlusion constraints

(when mixtures correspond to occlusion states)

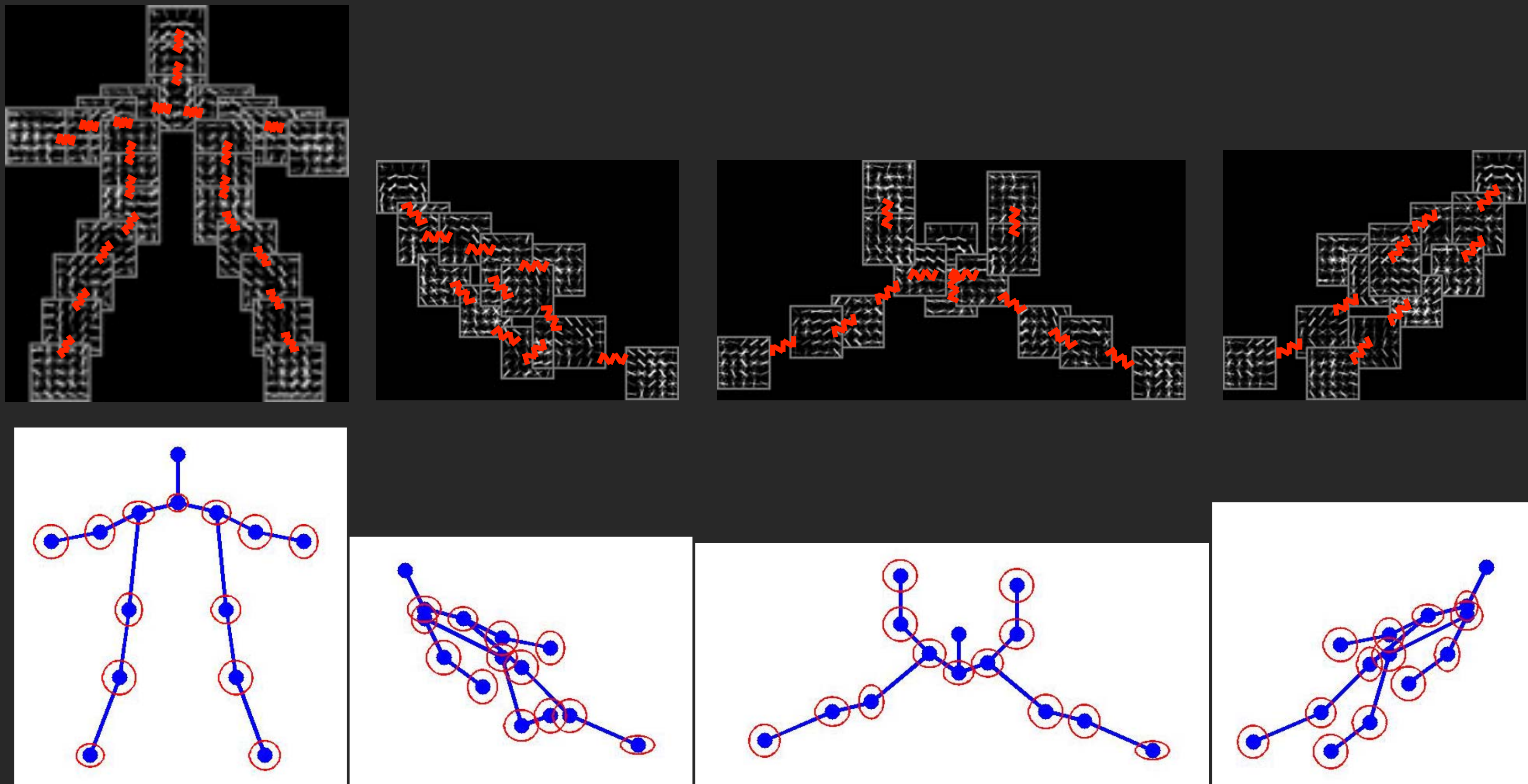


$$\boxed{\sum_{ij \in E} w_{ij}{}^{t_i, t_j} \cdot \psi(p_i, p_j) +} \boxed{\sum_{ij \in E} b_{ij}{}^{t_i, t_j}}$$

-If upper leg is occluded, model can learn to force lower-leg to be occluded

-If left and right hip are near each other, then both cannot be visible

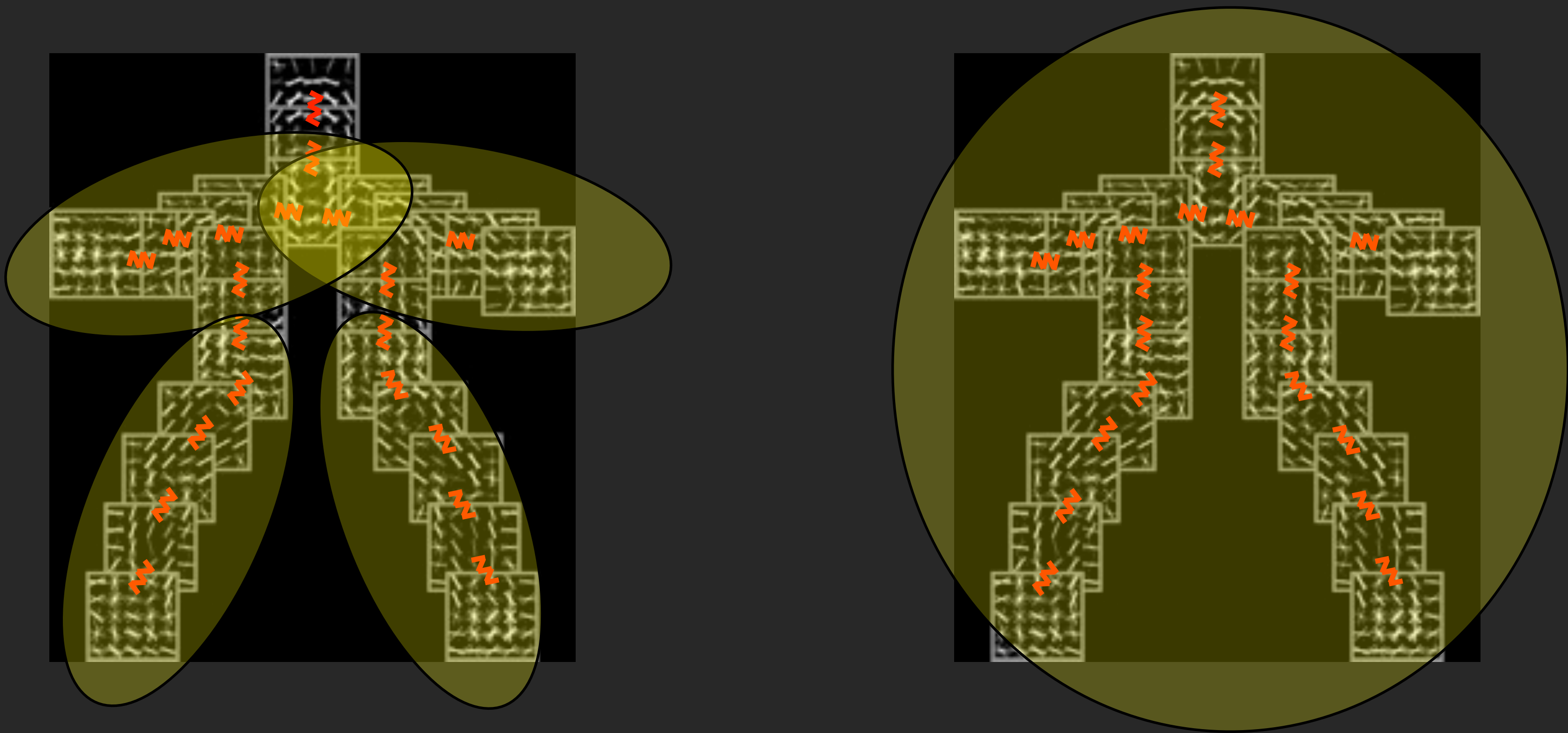
Exponential number of global mixtures



K parts, M local mixtures => K^M unique global mixtures

Not all combinations are equally likely;
“prior” given by co-occurrence model

Semi-global mixtures



Any connected sub-tree of parts can learn to behave like a rigid mixture

Local mixtures can represent (semi) global mixtures

cf. Sapp & Taskar, CVPR13

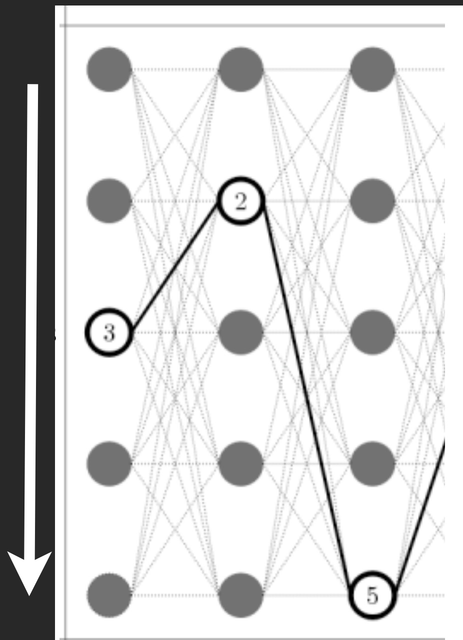
Inference

Consider "joint" domain of part location and mixture type: $z_i = (p_i, t_i)$

$$S(z) = \sum_{i} \phi_i(z_i) + \sum_{ij \in E} \psi_{ij}(z_i, z_j)$$

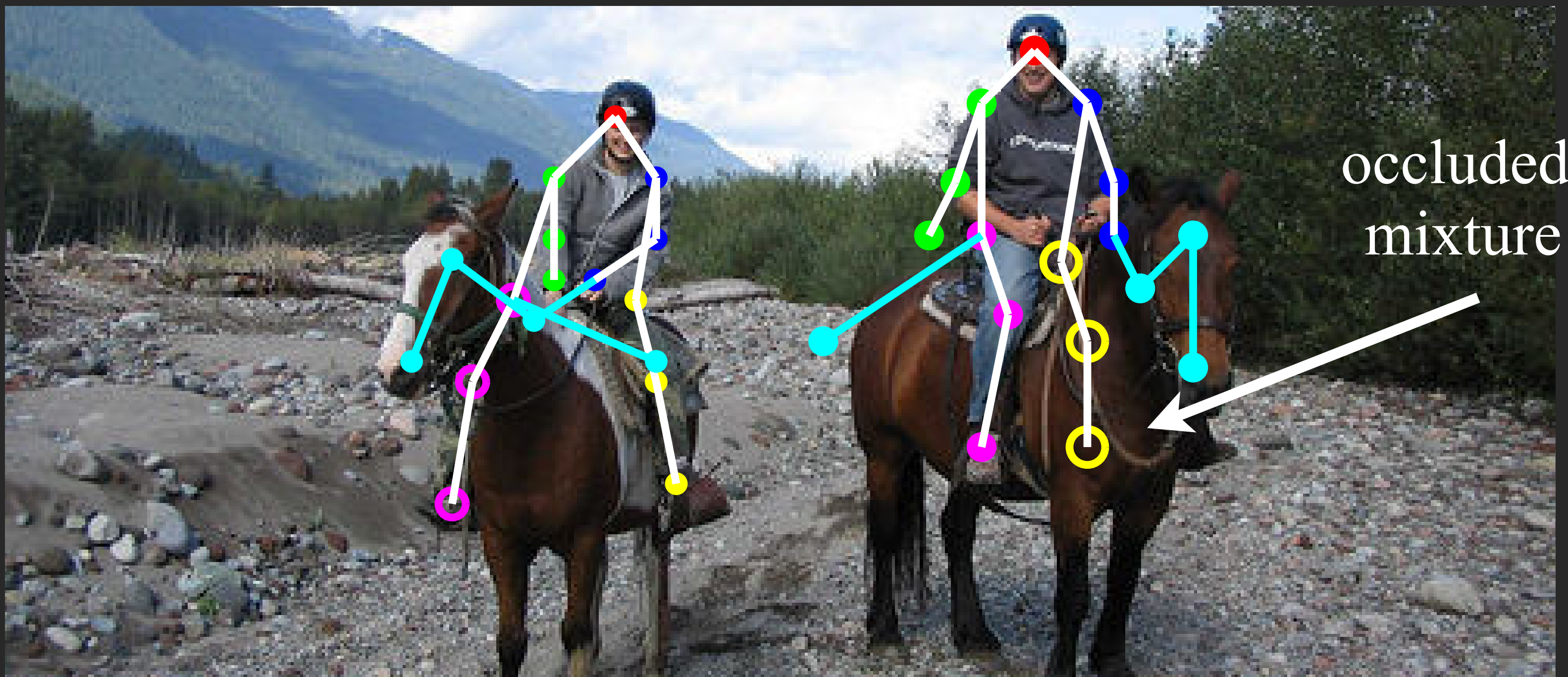
(simple discrete tree-MRF)

Pixel locations
and mixture types



head torso leg

Inference & Learning



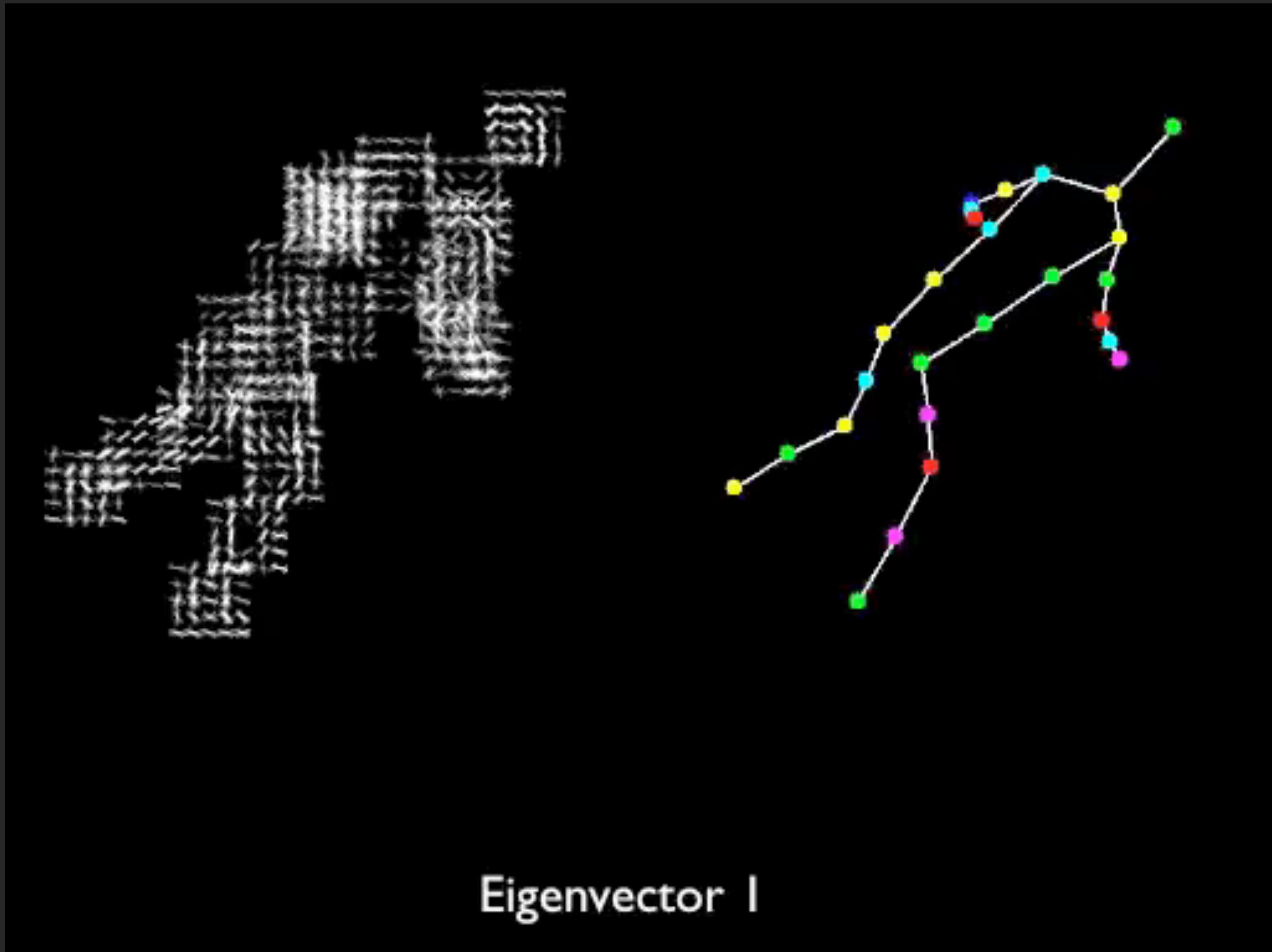
Inference: Infer part locations + mixtures with dynamic programming on trees

Learning: Tune linear parameters (including occlusion constraints) with SVM solver

Application: pose estimation

Yi & Ramanan CVPR11

Orientation mixtures

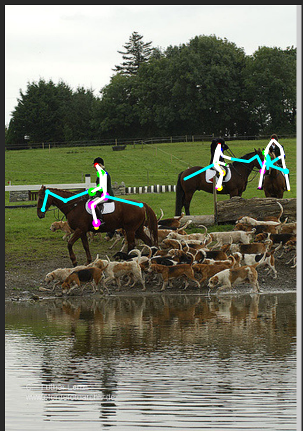


Part appearance (local mixture, denoted by color) depends on the location and appearance of other parts

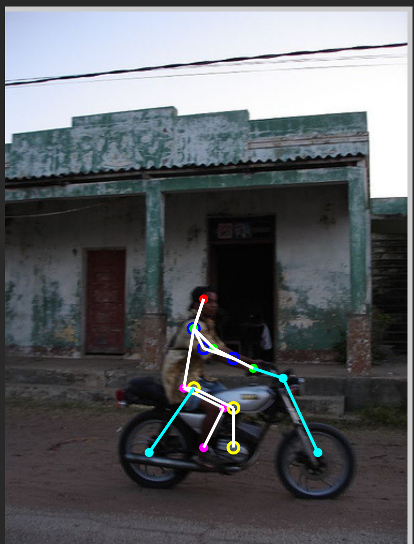
Application: human-object interactions

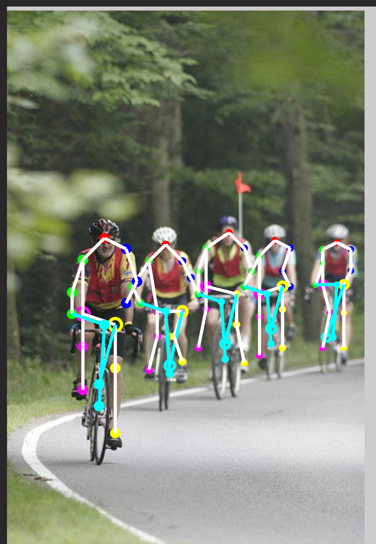
Desai & Ramanan ECCV12

Riding horse



Riding bike

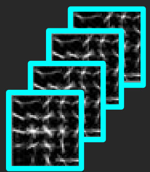




Riding Horse

Possible Criticisms

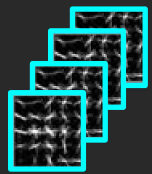
1. One should not score image evidence during occlusions



Possible Criticisms

1. One should not score image evidence during occlusions

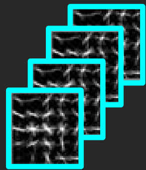
An occluded mixture template may learn all 0 weights; let the learning algorithm decide!



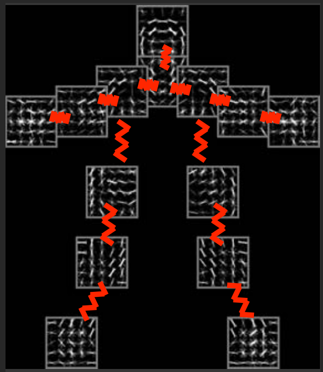
Possible Criticisms

1. One should not score image evidence during occlusions

   An occluded mixture template may learn all 0 weights; let the learning algorithm decide!



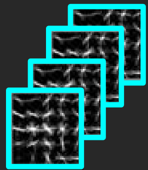
2. Small patches are not as discriminative as larger templates (visual phrases / poselets)



Possible Criticisms

1. One should not score image evidence during occlusions

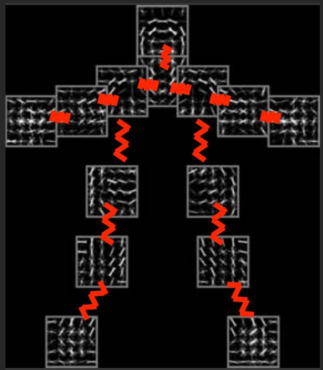
An occluded mixture template may learn all 0 weights; let the learning algorithm decide!



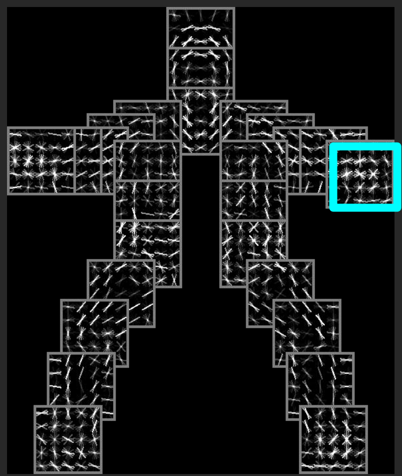
2. Small patches are not as discriminative as larger templates (visual phrases / poselets)

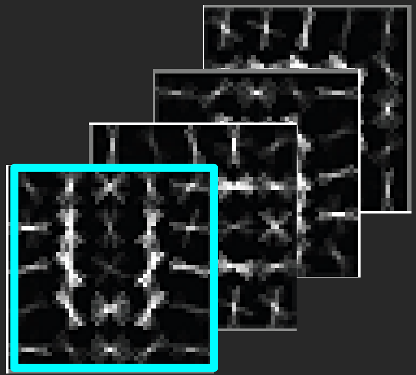
Any connected set of phraselets can learn to behave like a larger template (rigid springs)

"the whole is equal to the sum of its parts"!



Must we train parts jointly?

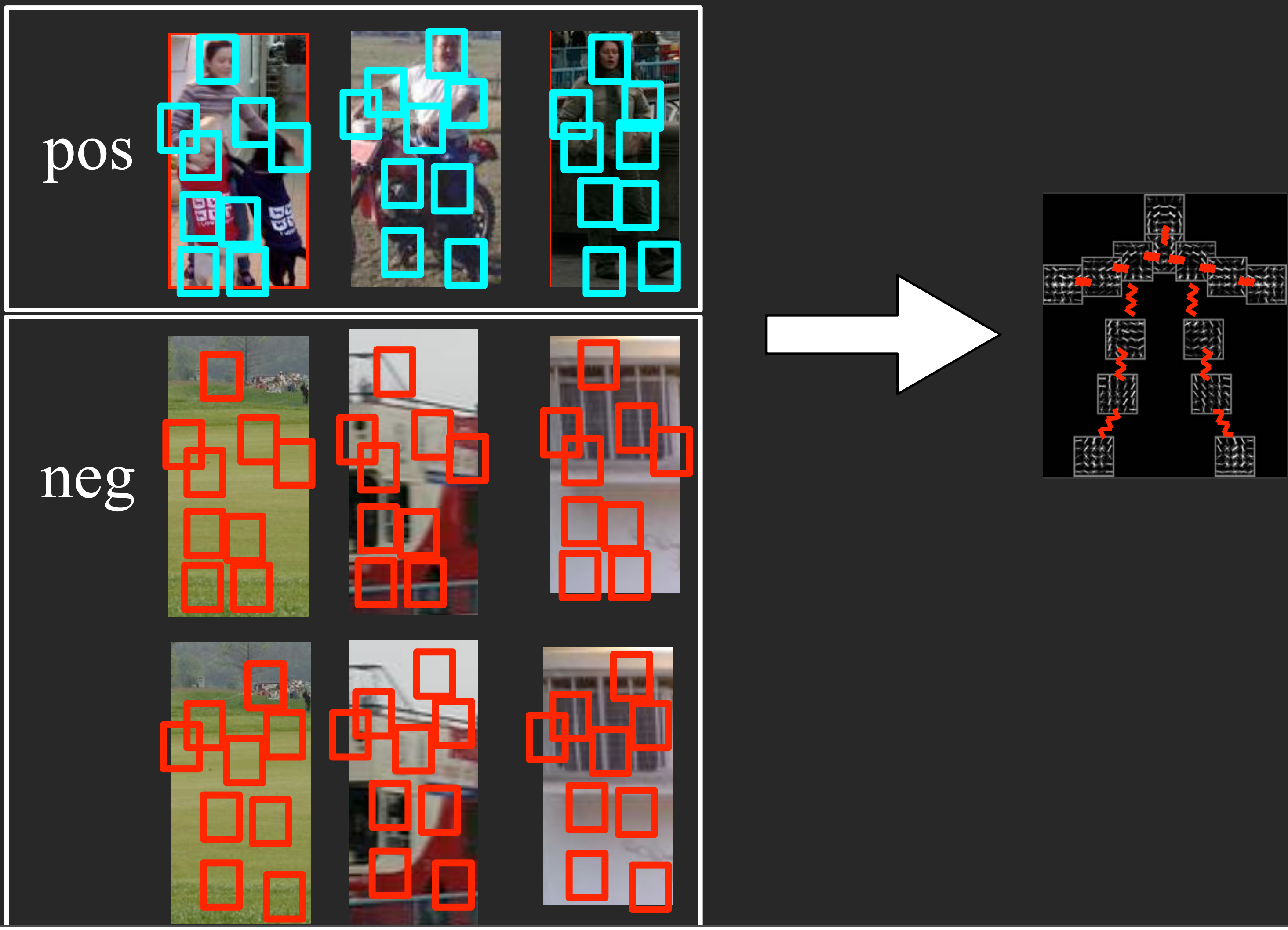




Joint	Indep
67.4	51.3

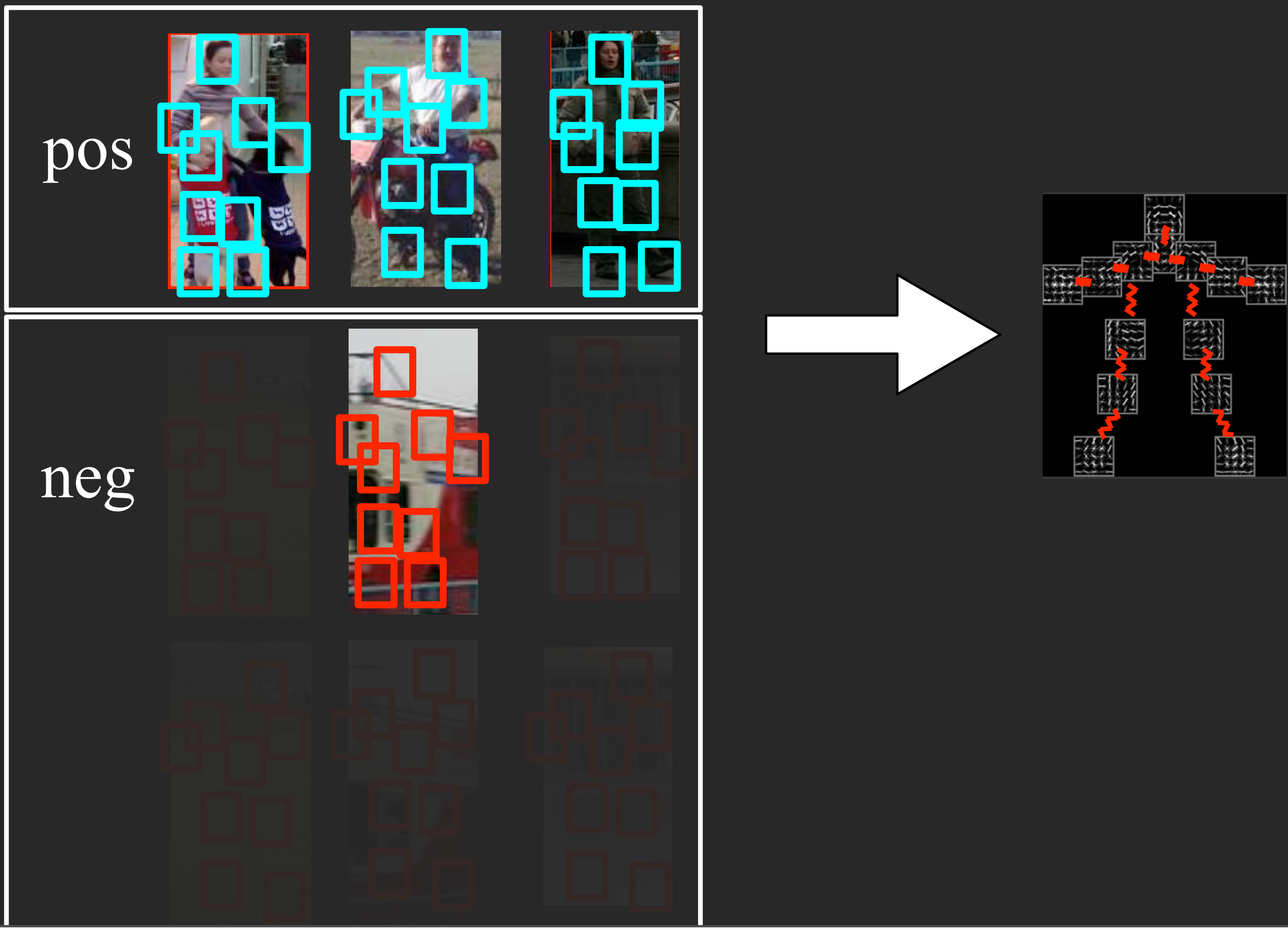
Why does joint training help?

Contextual learning: we need compete only against joint configurations of negatives that score above margin



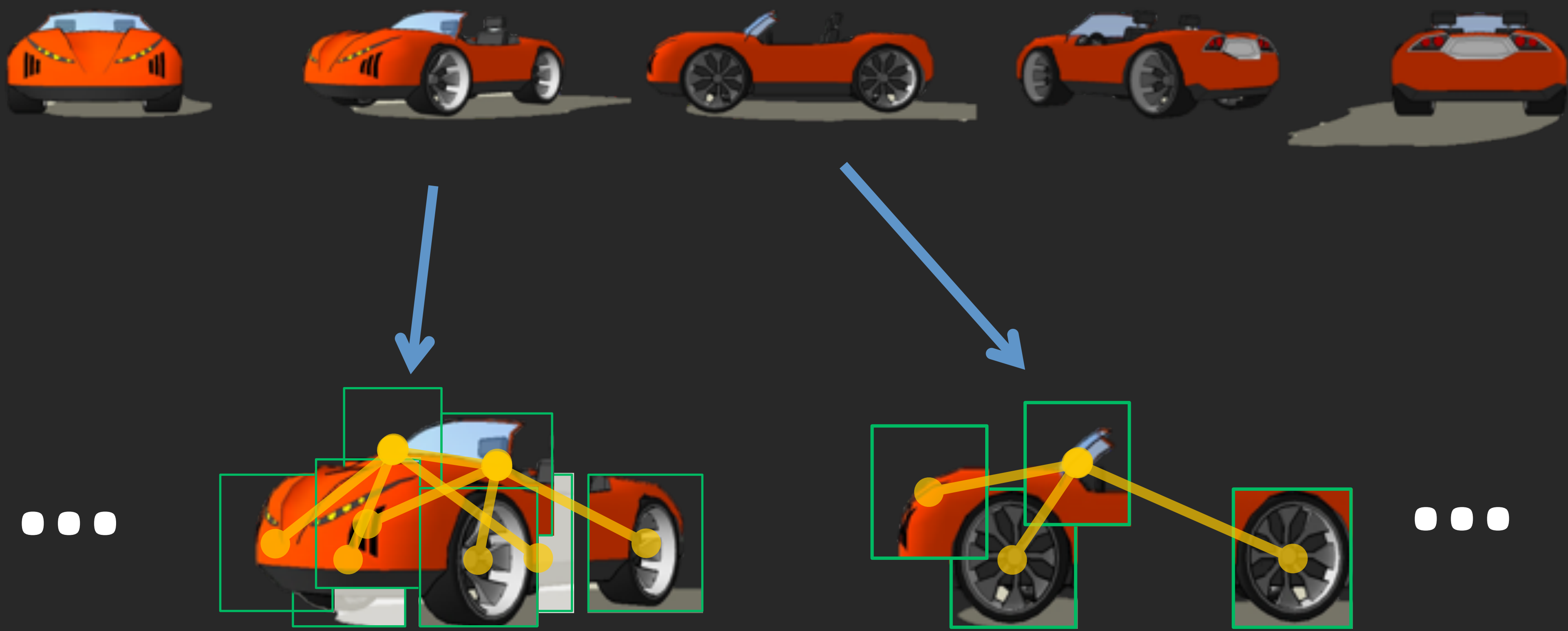
Why does joint training help?

Contextual learning: we need compete only against joint configurations of negatives that score above margin

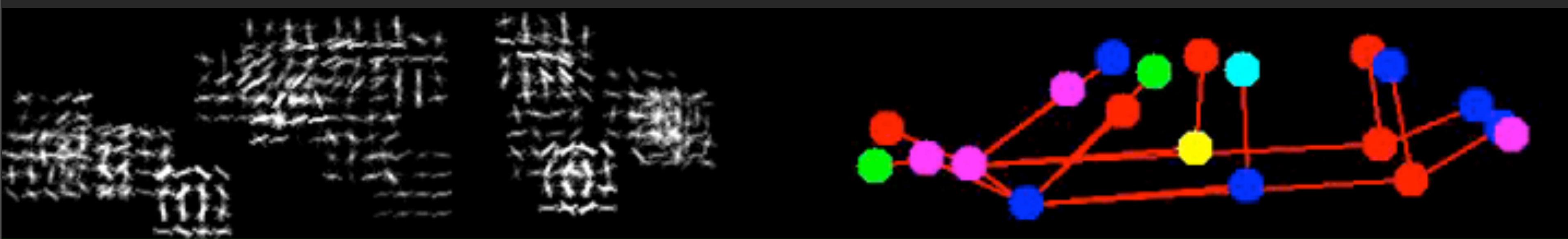


Case study 1: view-based models

Case study 1: view-based models

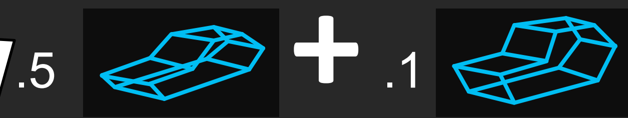


View-based local mixtures



Inferring 3D shape

Hejrati & Ramanan *NIPS* 12



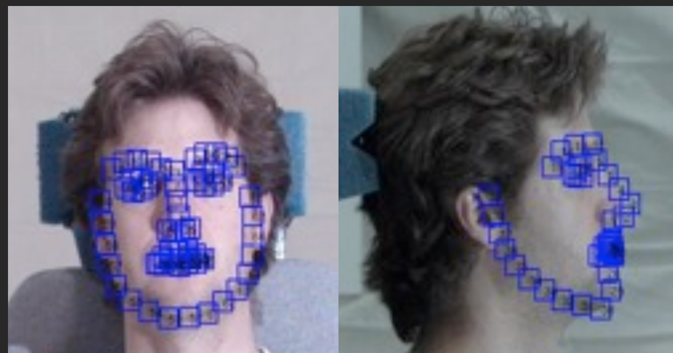
3D shape basis
(nonrigid SFM)

Results

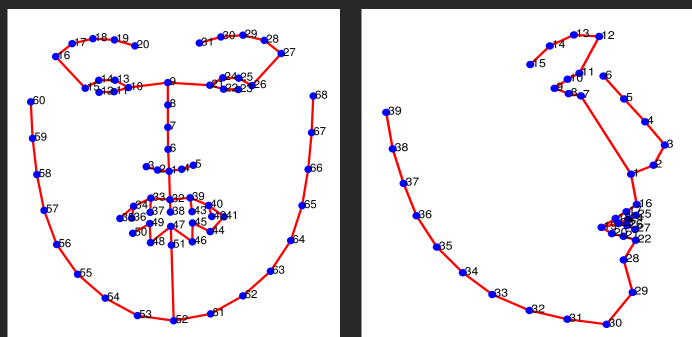
Hejrati & Ramanan NIPS 12

Case study 2: face detection (in-the-wild)

Learning

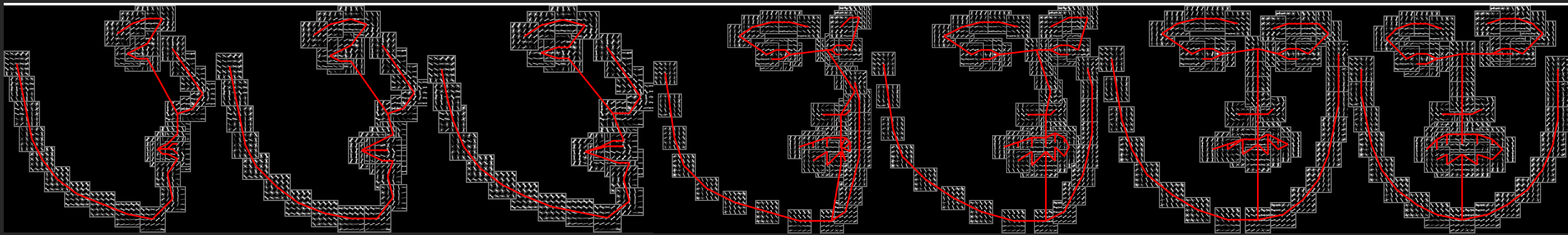


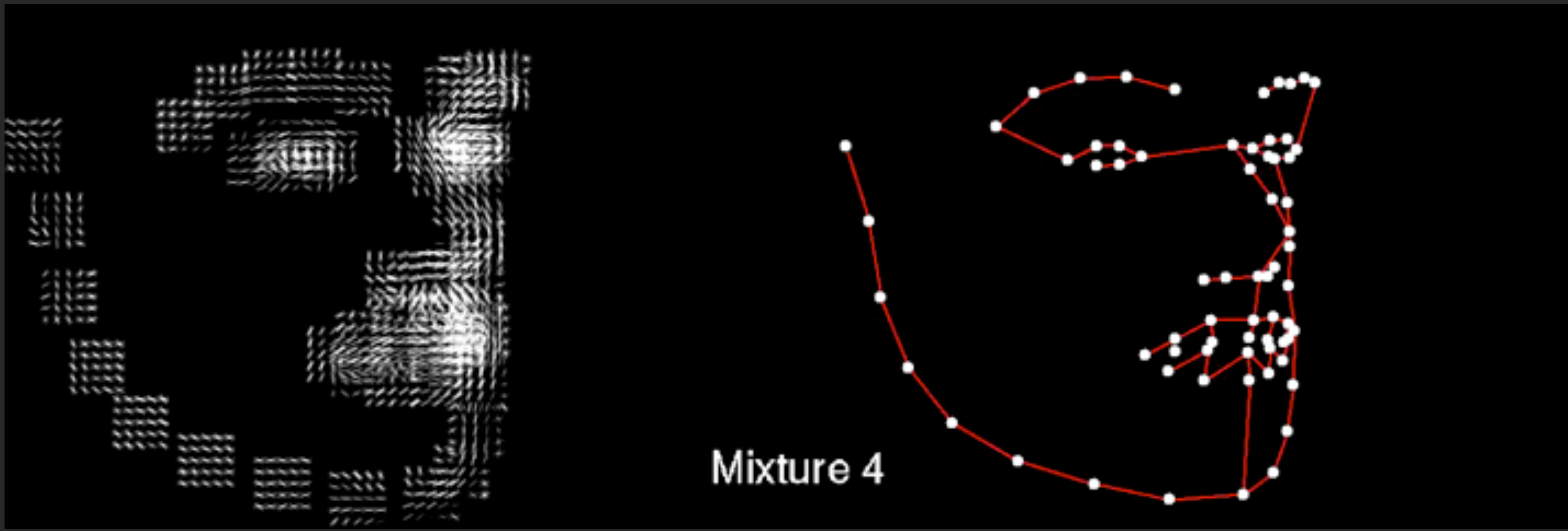
Fully-supervised dataset (CMU MultiPIE)



Chow-Liu algorithm

View-based model

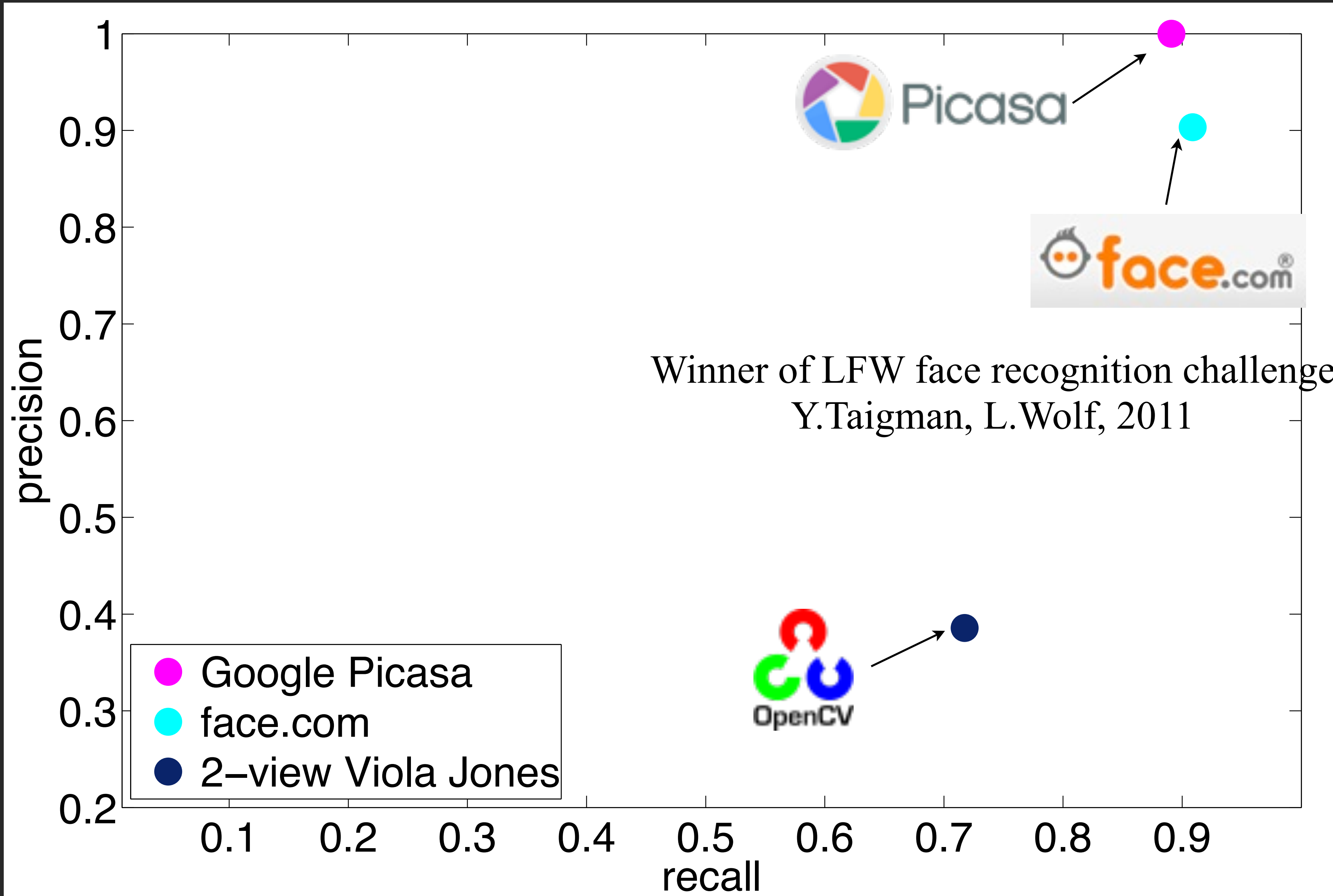




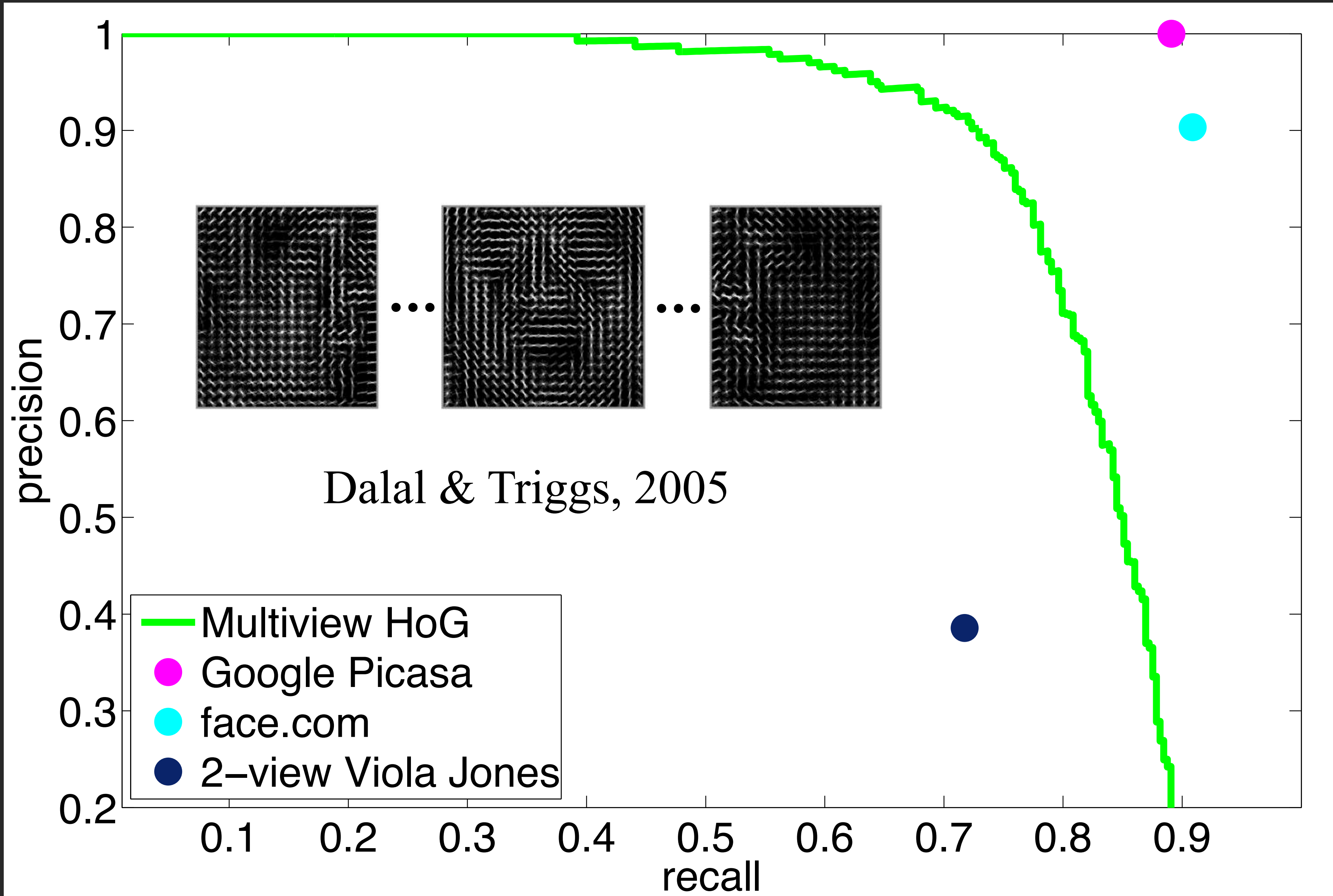
Qualitative Results

Zhu & Ramanan CVPR 12

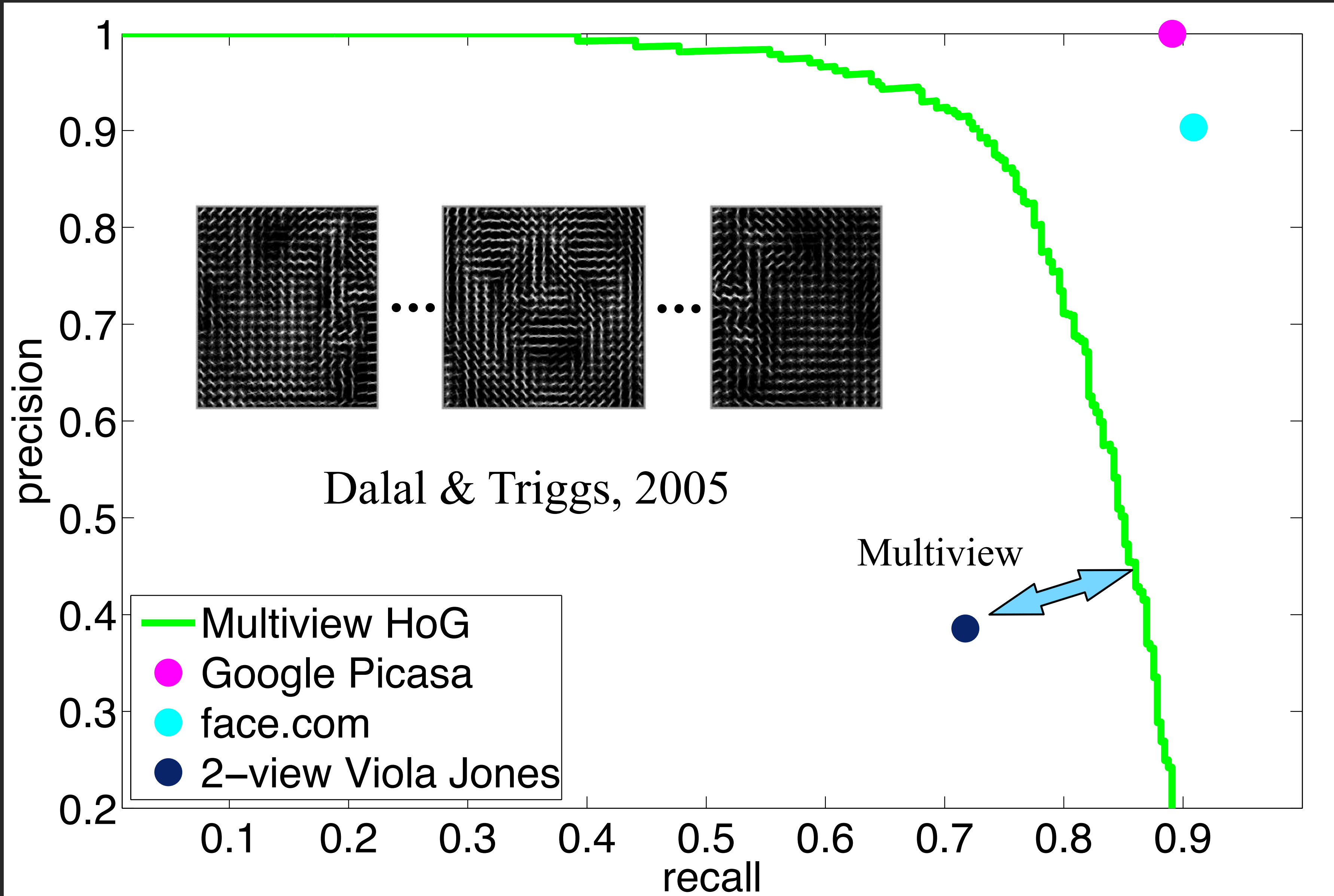
Detection results



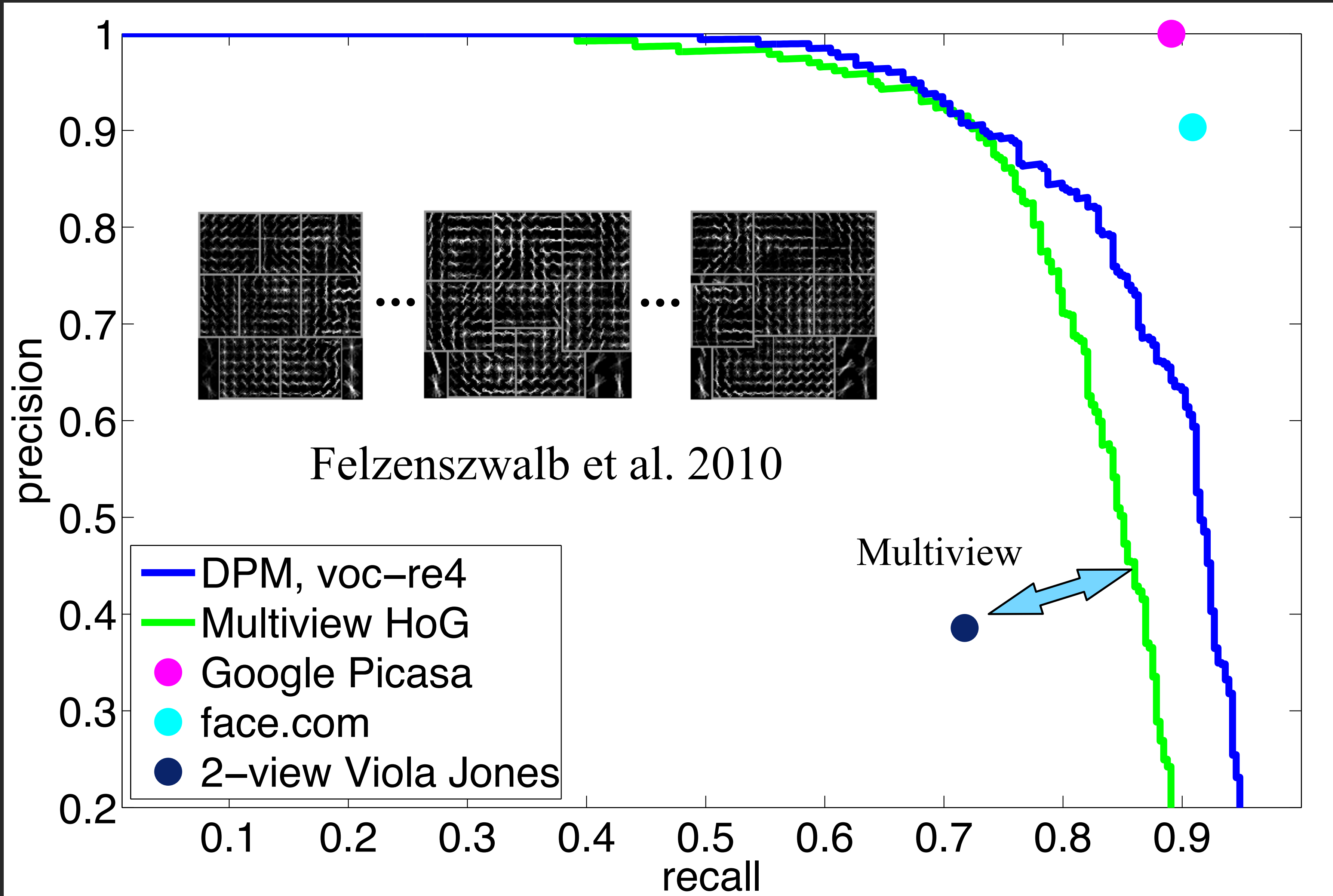
Detection results



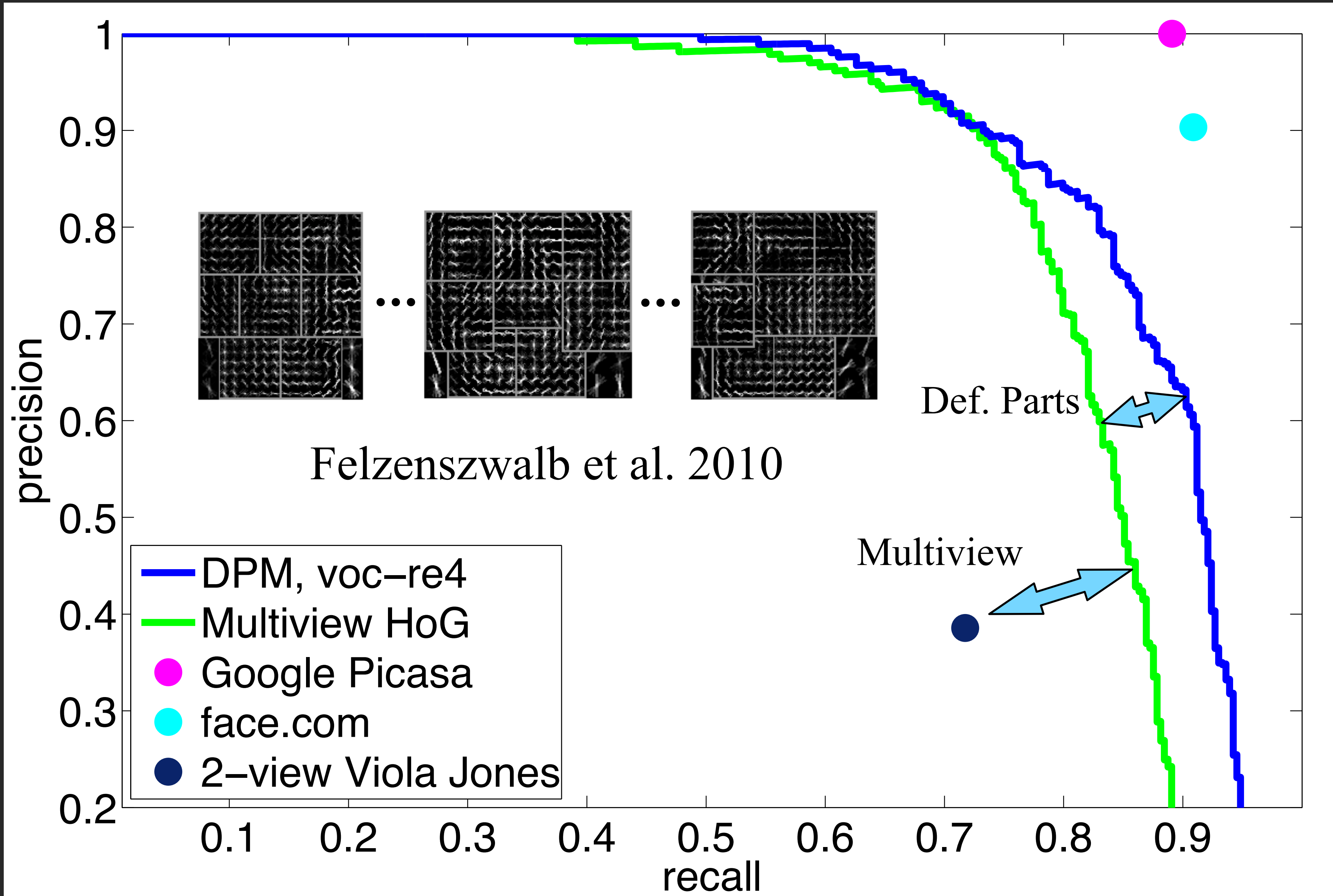
Detection results



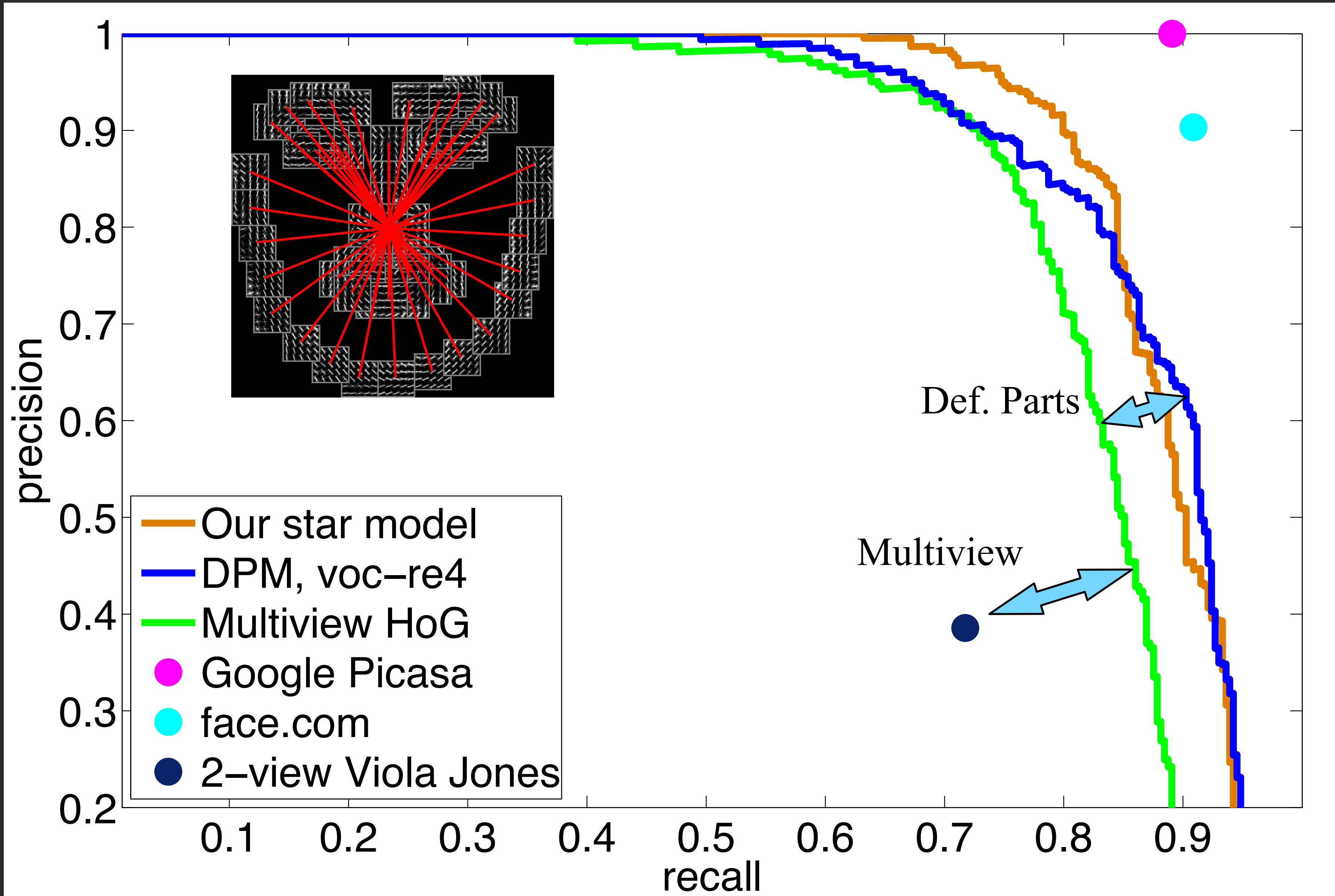
Detection results



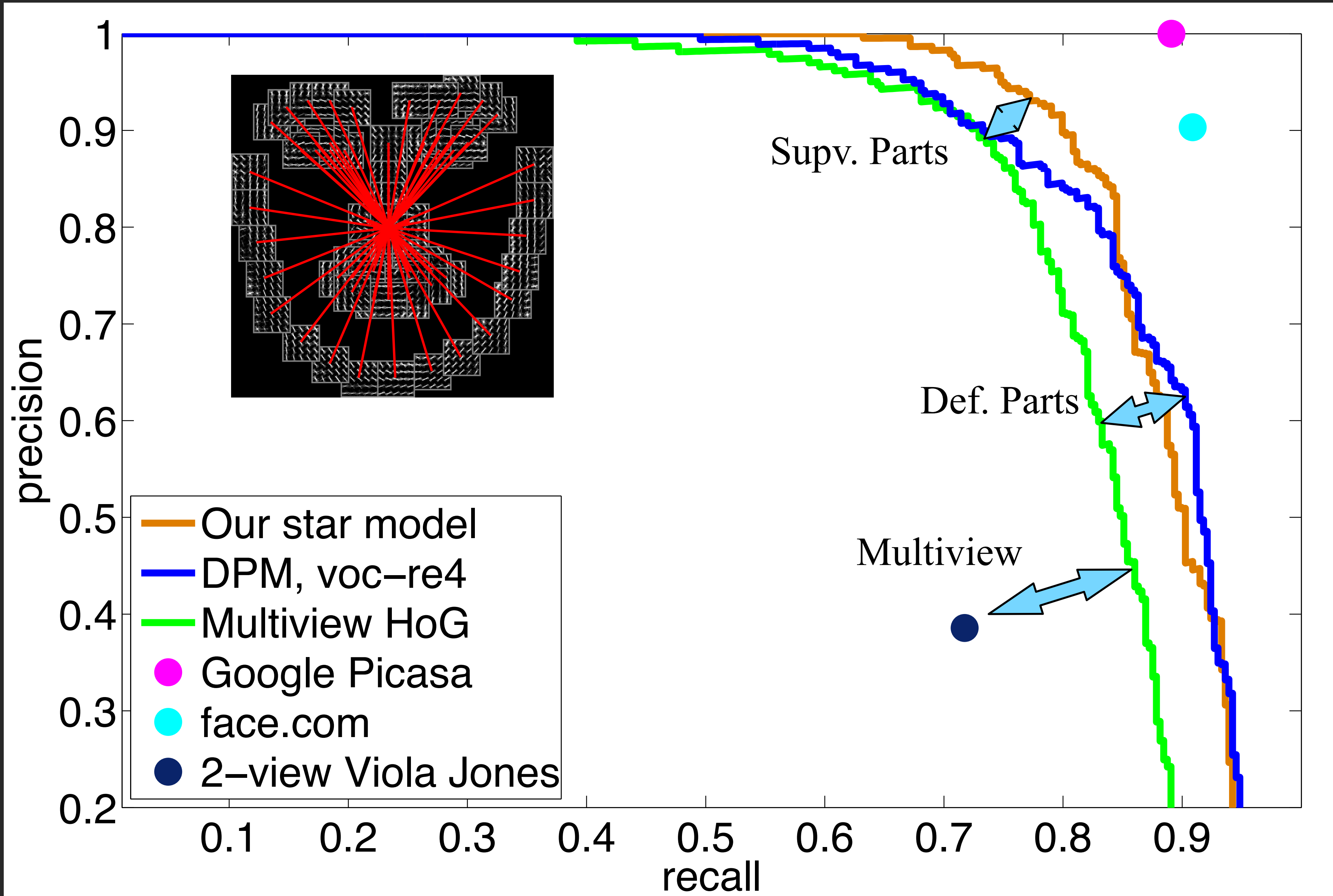
Detection results



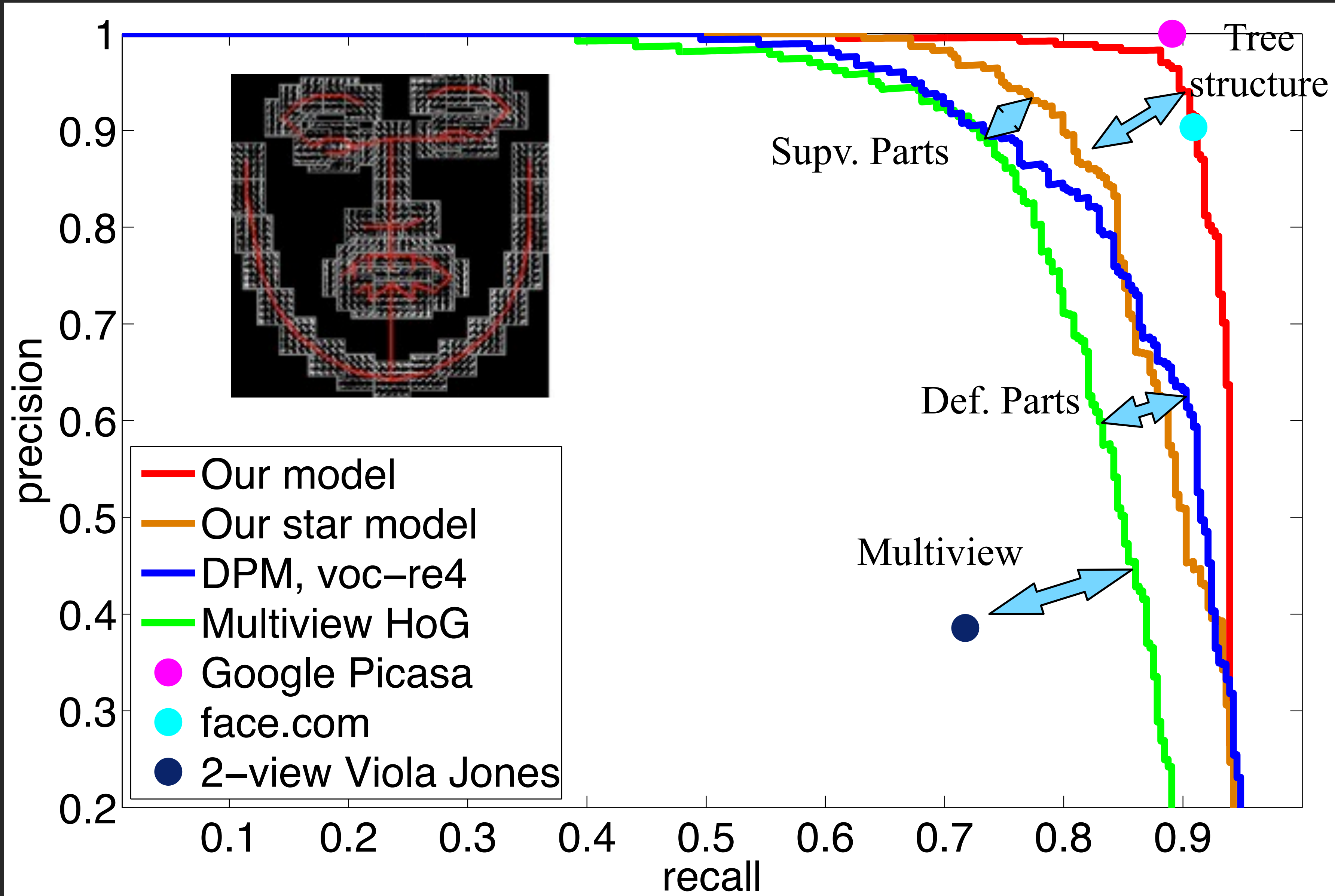
Detection results



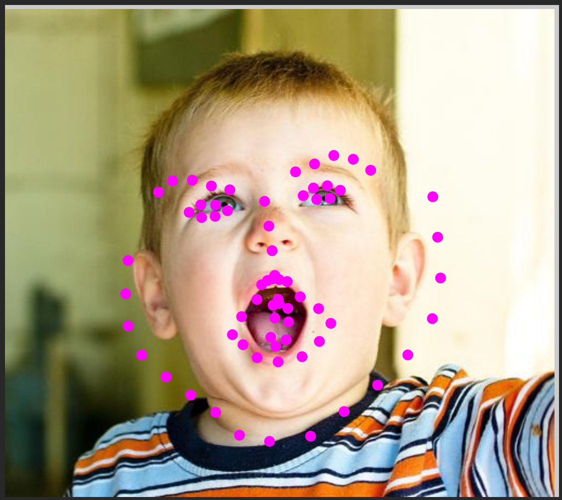
Detection results

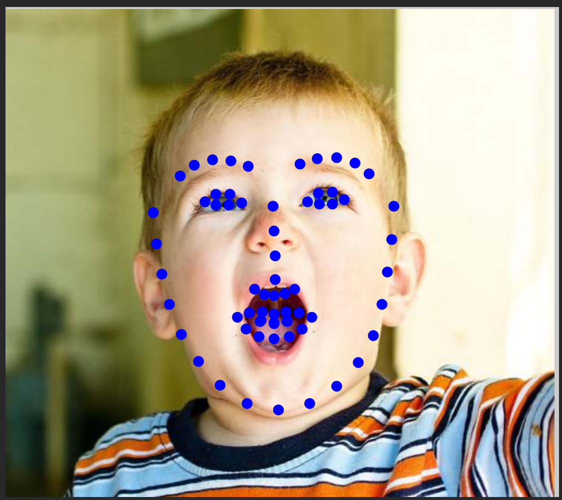


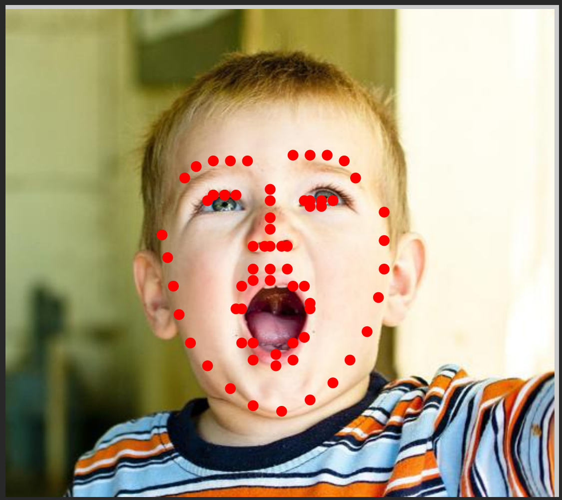
Detection results



Wild expression

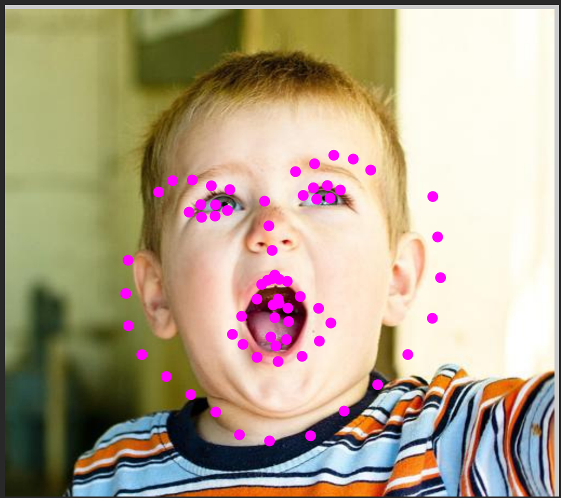




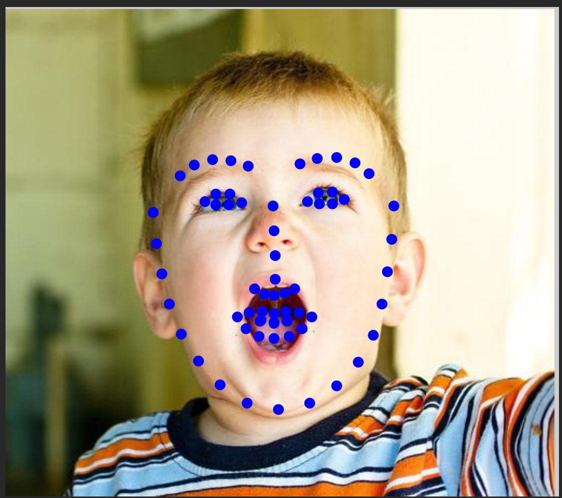


AAM

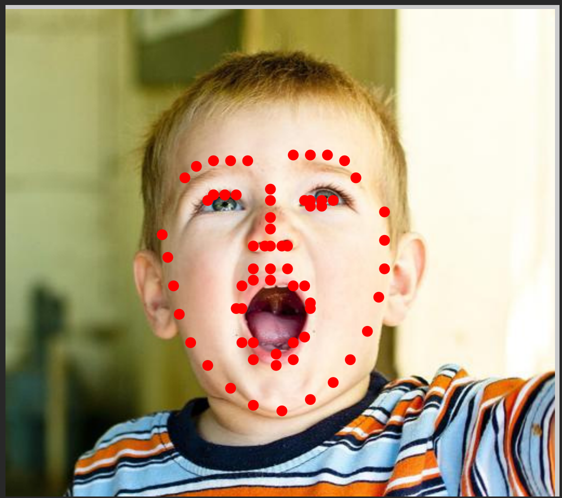
Wild expression



AAM

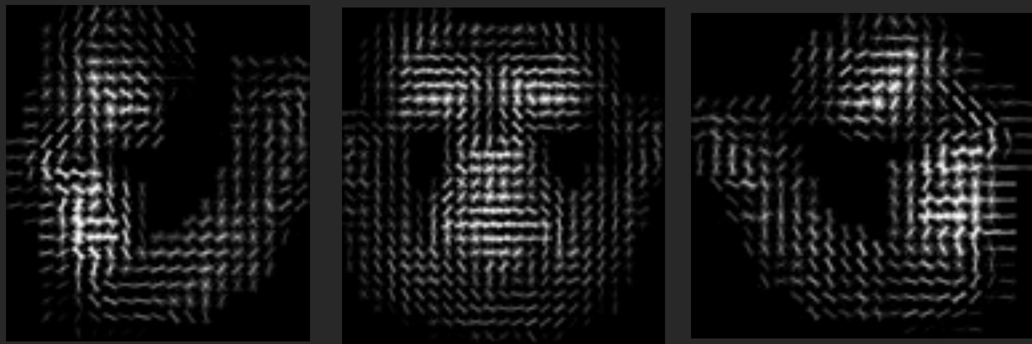


CLM



Our Model

DPMs vs explicit mixtures



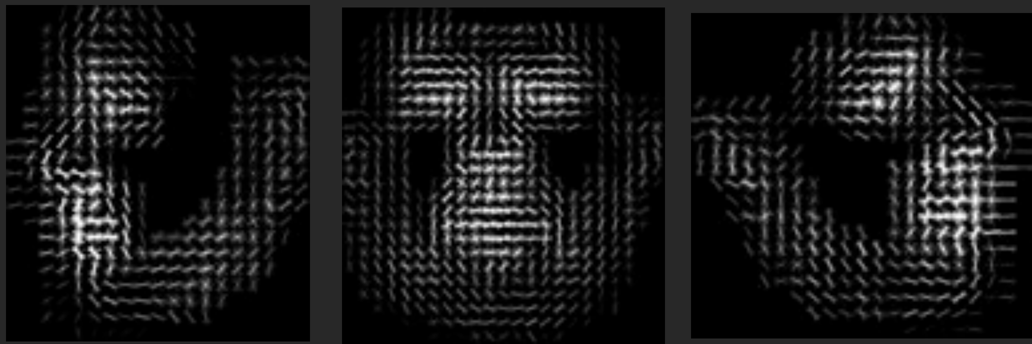
Mixtures of rigid templates

Part model

“Exemplar SVMs”

Malisiewicz et al ICCV 11

DPMs vs explicit mixtures



Mixtures of rigid templates

Part model

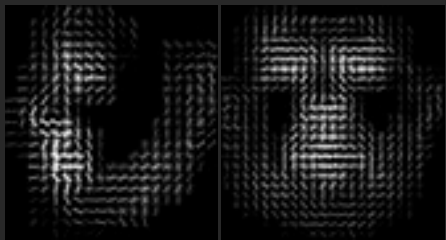
“Exemplar SVMs”

Malisiewicz et al ICCV 11

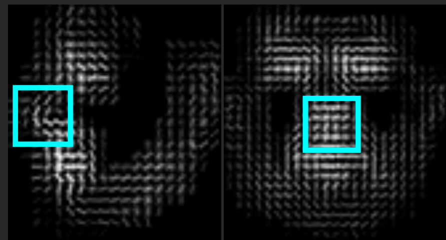
Compared to a mixture of exemplars, part models...

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

DPMs vs explicit mixtures



Mixtures of rigid templates



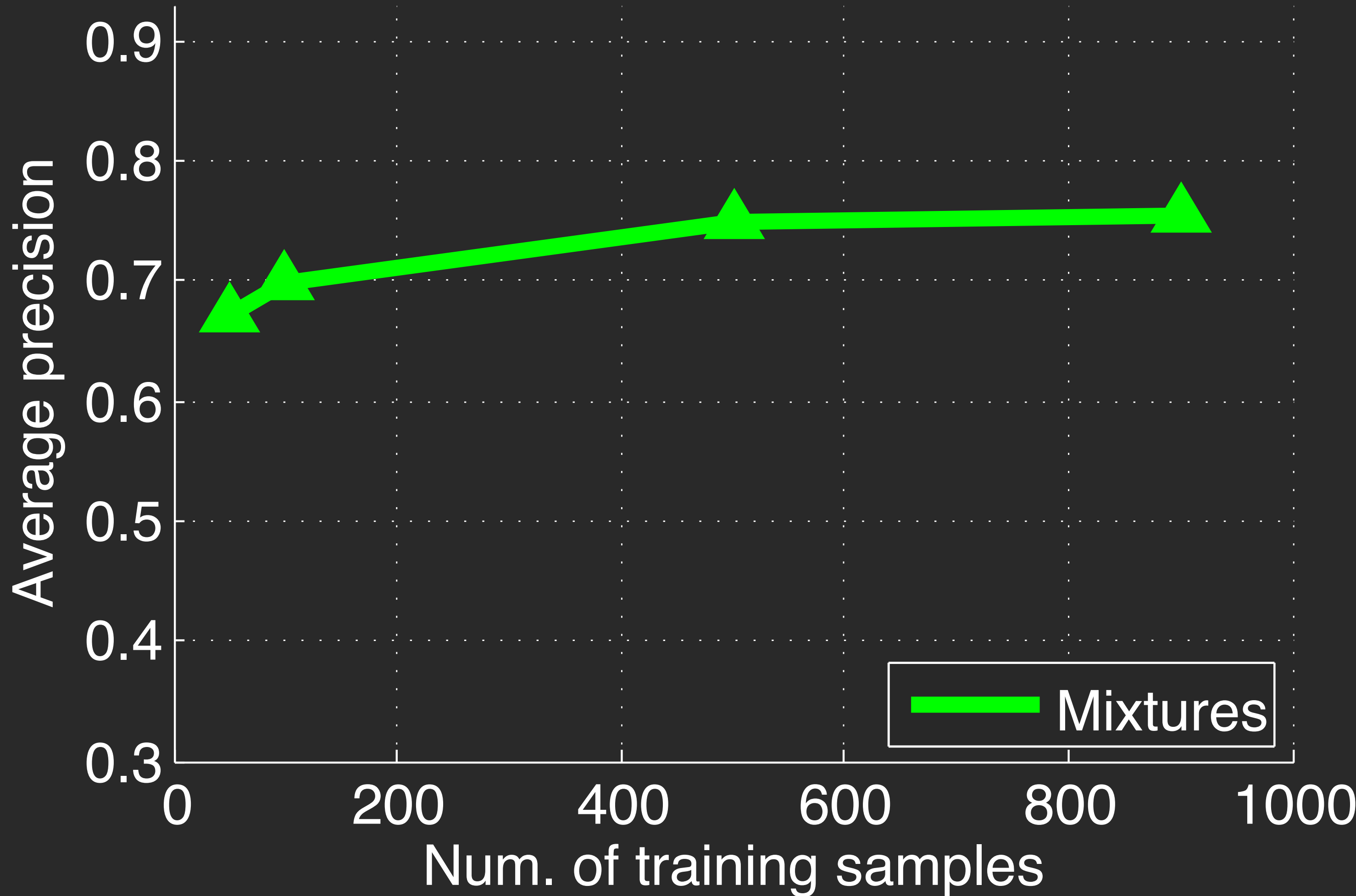
Mixtures of rigid templates with tied parameters (given by parts)

Part model

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

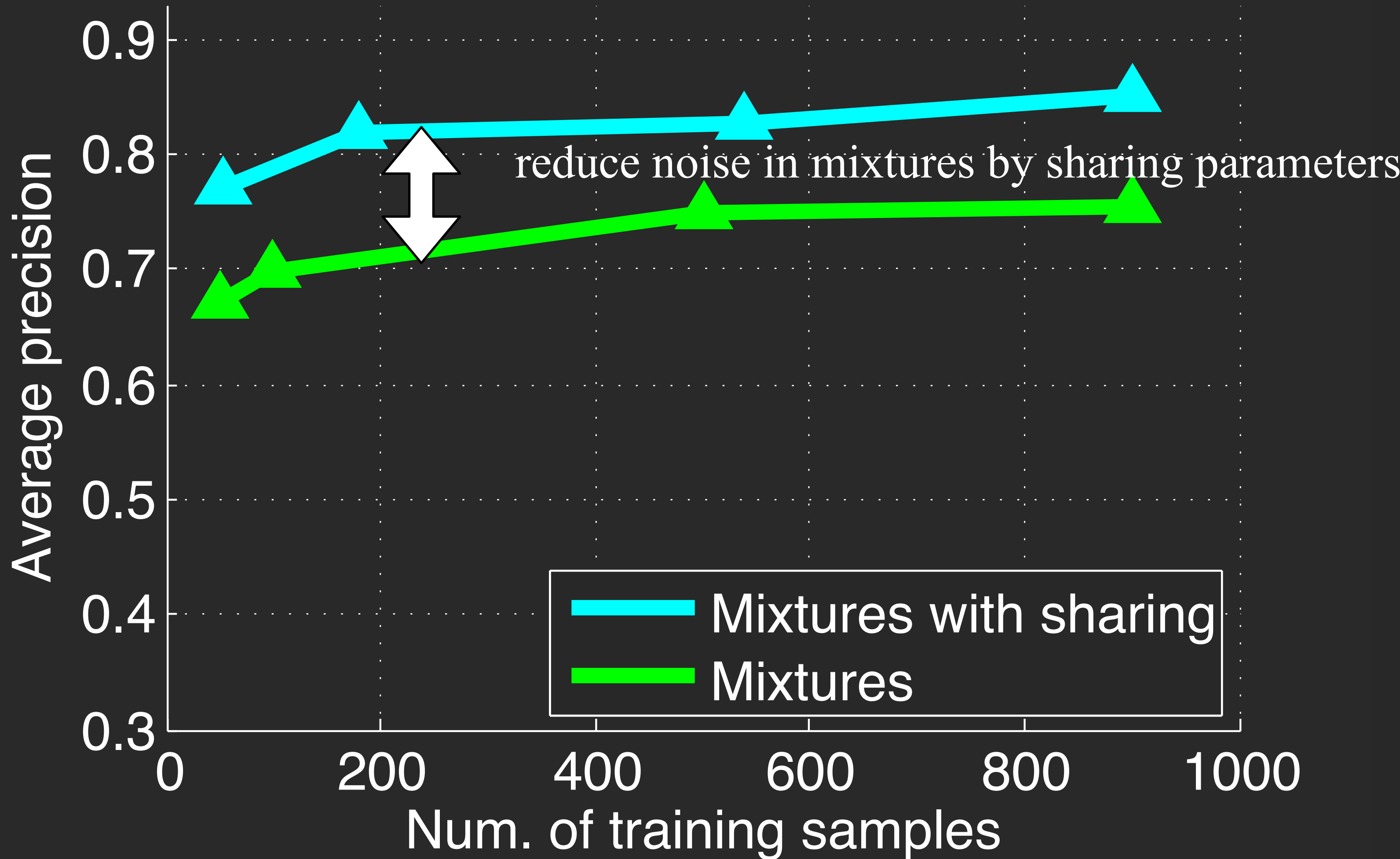
To examine (1) vs (2), lets define mixture of exemplars with sharing

An analysis of part models



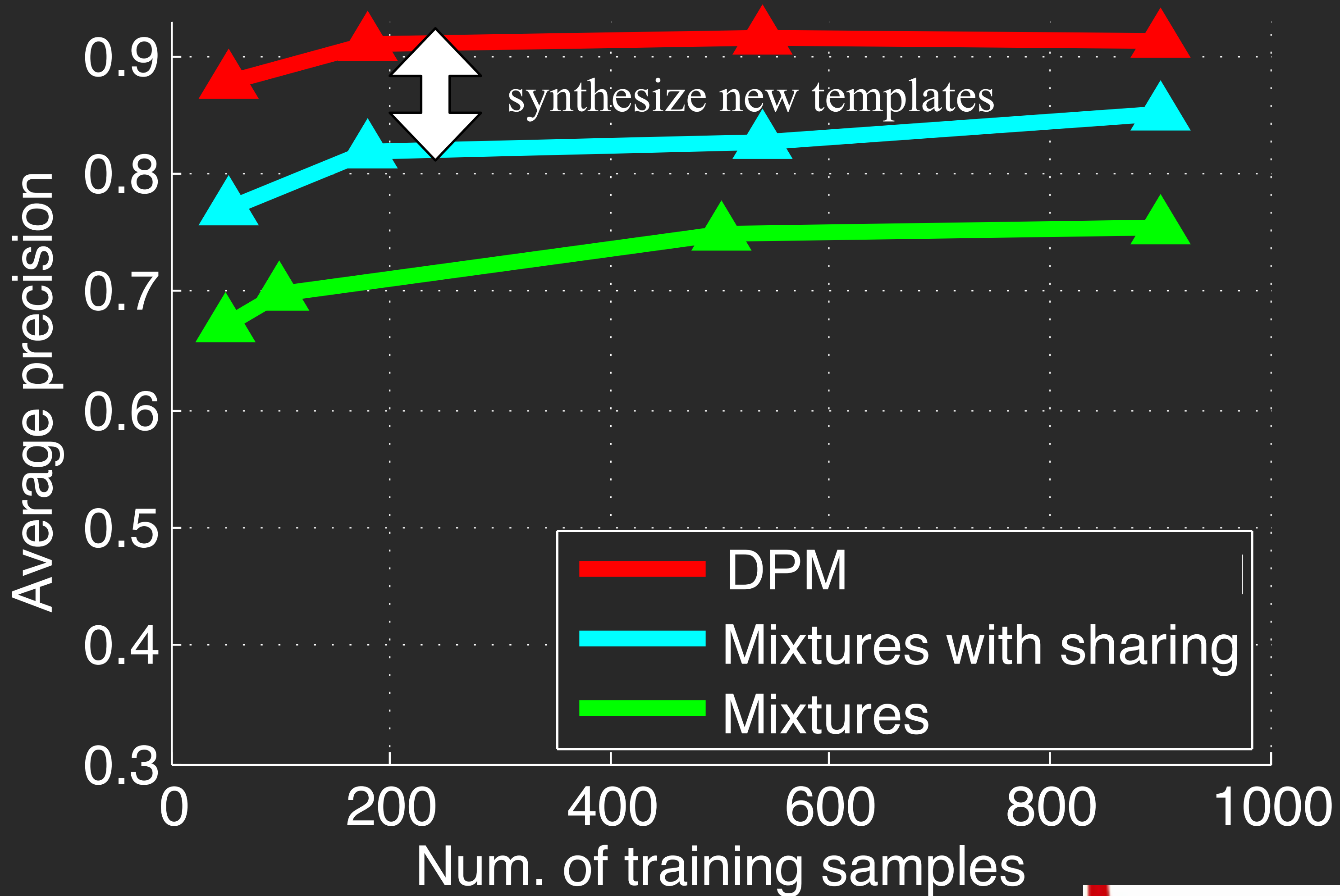
Zhu, Vondrick, Ramanan & Fowlkes,
“Do we need more training data or better models?”
BMVC 2012

An analysis of part models

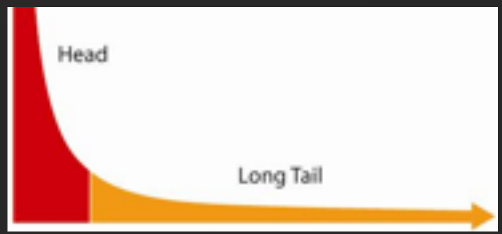


Zhu, Vondrick, Ramanan & Fowlkes,
"Do we need more training data or better models?"
BMVC 2012

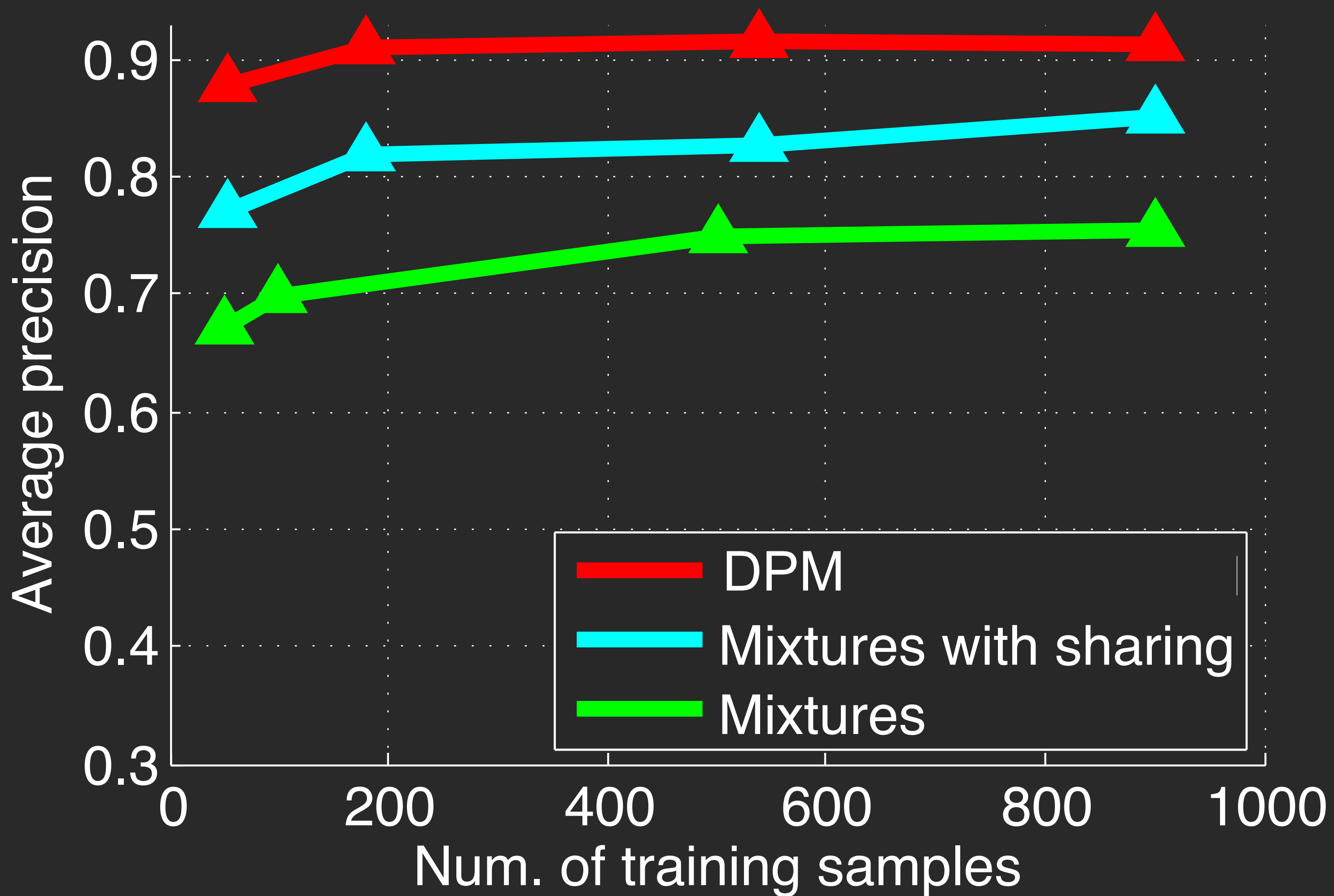
An analysis of part models



“Synthesis” of unseen (rare) templates is even more beneficial than sharing

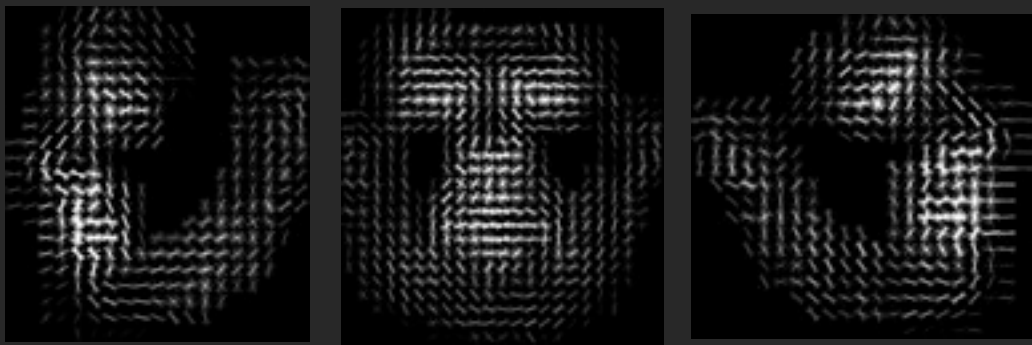


An argument against "big-data"



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

Strategic questions

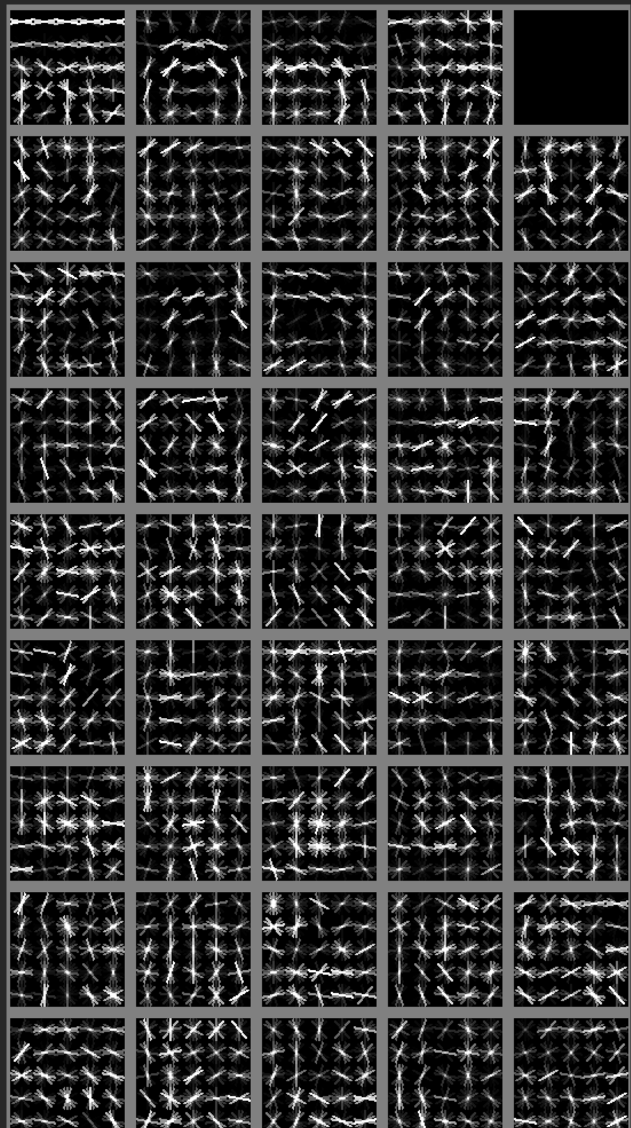


Given a collection of training images / templates, how do we share information between them and generate new unseen templates?

1. "Parts" = local quantized bits of templates
2. Define rules to create new unseen compositions of parts

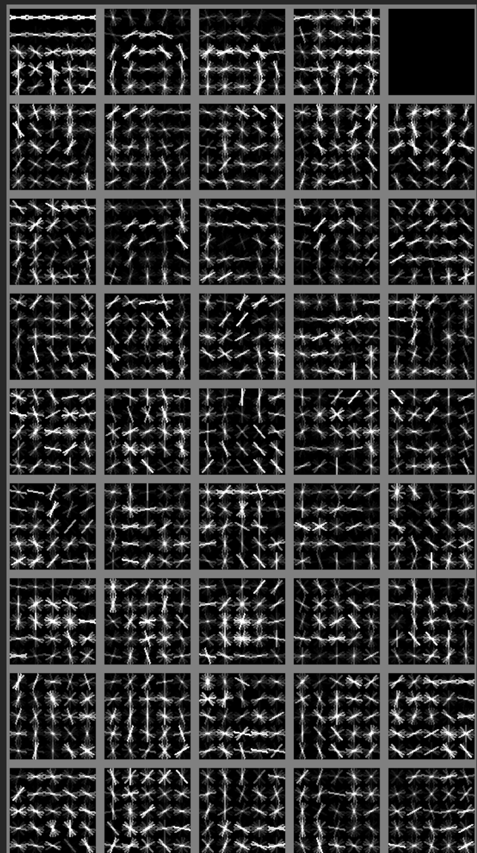
Other approaches...

How can we scale to thousands of parts?



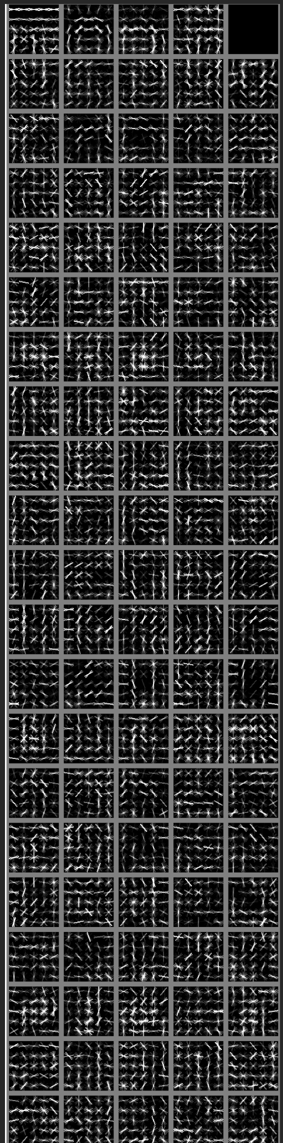
Nearest neighbor indexing (Google team, CVPR13)

How can we scale to thousands of parts?



"Parts" are simply linear filter banks.
Apply tools/representations from image processing

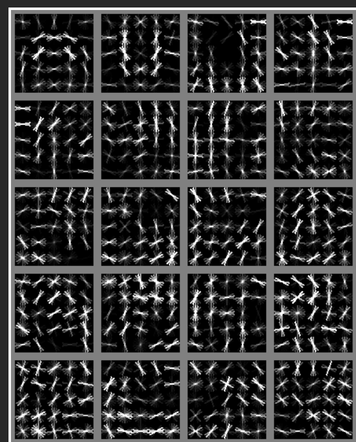
Steerable + separable basis



Freeman, Adelson, Perona

$$w_i = \sum_j s_{ij} b_j$$

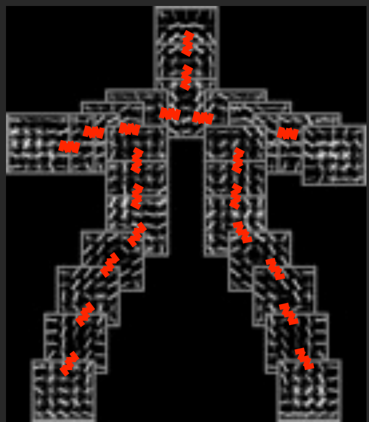
$\approx$ linear combinations of basis templates

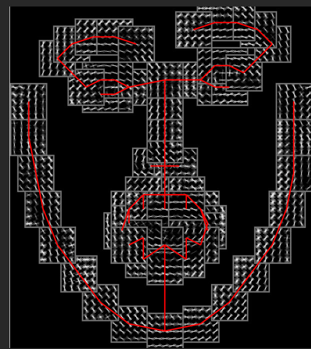


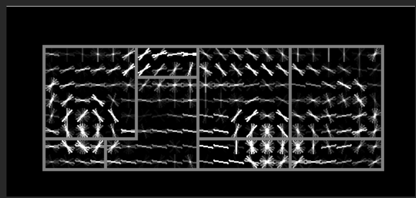
This can be implemented as a rank-restriction on original set of templates

Steerable (& separable) part models

Pirsiavash & Ramanan CVPR12

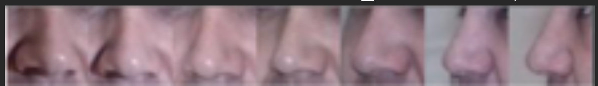






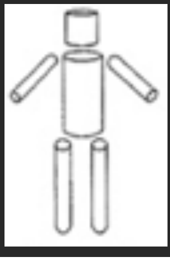
Models are 5-100X smaller & faster with near-equivalent performance

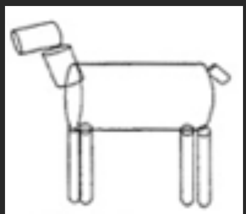
Share “soft” basis rather than fixed templates (across views/categories)

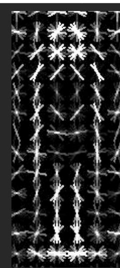


Philosophy: We should treat parameters w as spatial filters, not vectors

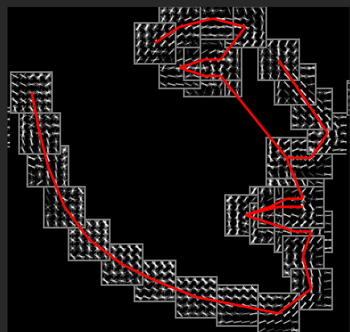
A look back



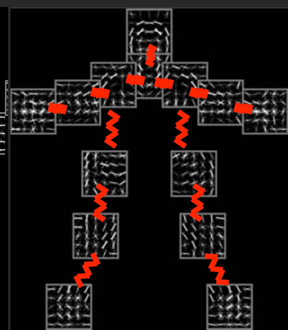




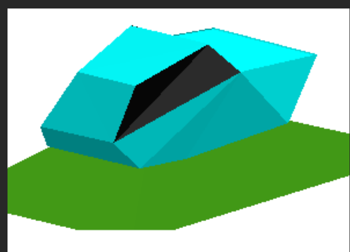
Geometric statistical models

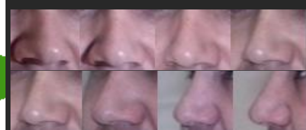


occlusion



articulation





viewpoint

steerability