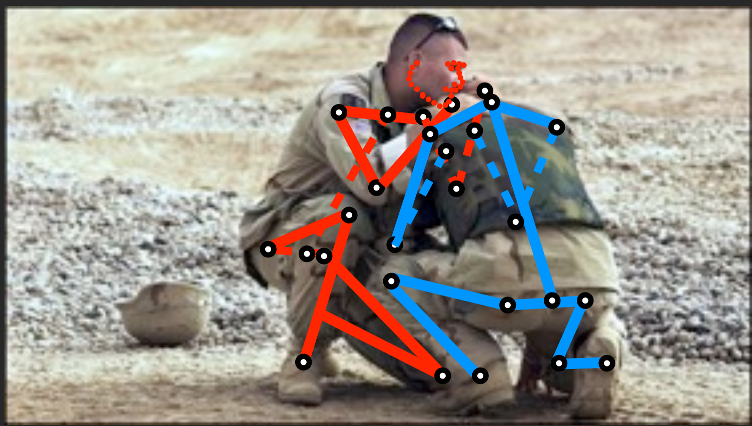
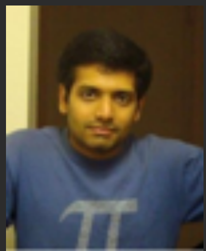
Human pose estimation

Deva Ramanan

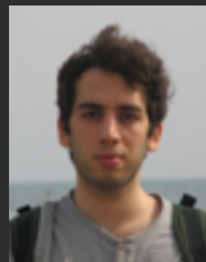
UC Irvine



Host of collaborators



Chaitanya Desai



Mohsen Hejrati

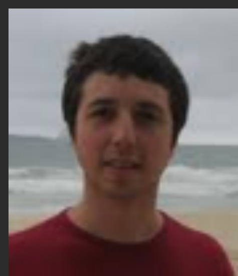
Hamed Pirsiavash

Dennis Park

Yi Yang

Xiangxin Zhu

David MxAllester



Pedro Felzenswzalb

Charless Fowlkes

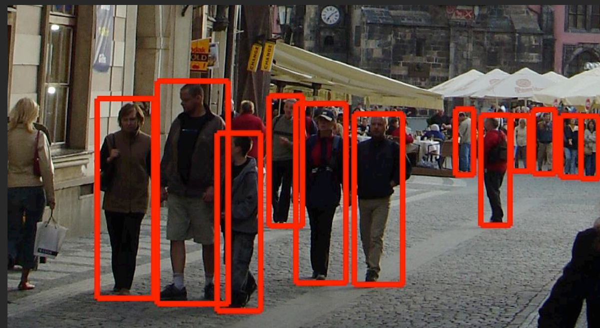
Preface

This talk will contain a mix of "standard" tutorial material and "speculative" opinions

Please interrupt with questions!

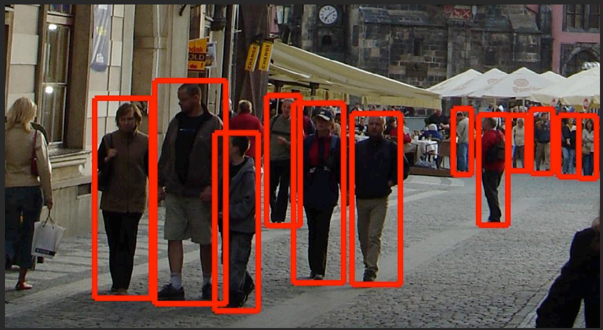
Pattern classification versus visual understanding

Yes/no
scanning window



Pattern classification versus visual understanding

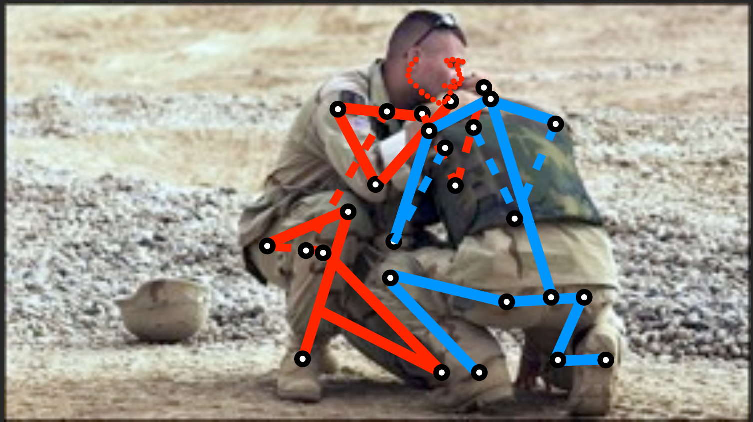
Yes/no
scanning window



vs

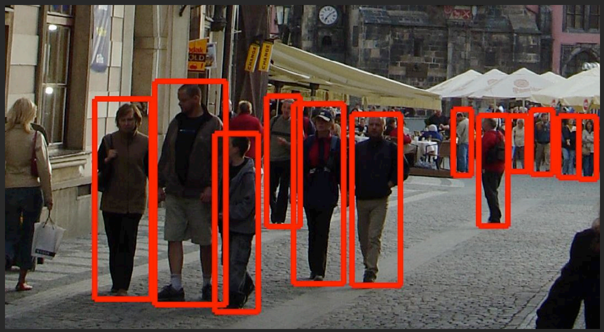
"In-the-wild"
pose estimation

Multiple bodies
Heavy occlusion
3D viewpoint



Pattern classification versus visual understanding

Yes/no
scanning window

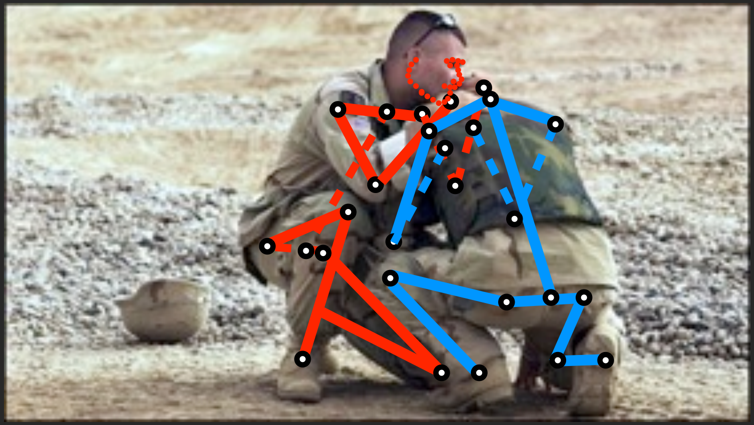


use tools
from here

vs

“In-the-wild”
pose estimation

Multiple bodies
Heavy occlusion
3D viewpoint



Why is finding people difficult?

variation in illumination

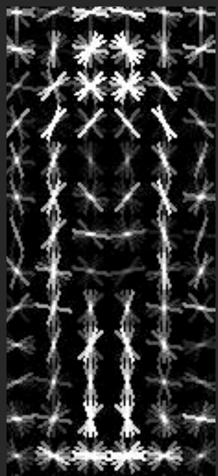
variation in appearance

variation in pose, viewpoint

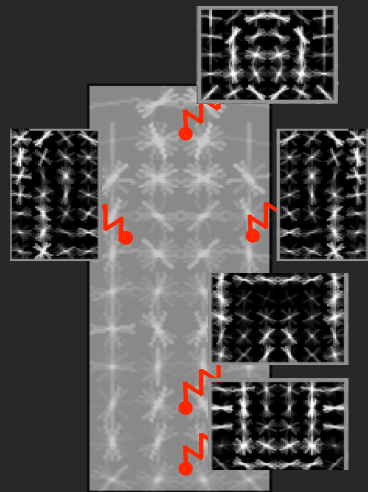
occlusion & clutter

Classic “nuisance factors” for general object recognition

Appearance Templates



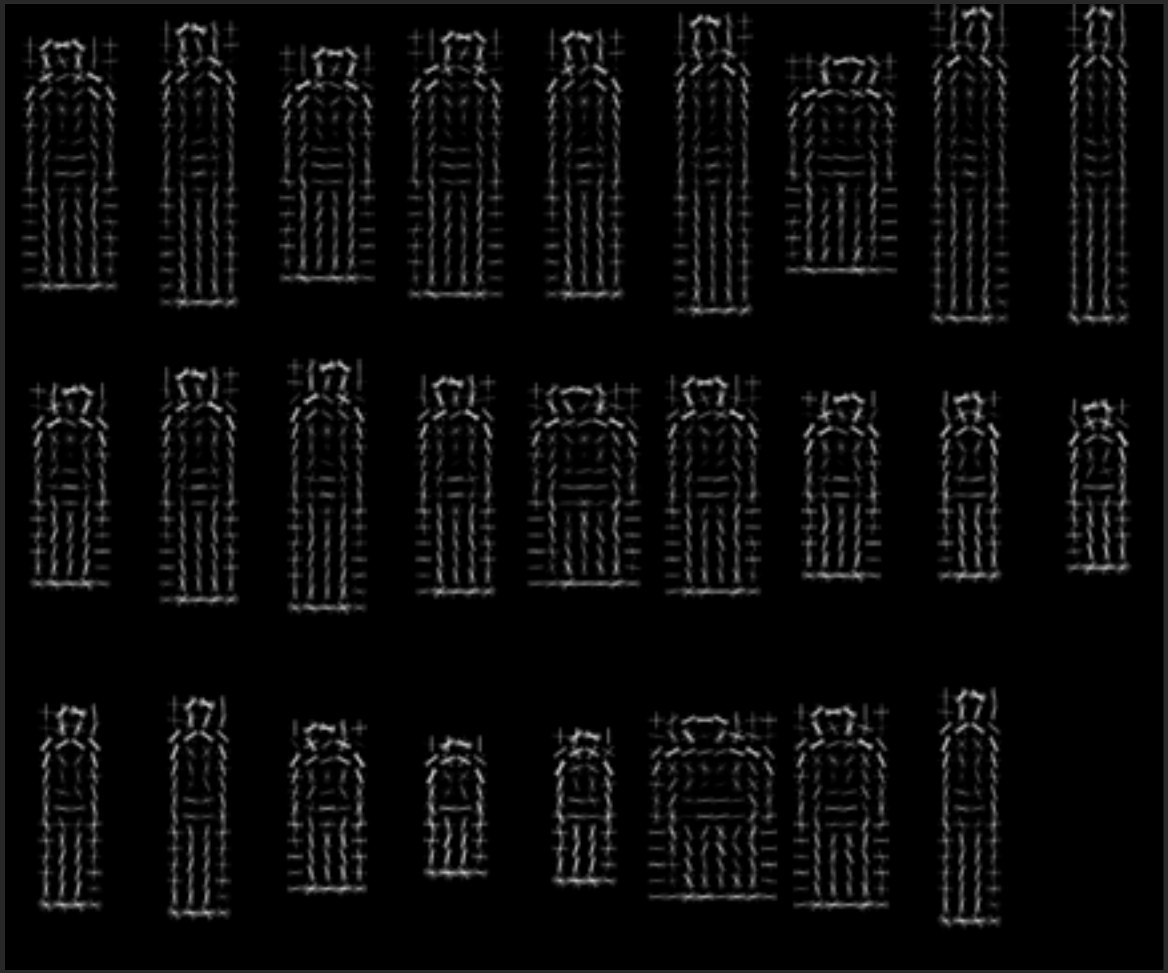
Rigid Templates

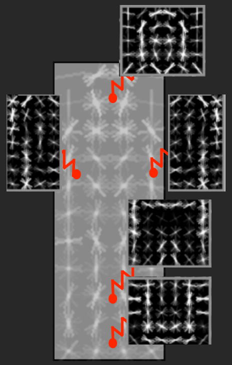


Quasi-rigid templates

Felzenswzalb, Girshick, McAllester, & Ramanan
PAMI 10

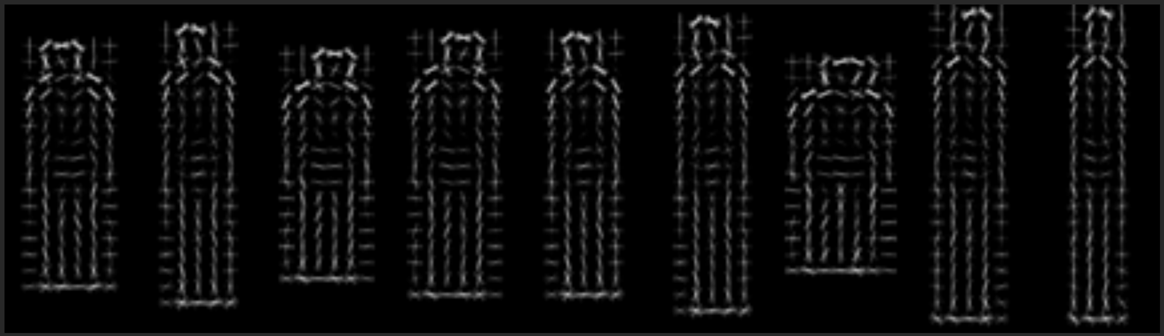
Why do parts help?

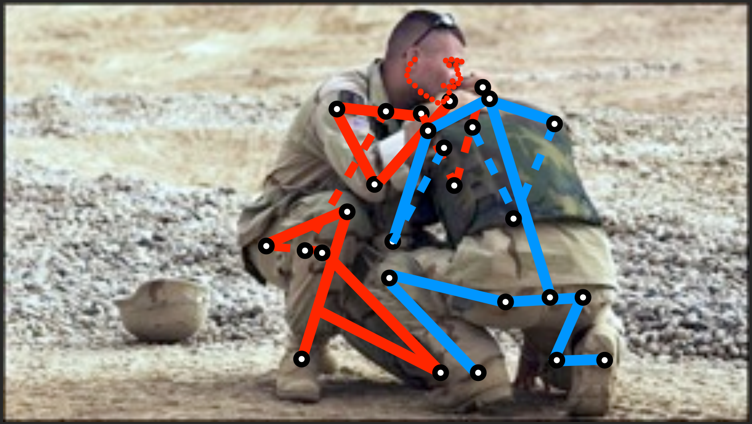




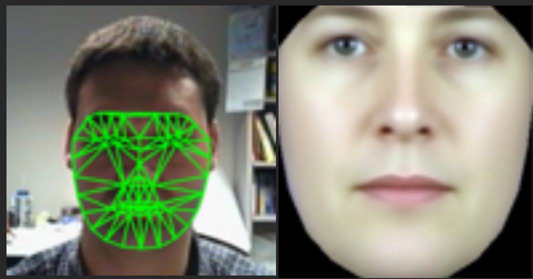
Star-models capture local affine (stretch,rotate,shear) deformations of template

Are quasi-rigid templates enough?





Deformable shape

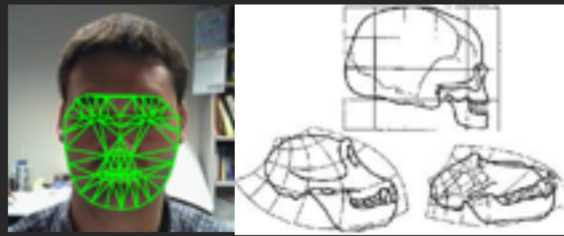


Factored models of elastic geometry + appearance

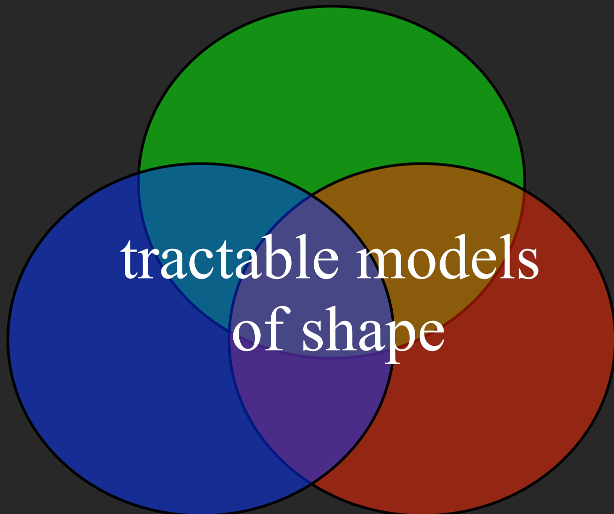
Successful in graphics, but why not vision?

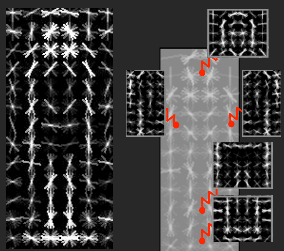
Too complex: inference is hard

Trifecta of shape



Flexible models

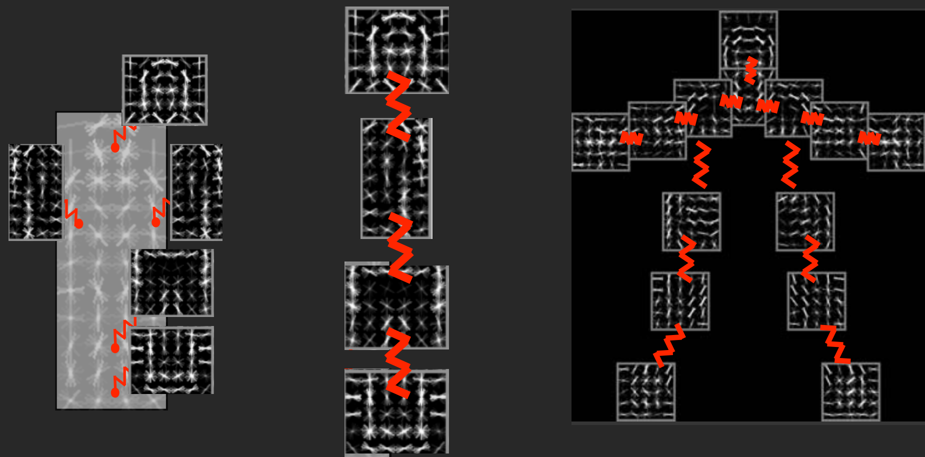




Quasi-rigid

Structureless

Tractable shape



Star Chain Tree

Increasingly flexible

Overview

Background: part models

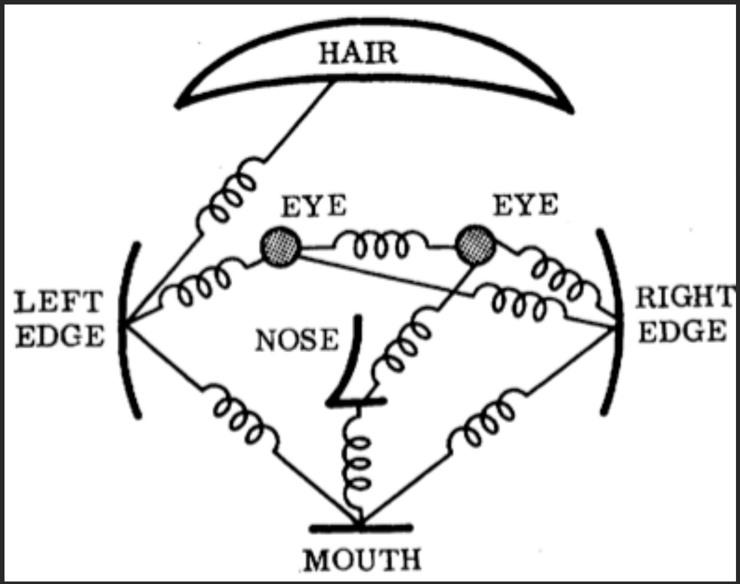
Articulation

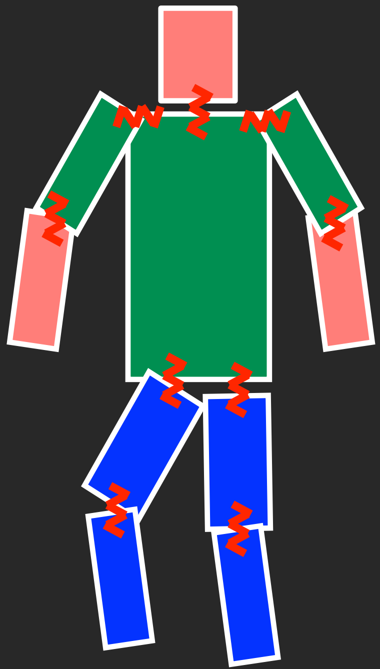
Occlusion

3D viewpoint

Extensions

Old idea: part models

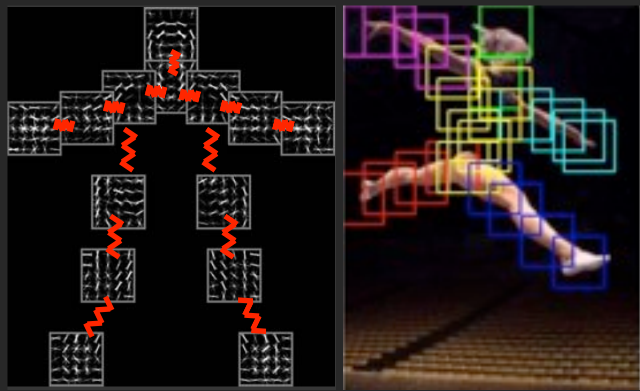




Model encodes local appearance + pairwise geometry
40 year history in vision

Pictorial Structures (Fischler & Elschlager 73, Felzenswalb and Huttenlocher 00)
Cardboard People (Yu et al 96)
Body Plans (Forsyth & Fleck 97)
Active Appearance Models (Cootes & Taylor 98)
Constellation Models (Burl et all 98, Fergus et al 03)

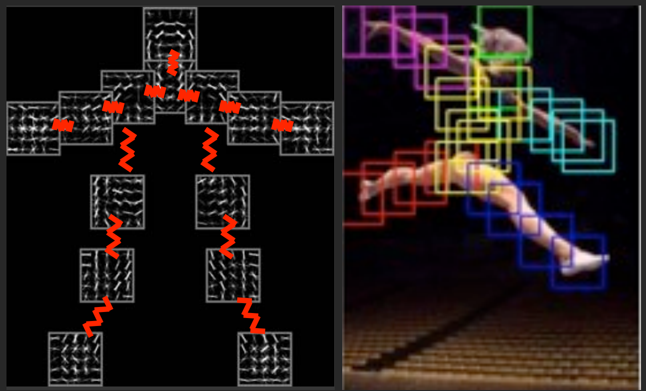
Background: deformable part models



$$S(x, p) =$$

x = image
$p_i = (x_i, y_i)$
$p = \{z_1, z_2 \ldots\}$

Background: deformable part models

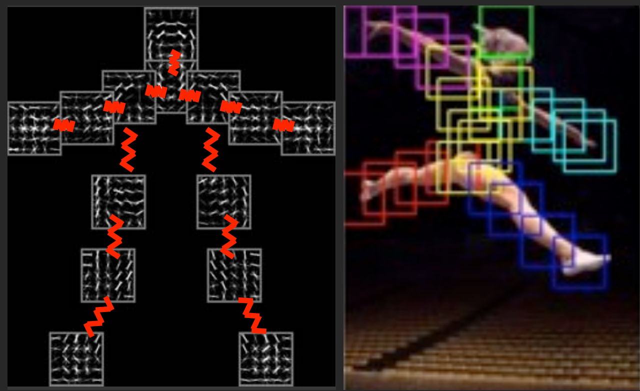


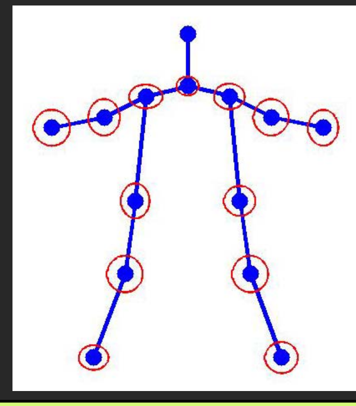
$$S(x,p) = \sum_i w_i \cdot \phi(x, p_i) +$$

x = image
$p_i = (x_i, y_i)$
$p = \{z_1, z_2 \ldots\}$

part template
scores

Background: deformable part models





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

x = image
$p_i = (x_i, y_i)$
$p = \{z_1, z_2 \ldots\}$

part template scores

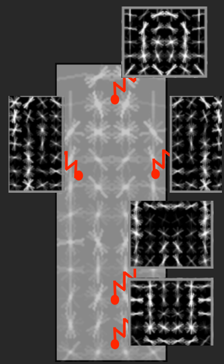
spring deformation model

$$\psi(p_i, p_j) = \begin{bmatrix} dx & dx^2 & dy & dy^2 \end{bmatrix}^T$$

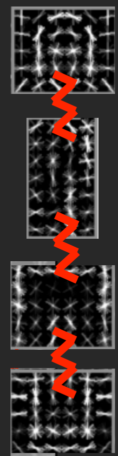
E = relational graph

Shape term

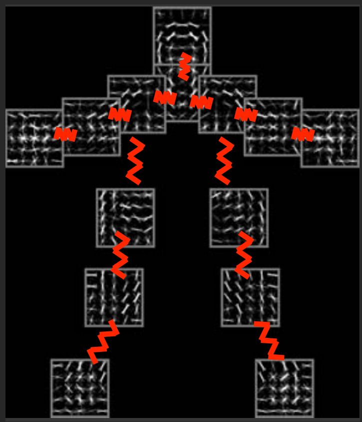
$$\sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j)$$



Star



Chain



Tree

Can make pairwise term image-independant $\psi(x, p_i, p_j)$

Sapp et al ECCV,CVPR,NIPS 2010
Tran & Forsyth ECCV 10

Shape term

$$\boxed{\sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j)} = (p - \mu)^T \Lambda (p - \mu)$$

where (μ, Λ) are functions/reparameterizations of $\{w_{ij}\}$

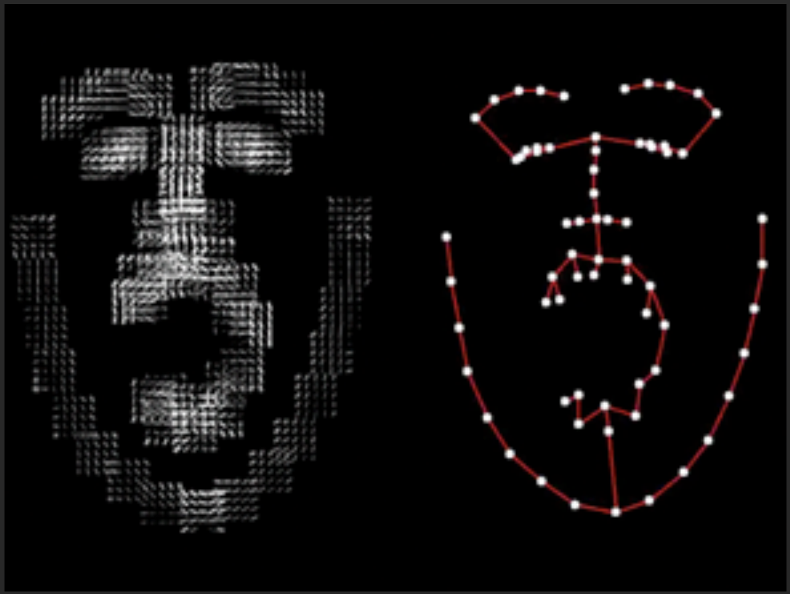
and Λ is the block-sparse inverse of a shape "covariance" matrix

Shape term

$$\boxed{\sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j)} = (p - \mu)^T \Lambda (p - \mu)$$

where (μ, Λ) are functions/reparameterizations of $\{w_{ij}\}$

and Λ is the block-sparse inverse of a shape "covariance" matrix



Lesson: stars don't deform that much, but trees do!

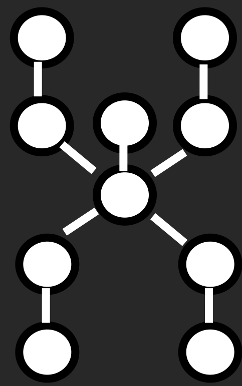
Shape term (derivation)

$$\sum_{ij \in E} a_{ij} dx^2 + b_{ij} dx + c_{ij} dy + d_{ij} dy^2 =$$

$$\sum_{ij \in E} \begin{pmatrix} p_i - \mu_i \\ p_j - \mu_j \end{pmatrix}^T \Lambda_{i,j} \begin{pmatrix} p_i - \mu_i \\ p_j - \mu_j \end{pmatrix} + \text{constant}, \qquad \text{where} \quad \Lambda_{i,j} = - \begin{bmatrix} a_{ij} & 0 & -a_{ij} & 0 \\ 0 & c_{ij} & 0 & -c_{ij} \\ -a_{ij} & 0 & a_{ij} & 0 \\ 0 & -c_{ij} & 0 & c_{ij} \end{bmatrix}$$

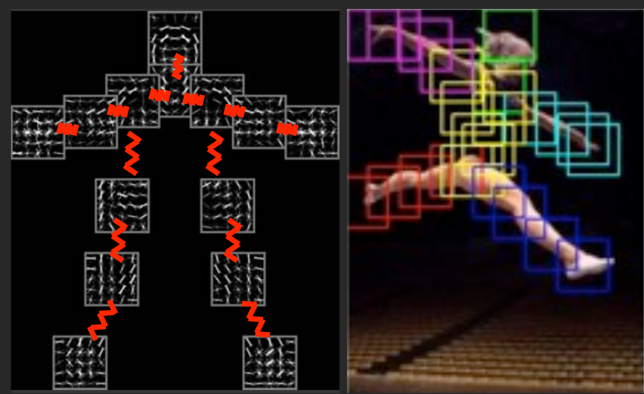
Inference: $\max_p S(x,p)$

Felzenszwalb & Huttenlocher IJCV 05

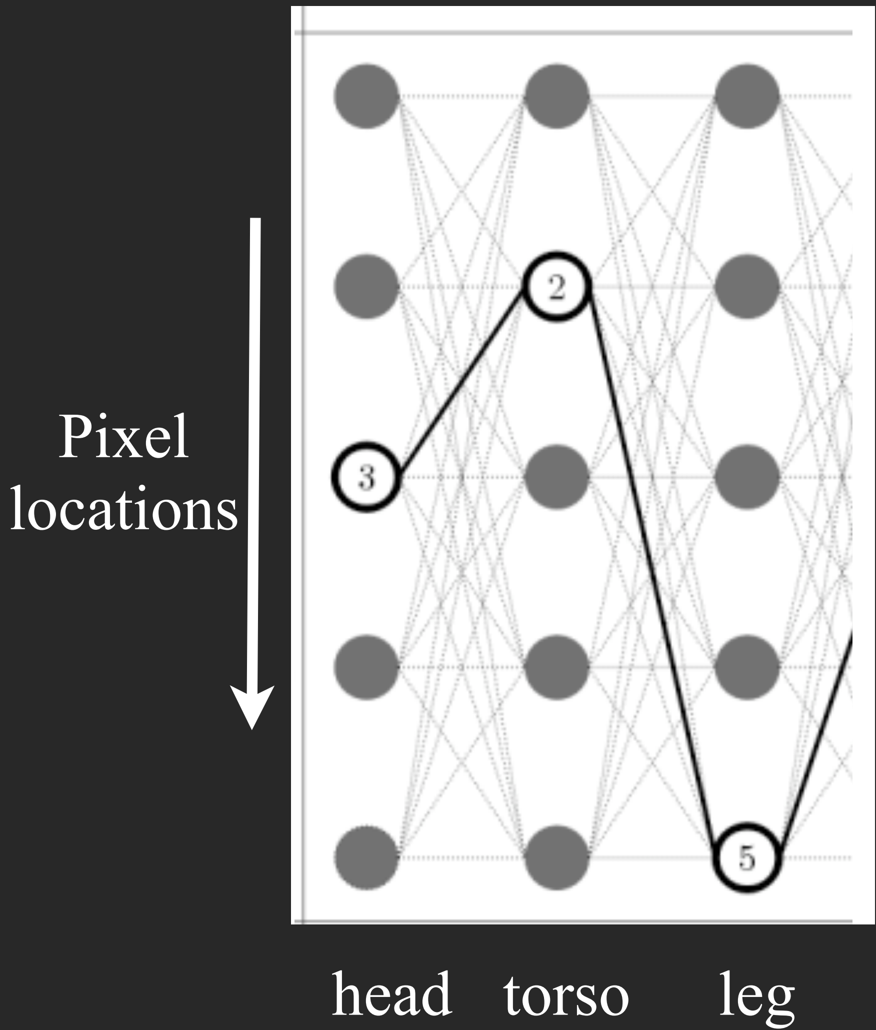


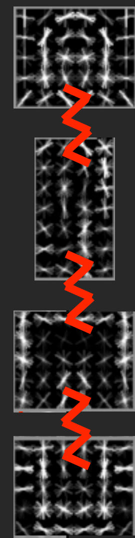
Markov model

- N candidate locations, K parts
- Dynamic programming reduces search from $O(N^k)$ to $O(KN^2)$ for trees
- For each candidate head, independently estimate best left and right arm
- In practice, no more expensive than scoring each part independently



Inference: $\max_{p} S(x,p)$





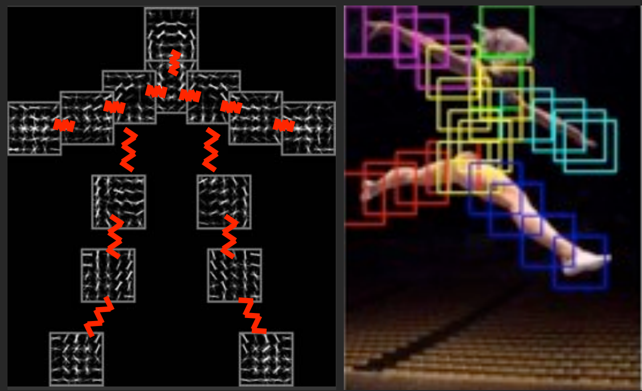
1) Initialize nodes with match score

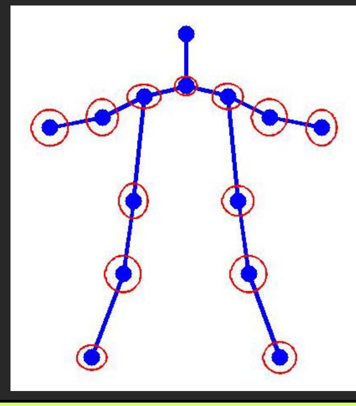
2) Initialize edges with spring score

3) Find best path from left to right

In practice, (1) is bottleneck

Background: linearly-parameterized deformable part models





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

x = image
$p_i = (x_i, y_i)$
$p = \{z_1, z_2 \ldots\}$

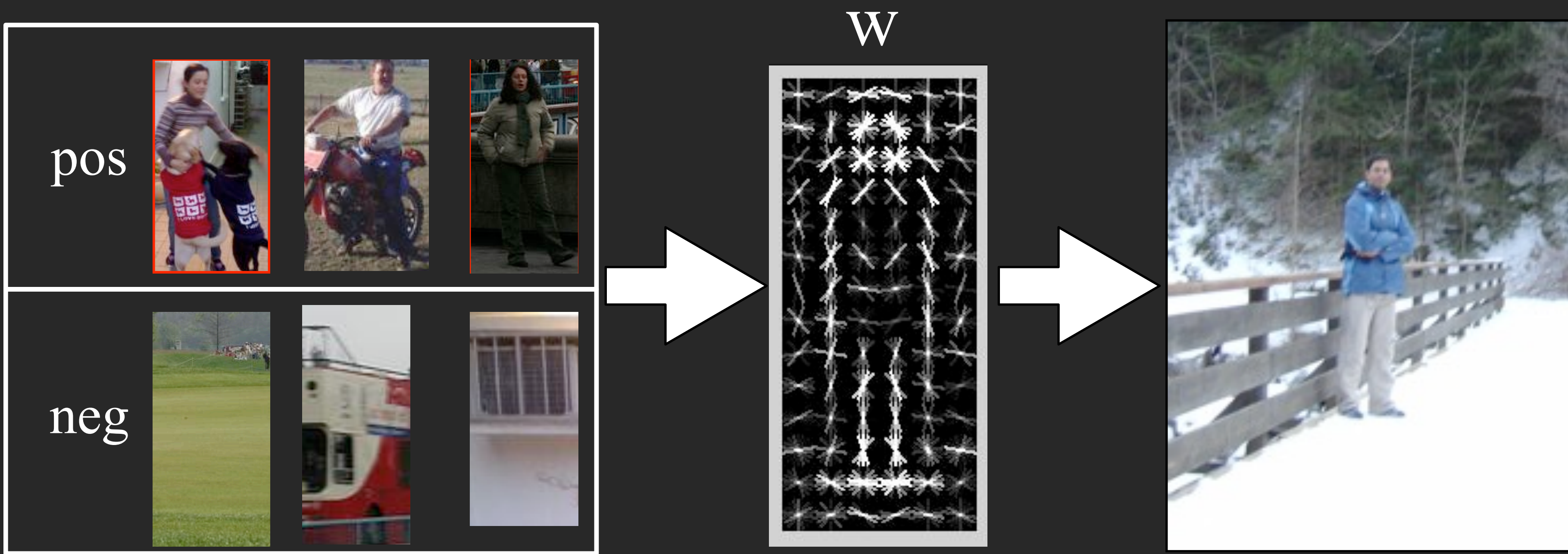
part template scores

spring deformation model

Score is linear in local templates w_i and spring parameters w_{ij}

$$S(x,p) = w \cdot \Phi(x,p)$$

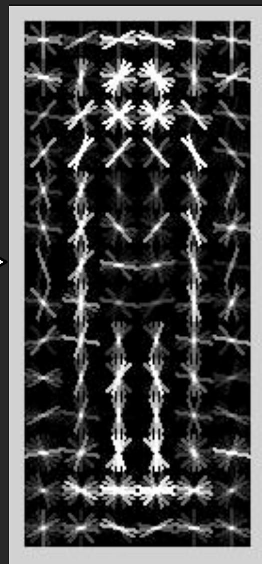
Learning linear parameters



Train ‘w’ with linear classifier (perceptron, SVM, regression, ...)

Learning linear parameters

w

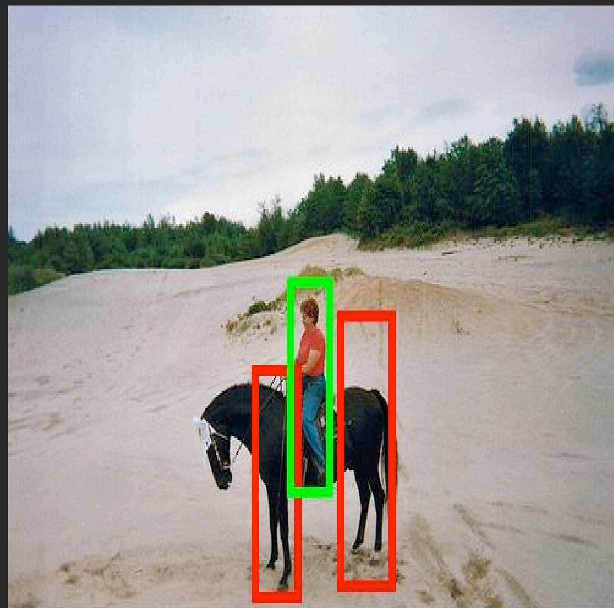


$$\min_{w} ||w||^2$$

$$\forall i \in pos, w \cdot x_i > 1$$

$$\forall i \in neg, w \cdot x_i < 1$$

Large-scale learning



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

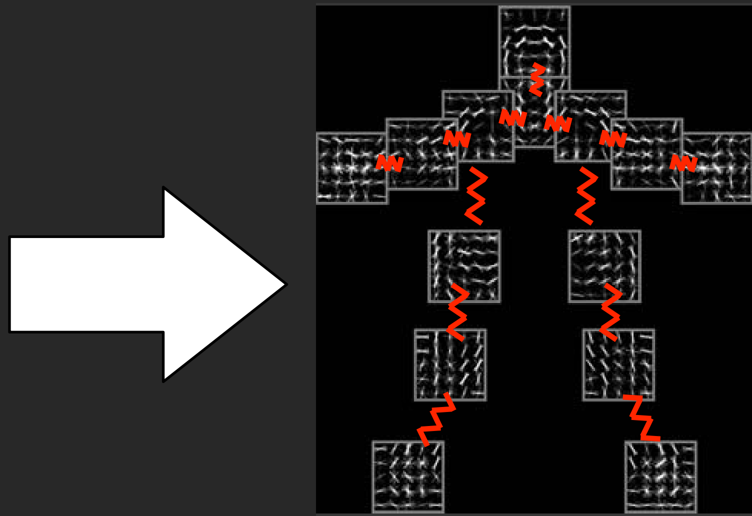
(hundreds of positives, hundreds of millions of negatives)

SVMs are attractive because they generate sparse learning problems

(One can solve problems that are too big to fit in memory)

Learning structured linear parameters

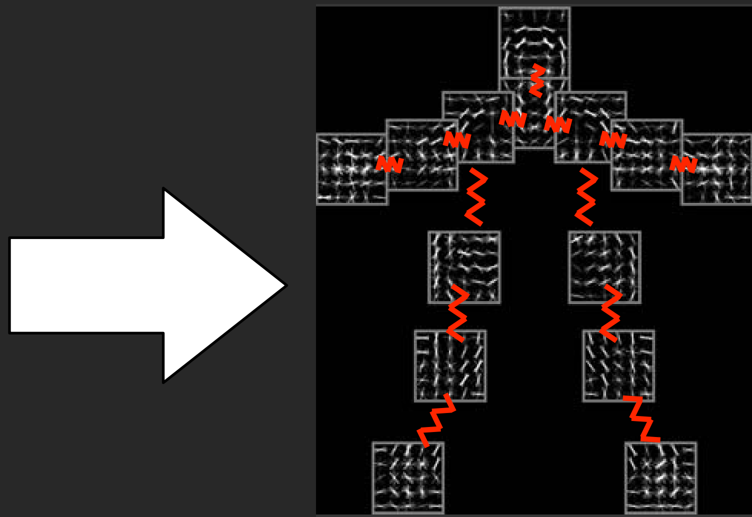
$$S(x, p) = w \cdot \Phi(x, p)$$



(Apply same sparse learning tricks to deal with exponential set of negatives!)

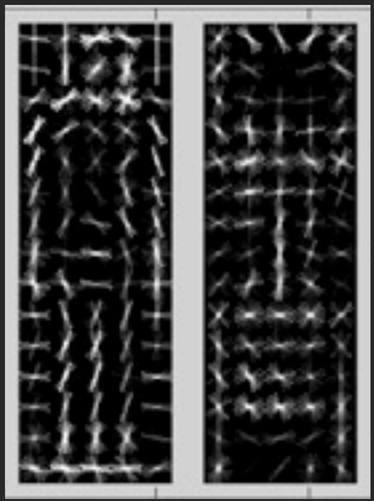
Learning structured linear parameters

$$S(x,p) = w \cdot \Phi(x,p)$$

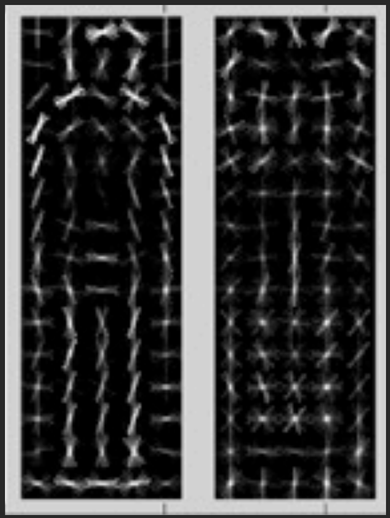


(Apply same sparse learning tricks to deal with exponential set of negatives!)

Perhaps we don't even need SVMs?



SVM



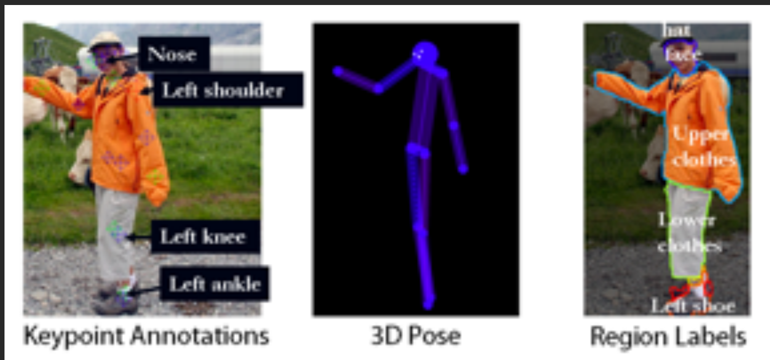
Gaussian model

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

Learn templates with simple statistical Gaussian models

Hariharan, Malik, Ramanan ECCV 12

Datasets



H3D Berkeley

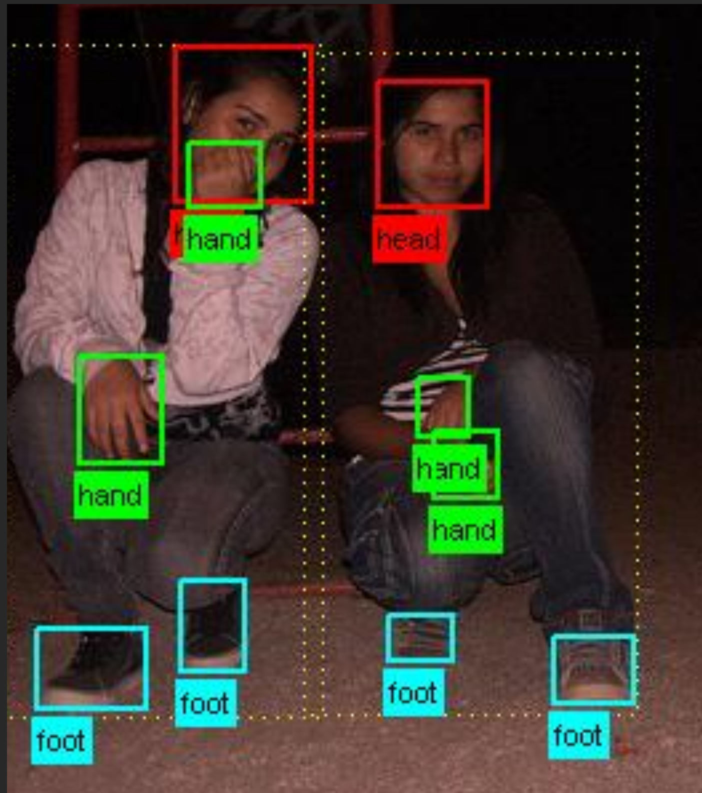
Buffy Oxford

Leeds Sports dataset

PASCAL Stickman

PASCAL Layout Competition

(the forgotten challenge)



	head	hand	foot
OXFORD_RANK_SLACK_RBF	72.9	26.9	4.1

(sole entry in 2011)

Makes use of DPM to detect candidate parts & reranks with SVM

Overview

Background: part models

Representations

Occlusion

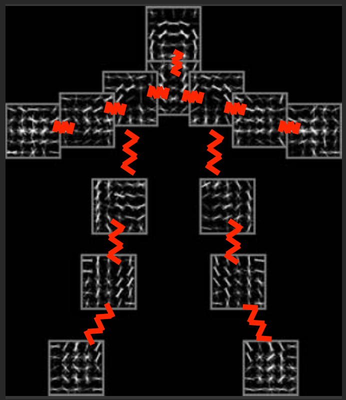
3D variation

Extensions

What’s wrong with part models?

(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)

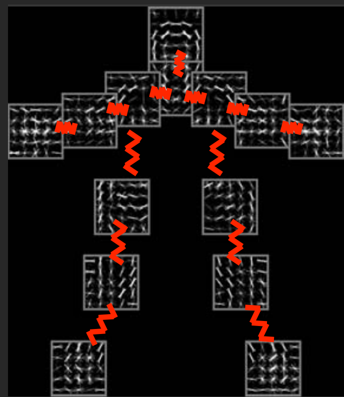


When does this fail?

What’s wrong with part models?

(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)



When does this fail?

Articulation

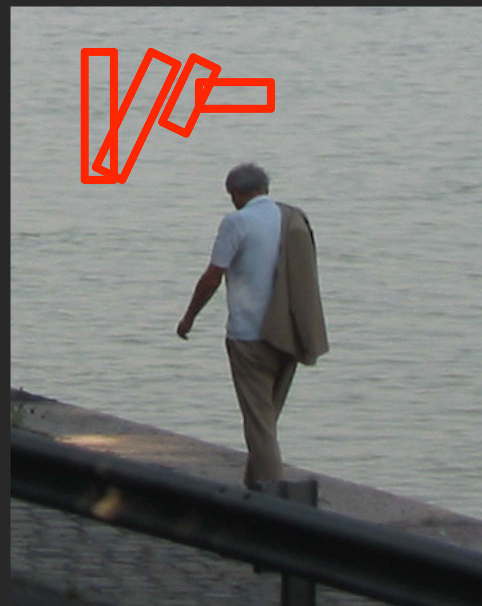
3D viewpoint

Occlusion

Modeling articulation

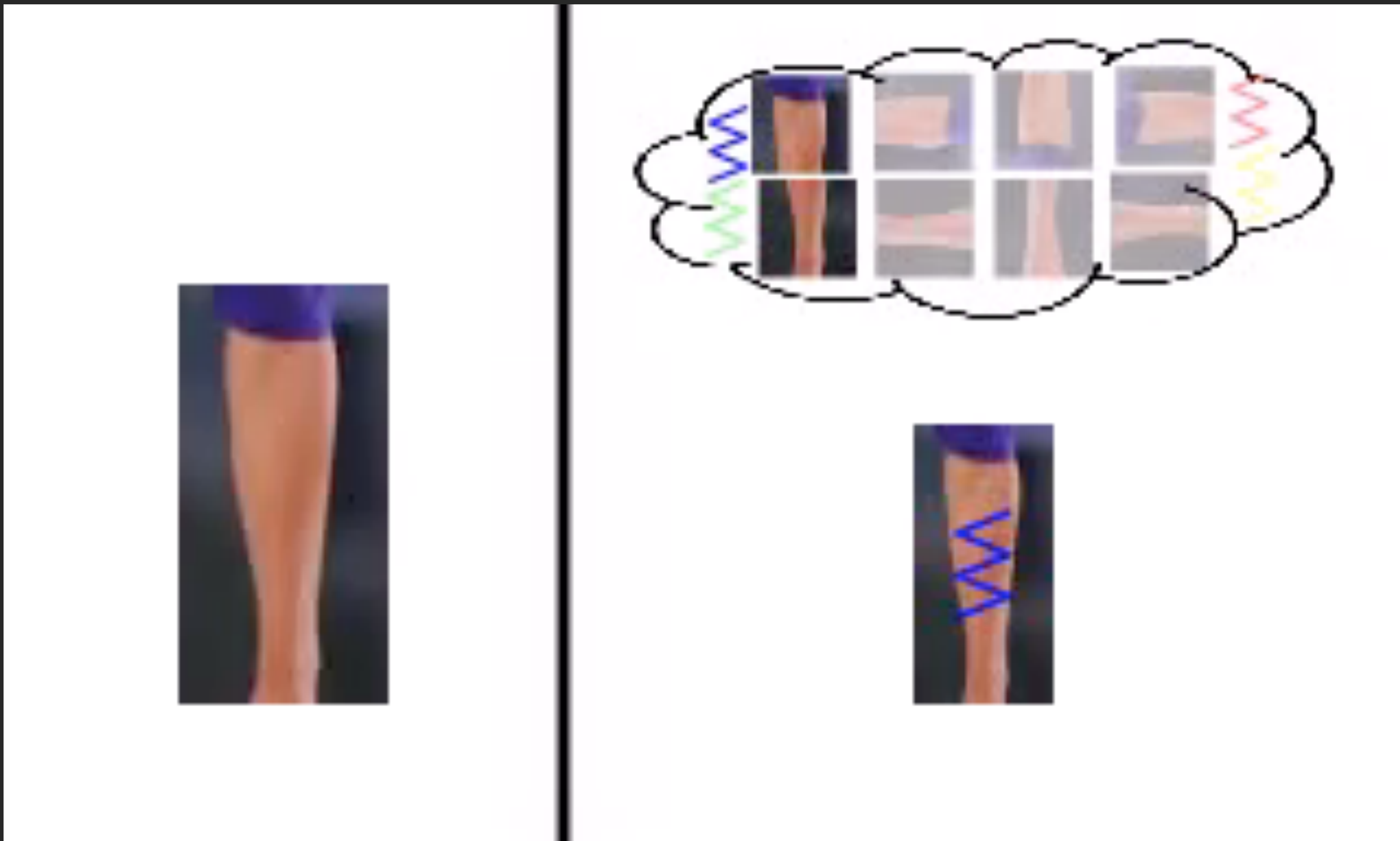
Enlarge state space of part location to include orientation and foreshortening

$(x_i, y_i) \Rightarrow (x_i, y_i, \theta_i, s_i)$



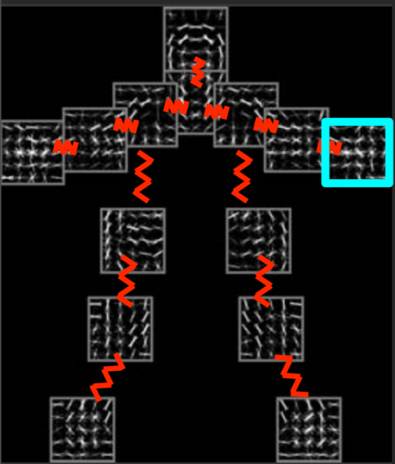
Problem: rather expensive and doesn't work well

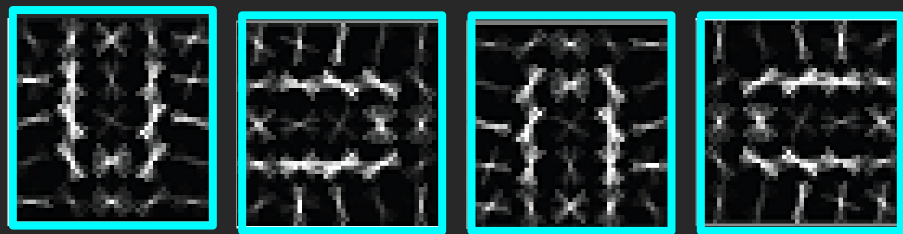
One solution: local mixtures of small patches



Any smooth spatial transformation is locally rigid

Local mixtures of parts





$$p_i = (x_i, y_i)$$

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

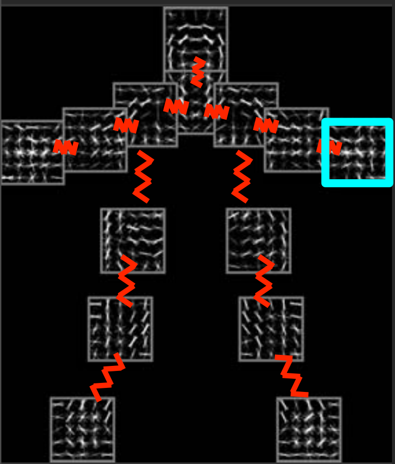
$$S(x,p,t) = \sum_i w_i^{t_i} \cdot \phi(x,p_i) + \sum_{ij \in E} w_{ij}^{t_i,t_j} \cdot \psi(p_i,p_j)$$

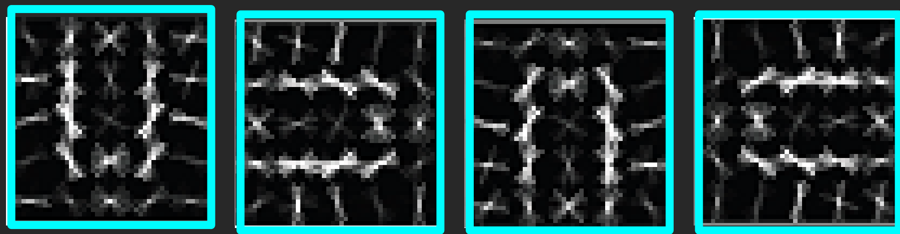
Each part has a position ‘p’ and mixture type ‘t’

Score local model with one of T templates

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

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

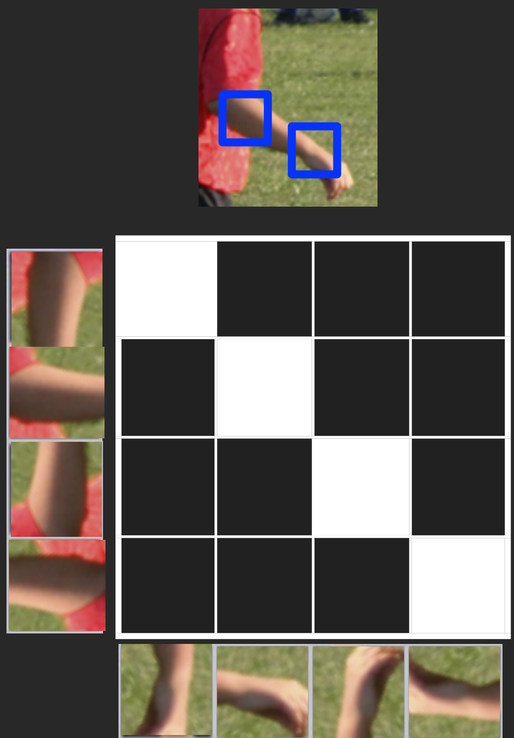
Score deformation with one of T^2 springs (interdependence of geometry + appearance)

Appearance relations

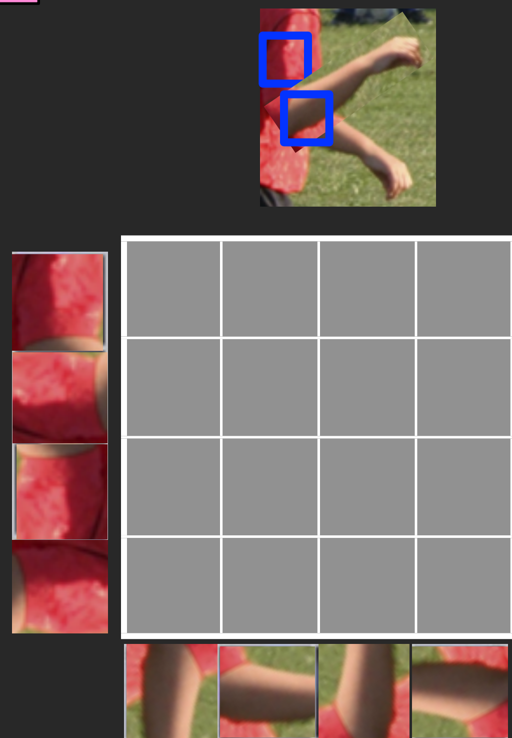
Spring co-occurrence prior

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

Rigid relation

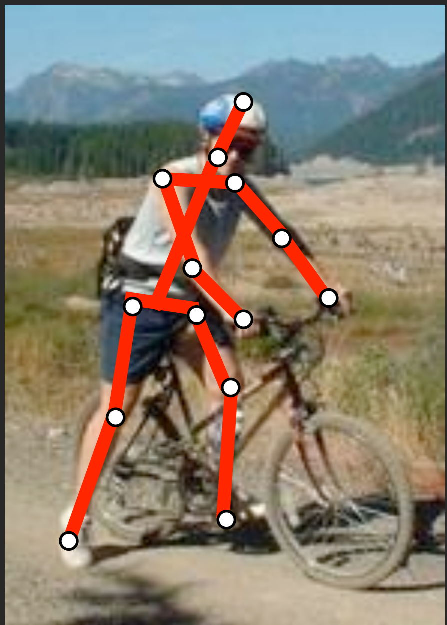


Flexible relation



Supervised learning

$$S(x, p, t) = w \cdot \Phi(x, p, t)$$



Given $\{x_n, p_n, t_n\}$, tune 'w' such that S(x,p,t) scores high on people and low on backgrounds

(structured prediction)

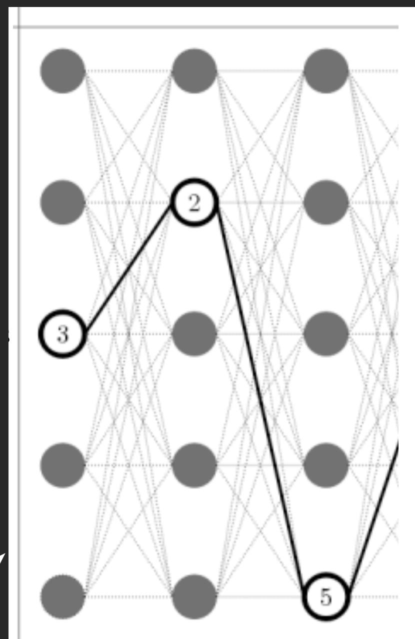
Inference

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

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

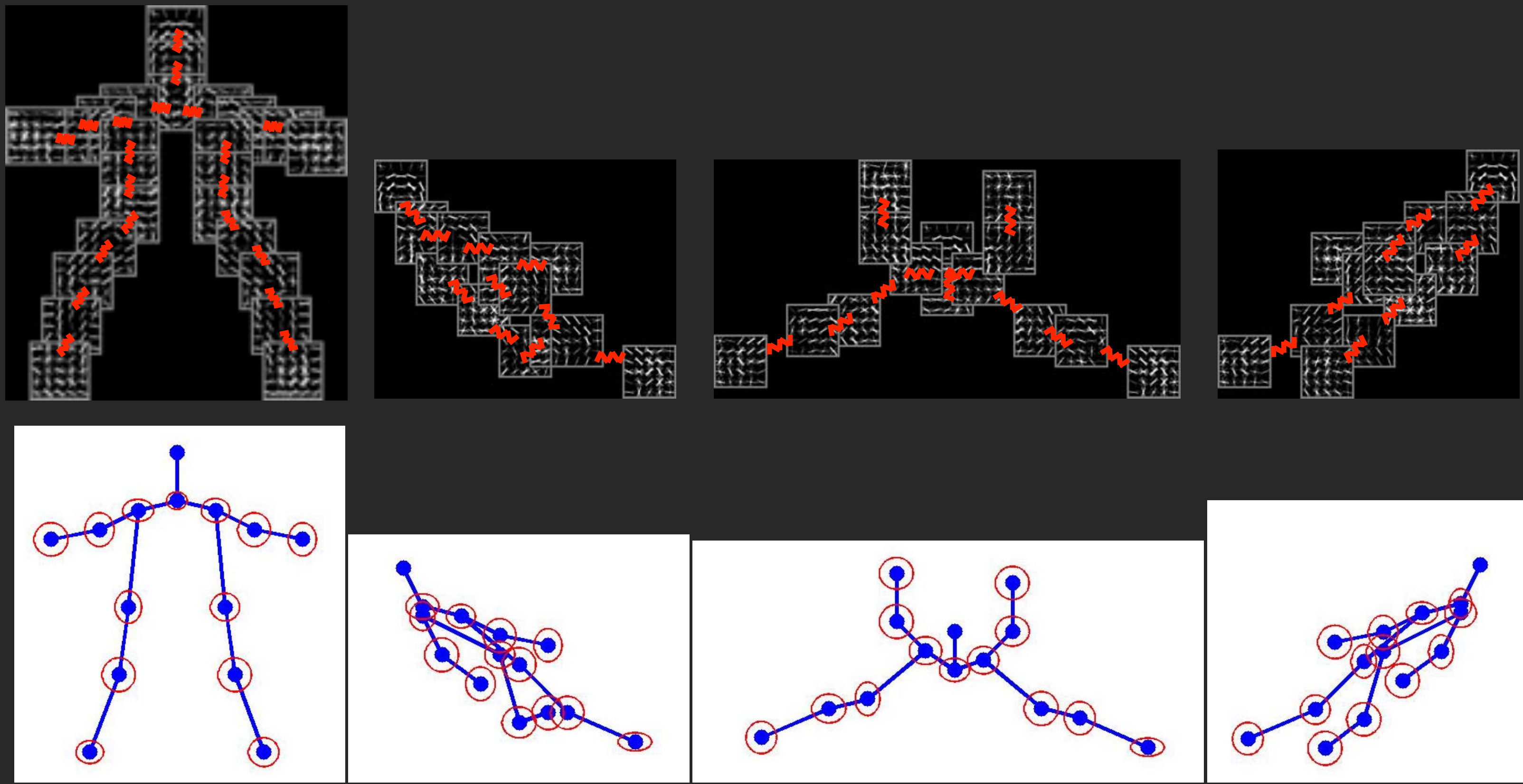
(simple discrete tree-MRF)

Pixel locations and mixture types



head torso leg

Exponential number of global mixtures



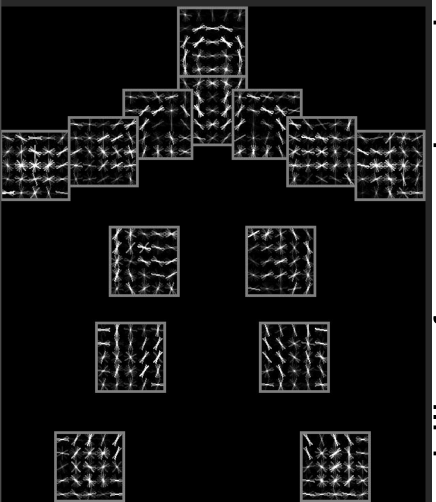
K parts, M local mixtures => K^M unique global mixtures

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

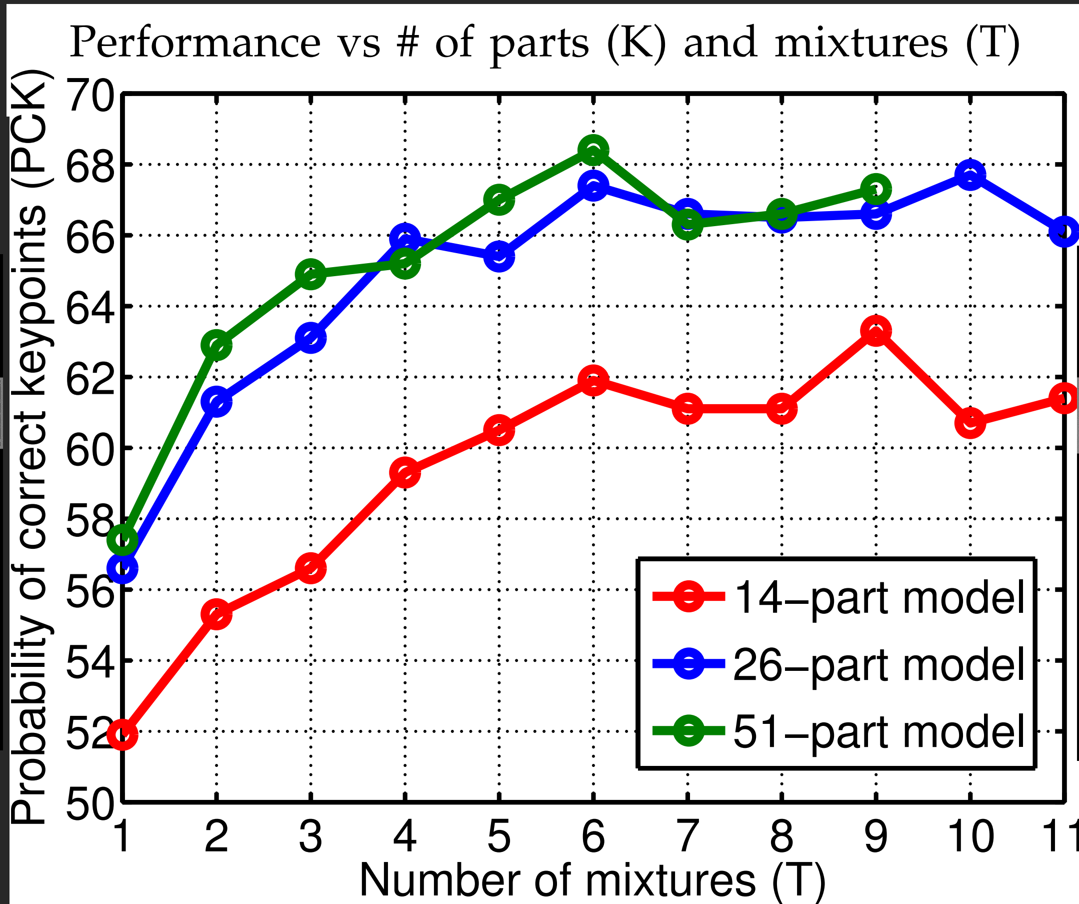
Qualitative Results

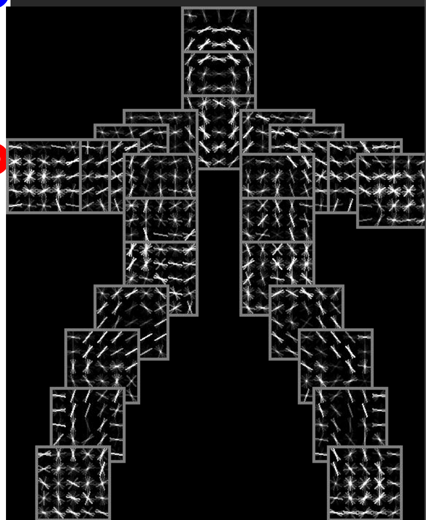
Yi & Ramanan CVPR11

Search over representations



14 parts





26 parts

Denser parts and more local mixtures help (up to a point)

Quantitative evaluation

% of correctly localized limbs

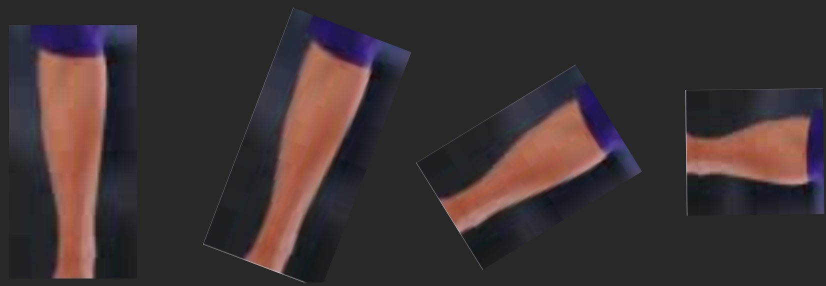
Image Parse Testset

Method	Torso	Head	Upper legs	Lower legs	Upper arms	Lower arms	Total
R [23]	52.1	37.5	31.0	29.0	17.5	13.6	27.2
ARS [1]	81.4	75.6	63.2	55.1	47.6	31.7	55.2
JEa [15]	77.6	68.8	61.5	54.9	53.2	39.3	56.4
SNH [29]	91.2	76.6	71.5	64.9	50.0	34.2	60.9
JEb [14]	85.4	76.1	73.4	65.4	64.7	46.9	66.2
Our Model	**97.6**	**93.2**	**83.9**	**75.1**	**72.0**	**48.3**	**74.9**

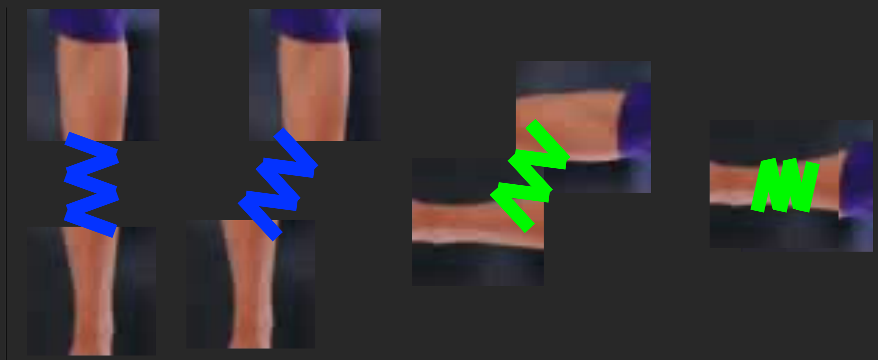
On-par with or outperforms previous work while being orders of magnitude faster (few seconds vs few minutes)

All previous work use explicitly articulated models

Model affine warps of templates with mixtures of pictorial structures



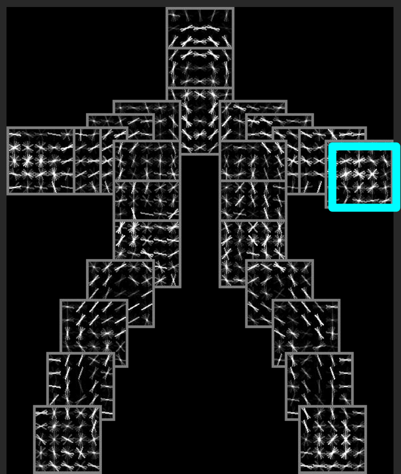
$\approx$

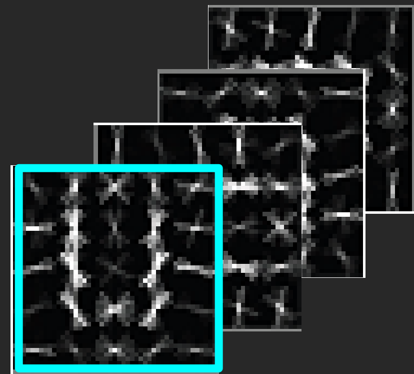


Faster run-time

(small templates + dynamic programming)

What makes it work better?

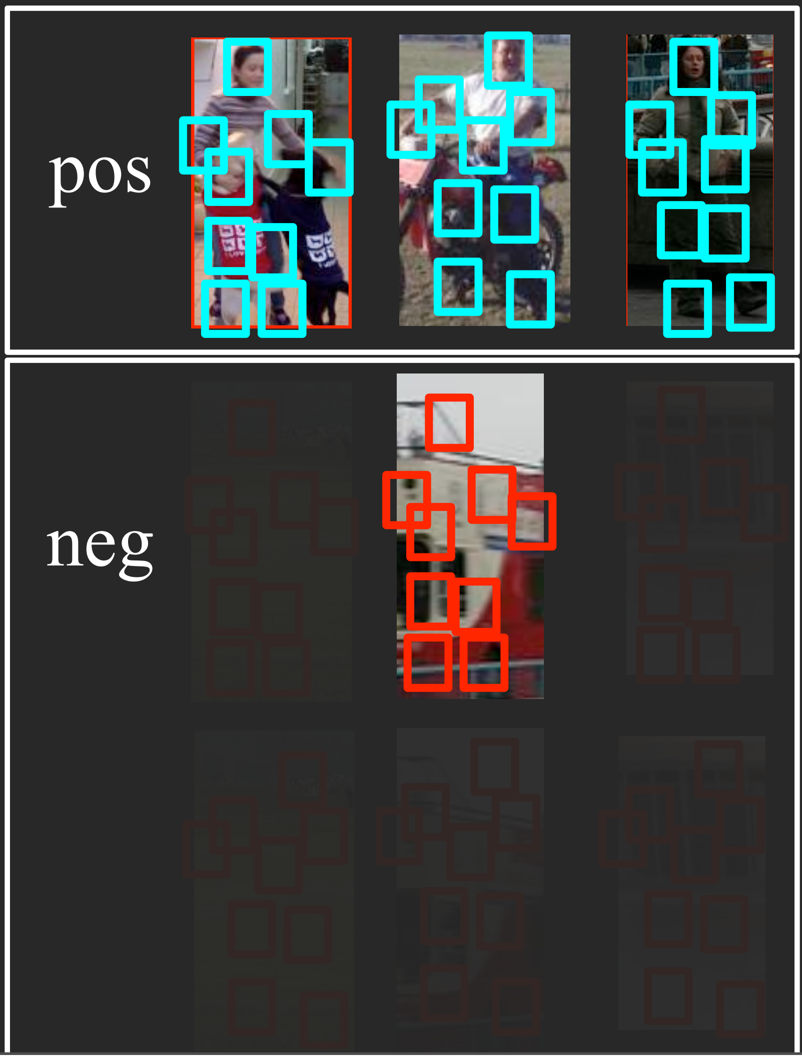


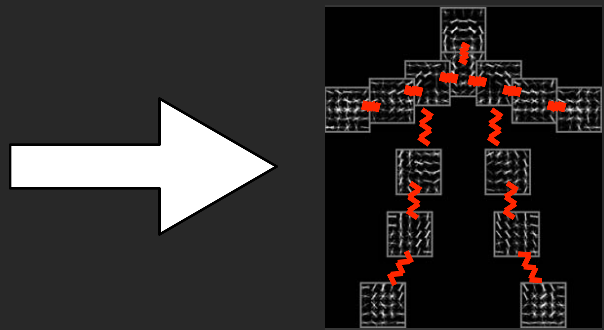


Joint	Indep	Indep+Invar
67.4	51.3	33.8

Why does joint training help?

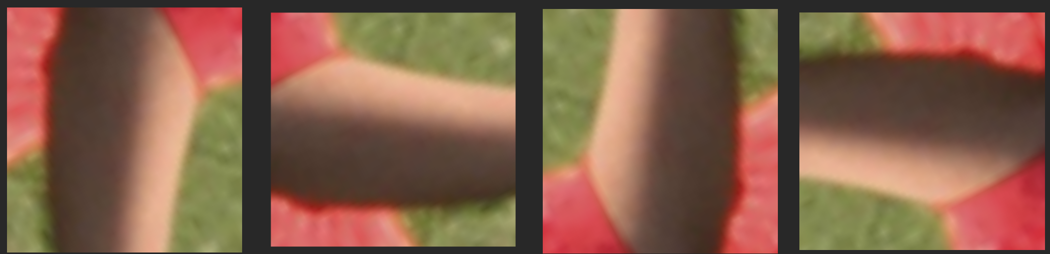
We need compete only against joint configurations
of negatives that score above margin





Why are parts not orientation-invariant?

Joint	Indep	Indep+Invar
67.4	51.3	33.8



Illumination (world is lit from above)
Occlusions (torsos tend to be upright)

Overview

Background: part models

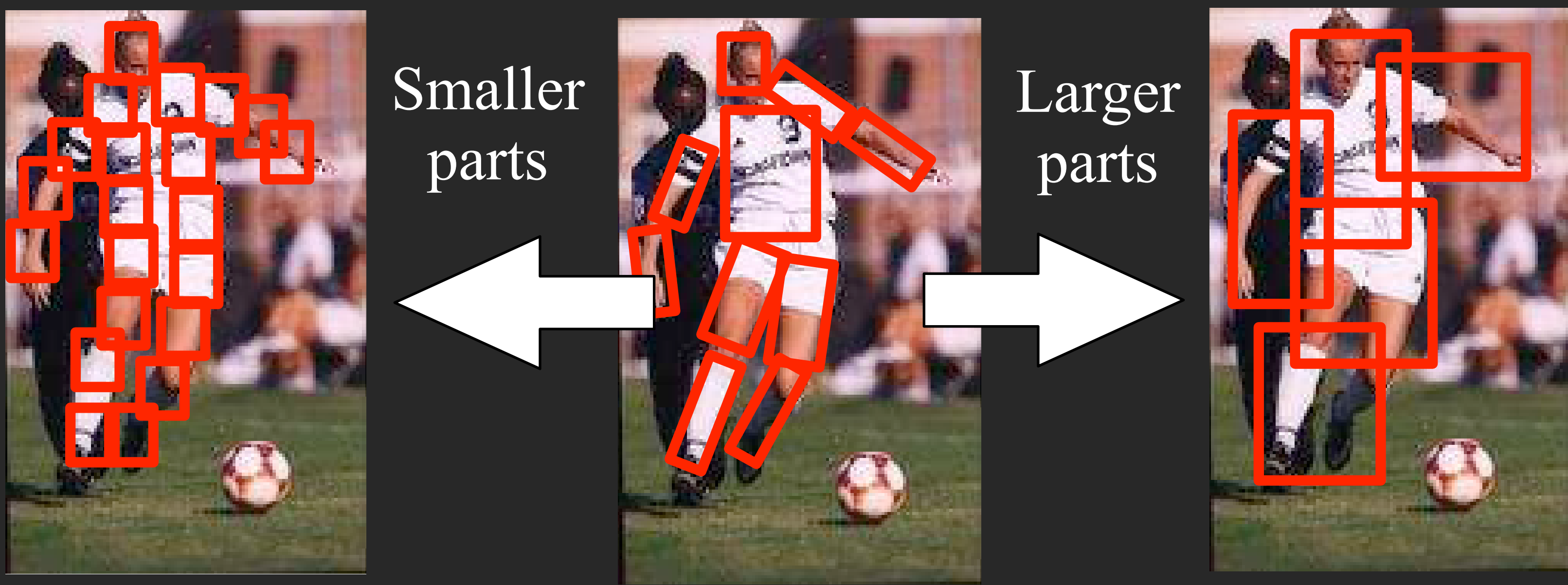
Articulation

Occlusion

3D variation

Extensions

Representations for human pose



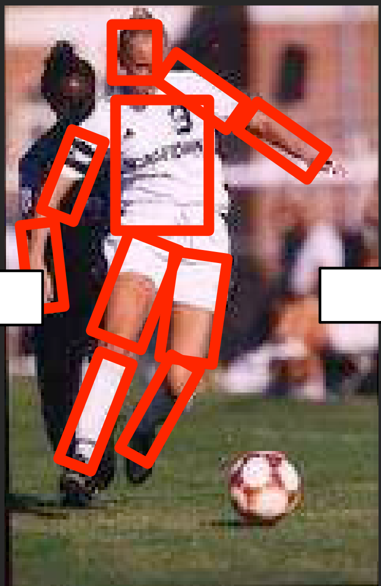
Patches

Skeleton

Poselets

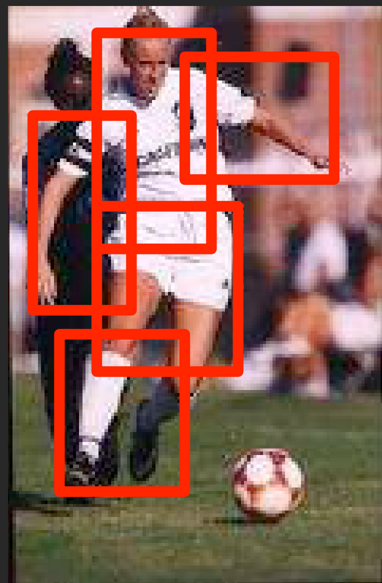
Representations for human pose

er
s



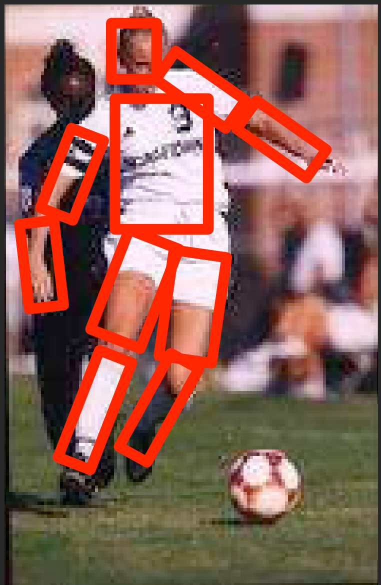
Skeleton

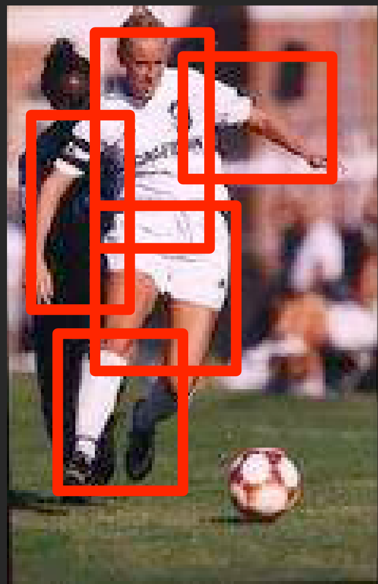
Larger
parts



Poselets

Global representations





Skeleton

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

Poselets

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

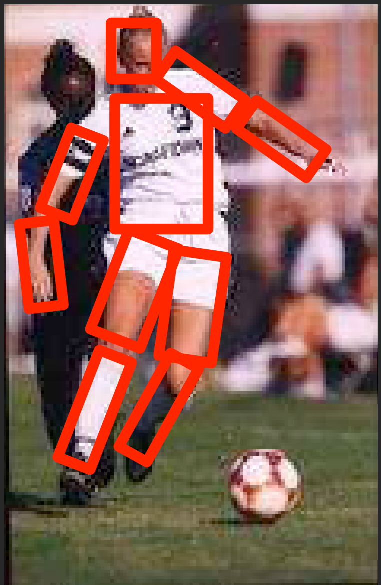
Exemplars

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

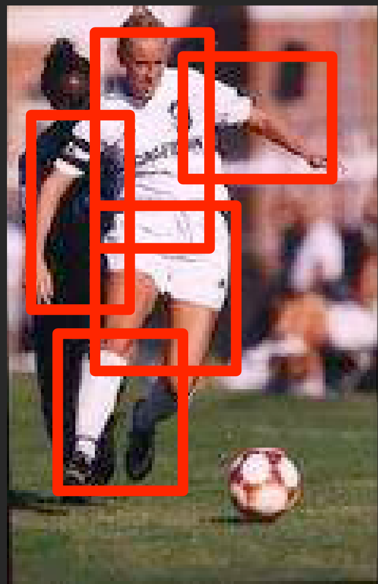
Visual Phrases

Sadeghi and Fahardi

Global representations



Skeleton



Poselets

Exemplars

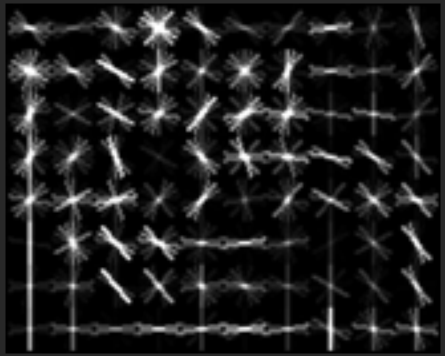
Visual Phrases

Insight from such global approaches (an opinion):

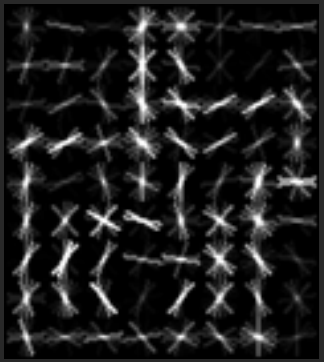
large composite templates better model occlusions and interactions

How to encode complex interactions?

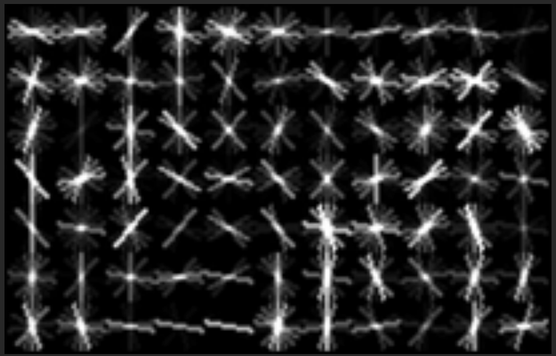
Visual Phrases
Sadeghi and Fahardi, CVPR 11



Person on jumping horse



Person on horse



Person standing next to horse

One may need lots of large composite templates

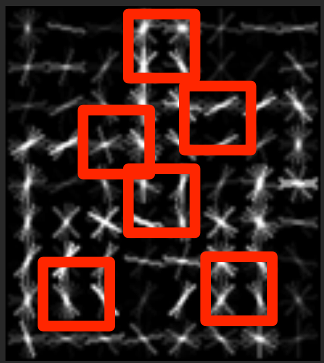
How to encode complex interactions?

Poselets

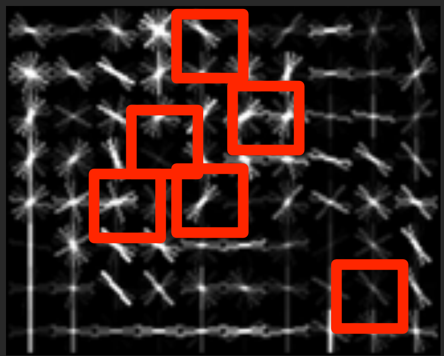
Bourdev & Malik ICCV09

One may need lots of large composite templates

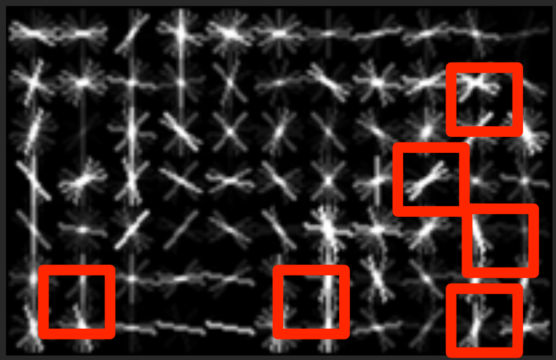
One take: visual "phraselets"



Person on horse



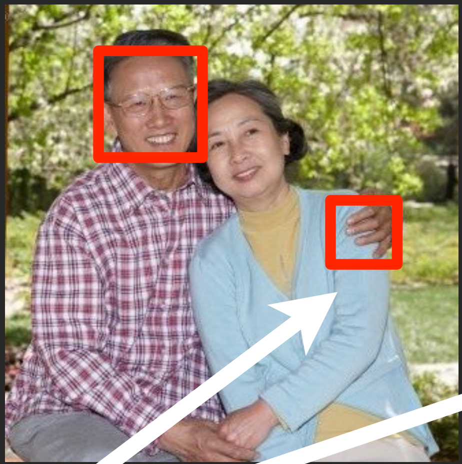
Person on
jumping horse



Person standing
next to horse

Break up visual composite into smaller
patches and reason about appearance relations

One take: visual “phraselets”



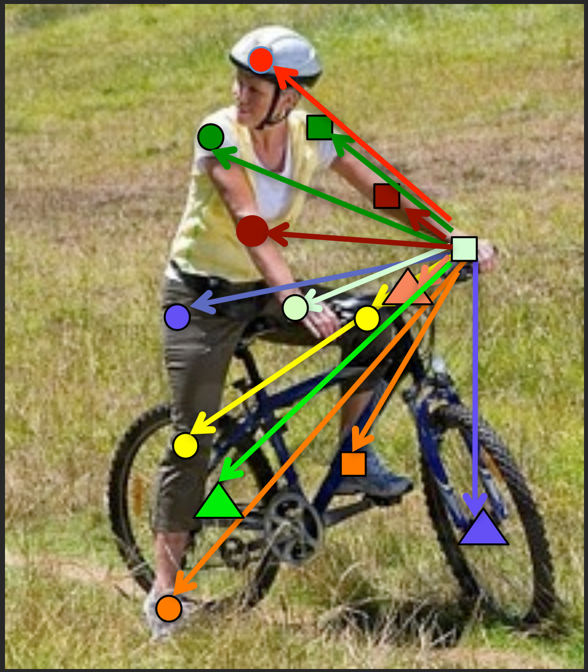
Hand looks different due to interactions with global geometry

We’ll encode such visual differences as local part mixtures

Learning phraselets

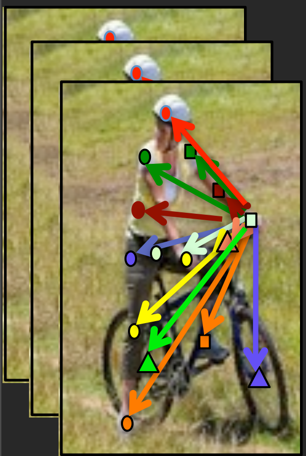
Define phraselets as commonly-occuring geometric configurations

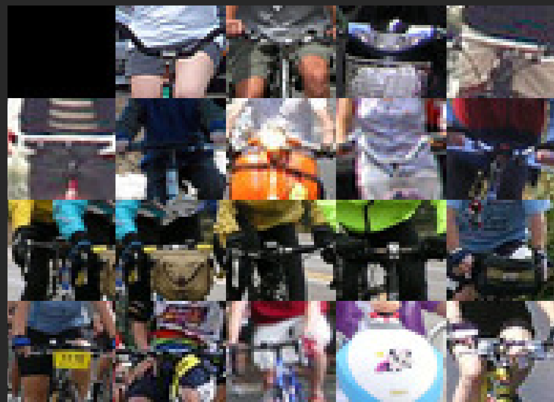
"Poselet-like clusters"



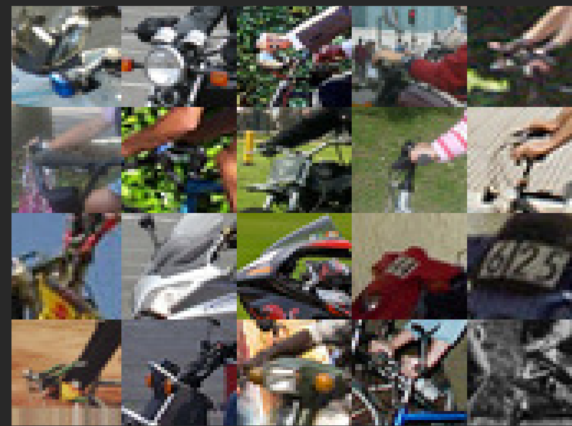
Given labelled training data, find clusters of keypoint configurations relative to each joint

Clusters

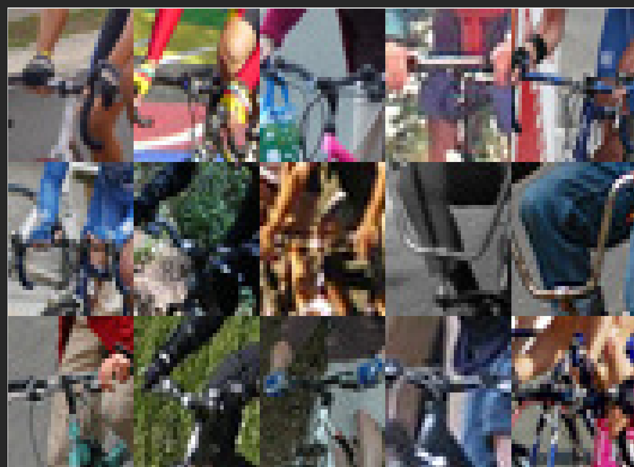




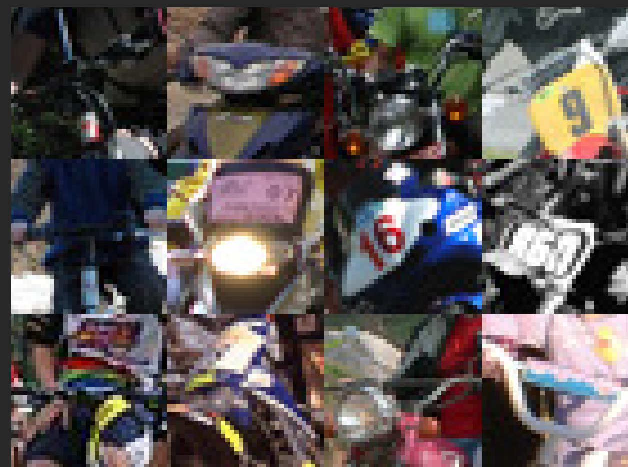
frontal



side-view



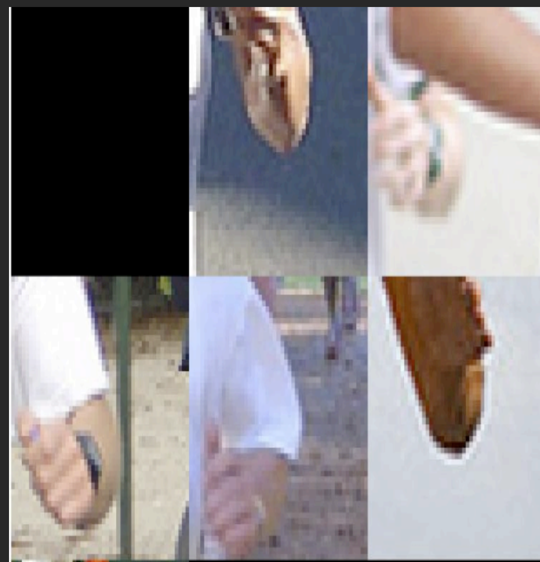
3/4 road bikes



3/4 motorbikes

Model occlusions with separate clusters

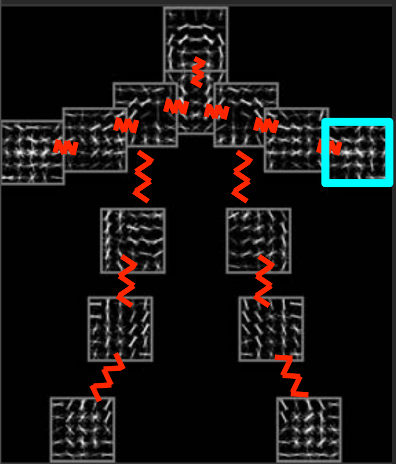
Visible left elbow

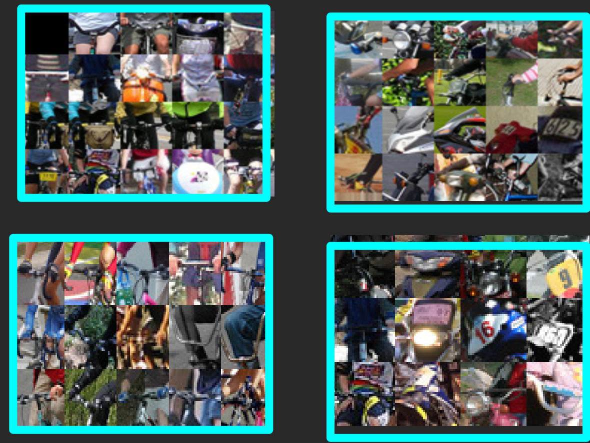


Occluded left elbow

Mixture label corresponds to visible/occlusion state

Local mixtures of phraselets





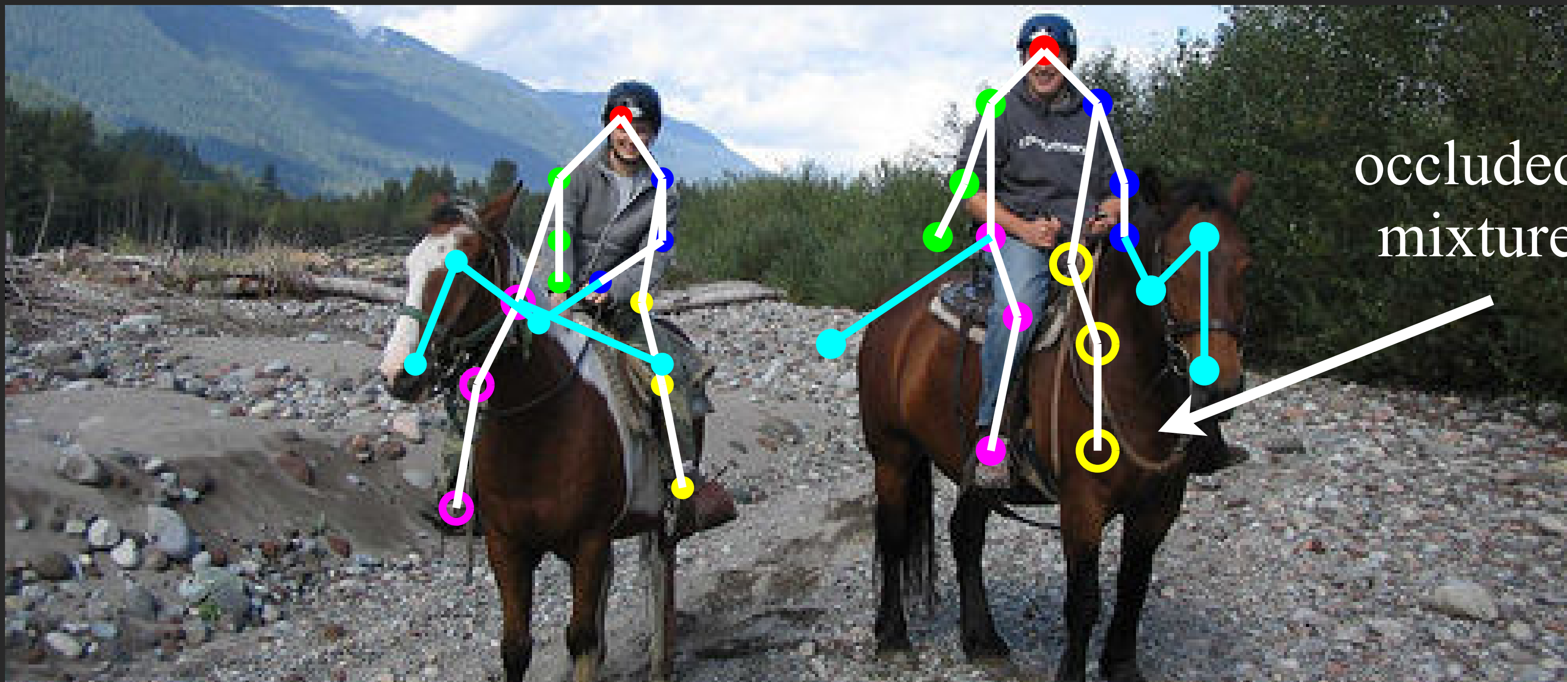
visible occluded

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

Relational model encodes that one (but not both) the left & right leg is occluded when they are nearby

Relational phraselets

Report back human+object part locations and mixture label

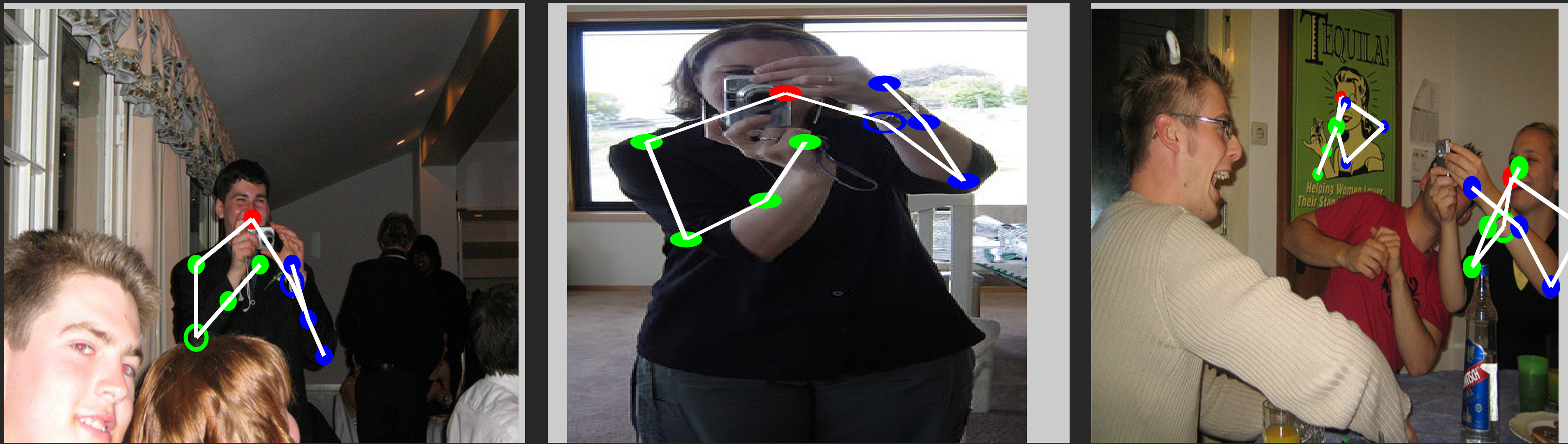


occluded
mixture

Desai and Ramanan ECCV 12

Results

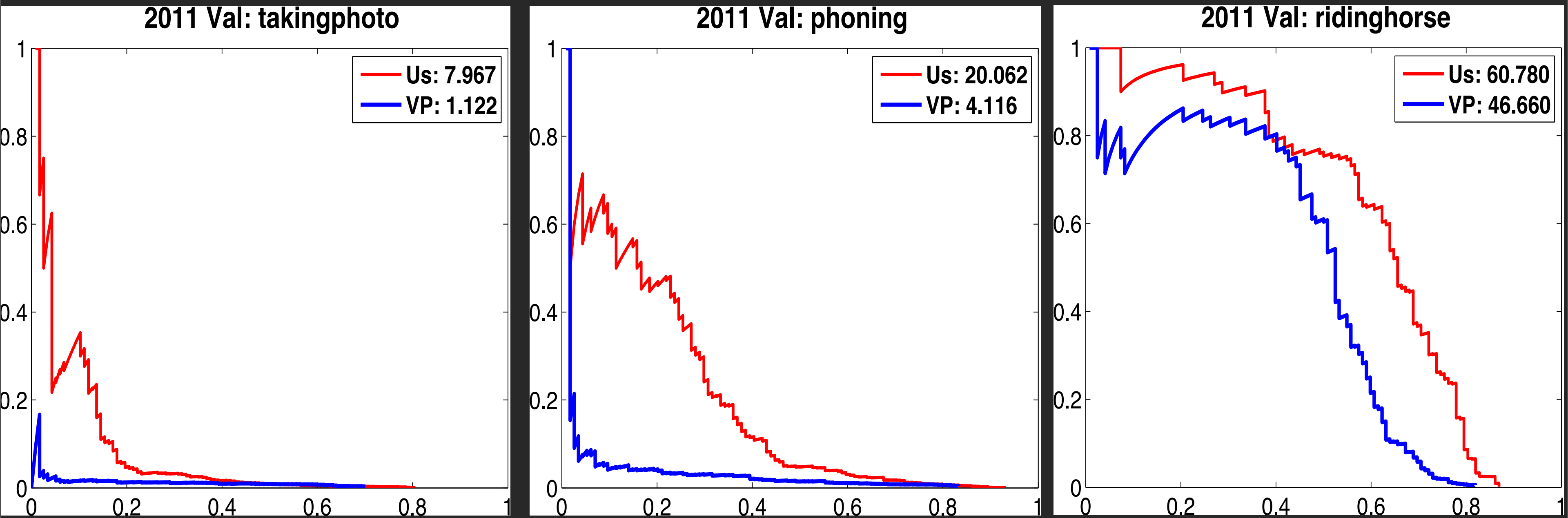
Person+phone



Person+bike

Results

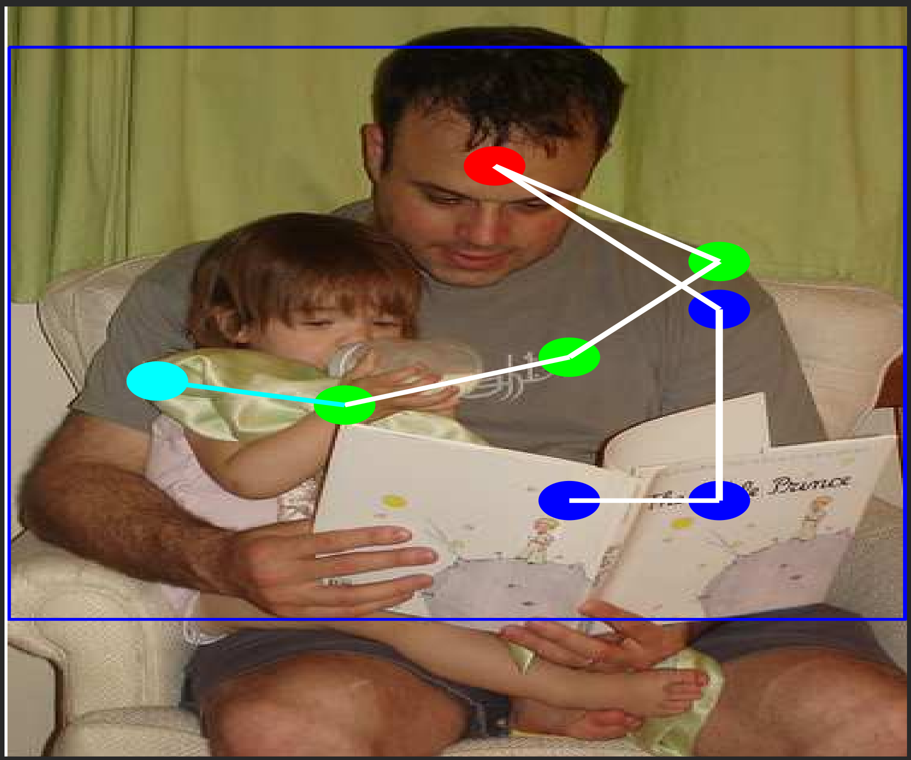
Detecting person-object interactions



Red line: Our compositional phraselet model
Blue line: DPM trained on person+object (visual phrase)

Top false positives

(penalized detections)

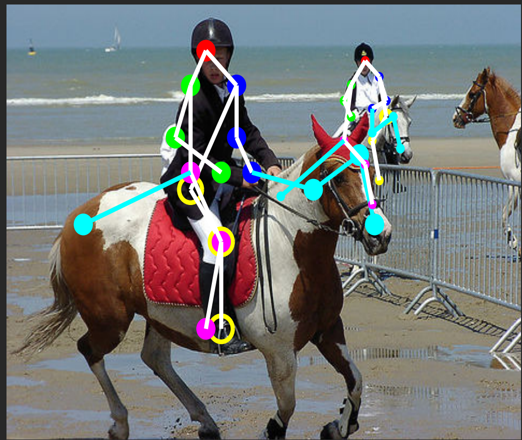


Using computer

Taking photo

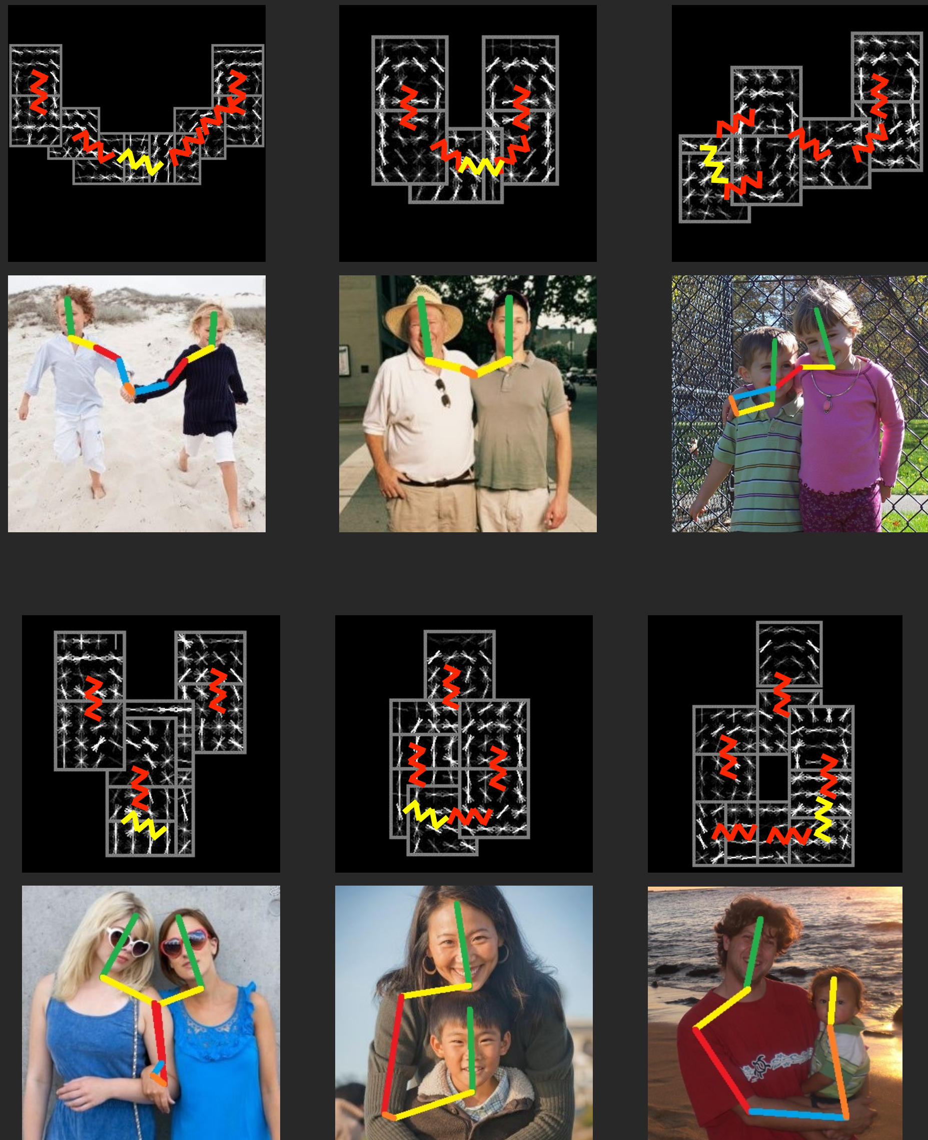
Action classification

	Run.	R. Bike	R. horse	Phoning	TakingPhoto	UsingComp.	Walk.
Us	81.6	**82.6**	84.7	**52.6**	25.3	**54.9**	56.4
Poselets	**83.1**	79.9	**87.6**	45.9	**26.2**	44.9	**66.6**


For action classification (given known bounding box), method near state-of-art

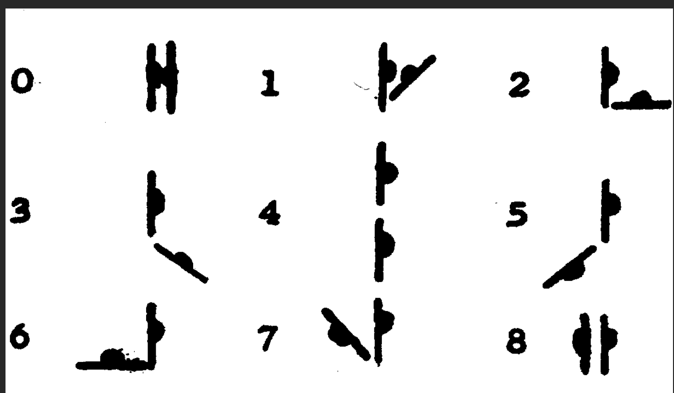
Person-person composites



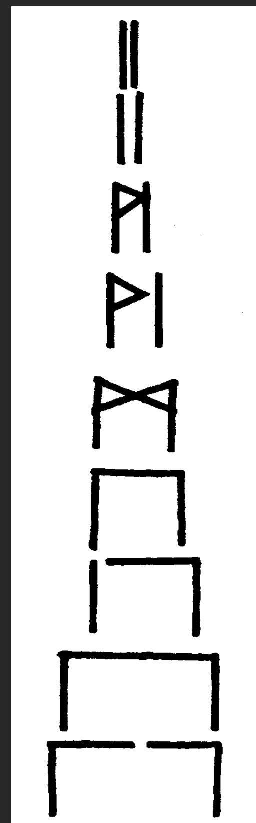
Yang et al. “Recognizing Proxemics in Personal Photo Collections” CVPR12

Proxemic analysis

Edward Hall "A system for the notation of proxemic behavior"
American Anthropologist 1963



Relative body orientation



Touching body parts
(heads, elbows, hands)

Multi-body pose estimation

Eichner & Ferrari. “We are Family: Joint Pose Estimation” ECCV 2010

Yang et al. “Recognizing Proxemics in Personal Photo Collections” CVPR12

Dataset statistics

(a) Image Statistics

No. Images	No. People	No. People Pairs
589	1207	1332

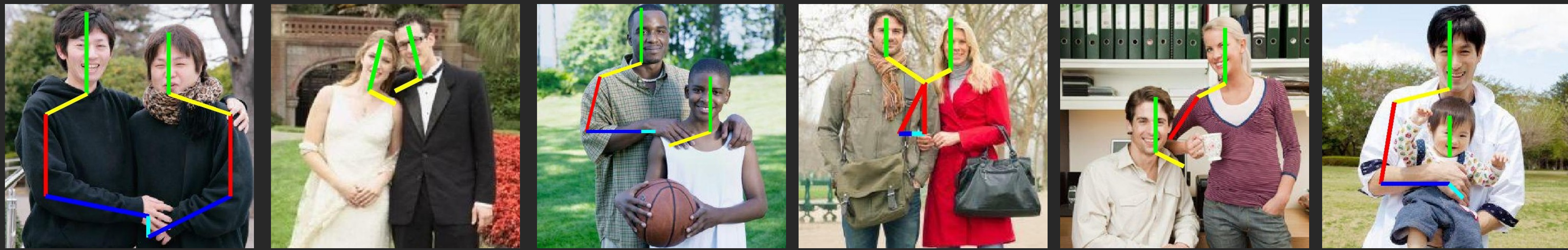
(b) Touch Code Statistics

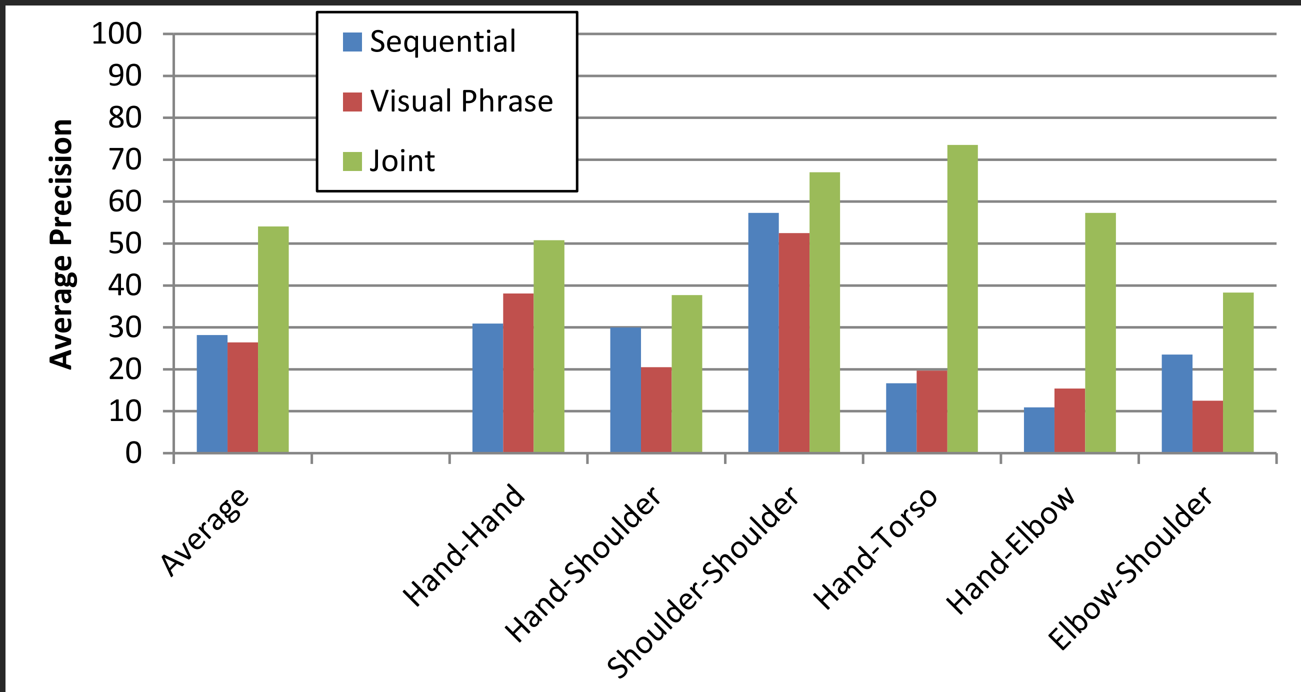
Hand-hand	Hand-shoul	Shoul-shoul	Hand-elbow	Elbow-shoul	Hand-torso
340	180	210	96	106	57
25.5%	13.5 %	15.8 %	7.2%	8.0%	4.3%

(c) Co-occurrence Statistics

0 Codes	1 Code	2 Codes	3+ Codes
531	626	162	13

Quantitative results

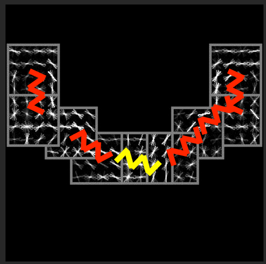
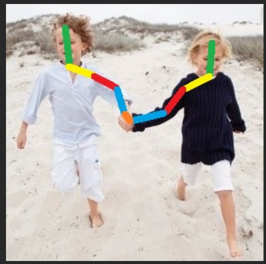




Sequential approach: (1) Estimate pose with single-body model
(2) Classify touch-code based on estimate pose

Yang et al. "Recognizing Proxemics in Personal Photo Collections" CVPR12

Quantitative results

Removing key spring (yellow) drops performance from 52% to 33%

Correct spatial structure is crucial
(e.g., difficult to reason about with a star model)

Overview

Background: part models

Occlusion reasoning

$\boxed{\text{3D variation}}$

Extensions

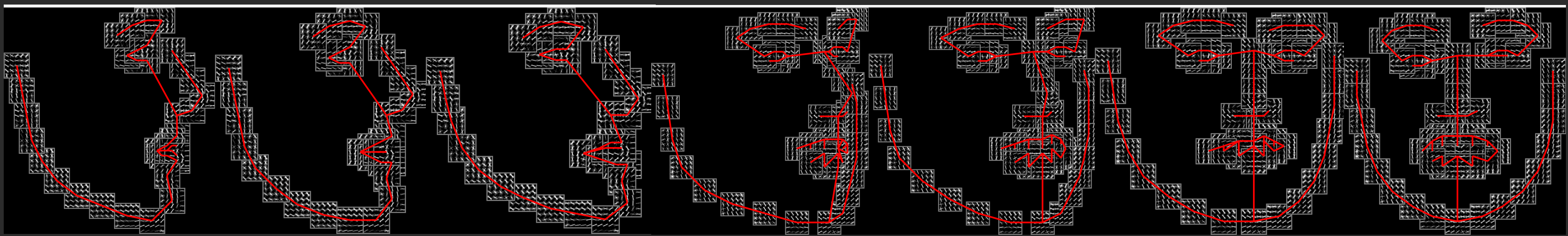
View-based models of faces

Ioffe & Forsyth, 2001

Everingham, Sivic, & Zisserman, 2006

Use global mixtures to capture topological changes due to viewpoint
Use common pool of parts

View-based models of faces

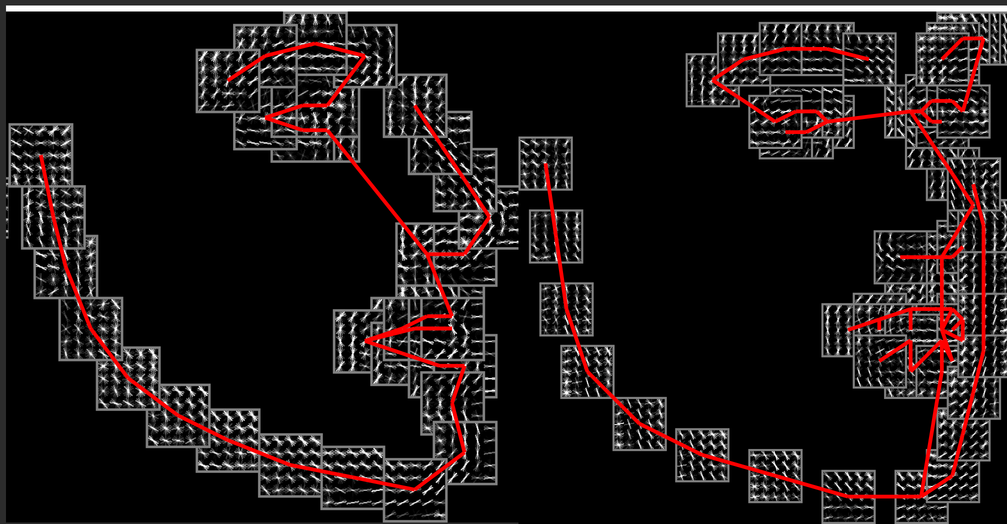


Ioffe & Forsyth, 2001

Everingham, Sivic, & Zisserman, 2006

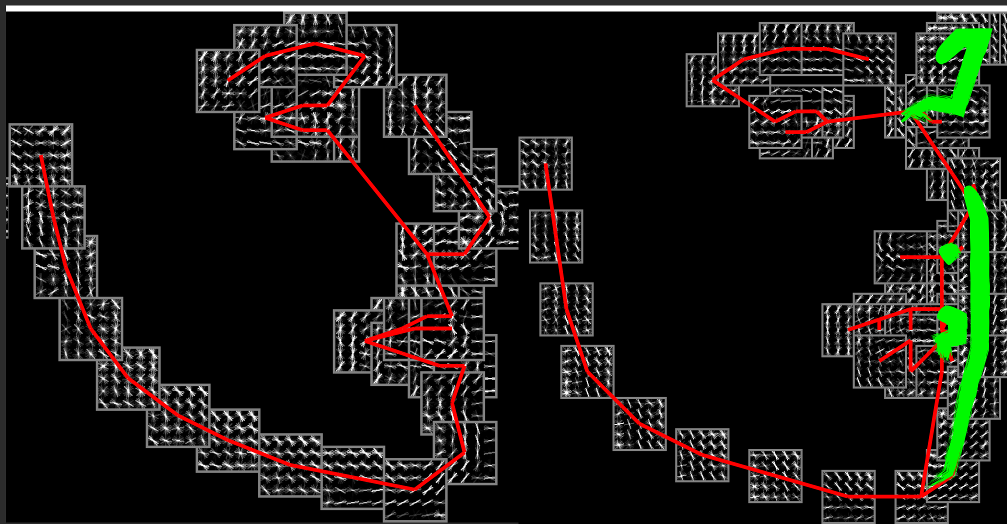
Use global mixtures to capture topological changes due to viewpoint
Use common pool of parts

View-based models



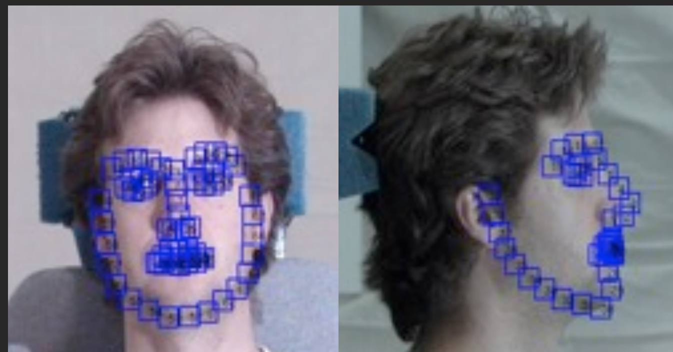
Model self-occlusion with missing branches

View-based models

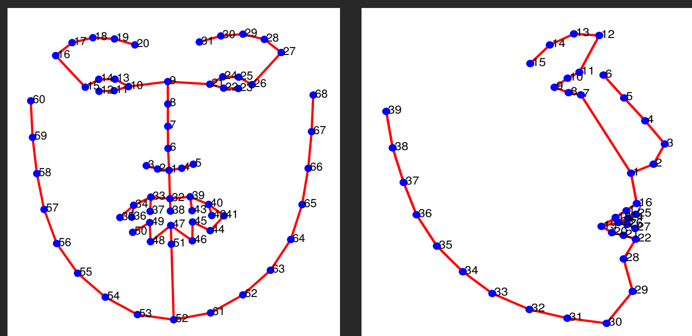


Model self-occlusion with missing branches

Learning

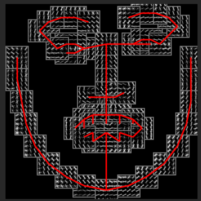


Fully-supervised dataset (CMU MultiPIE)



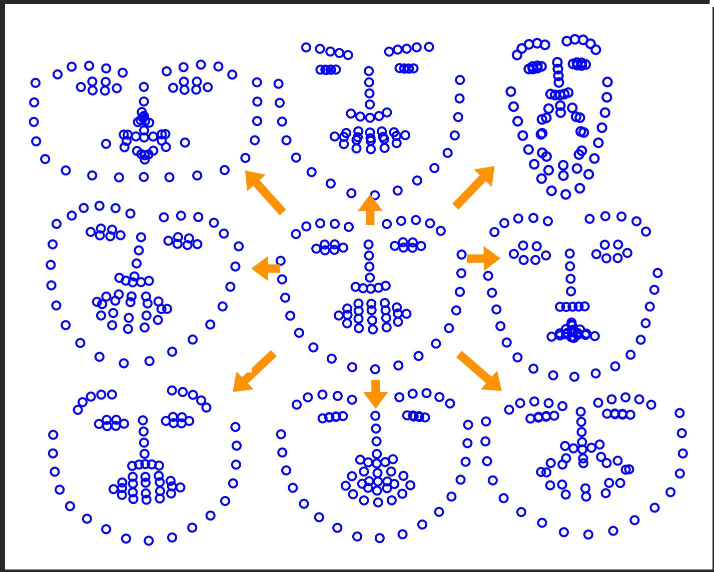
Chow-Liu algorithm

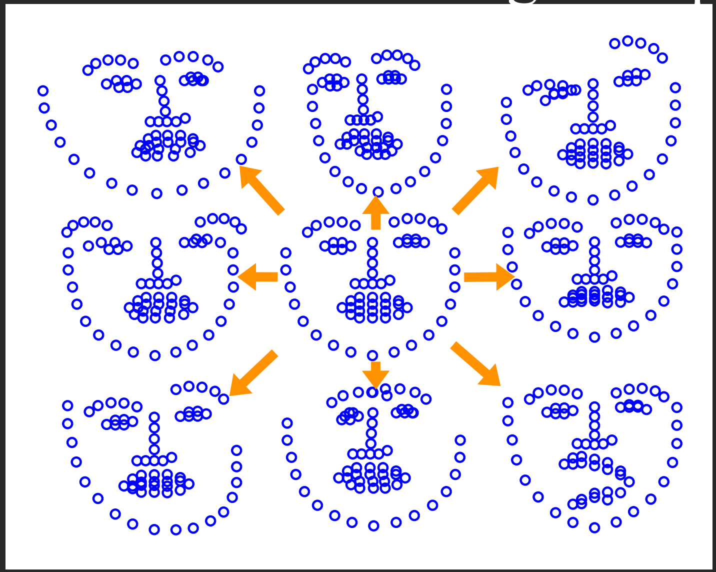
Global models of deformation



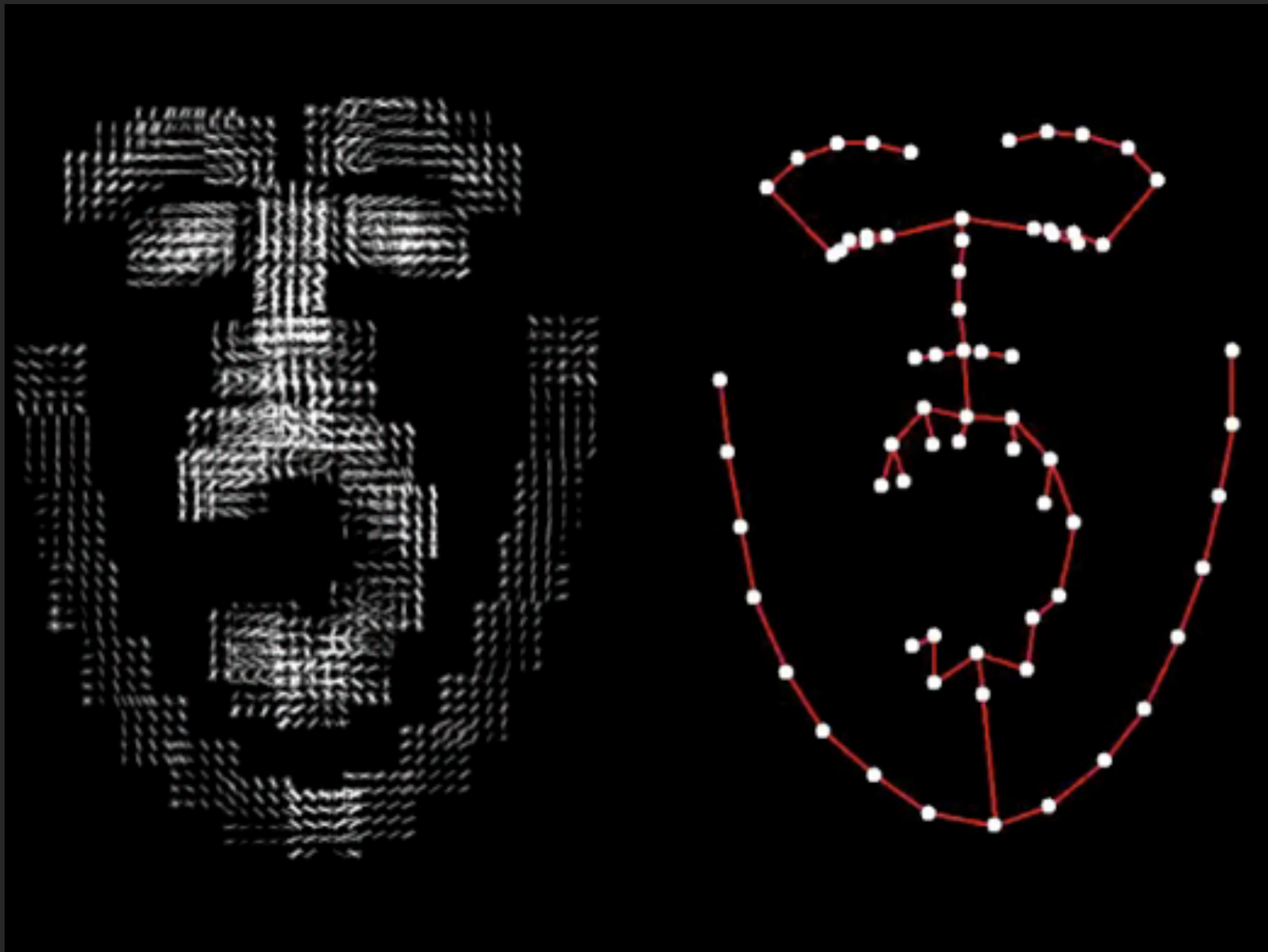
Full-covariance Gaussian shape

Tree-based max-margin shape

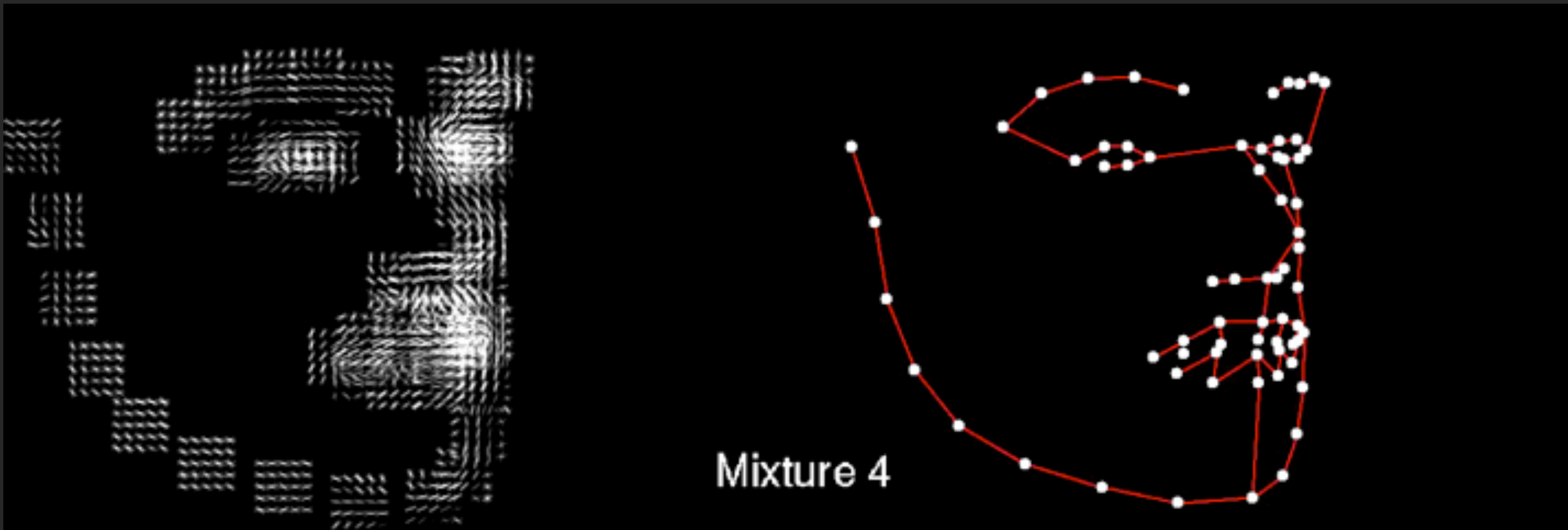




Learned appearance & deformation



Viewpoint variation



Global mixtures capture large viewpoint changes

Elastic springs capture small viewpoint changes

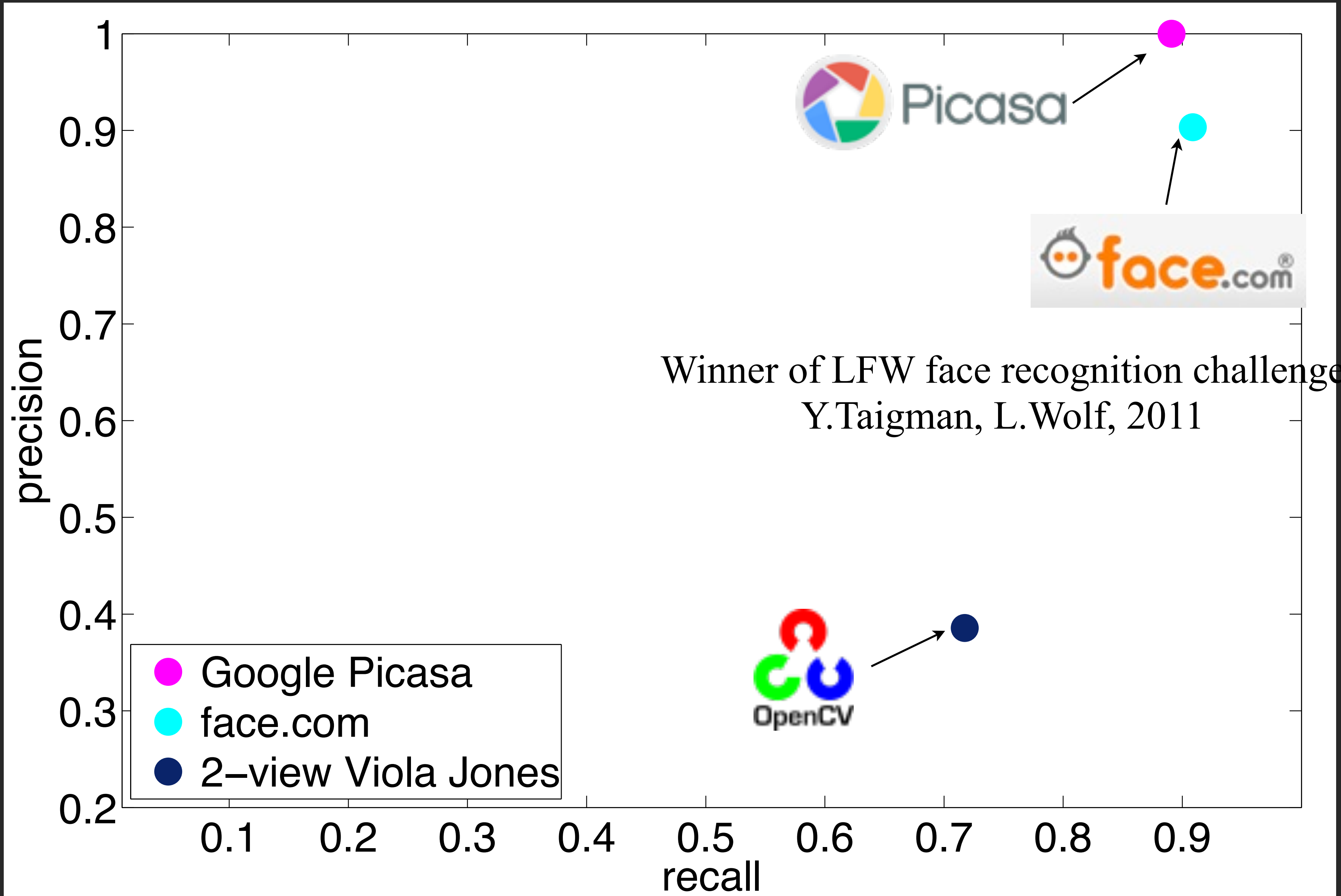
... all without explicit 3D reasoning

Evaluation on Flickr images

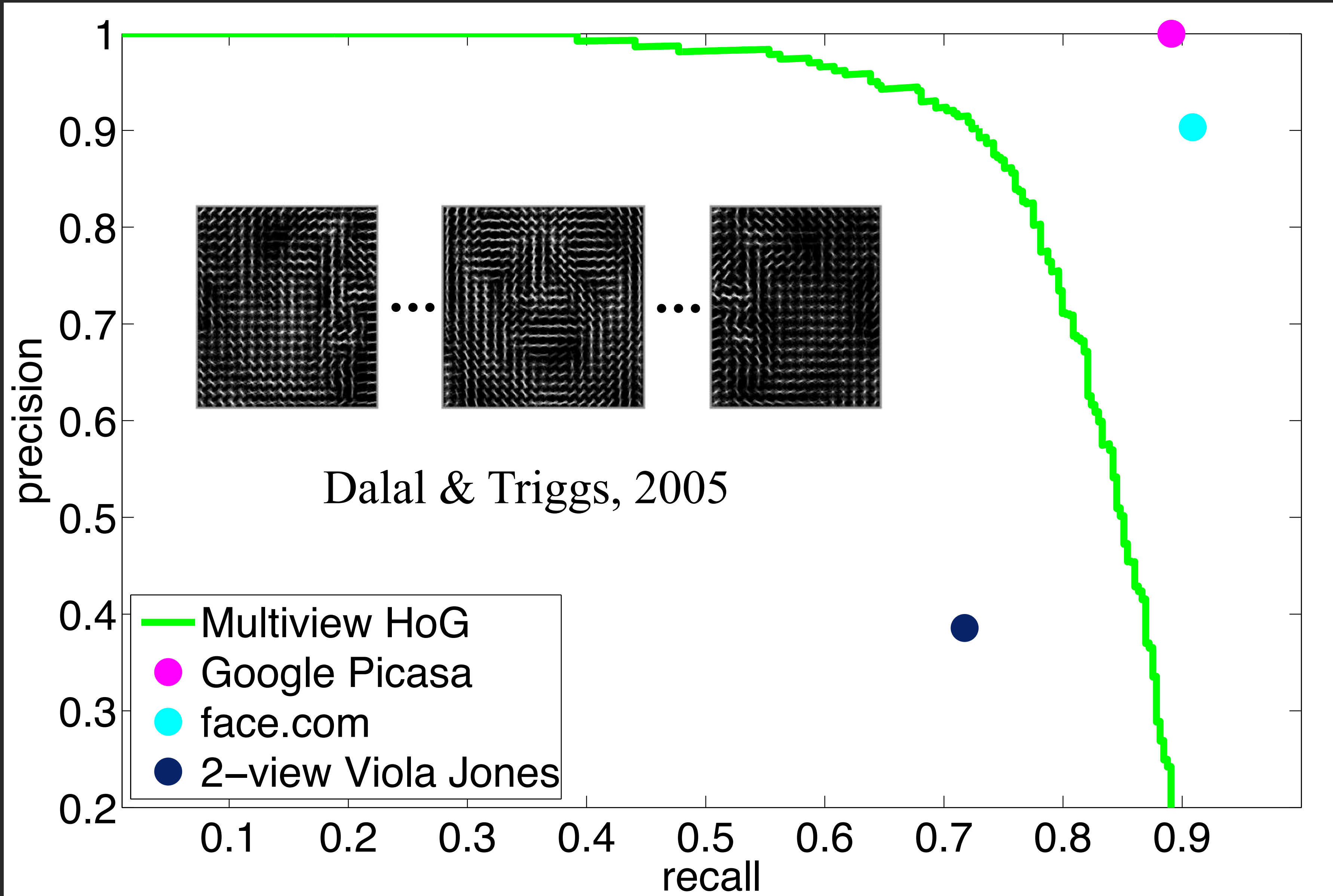
Qualitative results

Model simultaneously addresses face detection, pose estimation, and landmark localization

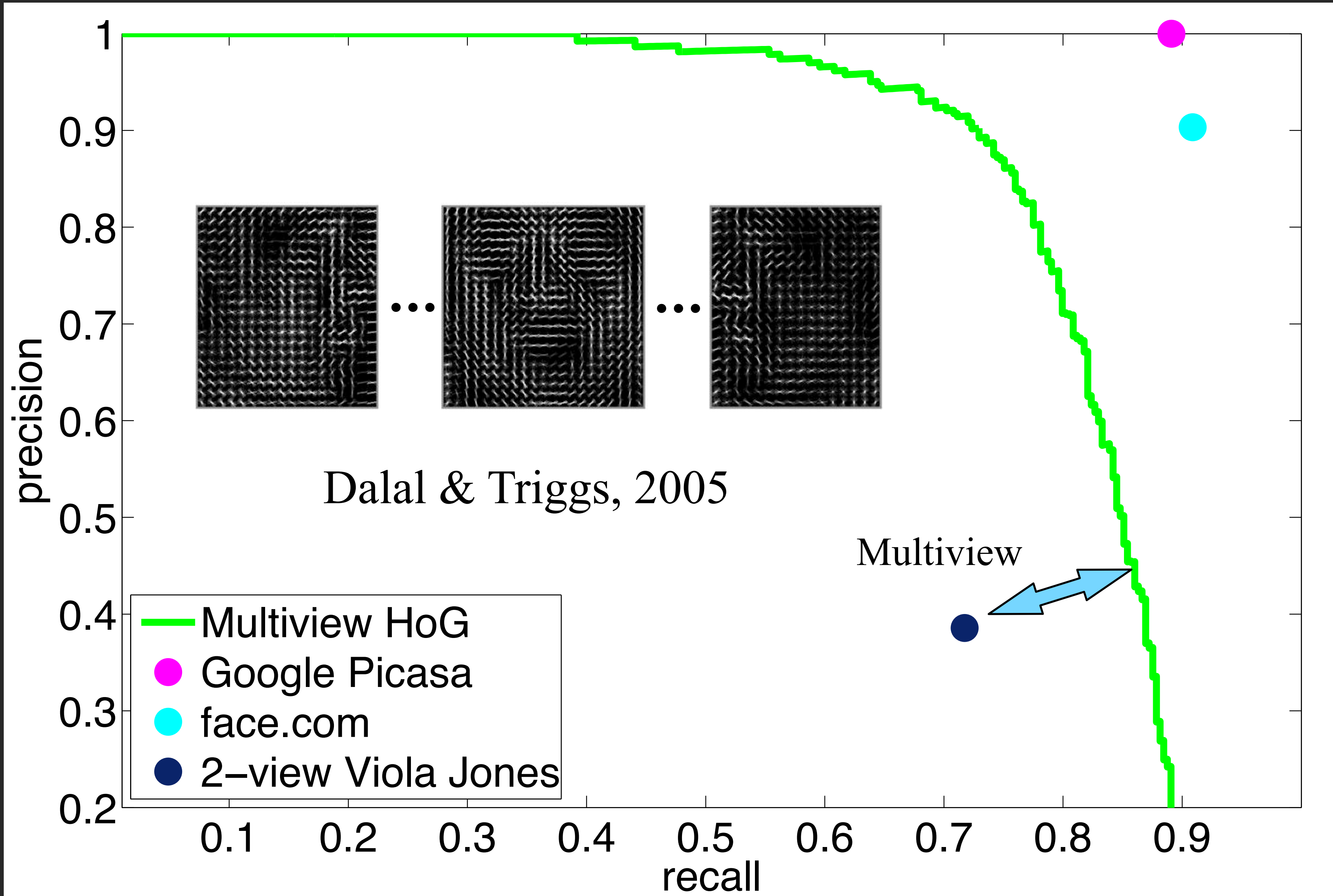
Detection results



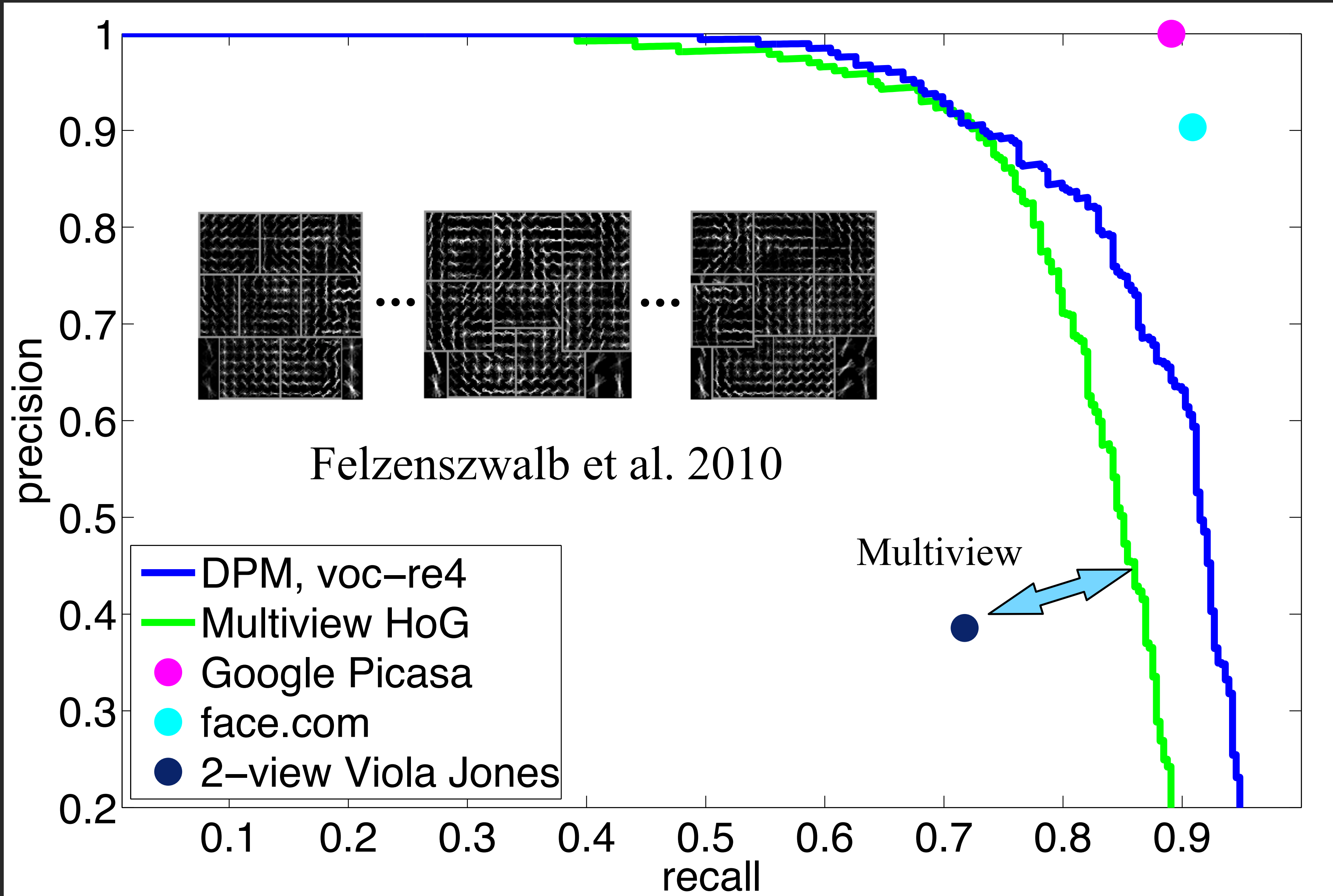
Detection results



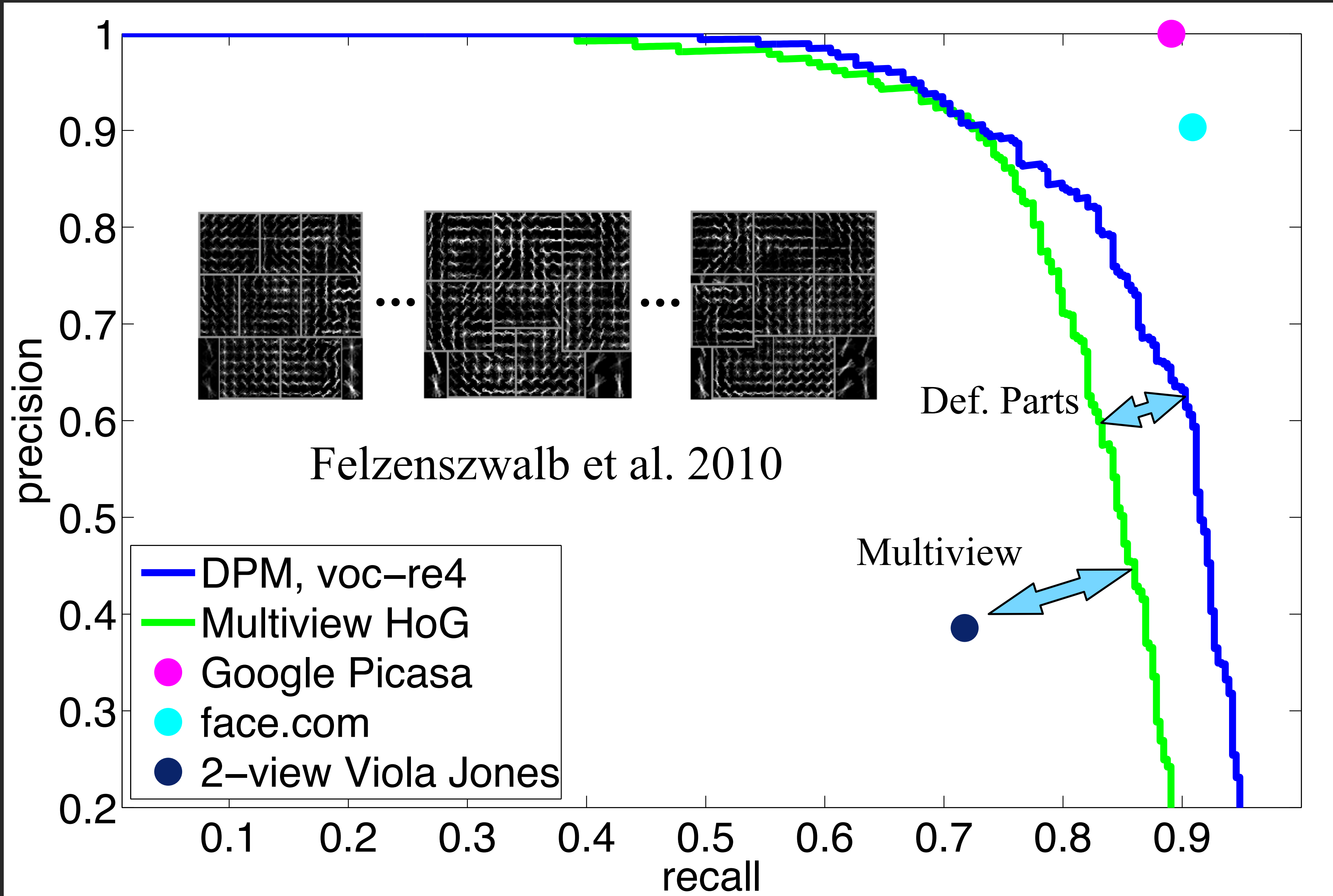
Detection results



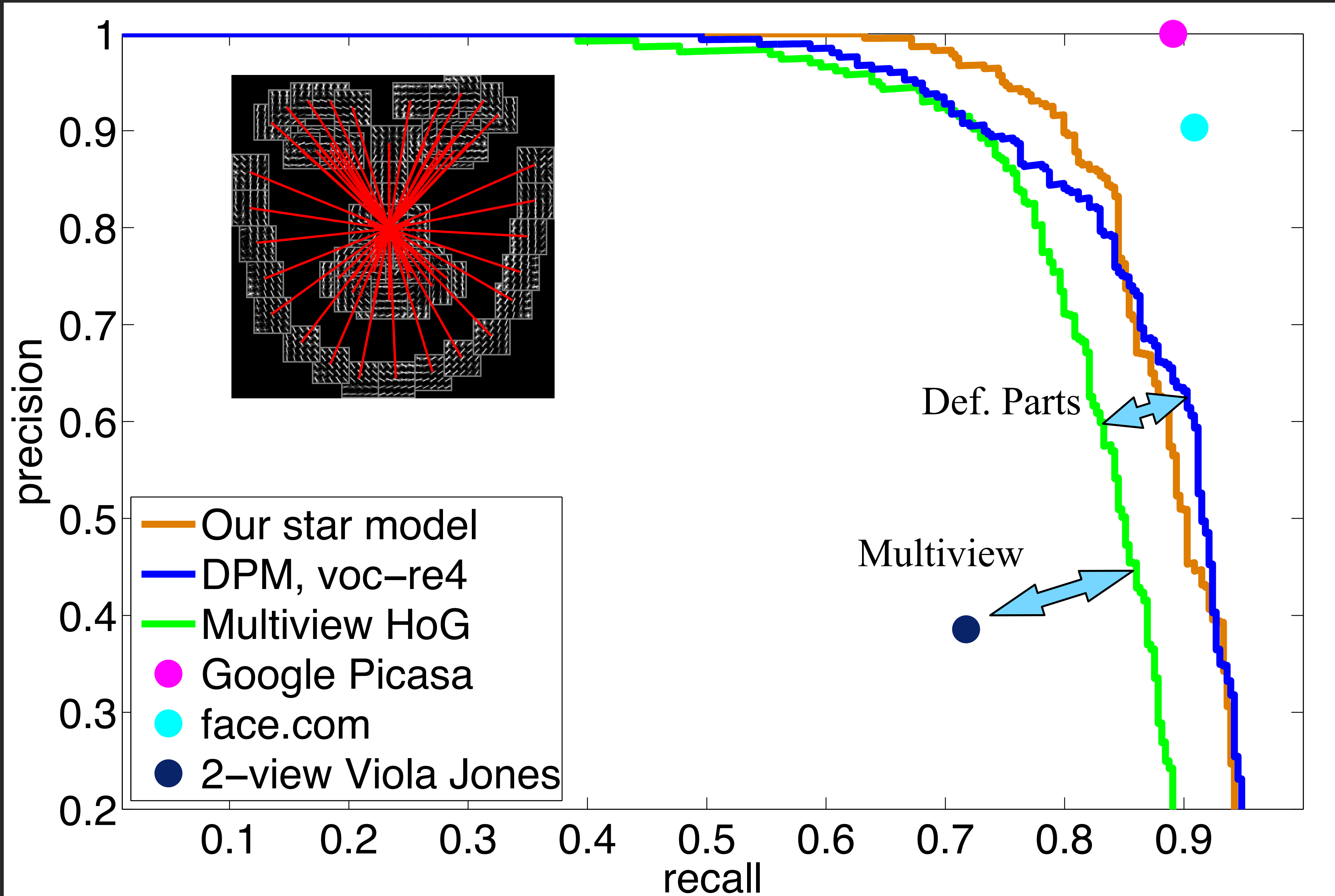
Detection results



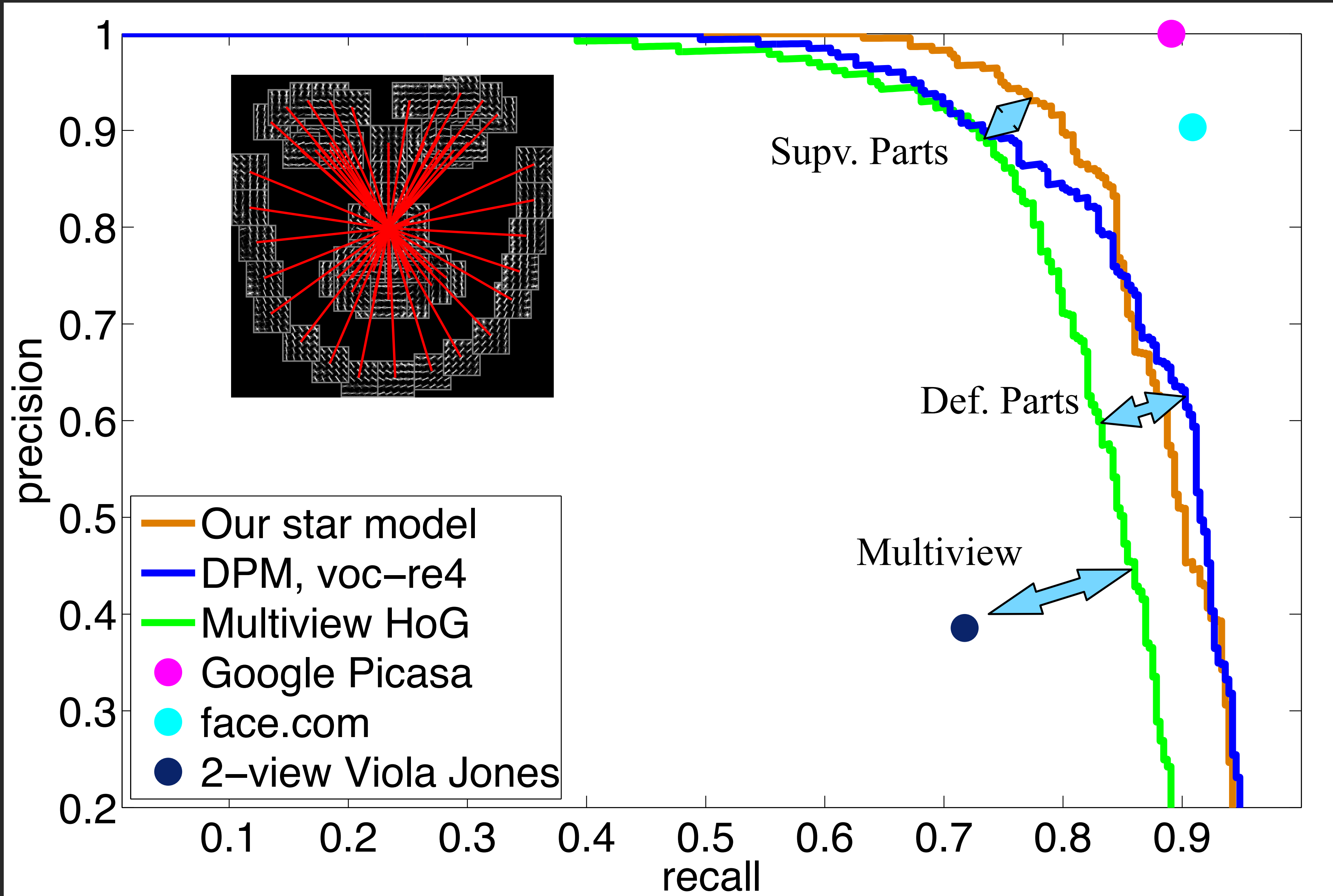
Detection results



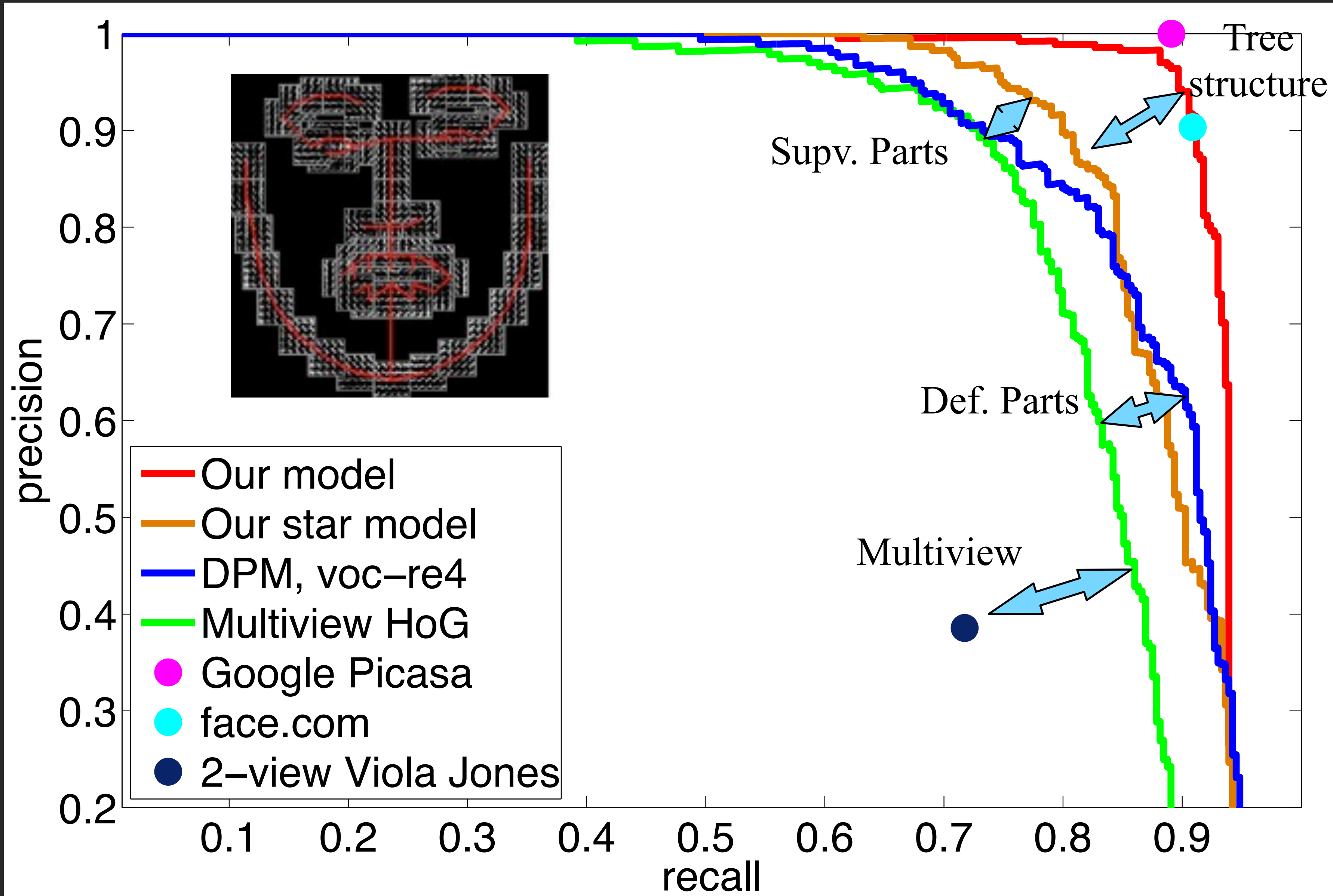
Detection results



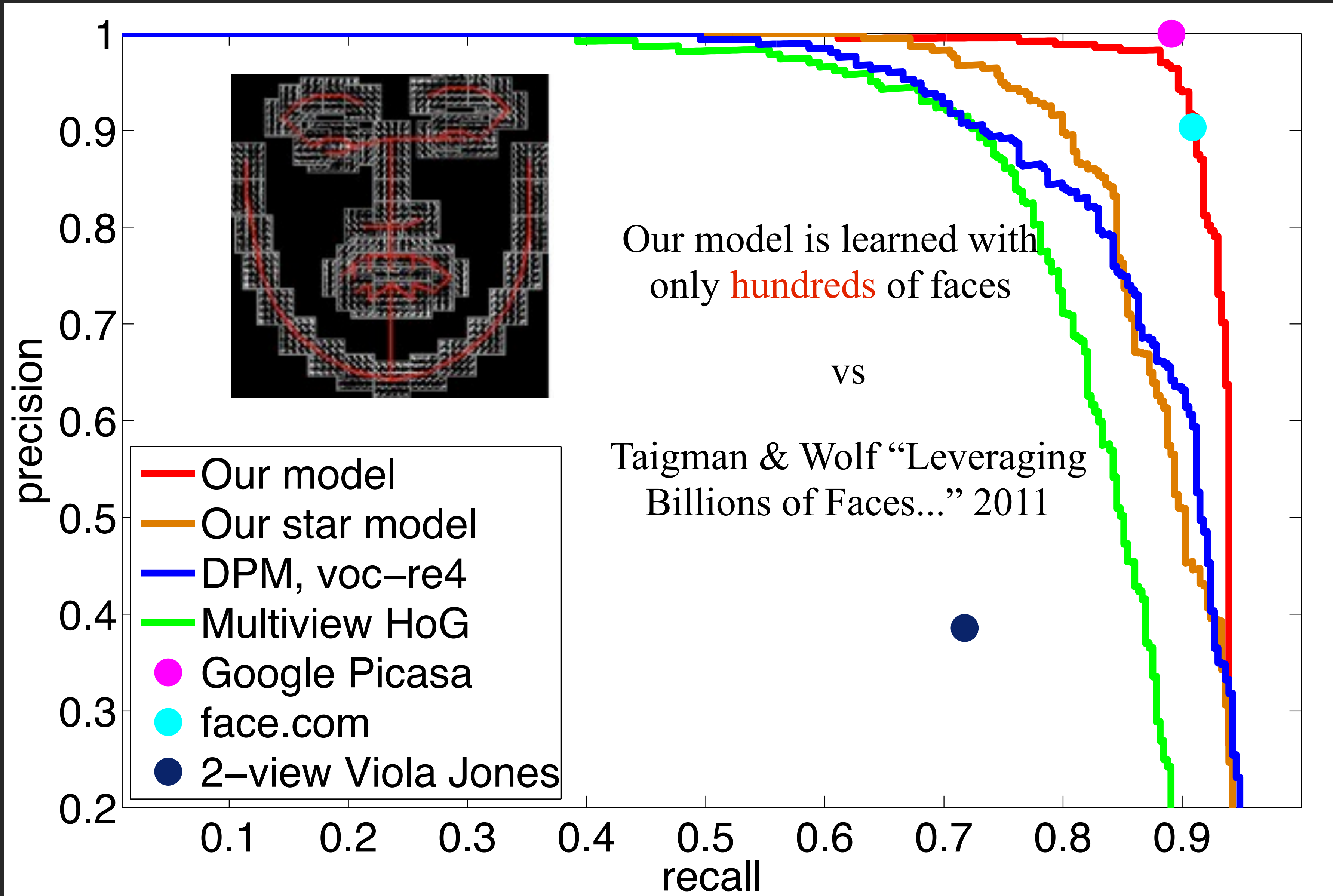
Detection results



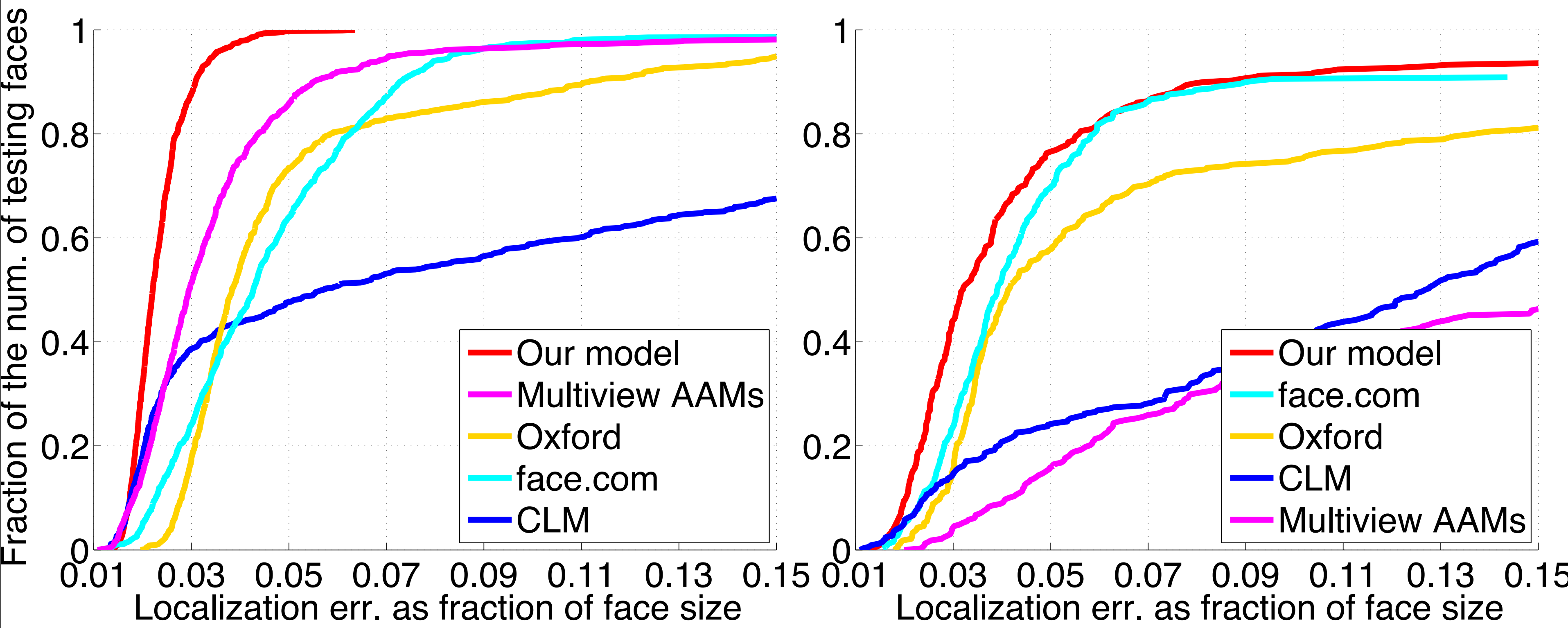
Detection results



Landmark localization



Landmark localization



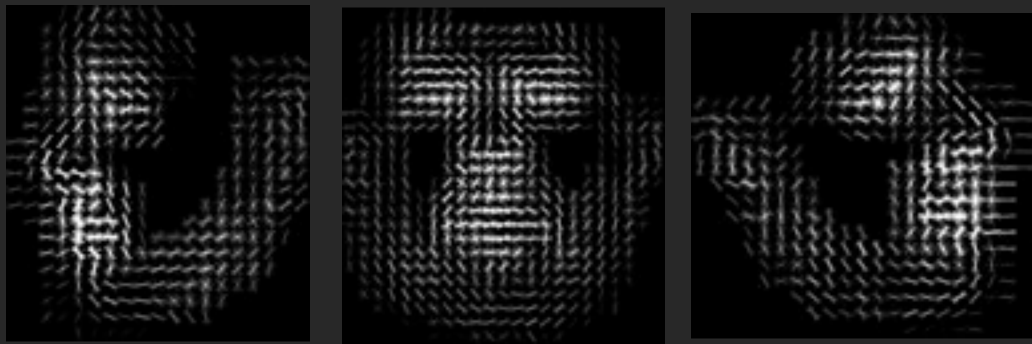
MultiPIE

Flickr

Baselines are initialized with ground truth detection on test images.

Our model naturally produces state-of-the-art pose and landmark estimates

A look back: why do part models help?



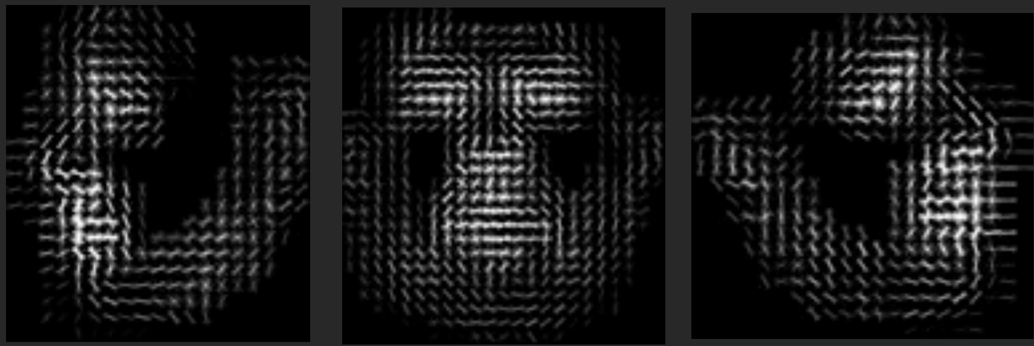
Mixtures of rigid templates

Part model

Consider a K-part model, with L discrete part locations

At run-time, part model = exponentially-large $O(L^K)$ mixture of rigid templates

A look back: why do part models help?



Mixtures of rigid templates

Part model

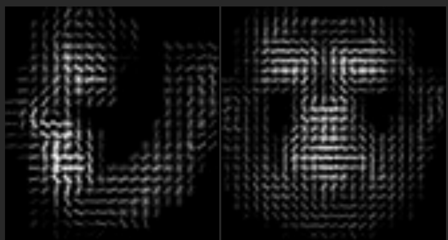
Consider a K-part model, with L discrete part locations

At run-time, part model = exponentially-large $O(L^K)$ mixture of rigid templates

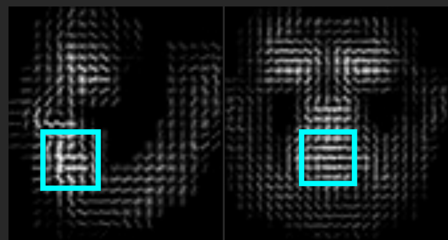
Compared to a mixture of exemplars (Malisiewicz et al), part models...

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

A look back: why do part models help?



Mixtures of rigid templates



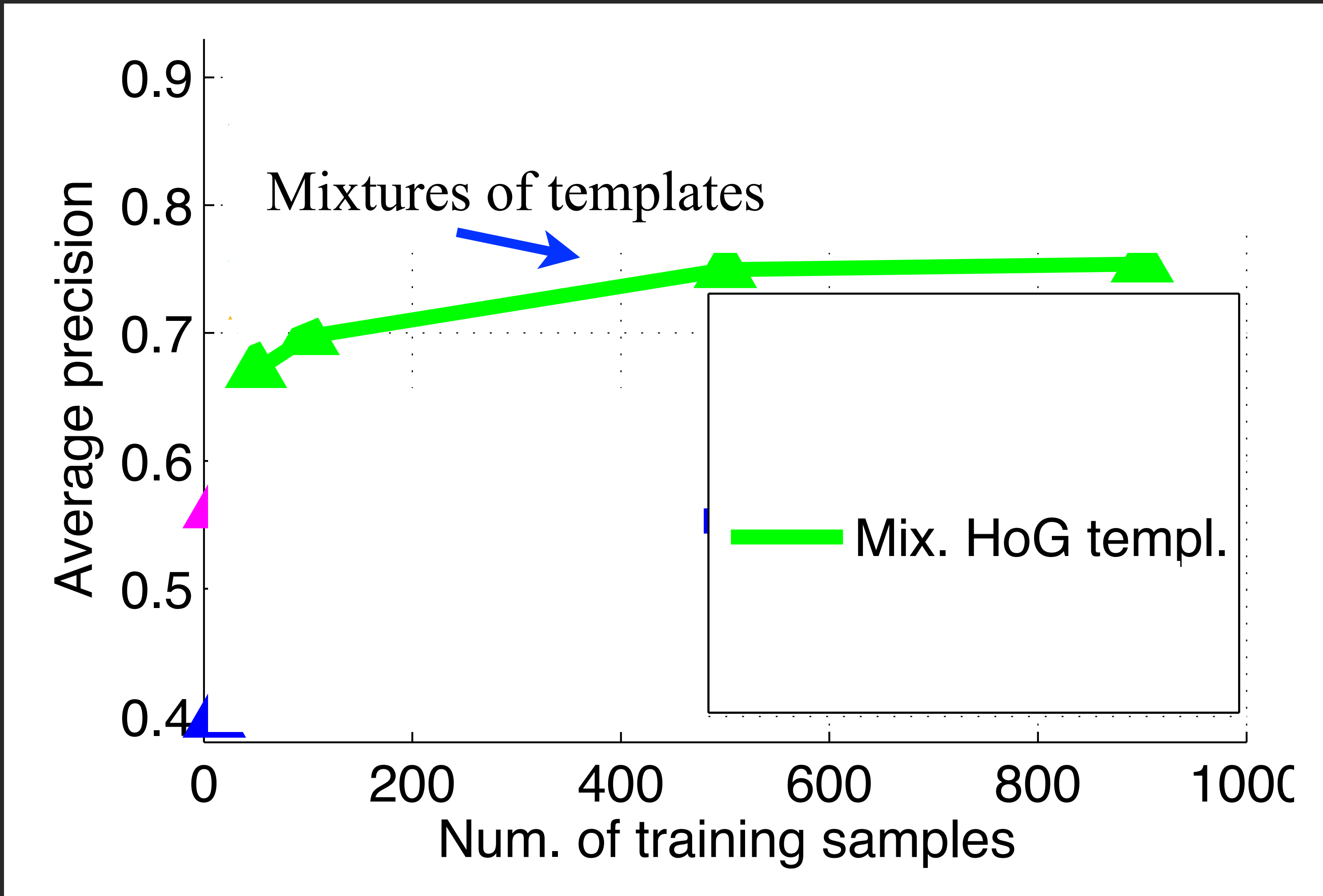
Mixtures of rigid templates with tied parameters (given by parts)

Part model

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

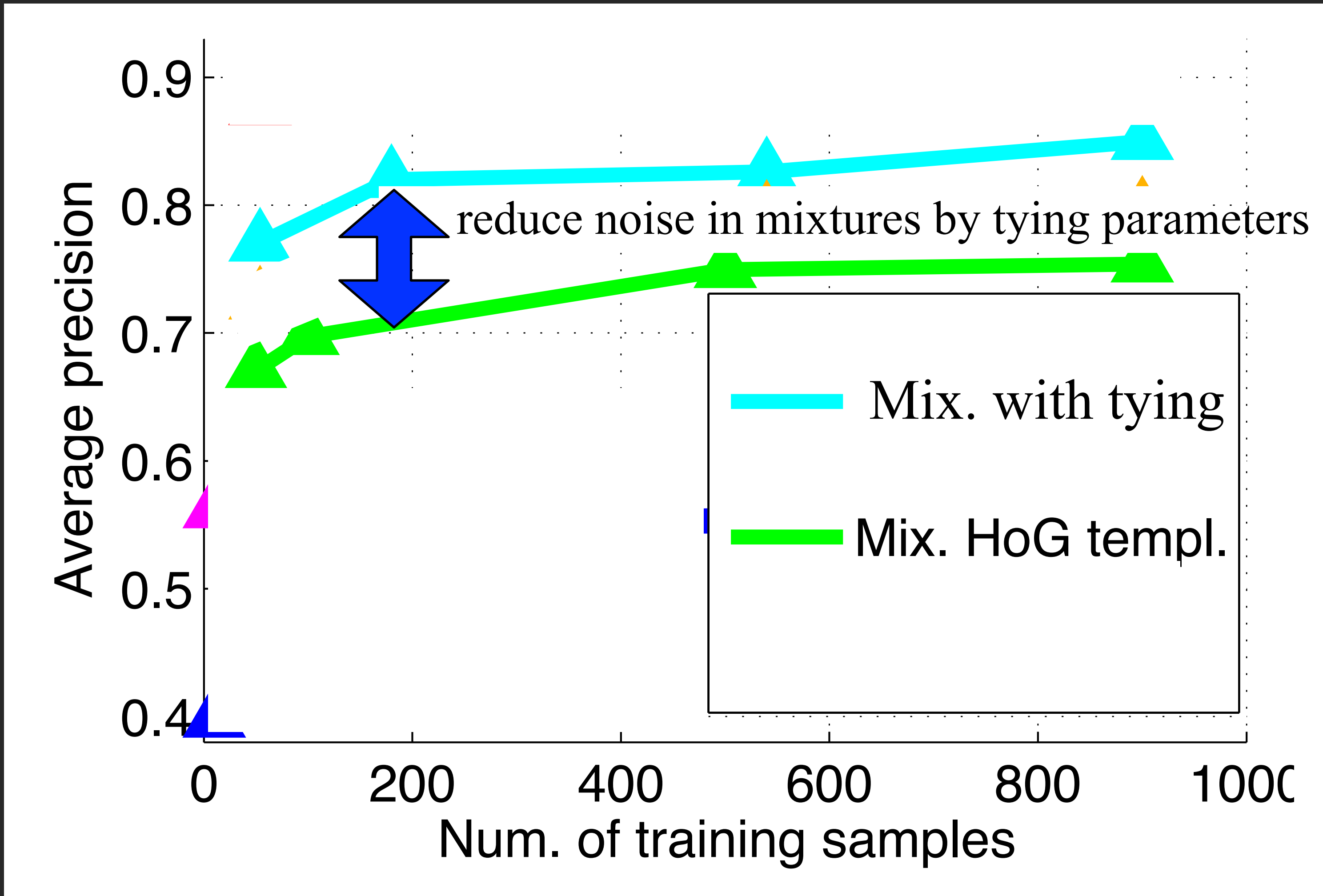
To examine (1) vs (2), lets define mixture of exemplars with sharing

An analysis of part models



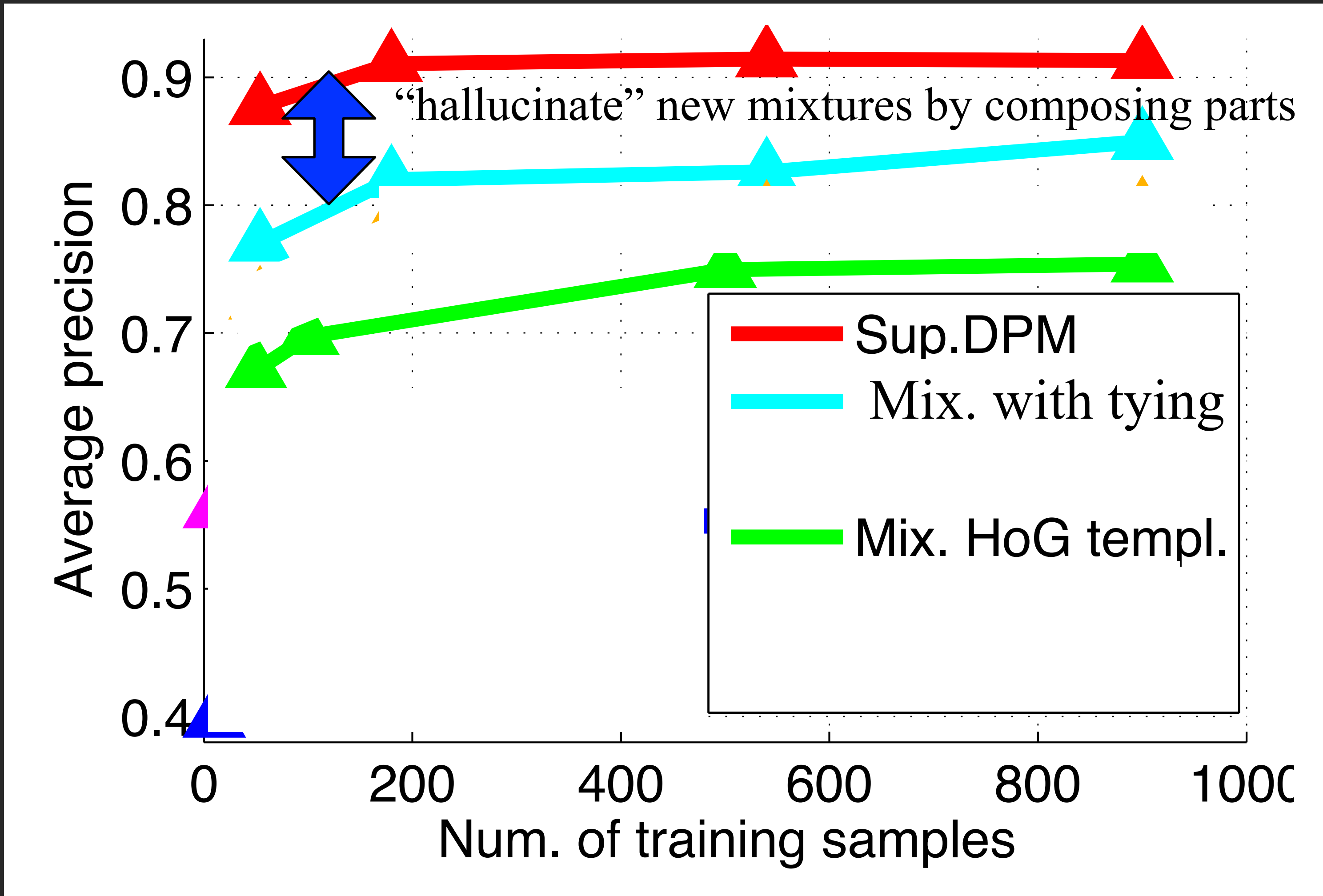
Zhu et al, BMVC 2012

An analysis of part models



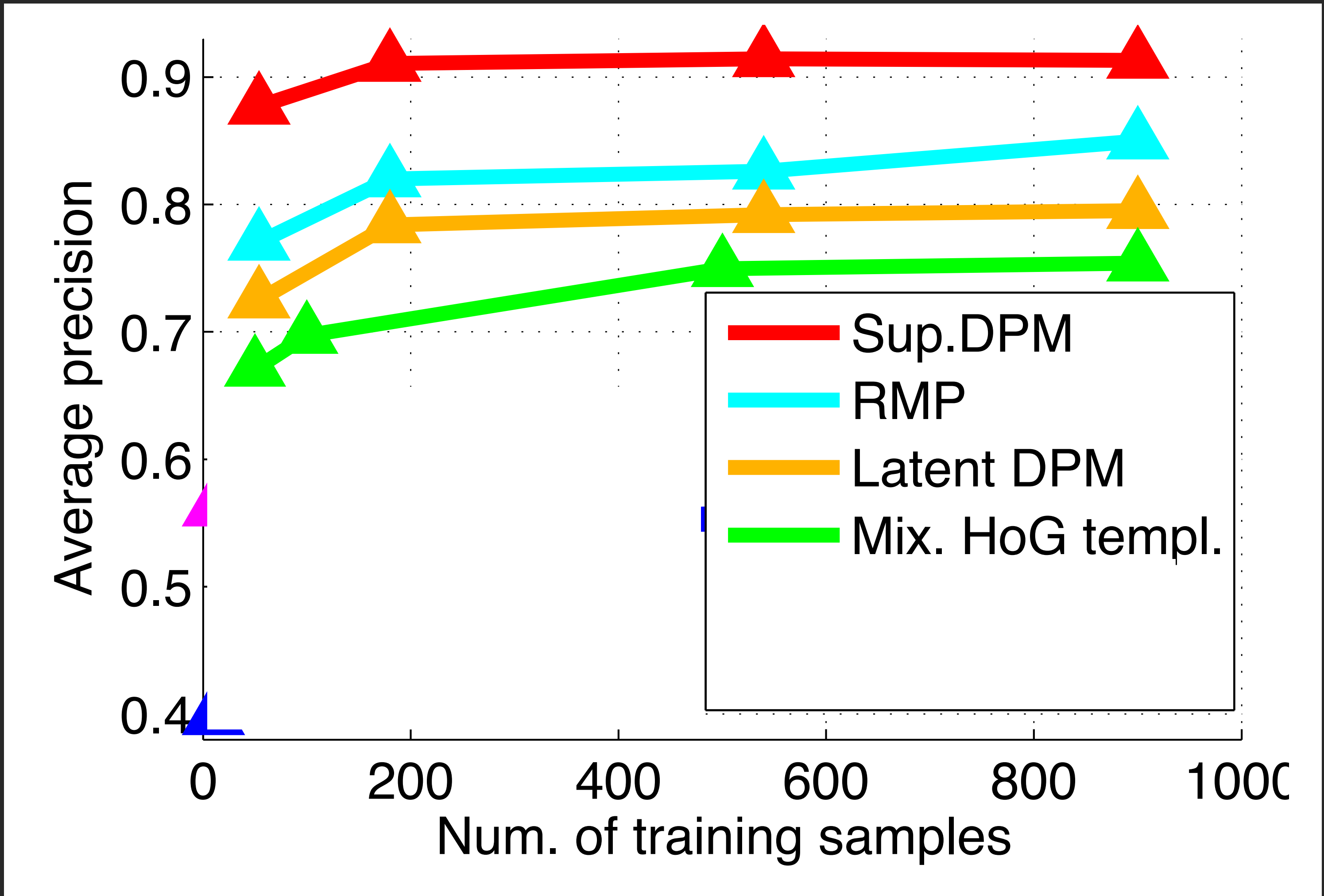
Zhu et al, BMVC 2012

An analysis of part models



Zhu et al, BMVC 2012

An analysis of part models



Zhu et al, BMVC 2012

Overview

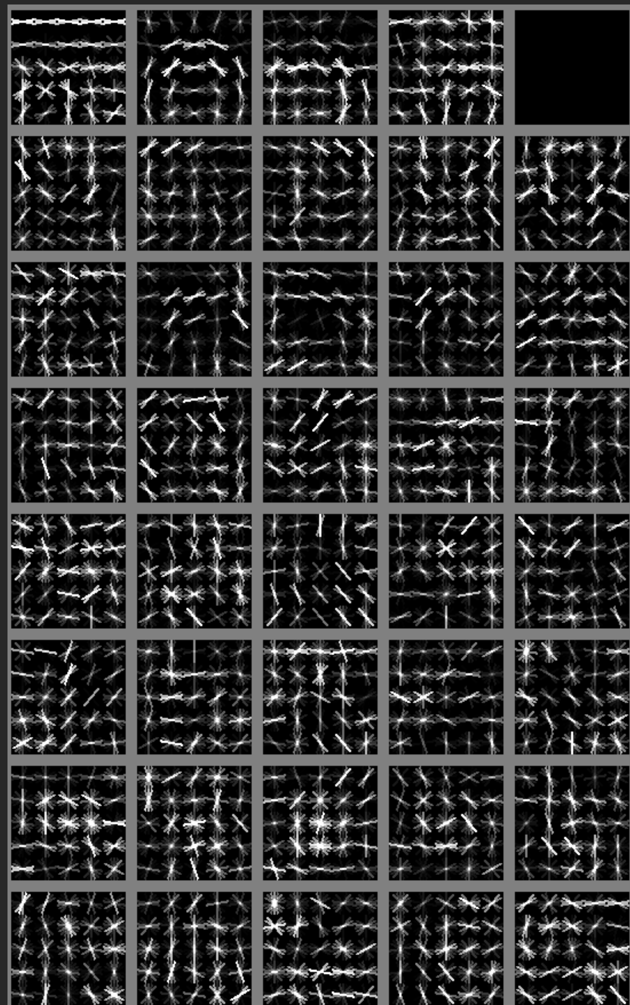
Background: part models

Occlusion reasoning

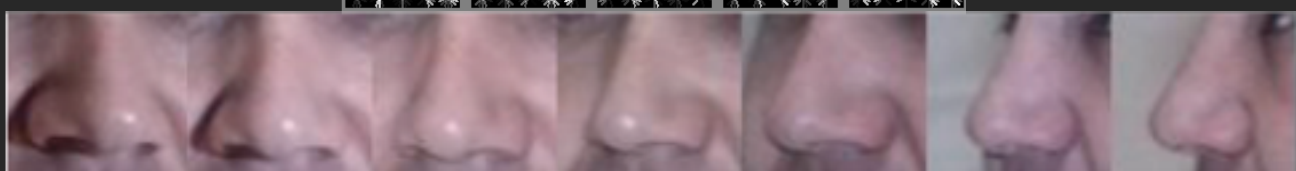
3D variation

Extensions

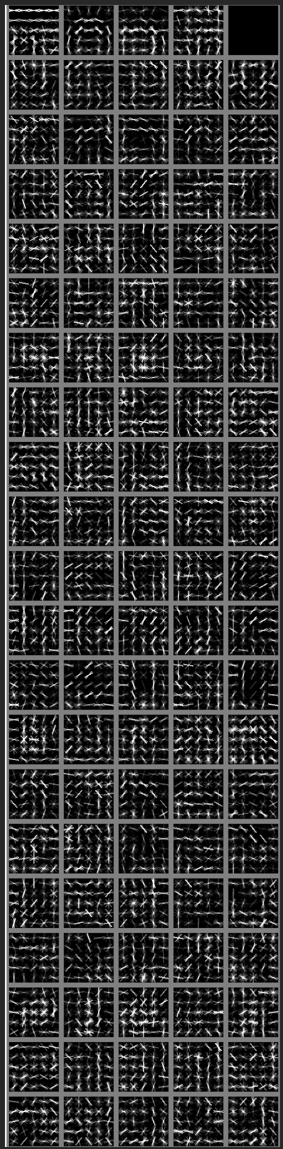
Challenges in scalability: Vocabularies of thousands of parts



Is there a more efficient representation?



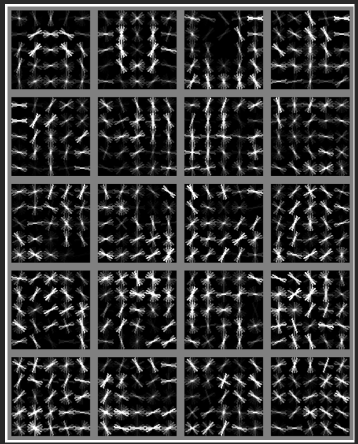
Steerable basis



Freeman, Adelson, Perona

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

$\approx$ linear combinations of basis templates

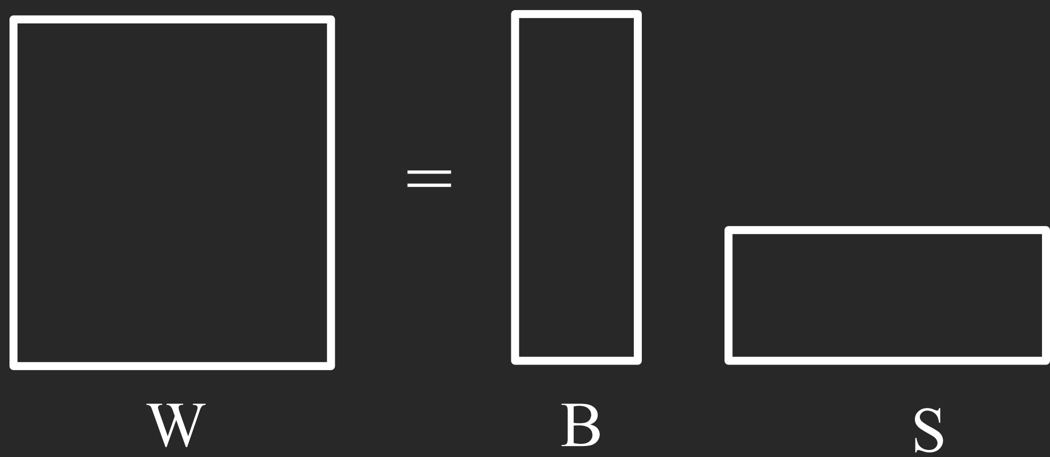


This can be implemented as a rank-restriction on original set of templates

Learning steerable part models

Learn vocabularies of thousands of parts

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



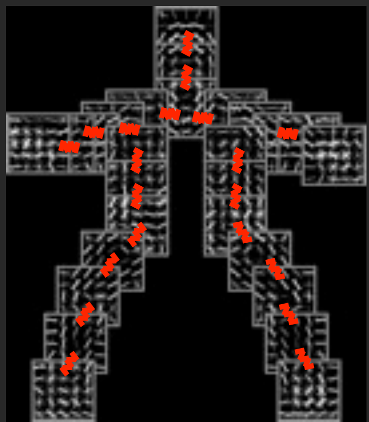
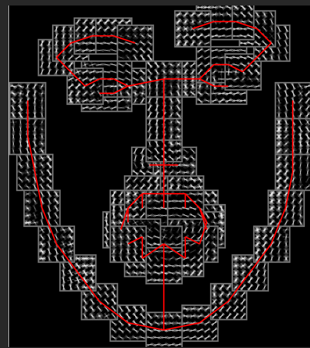
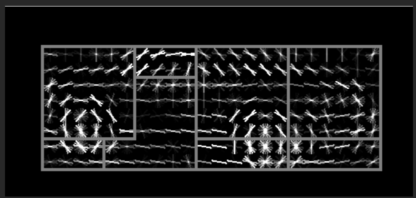
Basis

Steering Coefficients

Learn rank-constrained linear classifiers with off-the-shelf structural SVM solvers

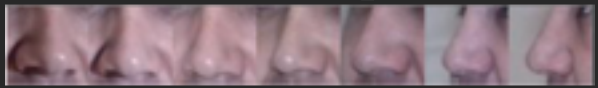
Steerable (& separable) part models

Pirsiavash & Ramanan CVPR12

Models are 10-100X 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

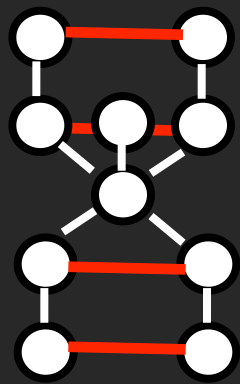
Non-tree constraints: occlusion

How to handle “loopy” constraints that arise from occlusion phenomena?

Sigal & Black CVPR 06

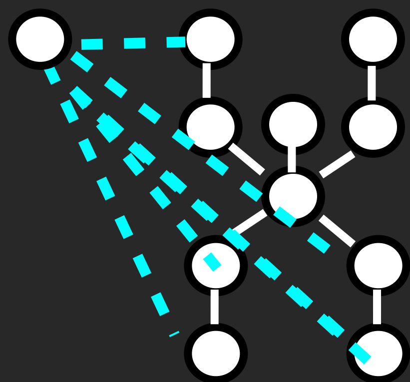
Non-tree constraints: appearance

Pairwise consistency (symmetry in appearance)



Tran & Forsyth
Mori & Malik

Global consistency (latent appearance)



Ramanan
Ferrari & Zisserman

Tools for inference on non-trees

One approach: apply standard approximate inference algorithms for Markov Random Feilds (MRFs)

Why is this hard?

1) Large discrete domains of variables (e.g., pixels in an image)

2) Continuous domains of variables (e.g., color and appearance)

Tools for inference on non-trees

One successful approach: use tree-like inference algorithms

Mixtures of trees (condition on mixture variable)

Ioffe & Forsyth, Johnson & Everginham, Lan & Huttenlocher, Wang & Mori

Loopy Belief Propagation (iteratively apply tree-based messages)

Sigal and Black

Dual Decomposition (break problem up into trees ensuring agreement)

Sapp et al, Kumar et al

Branch & Bound (use trees to generate strong lower bounds)

Tian and Scarloff, Nevatia

Sampling (importance sample from tree)

Felzenszwalb & Huttenlocher, Beuhler et al

N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex (loopy) model

Popular in speech, but why not vision?

N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex (loopy) model

Popular in speech, but why not vision?

N-best maximal decoding

Use max-marginals + NMS to compute the "next-best non-overlapping pose"

Park and Ramanan, ICCV11

Yadollahpour et al. ECCV12

N-best maximal decoding

Intuition: backtrack from all parts, not just root

(can we done without any noticeable increase in computation)

Park and Ramanan, ICCV 2011

N-best maximal decoding

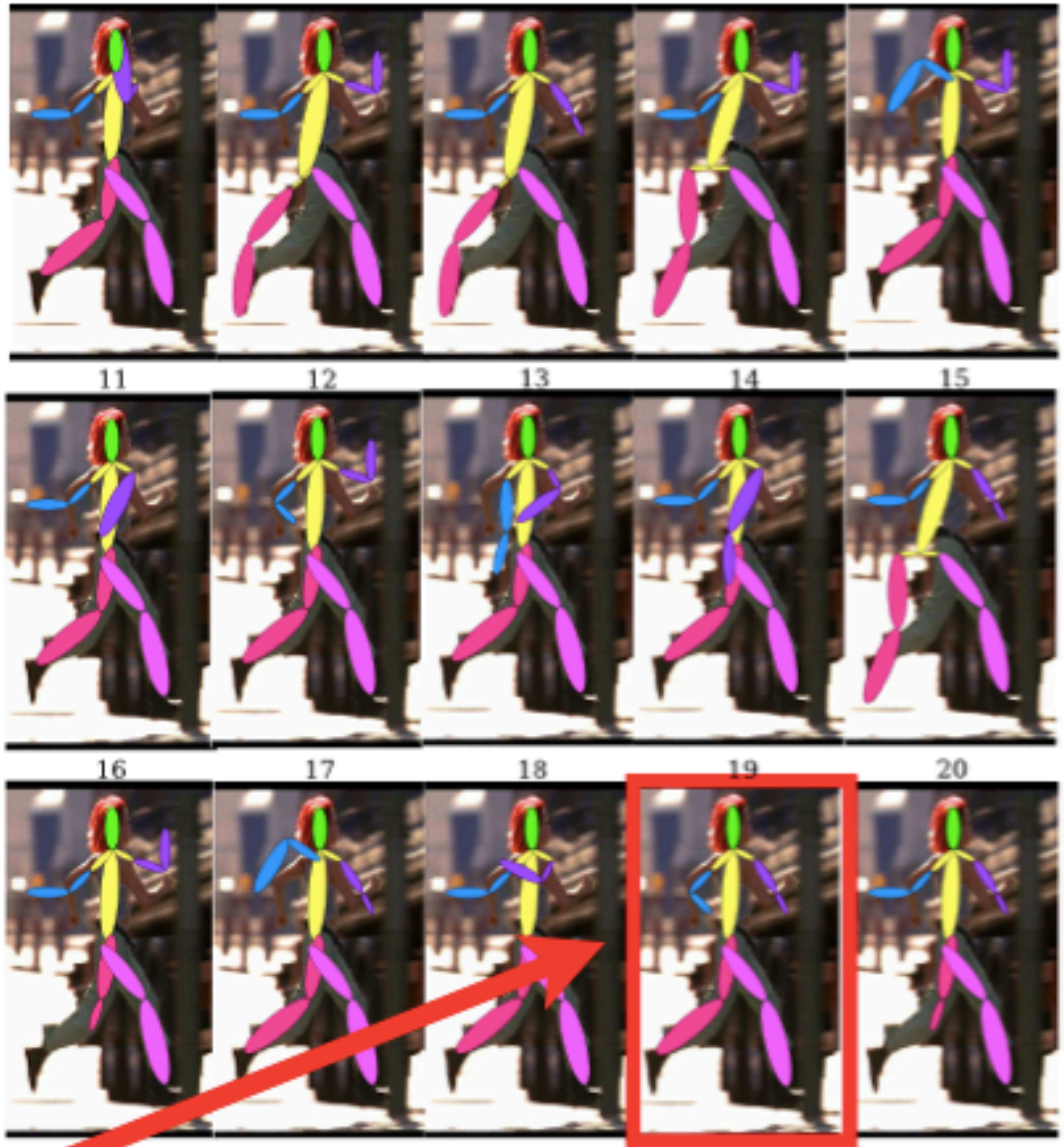
Philosophy: Delay hard decisions as much as possible

~~Candidate interest points~~

~~Candidate parts~~

Candidate poses

Maximal poses from a single frame



Correct one picked out by temporal context (tracker)

Evaluation

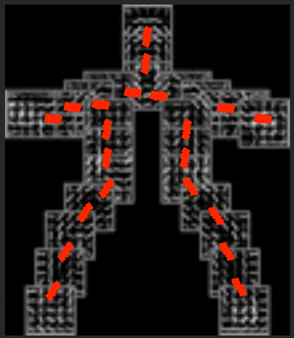
Percentage of correct frames

Algorithms	walking	pitching	lola1	lola2
noNMS	0.825	0.762	0.505	0.445
rootNMS	0.815	0.741	0.455	0.390
partNMS	0.825	0.762	0.515	0.420
MMsmpl	0.930	0.800	0.645	0.440
Nbest(all)	0.940	**0.800**	0.635	0.495
Nbest(limb)	**0.950**	0.797	**0.670**	**0.500**

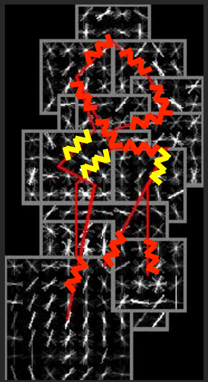
Outperforms standard approaches by 20%

Just as fast as finding single-best configuration

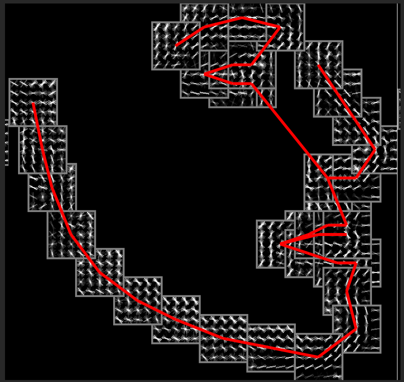
A look back



Articulation



Visual composites



3D aspect

Underlying theme: tractable, joint representations of shape and appearance

Thank you!