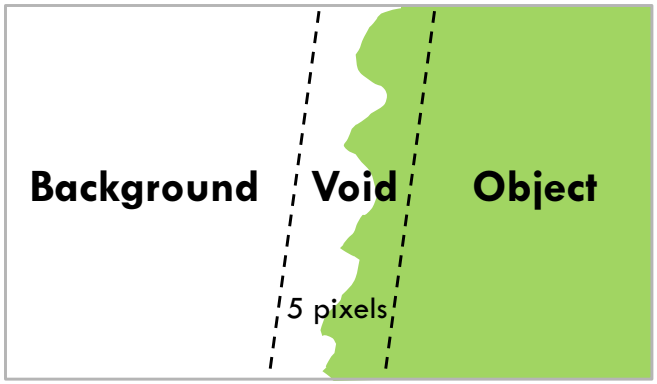
Segmentation challenge

- For each pixel, predict the class of the object containing that pixel or 'background'.

- Competition 5: Train on the supplied data
  - Which methods perform best given specified training data?
  - Can use bounding box data as well as seg. data
- Competition 6: Train on any (non-test) data
  - Available since VOC2009
  - Allows for use of own data

Annotation

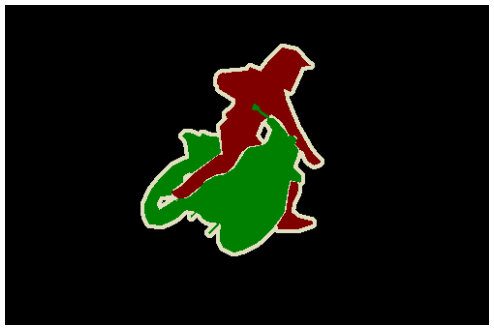
- Annotation in one session with written guidelines
  - Segmentation is 'refinement' of bounding box (but may go outside it)
  - Segmentation accurate to within 5-pixel boundary region which is marked 'void'



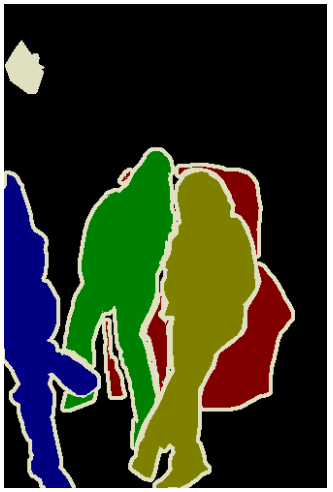
  - 1-pixel wide structures (whiskers, wires) can be ignored
  - Surface objects considered part of the object (e.g. items on a table)

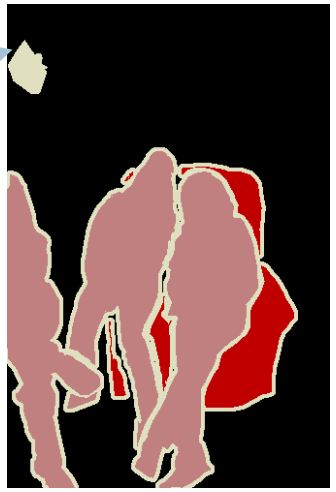
Example annotations

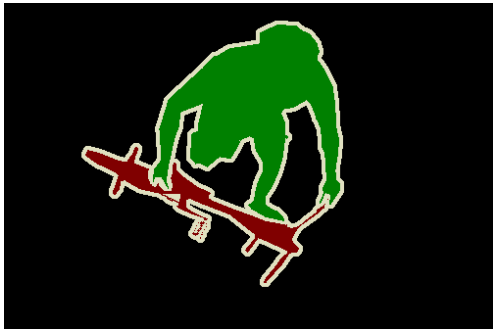
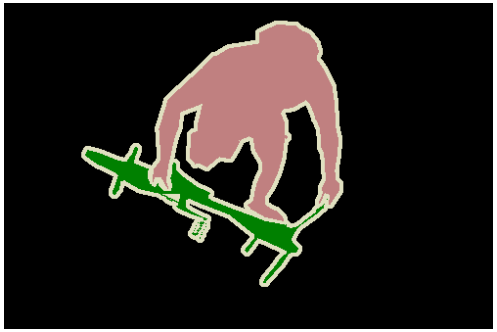
Image | **Object segmentation** | **Class segmentation**







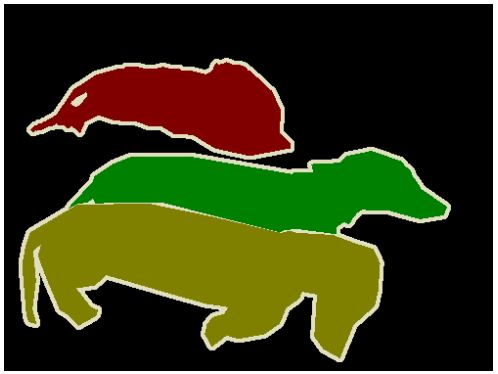
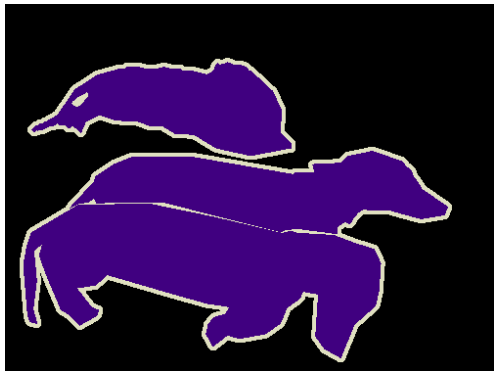




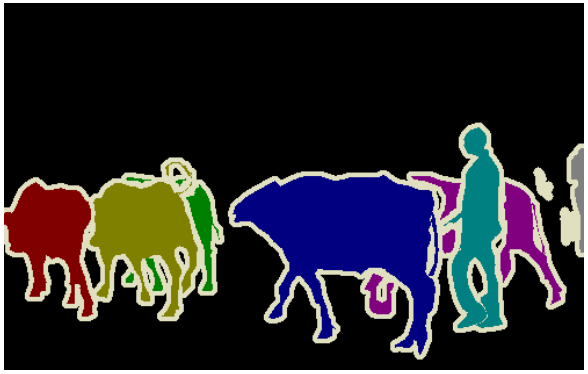
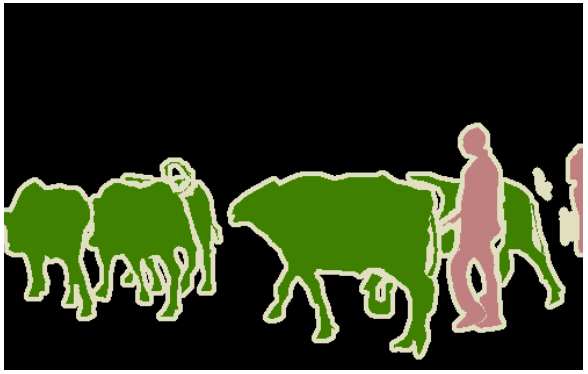
Example annotations

Image | **Object segmentation** | **Class segmentation**





Dataset statistics

- Contains VOC2008/9/10/11 data as subsets
- Around 40% increase in objects over VOC2011

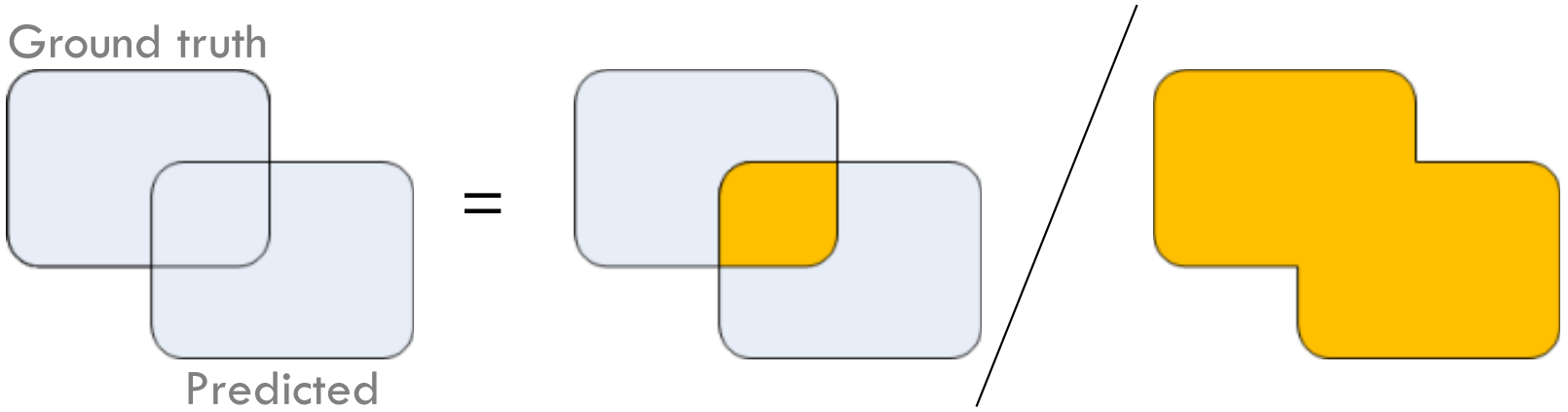
	Training		Testing	
Images	2,913	(2,223)	1,456	(1,111)
Objects	6,934	(5,034)	3,066	(2,028)

VOC2011 counts shown in brackets

- Training and test together make exactly 10,000 precisely segmented objects!

Evaluation Metric

Intersection/union of **class** labels $= \dfrac{\text{true pos. class}}{\text{true pos.} + \text{false pos.} + \text{false neg.}}$



- Metric chosen because:
  - Allows per-class participation
  - Penalises both over- and under-estimates
- Overall evaluation metric is average over all classes (including background)

Methods

- 5 entries (comp5) and 3 entries (comp6) from three institutions
- Top methods:
  - Run detection first and refine bounding boxes
    - Superpixel-based Markov Random Field
    - Colour and detector output used as unary
      [Segmentation over Detection by Coupled Global and Local Sparse Representations, ECCV 2012]
  - Extract multiple (overlapping) segmentations
    - Sample consistent (non-overlapping) tilings
    - Region descriptors passed through SVR + combined using second order pooling method to predict class

Example segmentations

Image

Ground truth

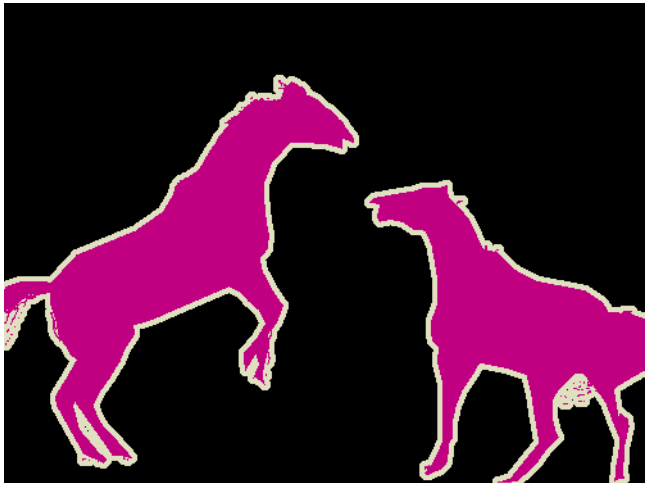
BONN_O2PCPMC_FGT_SEGM

NUS_DET_SPR_GC_SP

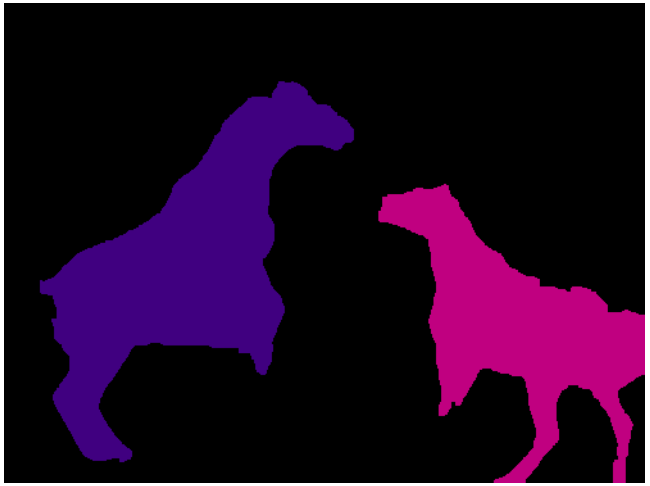
Example segmentations

Image

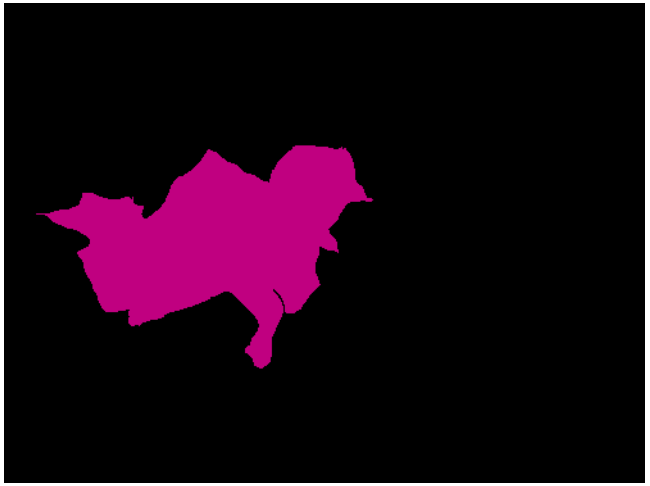
Ground truth



BONN_O2PCPMC_FGT_SEGM



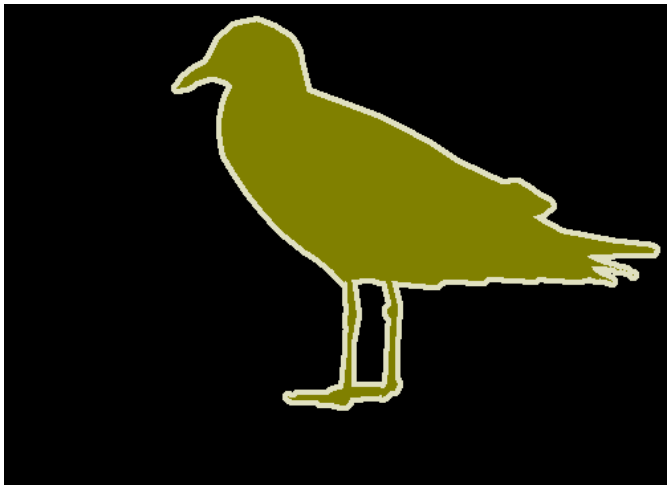
NUS_DET_SPR_GC_SP



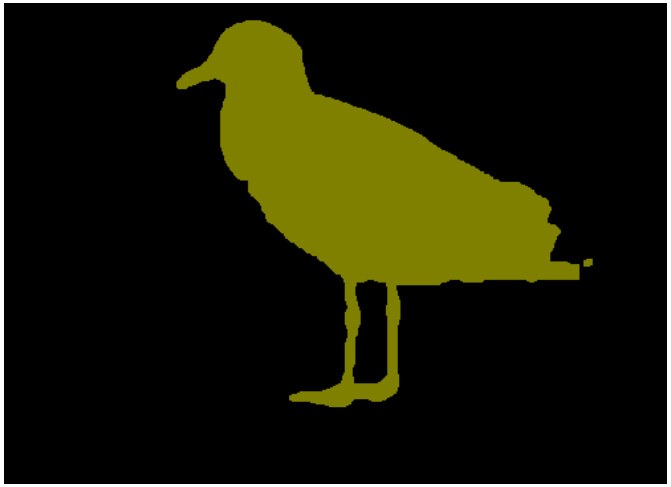
Example segmentations

Image

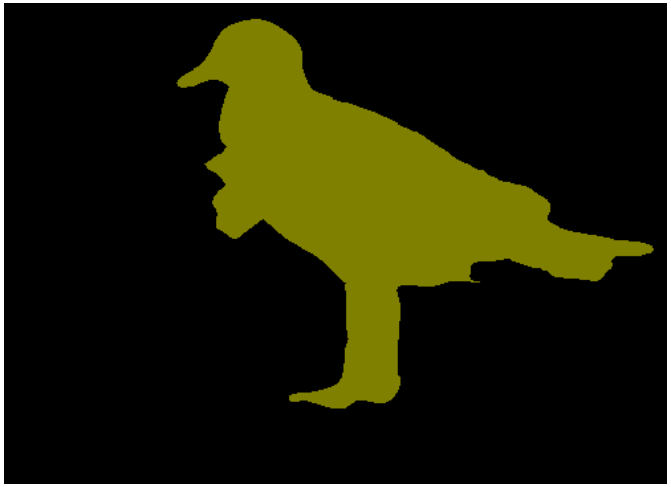
Ground truth



BONN_O2PCPMC_FGT_SEGM



NUS_DET_SPR_GC_SP



Example Segmentations

Image

Ground truth

BONN_O2PCPMC_FGT_SEGM

NUS_DET_SPR_GC_SP

Accuracy by class/method

Trained on VOC2011 data (comp5)

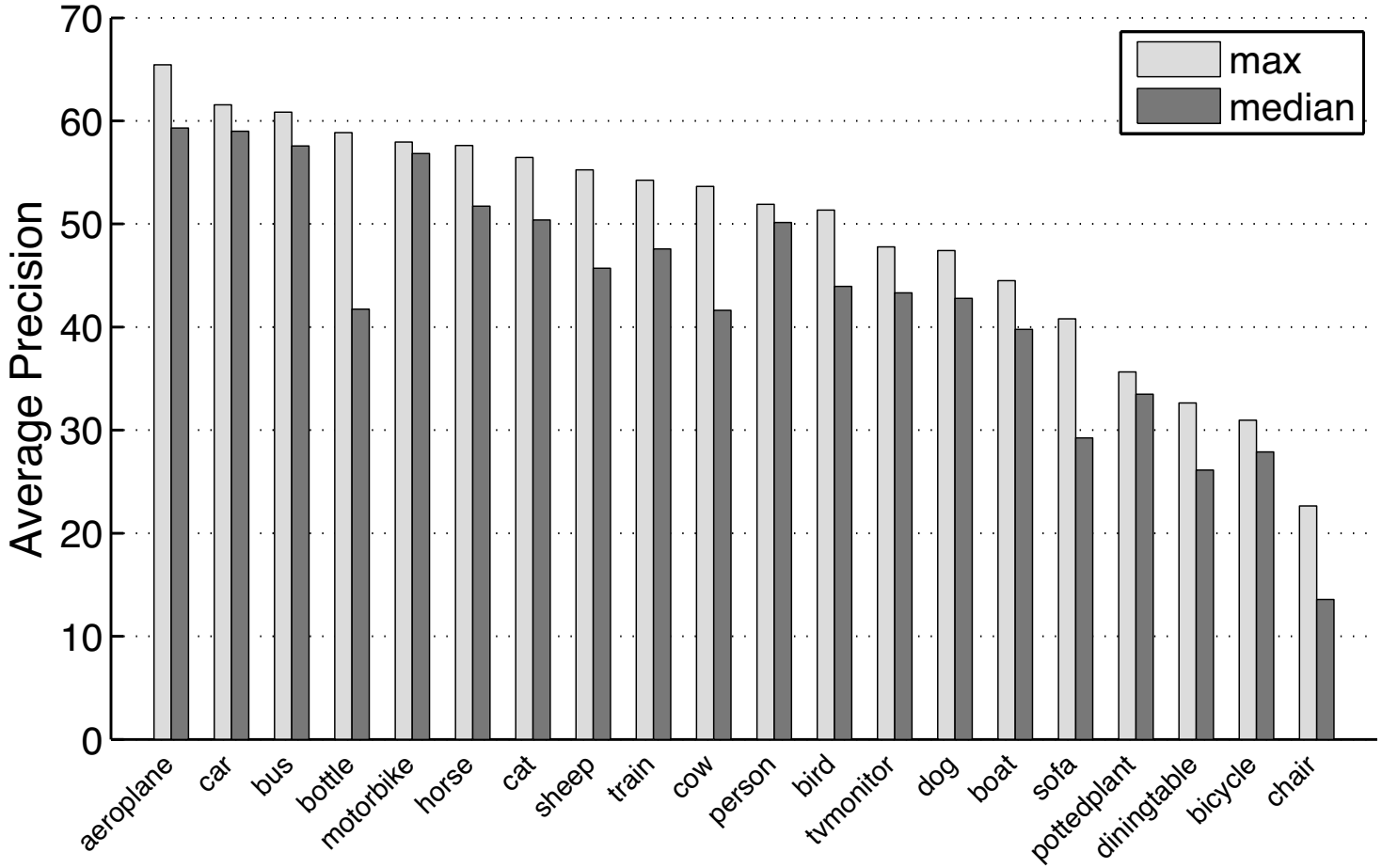
	[mean]	background	aeroplane	bicycle	bird	boat	bottle	Bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/Monitor
BONNGC_O2P_CSI	45.4	85.0	59.3	27.9	43.9	39.8	41.4	52.2	**61.5**	**B6.4**	13.6	44.5	26.1	42.8	51.7	57.9	51.3	29.8	45.7	28.8	49.9	43.3
BONN_CMBR_O2P_CPMC_LIN	44.8	83.9	60.0	27.3	46.4	40.0	41.7	57.6	59.0	50.4	10.0	41.6	22.3	43.0	51.7	56.8	50.1	33.7	43.7	29.5	47.5	44.7
BONN_O2PCPMC_FGT_SEGM	47.0	**85.1**	**65.4**	29.3	51.3	33.4	44.2	59.8	60.3	52.5	13.6	**53.6**	32.6	40.3	57.6	57.3	49.0	33.5	53.5	29.2	47.6	37.6
NUS_DET_SPR_GC_SP	47.3	82.8	52.9	**31.0**	39.8	**44.5**	**58.9**	**60.8**	52.5	49.0	**22.6**	38.1	27.5	**47.4**	52.4	46.8	51.9	**35.7**	**55.3**	**40.8**	**54.2**	**47.8**
UVA_OPT_NBNN_CRF	11.3	63.2	10.5	2.3	3.0	3.0	1.0	30.2	14.9	15.0	0.2	6.1	2.3	5.1	12.1	15.3	23.4	0.5	8.9	3.5	10.7	5.3

Trained on external data (comp6)

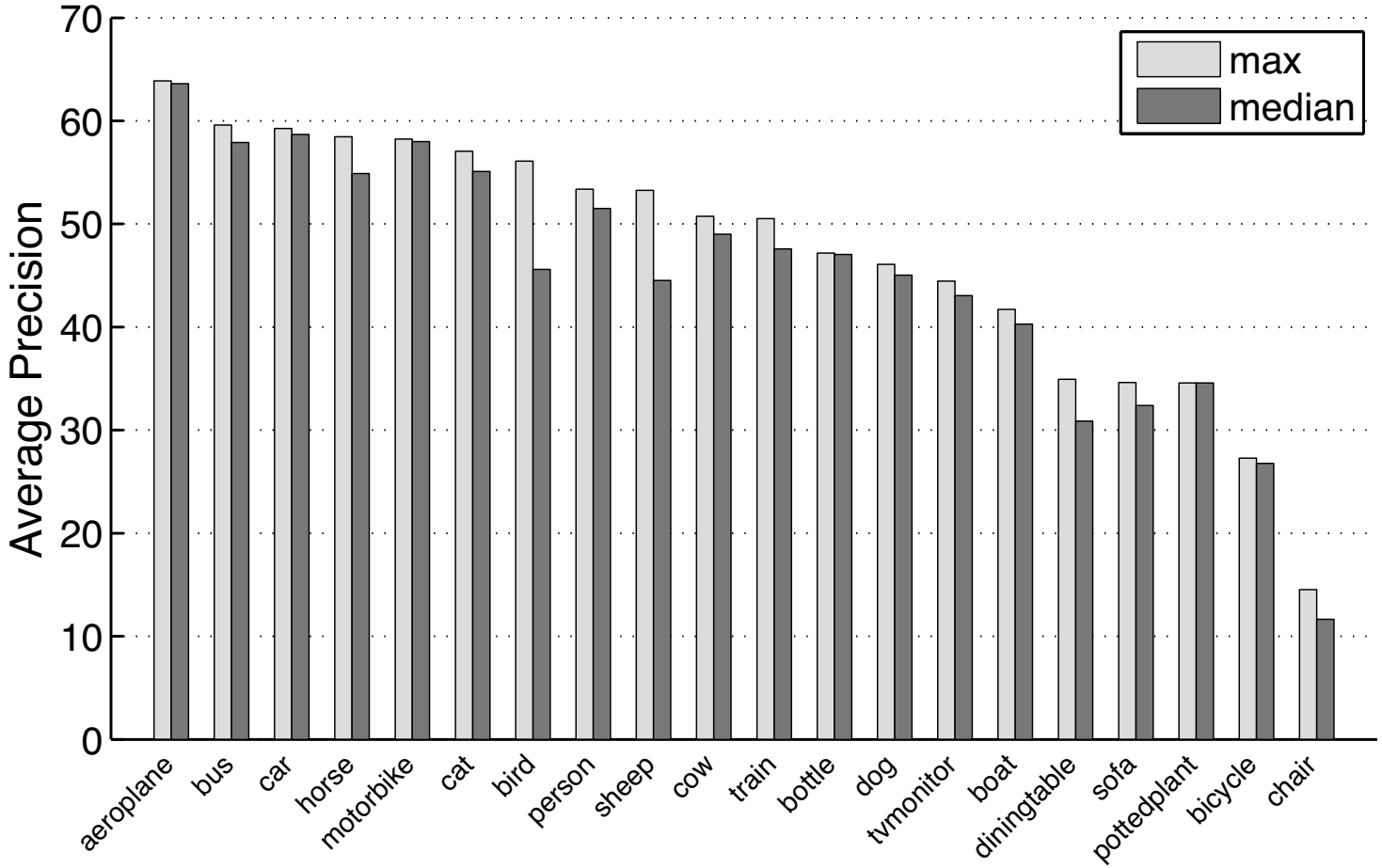
	[mean]	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/monitor
BONNGC_O2P_CSI	46.8	85.0	63.6	26.8	45.6	41.7	47.1	54.3	58.6	55.1	14.5	49.0	30.9	46.1	52.6	**58.2**	**53.4**	32.0	44.5	34.6	45.3	43.1
BONN_CMBR_O2P_CPMC_LIN	46.7	84.7	63.9	23.8	44.6	40.3	45.5	59.6	58.7	57.1	11.7	45.9	34.9	43.0	54.9	58.0	51.5	34.6	44.1	29.9	50.5	44.5
BONN_O2PCPMC_FGT_SEGM	**47.5**	85.2	63.4	27.3	**56.1**	37.7	47.2	57.9	59.3	55.0	11.5	50.8	30.5	45.0	**58.4**	57.4	48.6	34.6	**53.3**	32.4	47.6	39.2

- 4% improvement in mean accuracy over last year
- NUS_DET_SPR_GC_SP: 1st in 13 categories in comp5 (11 overall)
- BONN_O2PCPMC_FGT_SEGM: 1st in 9 categories in comp6 (4 overall)

Average precision by class (comp5)



Average precision by class (comp6)



Prizes

- Winner (comp5):
  - **NUS_DET_SPR_GC_SP**
    Wei Xia, Csaba Domokos, Jian Dong,
    Shuicheng Yan, Loong Fah Cheong
    Zhongyang Huang, Shengmei Shen
    National University of Singapore
    Panasonic Singapore Libraries
- Winner (comp6):
  - **BONN_(O2PCPMC_FGT_SEGM)**
    João Carreira, Adrian Ion, Fuxin Li,
    Cristian Sminchisescu
    University of Bonn

Challenge website

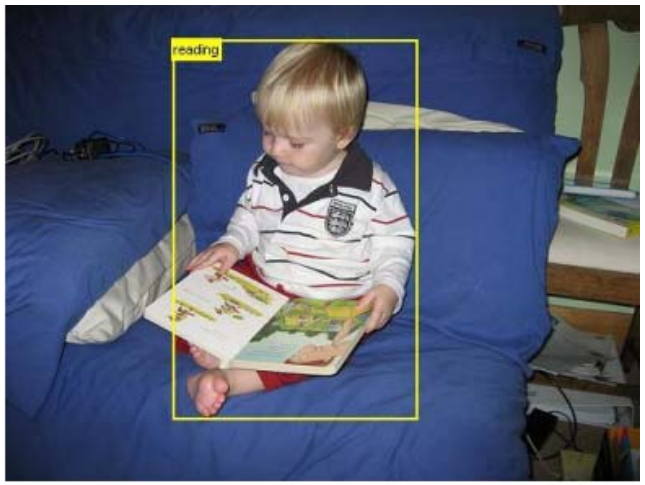
Complete results available for viewing at

http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/

Action classification taster challenge

- Given the bounding box of a person, predict whether they are performing a given action

Playing Instrument?



Reading?

- Encourage research on **still-image** activity recognition: more detailed understanding of image

Ten action classes + "Other"

Dataset statistics

- Around 90% increase in size over VOC 2011.

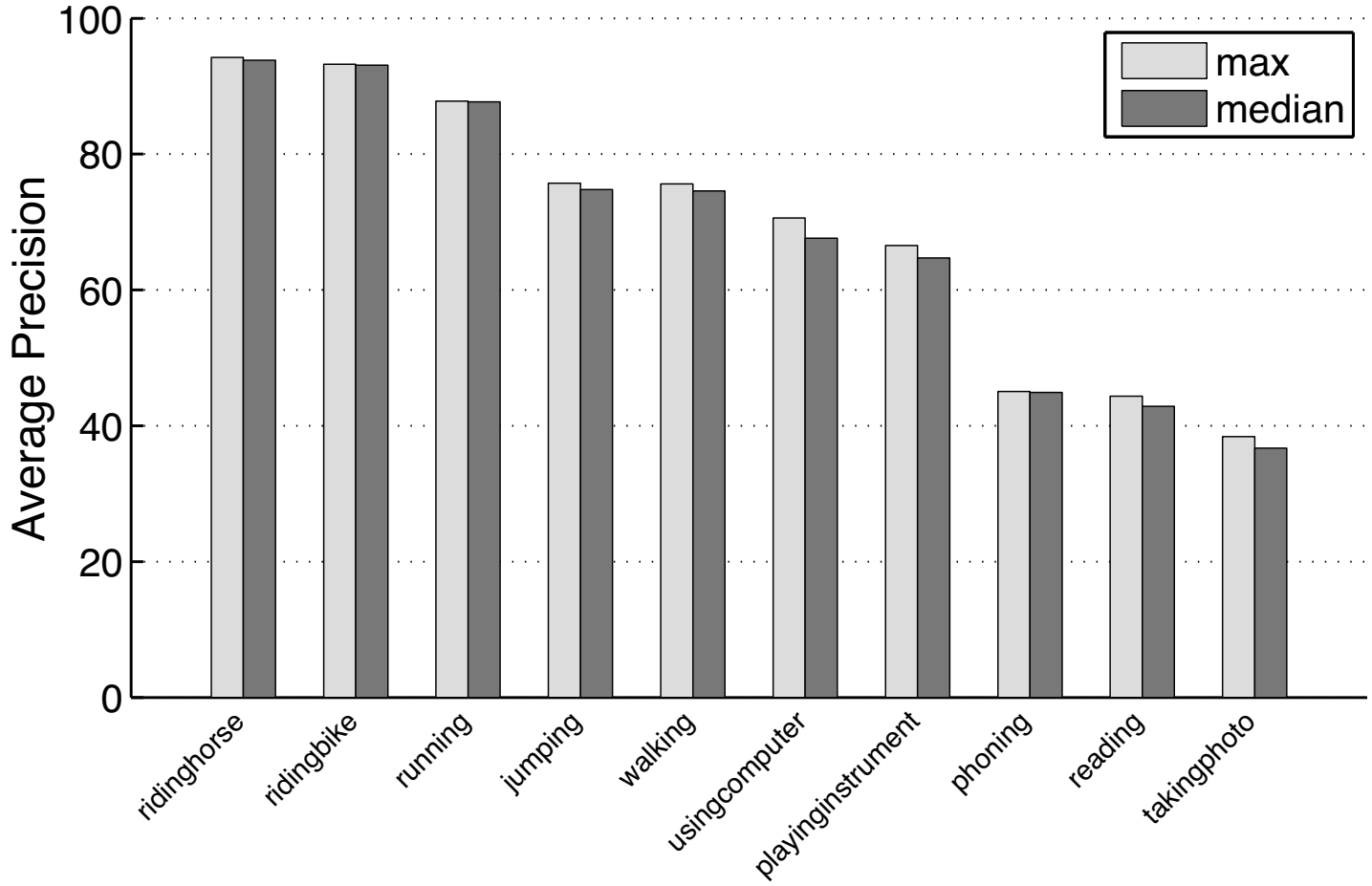
	Training	Testing
Actions	5303	5292

- Minimum ~400 people per action category
- Only subset of people are annotated
- Actions are not mutually-exclusive

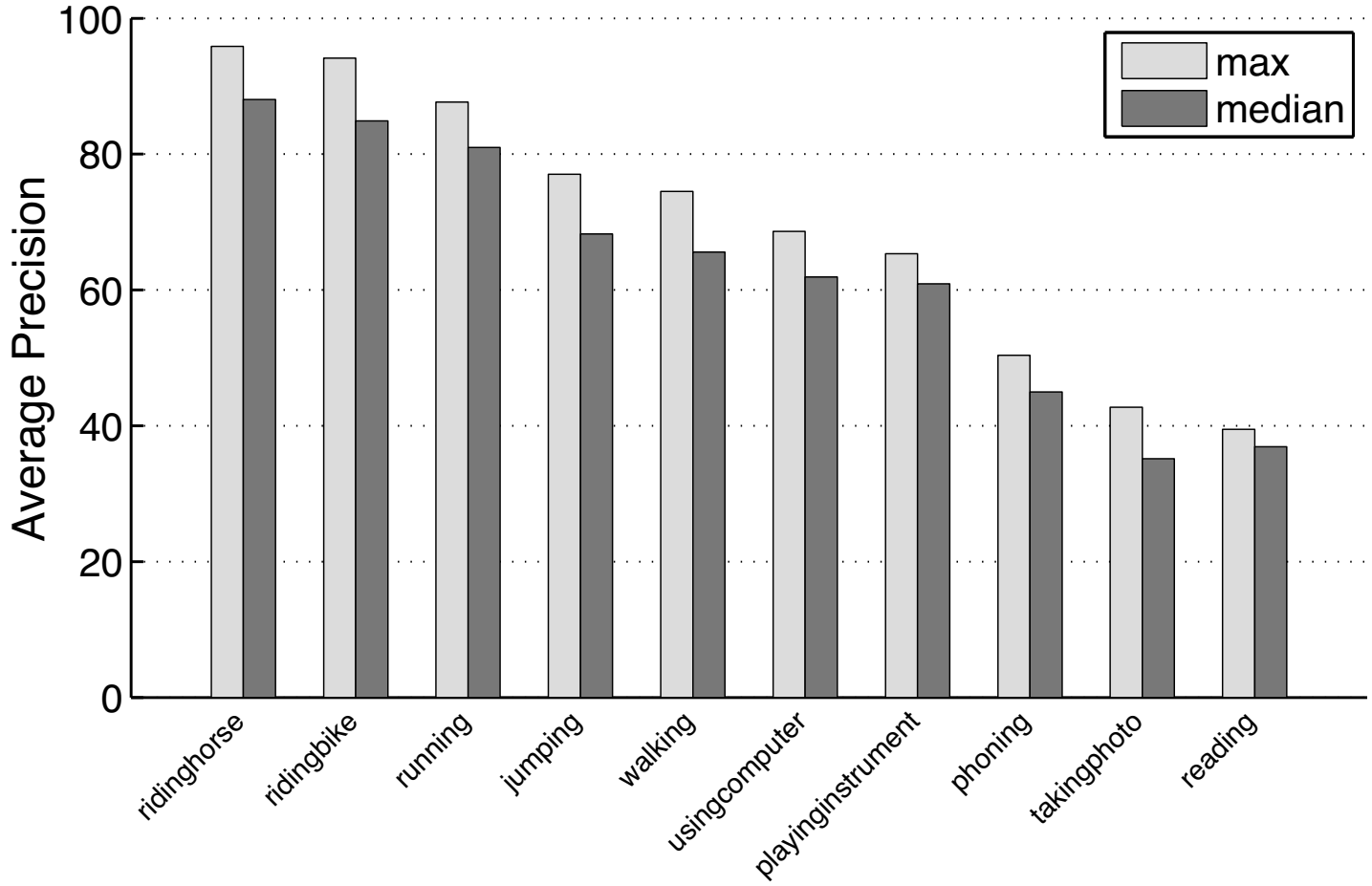
Submitted methods

- 4 methods, 4 groups
  - VOC2011: 10 methods, 6 groups…

Average precision by action (comp9)



Average precision by action (comp10)



AP by class and method

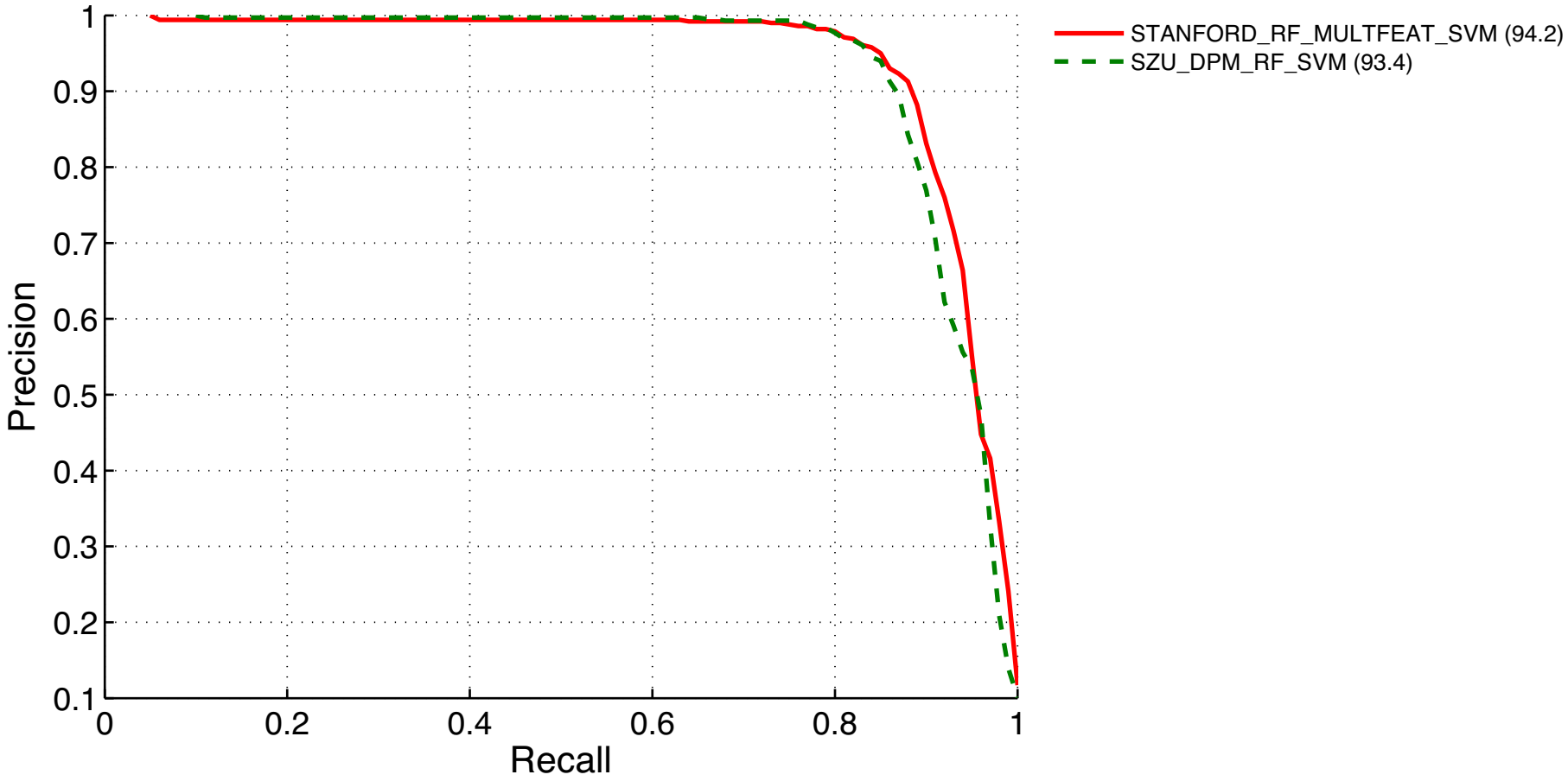
Trained on VOC 2012 data

	jumping	phoning	playinginstrument	reading	ridingbike	ridinghorse	running	takingphoto	usingcomputer	walking
STANFORD_RF_MULTFEAT_SVM	75.7	44.8	66.6	44.4	93.2	94.2	87.6	38.4	70.6	75.6
SZU_DPM_RF_SVM	73.8	45.0	62.8	41.4	93.0	93.4	87.8	35.0	64.7	73.5

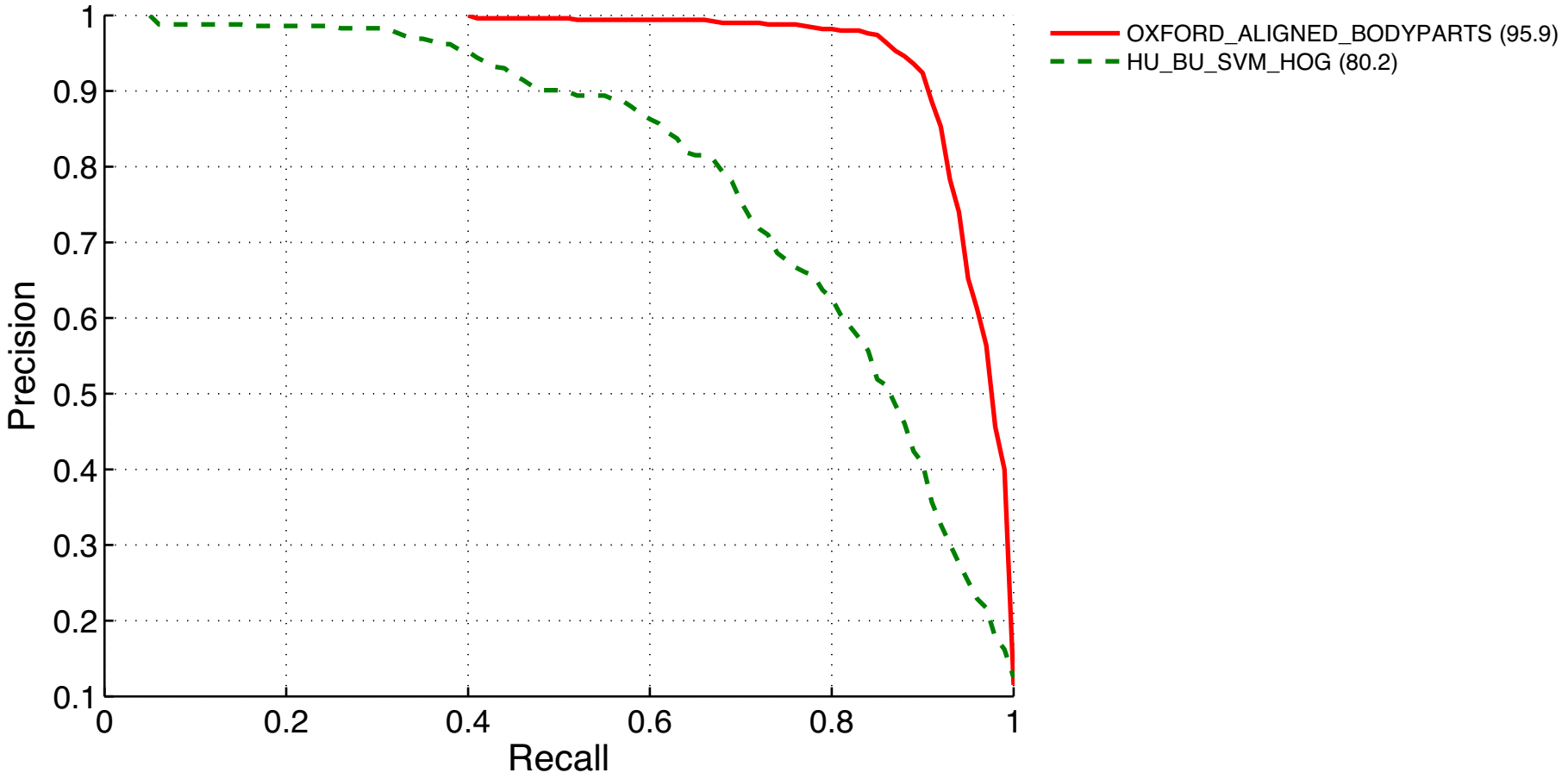
Trained on external data

	jumping	phoning	playinginstrument	reading	ridingbike	ridinghorse	running	takingphoto	usingcomputer	walking
HU_BU_SVM_HOG	59.4	39.6	56.5	34.4	75.7	80.2	74.3	27.6	55.2	56.6
OXFORD_ALIGNED_BODYPARTS	77.0	50.4	65.3	39.5	94.1	95.9	87.7	42.7	68.6	74.5

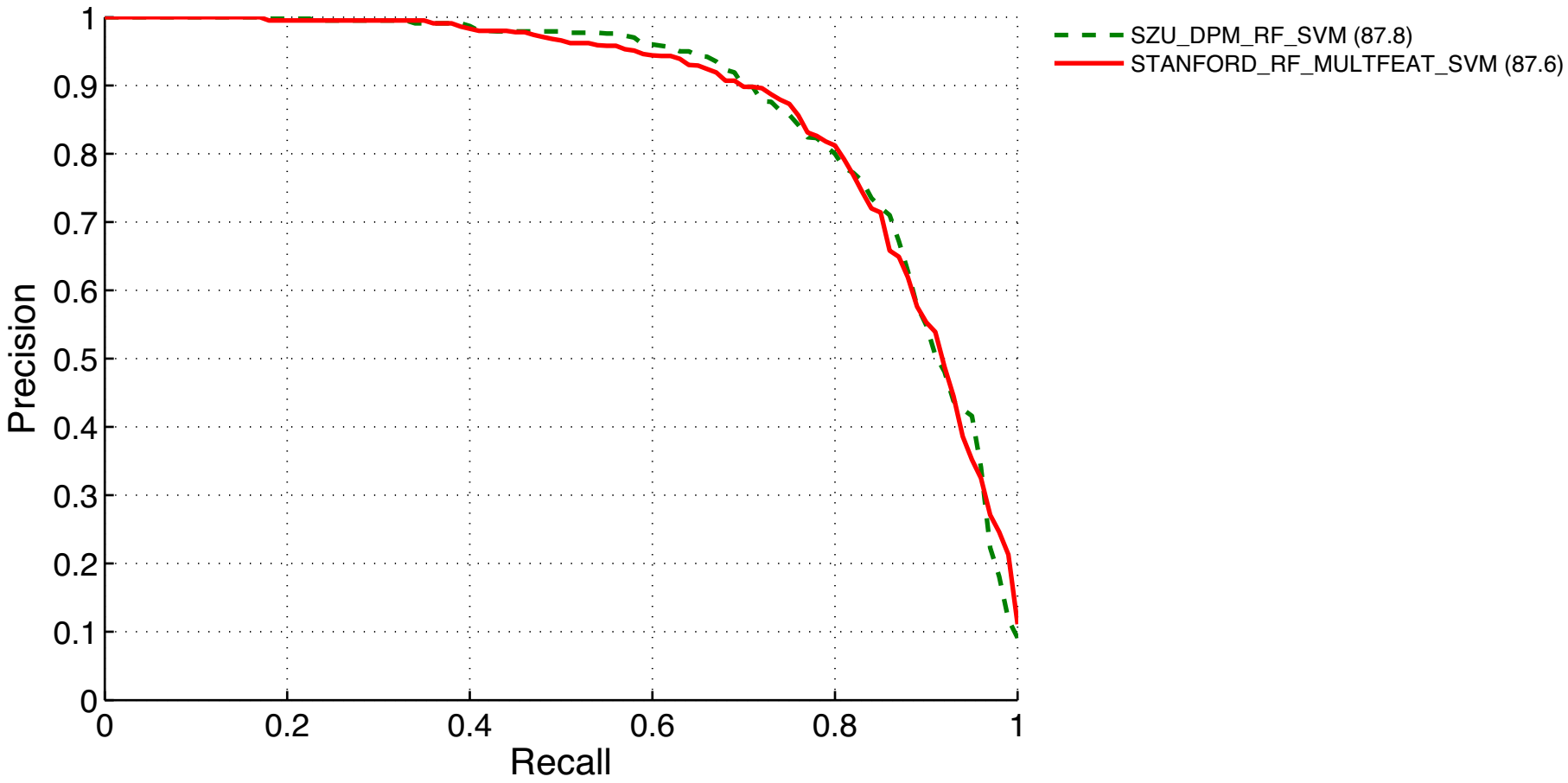
Precision/recall curves (riding horse)



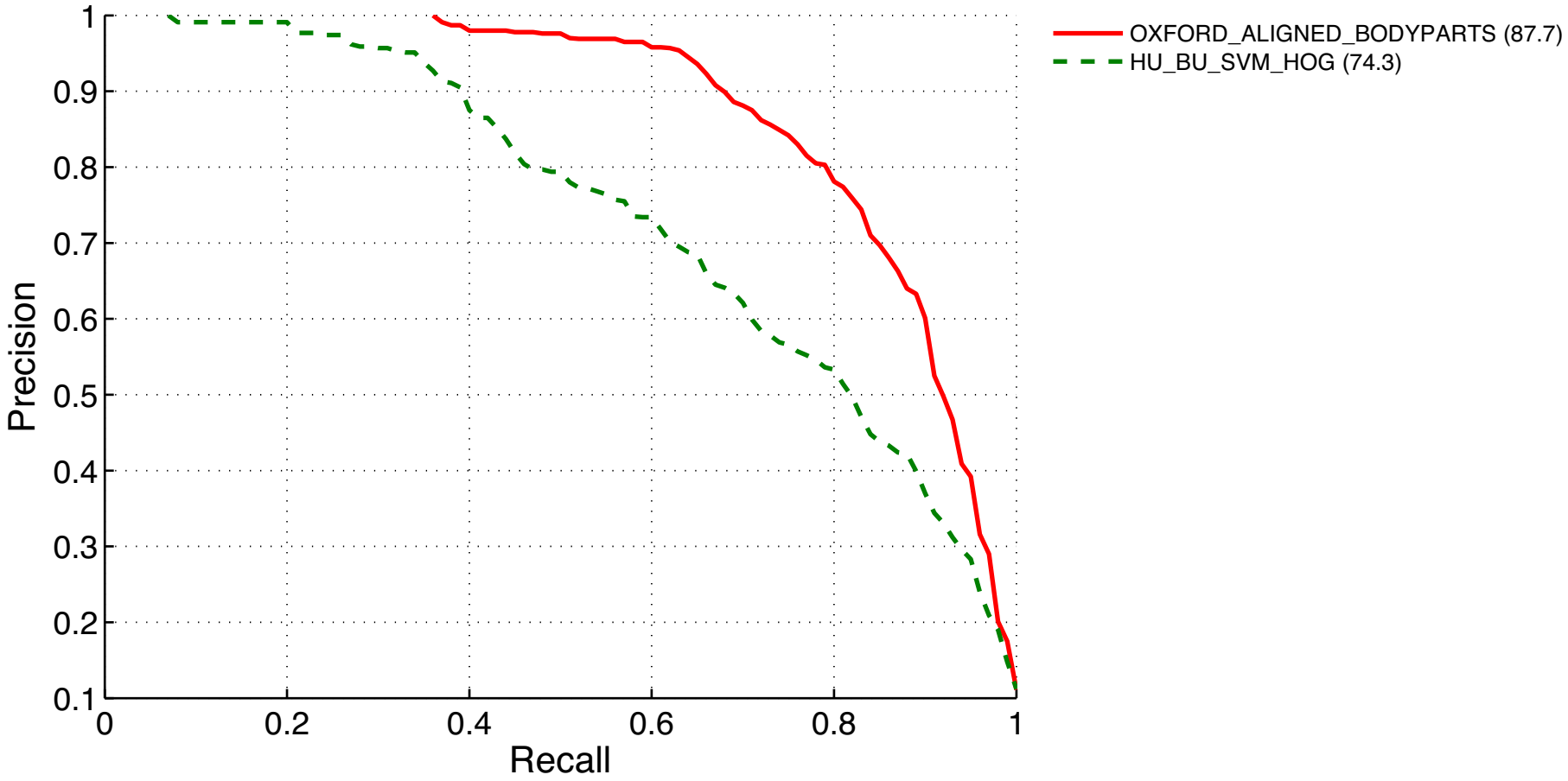
Precision/recall curves (riding horse)



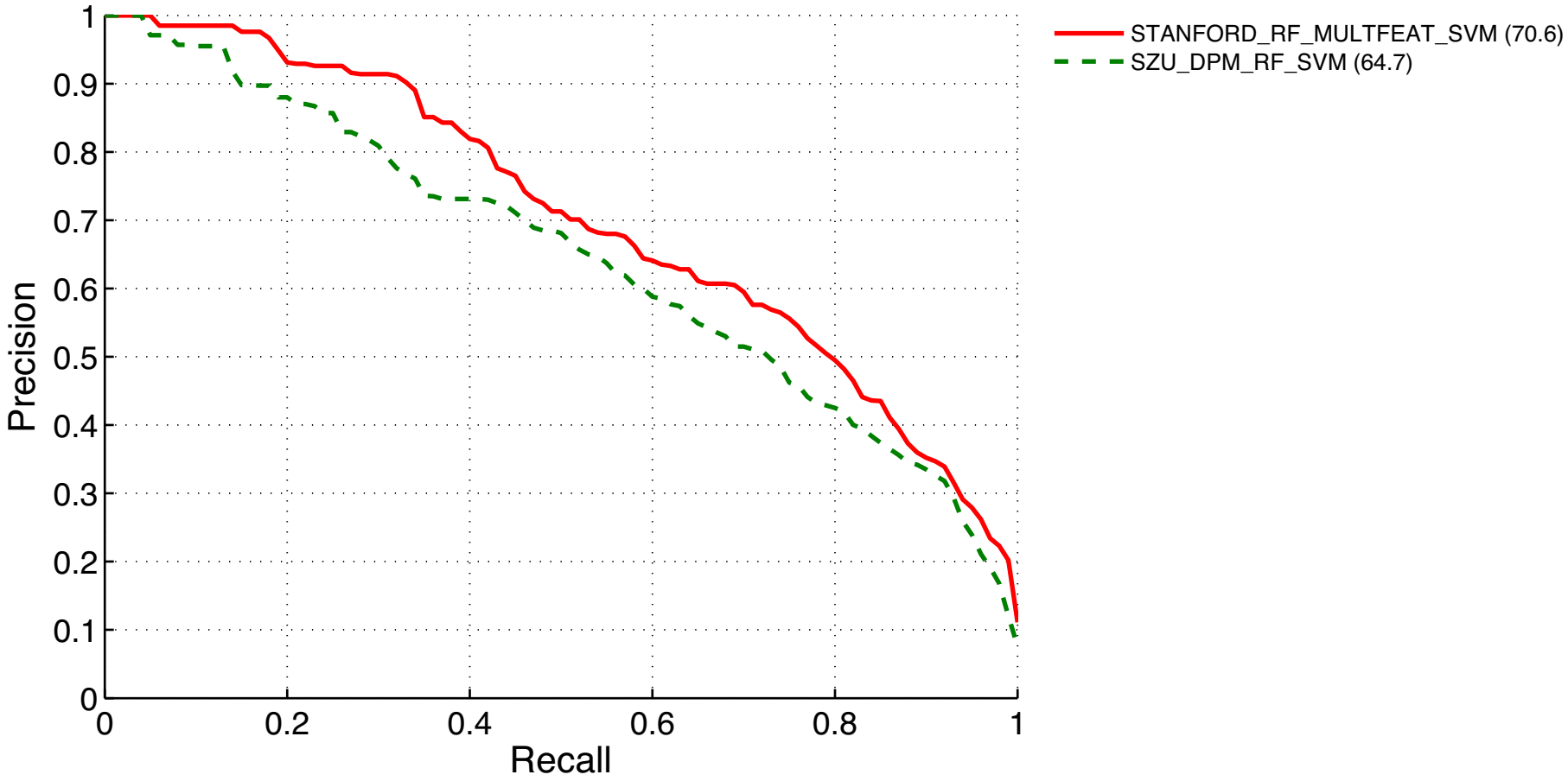
Precision/recall curves (running)



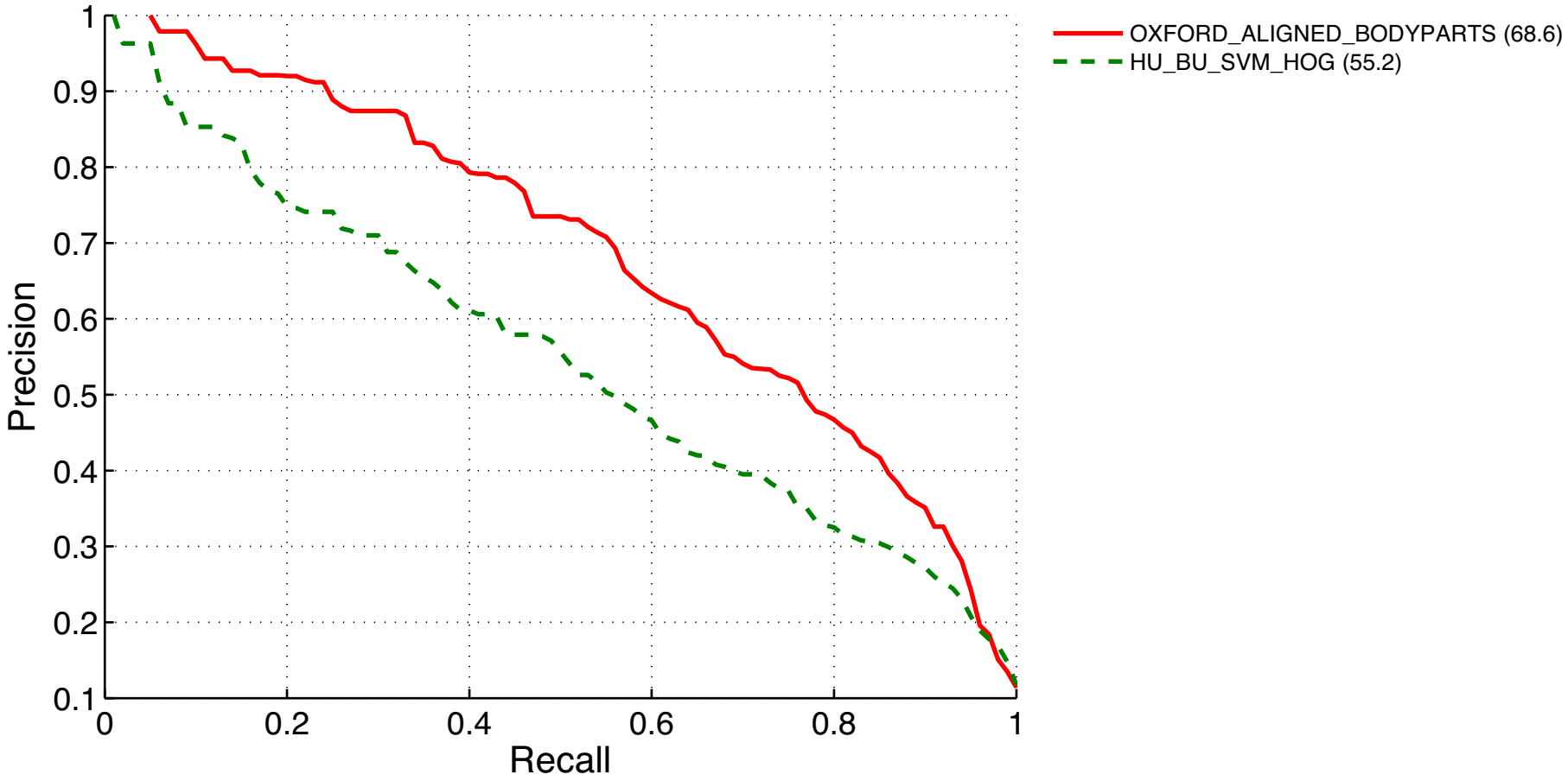
Precision/recall curves (running)



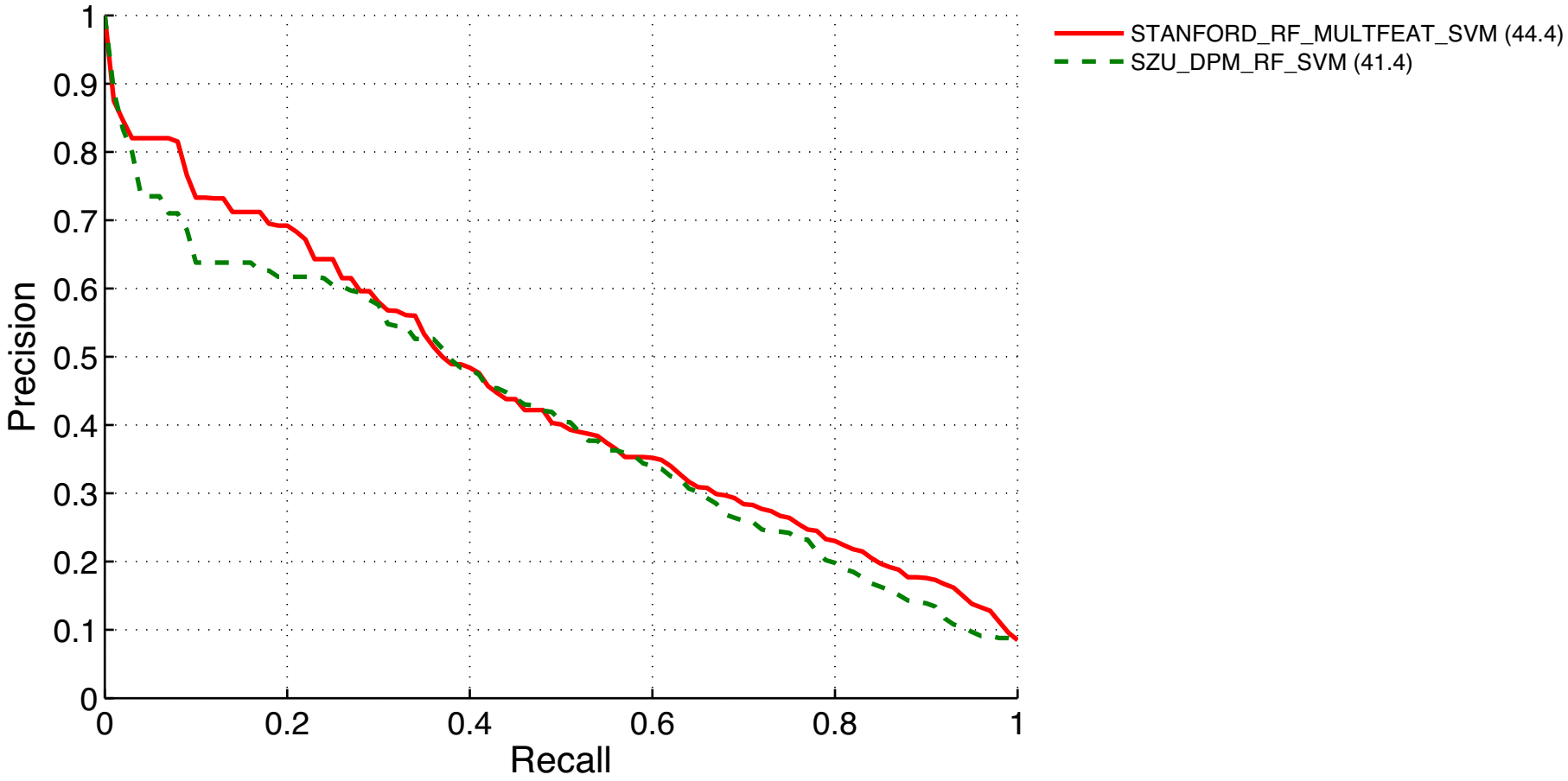
Precision/recall curves (using computer)



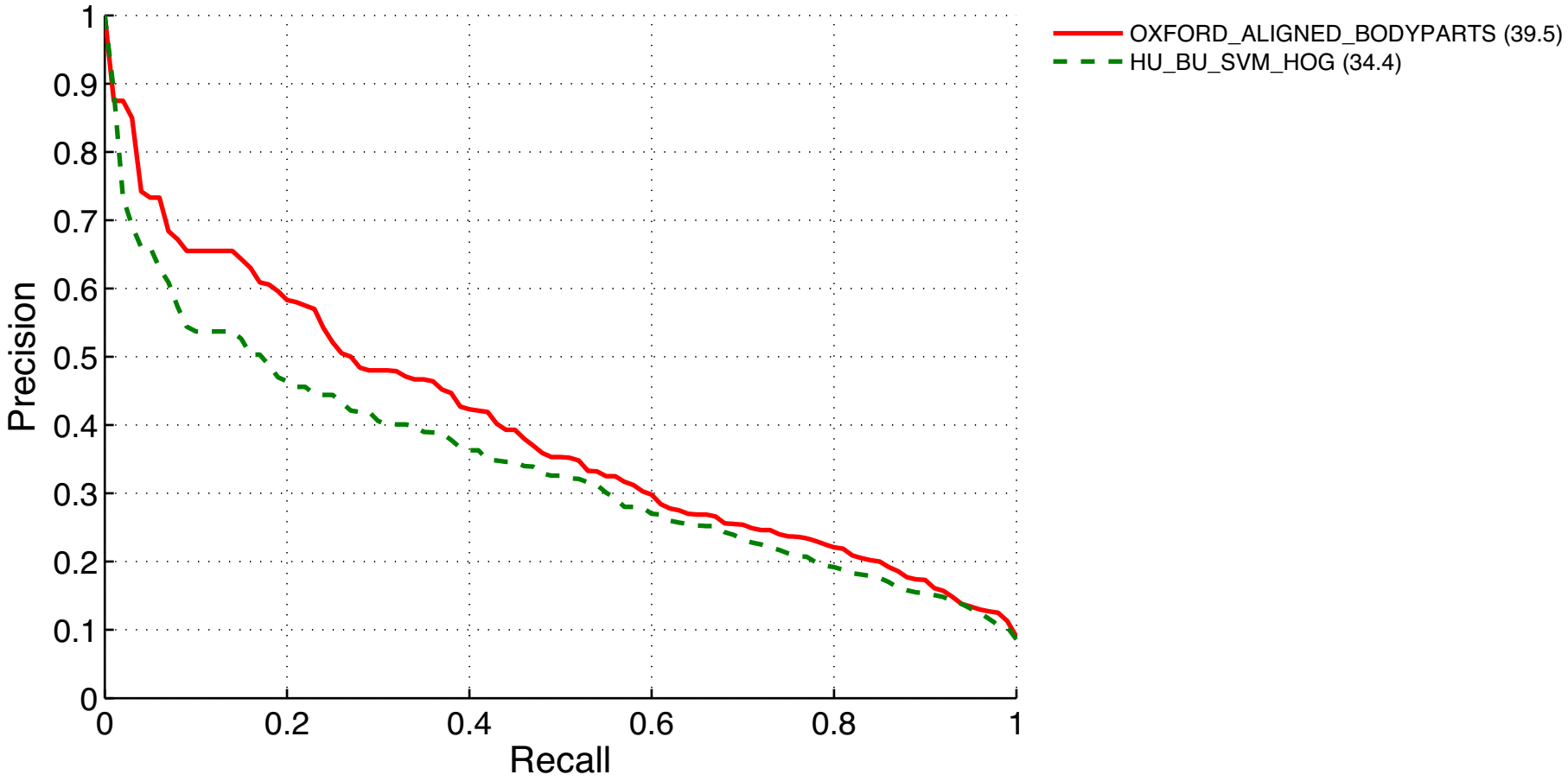
Precision/recall curves (using computer)



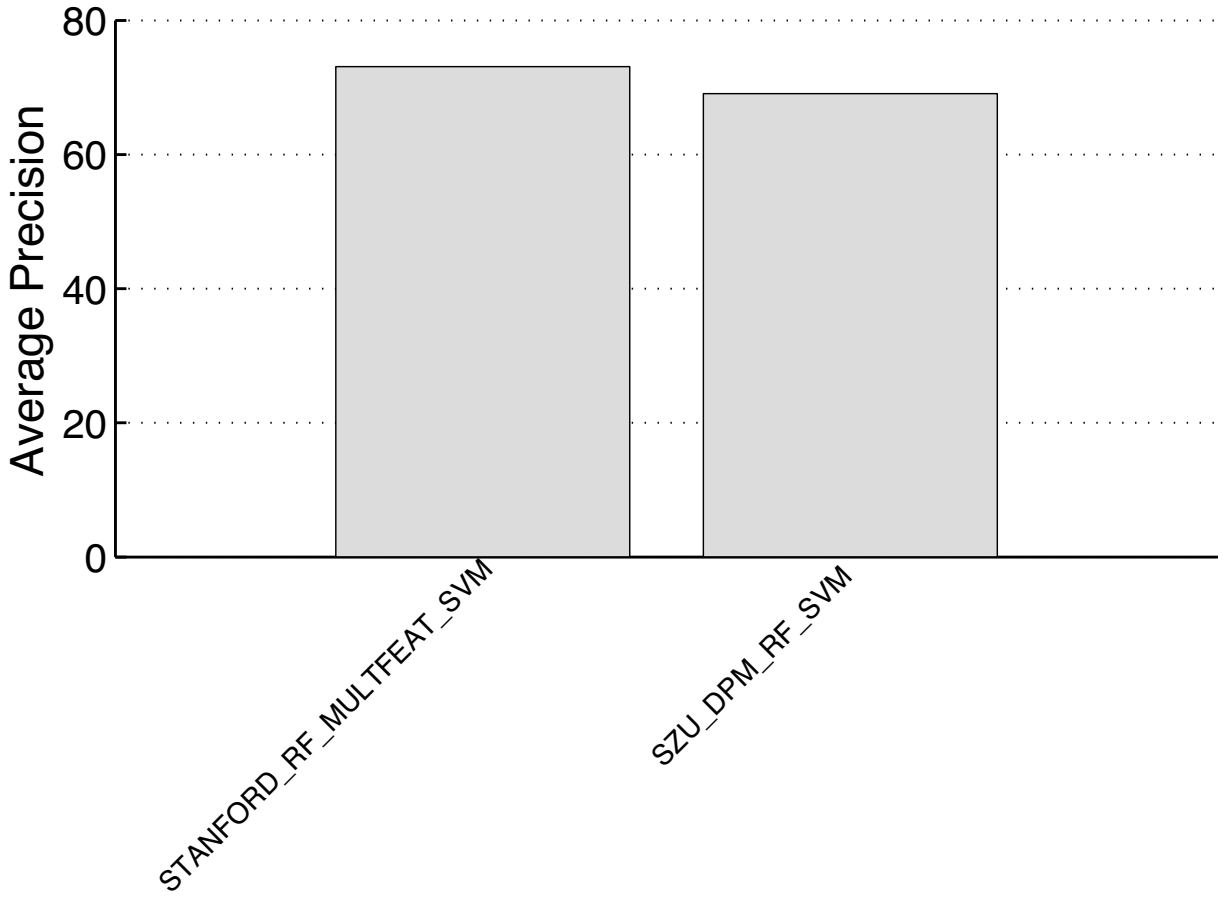
Precision/recall curves (reading)



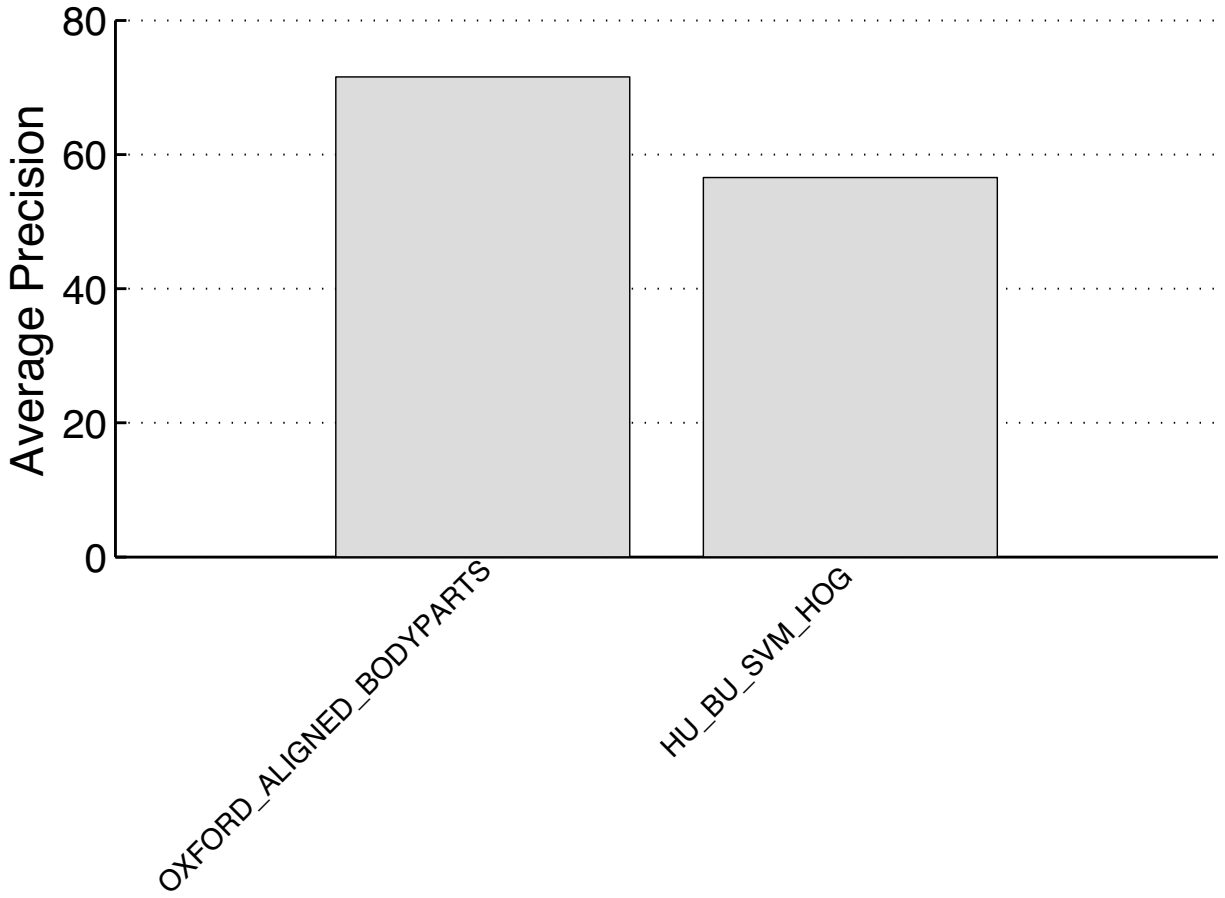
Precision/recall curves (reading)



Median average precision by method



Median average precision by method



Prizes

- Winner (comp9)
  - **STANFORD_RF_MULTFEAT_SVM**

    Aditya Khosla, Rui Zhang,

    Bangpeng Yao, Li Fei-Fei

    Stanford University

    MIT

- Winner (comp10)
  - **OXFORD_ALIGNED_BODYPARTS**

    Minh Hoai,

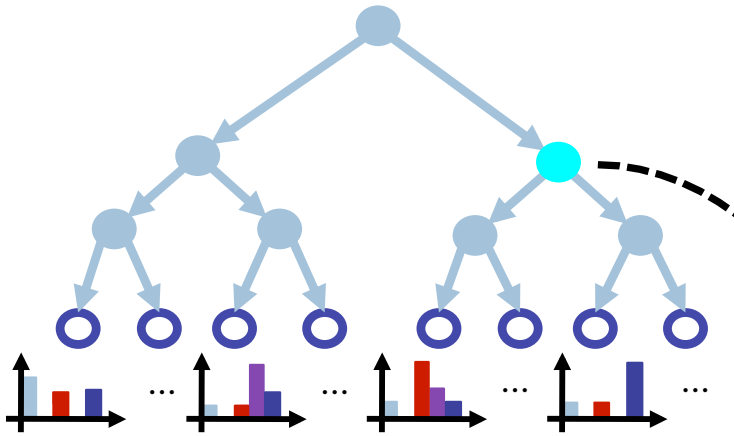
    Lubor Ladicky,

    Andrew Zisserman

    University of Oxford

Comp 9, bird's eye view

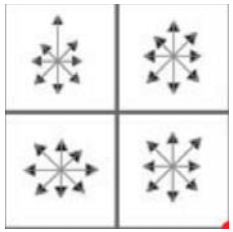
Discriminative Random Forest



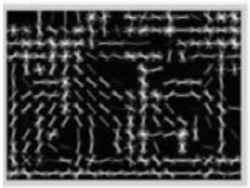
- Randomly generate a number of image regions.
- Random spatial pyramid for each region.
- Train an SVM for each region, select the best one.

[B. Yao*, A. Khosla*, and L. Fei-Fei, “Combining Randomization and Discrimination for Fine-Grained Image Categorization.” CVPR 2011]

Combining multiple features (more than '11)



Dense SIFT
[Lowe, 2004]

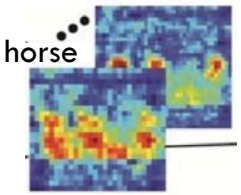


Dense HOG
[Dalal, 2005]

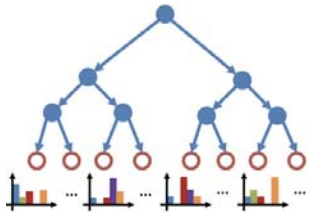
Color Names
[Van De Weijer et al, 2007]

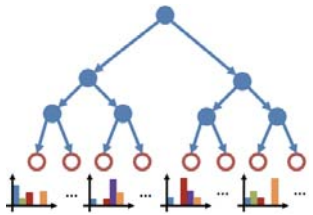
1	0	1
0		1
1	0	0

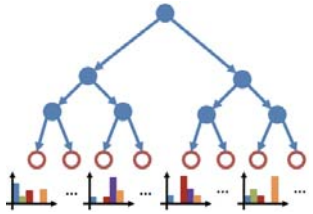
Quantized LBP

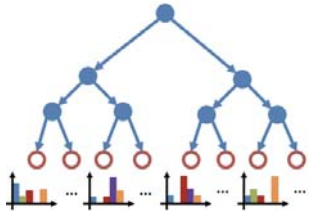


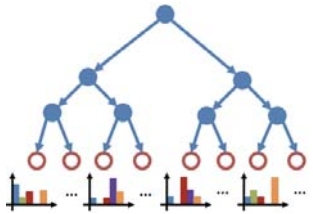
Object Bank
[Li et al, 2010]











w_1

w_2

w_3

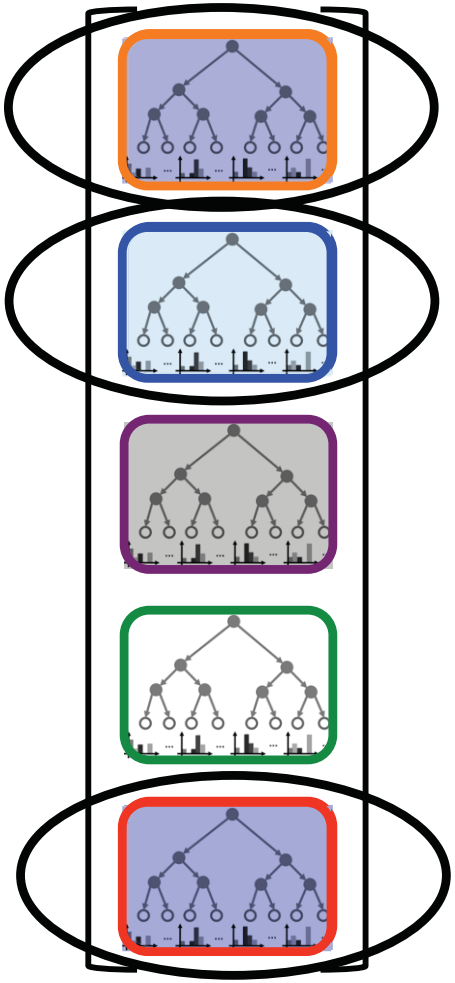
w_4

w_5

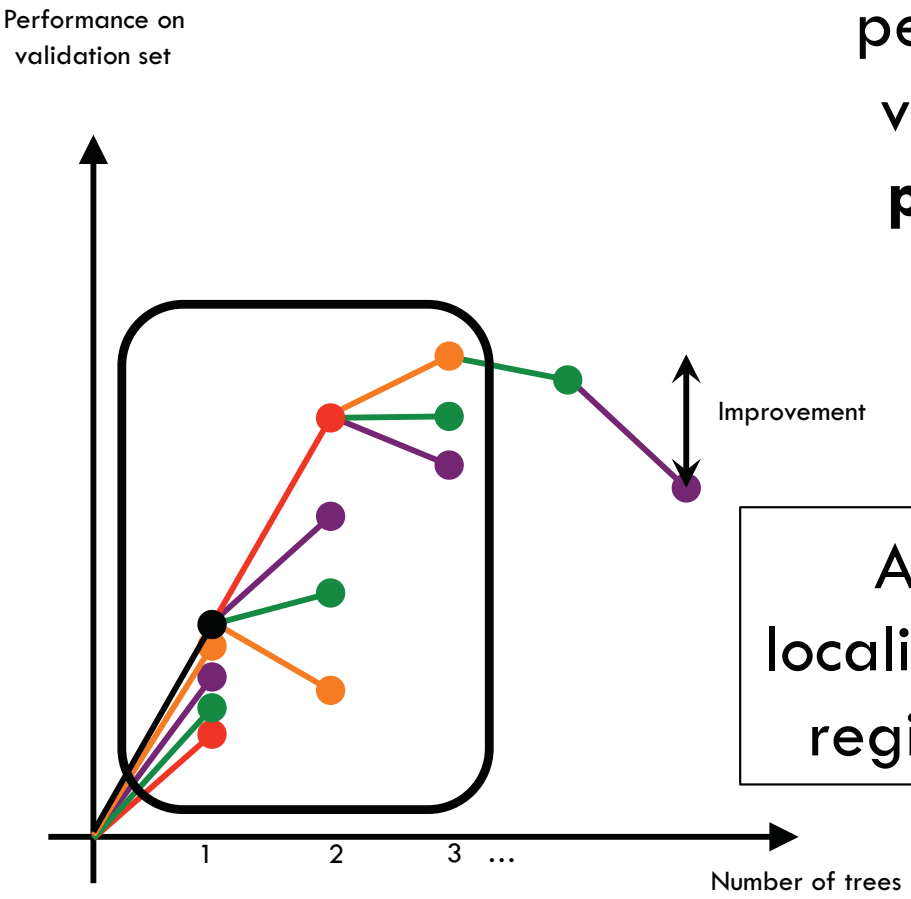
Action
Classification

$$w_1+w_2+w_3+w_4+w_5 = 1$$

Greedy Tree Selection (2nd change from '11)



Learned trees



Select trees with highest performance on the validation set in a **per-class manner**

Allows for better localization of important regions for each class!

Comp 10, bird's eye view

Action Recognition from Still Images by Aligning Body Parts

Minh Hoai, Lubor Ladicky, Andrew Zisserman

University of Oxford

Outline

Human focussed approach

- Bbox-aligned features
- Upper-body & hands
- Silhouette and segmentation-based features

Also use image classification and object detection

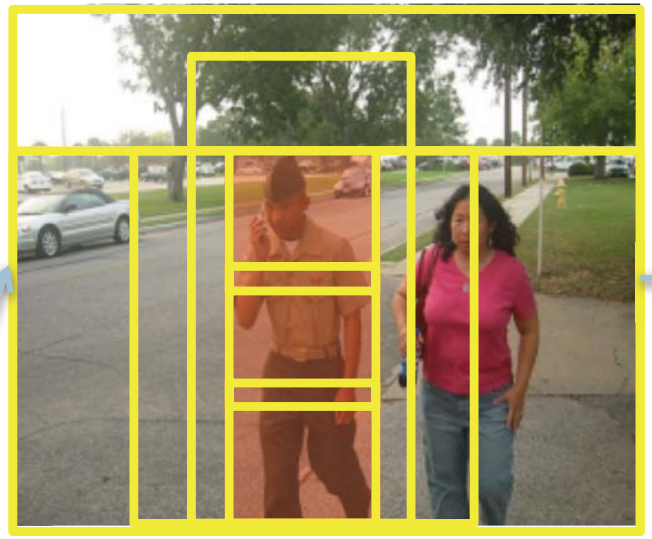
Final score: SVM with MKL

Bbox-aligned Features

Start with
groundtruth bbox

Get regions at
relative locations

Compute
features + kernels



Spatial pyramid of
Dense SIFTs

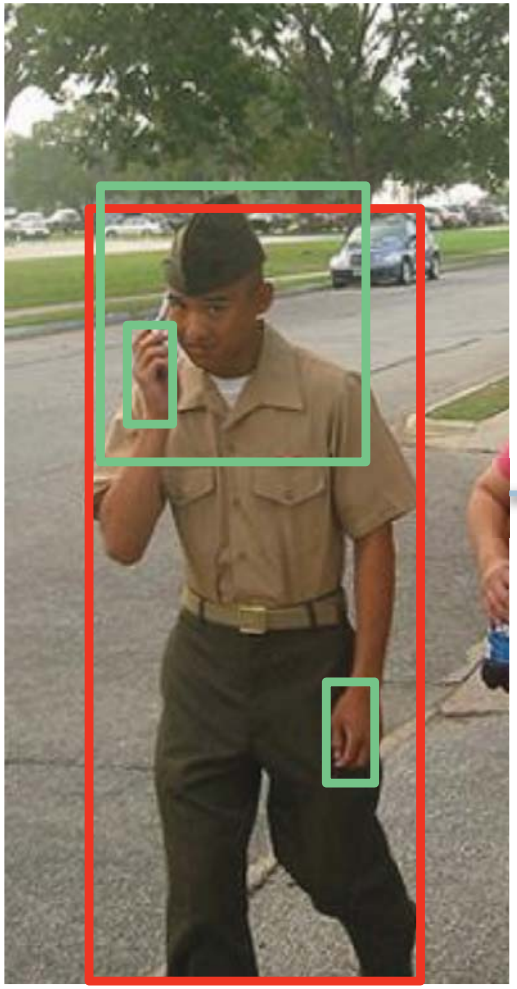
HOG
descriptors

Detection-aligned features

Detect upper-body and hands

Keep the upper-body and extend the hand regions

Compute features + kernels

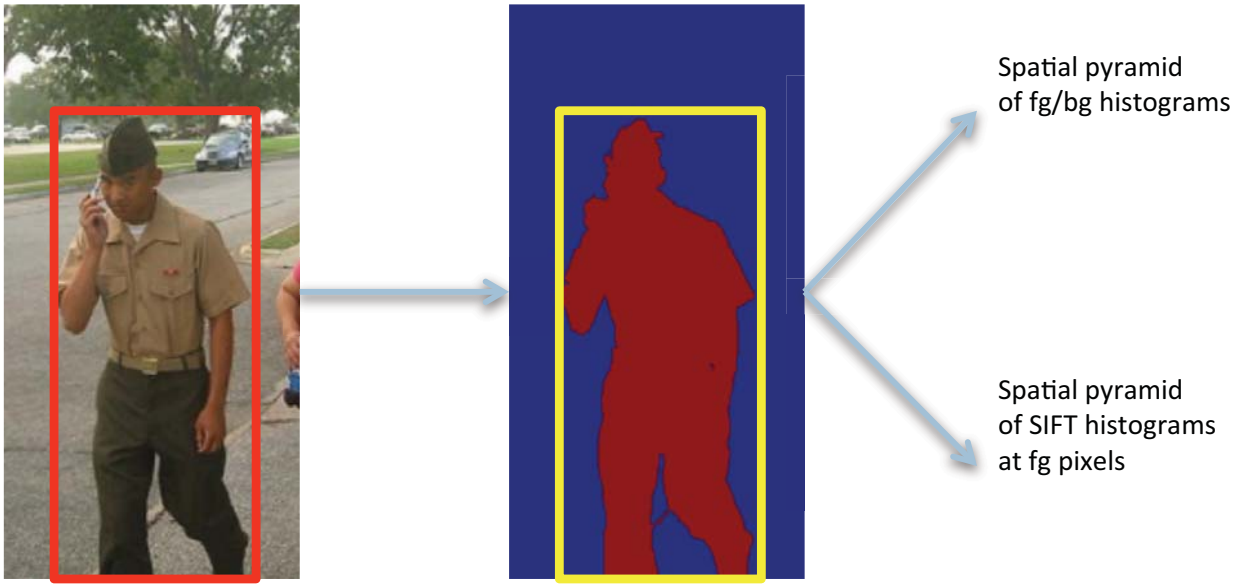


SIFT + HOG descriptors

Using Segmentation

Obtain the fg/bg segmentation

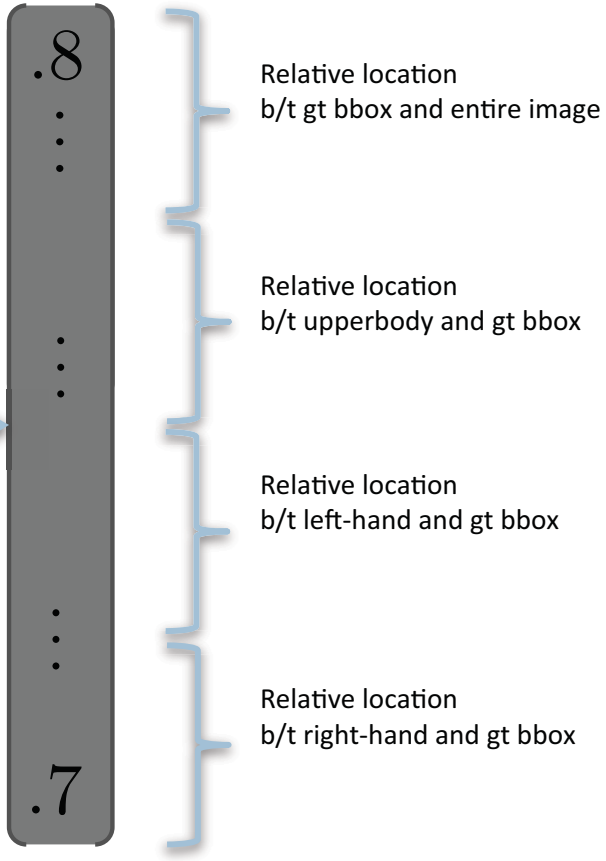
Compute features + kernels



Location Features

Start with ground truth (gt) bbox and upperbody, hand locations

Compute features + kernel

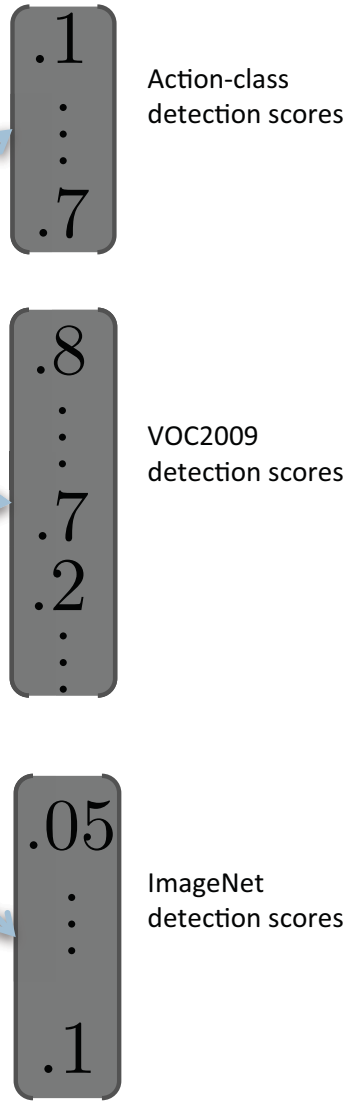


Latent SVM models

- Train an LSVM model for each action class:
  - 3 components (i.e., 6 if we count left-right mirror)
  - 8 deformable parts
- Obtain 20 pre-trained object detectors
  - LSVM models trained on VOC2009
  - Bundled with LSVM 4.0.1
- Obtain 16 musical instrument detectors
  - LSVM models trained on ImageNet

Using Detection Scores

Run detector and
record highest scores



Body layout challenge

- 1 submission, but only detecting heads
  - Hence no prize is awarded for this taster challenge this year