The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

Part 1 – Challenge & Detection Task

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman

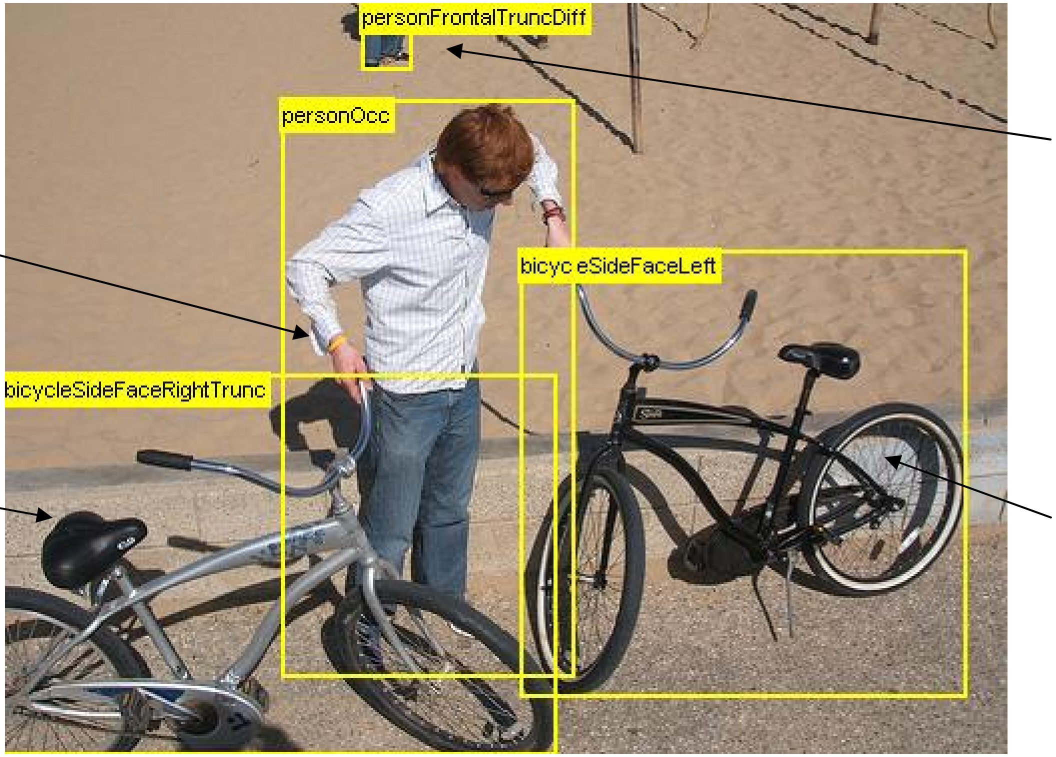
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Dataset: Collection

- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation
  - Queries
    - Keyword e.g. "car", "vehicle", "street", "downtown"
    - Date of capture e.g. "taken 21-July"
      - Removes "recency" bias in flickr results
    - Images selected from random page of results
      - Reduces bias toward particular flickr users
- 2008 dataset retained as subset of 2009
  - Assignments to training/test sets maintained

Dataset: Annotation

- Complete annotation of all objects
- Annotated over web with <u>written guidelines</u>
  - High quality (?)



Occluded
Object is significantly occluded within BB

Difficult
Not scored in evaluation

Truncated
Object extends beyond BB

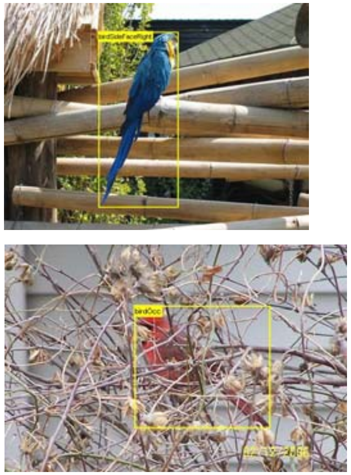
Pose
Facing left

Examples

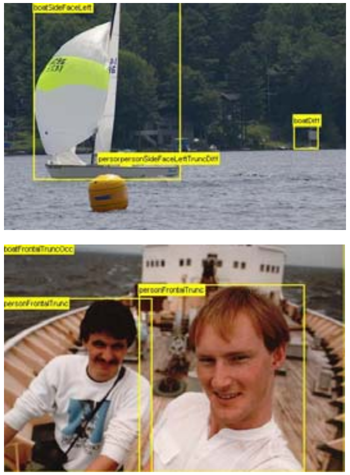
Aeroplane

Bicycle

Bird



Boat



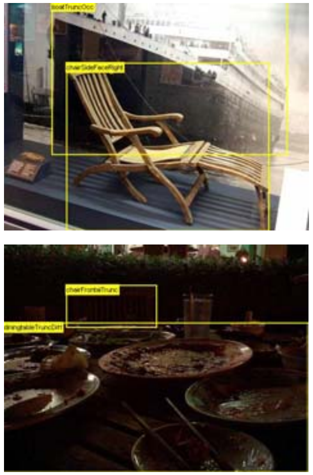
Bottle

Bus

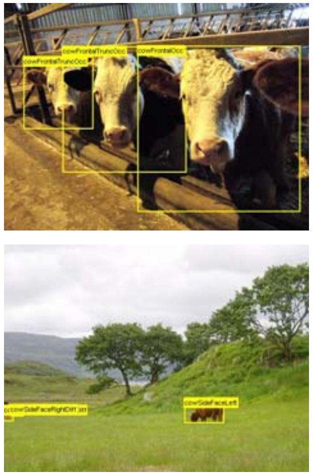
Car

Cat

Chair

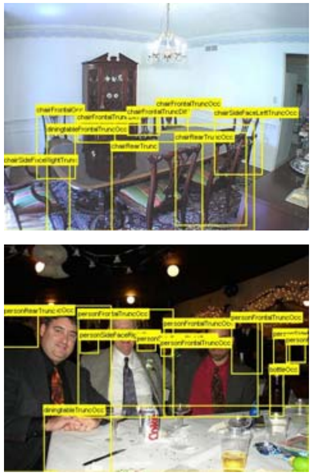


Cow



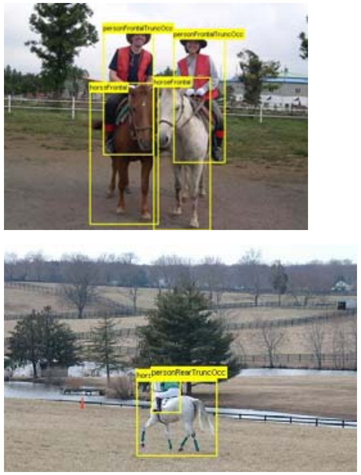
Examples

Dining Table



Dog

Horse

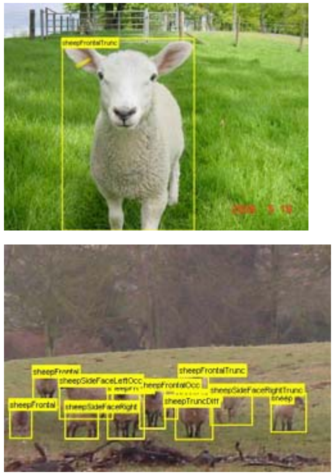


Motorbike

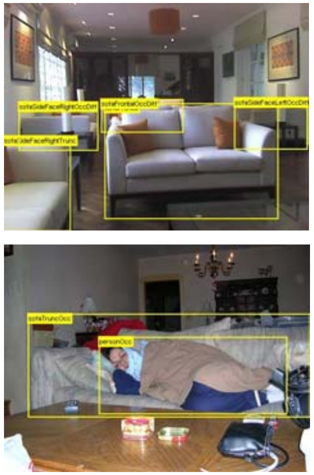
Person

Potted Plant

Sheep

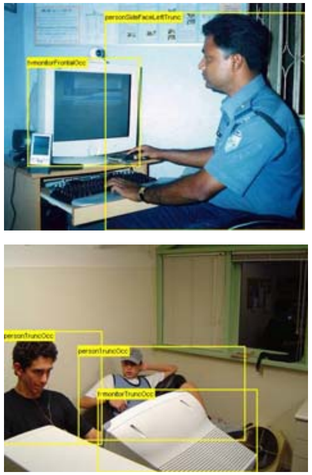


Sofa



Train

TV/Monitor



Dataset Statistics

	train		val		trainval		test	
	Images	Objects	Images	Objects	Images	Objects	Images	Objects
Aeroplane	201	267	206	266	407	533		
Bicycle	167	232	181	236	348	468		
Bird	262	381	243	379	505	760		
Boat	170	270	155	267	325	537		
Bottle	220	394	200	393	420	787		
Bus	132	179	126	186	258	365		
Car	372	664	358	653	730	1,317		
Cat	266	308	277	314	543	622		
Chair	338	716	330	713	668	1,429		
Cow	86	164	86	172	172	336		
Diningtable	140	153	131	153	271	306		
Dog	316	391	333	392	649	783		
Horse	161	237	167	245	328	482		
Motorbike	171	235	167	234	338	469		
Person	1,333	2,819	1,446	2,996	2,779	5,815		
Pottedplant	166	311	166	316	332	627		
Sheep	67	163	64	175	131	338		
Sofa	155	172	153	175	308	347		
Train	164	190	160	191	324	381		
Tvmonitor	180	259	173	257	353	516		
Total	3,473	8,505	3,581	8,713	7,054	17,218	6,650	16,829

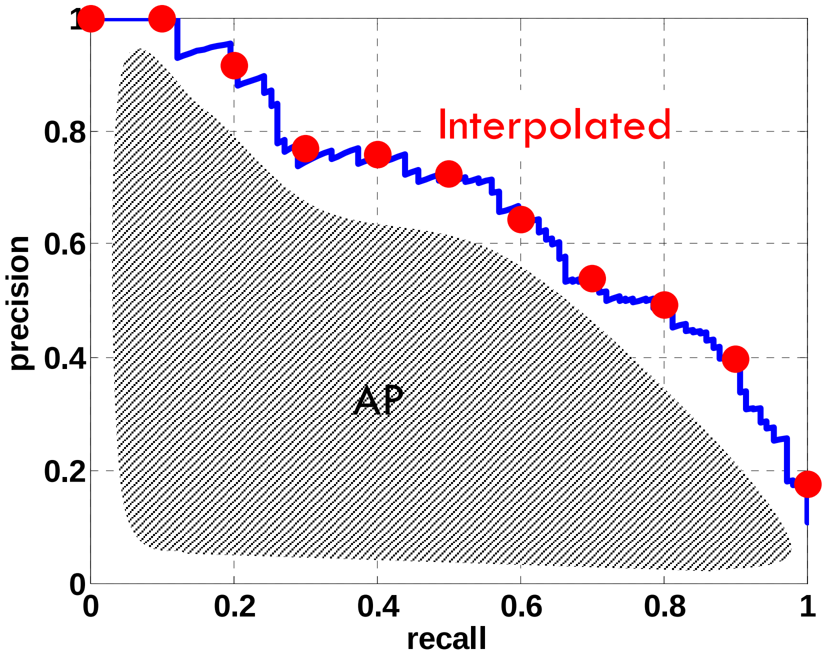
Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)

- Competition 3: Train on the supplied data
  - Which methods perform best given specified training data?
- Competition 4: Train on any (non-test) data
  - How well do state-of-the-art methods perform on these problems?

Evaluation

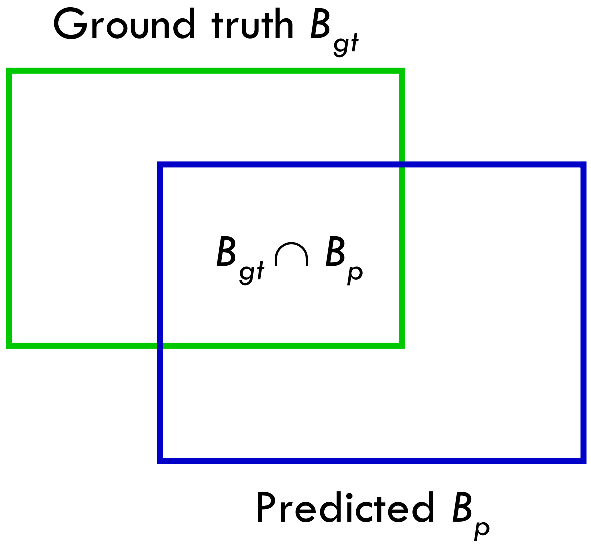
- Average Precision [TREC] averages precision over the entire range of recall
  - Curve interpolated to reduce influence of "outliers"



- A good score requires both high recall and high precision
- Application-independent
- Penalizes methods giving high precision but low recall

Evaluating Bounding Boxes

- Area of Overlap (AO) Measure



$$AO(B_{gt}, B_p) = \frac{|B_{gt} \bigcap B_p|}{|B_{gt} \bigcup B_p|}$$

- Need to define a threshold *t* such that $AO(B_{gt}, B_p)$ implies a correct detection: 50%

Participation

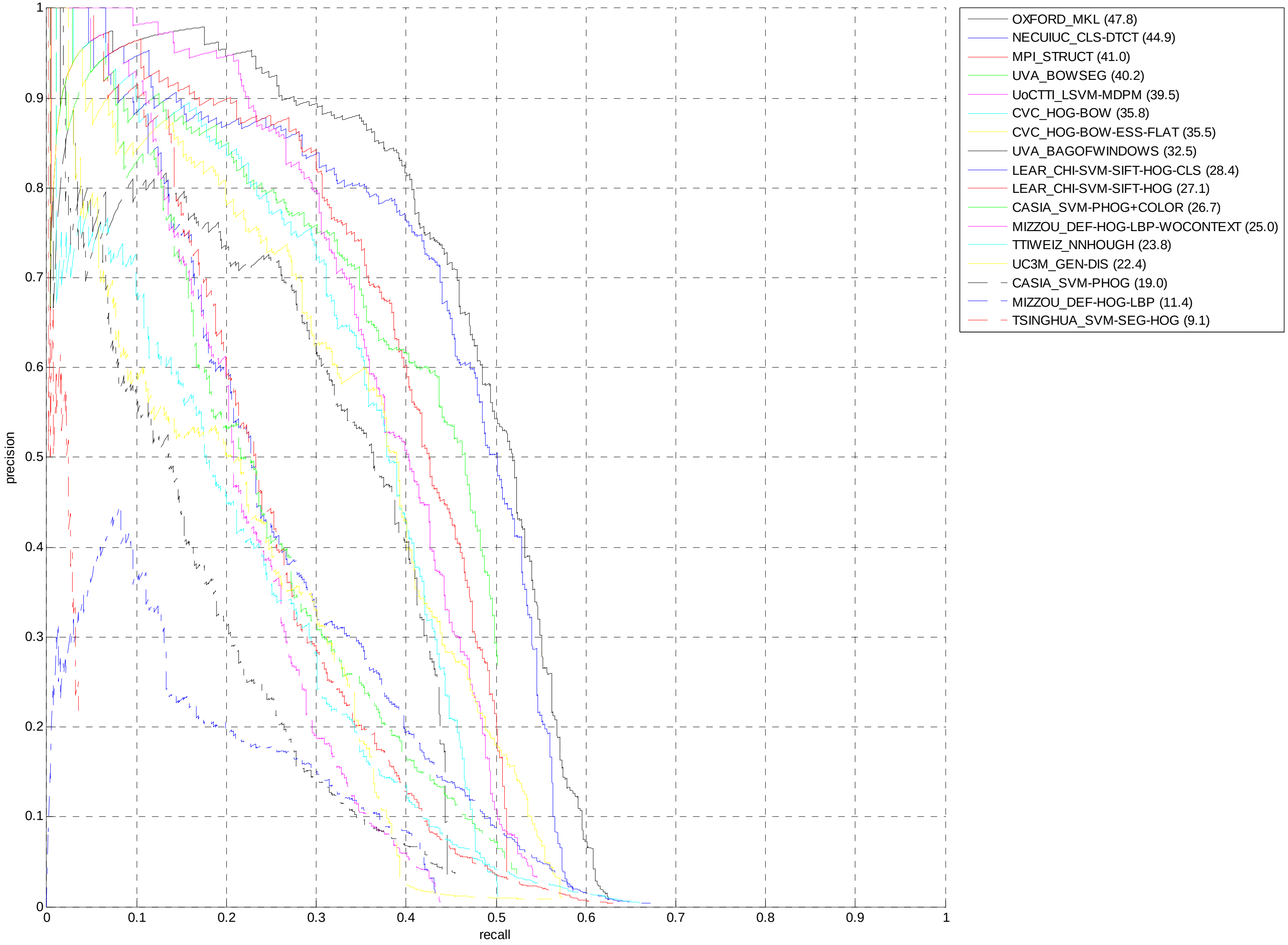
- 18 Methods, 12 Groups
- VOC2008: 8 Methods, 8 Groups
- 1 use of external data (BERKELEY_POSELETS)
- Wide variety of methods: sliding window, combination with whole-image classifiers, segmentation-based

AP by Method and Class

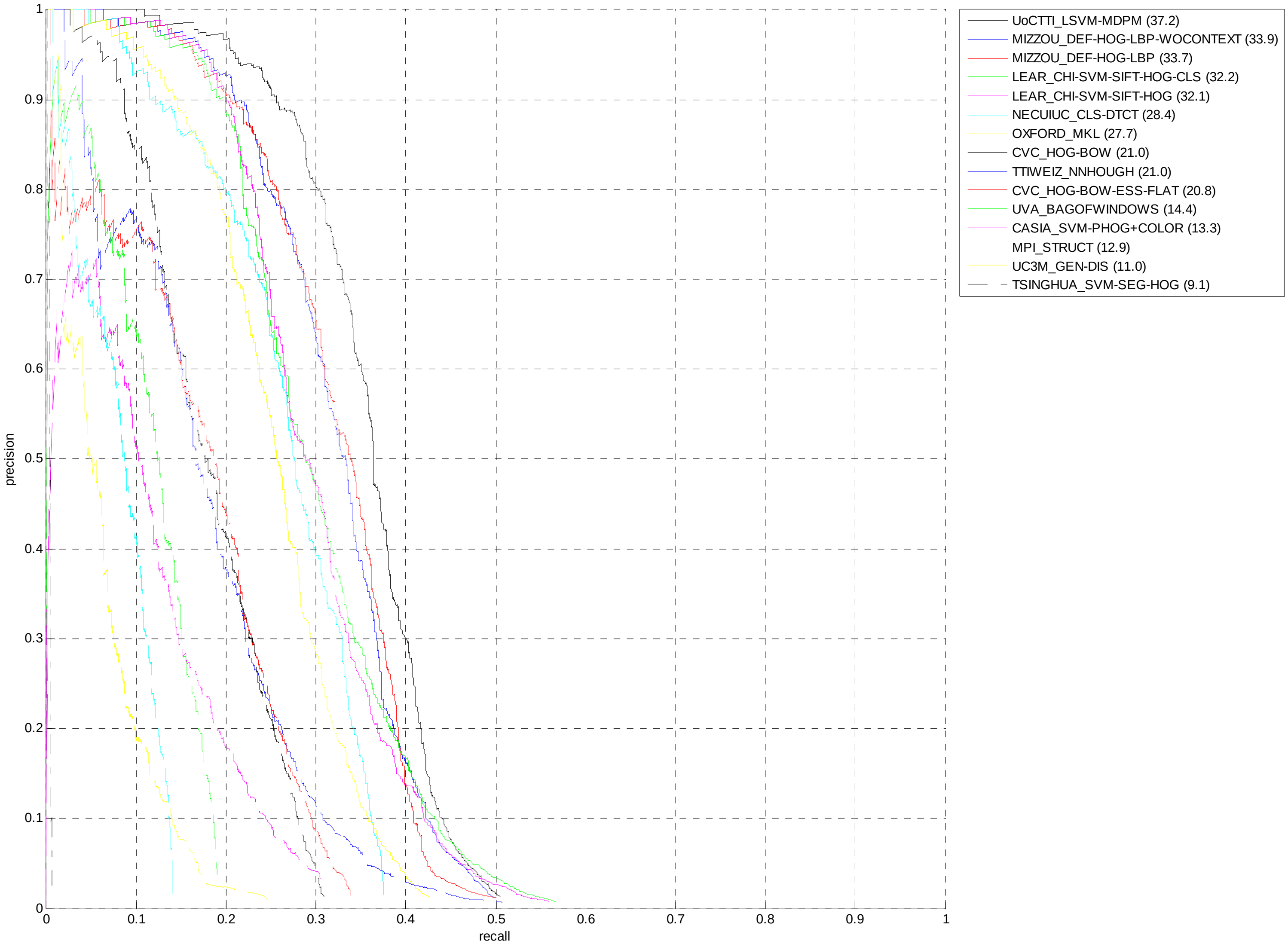
	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor
CASIA_SVM-PHOG	19.0	15.4	9.7	9.5	-	21.0	-	-	2.8	-	2.4	-	-	-	-	-	-	-	16.1	-
CASIA_SVM-PHOG+COLOR	26.7	20.5	10.2	10.2	9.5	26.6	13.3	12.7	9.5	7.6	10.2	11.1	16.6	22.1	15.8	9.4	4.2	10.1	25.3	16.1
CVC_HOG-BOW	35.8	27.6	10.2	10.1	17.2	32.1	21.0	18.9	13.0	10.9	17.1	14.2	24.5	28.8	18.0	10.3	16.0	13.1	25.9	27.3
CVC_HOG-BOW-ESS-FLAT	35.5	27.5	11.1	11.2	16.7	32.2	20.8	19.2	13.9	14.6	16.3	12.1	29.0	29.0	18.8	11.6	18.4	19.4	30.6	26.6
LEAR_CHI-SVM-SIFT-HOG	27.1	30.2	9.8	10.7	19.6	36.0	32.1	12.5	11.2	14.0	16.4	10.2	22.6	27.8	19.9	11.6	16.5	11.9	34.5	32.1
LEAR_CHI-SVM-SIFT-HOG-CLS	28.4	30.7	11.0	12.4	21.4	36.2	32.2	14.1	12.0	18.5	17.8	15.6	25.7	29.5	20.5	12.8	20.8	14.2	35.1	34.7
MIZZOU_DEF-HOG-LBP	11.4	27.5	6.0	11.1	27.0	38.8	33.7	25.2	15.0	14.4	16.9	15.1	36.3	40.9	37.0	13.2	22.8	9.6	3.5	32.1
MIZZOU_DEF-HOG-LBP-WOC	25.0	27.9	6.1	10.2	26.6	38.0	33.9	21.9	14.5	17.5	16.8	17.0	35.3	40.0	36.6	11.7	22.3	15.6	33.6	32.7
MPI_STRUCT	41.0	22.4	10.6	12.0	9.1	30.2	12.9	31.1	4.5	13.7	15.0	21.2	21.3	29.9	11.6	9.1	10.5	22.4	30.3	11.3
NECUIUC_CLS-DTCT	44.9	33.1	12.3	10.5	11.0	43.4	28.4	30.9	11.1	20.1	22.9	25.1	33.7	38.2	22.5	11.0	22.9	23.4	32.1	24.8
OXFORD_MKL	47.8	39.8	17.4	15.8	21.9	42.9	27.7	30.5	14.6	20.6	22.3	17.0	34.6	43.7	21.6	10.2	25.1	16.6	46.3	37.6
TSINGHUA_SVM-SEG-HOG	9.1	-	-	2.3	9.1	-	9.1	-	-	0.0	-	0.4	-	9.1	1.2	0.0	0.0	-	1.1	0.0
TTIWEIZ_NNHOUGH	23.8	24.0	-	-	-	21.9	21.0	-	-	14.3	-	-	19.6	24.0	-	-	-	-	-	23.2
UC3M_GEN-DIS	22.4	17.1	10.4	9.5	9.1	18.6	11.0	22.0	9.2	10.0	10.5	16.5	15.1	21.8	11.5	9.2	9.9	11.4	17.1	2.6
UoCTTI_LSVM-MDPM	39.5	46.8	13.5	15.0	28.5	43.8	37.2	20.7	14.9	22.8	8.7	14.4	38.0	42.0	41.5	12.6	24.2	15.8	43.9	33.5
UVA_BAGOFWINDOWS	32.5	23.7	10.6	8.4	3.2	28.2	14.4	33.7	1.2	13.2	16.3	23.2	24.6	30.7	13.1	4.5	9.3	28.0	29.0	9.5
UVA_BOWSEG	40.2	-	6.9	-	-	26.4	-	34.0	-	-	19.0	-	-	-	-	-	-	21.2	27.2	-

- Highlighted: 1st, 2nd or 3rd place by method
- Groups: LEAR, MIZZOU, MPI, NEC/UIUC, OXFORD, UoCTTI, UVA

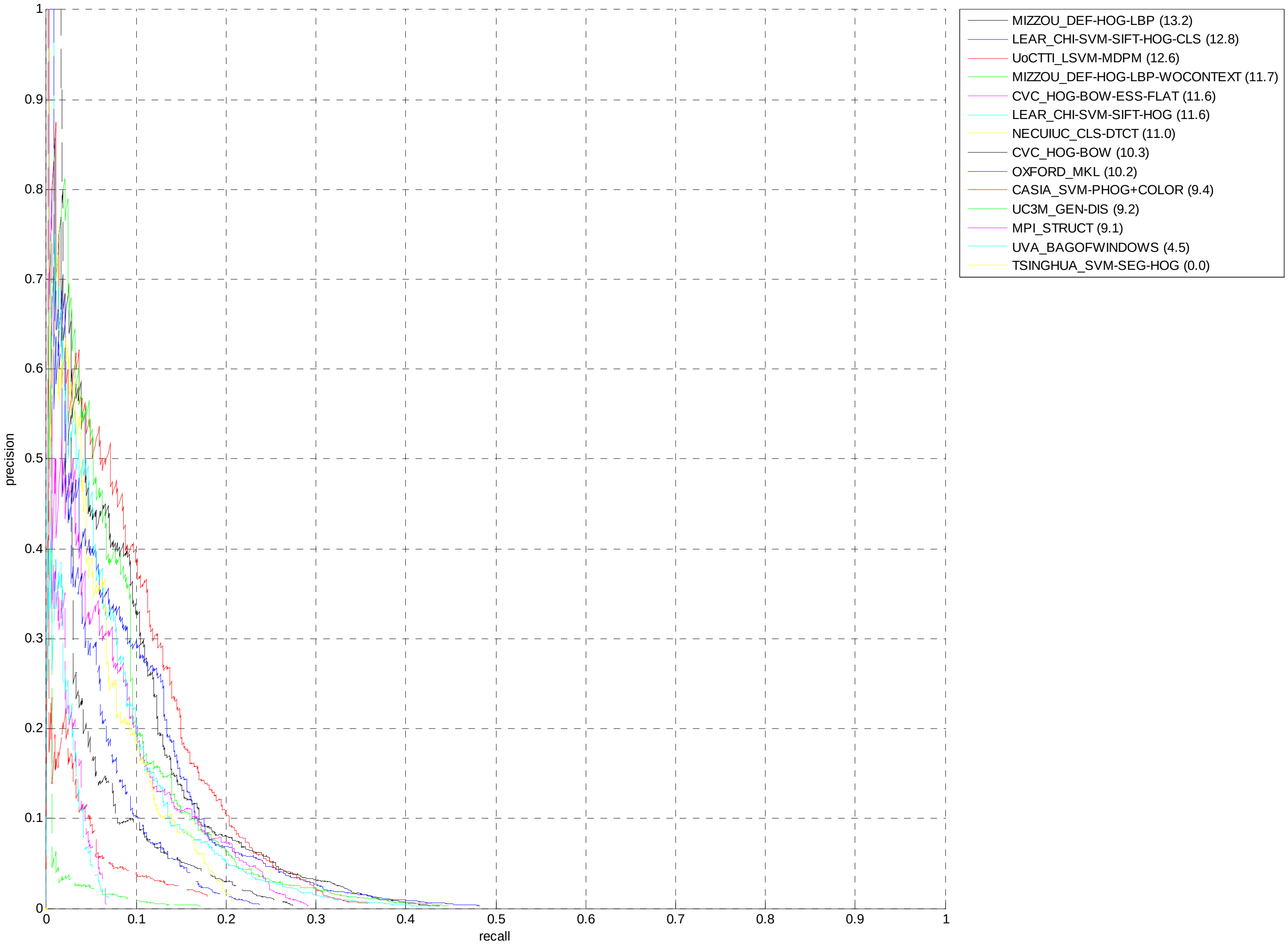
Precision/Recall - Aeroplane



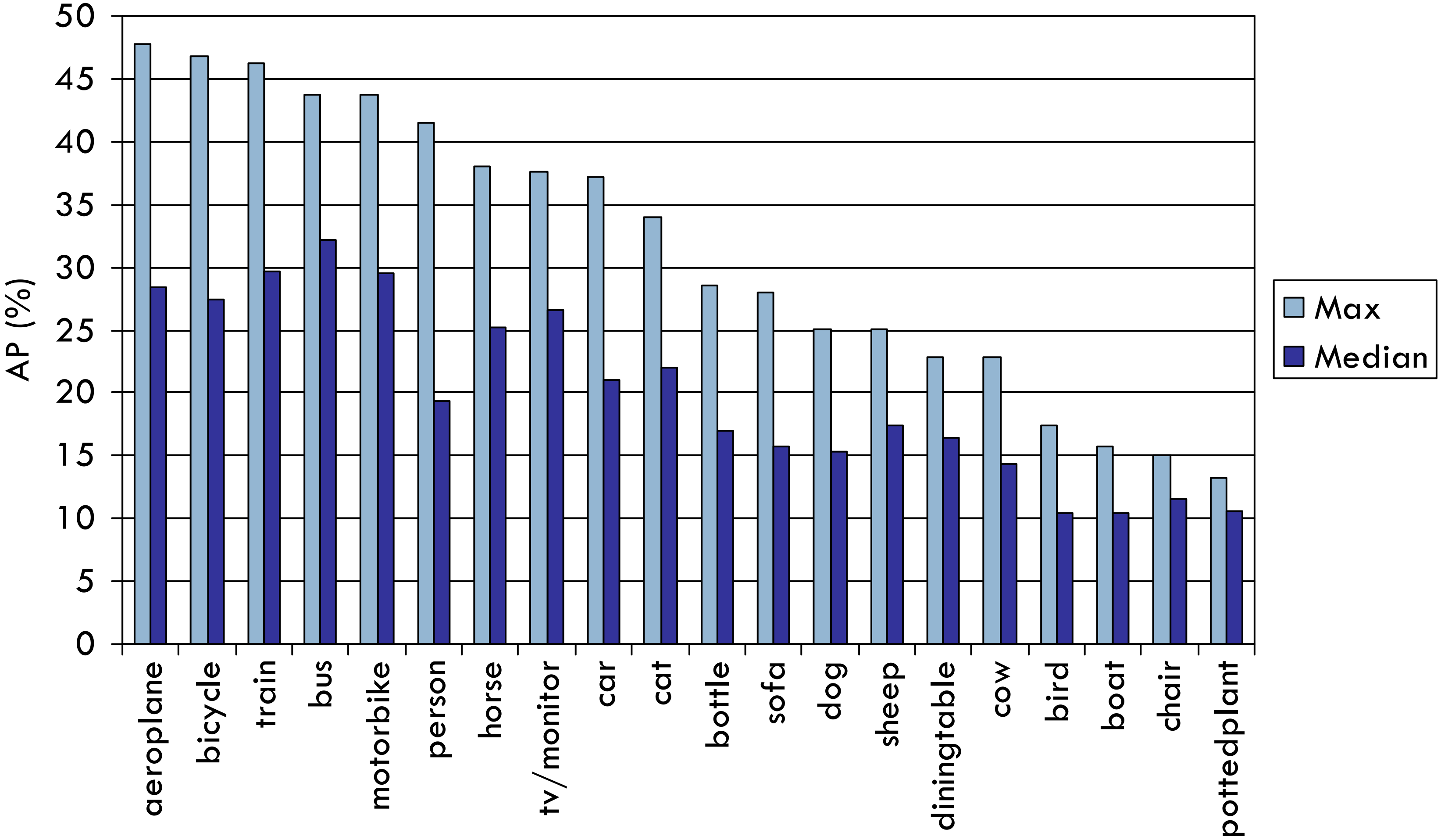
Precision/Recall - Car



Precision/Recall – Potted plant



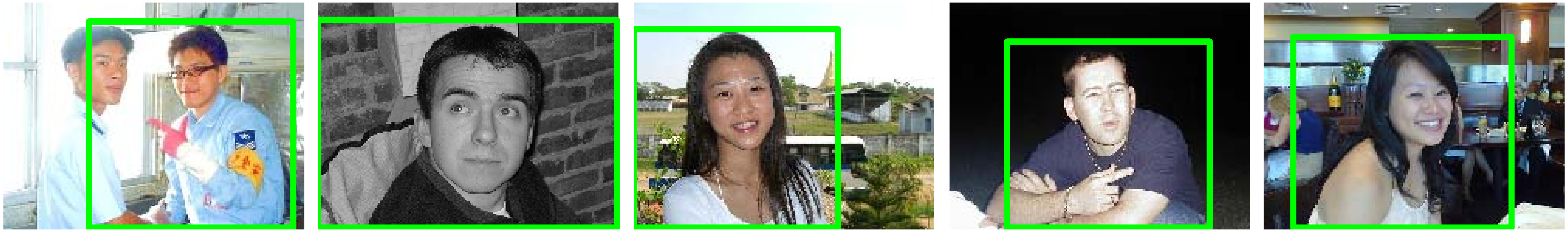
AP by Class



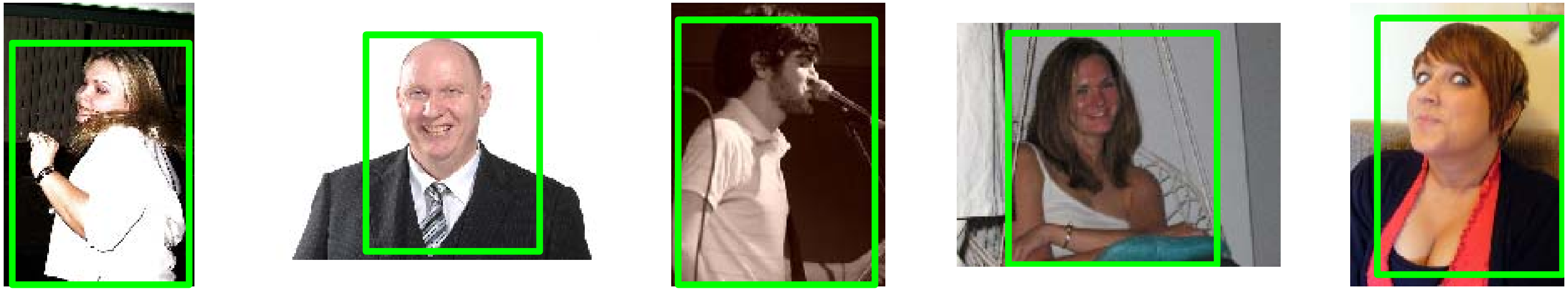
AP (%)
50
45
40
35
30
25
20
15
10
5
0
aeroplane
bicycle
train
bus
motorbike
person
horse
tv/monitor
car
cat
bottle
sofa
dog
sheep
diningtable
cow
bird
boat
chair
pottedplant
Max
Median

True Positives - Person

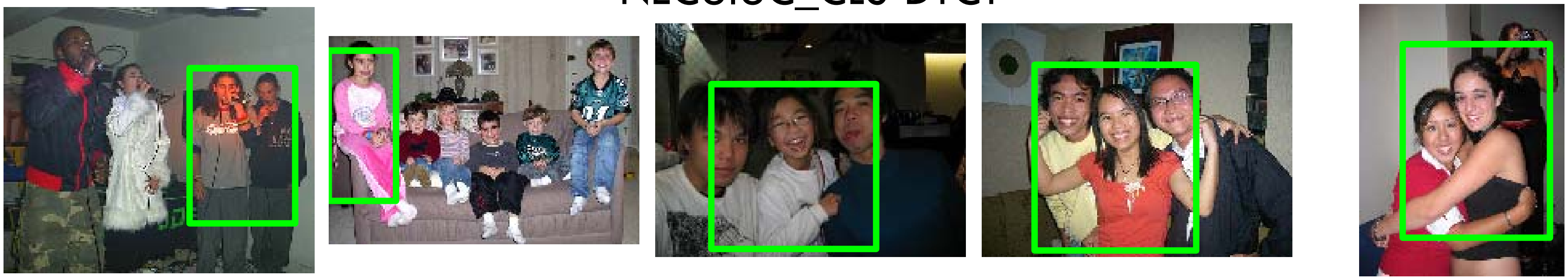
UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP

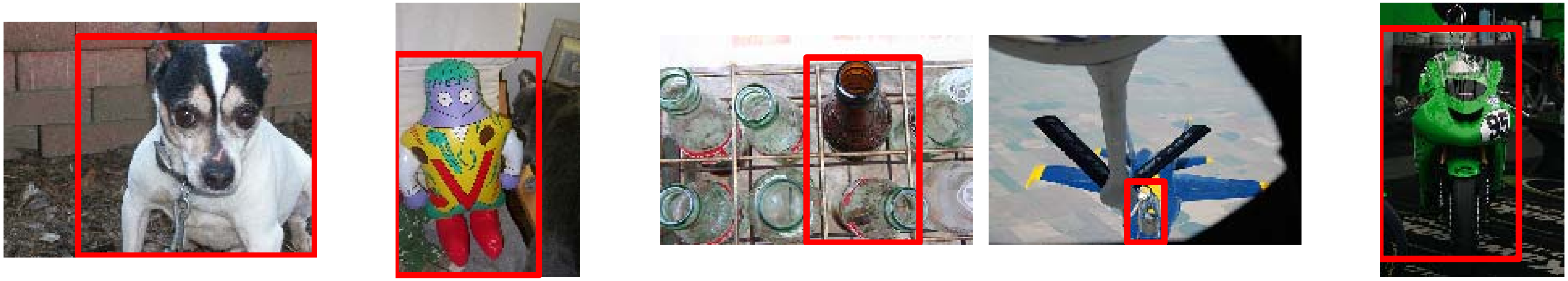


NECUIUC_CLS-DTCT

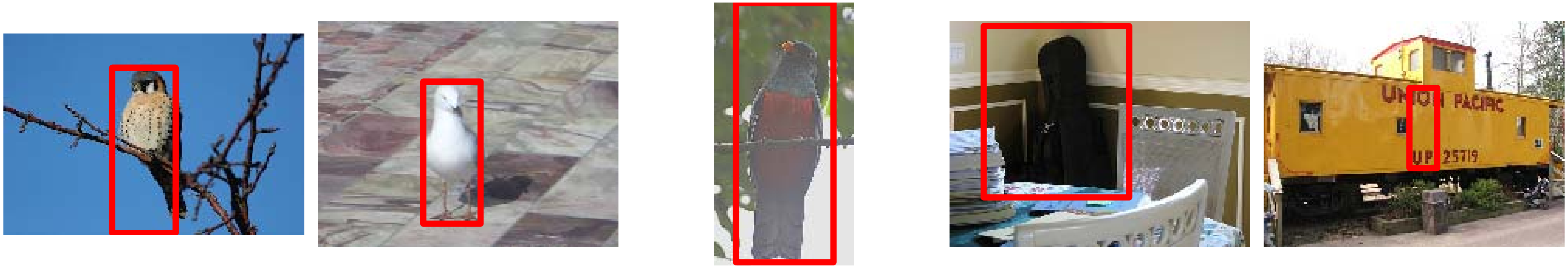


False Positives - Person

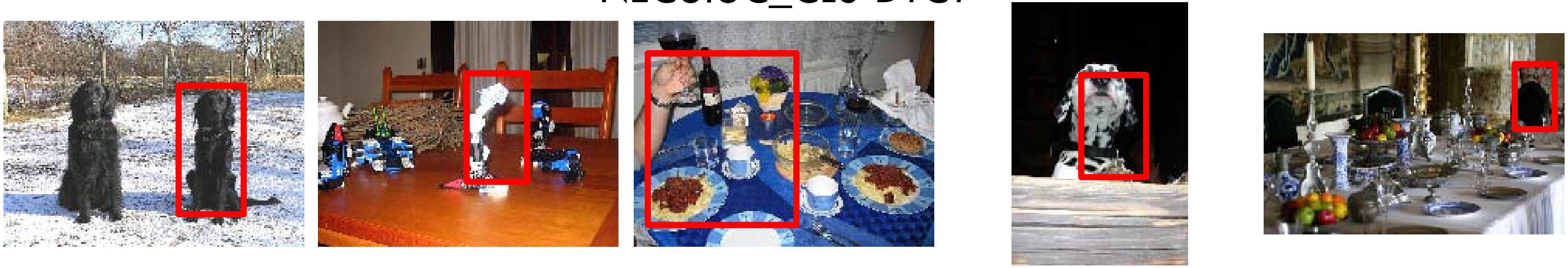
UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP

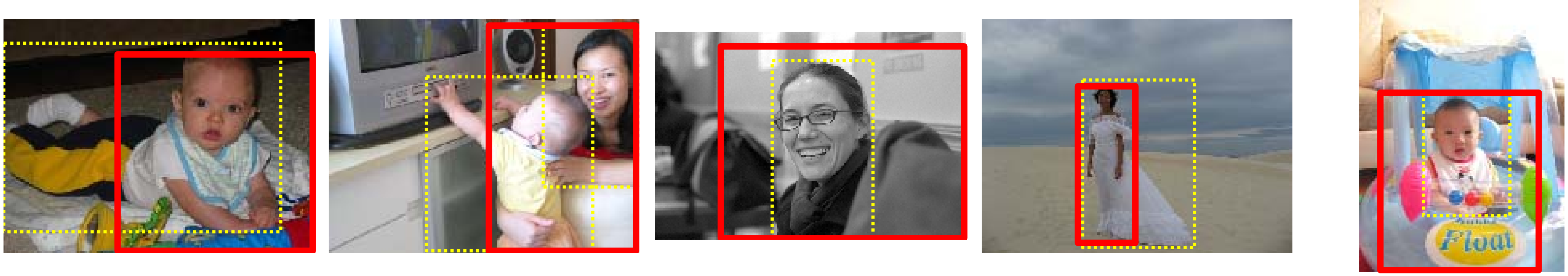


NECUIUC_CLS-DTCT

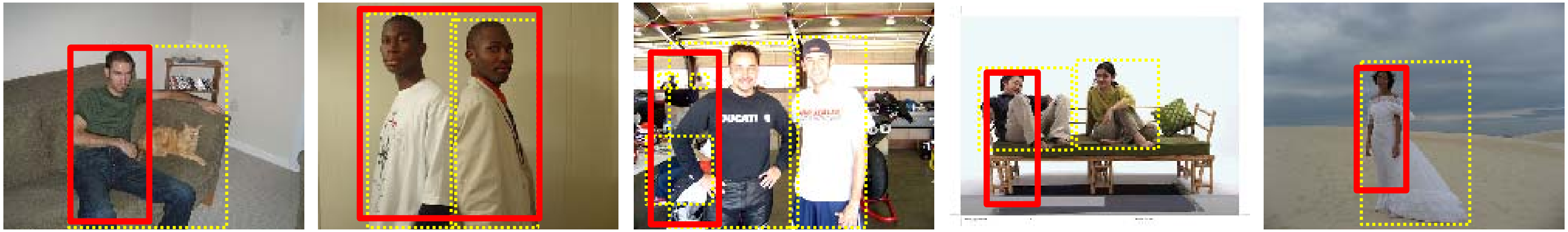


“Near Misses” - Person

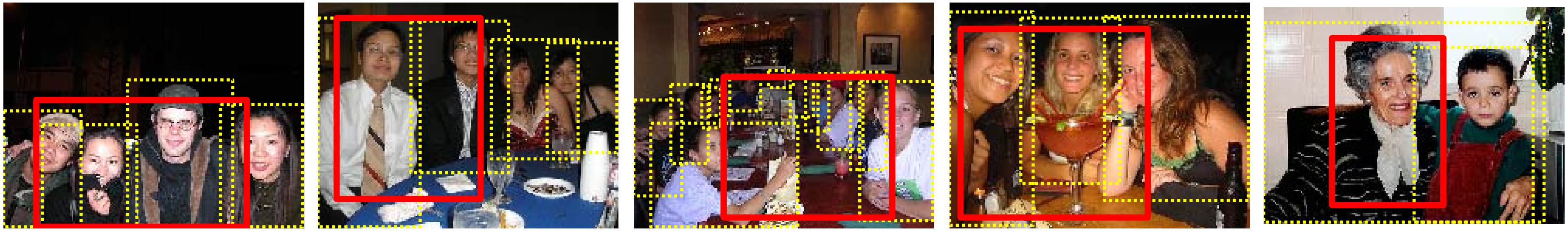
UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP

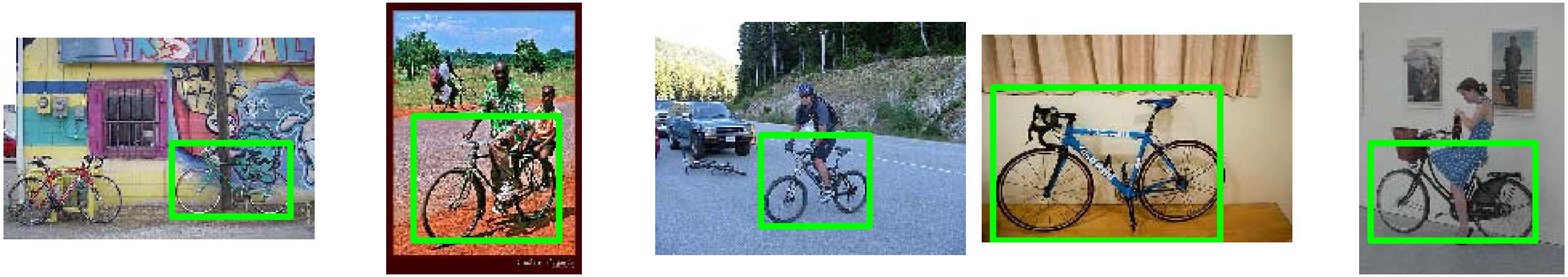


NECUIUC_CLS-DTCT

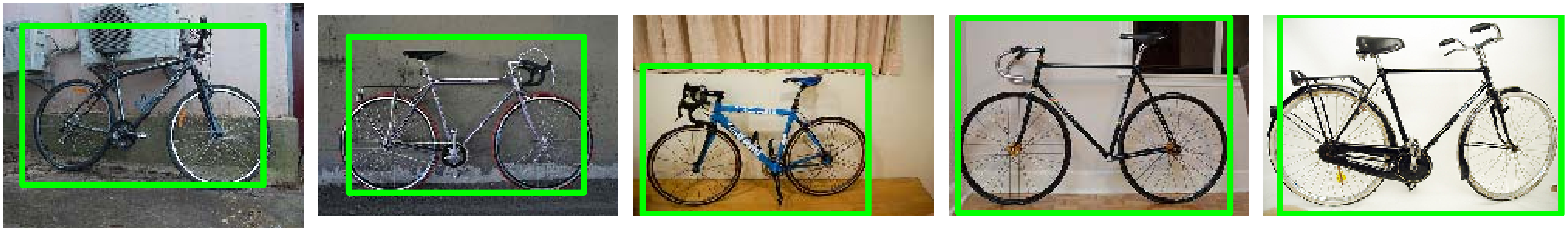


True Positives - Bicycle

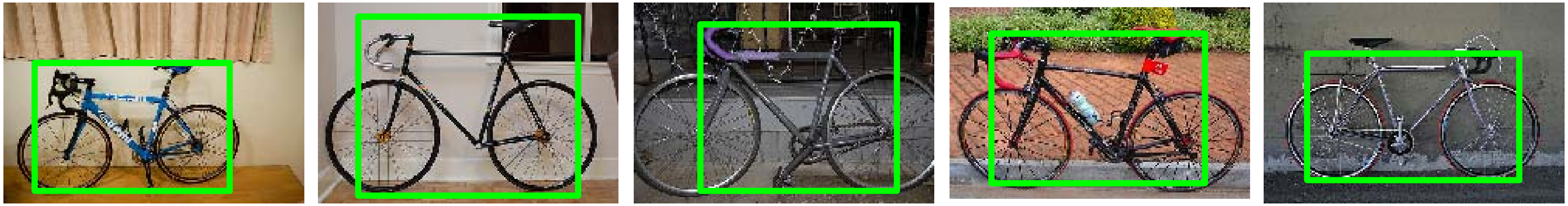
UoCTTI_LSVM-MDPM



OXFORD_MKL

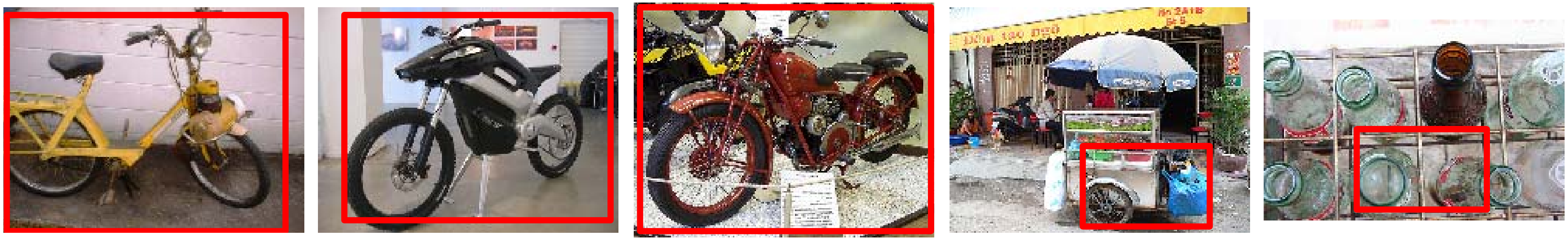


NECUIUC_CLS-DTCT

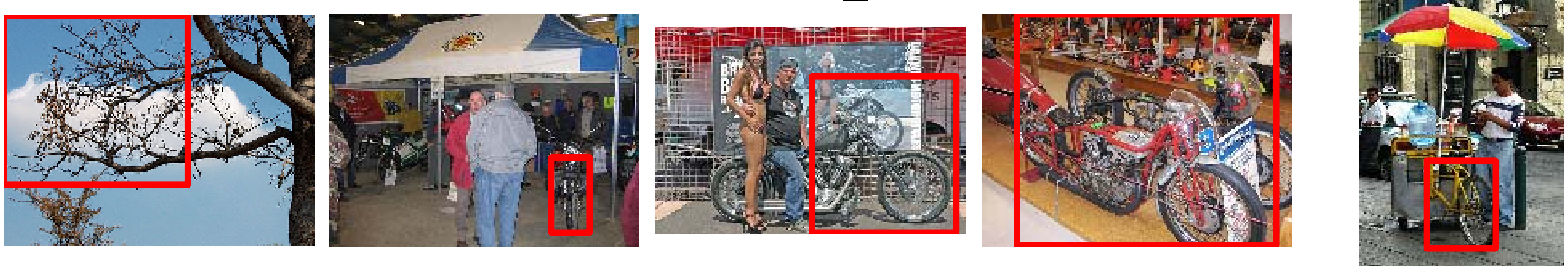


False Positives - Bicycle

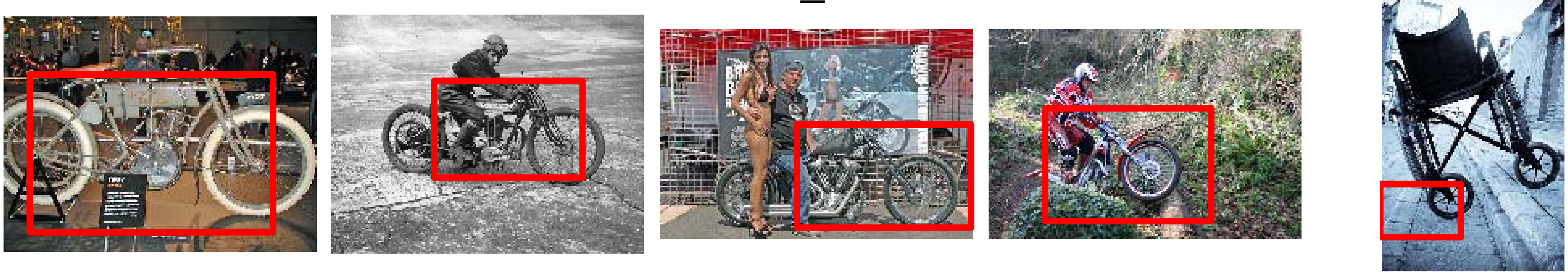
UoCTTI_LSVM-MDPM



OXFORD_MKL

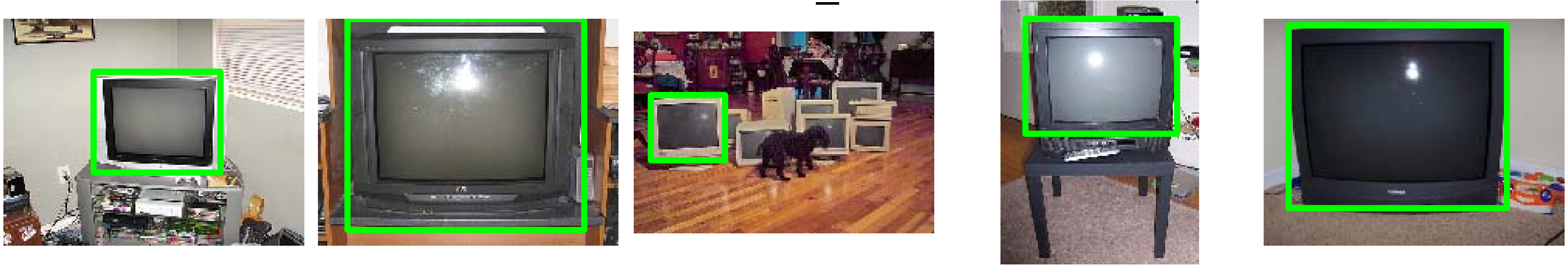


NECUIUC_CLS-DTCT



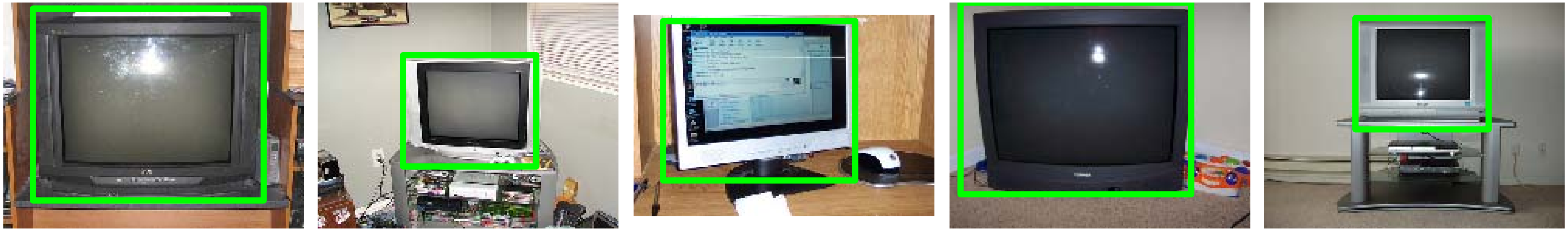
True Positives – TV/monitor

OXFORD_MKL



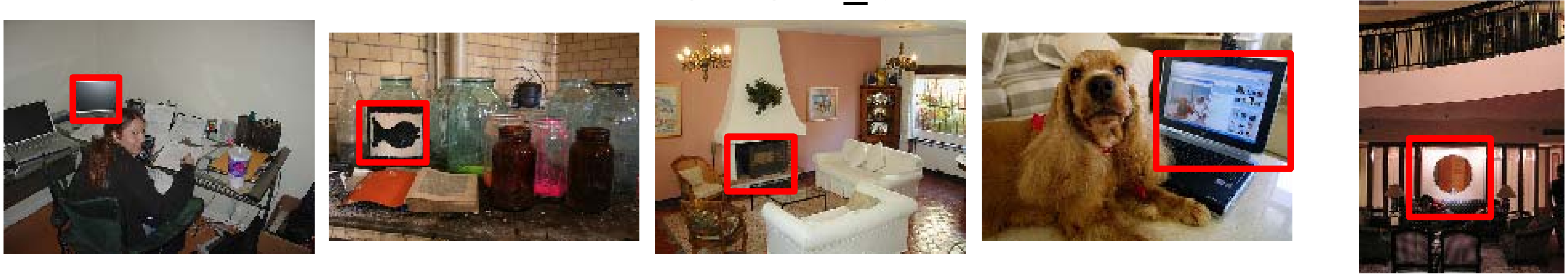
UoCTTI_LSVM-MDPM

LEAR_CHI-SVM-SIFT-HOG-CLS

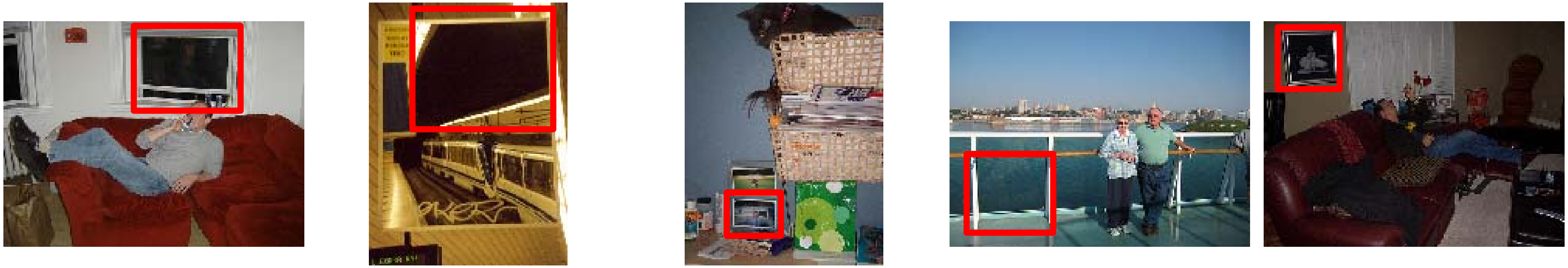


False Positives – TV/monitor

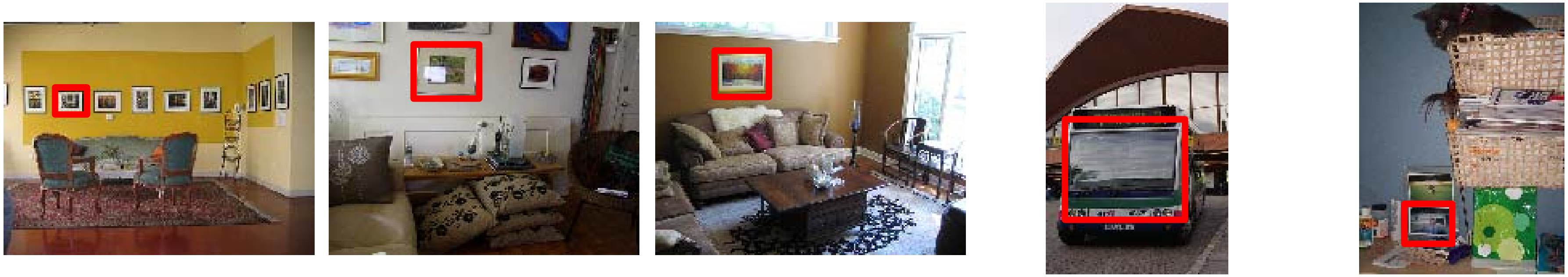
OXFORD_MKL



UoCTTI_LSVM-MDPM

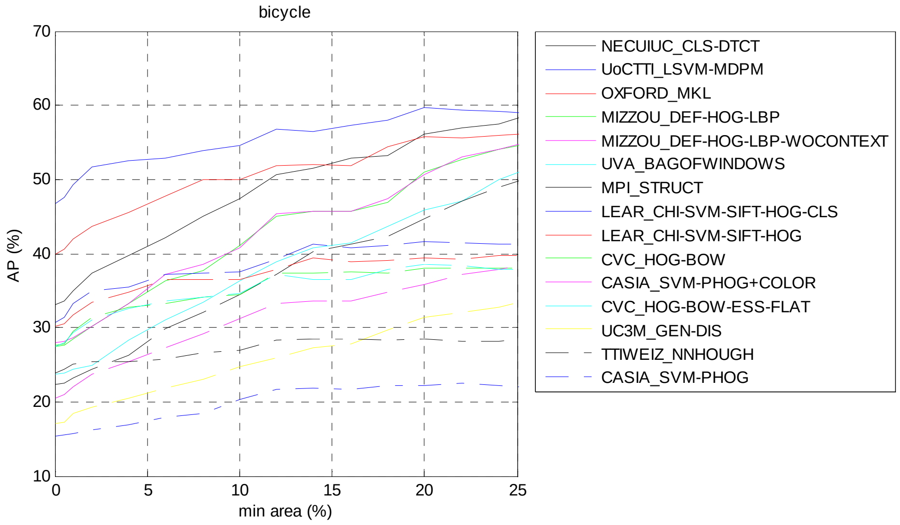


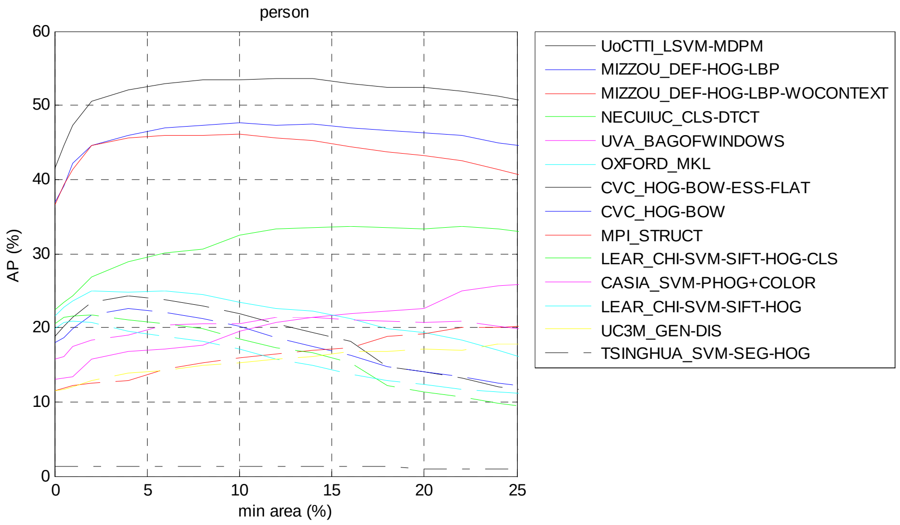
LEAR_CHI-SVM-SIFT-HOG-CLS



AP vs. Object Area

- Do these methods have a bias toward larger objects?

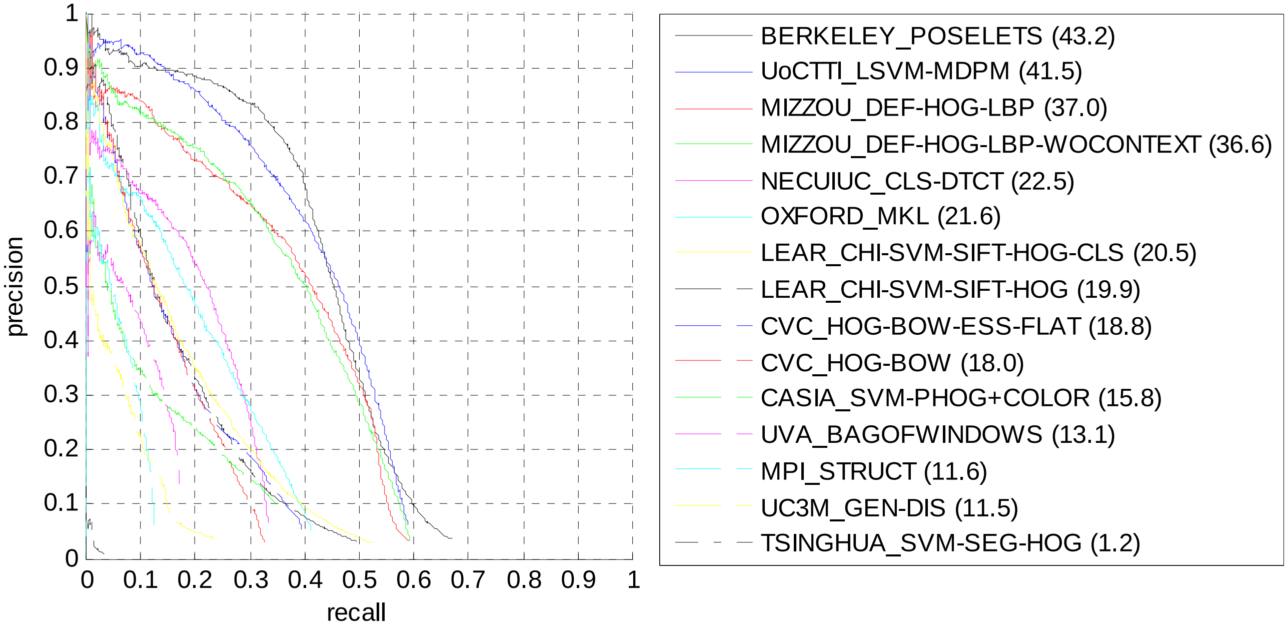




- Most methods show moderate preference for larger objects – use of bag of words stages and whole-image classifiers?
- For some objects accuracy reduces for large objects – occlusion?

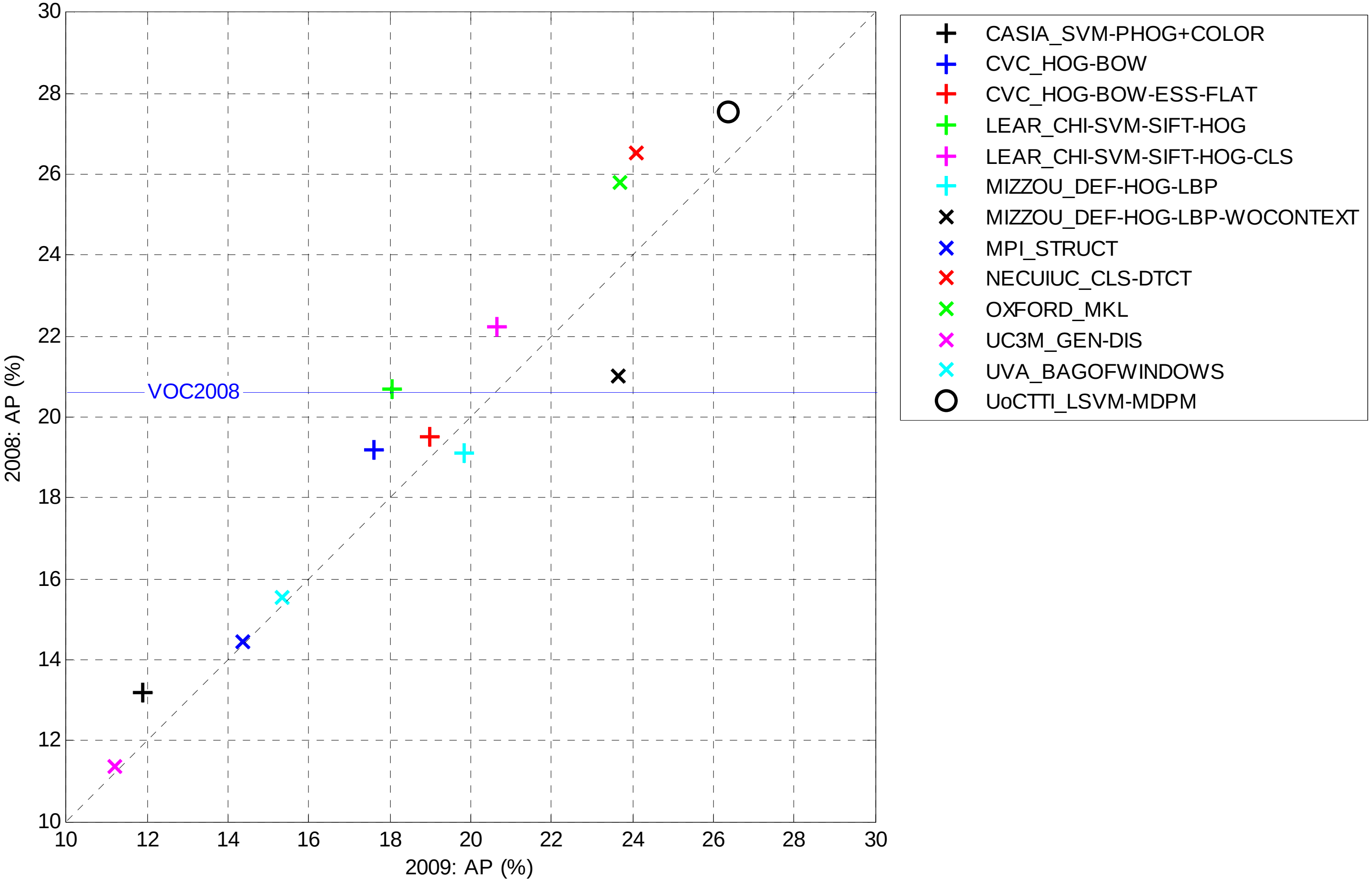
External Training Data

- BERKELEY_POSELETS method for "person" uses external training data based on 3D annotation



- Modest improvement over methods using VOC training data: 43.2% vs. 41.5% AP (UoCTTI)

VOC2008 vs. VOC2009 Test Data



- High correlation, generally better results on 2008
- Best methods are better than best 2008 result – better methods and/or advantage of more training data

Prizes

- Joint Winners:
  - **UoC/TTI Chicago**

    Pedro Felzenszwalb[1], Ross Girshick[1], David McAllester[2]
    [1]University of Chicago; [2]Toyota Technological Institute at Chicago
  - **Oxford/MSR India**

    Andrea Vedaldi[1], Varun Gulshan[1], Manik Varma[2], Andrew Zisserman[1]
    [1]University of Oxford; [2]Microsoft Research India