Regionlets for Generic Object Detection

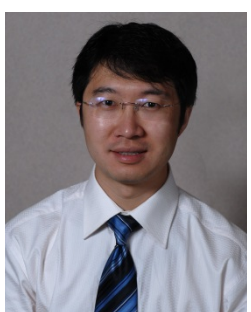
A test on ImageNet

Xiaoyu Wang †

Miao Sun ‡

Tianbao Yang †

Yuanqing Lin †



Tony X. Han ‡

Shenghuo Zhu †

† NEC Laboratories America

Relentless passion for innovation

‡ University of Missouri

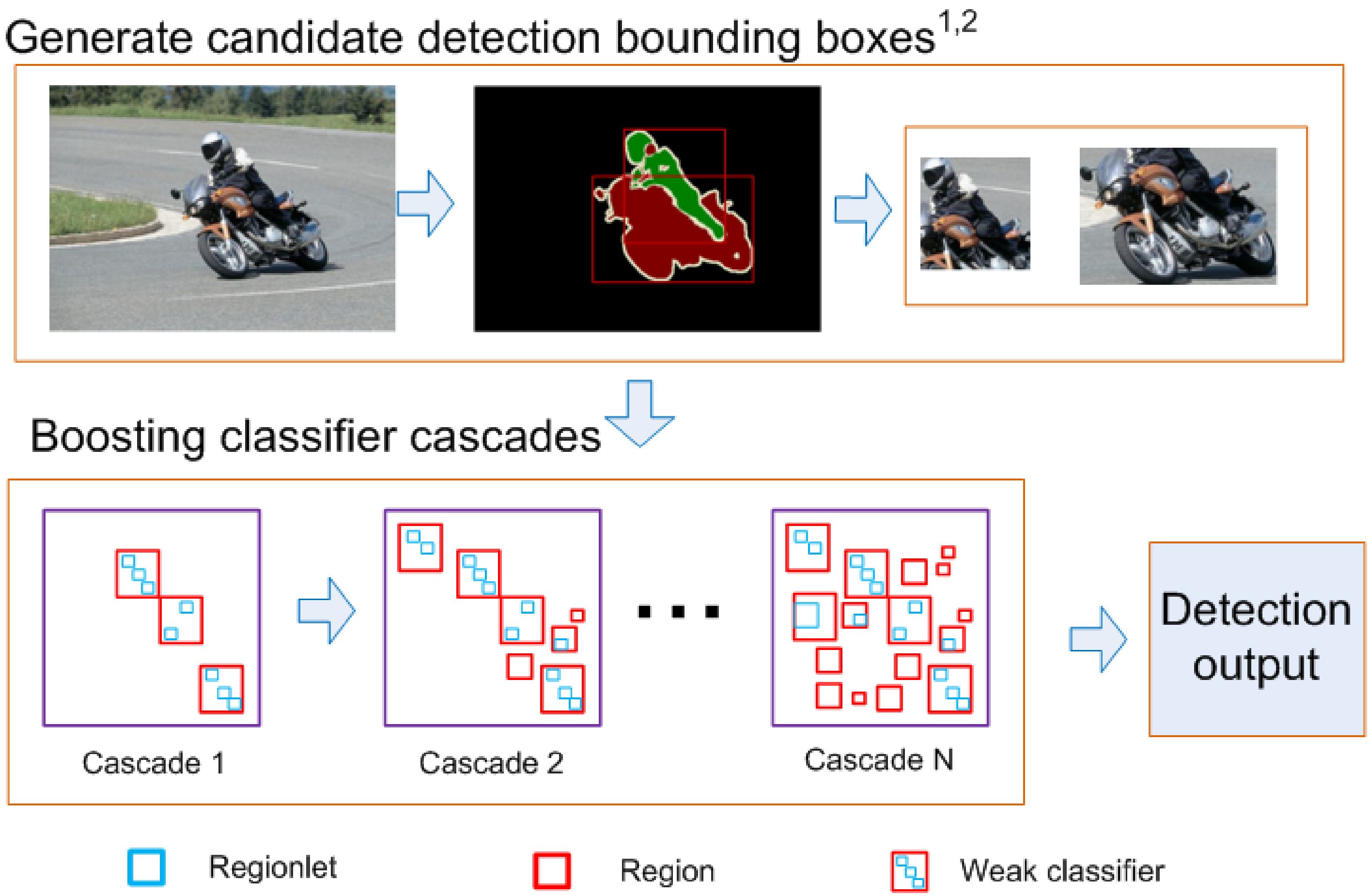
Introduction

- Generic object detection is challenging
  - Rich deformation
  - Arbitrary scales
  - Arbitrary viewpoints
- Limitations of current state of the art
  - Hand-crafted parameters to handle different degrees of deformation
  - Sub-optimal multiple scales/viewpoints handling

Motivation

- A flexible and general object-level representation
  - Data-driven deformation handling
  - Multiple scales/viewpoints handling using a single and flexible model (Detecting an object at its original scale and aspect ratio)
  - Fast and easy to be extended with different features

Detection Framework[3]

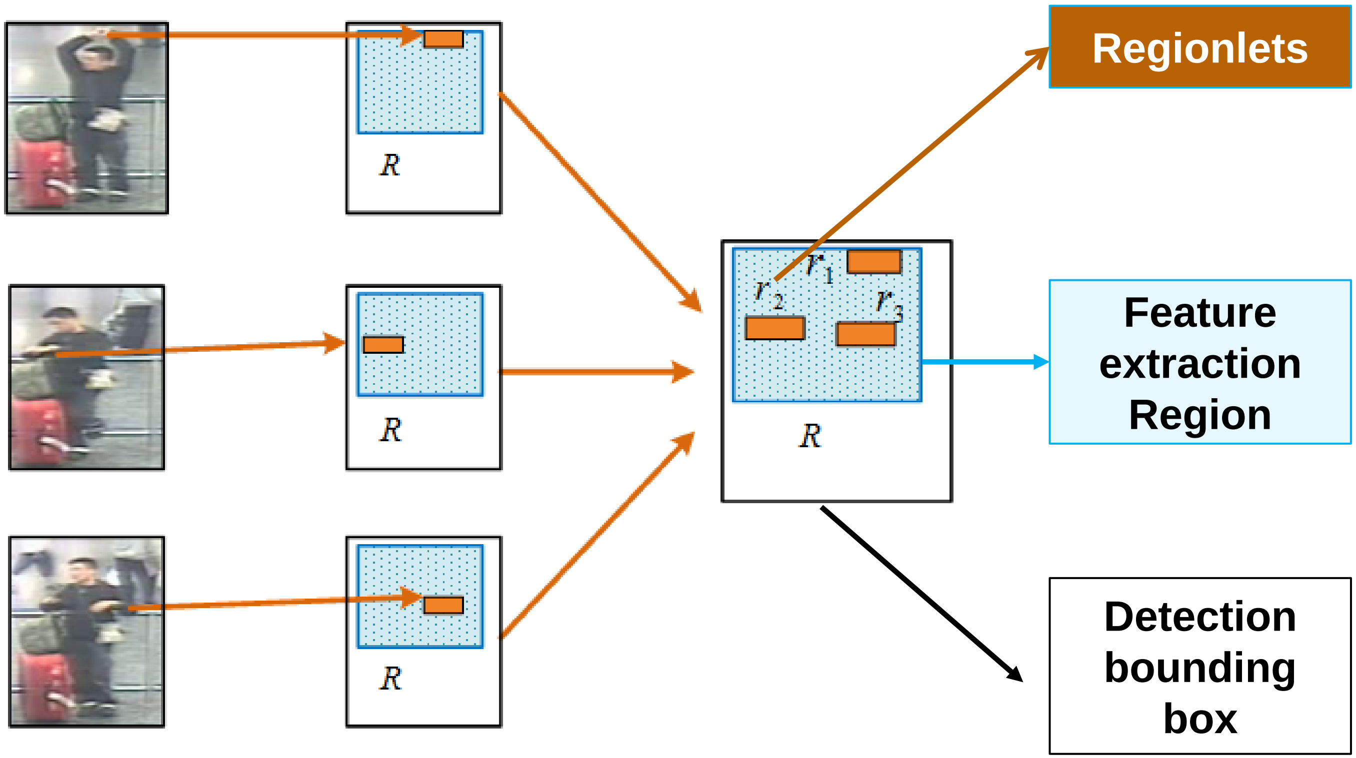


1. B. Alexe , et. al. What is an object? CVPR 2010
2. K. E. A. Van de Sande, et. al. Segmentation as selective search for object recognition. ICCV 2011
3. X. Wang, et. al. Regionlets for Generic Object Detection. ICCV 2013

Regionlet: Definition

- What is regionlet?
  - Region(R): Feature extraction region
  - Regionlet(r_1, r_2, r_3): A sub-region in a feature extraction area whose position/resolution are relative and normalized to a detection window

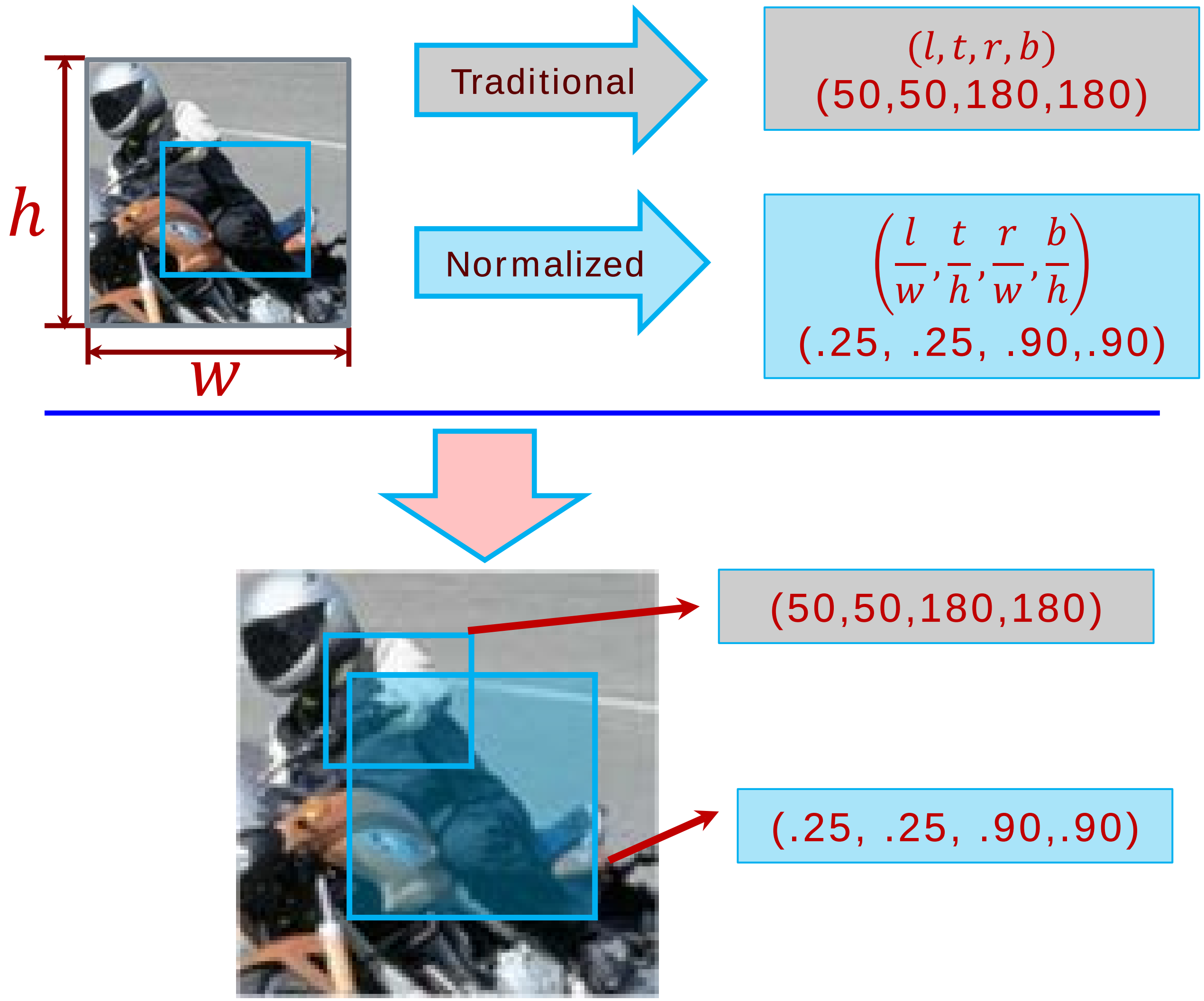
Figure 1



Regionlet: Definition(*cont.*)

- Relative normalized position

Figure 2



Regionlet: Feature extraction

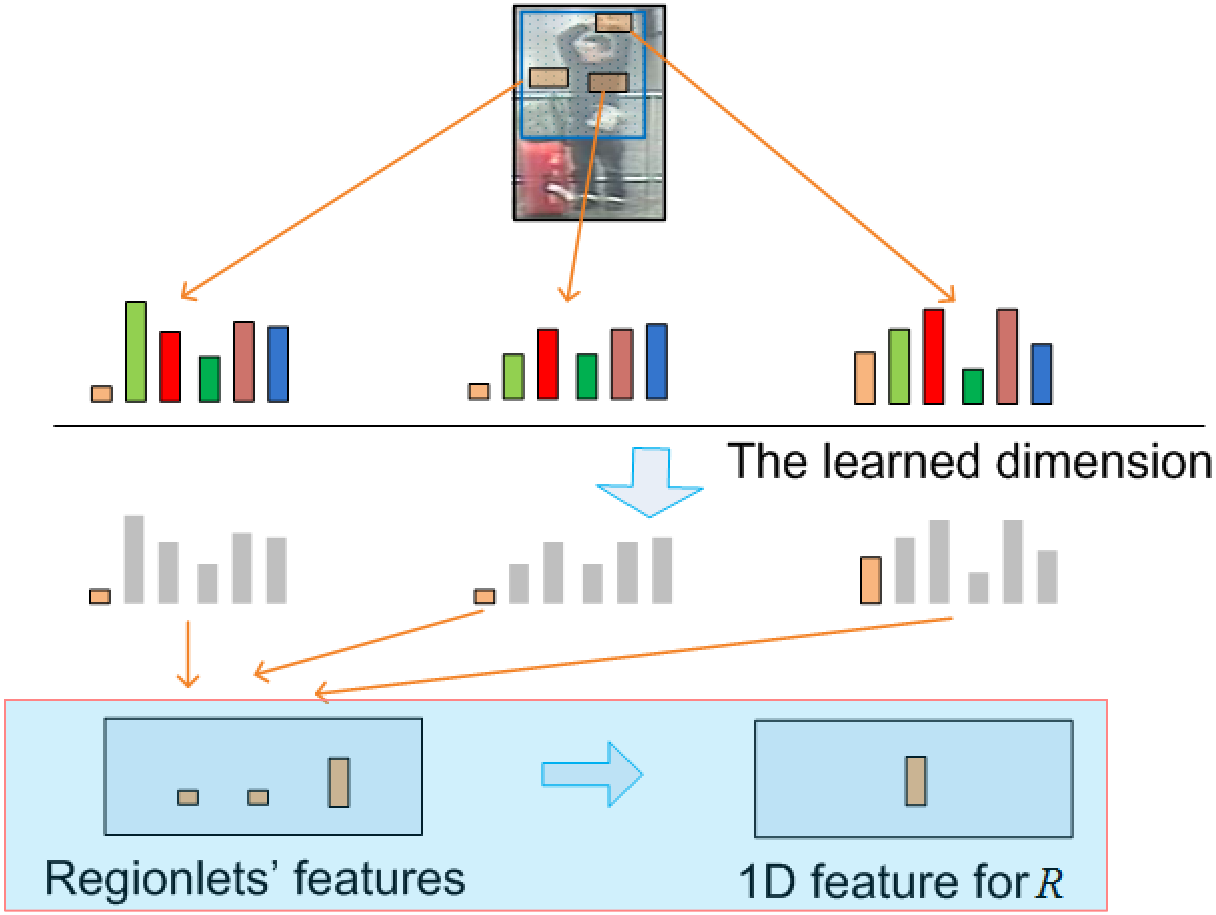


Figure 3

Non-local pooling

Could be SIFT, HOG, LBP , Covariance features, whatever feature your like!

Regionlets: Training

- Constructing the regions/regionlets pool
  - Uniformly sample the position/configuration space of regions/regionlets
- Learning realBoost[1] cascades
  - 16K region/regionlets candidates for each cascade
  - Learning of each cascade stops when the error rate is achieved (1% for positive, 37.5% for negative)
  - Last cascade stops after collecting 5000 weak classifiers
  - Result in 4-7 cascades
  - 2-3 hours to finish training one category on a 8-core machine

1. C. Huang, et. al. Boosting nested cascade detector for multi-view face detection. *ICPR*, 2004.

Deformation Handling

- Two-layers deformation handling
  - Data-driven feature extraction region
    - Larger region -> more robust to deformation
    - Small region -> finer spatial layout
  - Data-driven non-local max-pooling over regionlets
    - Permutation invariance among regionlets
    - Exclusive feature representation among regionlets

Scale/viewpoints Handling

- Arbitrary scale/viewpoints handling
  - Coordinates of regionlets are normalized in a model
  - Absolute regionlets coordinates are computed on the fly based on
    - The normalized coordinates
    - Resolution of the detection window

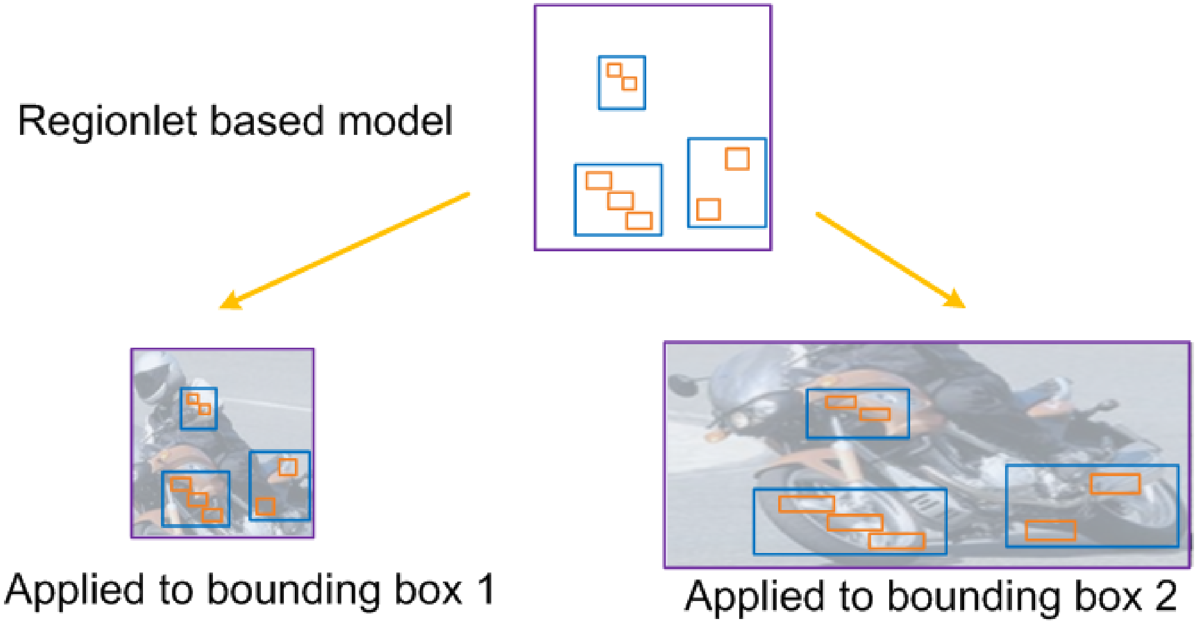


Figure 4

Experiments

- Datasets
  - PASCAL VOC 2007, 2010
    - 20 object categories
  - ImageNet Large Scale Object Detection Dataset
    - 200 object categories
- Investigated Features
  - HOG
  - LBP
  - Covariance
  - Deep Convolutional Neural Network (DCNN) feature

Regionlets on PASCAL

Table 1. Performance on the PASCAL VOC 2007 dataset (Evaluated using Average Precision or mean Average Precision: mAP, no DCNN feature, no outside data)

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM [12][1]	33.2	**60.3**	10.2	16.1	**27.3**	54.3	58.2	23.0	**20.0**	24.1	26.7	12.7	**58.1**	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
SS_SPM [25][2]	43.5	46.5	10.4	12.0	9.3	49.4	53.7	39.4	12.5	36.9	42.2	26.4	47.0	52.4	23.5	12.1	29.9	36.3	42.2	48.8	33.8
Objectness [3]	28.6	54.5	1.1	14.2	26.7	42.0	50.2	18.2	16.5	17.5	26.2	7.7	46.8	39.6	36.2	11.6	14.1	23.1	34.8	39.2	27.4
Regionlets-S	50.8	44.6	17.0	23.5	16.7	48.9	67.6	39.1	16.5	32.4	44.0	18.9	52.1	46.6	36.6	13.8	33.8	27.6	55.5	50.4	36.8
Regionlets-M	**54.2**	52.0	**20.3**	**24.0**	20.1	**55.5**	**68.7**	**42.6**	19.2	**44.2**	**49.1**	**26.6**	57.0	**54.5**	**43.4**	**16.4**	**36.6**	**37.7**	**59.4**	**52.3**	**41.7**

Table 2: Performance comparison with state of the art

	VOC 2007	VOC 2010	Results year
DPM(WC) [12]	35.4	33.4	2008
UCI_2009 [7]	27.1	N/A	2009
INRIA_2009 [13]	28.9	N/A	2009
NLPR(WC) [10]	N/A	36.8	2010
MITUCLA(WC) [10]	N/A	36.0	2010
UVA [10]	N/A	32.9	2010
MIT_2010 [31]	29.6	N/A	2010
Song *et al.* (WC) [23]	37.7	36.8	2011
Li *et al.* (WC) [18]	35.2	N/A	2011
SS_SPM [25]	33.8	34.1	2011
Cinbis *et al.* (WC) [5]	35.0	N/A	2012
Ours (Regionlets)	**41.7**	**39.7**	2013

Regionlets on PASCAL

- Regionlets with Deep CNN feature (outside data)

Table 3. Performance with Deep CNN feature

Deep CNN convolutional layer feature (outside data)	
CNN(ImageNet) + layer5 + SVM[1]	40.1%
CNN(ImageNet) + layer5 + Hand-crafted feature + Regionlets	**49.3%**

Deep CNN fine-tuned full connected layer feature (outside data)	
CNN(fine-tuned on PASCAL) + FC_7 + SVM[1]	48.0%

Will Regionlets model perform at 49.3% + 7.9% = 57.2% using fine-tuned full connected layer feature?

1. R Girshick, et. al. Rich feature hierarchies for accurate object detection and semantic segmentation. TR. 2013

Regionlets on ImageNet

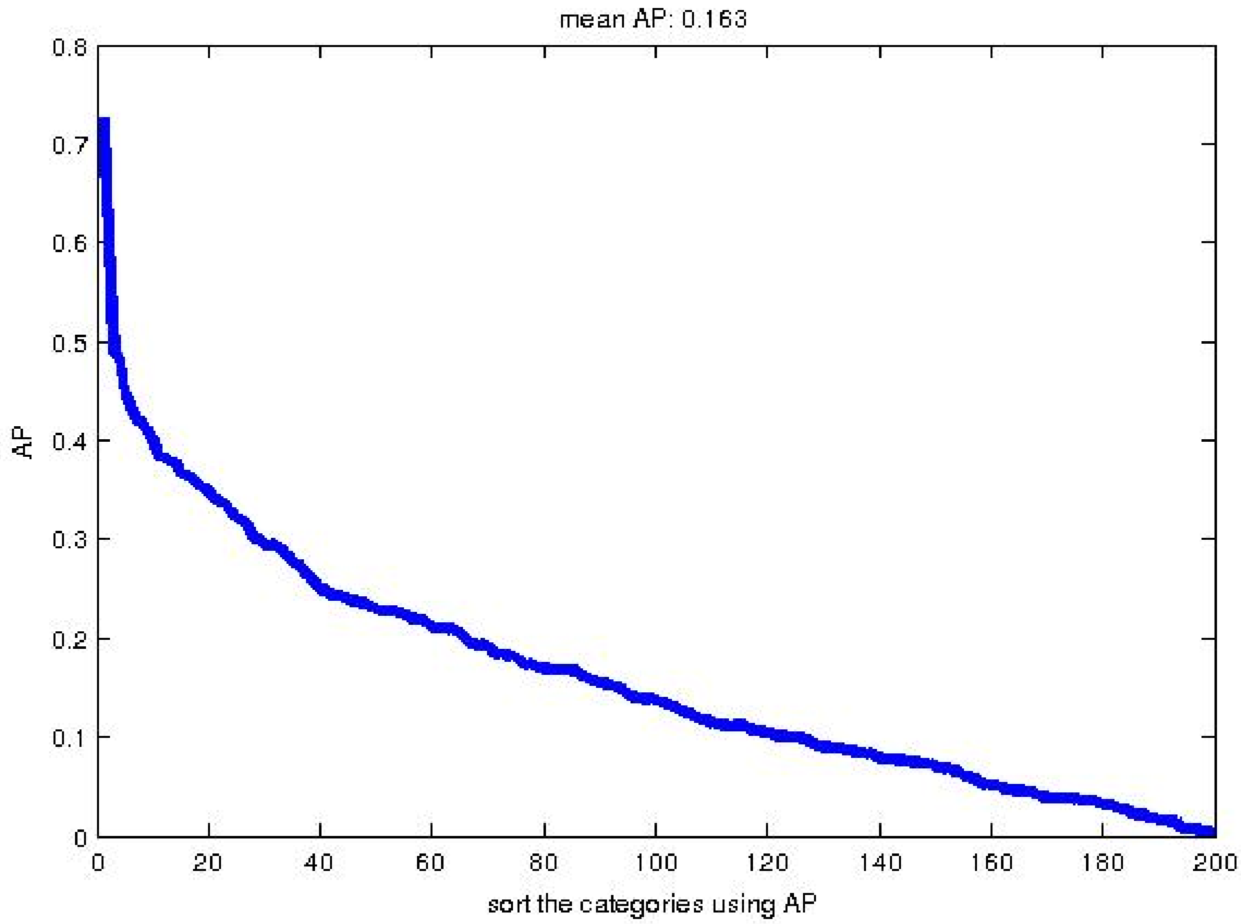
- ImageNet Challenge

Methods	mAP
UvA-EuVision	22.6% (with DCNN feature)
Regionlets with deep features[(1)]	20.9% (with DCNN feature)
Regionlets without deep features	**19.6% (no DCNN feature)**
OverFeat-NYU	19.4% (DCNN)
Toronto A	11.2% (N/A)
SYSU_Vision	10.5% (N/A)

(1) It's a preliminary result, we have a better performance now!

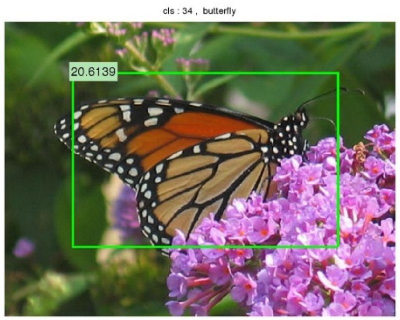
Regionlets on ImageNet

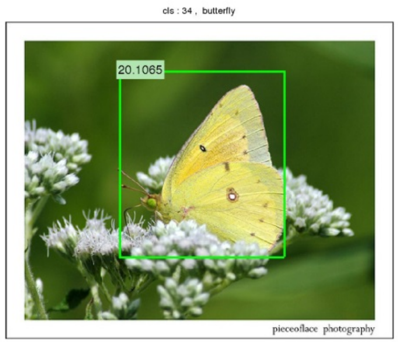
- Performance on the validation dataset

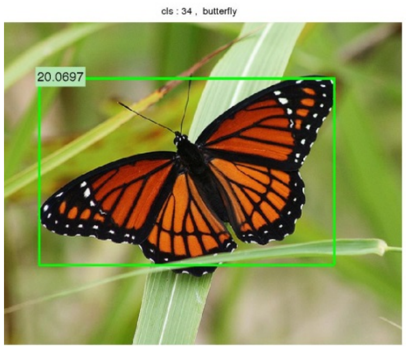


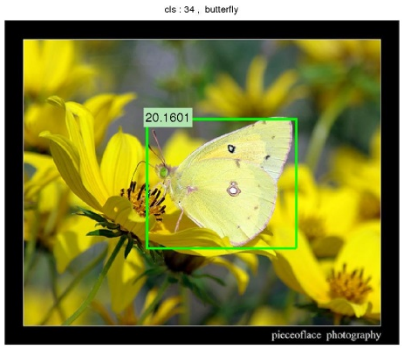
Regionlets on ImageNet

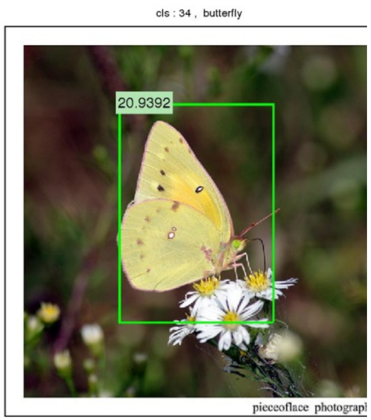
- Top 3 easiest categories: butterfly

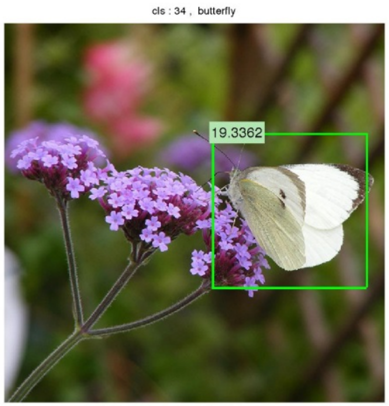


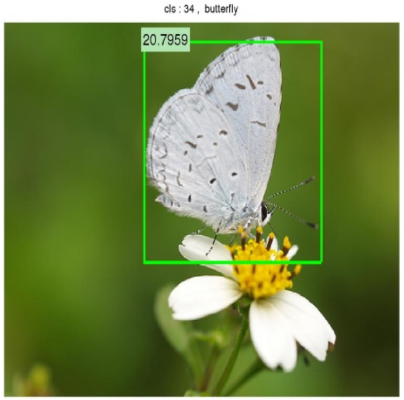






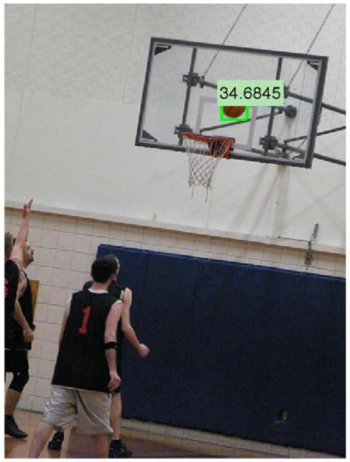


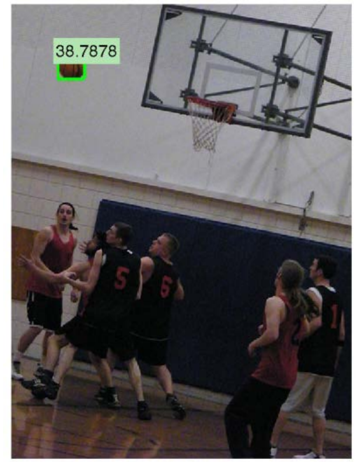




Regionlets on ImageNet

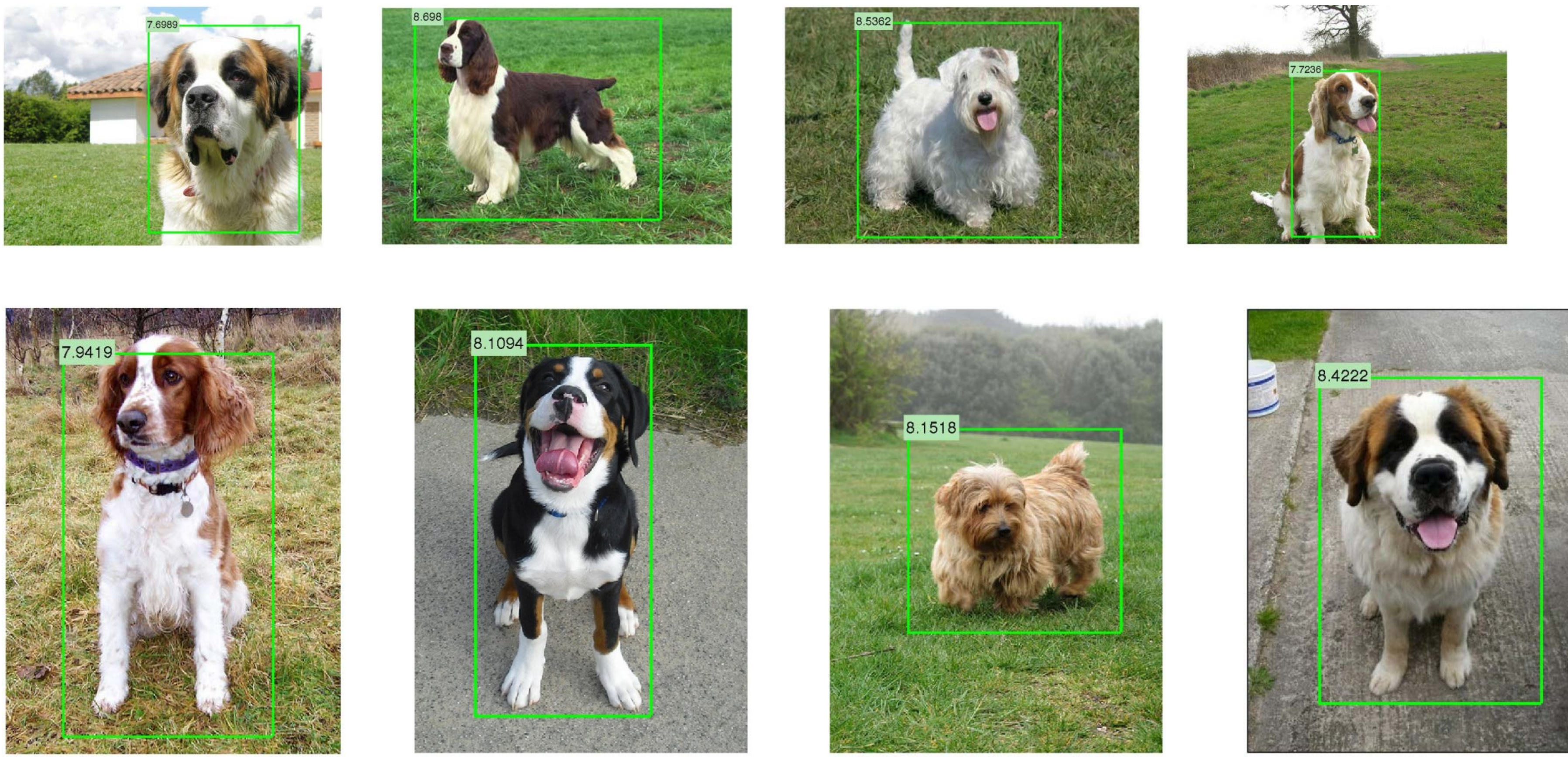
- Top 3 easiest categories: Basketball





Regionlets on ImageNet

- Top 3 easiest categories: Dog

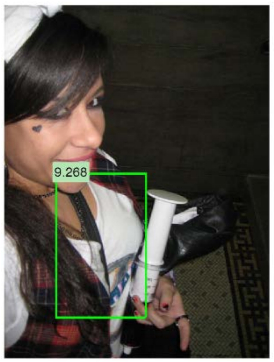


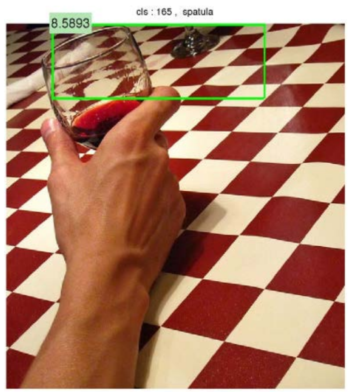
Regionlets on ImageNet

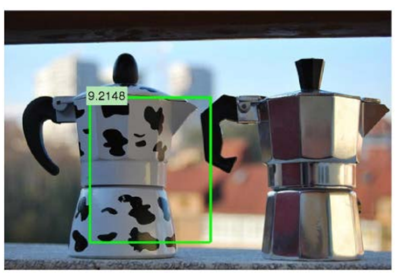
- Top 3 hardest categories: backpack

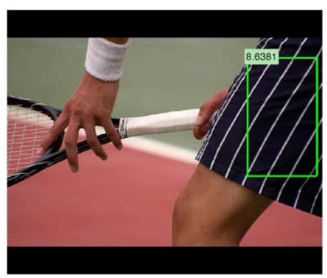
Regionlets on ImageNet

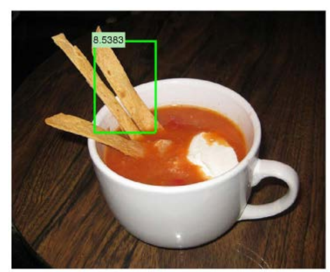
- Top 3 hardest categories: Spatula

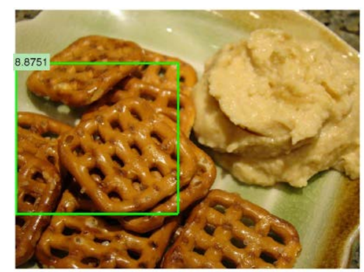






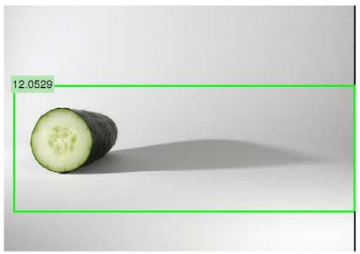


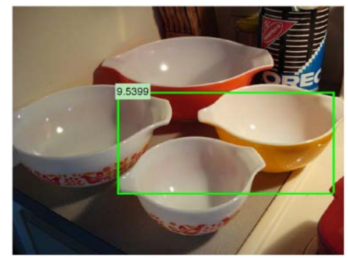


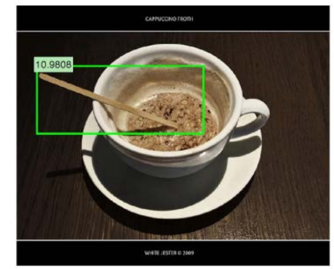


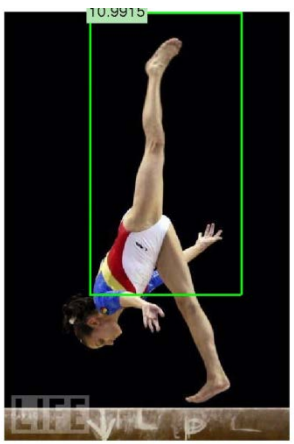
Regionlets on ImageNet

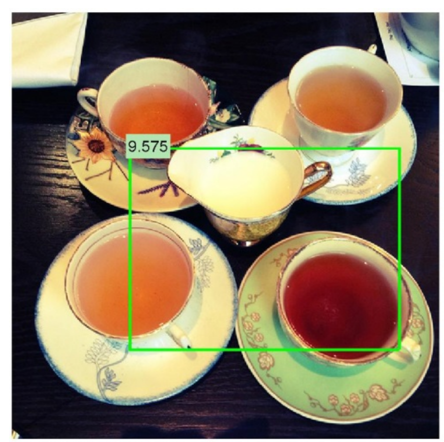
- Top 3 hardest categories: Ladle

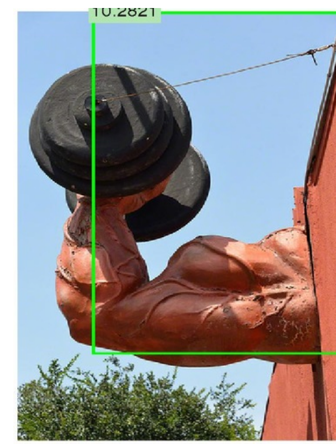












Conclusions

NEC Laboratories America
Relentless passion for innovation

- A new object representation for object detection
  - Non-local max-pooling of regionlets
  - Relative normalized locations of regionlets
  - Flexibility to incorporate various types of features
- A principled data-driven detection framework, effective in handling deformation, multiple scales, multiple viewpoints
- Superior performance with a fast running speed (.2 seconds per image)