NEC Laboratories America
Relentless passion for innovation

Regionlets for Generic Object Detection

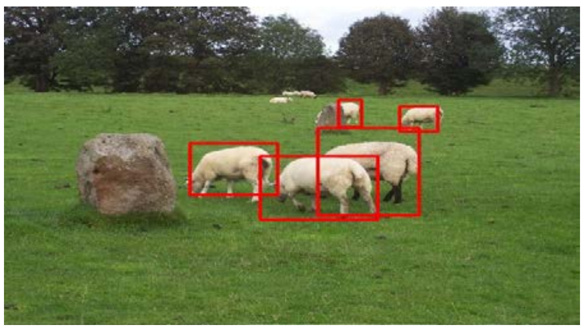
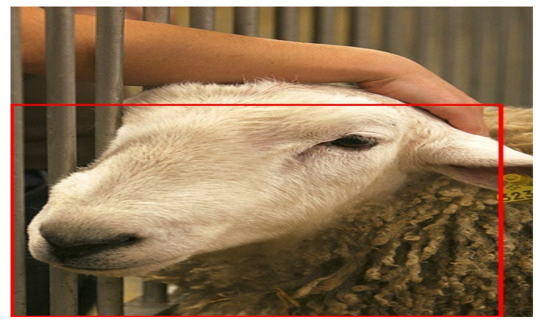
Xiaoyu Wang [†], Ming Yang [‡], Shenghuo Zhu [†] and Yuanqing Lin [†]

[†]NEC Labs America, Inc.
Cupertino, CA 95014, USA

[‡]Facebook, Inc.
Menlo Park, CA 94025, USA

Generic object detection


Train

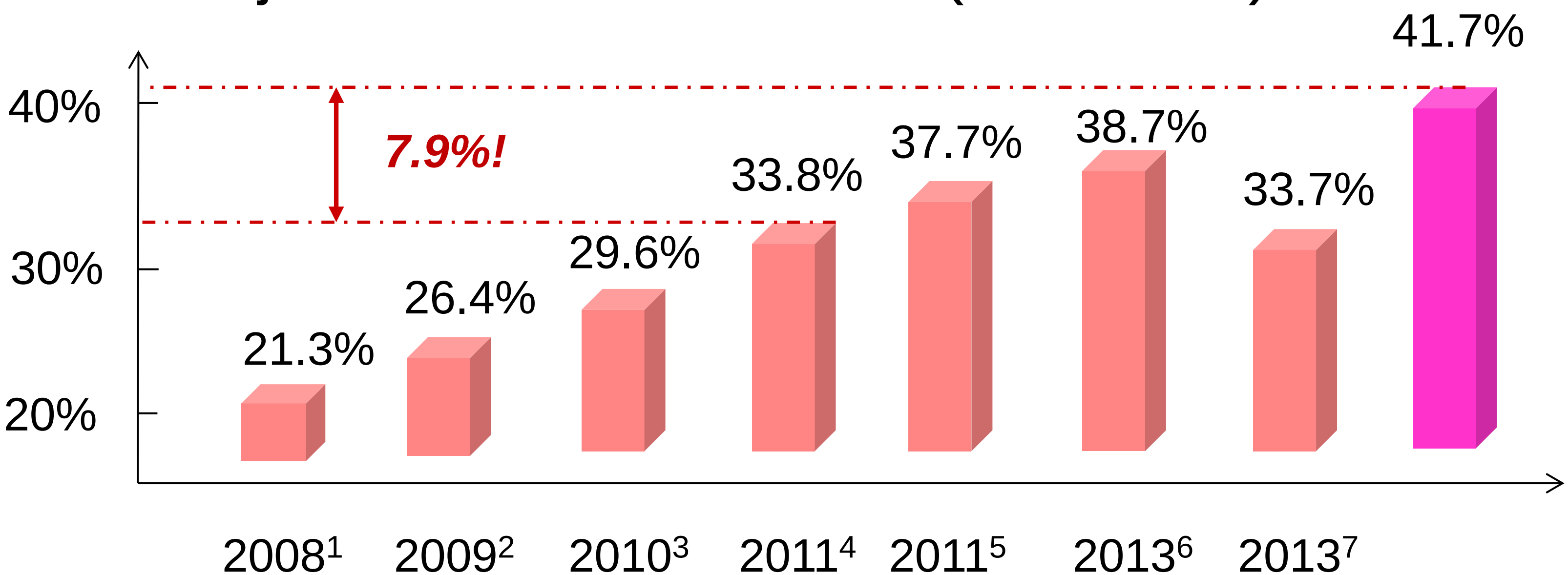
Sheep


Potted plant

State of the art

NEC Laboratories America
Relentless passion for innovation

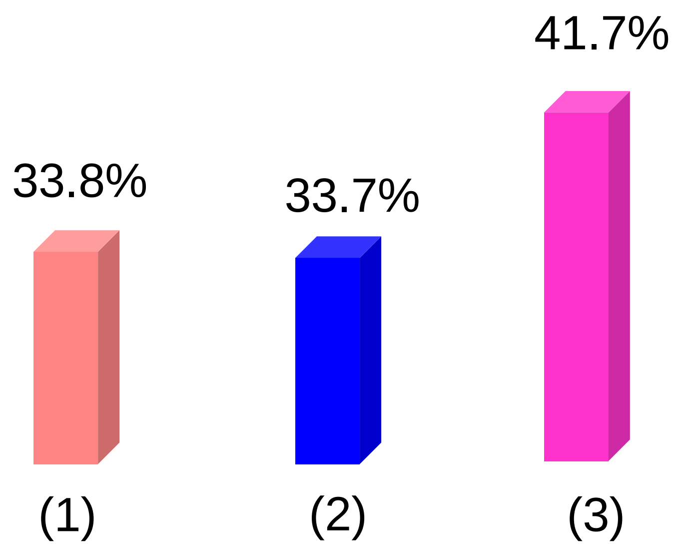
- Performance evolution on the PASCAL VOC 2007 object detection dataset(mean AP)



1. P. Felzenszwalb, et. al. A Discriminatively Trained, Multiscale, Deformable Part Model, CVPR 2008
2. A. Vedaldi, et. al. Multiple Kernels for Object Detection. ICCV 2009
3. L. Zhu, et. al. Latent hierarchical structural learning for object detection. CVPR 2010.
4. K. E. A. Van de Sande, et. al. Segmentation as selective search for object recognition. ICCV 2011
5. Z. Song, et. al. Contextualizing object detection and classification. CVPR, 2011
7. G. Chen, et. al. Detection Evolution with Multi-Order Contextual Co-occurrence, CVPR 2013
6. http://www.cs.berkeley.edu/~rbg/latent/ (DPM Release 5)

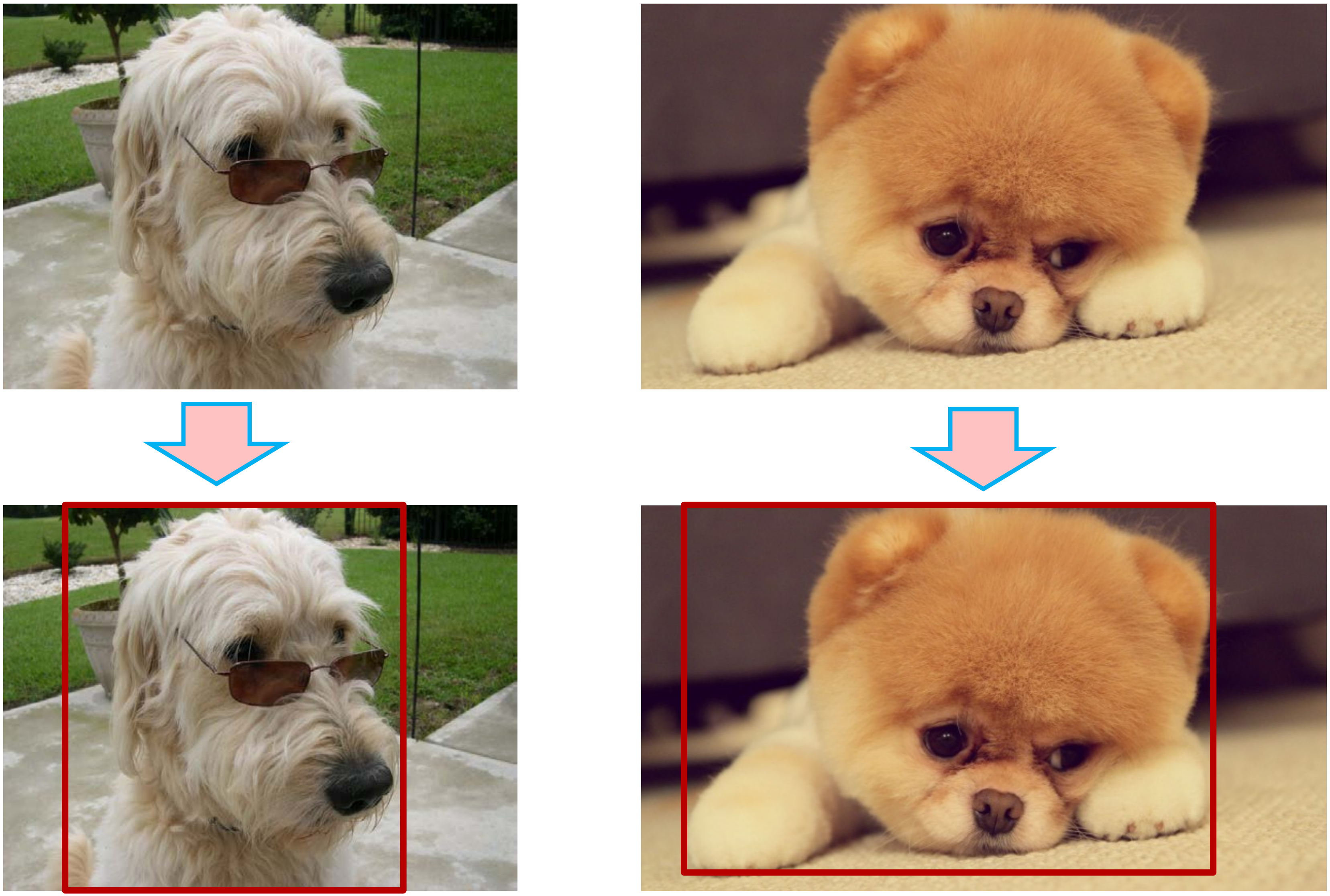
State of the art

- The two representative object detection frameworks

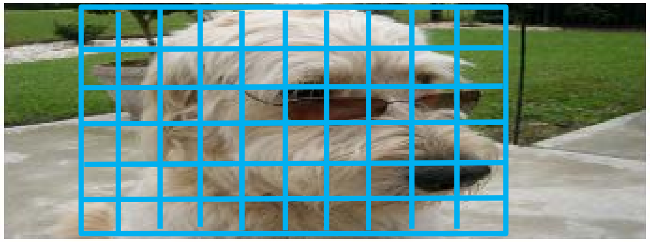


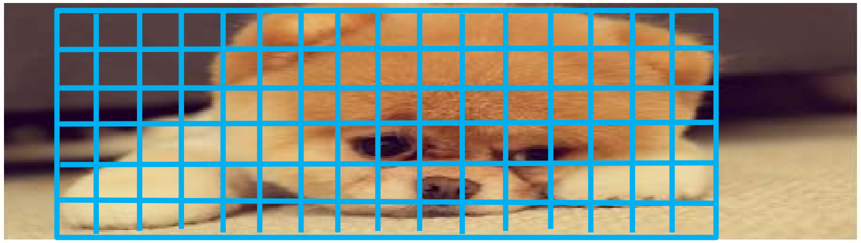
(1)	Scanning window with **D**eformable **P**art-based **M**odel (**DPM**)
(2)	**S**elective **S**earch with **S**patial **P**yramid **M**atching (**SS_SPM**)
(3)	Regionlets (No deep CNN feature yet 😊)

Object Detection



Review: Feature extraction

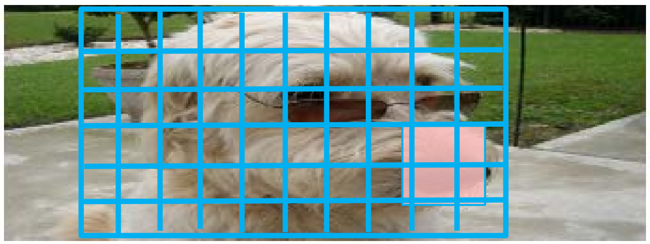


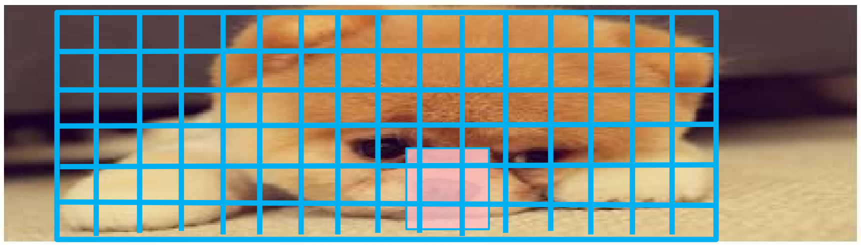


- Feature design
  - HOG
  - SIFT, and many others…
- Feature extraction
  - Densely extracted over N x N pixel cells

Review: Deformation handling

NEC Laboratories America
Relentless passion for innovation



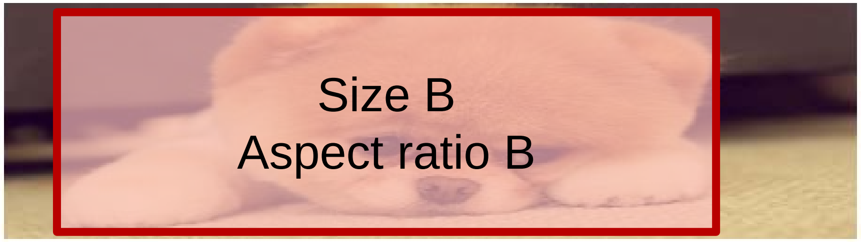


- Deformable Part-based Model (DPM)
  - Specify the number of deformable parts
- Spatial Pyramid Matching
  - Specify the number of pyramids to build

- *Do we have to pre-define model parameters to handle different degrees of deformation?*

Review: Multi scales/viewpoints

NEC Laboratories America
Relentless passion for innovation



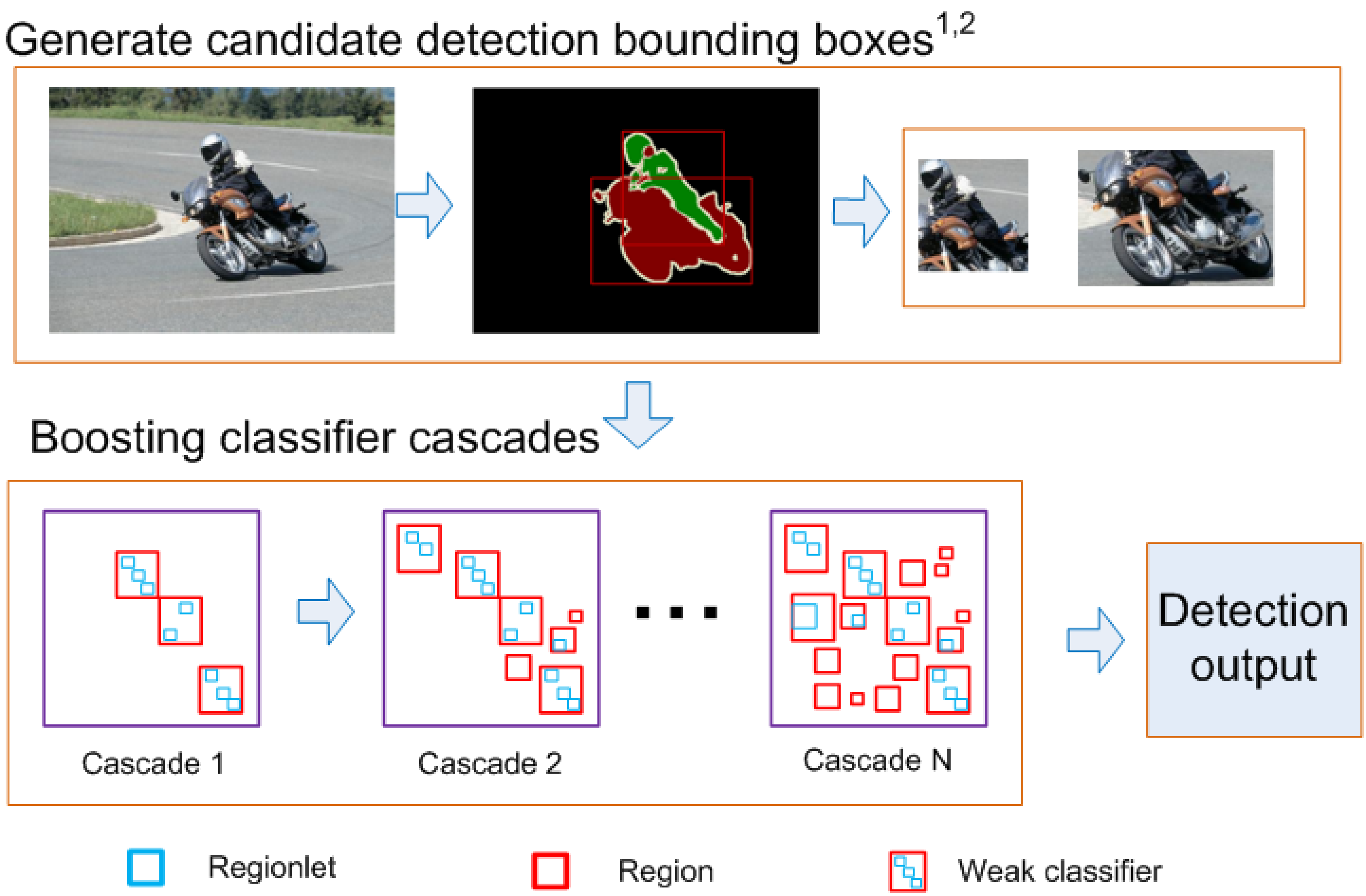
- DPM
  - Resize an image to detect objects at a fixed scale
  - Multiple models, each deals with one viewpoint
- Spatial Pyramid Matching
  - No need to resize the image
  - One model, a codebook is used to encode features
- *Can we learn a model that can be easily adapted to arbitrary scales and viewpoints?*

Motivation

- Motivation: A flexible and general object-level representation with
  - Hassle free deformation handling
  - Arbitrary scales and aspect ratio handling

Regionlets!

Detection framework



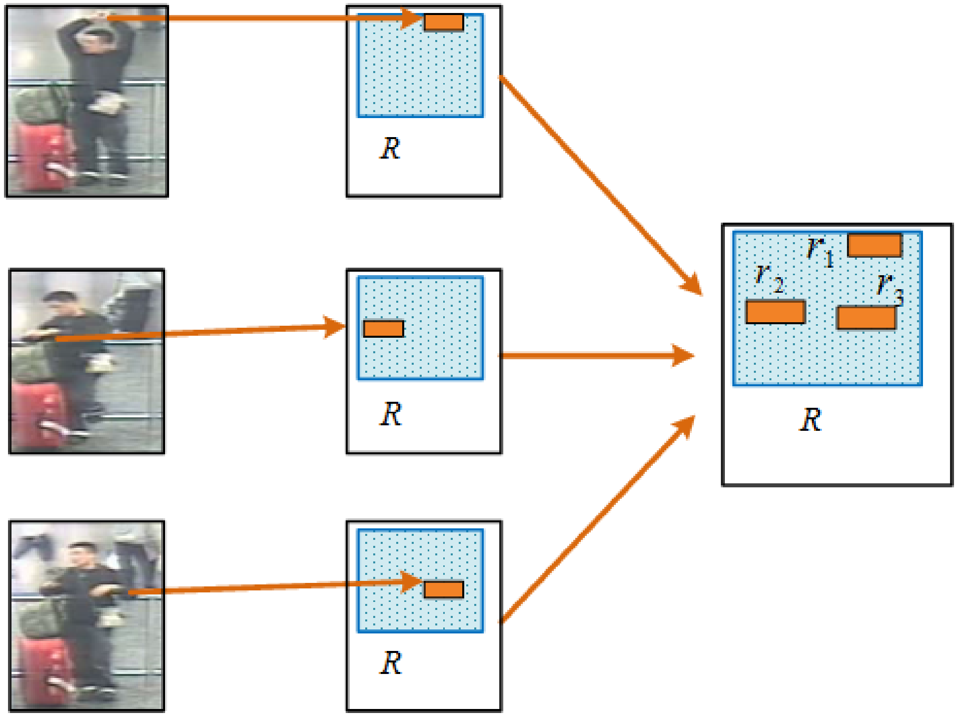
1. K. E. A. Van de Sande, et. al. Segmentation as selective search for object recognition. ICCV 2011
2. B. Alexe , et. al. Measuring the objectness of image windows. PAMI 2012

Regionlet: Definition

NEC Laboratories America
Relentless passion for innovation

- 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

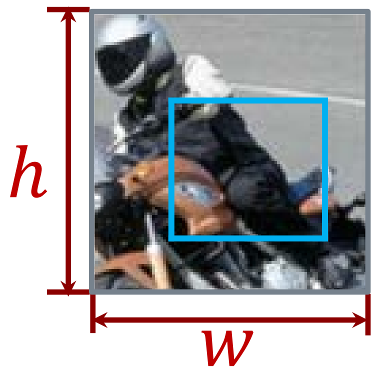
Figure 1



Regionlet: Definition(*cont.*)

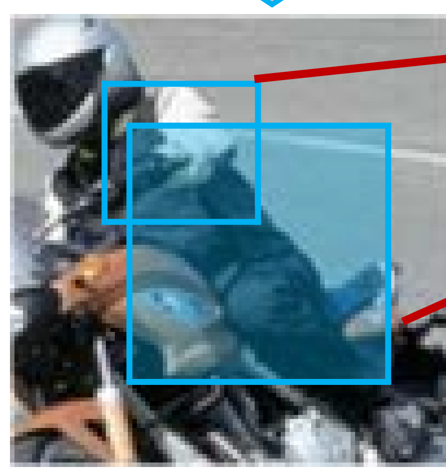
- Relative normalized position

Figure 2



Traditional → (l,t,r,b)
(50,50,180,180)

Normalized → $\left(\frac{l}{w},\frac{t}{h},\frac{r}{w},\frac{b}{h}\right)$
(.25, .25, .90,.90)



(50,50,180,180)

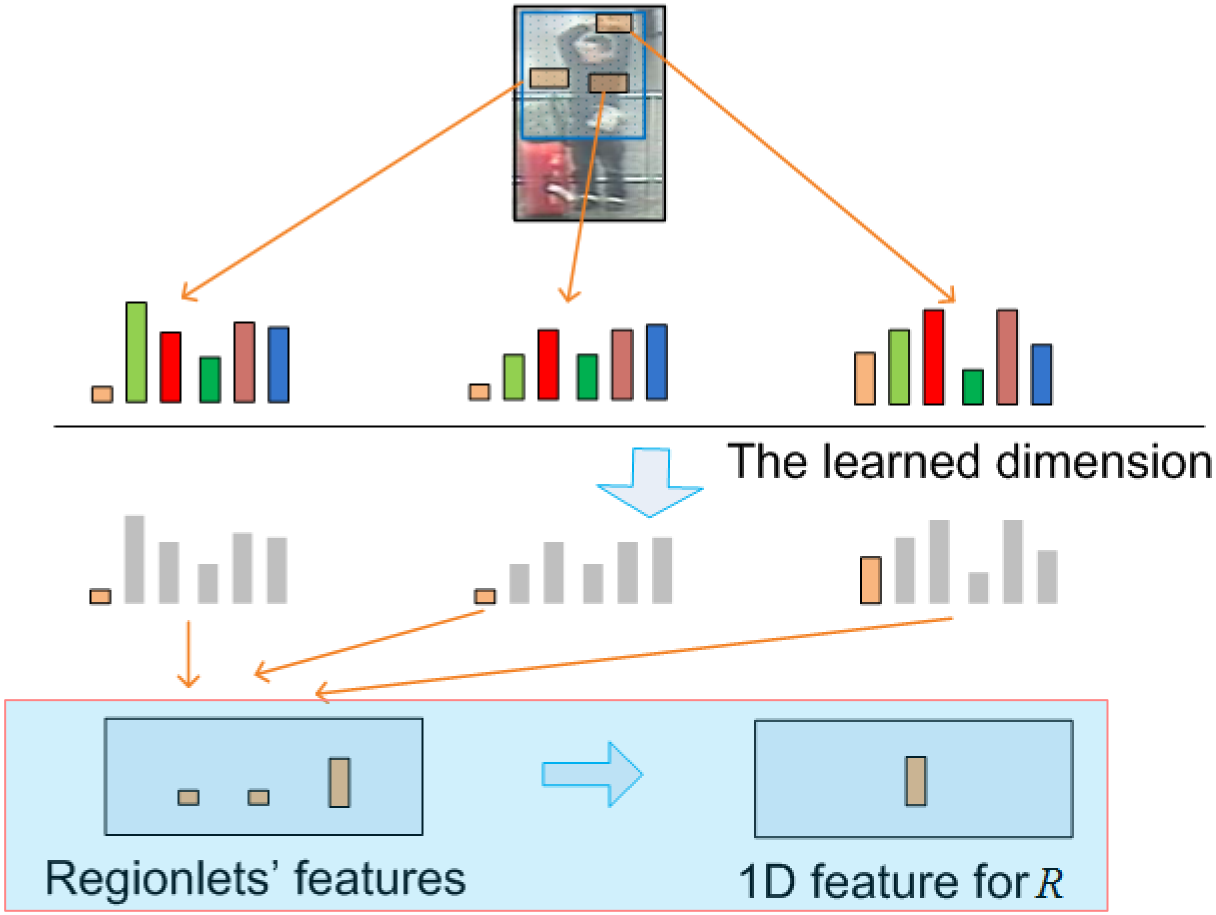
(.25, .25, .90,.90)

Regionlet: Feature extraction

NEC Laboratories America
Relentless passion for innovation

Figure 3

Non-local pooling



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

Regionlets: Training

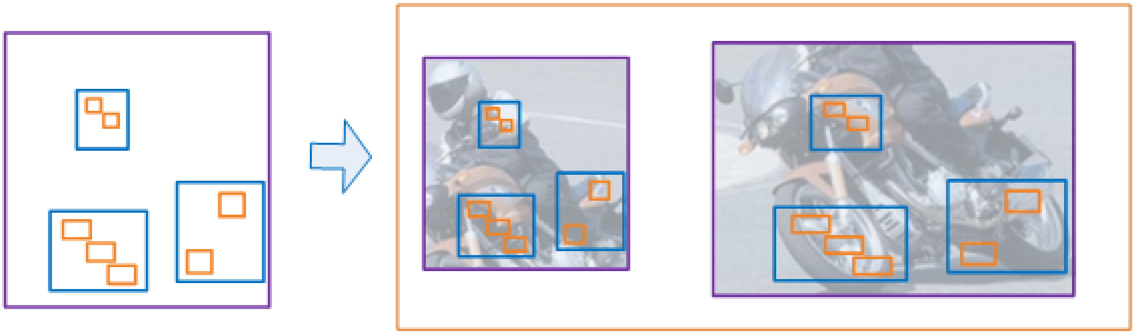
- Constructing the regions/regionlets pool
  - Small region, fewer regionlets -> fine spatial layout
  - Large region, more regionlets -> robust to deformation

- 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.

Regionlets: Testing

- No image resizing
- Any scale, any aspect ratio
- Adapt the model size to the same size as the object candidate bounding box



Regionlet based model

Applied to candidate boxes

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 (only for the ImageNet challenge)

Experiments: PASCAL VOC

NEC Laboratories America
Relentless passion for innovation

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

Experiments: ImageNet

NEC Laboratories America
Relentless passion for innovation

- 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(2)	**19.6% (no DCNN feature)**
OverFeat-NYU	19.4% (DCNN)
Toronto A	11.2% (N/A)
SYSU_Vision	10.5% (N/A)

(1) The result of using only a single method and single set of parameters, no context. No combining!

(2) The result of using traditional features only – no DCNN features were used.

Check our presentation at the ILSVRC2013 workshop for more details!

Running speed

NEC Laboratories America
Relentless passion for innovation

- 0.2 second per image using a single core if candidate bounding boxes are given, real time(>30 frames per second) using 8 cores 
- 2 seconds per image to generate candidate bounding boxes 
- 2-3 hours to finish training one category on a 8-core machine 

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