Accurate Object Detection with Location Relaxation and Regionlets Re-localization

Chengjiang Long, Xiaoyu Wang, Gang Hua, Ming Yang and Yuanqing Lin

NEC Laboratories America
Relentless passion for innovation

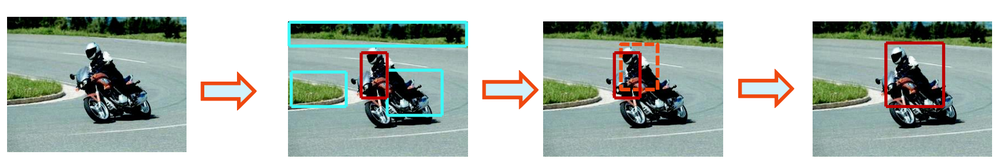
Problem & Background

1. To accurately localize the objects in the image → sliding window based detectors → computational cost is too high.
2. To reduce the computational cost → top-down or bottom-up approaches → none of these methods search for the object in the full continuous parameter space, *i.e.*, the center point, scale, and aspect ratio of the object.
3. Moreover, most of the existing approaches still rely on a classification model to localize the object → not necessarily optimize for accurate object localization.

Contributions

1. Proposed coarse detection plus supervised descent search in a fully parameterized location space for generic object detection which shows promising performance.
2. Proposed a novel Regionlets Re-localization method which complements the suboptimal object localization performance given by object detectors.
3. Our detection framework achieves the best performance on the PASCAL VOC 2007 dataset without using any outside data. It also demonstrates superior performance on our self-collected car dataset.

Pipeline



(a) A testing image (b) Coarse detections (c) Supervised descent search (d) Regionlets Relocalization

Bottom-up object proposal

We employ a segmentation based bottom-up scheme to generate our initial set of candidate searching locations.
Step 1. start with over-segments (*i.e.*, superpixels).
Step 2. hierarchically group these small regions to generate object hypotheses based on the characteristics like the size of the region, color histograms, and the texture.

Top-down Supervised Object Search

Ground truth object locations: $\{\mathbf{o}_*^i = (x_*^i, y_*^i, s_*^i, a_*^i)\}$
Starting locations: $\{\mathbf{o}_0^i = (x_0^i, y_0^i, s_0^i, a_0^i)\}$
$\Phi(.)$: the n-dimensional feature vector extracted from a location.
Train the project matrix $\mathbf{R}_0$ and the bias $\mathbf{b}_0$ by:

$$\arg\min_{\mathbf{R}_0,\mathbf{b}_0} \sum_i ||\Delta\mathbf{o}_{0*}^i - \Delta\mathbf{o}_0^i||^2$$
$$\Delta\mathbf{o}_{0*}^i = \mathbf{o}_*^i - \mathbf{o}_0^i$$
$$\Delta\mathbf{o}_0^i = \mathbf{R}_0^T\Phi(\mathbf{o}_0^i) + \mathbf{b}_0$$

The subsequent $\mathbf{R}_k$ and $\mathbf{b}_k$ ($k = 1, 2, \ldots$) can be learned iteratively. We update the new locations by the previous $\mathbf{R}_{k-1}$ and $\mathbf{b}_{k-1}$:

$$\mathbf{o}_k^i = \mathbf{o}_{k-1}^i + \mathbf{R}_{k-1}^T\Phi(\mathbf{o}_{k-1}^i) + \mathbf{b}_{k-1}$$

The optimal $\mathbf{R}_k$ and $\mathbf{b}_k$ can be learned by:

$$\arg\min_{\mathbf{R}_k,\mathbf{b}_k} \sum_i ||\Delta\mathbf{o}_{k*}^i - \Delta\mathbf{o}_k^i||^2$$

Given a testing image, we first apply the cascade regionlets detector to the coarse bottom-up object candidates → Object hypothesis which produce high detection scores are fed to the iterative supervised descent search process.

Regionlets Object Re-localization

Given the object location (l, t, r, b) detected by the object detector, and the ground truth object location (l^*, t^*, r^*, b^*)

$$l^* = l + w\Delta l_n, \qquad t^* = t + h\Delta t_n,$$
$$r^* = r + w\Delta r_n, \qquad b^* = b + h\Delta b_n.$$

Where $w = r - l + 1$, $h = b - t + 1$ are the detected bounding box width and height respectively.
$(\Delta l_n, \Delta t_n, \Delta r_n, \Delta b_n)$ are the relative location error between the ground truth and the current detection. We solve a support vector regression problem for each of the four coordinates respectively:

$$\min_V \left\{ \frac{\|V\|}{2} + C\sum_{m=1}^{M} \max(0, |\Delta L_m - V^T R_m| - \epsilon)^2 \right\}$$

Where $\Delta L = V^T R$, ΔL is either Δl_n, Δt_n, Δr_n, or Δb_n.
R_m is the feature extracted from regionlets.

Experimental Results

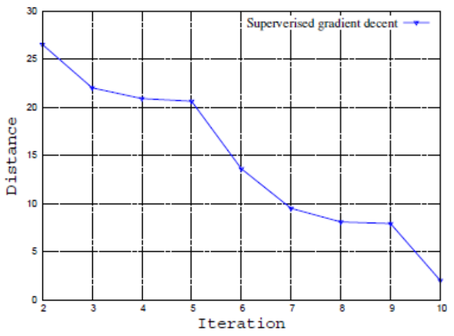
A. Effectiveness of detection with localization relaxation.
Performance comparison with the baselines on the PASCAL VOC 2007 dataset.

AP %	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
Regionlets [16]	53.1	49.5	16.7	25.9	16.3	49.8	64.2	37.9	16.7	39.3	44.7
LRS w/o aspect ratio	53.3	49.1	17.0	25.9	17.9	50.6	64.5	41.5	17.2	40.1	**46.8**
LRS w/ aspect ratio	**54.2**	**52.4**	**18.0**	**27.3**	**22.5**	**53.8**	**68.6**	**43.1**	**20.6**	**42.8**	45.6

	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Regionlets [16]	23.2	50.4	52.7	35.6	11.7	29.5	31.3	56.1	50.0	37.7
LRS w/o aspect ratio	25.0	51.6	53.3	36.6	13.0	29.6	34.4	55.6	50.5	38.7
LRS w/ aspect ratio	**26.2**	**56.2**	**57.2**	**42.7**	**16.0**	**37.0**	**38.7**	**57.1**	**51.7**	**41.6**

B. Understand the supervised decent search.

The trace of the searched bounding box center in supervised decent.



The distance between the searched bounding box center and the true object center in supervised descent.

C. Effectiveness of Regionlets Re-localization.
Performance on PASCAL VOC 2007 dataset.

AP %	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
LRS	54.2	52.4	18.0	27.3	22.5	53.8	68.6	43.1	20.6	42.8	45.6
LRS-RR	**55.8**	**53.5**	**22.1**	**28.8**	**25.1**	**54.1**	**71.5**	**45.9**	**22.3**	**45.7**	**50.6**

	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
LRS	26.2	56.2	**57.2**	42.7	16.0	37.0	38.7	57.1	51.7	41.6
LRS-RR	**29.6**	**58.4**	55.6	**49.0**	**17.6**	**41.1**	**42.4**	**59.5**	**54.2**	**44.1**

Comparison with state of the arts.

	VOC 2007	Results year
DPM(WC) [4]	35.4	2008
UCI_2009 [20]	27.1	2009
INRIA_2009 [21]	28.9	2009
MIT_2010 [2]	29.6	2010
Song *etal*(WC) [22]	37.7	2011
Li *etal*(WC) [23]	35.2	2011
SS_SPM [10]	33.8	2011
Cinbis *etal*(WC) [24]	35.0	2012
Regionlets [16]	41.7	2013
Ours(LRS + RR)	**44.1**	2014

Acknowledgement

NEC NSF National Institutes of Health Google Research Microsoft Research Stevens Institute of Technology