R-CNN for Small Object Detection

Chenyi Chen[1], Ming-Yu Liu[2], Oncel Tuzel[2], and Jianxiong Xiao[1]

[1] Princeton University, Princeton NJ, USA
[2] Mitsubishi Electric Research Labs (MERL), Cambridge MA, USA

Abstract. We investigate into a ConvNet-based co-occurrence model, which leverages the detection results of big objects to better localize the small objects. The spatial relation between big and small objects are formulated as learnable parameters integrated into an end-to-end training framework. However, the results show there is no improvement of performance over the automatic context learning approach.

1 Introduction

In the paper, we propose a straightforward automatic context learning approach that takes a 3x or 7x context region enclosing the proposal region as an additional input to the neural network. We argue that this method is simple yet effective. To demonstrate this, we also investigate into a ConvNet-based co-occurrence model, which leverages the detection results of big objects to better localize the small objects. The spatial relation between the big and small objects are learnt as convolutional filters through end-to-end training. We provide details of our implementation here.

2 Co-occurrence ConvNet

As we know, large objects are much easier to detect than small objects, and some small objects have strong co-occurrence spatial relation with big objects, thus we may use the more reliably detected big objects to improve the detection accuracy of small objects.

Fig. 1 shows an example. The initial detection of "mouse" contains a lot of false positives (shown in the lower left image with green boxes, the ground truth "mouse" locations are represented as red boxes). By knowing the locations of "monitor" and "keyboard" (indicated as blue and yellow boxes, respectively), and the spatial relation between "mouse" and "monitor", "mouse" and "keyboard" (upper left two heat maps), we can compute a corresponding probability map of finding a "mouse" in the input image (upper right image), and use the probability map to further remove false positives (lower right image). The spatial relation map represents the probability of having a small object at each location around the big object, where the location of the big object is fixed at the center of the map. For example, the mouse-keyboard spatial relation map indicates a mouse usually shows up on the right hand side of a keyboard, and

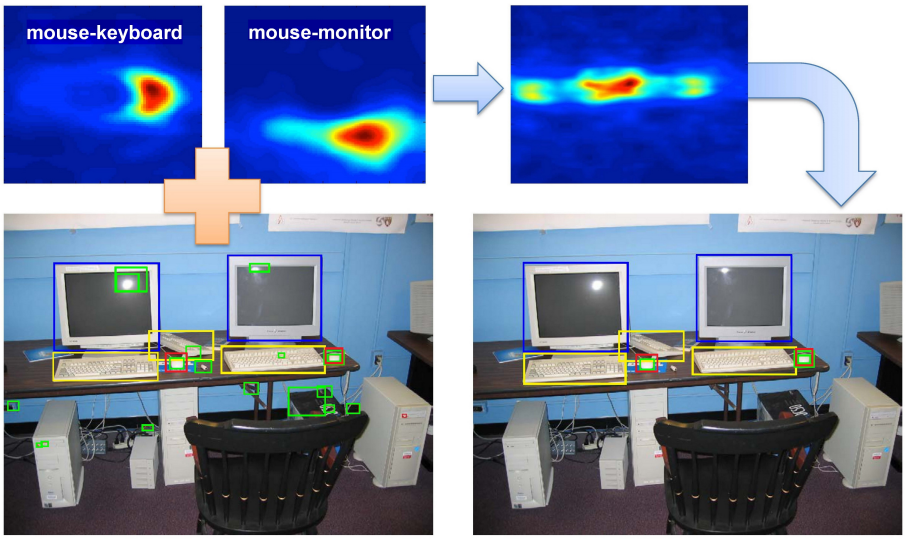


Fig. 1: **Example of how co-occurrence improves the accuracy of small object detection.** By knowing the location of monitor and keyboard, many high score false positives can be eliminated since those locations are not likely to have a mouse.

the mouse-monitor spatial relation map indicates a mouse usually presents at lower right to a monitor.

Such co-occurrence spatial relation maps can be learnt as convolutional filters in a ConvNet. We implement the ConvNet that re-scores the output of a object detection network by rejecting high score false positives and boosting low score true positives. The schematic of the network is shown in Fig. 2. In the figure, we use "mouse", "keyboard", and "monitor" as an example.

To test our co-occurrence approach, we choose three small objects: "mouse", "toilet paper", "faucet", and five big objects: "monitor", "keyboard", "toilet", "sink" and "night table". Positive spatial relation between small objects and big objects is learnt automatically through training, we do not manually pair any small objects and big objects. In our experiments, we use the ground truth bounding boxes of big objects in both the training and testing phase. The co-occurrence ConvNet has the following inputs:

1. **Small object region proposal score map**: the location of each non-zero point is the center of each small object region proposal (regardless of the box size). The value of the point is the classification score of that region proposal outputted by the object detection network.
2. **Big object location map**: the location of each non-zero point is the center of each big object bounding box (regardless of the box size), with the value being the detection score. In our current setting, we use ground truth bounding boxes, so the values of the non-zero points are always one.

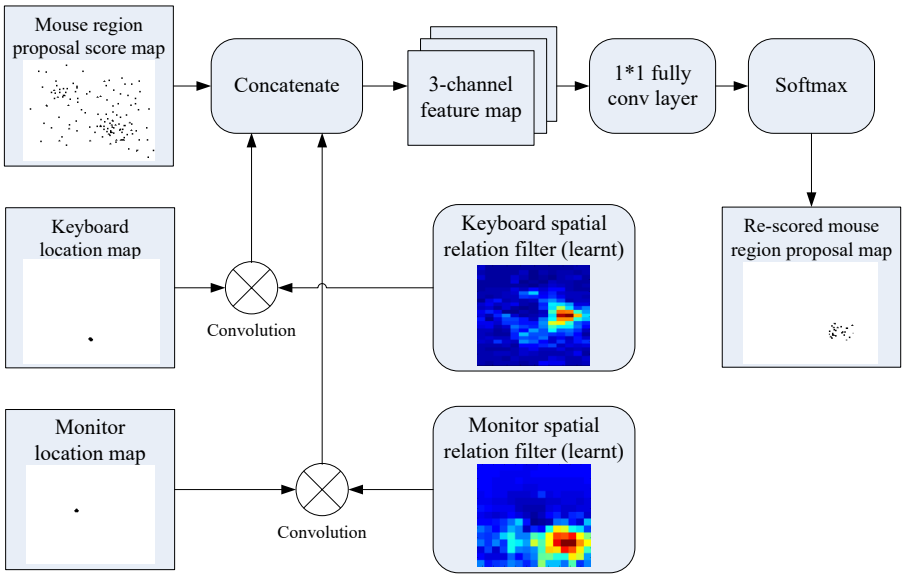


Fig. 2: **Co-occurrence ConvNet structure.** The classification scores of small object region proposals and the ground truth bounding boxes of the big objects are converted to input images of the co-occurrence ConvNet. Spatial relation filters convolve with the corresponding big object location maps, and the resulting feature maps concatenate with the small object score map. All these feature maps are used to produce a new score map for the small object region proposals.

To train a ConvNet for co-occurrence spatial relation, the filter size needs to be large enough to accommodate the space between objects. In our implementation, we choose $81 * 81$. Generally, with only a few thousand training samples, it is impossible to train a ConvNet with such large filters. However, as the spatial relation is based on statistics of data, the filter does not need to have high resolution. Thus, we implement a grid filter, which partition the filter area into a $n * n$ grid (e.g. $15 * 15$ grid for the $81 * 81$ filter). Within each block in the grid, the parameters are shared (updated together). So a $81 * 81$ filter actually only has $15 * 15 = 225$ trainable parameters, which can be successfully trained.

Objects in images have various scales, and in order to convolve all the images with only one filter, we need to explicitly handle scales. We select four different scales to cover the range of the object sizes, as illustrated in Fig. 3. The following pipeline is used to handle scales:

1. Re-scale the small object region proposals and the big object bounding boxes as if the original image is resized with the longest side to 240 (preserve the aspect ratio).
2. Partition the re-scaled region proposals and bounding boxes into four different scales, as listed in Table 1.

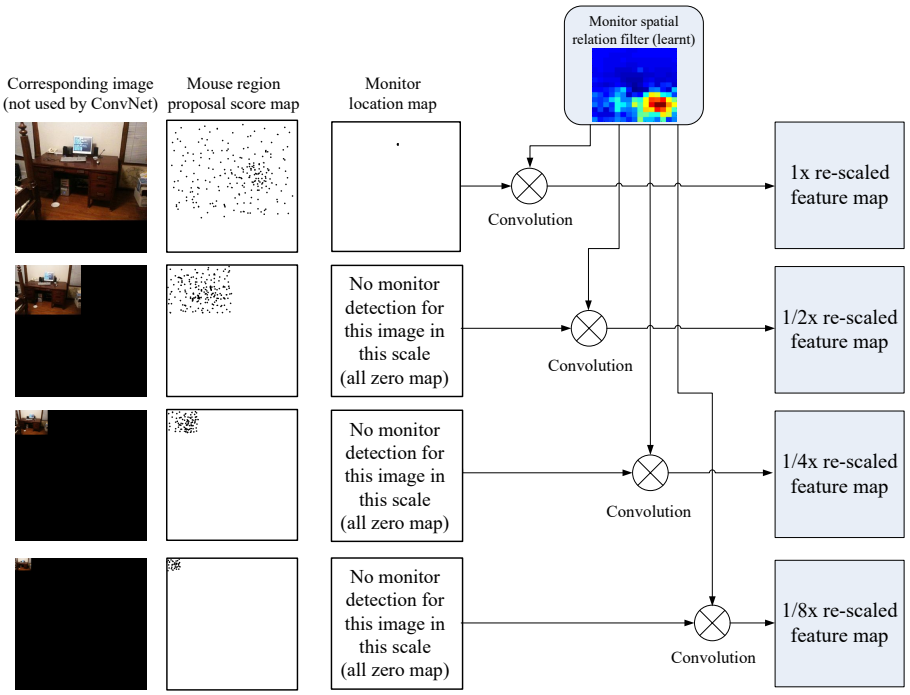


Fig. 3: **Handling various scales.** The small object region proposals and the big object bounding boxes are partitioned into four scales according to the size. Large region proposals and bounding boxes are down-sampled to match the size of the spatial relation filter.

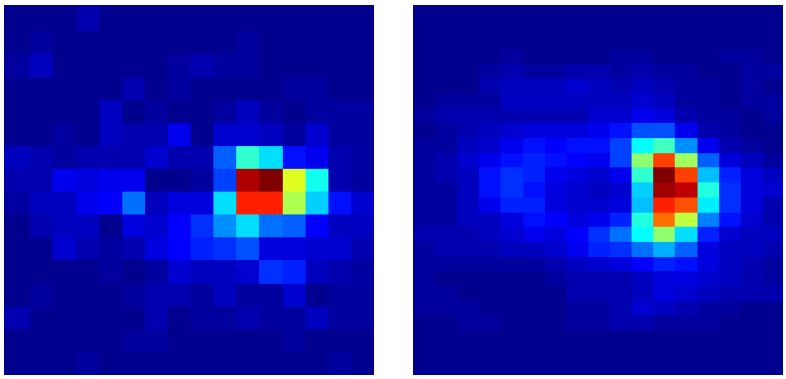
3. For large scales (e.g. scale 2 to scale 4), further down-sample the region proposals and bounding boxes to 1/2x, 1/4x, and 1/8x respectively to match the size of the spatial relation filter.

The ground truth locations of small objects are used as supervision to train the ConvNet. For each small object, we train a ConvNet to encode its co-occurrence spatial relation with all the five big objects. When we visualize the filters after sufficient training, we observe that the ConvNet does learn meaningful co-occurrence spatial relation between corresponding small objects and big objects, as shown in Fig. 4 to Fig. 7, where the small object region proposal score map is produced by Full AlexNet (R-CNN). The filters of non-related objects (e.g. "mouse" and "toilet", "faucet" and "keyboard") are close to all-zero maps.

We train the co-occurrence ConvNet on top of the results of various object detection networks, and find the co-occurrence spatial relation is not very helpful in improving the average precision. In Table 2, we list the average precision computed using $IoU \geq 0.5$ and $IoU \geq 0.3$ criteria. From the results we observe that the co-occurrence approach has no advantage over the simple end-to-end trainable ContextNet approach.

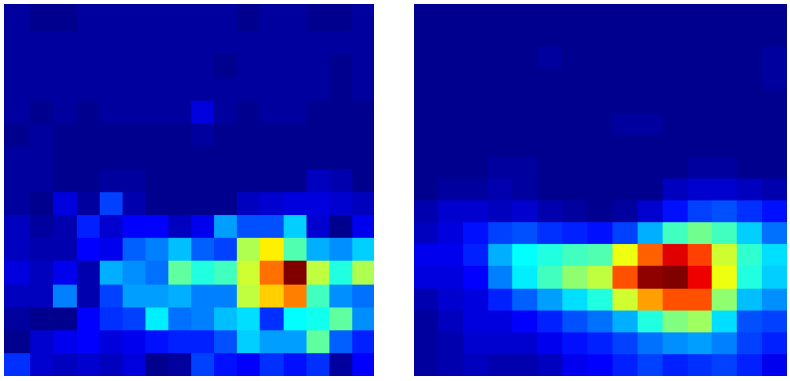
Table 1: **Four scales of the object size.** The size is counted as the longest edge of the re-scaled box, in pixel.

Category	scale 1	scale 2	scale 3	scale 4
mouse	≤ 10	$11 \sim 20$	$21 \sim 40$	≥ 41
toilet paper	≤ 10	$11 \sim 20$	$21 \sim 40$	≥ 41
faucet	≤ 14	$15 \sim 28$	$29 \sim 56$	≥ 57
monitor	≤ 43	$44 \sim 86$	$87 \sim 172$	≥ 173
keyboard	≤ 34	$35 \sim 68$	$69 \sim 136$	≥ 137
toilet	≤ 45	$46 \sim 90$	$91 \sim 180$	≥ 181
sink	≤ 43	$44 \sim 86$	$87 \sim 172$	≥ 173
night table	≤ 40	$41 \sim 80$	$81 \sim 160$	≥ 161



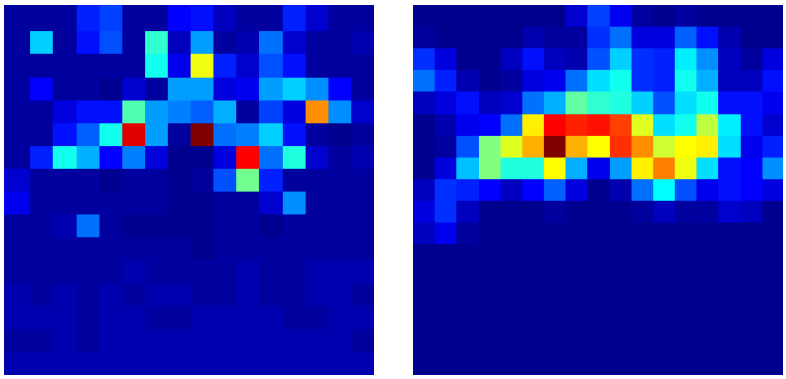
(a) Learnt by the ConvNet (b) Manually computed

Fig. 4: Spatial relation filter of "mouse" and "keyboard".



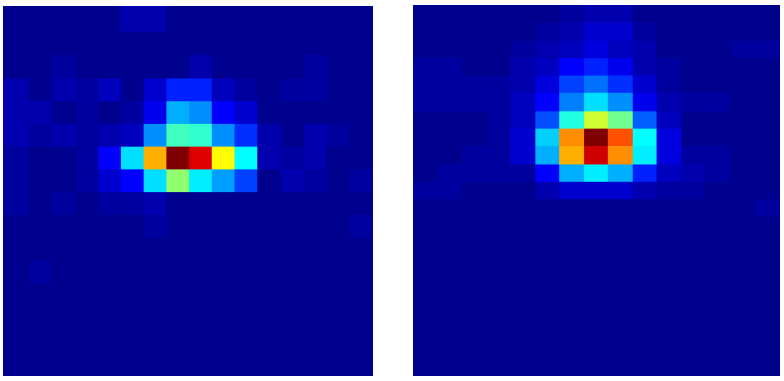
(a) Learnt by the ConvNet (b) Manually computed

Fig. 5: Spatial relation filter of "mouse" and "monitor".



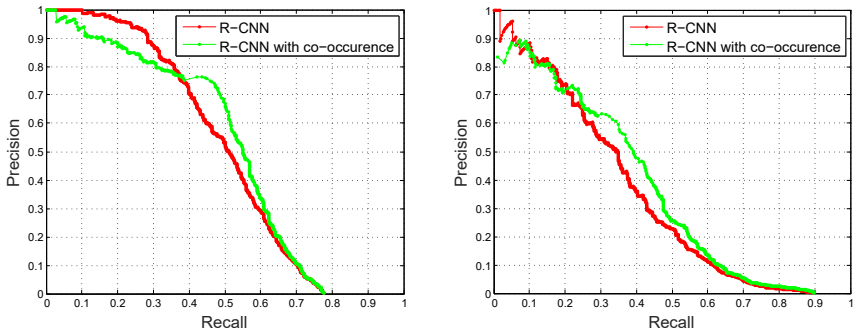
(a) Learnt by the ConvNet (b) Manually computed

Fig. 6: Spatial relation filter of "toilet paper" and "toilet".



(a) Learnt by the ConvNet (b) Manually computed

Fig. 7: Spatial relation filter of "faucet" and "sink".



(a) mouse (b) faucet

Fig. 8: **Precision-recall curves of the co-occurrence ConvNet.** The ConvNet is built on top of the Full AlexNet (R-CNN)'s results, and the curves are computed using $IoU \geq 0.3$.

Table 2: **Original results of various object detection networks and re-scored results of the co-occurrence ConvNet.** The average precision is computed using $IoU \geq 0.5$ (upper table) and $IoU \geq 0.3$ (lower table), respectively.

$IoU \geq 0.5$	mouse	toilet paper	faucet
Partial AlexNet	29.8	15.5	6.7
Partial AlexNet + co-occurrence	30.9	12.7	12.1
Full AlexNet (R-CNN)	42.9	26.7	18.6
Full AlexNet + co-occurrence	33.6	22.1	18.8
ContextNet (AlexNet, 7x)	**48.4**	**30.4**	**20.5**
ContextNet + co-occurrence	30.0	21.1	17.2

$IoU \geq 0.3$	mouse	toilet paper	faucet
Partial AlexNet	36.7	28.7	18.4
Partial AlexNet + co-occurrence	43.9	32.4	29.2
Full AlexNet (R-CNN)	49.5	34.4	34.0
Full AlexNet + co-occurrence	49.8	39.9	36.4
ContextNet (AlexNet, 7x)	**54.9**	**49.2**	**38.3**
ContextNet + co-occurrence	47.5	42.8	41.0

We investigate the results and find the following two reasons that could possibly explain why the co-occurrence model degrades the performance:

1. The co-occurrence ConvNet re-scores lots of proposals covering the same ground truth to high scores, such overlapping proposals may not be eliminated by non-maximum suppression, thus reduces the precision.
2. The co-occurrence ConvNet re-scores false positives that match the spatial relation (e.g. headset next to the keyboard) to high scores.

The precision-recall curves of "mouse" and "faucet" are shown in Fig. 8. Based on these curves, we observe that the co-occurrence approach almost always deteriorate the performance in low recall - high precision range. In that range, the object detection network usually has very reliable performance, so the co-occurrence ConvNet may mostly reject true positives and boost false positives. In contrast, using the context region as input to the object detection network is a more convenient and effective way. As we use a large context patch, the network automatically learns all kinds of context information (possibly also including co-occurrence) through training. While in the co-occurrence ConvNet setting, we need to manually select a list of big objects and constraint the context information as co-occurrence spatial relation only. The finding is reasonable, since research has already demonstrated that deep features automatically learnt by ConvNet is superior than manually engineered ones.