MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

R-CNN for Small Object Detection

Chen, C.; Liu, M.-Y.; Tuzel, C.O.; Xiao, J.

TR2016-144 November 2016

Abstract

Existing object detection literature focuses on detecting a big object covering a large part of an image. The problem of detecting a small object covering a small part of an image is largely ignored. As a result, the state-of-the-art object detection algorithm renders unsatisfactory performance as applied to detect small objects in images. In this paper, we dedicate an effort to bridge the gap. We first compose a benchmark dataset tailored for the small object detection problem to better evaluate the small object detection performance. We then augment the state-of-the-art R-CNN algorithm with a context model and a small region proposal generator to improve the small object detection performance. We conduct extensive experimental validations for studying various design choices. Experiment results show that the augmented R-CNN algorithm improves the mean average precision by 29.8% over the original R-CNN algorithm on detecting small objects.

ACCV

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2016
201 Broadway, Cambridge, Massachusetts 02139

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

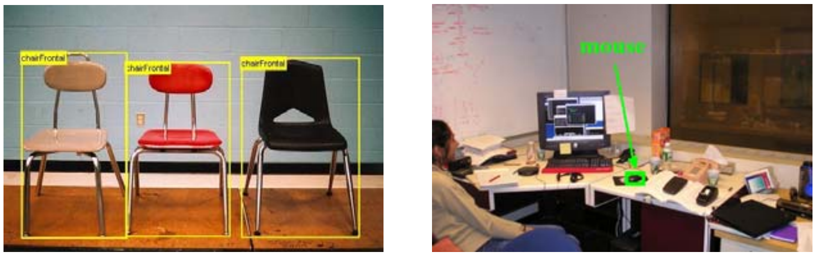
Abstract. Existing object detection literature focuses on detecting a big object covering a large part of an image. The problem of detecting a small object covering a small part of an image is largely ignored. As a result, the state-of-the-art object detection algorithm renders unsatisfactory performance as applied to detect small objects in images. In this paper, we dedicate an effort to bridge the gap. We first compose a benchmark dataset tailored for the small object detection problem to better evaluate the small object detection performance. We then augment the state-of-the-art R-CNN algorithm with a context model and a small region proposal generator to improve the small object detection performance. We conduct extensive experimental validations for studying various design choices. Experiment results show that the augmented R-CNN algorithm improves the mean average precision by 29.8% over the original R-CNN algorithm on detecting small objects.

1 Introduction

We have witnessed several breakthroughs in the field of visual object detection in the past decade, demonstrated by the ever-increasing performance improvement on the PASCAL VOC [1]. However, the object detection problem still remains largely unsolved as none of the state-of-the-art object detectors is close to perfect. Moreover, the performance on the PASCAL VOC can be misleading due to the dataset bias as pointed out by Torralba and Efros [2]. It is expected that when the application domain has a very different bias to the one in the PASCAL VOC, the performance of the state-of-the-art detectors for the PASCAL VOC would degrade significantly.

In this paper, we study the small object detection problem. By small objects, we refer to objects with smaller physical sizes in the real world. We also limit our interest to the small objects that each occupies a small part of an image. This means that comparing to the PASCAL VOC where the majority of objects are big in the real world and each occupies a large portion of an image, we are considering an application domain with a selection bias toward small objects as shown in Fig. 1.

It is true that one can always have a higher resolution image or take a closer snapshot of a small object in order to detect it. But the low-resolution inputs for small objects is deeply embedded in the nature of visual perception, and a robust vision system should be able to deal with it. For example, the physical



(a) Typical objects in PASCAL (b) Small objects in the real world

Fig. 1: Detecting small objects with low-resolution inputs.

size of a typical desk and monitor is many times bigger than a mouse. As a human, when we see a desk with a monitor and a mouse, we recognize all of them in one shot. We do not look particularly closer to the mouse to put a large image at the center of our retina. It is desirable that a computer vision system possesses a similar capability.

Moreover, detecting small objects is itself an intriguing problem due to several unique challenges. First, there are much more possibilities for the locations of small objects. The precision requirement for accurate localization is several magnitudes higher than that for typical PASCAL VOC objects. Second, there are much fewer pixels available for small objects, which means much weaker signal for the detector to utilize. Third, there are only limited prior knowledge and experiences in this area since most of the prior works are tuned for the big object detection problem. Practically, there is no benchmark dedicated to such a task[3]. In fact, we do not have much understanding on how difficult the small object detection task is or how well existing object detectors work. In order to better assess the performance of an algorithm for the small object detection problem, we establish a small object detection benchmark.

The R-CNN algorithm [3, 4], which extracts discriminative features using deep convolutional neural network from region proposals, has been established as the state-of-the-art approach for object detection as supported by the achieved impressive performance on the PASCAL VOC benchmark. In this paper, we extend the R-CNN algorithm to deal with the small object detection problem. Specifically, we propose a region proposal network tailored for capturing the "objectness" for small objects in order to obtain a small set of proposals while still keeping a high recall rate. We also propose a way to encode the context information from the surrounding areas of an object proposal. We show that the extended R-CNN algorithm achieves a mean Average Precision (mAP) of 23.5% on the benchmark dataset, which significantly outperforms a mAP of 18.1% achieved by the original R-CNN algorithm. We also present extensive experimental evaluations on various design choices for understanding their impacts to the small object detection performance.

---

[3] Although standard datasets such as the Microsoft COCO contains several "small" object categories, many of the instances of the objects in the "small" object categories occupy a large part of an image.

1.1 Related Work

Earlier work on small object detection is mostly about detecting vehicles using hand-crafted features and shallow classifiers in aerial images [5, 6]. In this paper, we cover a diverse set of small objects in the daily life and augment the state-of-the-art R-CNN algorithm for detecting them. [7] analyzes the influences of object characteristics on the performance of multiple detectors, with "object size" among the characteristics being studied. The results reveal that the detection accuracy drops as the object size becomes smaller, which provides some initial insight into the small object detection problem.

The PASCAL VOC [1] is the most widely used benchmark dataset for general object detection. It contains 20 object categories including "cow", "vehicle", and "dog". The object instances in the PASCAL VOC are usually large. Many of them occupy a major portion of the image. Our focus is on small objects where the object instance should only occupy a small portion of the image. In this sense, directly using the PASCAL VOC dataset is inappropriate. Microsoft COCO dataset [8] is proposed to advance the object detection techniques by placing it in the context of scene understanding, and the dataset contains many categories of small objects. To better represent the problem, we compose our small object detection dataset by using a subset of images from both the COCO dataset and the Scene UNderstanding database (SUN) [9], which also contains a large amount of small objects in various scenes.

[3, 4] propose the R-CNN algorithm, which combines convolutional neural networks with bottom-up region proposals [10] for object detection. R-CNN significantly outperforms conventional approaches on the PASCAL VOC dataset and establishes the new state-of-the-art in object detection research. Recently, some work improves the region proposal generation part of R-CNN and obtain faster computation speed and more accurate detection performance. [11] generates region proposals using edge cues. [12] computes "objectness" of region proposals based on a convolutional neural network. The MultiBox method [13] directly predicts a set of class-agnostic bounding boxes along with a single objectness score for each box, the method is not translation-invariant. [14] propose a translation-invariant Region Proposal Network (RPN) that shares convolutional layers with the detection network and achieve faster computation speed and better performance. The above algorithms are designed for detecting large objects in the PASCAL VOC. We focus on the small object detection problem and systematically study the applicability of the R-CNN style algorithms for detecting small object in the image.

Generally, context is useful for improving the object detection performance in natural scenes [15, 16]. Based on R-CNN, [17] proposes a pipeline for action recognition using more than one regions. [18] proposes a multi-region object detection system that can steering the ConvNet to focus on different regions of the object. [19] use both segmentation and context to improve object detection accuracy. [20] studies the role of context in existing object detection approaches and further proposed a model that exploits both the local and global context. In this work, we also leverage the context information to get better performance.

Many researches have been shown to improve the localization accuracy of object detectors. [21] introduces a Bayesian optimization-based algorithm that iteratively searches for better bounding boxes for object detection. [22] casts object detection as an iterative classification problem and proposed AttentionNet which achieves more accurate localization. [23] and [24] propose object detection pipelines that completely eliminate region proposal generation stage by predicting category scores and bounding box locations altogether from feature maps. [25] shows the overall performance of object detection can also be improved by using image renderings for data augmentation.

1.2 Contributions

This paper makes the following contributions:

1. We propose a dataset containing diverse small objects to facilitate the study of the applicability of state-of-the-art deep learning-based object detectors for detecting small objects in the image.
2. From systematic experiment design and performance comparison, we augment the R-CNN algorithm, which boosts the small object detection performance by 29.8% on the benchmark dataset.

2 Small Object Dataset

We compose our dataset for the small object detection problem by using a subset of images from both the Microsoft COCO and SUN datasets. We call the dataset the "small object dataset". We manually select ten small object categories where the largest physical dimension of instances in the categories are smaller than 30 centimeters. The selected object categories are "mouse", "telephone", "switch", "outlet", "clock", "toilet paper", "tissue box", "faucet", "plate", and "jar". A small object is not necessarily small in the image. For instance, the "tissue box" may occupy a large portion of an image. We use the ground truth bounding box locations in the COCO and SUN datasets to prune out big object instances and compose a dataset containing purely small objects with small bounding boxes.

The statistics of the small object dataset is shown in Table 1. It contains about 8,393 object instances in 4,925 images. The "mouse" category has the largest number of object instances: 2,137 instances in 1,739 images. The "tissue box" category has the smallest number of instances: 103 instances in 100 images. All the object instances in our dataset are small. Median of relative areas (the ratio of the bounding box area over the image area) of all the object instances in the same category is between 0.08% to 0.58%. This corresponds to 16×16 to 42×42 pixel^2 areas in a VGA image. As a comparison, median of relative areas of object categories in the PASCAL VOC dataset is between 1.38% to 46.40%, as shown in Table 2. Even the smallest object category is much larger than the biggest object category in our dataset.

Our small object dataset is considered more challenging than the PASCAL VOC in at least two ways: First, the appearance cue available for distinguishing a small object from background clutters is much less due to the small size. Second,

Table 1: **Statistics of our small object dataset.** Relative area (%) of each instance is computed as the ratio of the bounding box area over the image area.

Category	mouse	telephone	switch	outlet	clock	toilet paper	tissue box	faucet	plate	jar
Number of images	1739	345	425	916	746	157	100	1094	419	252
Number of instances	2137	363	487	1210	814	175	103	1388	1005	711
Median relative area	0.35	0.38	0.08	0.08	0.25	0.40	0.58	0.43	0.37	0.29
Median top-10% area	2.76	1.99	0.33	0.37	1.92	1.43	1.94	2.02	2.40	1.57

Table 2: Median relative area (%) of the object categories in the PASCAL VOC.

Category	cat	sofa	train	dog	table	mbike	horse	bus	aero	bicycle
Median relative area	46.40	33.87	32.33	30.96	23.73	23.69	23.15	23.04	22.83	14.38
Category	person	bird	cow	chair	tv	boat	sheep	plant	car	bottle
Median relative area	8.14	8.03	6.68	6.09	5.96	3.82	3.34	2.92	2.79	1.38

the number of bounding box hypotheses for a small object in an image is much larger than that for a big object in the PASCAL VOC.

During evaluation, the small object dataset is split into two subsets: one for training and the other for testing. The number of object instances per category in the training set is roughly two times the corresponding number in the testing set. There are no common images between the two sets.

Performance metric: we use the standard performance metric for comparing various object detection algorithms. An object bounding box hypothesis is considered as a true detection if its overlap ratio with the ground truth bounding box is greater than 0.5, where the overlapping ratio is measured using the Intersection over Union (IoU) measure. The detection algorithm returns a confidence score for each object bounding box hypothesis. We vary the threshold and compute the precision recall curve for each object. We then use the average precision of the curve to report the performance of the detector for an object category. The performance of the detector for the entire dataset is measured using the mean Average Precision (mAP) score.

3 R-CNN for Small Object Detection

The R-CNN algorithm [3] has been established as the de facto algorithm for deep learning-based object detection. It significantly outperforms conventional approaches in the PASCAL VOC by capitalizing the following two insights: First, it uses object proposals rather than sliding windows. Before the R-CNN, most object detectors such as DPM adopt a image pyramid plus sliding window approach [26] to generate potential object locations and handle various scales. In the R-CNN pipeline, a fixed number (e.g. 2000) of boxes are proposed per image which most likely contain the target objects. The problem of various scales is also handled automatically by the proposal generation. Fewer but better proposals contribute a lot to the good performance of the R-CNN. Second, it leverages ImageNet pre-trained deep neural network models, which is then fine-tuned using the PASCAL VOC. The pre-training process is proven to be crucial to the performance. Without the pre-training process, the R-CNN works poorly.

Given the region proposals, training an R-CNN object detector generally composing two major steps: supervised pre-training and domain-specific fine-tuning. During supervised pre-training, ImageNet data are used to train the entire network from scratch. In the domain-specific fine-tuning, the weights of the network are initialized by the pre-trained model and trained by the domain-specific data (for example, PASCAL VOC). Training images for the ConvNet are region proposal patches being resized and warped to the required resolution (e.g. 227×227). Both the positive and negative patches are sampled from the region proposals according to certain overlap thresholds.

In the following sections, we investigate into various necessary changes for successfully extending the R-CNN algorithm for small object detection. We follow the same procedure to train our small object detection networks, but based on the nature of the problem, in the domain-specific fine-tuning stage, we only sample the negative patches from the region proposals. The positive patches are generated by randomly deviating from the ground truth box. We also try to balance the positive patches of each category by sampling complementary number of positive patches per category per instance.

The Fast R-CNN algorithm [4] simplifies the R-CNN pipeline by proposing a *ROIPooling* layer that crops the proposals from the feature map instead of the input image. Although the Fast R-CNN reduces the time cost and further improves the performance on PASCAL VOC, the core idea of R-CNN is intact. Adding the *ROIPooling* leads to the primary difference between the two methods: in R-CNN, all the proposal boxes (even small ones) are resized to a canonical size, this means that full feature map is generated for each proposal box at the last pooling layer. However, in Fast R-CNN, a small proposal box gets mapped to only a small map (sometimes 1*1*n) at the last pooling layer. Such a small feature map may lack necessary information for the classification step, adding unnecessary uncertainty into the study. Thus, we feel that the R-CNN is more suitable than the Fast R-CNN algorithm in this case. Moreover, as we do not have much knowledge about how the deep learning-based method works on small objects, the original R-CNN pipeline provides a more convenient way to better understand the problem. For example, it is more convenient to visualize the neuron responses of the R-CNN than the Fast R-CNN. By working with proposal patch input, analyzing the effects of up-sampling and context is also easier. Thus in this paper, we choose to follow the original R-CNN pipeline.

Moreover, in our work, we do not implement bounding box regression. Although bounding box regression is proven as an effective way to increase the localization accuracy, it is not a major issue for small object detection. We believe the challenges come from the region proposal generation and classification, while bounding box regression will be less useful on poor proposal and classification results. So in this paper, we will only focus on generating better region proposals and searching for stronger classifiers.

For all the experiments, our training pipeline consists of two stages: in the first stage, the weights of the ConvNets are initialized with corresponding ImageNet pre-trained models. We then fix the convolutional layers and only update fully connected layers for 8000 iterations with a learning rate of 0.0005. In the second stage, all the layers are updated with a learning rate of 0.00005. We use stochastic gradient descent with momentum of 0.9 for optimization, the batch size is 100. The training is terminated after 80000 iterations.

3.1 Small Proposal Generation

Selective search and edge box are two popular choices for object proposal generation. They use mid-level image cues, such as segments and contours and are object category-agnostic. While the selective search and edge box work well for generating proposals for big objects in the PASCAL VOC. We empirically find them rendering unsatisfactory results for generating small object proposals even after an exhaustive search of the algorithm parameter space. With 2000 object proposals per image, the typical recall rate is lower than 60%, leading to poor performance for detecting small objects using R-CNN. Further investigation shows that both of the algorithms favor salient objects with closed contours and distinctive colors. Since the nature of the small objects are non-prominent, they are non-ideal for small object proposal generation.

The Region Proposal Network (RPN) [14] is the current state-of-the-art method for proposal generation. It attaches nine anchor boxes - derived from three different aspect ratios at three different scales - to each spatial dimension of the feature map from the *conv5_3* layer of the VGG16 network [27] for region proposal classification and bounding box regression. The three aspect ratios used are 0.5 (landscape), 1 (square), and 2 (portrait), and the areas of the square shape bounding boxes at the three scales are 128^2, 256^2, and 512^2 pixel^2, respectively. The RPN achieves good performance for big object proposal generation. But we find that directly applying the RPN to the small object proposal generation results in poor performance. Several modifications are necessary as described below.

We first notice that the RPN anchor boxes are too large. Even the smallest anchor box is much bigger than most instances in our small object dataset. Based on the statistics of the small object size in the dataset, we choose 16^2, 40^2, and 100^2 pixel^2 for the square shape anchor box sizes. For the aspect ratios, we keep the original values used in the original paper. We further notice that the stride length of the *conv5_3* feature map, which is 16 pixels, is too large. It is larger than most of the “switch” and “outlet” objects in our dataset. The other candidate feature maps for attaching the anchor boxes are *conv2_2*, *conv3_3* and *conv4_3*. We empirically compare the performance and find that *conv4_3* renders the best performance for small object proposal generation. The *conv4_3* feature map has a theoretical receptive field of 92x92 pixel^2, which appears to be more appropriate than 196x196 pixel^2 from the *conv5_3* feature map.

For benchmarking the performance of deep learning for small object detection, we also apply the Deformable Part Model (DPM) [28] detector to detect the small object. The DPM detector was the state-of-the-art algorithm on the PASCAL VOC dataset before the R-CNN algorithm. The DPM detector is based on the Histogram of Oriented Gradient (HOG) features and latent support vector

Table 3: Recall rate (%) of the region proposal generation methods.

Method	mouse	tel.	switch	outlet	clock	t. paper	t. box	faucet	plate	jar	average
DPM, 300 prop. per category	70.9	58.0	70.5	80.9	79.1	**86.6**	76.2	69.3	58.0	**63.4**	71.3
RPN original, 300 prop.	85.0	63.4	78.7	73.1	66.0	76.1	50.0	76.0	58.6	31.8	65.9
RPN modified, 300 prop.	**88.4**	**82.4**	**80.9**	**83.1**	**86.9**	83.6	**88.1**	**86.4**	**71.9**	58.9	**81.1**
DPM, 500 prop. per category	73.2	61.8	74.3	82.2	82.5	86.6	78.6	73.9	62.2	**72.9**	74.8
RPN original, 500 prop.	85.7	64.9	79.2	74.7	68.4	77.6	57.1	78.0	61.4	38.2	68.5
RPN modified, 500 prop.	**89.9**	**86.3**	**82.0**	**84.2**	**88.9**	**91.0**	**90.5**	**89.8**	**76.4**	67.1	**84.6**
DPM, 1000 prop. per category	76.5	67.2	78.7	84.2	86.9	89.6	81.0	79.7	68.3	**81.7**	79.4
RPN original, 1000 prop.	87.0	70.2	79.8	75.6	71.7	82.1	66.7	80.9	66.4	46.2	72.7
RPN modified, 1000 prop.	**92.4**	**93.1**	**83.6**	**86.0**	**90.2**	**97.0**	**92.9**	**93.3**	**82.5**	76.4	**88.7**
DPM, 2000 prop. per category	80.2	72.5	82.0	86.2	89.9	92.5	83.3	83.3	73.9	**87.8**	83.2
RPN original, 2000 prop.	87.7	75.6	80.3	76.0	75.1	89.6	76.2	84.0	69.4	54.6	76.9
RPN modified, 2000 prop.	**94.1**	**94.7**	**85.3**	**87.1**	**90.9**	**97.0**	**97.6**	**95.3**	**86.1**	85.2	**91.3**

machine. To accommodate the small object size, we down-sample the root and part template sizes of the DPM detector by half. The DPM is a category-specific object detector. We train a DPM detector for each class.

Evaluation: in Table 3, we compare the recall rate of the proposal generation methods for the small object detection problem. Specifically, we compare the recall performance of using the DPM detector, the original RPN, and the proposed modification of RPN. We vary the number of proposals per image and show the recall numbers. The DPM is category-specific. We use the top scored bounding boxes from all the classes for computing the recall rate. The effective number of bounding boxes are 10 times the number of the RPN. As discussed, the modified RPN renders the best recall performance. For 2000 proposals, the recall rate for the “tissue box” is about 97.6%. The recall rate for the “jar” is the worst. It is 85.2% with 2000 proposals. However, this is still much better than 54.6% achieved by the original RPN method. From the table, we also find that the original RPN algorithm renders worse performance than the DPM algorithm. The proposed modification of the RPN algorithm considers the nature of small object and largely improve the performance. Overall, the proposed modification achieves an average recall rate of 91.3%, which is relatively 19% better than the original RPN method.

3.2 Up-sampling

The first question encountered as extending the R-CNN algorithm to the small object detection is whether to aggressively up-sample the image or not. Unlike the objects in the PASCAL VOC, the bounding boxes of the small objects in our dataset are very small. In Table 4, we show the median bounding box size (square root of the box area) of the objects per category and the corresponding up-sampling ratios required to match the input size (227 × 227 in this case) of the deep convolutional neural networks. We find that, generally, 6 to 7 times up-sampling is required, which will introduce a large amount of up-sampling artifacts. One way to reduce the artifacts is to use low resolution small input patches with a ConvNet deviated from the standard pre-trained models. For example, we can exclude the pre-trained weights in the last few fully connected layers and only use the convolution layers. However, using small patches as input may create other disadvantages:

Table 4: **Up-sampling effects.** Both networks are trained and tested with DPM proposals, 500 per image per category.

	mouse	tel.	switch	outlet	clock	t. paper	t. box	faucet	plate	jar	average
Partial AlexNet	29.8	3.1	5.3	18.0	19.6	15.5	1.9	6.7	**5.4**	2.0	10.7
Full AlexNet	**42.9**	**7.7**	**9.4**	**22.7**	**28.2**	**26.7**	**15.7**	**18.6**	**5.4**	**3.4**	**18.1**
Median size	32.4	54.0	25.5	25.8	38.5	73.1	90.0	50.8	39.2	29.4	45.9
Up-sampling ratio	7.0	4.2	8.9	8.8	5.9	3.1	2.5	4.5	5.8	7.7	5.8

1. The receptive field over small patch is larger than the same receptive field over large patch. This means given a small patch, the network can only look at the object in a coarse scale, thus possibly loses useful information regarding the parts of the object.
2. Small input patch produces lower dimensional feature vector, thus the size of the vector may not be large enough to accommodate all the crucial information.
3. Since all the fully connected layers need to be trained from scratch, we only utilize the partial strength of the pre-trained models.

To answer this question. We design an experiment comparing the two solutions using the following two networks:

1. Partial AlexNet [29]: Using *conv1* to *pool5* layers from the AlexNet. The object proposals are re-scaled to 67×67. The *pool5* layer produces a $1 \times 1 \times 256$ feature vector, which is used to get the final classification scores.
2. Full AlexNet: Using the entire AlexNet structure. The object proposals are up-sampled to 227×227 and contains a large amount of artifacts.

The results are shown in 4. From the table, we found that although with the up-sampling artifacts. The full AlexNet still renders much better performance. So in our following experiments, we will only use the aggressively up-sampled proposal patches as input.

3.3 Context

Context is an important cue for object detection. We expect that it will be even more important for small object detection, since small objects are simple in shape and usually only cover a small image region. The feature extracted from the proposal region is less discriminative, so when only given the proposal region, it can be very difficult to recognize, even for human beings.

We investigate into several methods for incorporating context information to boost small object detection performance, and based on the R-CNN algorithm, we propose a simple method that works quite well. When given an object proposal in an image, in addition to cropping the proposal region, we crop the corresponding context region enclosing the proposal region, with the center coinciding with the center of the proposal region. The context region is set to be several times larger than the proposal region. We then feed both regions into a neural network. The neural network consists of three sub-networks where the

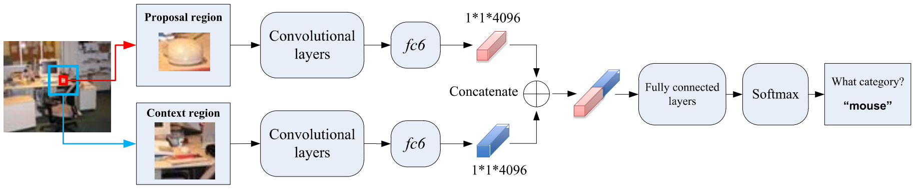


Fig. 2: **ContextNet: the neural network for integrating context information.** The two front-end sub-networks take proposal region patches and context region patches as input respectively, the back-end sub-network takes in the concatenation of the two feature vectors and computes the final classification score.

Table 5: **Results of ContextNet.** All the networks are trained (2000 per image) and tested (500 per image) with modified RPN proposals.

Method	mouse	tel.	switch	outlet	clock	t. paper	t. box	faucet	plate	jar	mAP
Baseline AlexNet	48.2	10.6	8.9	21.4	32.3	**34.1**	**23.0**	25.1	6.7	3.6	21.4
ContextNet (AlexNet, 3x)	54.8	9.1	12.8	**30.7**	28.5	28.4	18.6	**30.8**	**10.6**	**6.4**	23.1
ContextNet (AlexNet, 7x)	**56.4**	**12.2**	**12.9**	26.3	**32.7**	34.0	18.7	26.8	9.9	4.6	**23.5**

first one takes the proposal region as input, the second one takes the context region as input, and the last one takes the concatenation of the outputs of the others as input and computes the final classification score. We call this neural network ContextNet, and the structure is shown in Fig. 2.

The two front-end sub-networks have identical structure. Each consists of a few convolutional layers followed by one fully connected layer, which are derived from the first six layers of the AlexNet (or the equivalent layers of VGG16). Input image regions to the two sub-networks are resized to 227×227 (224×224 for VGG16) patches. Each of the front-end sub-networks outputs a 4096 dimensional feature vector. The back-end sub-network consists of two fully connected layers and outputs the predicted object category label. During training, the front-end sub-networks are initialized using the ImageNet pre-trained model. However, the weights of the two sub-networks evolve separately - the weights are not shared.

Evaluation: we evaluate the performance of the AlexNet-based ContextNet with two variants: the 3x and 7x models. The context region of the 3x model is three times larger than the proposal region in both height and width dimension. The 7x model is defined in a similar way and it uses a very larger context region. We also include the AlexNet R-CNN model as the baseline.

The performance is shown in Table 5. We find that the neural network with context integration achieves better performance than the baseline model. The improvement with the 7x model is slightly better than that with the 3x model. Overall, the relative mAP improvement over the baseline are 7.9% and 9.8% for the 3x and 7x models, respectively. We also investigate a ConvNet-based co-occurrence model, which leverages the detection of big objects to better localize the small objects. The spatial relation between the big and small objects are posed as learnable parameter integrated into an end-to-end training framework. However, we find this method is only effective when attached to the Baseline AlexNet, it does not make any improvement when attached to both ContextNets.

Table 6: **Results of DPM, AlexNet R-CNN, and VGG16 R-CNN.** The AlexNet in row 2 is trained and tested with DPM proposals, 500 per image per category. The AlexNet in row 3 and the VGG16 in row 4 are trained (2000 per image) and tested (500 per image) with modified RPN proposals.

Method	mouse	tel.	switch	outlet	clock	t. paper	t. box	faucet	plate	jar	mAP
DPM	18.9	0.3	1.9	23.0	9.1	18.3	2.0	5.7	2.4	0.4	8.2
DPM prop. + AlexNet	42.9	7.7	9.4	22.7	28.2	26.7	15.7	18.6	5.4	3.4	18.1
RPN prop. + AlexNet	48.2	10.6	8.9	21.4	**32.3**	**34.1**	23.0	25.1	6.7	3.6	21.4
RPN prop. + VGG16	**56.8**	**16.4**	**14.2**	**31.1**	31.9	29.4	**23.4**	**31.3**	**9.3**	**4.2**	**24.8**

Table 7: **Results of ContextNet.** Both networks are trained (2000 per image) and tested (various) with modified RPN proposals.

Method	mouse	tel.	switch	outlet	clock	t. paper	t. box	faucet	plate	jar	mAP
AlexNet, 7x, 300 prop.	**56.9**	**12.4**	**13.6**	**28.0**	32.4	35.6	17.9	**27.2**	9.8	**5.1**	**23.9**
AlexNet, 7x, 500 prop.	56.4	12.2	12.9	26.3	**32.7**	34.0	18.7	26.8	**9.9**	4.6	23.5
AlexNet, 7x, 1000 prop.	55.4	11.2	11.4	25.7	29.5	**37.6**	18.5	25.7	9.1	4.2	22.8
AlexNet, 7x, 2000 prop.	54.9	10.9	10.9	24.6	29.8	35.0	**19.5**	24.8	8.4	3.9	22.3
VGG16, 7x, 300 prop.	**60.6**	13.7	**21.5**	**41.5**	**37.7**	33.3	**22.0**	30.3	15.8	7.2	**28.4**
VGG16, 7x, 500 prop.	60.2	14.0	20.0	40.7	36.4	**35.7**	20.4	**31.4**	**16.0**	**7.7**	28.3
VGG16, 7x, 1000 prop.	59.6	**14.6**	18.9	39.9	36.2	34.9	18.7	30.9	15.3	7.4	27.6
VGG16, 7x, 2000 prop.	58.4	13.7	18.1	38.2	33.6	33.0	18.5	30.1	14.0	7.1	26.5

3.4 Summary

In Table 6, we list the average precision of our R-CNN models on small object dataset, we also list the DPM as a baseline. Not surprising at all, DPM is significantly outperformed by all the deep learning-based models. And deeper network (VGG16) has superior performance over shallow network (AlexNet).

To demonstrate the influence of region proposal quality on the final average precision, we compare two AlexNet models: one using the DPM detection outputs as proposals, and the other use the modified RPN proposals. From Table 3, we know the modified RPN proposals have much higher recall rate than the DPM proposals, and consequently, the AlexNet trained on modified RPN proposals performs much better (Table 6).

Fewer proposals: in Table 7, we show the average precision of the ContextNet using 7x context region on different number of proposals per image. We find it achieves higher average precision on a smaller number of proposals. Small object detection is very vulnerable to false positives. Using a smaller number of proposals eliminates a large amount of potential false positives and improves the average precision. 300 proposals per image produces the best performance.

Stronger pre-trained model: we also experiment with replacing the AlexNet with the VGG16 net to verify if the performance boost in the big object detection due to the stronger pre-trained model is also true for small object detection. The results are shown in Table 7. From the table, we find that the stronger pre-trained model leads to improved performance for all the proposal numbers.

In Fig. 3, we show the detection results of the ContextNet (AlexNet, 7x) model on several images in the testing set. We use a fix threshold and show the output bounding boxes after non-maximum suppression. Since the target objects are too small for visualization, we put a zoom-in window to highlight the output bounding boxes. From the figure, one can see that the small object

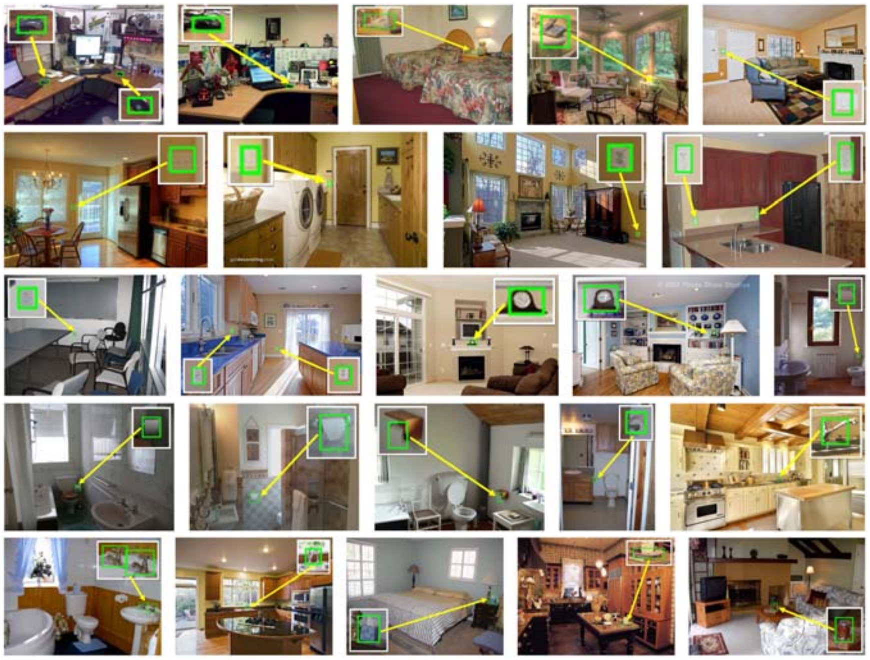
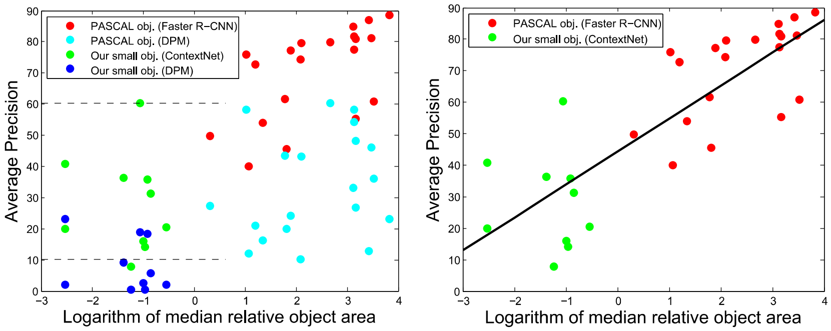


Fig. 3: Examples of the detection results on some testing images.



(a) (b)

Fig. 4: **Comparison of methods on small objects and PASCAL.** In both (a) and (b), a marker represents the mAP of a detector on an object category. Specifically, red represents Faster R-CNN on PASCAL objects, green represents our ContextNet on our small objects, light and dark blue represent DPM on PASCAL objects and our small objects, respectively.

detector works well on many categories. It can detect object instances with very low resolution.

As one of the major purposes of this paper is to study the applicability of the state-of-the-art object detection algorithms to the small object detection problem, by summarizing the findings, we now can answer this question. Our answer

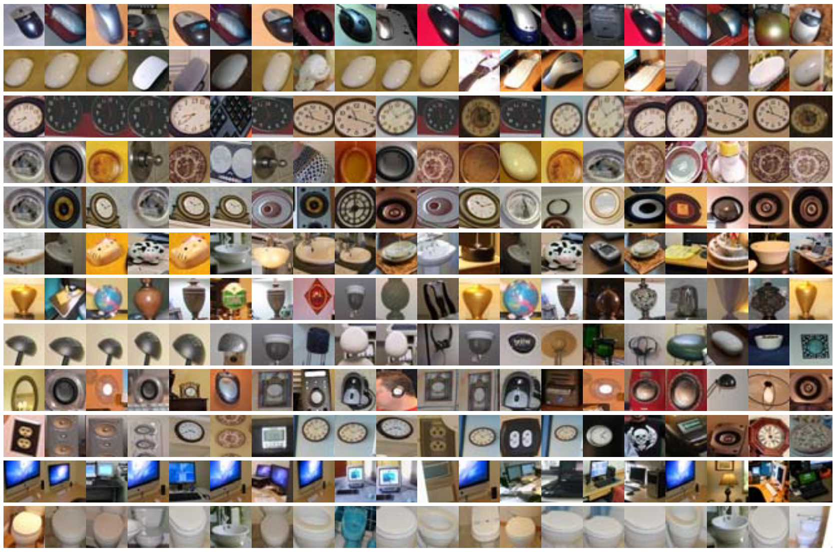


Fig. 5: The proposal patches that have the largest excitation to the neurons in *fc6* of proposal sub-network. Please refer to the main text for further discussions.

is based on two observations: 1) before the R-CNN algorithm, the state-of-the-art object detector on PASCAL VOC was DPM. Since our work is a preliminary stage of small object detection, we think it is comparable to DPM on PASCAL. Shown in Fig. 4a, the average precision of our best model, e.g. ContextNet (VGG16, 7x), on small object categories is distributed in the same range (indicated by the black dashed lines) as that of DPM on PASCAL. Numerically, on the small object dataset, our deep learning-based algorithm (mAP 28.3) has close performance to the DPM on PASCAL (mAP 33.7). 2) on PASCAL VOC, the R-CNN style algorithm improves the mAP of DPM from 33.7 to 70.4. While on the small object dataset, our best model improves the mAP of DPM from 8.2 to 28.3, which indicates the deep learning models are still very effective on small objects. Thus, we think they are applicable to small object detection problem.

4 Visualization

We visualize the neurons in our ContextNet (AlexNet, 7x) model to better understand what the network learns as learning to detect small objects. We plot the training patches that excite each neuron in the $fc6$ layer most for both the proposal and context front-end sub-networks.

In Fig. 5, we display the top 20 image patches with the highest response to several neurons in the proposal sub-network. We find that the patches are dominated by mouse and round shape objects (e.g. row 1 to row 5). This partially explains why the network performs better for the "mouse" and "clock" categories. We also find the neurons in row 2 fire when seeing Apple mouses or similar shapes, while those in row 9 response to oval pattern. In row 10, we can

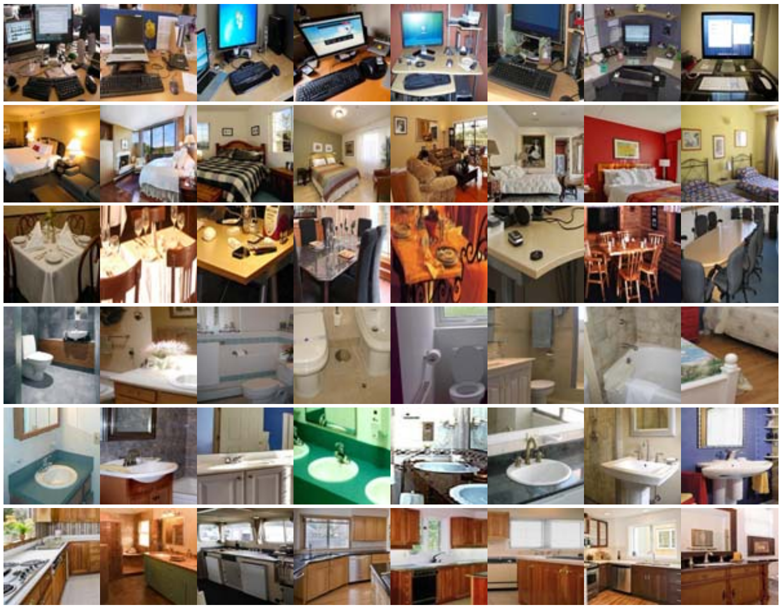


Fig. 6: The context patches that have the largest excitation to the neurons in *fc6* of context sub-network. Please refer to the main text for further discussions

see outlet patches are mixed with speaker and clock patches. The neurons in row 11 and row 12 correspond to a monitor detector and a toilet detector. This is surprising since our dataset does not contain these two object category labels. The figure also suggest that there is not much high-level features to distinguish small objects. Hence, the network relies on basic shape patterns to detect small objects (e.g. row 6 to row 8).

In Fig. 6, we display the top 8 image patches with the highest response to several neurons in the context sub-network. Since the 7x context region covers a large image area, the context patches fire for the same neuron have more diverse patterns. As expected, strong scene-specific patterns exist on many neurons. The neuron in row 1 looks at computers, and the neuron in row 2 evolves for bedroom scene. The neurons in row 3, 4, and 5 respond to tables, toilets and sinks, respectively. The neuron in row 6 activates on kitchen scene. These neurons provide context information to resolve the ambiguity in the proposal patches.

5 Conclusion

We extended the state-of-the-art R-CNN algorithm to deal with the small object detection problem. We composed a small object dataset to facilitate the study. Through detailed experimental validation and analysis, we found that, with a carefully designed region proposal network and context modeling, the deep learning-based object detection algorithm achieves similar performance improvement over the conventional approach for small object detection as it did for big object detection.

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88** (2010) 303–338
2. Torralba, A., Efros, A., et al.: Unbiased look at dataset bias. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (2011) 1521–1528
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (2014) 580–587
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
5. Kembhavi, A., Harwood, D., Davis, L.S.: Vehicle detection using partial least squares. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 1250–1265
6. Morariu, V., Ahmed, E., Santhanam, V., Harwood, D., Davis, L.S., et al.: Composite discriminant factor analysis. In: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. (2014) 564–571
7. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Computer Vision–ECCV 2012. Springer (2012) 340–353
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014. Springer (2014) 740–755
9. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: Exploring a large collection of scene categories. International Journal of Computer Vision (2014) 1–20
10. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International Journal of Computer Vision **104** (2013) 154–171
11. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014. Springer (2014) 391–405
12. Kuo, W., Hariharan, B., Malik, J.: Deepbox: Learning objectness with convolutional networks. arXiv preprint arXiv:1505.02146 (2015)
13. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (2014) 2155–2162
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015) 91–99
15. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A., Hebert, M., et al.: An empirical study of context in object detection. In: Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on. (2009) 1271–1278
16. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M., et al.: Context-based vision system for place and object recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2003) 273–280
17. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. arXiv preprint arXiv:1505.01197 (2015)
18. Gidaris, S., Komodakis, N.: Object detection via a multi-region & semantic segmentation-aware cnn model. arXiv preprint arXiv:1505.01749 (2015)

19. Zhu, Y., Urtasun, R., Salakhutdinov, R., Fidler, S.: segdeepm: Exploiting segmentation and context in deep neural networks for object detection. arXiv preprint arXiv:1502.04275 (2015)
20. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., et al.: The role of context for object detection and semantic segmentation in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (2014) 891–898
21. Zhang, Y., Sohn, K., Villegas, R., Pan, G., Lee, H.: Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. arXiv preprint arXiv:1504.03293 (2015)
22. Yoo, D., Park, S., Lee, J.Y., Paek, A., Kweon, I.S.: Attentionnet: Aggregating weak directions for accurate object detection. arXiv preprint arXiv:1506.07704 (2015)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640 (2015)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.: Ssd: Single shot multibox detector. arXiv preprint arXiv:1512.02325 (2015)
25. Pepik, B., Benenson, R., Ritschel, T., Schiele, B.: What is holding back convnets for detection? In: Pattern Recognition. Springer (2015) 517–528
26. Liu, M.Y., Mallya, A., Tuzel, O., Chen, X.: Unsupervised network pretraining via encoding human design. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2016) 1–9
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010) 1627–1645
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2012) 1097–1105