An Empirical Study of Context in Object Detection

Santosh K. Divvala  
Carnegie Mellon University

Derek Hoiem  
University of Illinois at Urbana-Champaign

James H. Hays  
Carnegie Mellon University

Alexei Efros  
Carnegie Mellon University

Martial Hebert  
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/robotics

Part of the Robotics Commons

Published In  
An Empirical Study of Context in Object Detection

Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, Martial Hebert

1Carnegie Mellon University.  
\{santosh, jhhays, efros, hebert\}@cs.cmu.edu  

2University of Illinois Urbana-Champaign.  
dhoiem@cs.uiuc.edu

Abstract

This paper presents an empirical evaluation of the role of context in a contemporary, challenging object detection task – the PASCAL VOC 2008. Previous experiments with context have mostly been done on home-grown datasets, often with non-standard baselines, making it difficult to isolate the contribution of contextual information. In this work, we present our analysis on a standard dataset, using top-performing local appearance detectors as baseline. We evaluate several different sources of context and ways to utilize it. While we employ many contextual cues that have been used before, we also propose a few novel ones including the use of geographic context and a new approach for using object spatial support.

1. Introduction

There is a broad agreement in the community about the valuable role that context plays in any image understanding task. Numerous psychophysics studies (see [29] for an overview) have shown the importance of context for human object recognition. Several recent computer vision approaches have demonstrated that the use of context improves recognition performance [4, 11, 14, 17, 24, 26, 32, 35, 39, 41]. Yet, in practice, when a high-performance recognition system is required (e.g., for commercial deployment or to enter a recognition competition), people almost always revert to the tried-and-true local sliding window approaches [5, 7].

Why such a disconnect? We believe there are two reasons. First, in all the previous work on context, every approach reported results only on its own, home-grown dataset. Because of this lack of standardization, it becomes very difficult to compare the different approaches to each other, and to the standard non-contextual baseline methods. Second, there is very little agreement in the literature about what constitutes “context”, with poor differentiation between very simple types of context (e.g., using a slightly larger local window) and ones that are much more involved. As a result, it is unclear which, if any, of the contextual approaches might be worthwhile for any given task, and how much of an increase in performance are they likely to produce.

In this work, our goal is to bring context into the mainstream of object detection research by providing an empirical study of the different types of contextual information on a standard, highly regarded test set. This provides us a basis for assessing the inherent limitations of the existing paradigms and also the specific problems that remain unsolved. Our main contributions are as follows: 1) Objective evaluation of context in a standardized setting. We have chosen to participate in the PASCAL VOC Detection Challenge [6] – by far the most difficult, of all object detection datasets. As our baseline local detector, we choose from amongst the top-performing detectors in this challenge. Our results demonstrate that carefully used contextual cues can not only make a very good local detector perform even better but also change the typical error patterns of the local detector to more meaningful and reasonable errors. 2) Evaluation of different types of context. In this study, we look at several sources of contextual information, as well as different ways of using this information to improve detection performance. 3) Novel algorithms. While we employ several contextual cues that have been used before, we also propose a few new approaches, including the use of geographic context and a new approach for using object spatial support.

1.1. Sources of Context

While the term “context” is frequently used in computer vision, it lacks a clear definition. It is vaguely understood as “any and all information that may influence the
way a scene and the objects within it are perceived" [38].
Many different sources of context have been discussed in
the literature [2, 29, 38] and others are proposed here (see
Table 1 for summary). The most common is what we
broadly term local pixel context, which captures the basic
notion that image pixels/patches around the region of interest
carry useful information. The classic trick of increasing
the size of a scanning-window detector to include surrounding
pixels [5, 41] is one simple application, as are more involved
MRF/CRF-based methods, such as [4, 20, 35]. Image segmentation, object boundary extraction, and various object shape/contour models are also examples of local pixel context, as they use the object’s surroundings to define its shape/boundary [31]. 2D scene gist uses global statistics of an image to capture the “gist” of the visual experience [28, 32]. Geometric context aims to capture the coarse 3D geometric structure of a scene, or the “surface layout” [16], which can be used to reason about supporting surfaces [17], occlusions [15], contact points, etc. Semantic context might indicate the kind of event, activity, or other scene category being depicted [1, 22, 28]. It also may indicate the presence and location (spatial context) of other objects and materials [10, 11, 12, 37]. Photogrammetric context describes various aspects of the image capturing process, such as intrinsic camera parameters i.e., focal length, lens distortion, radiometric response [24], as well as extrinsic i.e., camera height and orientation [17]. Illumination context captures various parameters of scene illumination, such as sun direction [21], cloud cover, shadow contrast, whereas weather context would describe meteorological conditions such as current/recent precipitation, wind speed/direction, temperature, season as well as conditions of fog and haze [27]. Geographic context might indicate the actual location of the image (e.g. GPS), or a more generic terrain type (e.g., tundra, desert, ocean), land use category (e.g. urban, agricultural), elevation, population density, etc. [13]. Temporal context would contain temporally proximal information, such as time of capture [9], nearby frames of a video (optical flow), images captured right before/after the given image, or video data from similar scenes [23]. Finally, there is what we broadly term the cultural context, a largely neglected aspect of context modeling. Its role is to utilize the multitude of biases embedded in how we take pictures (framing [36], focus, subject matter), how we select datasets [30], how we gravitate towards visual clichés [34], and even how we name our children [8]!

1.2. Use of Context for Object Detection

While in the previous section we cataloged the many possible sources of context that could be available to a vision system, what we are primarily interested in this paper is how context can be used for the task of object detection. Let us now consider the different aspects of an object detection architecture to see how contextual information could be useful in each.

<table>
<thead>
<tr>
<th>Local Pixel Context</th>
<th>window surround, image neighborhoods, object boundary/shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Scene Gist Context</td>
<td>3D scene layout, support surface, surface orientations, occlusions, contact points, etc.</td>
</tr>
<tr>
<td>3D Geometric Context</td>
<td>3D scene layout, support surface, surface orientations, occlusions, contact points, etc.</td>
</tr>
<tr>
<td>Semantic Context</td>
<td>event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords</td>
</tr>
<tr>
<td>Photogrammetric Context</td>
<td>camera height, orientation, focal length, lens distortion, radiometric response function</td>
</tr>
<tr>
<td>Illumination Context</td>
<td>sun direction, sky color, cloud cover, shadow contrast, etc</td>
</tr>
<tr>
<td>Weather Context</td>
<td>current/recent precipitation, wind speed/direction, temperature, season, etc</td>
</tr>
<tr>
<td>Geographic Context</td>
<td>GPS location, terrain type, land use category, elevation, population density, etc.</td>
</tr>
<tr>
<td>Temporal Context</td>
<td>nearby frames (if video), temporally proximal images, videos of similar scenes, time of capture</td>
</tr>
<tr>
<td>Cultural Context</td>
<td>photographer bias, dataset selection bias, visual clichés, etc</td>
</tr>
</tbody>
</table>

Table 1. Taxonomy of sources of contextual information.

Object Presence. Many objects have typical environments, such as toasters in kitchens or moose in woodlands. The appearance of the scene (gist context), its layout (geometric context), scene or event category/the presence of other objects (semantic context), previous scenes (temporal context) can all help in predicting the presence of an object. Moreover, some objects tend to appear in certain parts of the world (geographic context), and some objects are more likely to be photographed than others (cultural context). Object presence is roughly equivalent to the probability constraint proposed by Biederman [2].

Object Appearance. The color, brightness, and shading of an object will depend on scene illumination (illumination context) and weather (weather context). Camera parameters such as exposure and focal length (photogrammetric context) can help explain intensity and perspective effects.

Object Location. 3D physical constraints, such as objects requiring a ground plane or some other support surface, help to determine likely locations of objects in the scene (geometric context). Moreover, some objects are likely to appear near others, such as people near other people, or in particular relations to objects or materials, such as cars on the road, squirrels in trees, grass below sky, etc (semantic context). Presence of an object at a particular location in nearby scenes can help predict its location in a future scene (temporal context). Photographer biases (cultural context) often provide useful information, such as an object being centered in the image due to photographer framing and its bottom position to be towards the bottom of the image due to roughly level imaging. Object location is roughly equivalent to Biederman’s support and location...
constraints [2].

**Object Size.** Given object presence and location, its size in the image can be estimated. This requires knowing either camera orientation and height above the supporting surface (photogrammetric context), or relative sizes of other known objects in the scene (semantic context) and their geometric relationships (geometric context). Object size is roughly equivalent to Biederman’s size constraint [2].

**Object Spatial Support.** Given object presence, location and size in the image, its spatial support can be estimated in order to: 1) better localize a bounding box; 2) perform more accurate non-max suppression and multiple object separation (by using segment overlap instead of bounding box overlap); 3) estimate a more precise object shape and appearance model. Estimating the spatial support of an object can be assisted by a number of contextual cues. Local image evidence, such as contours/edges, areas of similar color or texture, etc (local pixel context), occlusion boundaries and surface orientation discontinuities (geometric context), as well as class-specific shape prior (semantic context) can all provide valuable information. This use of context is roughly equivalent to Biederman’s interposition constraint [2].

### 2. Approach

In the previous section, we generated a full wish list of contextual cues and their uses that can potentially benefit object detection. In designing our approach, we picked the context cues which could not only be reliably learned given the available data, but also fit the “plug-and-play” philosophy of taking an off-the-shelf local detector and adding contextual information to it. Therefore, in this work, we have used local pixel context, 2D scene gist, 3D geometric, semantic, geographic, photogrammetric and, to a limited extent cultural context cues, while finding that we did not have good training data for the others. Based on these available context sources, we have implemented object presence, location, size, and spatial support uses of context.

#### 2.1. Local Appearance Detectors

To fairly evaluate the role of context, we need to start with a good local detector. Amongst the top-performing PASCAL [6] detectors, we use the UoCTTI [7] detector which was the only publicly available one. Qualitatively, we have observed that the detector achieves substantially better results than that suggested by the raw performance numbers. This is because, although the detector does a fair job in detecting the presence of an object correctly, it often makes mistakes in localizing it, partially due to the fixed aspect ratio of the bounding box and multiple firings on the same object. Thus, some false positives are due to mistakes in the appearance model but others are due to poor localization. We attempt to overcome these problems by augmenting the detector with contextual information.

#### 2.2. Object Presence

To predict the likelihood of observing an object \( o \) given the image \( I \), i.e., \( P(o|I) \), we use the 2D scene gist, 3D geometric, semantic and geographic contexts. The 2D scene gist of an image is computed in the standard way as described in [28]. The geometric context for an image is computed as a set of seven geometric class (ground, left, right, center, sky, solid, porous) confidence maps as described in [16]. These confidence maps are re-sized to 12×12 grids and vectorized to serve as a coarse “geometric gist” descriptor. We use logistic regression [19] to train two separate object presence classifiers based on each descriptor. The use of these descriptors for scene classification has become fairly standard in literature and has shown good results. However, our use of geographic and semantic information is a novel contribution.

For the geographic context, we follow the approach of [13], estimating geographic properties for a novel image by finding matching scenes within a database of approximately 6 million geotagged Flickr photographs (excluding images that overlap with the VOC dataset and photographers). We compute 15 geographic properties such as land cover probability (e.g., ‘forest’, ‘cropland’, ‘barren’, or ‘savanna’), vegetation density, light pollution, and elevation gradient magnitude. We train a logistic regression classifier based on these geographic properties. Object class occurrence is correlated with geography (e.g., ‘boat’ is frequently...
found in water scenes, ‘person’ is more likely in high population density scenes) but the relationship is often weak. For instance, the ten indoor object classes in the VOC dataset cannot be well distinguished by geography.

For semantic context, we use the keywords associated with matching scenes in the im2gps dataset [13] to predict object occurrence. The 500 most popular words appearing in Flickr tags and titles were manually divided into categories corresponding to the 20 VOC classes and 30 additional semantic categories. For instance, ‘bottle’, ‘beer’, and ‘wine’ all fall into one category, while ‘church’, ‘cathedral’, and ‘temple’ fall into another category. For a novel instance, the ten indoor object classes in the VOC dataset are modeled as whole image descriptors. We train a separate logistic regression classifier on the VOC’08 trainset for each of the above feature descriptors. This regression task is reformulated as a series of classification tasks [26], where we first cluster object sizes (using K-means) into five clusters $s_1, s_2, s_3, s_4, s_5$ and then train a separate classifier for each size (i.e., $size < s_2$, $size < s_3$, $size < s_4$, $size < s_5$). The object sizes for training classifiers are calculated using the ground-truth annotations provided in the VOC’08 dataset. The predictions from individual classifiers are combined using another logistic regression classifier trained on the VOC’08 validation set. At testing, we calculate $P(size = k) = P(size < k+1) \times (1 - P(size < k))$, with $\sum_k P(size = k) = 1$ and compute the expected object size as $\sum_k P(size = k) \times center(k)$.

### 2.5. Combining Contexts

The task here is to combine the object detection results with the various context uses, so as to rescoring detection hypotheses that do not agree with the object presence, location and size context predictions to a lower value. Detections that occur at unusual poses should have significantly high score from the base detector for them to be selected in this scheme [26]. First we retrieve the top 100 detections (after non-max suppression) per image for all the training images. For each detection, we retrieve: 1) object presence estimates in terms of the scene gist, geometric context, geographic and semantic context classifier confidences; 2) object location estimates in terms of the confidence of the grid in which the bottom center of the bounding box occurs and also the max confidence in its neighborhood; 3) object size estimates in terms of the predicted height and the negative absolute difference between the bounding box height and the predicted height. We train a logistic regression classifier using the above features on the VOC’08 validation set. We consider a detection hypothesis to be positive if there is at least 50% overlap with a true detection. If any of the above context features are assigned a negative weight during the training process, we retrain the classifier again after setting those features to zero. While testing, we retrieve the top 500 detections for every image (obtained using [7]) and rescoring these detections using the above classifier. These rescoring detections are used by the object spatial support context described in Section 2.6.

In all cases, we evaluate different classifiers for modeling the various contexts and also for combining them - kNN, SVM (linear and RBF) [18], logistic regression (L1 and L2). We found L1-regularized logistic regression to perform at least as well as other.
2.6. Object Spatial Support

The task here is to compute the object’s spatial support given an (often poorly localized) candidate detection and its confidence. This is a much easier problem than the general segmentation problem because the type of object and its rough location in the image is known. We implement a simple segmentation approach based on graph cuts.

**Unary Potential:** Our unary features model the object class appearance, a position/shape prior, and the object instance appearance. For class appearance, we compute K-means clustered \( \text{L*a*b*} \) color (\( K=128 \)) and texton \( \text{[40]} \) (\( K=256 \)) histograms, geometric context confidences \( \text{[16]} \) and the probability of background confidences (trained using \( \text{[16]} \) on LabelMe \( \text{[33]} \) examples), quantized to ten values. The features are the class-conditional log-likelihood ratios i.e., \( P(\text{feature} \mid \text{object})/P(\text{feature} \mid \text{background}) \) given the quantized value, as estimated on the segmentation ground-truth in the VOC’08 trainset. The position/shape prior is computed as the log-likelihood ratio for each pixel given its location with respect to the location and scale of the bounding box. The object instance appearance is modeled by taking the log ratio of the histograms computed within and outside the bounding box. Altogether, this gives us thirteen features (class appearance: color, texture, seven geometric classes, probability of background; location/shape prior; instance appearance: color, texture), plus a prior.

**Pairwise Potential:** The pairwise potentials are modeled using probability of boundary (Pb) \( \text{[25]} \) and probability of occlusion \( \text{[15]} \) confidences. They are set to be the negative log-likelihood of boundary, and separate weights are learned for horizontal, vertical, and diagonal neighbors (eight-connected neighborhood).

**Learning:** Unary and pairwise potentials are learned together using pseudo-likelihood, maximizing the likelihood of a pixel given the ground truth values of its immediate neighbors. After learning the potentials, we make small adjustments to them (specifically the unary prior and shape/position) for each object to give good results on the validation set (as the automatically learned prior weight tends to lead to under-segmentation).

**Inference:** Each candidate detection is segmented using graph cuts \( \text{[3]} \), after resizing the image so that the object length is 100 pixels. (The resizing is important to achieve good segmentations for objects of different sizes). For computational reasons, only post-context detections that are above a threshold (corresponding to 0.025 precision in validation) are processed. See Figure 4 for an illustration.

After segmenting an object, we represent its appearance with histograms of \( K \)-means quantized color, texture and HOG features \( \text{[5, 7]} \) (\( K=128, 256, 1000 \) respectively), and a measure of segmentation quality (defined as the difference between the energy of the graph cut solution and the energy of all pixels labeled as background, normalized by the number of object pixels). A classifier on these segment-based features is trained using a linear SVM \( \text{[18]} \) for each object class. When testing, we reclassify the object based on the features computed within the segment and assign the final detection score as a linear combination of the original score and this segment-based score. This is similar to the segmentation-based verification strategy of Ramanan \( \text{[31]} \), who instead uses the pixels of the segmentation mask as features.

Beyond rescoring, we also use the computed spatial support to improve non-maximum suppression and localization. If two candidate detections yield segmentations with pixel overlap (intersection over union) of at least 0.5, the candidate with the lower score is removed. A new bounding box is estimated by taking a weighted average of the original bounding box and a tight fitting box around the segment. The box is then adjusted by a fixed percentage of width or height to account for bias (e.g., consistently undersegmenting the legs of chairs). Parameters are learned on the validation set. For few classes (sofa, bicycles), the spatial support cannot be reliably estimated, resulting in a decrease in performance. To avoid this, a per-class parameter is learned on the validation set to decide if the rescoring/improved localization step is applied during the testing phase.

3. Experimental Results and Analysis

The PASCAL 2008 dataset \( \text{[6]} \) consists of roughly 10,000 images (50% test, 25% train, 25% validation) containing more than 20,000 annotated objects from 20 classes. The images span the full range of consumer photographs, including indoor and outdoor scenes, close-ups and landscapes, and strange viewpoints. The dataset is extremely challenging due to the wide variety of object appearances and poses and the high frequency of major occlusions.

**Per-Class Detection Results.** Table 2 displays the detection results obtained on the VOC’08 test set with and without using context. The results are reported using the average precision (A.P.) metric, which is the standard mode of evaluation in the PASCAL VOC challenge. Our experiments show the importance of reasoning about an object within the context of the scene, as we are able to boost the average precision of the original UoCTTI’07 detector from 18.2 to 22.0. The table includes a comparison with the recently released UoCTTI’08 to demonstrate the generalizability of
agreement. Context is particularly helpful when the objects have impoverished appearance i.e., when they are small and occluded. In this case when the overlap is under 0.25 (poor localization); and 3) ‘Bgnd’ denotes the background. Observe that there are much fewer detections on background upon adding contextual information. Further the remaining mistakes that occur after adding context are more reasonable where the confusions are between similar classes such as bicycles getting confused with motorbikes, buses with cars, cows with horses and sheep etc.

### Analysis of sources and uses of context

We measured the influence of each of the individual sources of context for the tasks of object presence, location and size estimation. For object presence (“Does the object appear in the image?”), the mean A.P. across 20 classes using individual cues was as follows: Semantic (25.6%), Gist (23.9%), Geometric (21.5%) and Geographic (15.1%), while using all the cues gave 31.2%. For object location (“In which of the 25 grids is the bottom of the object located?”), the mean A.P. across 20 classes was: Gist (3%), and Geometric (2.5%), while using both cues gave 6.5%. Finally for object size estimation, the average prediction error i.e., $\frac{\sum_{\text{instances}}|\log(\text{trueHeight/predictedHeight})|}{\text{#instances}}$ across 20 classes was: Photogrammetric (1.08), Gist (1.16) and Geometric (1.18) while using all the cues gave an error of 1.086. The baseline error of simply predicting the mean object height is 1.22. To analyze the importance of the uses of context i.e., object presence, location and size, we run our detection experiments in a leave-one-out methodology. The mean A.P. across 20 classes for each of the case is as follows: 1) excluding object presence - 19.8%; 2) excluding object location - 20.2%, 3) excluding object size - 19.2%, 4) excluding all the three (i.e., simply running the base detector) - 18.5%, and 5) including all the three - 20.5%. Thus we observe that the object size context is the strongest, while object location is our weakest context use.

### Change in Accuracies with respect to size and occlusion

We also analyzed the change in accuracies as a function of two different object characteristics/types, namely occlusion and size (Table 3). The type ‘occluded’, ‘non-occluded’ and ‘difficult’ are as defined in the PASCAL annotations. The type ‘small’/‘large’ refers to the object instances that were lesser/greater than the median object area in the image. Context is particularly helpful when the objects have impoverished appearance i.e., when they are small and occluded in the image.

We also analyzed at the results by segregating objects into man-made vs. natural object categories. In this case,
Figure 5. Confusion matrices (a) Without Context (b) With Context (c) Change in confusions i.e., (b-a) quantized into three values - white indicates positive change, black indicates negative change, and gray indicates negligible change (within +/- 0.05). Observe that many fewer extra detections, localization errors, and background detections occur upon the addition of contextual information. Further, the remaining errors made are more reasonable – cows getting confused with horses, cats confused with dogs etc.

Figure 7. Images in which addition of context had the largest decrease in the top detection confidence. (Red Dotted: Detector, Green Solid: Detector+Context.) Performance is hurt mostly in cases when the objects occur outside their typical context.

Figure 8. Mistakes/Errors made despite augmenting a top-performing object detector with several contextual cues. Such scenarios present a challenge to existing detection algorithms.

we observed that for natural objects (i.e. bird, cat, cow, dog, horse, person, sheep) the improvement in A.P. is 2.1 (from 14.4 to 16.5), while for man-made objects (i.e. aeroplane, bicycle, boat, bottle, bus, car, chair, diningtable, motorbike, pottedplant, sofa, train, tvmonitor), it is 0.8 (from 20.2 to 21.0).

Qualitative Analysis. Figure 6 displays some of the qualitative results showing the largest increases and decreases in detection confidences after adding contextual information. Although context almost always helps in improving the detector performance, there are certain scenarios where it hurts. Figure 7 displays some cases where the addition of context leads to some of the original highly confident detections being discarded. Finally in Figure 8, we display the mistakes/errors that still occur despite augmenting a top-performing detector with several contextual cues. Most errors are amongst classes that share similar contexts, e.g., cats confused with dogs, airplanes confused with birds etc. Such confusions are subtle and present a challenge to the existing detection algorithms. We believe a more object specific appearance model would be required to avoid such errors.

4. Discussion

In this paper, we have presented an empirical analysis of the role of context for the task of object detection.

By achieving substantial gains on the challenging PASCAL VOC dataset, we have reaffirmed that contextual reasoning is a critical piece of the object recognition puzzle. Context not only reduces the overall detection errors, but, more importantly, the remaining errors made by the detector are more reasonable. Many sources of context provide a large benefit for recognizing a small subset of objects, yielding a modest average improvement. This highlights the importance of evaluation on many object types as well as the need to include many types of contexts if good performance is desired for a wide range of objects.

Several issues remain to be explored for making context an integral part of object detectors. In this work, we have performed simple implementations of different context sources and uses. Each of these could be improved with further study. Further we have used a naive combination scheme to combine the various contexts. A more sophisticated scheme would offer better gains. Finally, an iterative feedback-based framework connecting the detector and the various contexts together is worth exploring.

Acknowledgments. We thank the PASCAL VOC organizers (Mark Everingham) for evaluating our results on the VOC 2008 testset. This research was supported in part by the NSF Grant IIS-0745636, IIS-0546547 and Kodak. DH was supported by a Beckman Fellowship.
Figure 6. Images for the bike, diningtable, and train classes for which the best detections had the largest increase and decrease in confidence with the addition of context. In these cases the local appearance and global context disagree most strongly. When the addition of context increases confidence (left) it is because a detection is in a reasonable setting for the object class, even if the local appearance does not match well (motorbikes on top row share context with bicycles). When the addition of context decreases confidence (right) it is typically pruning away spurious detections that had high confidence scores from the local detector. (Red Dotted: Detector, Green Solid: Detector+context)

References