An Empirical Study of Context in Object Detection

Santosh K. Divvala[1], Derek Hoiem[2], James H. Hays[1], Alexei A. Efros[1], Martial Hebert[1]

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

[2]University 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 pro-
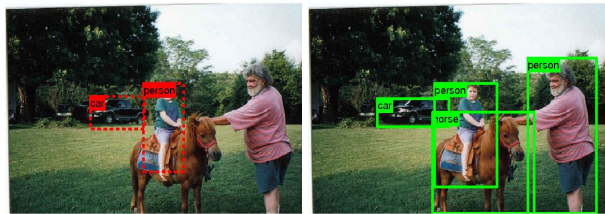

Figure 1. On the challenging PASCAL VOC dataset, even the best local-window detectors [7] often have problems with false positives, poor localization, and missed detections (left). In this paper, we enhance these detectors using contextual information (right). Only detections above 0.5 precision are shown. (Red Dotted: Detector, Green Solid: Detector+Context)

duce.

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, dessert, 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.

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

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.
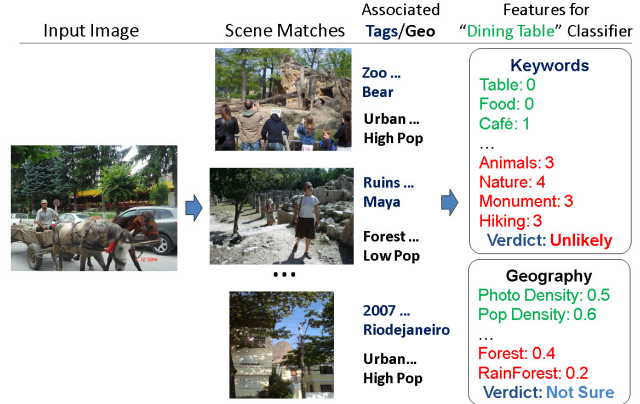

Figure 2. Geographic and Semantic (keyword) context: Geographic properties and keywords associated with the scene can help predict object presence in an image. The base detector finds a dining table in this input image (see figure 6), while the context indicates that a dining table is unlikely.

In this work, we use the detector trained on the VOC'07 trainval set, and use the VOC'08 trainval set for learning the context classifiers (described below). This ensures that the baseline detector and context are trained on different datasets to avoid overfitting. To help ensure that few true detections are missed by the detector, we reduce the threshold for detection such that there are at least 1000 detections per image per object .

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
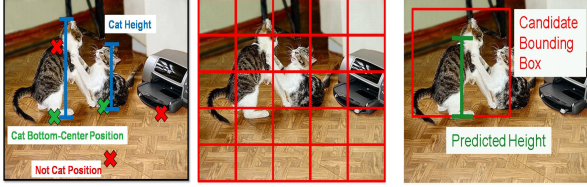
| Location | Size |

Figure 3. Object properties such as bottom-center position and height are used for modeling object location (see section 2.3) and object size (see section 2.4) respectively.

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 image we build a histogram of the keyword categories that appear among the 80 nearest neighbor scenes. We use logistic regression to predict object class based on this histogram. Keywords from Internet images are very noisy and sparse (the im2gps database averages just one relevant keyword per image), but they are quite discriminative when they do occur. All the above classifiers are trained on the VOC'08 trainset.

2.3. Object Location

The goal is to predict *where* the object(s) are likely to appear in an image given that there is at least one object occurring in the image i.e., $P(x|o, I)$. To train this location predictor, we divide the image into $n \times n$ grids ($n = 5$) and train for each grid, two separate logistic regression classifiers [19], one each for the whole image scene gist and the whole image 3D geometric context descriptors as described earlier. The classifiers are trained using the VOC'08 trainset. A grid is labeled as a positive example if the bottom mid-point $(\frac{x_{left}+x_{right}}{2}, y_{bottom})$ of a bounding box falls within it (See figure 3). We then combine the predictions of the above two classifiers using another logistic regression classifier trained on the VOC'08 validation set. For some classes, a few grid cells end up having no (or very few) positive examples (e.g., dining tables never occur in the (1,1) grid). No classifiers were trained for such grid cells and the confidence of finding an object in this location was set to a minimum value while testing.

2.4. Object Size

The idea here is to predict the size (as log pixel height) of an object, given its location in the image i.e., $P(h|x, o, I)$

as illustrated in figure 3. This is learned using three types of contextual cues: 1) photogrammetric context modeled in terms of viewpoint estimates [17] (relative y-value) and the object depth [15] (value at the bottom mid-point of an object bounding box); 2) 2D scene gist; and 3) 3D geometric contexts (the latter two 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 $P(size = k)$ as $P(size < k+1)*(1 - P(size < k))$, with $\sum_k P(size = k) = 1$ and compute the expected object size as $\sum_k P(size = k) * center(k)$.

2.5. Combining Contexts

The task here is to combine the object detection results with the various context uses, so as to rescore those 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 [19] 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 rescore these detections using the above classifier. These rescored 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.
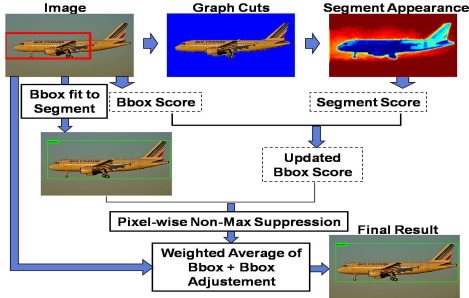
Figure 4. Modeling Object Support (see section 2.6).

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 L*a*b* color (K=128) and texton [40] (K=256) histograms, geometric context confidences [16] and the probability of background confidences (trained using [16] on LabelMe [33] examples), quantized to ten values. The features are the class-conditional log-likelihood ratios i.e., P(feature | object)/P(feature | 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) [25] and probability of occlusion [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 [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 [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 [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 [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 [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

Objects	UoCTTI 2007	+Context		UoCTTI 2008	+Context	
		+Scene	+Scene +Support		+Scene	+Scene +Support
plane	18.8	21.3	**34.5**	28.7	26.8	**32.7**
bike	**33.5**	31.7	32.7	**44.6**	42.9	42.9
bird	9.3	9.9	**12.3**	0.5	**5.0**	5.0
boat	10.4	10.6	**11.0**	12.6	**13.1**	13.1
bottle	22.9	**23.2**	22.4	**28.8**	27.8	27.8
bus	**19.2**	17.7	18.5	22.7	**23.9**	23.9
car	25.1	26.0	**27.8**	31.9	31.6	31.6
cat	6.7	15.8	**21.6**	14.4	18.1	**19.8**
chair	13.3	**14.1**	8.8	15.9	**17.4**	17.4
cow	**16.6**	14.7	14.1	**14.4**	12.3	12.3
dtable	15.0	**18.4**	15.2	12.0	**21.4**	21.4
dog	6.3	7.9	**17.8**	**11.4**	7.7	9.4
horse	24.6	26.6	**27.4**	34.3	**35.7**	35.7
mbike	32.7	34.0	**40.9**	**37.7**	37.1	37.1
person	26.4	28.7	**37.4**	36.6	**39.5**	39.5
pplant	**11.2**	10.8	11.2	8.6	**12.6**	12.6
sheep	10.9	**12.0**	7.0	12.1	**13.5**	13.2
sofa	11.6	**13.7**	13.5	15.0	**15.8**	15.8
train	16.0	17.6	**28.2**	30.1	31.4	**32.2**
tv	32.9	33.3	**38.5**	34.7	**35.2**	35.2
Mean	18.2	19.4	**22.0**	22.4	23.4	**23.9**

Table 2. Detection Results on PASCAL VOC 2008 testset. The first column is the average precision (A.P.) obtained using the base detector. The second and third column show the A.P. obtained upon the addition of the scene context (object presence, location and size) and the spatial support context. Context aids in improving the detection results for many object classes.

Type	Mean A.P.		Most Improved	Least Improved
	w/o	w/		
Small	6.7	12.0	planes (5.4 to 24.8)	pplant (10.3 to 5.9)
Large	9.3	9.7	dtable (4.5 to 9.3)	sheep (5.4 to 0.7)
Occluded	4.8	7.5	cat (3.1 to 13.8)	mbike (18.7 to 16.5)
Non-Occluded	10.4	11.5	dog (2.5 to 7.4)	chair (12.5 to 5.1)
Difficult	0.2	0.3	dtable (0.3 to 2.9)	chair (2.2 to 0.1)

Table 3. Average Precision w.r.t. two object types, Size and Occlusion. For each type, we display the mean A.P. across all object instances without ('w/o') and with ('w/') context along with most/least improved classes. Context particularly helps when objects have impoverished appearance.

our results. We also display the relative improvement obtained by the scene context (presence, location and size), and the spatial support context. We observe that both pieces of information contribute towards the increase in performance (however they cannot be compared on an absolute scale as the output of one process is the input to the other). Notice that for many classes there is a large improvement (e.g., airplane, cat, person, and train), while for some (e.g., bicycles, cows) there is a small drop in performance indicating that the benefit of context varies per class. It must be noted that our numbers cannot be directly compared to the official PASCAL VOC 2008 challenge rankings as our approach involves the usage of external datasets (VOC 2007 and Flickr images). Comparing the results obtained using the two different detectors reveals similar performance by our contextual information in either case. Therefore the rest of our analysis is conducted using the UoCTTI'07 detector on the VOC'08 validation set.

Change in Confusion matrices. Figure 5 displays the change in the types of mistakes that are made after adding contextual cues. The confusion matrix is computed as usual, except that we include three new classes: 1) 'extraDet' addresses the scenario in which the overlap of a box is greater than 0.5 on an already detected object (extra detection); 2) 'poorLoc' includes scenarios in which overlap is between 0.25 and 0.5 (poor localization); and 3) 'Bgnd' denotes the case when the overlap is under 0.25 (fired on the background). Observe that there are much fewer extra detections (better non-max suppression), fewer localization errors, and

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., $\frac{\sum |log(trueHeight/predictedHeight)|}{\#instances}$ 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,
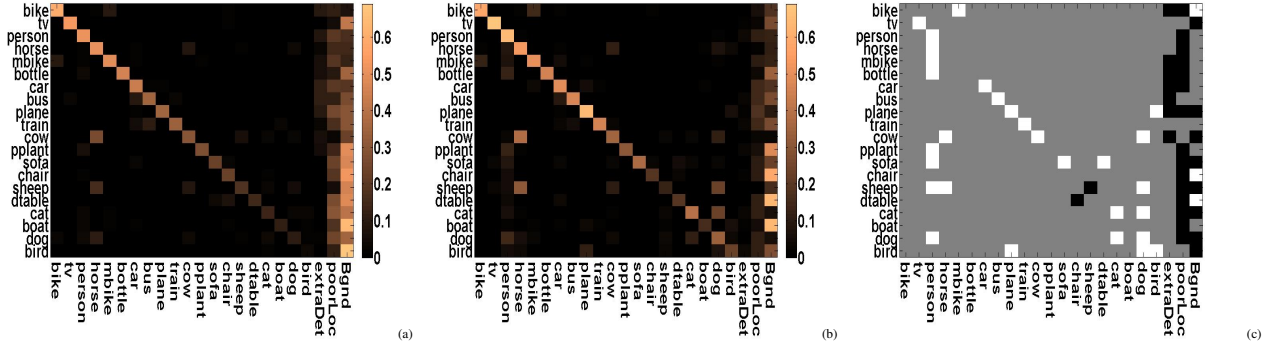
Figure 5. Confusion matrices (a) Without Context (b) With Context (c) Change in confusions i.e., (b-a) quantized into three values - white indicates positive change, black indicates negative change, and gray indicates negligible change (within +/- 0.05) . Observe that many fewer extra detections, localization 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.
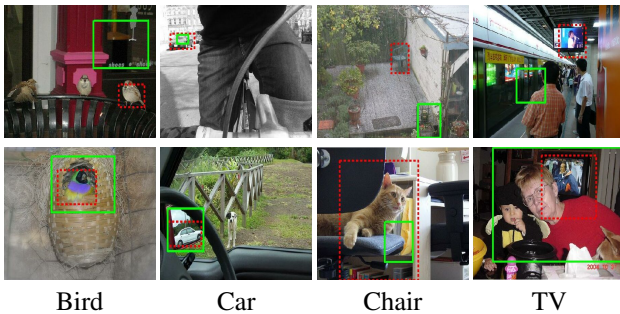

| Bird | Car | Chair | TV |

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.

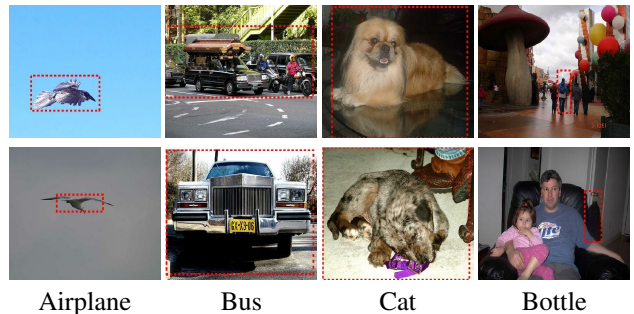| Airplane | Bus | Cat | Bottle |

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.
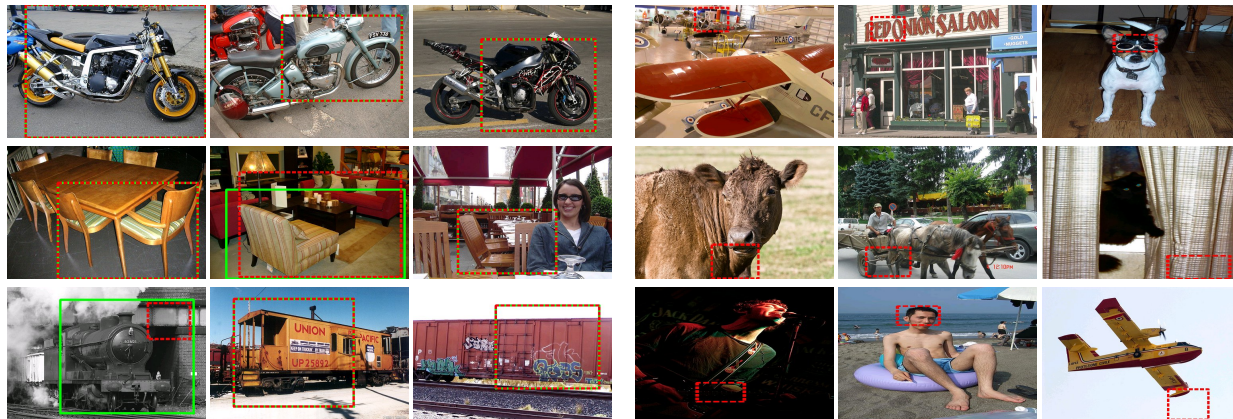
| Largest increase in confidence | Largest decrease in confidence |

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

[1] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, 2004.

[2] I. Biederman. On the semantics of a glance at a scene. In M. Kubovy and J. R. Pomerantz, editors, *Perceptual Organization*, chapter 8. Lawrence Erlbaum, 1981.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[4] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc. ECCV*, 2004.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[6] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, , and A. Zisserman. The pascal visual object classes challenge results, 2008. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *CVPR*, June 2008.

[8] A. Gallagher and T. Chen. Estimating age, gender and identity using first name priors. In *CVPR*, 2008.

[9] A. Gallagher, C. Neustaedter, J. Luo, L. Cao, and T. Chen. Image annotation using personal calendars as context. In *ACM Multimedia*, 2008.

[10] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Technical Report UCSD CS2008-0928*, 2008.

[11] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[12] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

[13] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. *CVPR*, 2008.

[14] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *Proc. ECCV*, 2008.

[15] D. Hoiem, A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. *ICCV*, 2007.

[16] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1), 2007.

[17] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1), 2008.

[18] T. Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press*, 1999.

[19] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. In *Journal of Machine Learning Research*, pages 1519–1555, June 2007.

[20] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proc. ICCV*, 2005.

[21] J.-F. Lalonde, S. G. Narasimhan, and A. A. Efros. What does the sky tell us about the camera? In *ECCV*, 2008.

[22] L.-J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *ICCV*, 2007.

[23] C. Liu, J. Yuen, A. B. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008.

[24] J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum. In *IEEE Singal Processing Magazine*, 2006.

[25] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *Proc. CVPR*, 2008.

[26] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In *Proc. NIPS*. MIT Press, 2003.

[27] S. Narasimhan and S. Nayar. Vision and the atmosphere. In *IJCV*, 2002.

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[29] A. Oliva and A. Torralba. The role of context in object recognition. *Trends Cogn Sci*, November 2007.

[30] J. Ponce and et al. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*. Springer-verlag LNCS, 2006.

[31] D. Ramanan. Using segmentation to verify object hypotheses. In *CVPR*, 2007.

[32] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[33] B. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. In *IJCV*, 2007.

[34] J. Salavon. 100 special moments. http://salavon.com/SpecialMoments/SpecialMoments.shtml.

[35] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[36] I. Simon and S. M. Seitz. Scene segmentation using the wisdom of crowds. In *ECCV*, 2008.

[37] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *Proc. CVPR*, 2003.

[38] T. M. Strat. Employing contextual information in computer vision. In *In Proceedings of ARPA Image Understanding Workshop*, 1993.

[39] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[40] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62(1–2):61–81, Apr. 2005.

[41] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 2006.