

Contextual Object Detection with a Few Relevant Neighbors

Ehud Barnea and Ohad Ben-Shahar

Dept. of Computer Science, Ben-Gurion University
Beer-Sheva, Israel
`{barneah, ben-shahar}@cs.bgu.ac.il`

Abstract. A natural way to improve the detection of objects is to consider the contextual constraints imposed by the detection of additional objects in a given scene. In this work, we exploit the spatial relations between objects in order to improve detection capacity, as well as analyze various properties of the contextual object detection problem. To precisely calculate context-based probabilities of objects, we developed a model that examines the interactions between objects in an exact probabilistic setting, in contrast to previous methods that typically utilize approximations based on pairwise interactions. Such a scheme is facilitated by the realistic assumption that the existence of an object in any given location is influenced by only few informative locations in space. Based on this assumption, we suggest a method for identifying these relevant locations and integrating them into a mostly exact calculation of probability based on their raw detector responses. This scheme is shown to improve detection results and provides unique insights about the process of contextual inference for object detection. We show that it is generally difficult to learn that a particular object *reduces* the probability of another, and that in cases when the context and detector strongly disagree this learning becomes *virtually impossible* for the purposes of improving the results of an object detector. Finally, we demonstrate improved detection results through use of our approach as applied to the PASCAL VOC and COCO datasets.

Keywords: Context · Object detection.

1 Introduction

The task of object detection entails the analysis of an image for the identification of all instances of objects from predefined categories [7,11]. While most methods employ local information, in particular the appearance of individual objects [5], the contextual relations between objects were also shown to be a valuable source of information [23,26]. Thus, the challenge is to combine the local appearance at each image location with information regarding the other objects or detections.

Even when focusing on how context could influence the detection of a single object, one immediately realizes that the difficulty stems from the varying number of objects that can be used as predictors, their individual power of prediction, and more importantly, their *combined* effect as moderated by the complex

interactions between the predictors themselves. Unfortunately, however, most previous works that employ relations between objects and their context focused on pairwise approximations [30,26,1], assuming that the different objects that serve as sources of contextual information do not interact among themselves.

More current detectors based on convolutional neural networks are also able to reason about context since the receptive field of neurons grows with depth, eventually covering the entire image. However, the extent to which such a network is able to incorporate context is still not entirely understood [19]. To include more explicit contextual reasoning in the detection process, several approaches suggested to include additional layers such as bidirectional recurrent neural networks (RNNs) [2], or attention mechanisms [17]. These methods have shown to improve detection results, but the types of contextual information they can encode remains unclear. Additionally, such networks are not able to reason about object relations in a manner invariant to viewpoint, requiring training data in which all meaningful relations between all groups of objects are observed from all relevant viewpoints.

As mentioned above, some complications in addressing the full fledged contextual inference problem may emerge from attempting to model the relations between all (detected or predicted) objects to *all* other detections or image locations. But in reality, such comprehensive contextual relations are rarely observed or needed, as the existence of objects (or lack thereof) is correlated to just *few* other locations in space (and thus to the detected objects in those locations). This notion is exemplified in Figure 1. In this paper we employ exactly this assumption to calculate a score for any given query detection. To do so, we first define the relevant context for that detection as the (few) other most informative detections, and then we use only these detections to calculate (in a closed form fashion) the probability that the query detection is indeed an object.

The suggested approach, facilitated by the decision to employ only few informative detections, provides several contributions. First, it is shown to improve the results of state-of-the-art object detectors. Second, unlike previous methods that require costly iterative training procedures, training our model is as quick and simple as just counting. Third, we represent object relations in a framework that allows to incorporate scale-invariant relations, reducing the number of needed examples, thus simplifying the training phase even further. Finally, using the derived calculation we observe various aspects and obtain novel insights related to the contextual inference of objects. In particular, we show that the effect of context is relative to the prior probability of the query object. As we show, this typically small quantity makes it difficult to infer when an object *reduces* the probability of another, and it practically prohibits the *improvement* of detection probability when the context *strongly disagrees* with the raw detector result. These and further observations and insights are analyzed in our results.

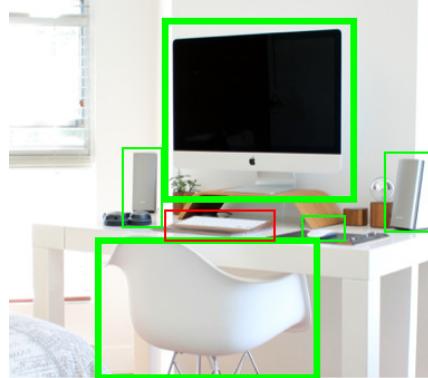


Fig. 1. An example scene with overlaid detections where thicker boxes represent higher confidence. Employing the green box detections as the context of a query red box detection can affirm (or reject) the presence of a keyboard. In this case, the contextual information supplied by the low confidence speakers and mouse detections may become redundant once the confidently detected chair and monitor have been accounted for. In this paper we propose a rigorous model that makes a decision for each query detection by identifying and using only its most informative detections for a more tractable decision.

2 Relevant Work

Valuable information regarding image or scene elements may be obtained by examining their context. Indeed, different kinds of context were employed for various inference problems [31,27,12,13,20]. In this work we focus on employing long-range spatial interactions between objects. This problem has also gained attention, where the standard framework is that of employing fully-connected Markov random fields (MRF) and conditional random fields (CRF) with pairwise potentials [30,6,21,9]. This pairwise assumption is at the heart of the decision process, as it entails that a decision about an object is made by employing information supplied by its neighbors but without considering interactions *between* these neighbors. Other notable works make such assumption, where all detections affect each other in a voting scheme that is weighted by confidence [26], as a sum weighted by the probability of an object given by the detector [23], or a more complex voting mechanism that favors the most confident hypotheses and gives higher relevance to the object relations observed during training [24]. Yet another popular tool for reasoning with many elements is to employ linear classifiers [8,32] or more complicated set classifiers [4] in a pairwise interaction scheme.

Seeking more accurate representations, schemes that incorporate higher order models have been suggested for different kinds of problems in computer vision other than contextual object detection [14]. One such noteworthy model was suggested for objects that were not detected at all, where new detection

hypotheses are generated by sampling pairwise and higher order relations using methods from topic modeling [22]. While this method was facilitated by context, it did not play a role in re-scoring existing or new detections. Other higher order models include neural networks that implicitly or explicitly reason about context as discussed in the Introduction.

A fundamental aspect of our work deals with finding the most relevant set of location variables for a prediction about another location. This problem can be abstracted as finding the structure of a graph (even if only locally) and algorithms for doing so can be grouped to roughly three types [15]. *Constraint-based methods* make local decisions to connect dependent variables, *score-based methods* penalize the entire graph according to an optimization criterion, and *model averaging methods* employ multiple graph structures. While our problem is better related to the constraint-based approach, most algorithms in this class seek to find the structure without considering current beliefs, a set of measures that change dramatically after observing the detections. To better support such cases, Chechetka and Guestrin [3] proposed to learn evidence-specific structures for CRFs, in which a new structure is chosen based on given evidence. This approach, however, is limited to trees. In a different approach, contextual information sources were dynamically separated to those that accept or reject each detection [33], but in a non-probabilistic framework. To facilitate structure learning in our extended graph configurations we therefore propose a different regime based on local structure exploration for each variable during the process of belief propagation. As will be discussed, our computational process is inspired by Komodakis and Tziritas [16] since it prioritizes variables for the message passing process according to their confidence regarding the labels they should be assigned.

3 Suggested Approach

The input to our algorithm is a set of detections $\mathcal{Y} = \{Y_1, Y_2, \dots, Y_n\}$, such that each detection $Y_i = (t_i, l_i, s_i, c_i)$ comprises type t_i , location l_i , size s_i , and confidence c_i . A random variable X_i is created for each detection Y_i , denoting the probability of having an object of type t_i at location l_i with size s_i . For the sake of brevity, in the remainder of this text we refer to X_i as representing an empty location if it is indeed empty, or as a location containing an object of different type or size.

Our goal is to calculate a new confidence for each location variable X_i using detections \mathcal{Y} . To do so, we calculate the probability $P(X_i|\mathcal{Y})$ in a belief propagation process, where the context of X_i is dynamically selected as the most informative *small* set \mathcal{N}_i of location variables $\mathcal{X}_{j \neq i}$, which is the set of all location variables except X_i . The initial beliefs are determined according to the detector, followed by iterations in which an updated belief is calculated for each X_i by identifying the best set \mathcal{N}_i and employing the current belief of its variables for a decision about X_i . Alas, it turns out that this calculation can be very sensitive and produce problematic results when the detector and context

strongly disagree. We therefore identify these cases first, then calculate $P(X_i|\mathcal{Y})$ accordingly.

3.1 Calculation of object probability

We assume that location and detection variables are connected according to the graph structure shown in Figure 2. More formally, we assume that detection Y_i directly depends only on the existence of an object at X_i , so

$$P(Y_i, \mathcal{Y}_{j \neq i} | X_i) = P(Y_i | X_i) P(\mathcal{Y}_{j \neq i} | X_i) . \quad (1)$$

We further assume that X_i directly depends only on its detection Y_i and on a small set \mathcal{N}_i of location variables. Therefore

$$P(\mathcal{Y}_{j \neq i} | X_i, \mathcal{N}_i) = P(\mathcal{Y}_{j \neq i} | \mathcal{N}_i) . \quad (2)$$

We note that variables in the set \mathcal{N}_i may or may not directly depend on each other.

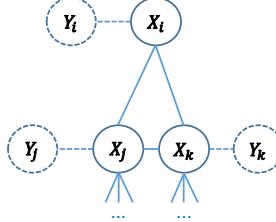


Fig. 2. Assumed graph structure when calculating $P(X_i|\mathcal{Y})$ in an iteration in which $\mathcal{N}_i = \{X_j, X_k\}$ is the set of most relevant neighbors for the location variable X_i . Each detection variable from \mathcal{Y} is associated with a single location variable, the query X_i and the variables in \mathcal{N}_i form a clique, and the rest of the variables are connected to X_i only via \mathcal{N}_i . We note that different graph structures are used for each X_i in each iteration according to current beliefs.

Employing these assumptions and the set \mathcal{N}_i (identified as described in Sec. 3.4), we calculate $P(X_i|\mathcal{Y})$ in the following way:

$$P(X_i|\mathcal{Y}) = \sum_{\mathcal{N}_i} P(X_i, \mathcal{N}_i | \mathcal{Y})$$

Applying Bayes' rule we first obtain

$$P(X_i|\mathcal{Y}) = \sum_{\mathcal{N}_i} P(\mathcal{Y}|X_i, \mathcal{N}_i) P(X_i, \mathcal{N}_i) \frac{1}{P(\mathcal{Y})} .$$

Employing Eq. 1 entails

$$\begin{aligned} P(X_i|\mathcal{Y}) &= \sum_{\mathcal{N}_i} P(Y_i|X_i, \mathcal{N}_i) P(\mathcal{Y}_{j \neq i}|X_i, \mathcal{N}_i) P(X_i, \mathcal{N}_i) \frac{1}{P(\mathcal{Y})} \\ P(X_i|\mathcal{Y}) &= \sum_{\mathcal{N}_i} P(Y_i|X_i) P(\mathcal{Y}_{j \neq i}|X_i, \mathcal{N}_i) P(X_i, \mathcal{N}_i) \frac{1}{P(\mathcal{Y})}. \end{aligned}$$

Employing Eq. 2 provides

$$P(X_i|\mathcal{Y}) = \frac{1}{P(\mathcal{Y})} P(Y_i|X_i) \sum_{\mathcal{N}_i} P(\mathcal{Y}_{j \neq i}|\mathcal{N}_i) P(X_i, \mathcal{N}_i),$$

and applying Bayes' rule again provides

$$P(X_i|\mathcal{Y}) = \frac{P(Y_i|X_i)}{P(\mathcal{Y})} \sum_{\mathcal{N}_i} P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) \frac{P(\mathcal{Y}_{j \neq i})}{P(\mathcal{N}_i)} P(X_i, \mathcal{N}_i),$$

which results in

$$P(X_i|\mathcal{Y}) = \frac{P(\mathcal{Y}_{j \neq i})}{P(\mathcal{Y})} P(Y_i|X_i) \sum_{\mathcal{N}_i} P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) P(X_i|\mathcal{N}_i), \quad (3)$$

an expression reminiscent of the belief propagation process suggested by Pearl [25]. This is further developed by applying Bayes' rule once more:

$$P(X_i|\mathcal{Y}) = \frac{P(Y_i) P(\mathcal{Y}_{j \neq i})}{P(\mathcal{Y})} P(X_i|Y_i) \sum_{\mathcal{N}_i} P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) \frac{P(X_i|\mathcal{N}_i)}{P(X_i)}.$$

Finally, we denote the first term with ξ , which is a normalizing constant and need not be explicitly calculated. We thus obtain:

$$P(X_i|\mathcal{Y}) = \xi \cdot P(X_i|Y_i) \sum_{\mathcal{N}_i} P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) \frac{P(X_i|\mathcal{N}_i)}{P(X_i)} \quad (4)$$

an expression we assert is more informative than Eq. 3, as it is now apparent that the way in which the context \mathcal{N}_i affects X_i is relative to its prior probability $P(X_i)$. We note that instead of calculating ξ , we normalize the values of $P(X_i|\mathcal{Y})$ calculated with $\xi = 1$, so that their sum equals to 1 as in a standard belief propagation process [25].

As can be seen, apart from $P(\mathcal{N}_i|\mathcal{Y}_{j \neq i})$ this expression contains only functions over a small number of variables (for small sizes of \mathcal{N}_i). We therefore restrict the size of \mathcal{N}_i to the maximum that still enables to properly represent and learn these functions, as employing more neighbors would require more memory and more training examples.

We are left with the term $P(\mathcal{N}_i|\mathcal{Y}_{j \neq i})$, which is more complicated to calculate. This term is the (joint) belief of variables in \mathcal{N}_i given all detections but Y_i , and can be seen as a weighting factor to the extent we are confident about

assignments for \mathcal{N}_i . We therefore suggest to approximate it as the product of individual beliefs of variables in \mathcal{N}_i :

$$P(\mathcal{N}_i | \mathcal{Y}_{j \neq i}) = \prod_{X_j \in \mathcal{N}_i} P(X_j | \mathcal{Y}_{j \neq i}) . \quad (5)$$

With this approximation, the representation of $P(X_i | \mathcal{Y})$ based on Eqs. 4 and 5 now consists of simple functions (that are easily measured from data), and terms of the form $P(X_j | \mathcal{Y}_{j \neq i})$, which can be seen as the messages in a standard belief propagation process.

Finally, given detections \mathcal{Y} provided by a base detector applied to an image, a new confidence is calculated for each detection X_i . Its most informative neighbors \mathcal{N}_i are identified as explained in Sec. 3.4, and used for the calculation of the probability $P(X_i | \mathcal{Y})$ in a belief propagation process. The confidence assigned to X_i is then $P(X_i = \text{True} | \mathcal{Y})$.

3.2 Scale invariant representation

The term $P(X_i | \mathcal{N}_i)$ in Eq. 4 represents relations between several locations. Such a term must be calculated for each set of locations, requiring to observe many examples of object groups in each location. To reduce this complexity we make the (very reasonable) assumption that relations between objects are independent of the viewer, and suggest a representation that is invariant to different object scales.

Similar to the spatial features employed by Cinbis and Sclaroff [4], we represent the spatial relation $P(X_i | \mathcal{N}_i)$ with respect to the size of a reference object $X_j \in \mathcal{N}_i$. The relative location of X_i (and any other non-reference object $X_k \in \mathcal{N}_i$) is represented as

$$\frac{l_i - l_j}{s_j f_j} , \quad (6)$$

where l_i, l_j are the locations of the center points of X_i and X_j , s_j is the height of X_j , and $f(t_j)$ is a scaling factor for s_j according to t_j , the type of X_j assigned by the base detector. Object scales are also represented relative to the reference X_j

$$\log\left(\frac{s_i}{s_j}\right) . \quad (7)$$

Using this representation, the probability $P(X_i | \mathcal{N}_i \setminus X_j, X_j = \text{True})$ is measured for any assignment to X_i and to the variables in \mathcal{N}_i except X_j , which contains an object used as a reference frame. Specifically, we count the occurrences of objects of each type in each location and scale relative to reference objects of type t_j that appear in training data.

A non-parametric representation of $P(X_i | \mathcal{N}_i)$ requires a value for every possible assignment of the variables. The described method for measuring $P(X_i | \mathcal{N}_i)$ requires at least one variable $X_j \in \mathcal{N}_i$ that is not empty (*i.e.*, contains an object) with which to construct a reference frame. However, in some assignments

there may be no such X_j . Hence, for assignments in which all the variables in \mathcal{N}_i are empty while X_i is not, we simply use X_i as reference:

$$P(X_i|\mathcal{N}_i) = P(X_k|X_i, \mathcal{N}_i \setminus X_k) \frac{P(X_i|\mathcal{N}_i \setminus X_k)}{P(X_k|\mathcal{N}_i \setminus X_k)} , \quad (8)$$

where X_k is arbitrarily picked from \mathcal{N}_i . Notice how the terms of the quotient operate on one less variable.

Finally, the probability for the assignment in which all the variables are empty is calculated by subtracting the probability of the complementary event from one.

3.3 The implications of high probability derivative

To better understand the way contextual information is combined with the detector response we revisit Eq. 4 and examine its behavior when the context is known. Thus, for some assignment to the members of \mathcal{N}_i we assume that $P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) = 1$, and so Eq. 4 reduces to:

$$P(X_i|\mathcal{Y}) = \xi \cdot P(X_i|Y_i) \frac{P(X_i|\mathcal{N}_i)}{P(X_i)} . \quad (9)$$

A graph of Eq. 9 for different detector responses is presented in Figure 3. As can be seen, the addition of context strengthens a detection when the context-based probability is bigger than the prior, and weakens it when the opposite occurs. The red and blue curves, representing especially strong and weak detections respectively, *exhibit large derivative regions where the detector and context strongly disagree*. It may also be the case when the detector is confident and the context is independent, *i.e.*, $P(X_i|\mathcal{N}_i) = P(X_i) = 0.02$ in the case of Fig. 3. In these cases, the overall probability greatly changes with small perturbations of the context-based probability $P(X_i|\mathcal{N}_i)$. Because this quantity is measured from data, great errors are to be expected when the number of samples does not suffice.

Owing to this, many cases are indeed observed where detections are incorrectly assigned with low probabilities despite a confident detector and when the context is seemingly independent. Failing to address such cases leads to poor results, as we show in Section 4, and a specific case can be seen in Figure 4. Hence, to handle such cases, for each assignment to the members of \mathcal{N}_i we calculate the derivative of $P(X_i|\mathcal{Y})$ at $x = P(X_i|\mathcal{N}_i)$ and if it exceeds a threshold we ignore the context by setting $P(X_i|\mathcal{N}_i) = P(X_i)$, essentially assuming that X_i and \mathcal{N}_i are independent, and thus, that the context has no effect.

Another way to identify these cases is to estimate the number of samples needed to ensure a low error. Let $p^* = P(X_i|\mathcal{Y})$ and $h^* = P(X_i|\mathcal{N}_i)$, as depicted in Figure 5. If we allow a maximal error of ϵ , we require that:

$$|p - p^*| < \epsilon ,$$

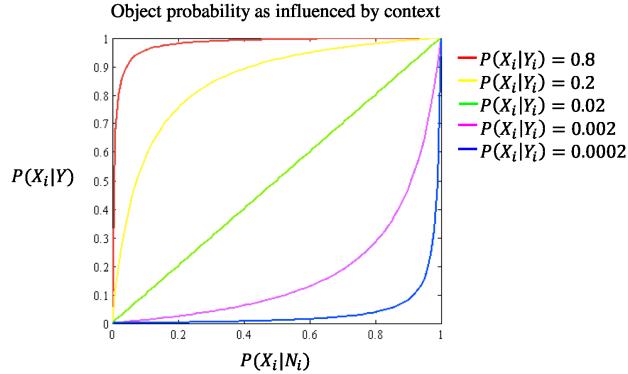


Fig. 3. The overall object probability $P(X_i|\mathcal{Y})$ as a function of the context-based probability $P(X_i|\mathcal{N}_i)$ for some variable assignment. Each curve represents a different value of local detector response $P(X_i|Y_i)$, for a relatively large prior probability of $P(X_i) = 0.02$. This visualizes the way in which both sources of information are combined for a final decision. See the text for further analysis.

where p is the value calculated using measured data. For h , the measured value of $P(X_i|\mathcal{N}_i)$, to provide an error that does not exceed ϵ , it is required to stay within the limits of h_1 and h_2 , that are the values of $P(X_i|\mathcal{N}_i)$ that correspond to $P(X_i|\mathcal{Y}) = p^* - \epsilon$ and $P(X_i|\mathcal{Y}) = p^* + \epsilon$ respectively. Thus:

$$|h - h^*| < \min(|h^* - h_1|, |h^* - h_2|) = \epsilon_h ,$$

where ϵ_h is the allowed measurement error for h .

We then employ Hoeffding's inequality [29] to estimate m , the number of samples Z_i required for h :

$$h = \frac{1}{m} \sum_{i=1}^m Z_i .$$



Fig. 4. A recurring case in which a confident detection (in the left) is assigned with a low probability. The probability of the left detection is calculated using the detector confidence and the detection in the right. While it is reasonable that the context affects the detection, it should not do so to the extent of nullifying a confident detection.

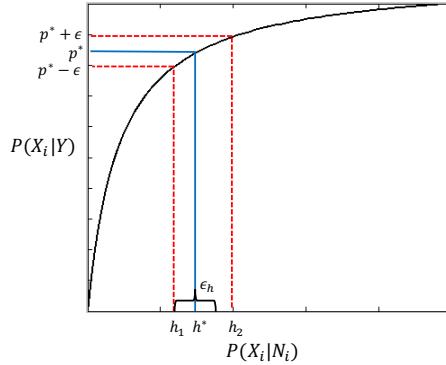


Fig. 5. The overall probability $P(X_i|\mathcal{Y})$ and its error in relation to h^* , the value of $P(X_i|\mathcal{N}_i)$ we measure from data. See the text for details.

We assume that the indicator variables Z_i are sampled *i.i.d* and note that their expectation is equal to h^* . Therefore, the probability of measuring h with large overall error can be expressed as:

$$P(|p - p^*| > \epsilon) < P(|h - h^*| > \epsilon_h) < 2e^{-2m\epsilon_h^2} = \delta .$$

Finally, to guarantee a maximal error of ϵ with probability δ , the number of required samples m is:

$$m > \frac{\ln(\frac{2}{\delta})}{2\epsilon_h^2} . \quad (10)$$

This expression enables to calculate the needed number of samples for $P(X_i|\mathcal{N}_i)$ in each case and then to employ it only when enough data is provided (even when the derivative is high). In our experiments the derivative was used to identify problematic cases, while m is used for discussion in Section 4.

3.4 Identification of relevant detections

The described algorithm requires to identify the most relevant set of locations $\mathcal{N}_i \in \mathcal{S}$ to use as context for X_i , where \mathcal{S} is the set of subsets of $\mathcal{X}_{j \neq i}$. To limit the size of \mathcal{N}_i , we consider only subsets for which the cardinality is equal or smaller than a predefined number that is given as a parameter of the model.

The set \mathcal{N}_i was used for the assumption that X_i does not depend on other locations (or their detections) given \mathcal{N}_i :

$$P(X_i|\mathcal{N}_i, \mathcal{Y}_{j \neq i}) = P(X_i|\mathcal{N}_i) . \quad (11)$$

Thus, the most suitable \mathcal{N}_i would be

$$\arg \min_{\mathcal{N}_i \in \mathcal{S}} \sum_{X_i, \mathcal{N}_i} |P(X_i|\mathcal{N}_i, \mathcal{Y}_{j \neq i}) - P(X_i|\mathcal{N}_i)| . \quad (12)$$

However, this calculation requires a function over many variables $\mathcal{Y}_{j \neq i}$, which seems as complicated as calculating $P(X_i|\mathcal{Y})$. We hence suggest a different way to determine \mathcal{N}_i . Basically, we would like to employ the set \mathcal{N}_i that would be the best predictor for X_i using our current beliefs for the different variables. For this reason, we pick \mathcal{N}_i as those variables that are the least independent of X_i and weigh them according to the current beliefs

$$\arg \max_{\mathcal{N}_i \in \mathcal{S}} \sum_{X_i, \mathcal{N}_i} |P(X_i|\mathcal{N}_i) - P(X_i)| \cdot P(\mathcal{N}_i|\mathcal{Y}_{j \neq i}) , \quad (13)$$

where $P(\mathcal{N}_i|\mathcal{Y}_{j \neq i})$ is calculated as in Eq. 5.

4 Results and Discussion

We evaluate the proposed approach using the PASCAL VOC 2007 dataset where initial detections are provided by the Fast R-CNN detector [10]. Training and validation set objects are used to measure the probability of an object given its context $P(X_i|\mathcal{N}_i)$ as described in Sec. 3.2. For the probability of an object given its detection $P(X_i|Y_i)$ we simply use the confidence c_i provided by the base detector. To determine $P(X_i)$, we assume that the prior probability of an object is fixed regardless of image location and size, but depends on the object type. The value of $P(X_i)$ for each type is found by an exhaustive search to maximize the method's average precision (AP) on the training and validation set.

The results of our approach are summarized in Table 1, and specific examples including the detections identified as most informative can be seen in Figure 6. Included in the table are the base Fast R-CNN detector, our proposed model with two most informative detections and *no* treatment for large derivatives (dFNM), our model with randomly selected context (rFNM), and finally, our contextual inference model *with* treatment for large derivatives and two most informative detections (FNM). We note that it is also possible for a single detection to be identified as the most informative, and that employing two detections to reason about a third constitutes a triple-wise model. In all variants of our computational model the results are reported after a single iteration of belief propagation, as usually just two or three iterations were needed for convergence.

As can be seen, our suggested approach (FNM) improves detection results and greater improvement is observed for the detection of chairs, sheep, televisions and bottles. The model in which large derivative regions are not handled (dFNM) is worse than the detector alone, which fits the analysis performed in Sec 3.3.

For additional comparison, we test the ability of our model to improve detection results of the Faster R-CNN detector [28] over the COCO dataset [18]. In this case the suggested model (FNM) improved the mAP from 66.5 to 66.9 where the largest improvement of 2.6 was observed for sheep. We also compare our results to the ION contextual detector [2] that employs the same base detector and dataset (PASCAL 2007) and reports the results provided by the added context layers (but without additional components that are unrelated to context). In a nutshell, both ION and our approach provide a comparable improvement, where

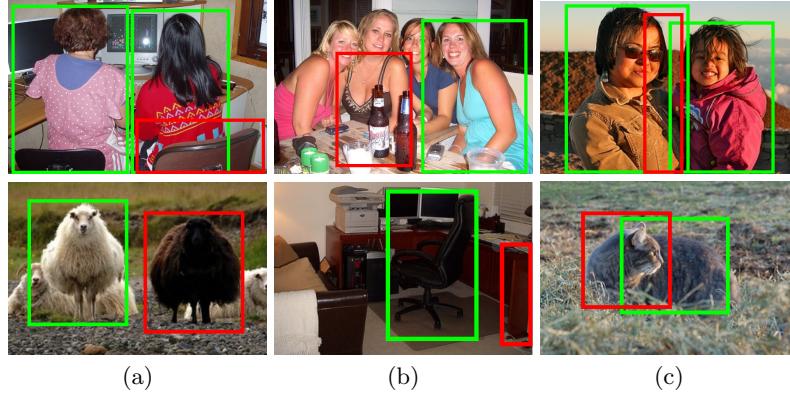


Fig. 6. Example detections (red) with their most informative neighbors (green). The confidence of correct chair/sheep detections in (a) increased following the use of our context model, and the confidence of the incorrect bottle (top) and tv/monitor (bottom) detections in (b) decreased. The hard to detect chair/sheep in (a) gain a boost in confidence due to the favorable context in which they appear. Similarly, The false detections in (b) are weakened due to their unusual location and size in relation to the other objects that serve as their context. Detections in (c) represent recurring mistakes due to localization errors, in which an increased confidence was provided for a false person detection (top) and a false cat detection (bottom). In both cases, object parts that were mistaken for full fledged objects are strengthened due to the use of context.

the mAP of the base detector was increased by 0.98 and 0.52, respectively, and our model provided a larger improvement than ION on 8 out of 20 object categories using a *significantly simpler and quicker* model. Moreover, please recall that our computational approach can also be applied to the results provided by networks with context such as ION for an additional boost.

4.1 Impracticality of learning certain properties

The graph of Eq. 9 in Fig. 3 sheds light on the way information supplied by the local detector is combined with the context. As can be seen, the contextual information can either increase or decrease the probability of an object, where different probabilities of the local detector response affect the rate of change.

Using Eq. 10, we examine the difficulty of learning different properties with regard to the number of samples needed to stay within the limits of an allowed error. In this section we show the impracticality of learning certain properties even under modest error requirements. We first examine the requirements for learning relations that decrease the probability of objects without necessarily seeking to improve detections. As a test case, we set $P(X_i) = 0.02$, which is a relatively high prior probability for location X_i to contain an object. So, to decide with high certainty ($\delta = 0.1$) whether an assignment to the members of \mathcal{N}_i reduces the probability of X_i , that is $P(X_i|\mathcal{N}_i) < P(X_i)$, an ϵ_h error of at

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
FRCNN	72.15	72.51	61.19	42.98	28.67	69.89	70.67	81.14	36.15	69.61	55.90	77.31	76.69	69.01	64.55	28.24	57.35	59.97	75.58	62.58	61.61
dFNM	72.08	69.82	61.93	41.79	25.12	67.73	69.73	80.84	39.00	69.22	54.30	74.91	76.38	66.41	62.83	28.20	59.33	57.46	74.94	59.30	60.57
rFNM	72.11	72.57	61.21	43.20	28.56	69.69	70.73	81.38	33.74	69.63	55.47	77.29	76.77	68.86	64.56	28.46	57.44	60.13	76.10	62.71	61.53
FNM	72.29	72.63	61.68	43.15	29.64	69.81	70.76	81.36	39.00	69.73	56.01	77.63	76.78	68.93	64.85	28.90	59.46	60.09	76.26	63.58	62.13

Table 1. Detection results in average precision (AP) of the base detector Fast R-CNN (FRCNN) and different variants of our model on all objects in the PASCAL VOC 2007 test set. Employing our model without treatment for large derivatives (dFNM) hurts detection results, and when large derivatives are treated (FNM) our model provides improved detection results. This improvement is not observed with randomly selected context (rFNM). The most significant increase is obtained for object categories such as chairs, sheep, televisions, and bottles, which benefit more from the use of context.

most 0.02 is required. According to Eq. 10, this requires at least 3745 samples of that relation (or even more for the average prior probability).

Seeking to improve detection results, let us examine one test case in which the context-based probability $P(X_i|\mathcal{N}_i)$ is half of the prior probability $P(X_i)$:

$$P(X_i|\mathcal{N}_i) = \frac{P(X_i)}{2} = 0.01 .$$

For confident detections $P(X_i|Y_i) = 0.8$, the overall probability of X_i to contain an object in this case is $P(X_i|\mathcal{Y}) = 0.6644$. If we require a modest overall error of at most $\epsilon = 0.1$, then according to the construction in Section 3.3, a measurement error less than $\epsilon_h = 0.0034$ is required. In this case, for a high certainty ($\delta = 0.1$) the number of required samples is 127,095. Similarly, for more accurate results ($\epsilon = 0.05$), as much as 400,048 samples are required. Of course, in more extreme cases many more samples would be required. The need to collect datasets that large renders such relations impractical to learn by observing object occurrences.

The suggestion here was to handle the cases when the detector and context strongly disagree by assuming independence from context $P(X_i|\mathcal{N}_i) = P(X_i)$ according to the derivative. This decision to ignore the context (instead of the detector) corresponds to similar processes we observe in the human visual system, as exemplified in Figure 7. And yet, it is important to note that there are indeed cases in which our contextual computation successfully reduces the probability of an object. Two such cases can be seen in Figure 6b.

Also important is the rate of change of the graph in Fig. 3 for different detector responses $P(X_i|Y_i)$ depicted by the differently colored curves. The central curve, calculated for $P(X_i|Y_i) = 0.02$, behaves as a straight line. The red and yellow curves behave similarly to the blue and magenta ones. However, the scale of the latter curves is significantly smaller, reducing the high derivative cases to those in which the detector is extremely confident that an object is not present. From this we conclude that it is generally simpler to learn relations that increase the probability of an object in comparison to those that decrease its probability.

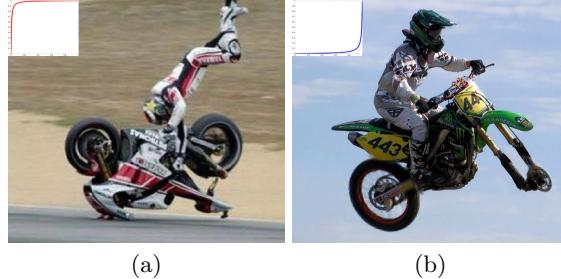


Fig. 7. Examples of disagreement between the context and the detector (based on local appearance) where in both cases the human visual system seems to resolve this conflict by ignoring the context. The insets show the corresponding situation of the conditional probability as discussed in Fig. 3. **(a)** Most observers do not have a problem to immediately spot the rider despite its improbable configuration relative to the motorbike. **(b)** When looking directly at the expected position of the front wheel, most observers report its absence even though motorbikes are mostly observed to have two wheels.

5 Conclusions

The problem of including context in object detection is important but difficult, as decision over many locations is needed. We have suggested to employ only a small number of locations for a more accurate decision, and presented a method for the identification of the most informative set of locations, and a formulation that employs it to infer the probability of objects at different locations in a probabilistic fashion. Key benefits of our computational approach is how it facilitates better understanding of certain aspects of the problem, and in particular it allowed to conclude that it is impractical to infer relations that decrease or increase the probability of detections when the detector and the context strongly disagree, or that in general it is more difficult to infer conditions that reduce the probability of an object rather than relations that increase it. Finally, we have demonstrated how our approach improves detection results using a model that is quick and simple to train and to employ for context-based inference.

Acknowledgments

This research was supported in part by Israel Ministry of Science, Technology and Space (MOST Grant 54178). We also thank the Frankel Fund and the Helmsley Charitable Trust through the ABC Robotics Initiative, both at Ben-Gurion University of the Negev, for their generous support.

References

1. Arbel, N., Avraham, T., Lindenbaum, M.: Inner-scene similarities as a contextual cue for object detection. arXiv preprint (2017)

2. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: CVPR. pp. 2874–2883 (2016)
3. Chechetka, A., Guestrin, C.: Evidence-specific structures for rich tractable crfs. In: Advances in Neural Information Processing Systems. pp. 352–360 (2010)
4. Cinbis, R.G., Sclaroff, S.: Contextual object detection using set-based classification. In: ECCV. pp. 43–57. Springer (2012)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. pp. 886–893 (2005)
6. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. Int. J. Comput. Vision **95**(1), 1–12 (2011)
7. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multi-scale, deformable part model. In: CVPR. pp. 1–8 (2008)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
9. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR. pp. 1–8 (2008)
10. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
12. Heitz, G., Koller, D.: Learning spatial context: Using stuff to find things. In: ECCV. pp. 30–43 (2008)
13. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. Int. J. Comput. Vision **80**(1), 3–15 (2008)
14. Kohli, P., Rother, C.: Higher-order models in computer vision. In: Image Processing and Analysing with Graphs: Theory and Practice, chap. 3, pp. 65–92. CRC Press (2012)
15. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
16. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. IEEE Trans. Pattern Anal. Mach. Intell. **16**(11), 2649–2661 (2007)
17. Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J., Yan, S.: Attentive contexts for object detection. IEEE Transactions on Multimedia **19**(5), 944–954 (2017)
18. 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: ECCV. pp. 740–755. Springer (2014)
19. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: NIPS. pp. 4898–4906 (2016)
20. Mairon, R., Ben-Shahar, O.: A closer look at context: From coxels to the contextual emergence of object saliency. In: ECCV. pp. 708–724. Springer (2014)
21. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR. pp. 891–898 (2014)
22. Oramas, J., Tuytelaars, T.: Recovering hard-to-find object instances by sampling context-based object proposals. CVIU (2016)
23. Oramas M, J., De Raedt, L., Tuytelaars, T.: Allocentric pose estimation. In: ICCV (2013)
24. Oramas M, J., De Raedt, L., Tuytelaars, T.: Towards cautious collective inference for object verification. In: WACV (2014)

25. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
26. Perko, R., Leonardis, A.: A framework for visual-context-aware object detection in still images. *CVIU* **114**(6), 700–711 (2010)
27. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV. pp. 1–8. IEEE (2007)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
29. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge University Press (2014)
30. Torralba, A., Murphy, K.P., Freeman, W.T.: Contextual models for object detection using boosted random fields. In: NIPS. pp. 1401–1408 (2004)
31. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: ICCV. vol. 1, pp. 763–770. IEEE (2001)
32. Wolf, L., Bileschi, S.: A critical view of context. *Int. J. Comput. Vision* **69**(2), 251–261 (2006)
33. Yu, R., Chen, X., Morariu, V.I., Davis, L.S.: The role of context selection in object detection. British Machine Vision Conference **abs/1609.02948** (2016)