Neurocomputing 000 (2017) 1–8

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep feature based contextual model for object detection

Wenqing Chu, Deng Cai*

The State Key Lab of CAD & CG, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 23 March 2017
Accepted 17 September 2017
Available online xxx

Communicated by Steven Hoi

Keywords:
Object detection
Context information
Conditional random field

ABSTRACT

One of the most active areas in computer vision is object detection, which has made significant improvement in recent years. Current state-of-the-art object detection methods mostly adhere to the framework of the regions with convolutional neural network (R-CNN). However, they only take advantage of the local appearance features inside object bounding boxes. Since these approaches ignore the contextual information around the object proposals, the outcome of these detectors may generate a semantically incoherent interpretation of the input image. In this paper, we propose a novel object detection system which incorporates the local appearance and the contextual information. Specifically, the contextual information comprises the relationships among objects and the global scene based contextual feature generated by a convolutional neural network. The whole system is formulated as a fully connected conditional random field (CRF) defined on object proposals. Then the contextual constraints among object proposals are modeled as edges naturally. Furthermore, a fast mean field approximation method is utilized to infer in this CRF model efficiently. The experimental results demonstrate that our algorithm achieves a higher mean average precision (mAP) on PASCAL VOC 2007 datasets compared with the baseline algorithm Faster R-CNN.

1. Introduction

Object detection is one of the fundamental problems in computer vision. It plays an important role in many real-world applications such as image retrieval, advanced driver assistance system and video surveillance. This problem is very difficult due to the huge intra-class variations of semantic objects of a certain class in digital images. Their appearances vary dramatically from changes in different illuminations, view points, nonrigid deformations, poses, and the presence of occlusions. For instance, there is a large amount of partial occlusions between pedestrians standing next to each other in a crowded street which leads to an enormous challenge for detecting all the persons.

In the past few years, remarkable progress has been made to boost the performance on the object detection problem. A common pipeline to address this problem consists of two main steps: (1) object proposal generation, (2) class-specific scoring and bounding box regression. There is a significant body of methods for generating object proposals such as [1–6] or just a sliding window fashion [7]. Then some specific feature of the object bounding box is extracted and some classifier is applied for efficient and accurate object classification, in which the representative methods include AdaBoost algorithm [8], DPM models [7] and deep CNN models [9]. However, most state-of-the-art detectors like Faster R-CNN [10] only consider the object proposals individually without taking the contextual information around the object bounding boxes into account.

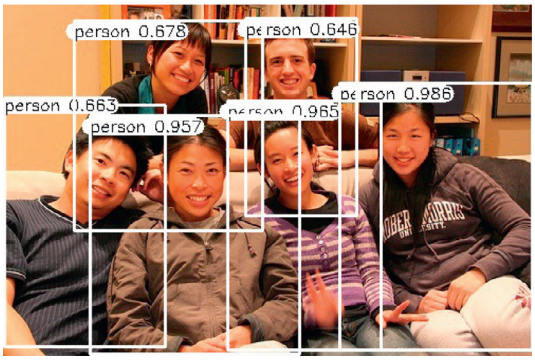
In the real world, there exists a semantic coherent relationship between the objects in terms of the relative spatial location and the co-occurrence probability [11,12]. In some situations, contextual information among objects in the input image can provide a more valuable cue for dealing with the detection than the local appearance evidence of the object candidates. In addition, the global contextual information based on scene understanding can also assist the detector to rule out some false alarms.

To be more clear, we show some results of the dominant Faster R-CNN [10]. In Fig. 1(a), the Faster R-CNN employs the CNN model to classify the two object proposals individually, whereas boats and trains stand little chance of co-occurrence in the input image. That common sense means we could decrease the probability that the object proposal is a boat. In Fig. 1(b), since one of the persons is occluded by the sofa and thus the Faster R-CNN recognizes the object proposal as a person with a low confidence score of 0.646. However, if we can exploit the contextual information that there seems to be a lot of person object candidates around the occluded bounding box, we can raise our confidence that the category of this object candidate behind the sofa is a person. In addition,

* Corresponding author.
E-mail addresses: wqchu16@gmail.com (W. Chu), dengcai@cad.zju.edu.cn (D. Cai).

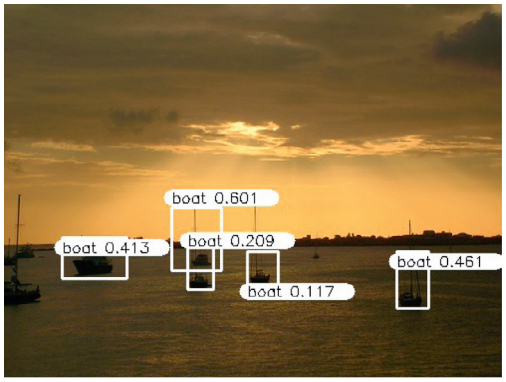
https://doi.org/10.1016/j.neucom.2017.09.048
0925-2312/

(a) Boat and Train

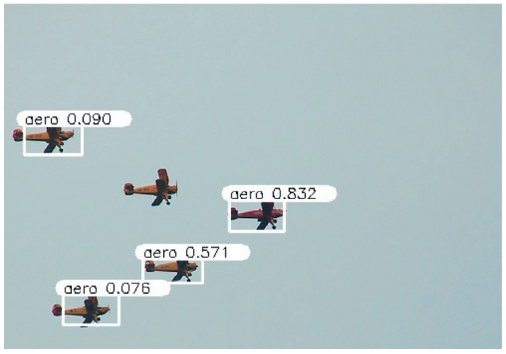


(b) Partial Occlusions between persons

Fig. 1. Object-level contextual information.



(a) Boats in Lake



(b) Aeroplanes in air

Fig. 2. Image-level contextual information.

the image-level contextual information can also support object detection. If we can recognize the scene in Fig. 2(a) as a lake or sea, then we can make a better judgment that the object proposals are likely to be tiny boats. Similarly, the spectacle in Fig. 2(b) looks like sky which enhances the probability of the presence of airplanes. In a word, the contextual information implied by the input image can be a strong clue for recognizing the object candidates which are ambiguous because of low resolution and variation in pose and illumination.

In the history of computer vision, a number of approaches [13–22] have exploited contextual information in order to boost the performance on the object detection problem. Nevertheless, most of these methods leverage hand-crafted features such as Gabor [23], Gist [13] or HOG [24–26] to represent the input image. Recently, the convolutional neural networks (CNN) have achieved great success in computer vision tasks such as image classification [27], which inspired us to employ the powerful CNN model to devise a novel contextual model.

In this paper, we propose a novel object detection framework which can exploit both the object-level and image-level contextual information. At first, it is obvious that the objects present in the same image conform to a semantic coherent relationship in terms of the relative spatial location and the co-occurrence probability. As in [17], inter-object constraints vary greatly from changes in different categories and locations, which can be learned from the statistical summary of the datasets. In addition, we also generate the global scene descriptor using a CNN model which is trained for the scene understanding task. Then the global scene information of the input image is fed into a logistic regression model. And the model will predict the probability that how much the instances of a certain category are likely to occur in the image. To leverage the local appearance information, we apply the popular Faster R-CNN [10] method to each input image and obtain a pool of object proposals with the corresponding scores and locations for each category. After that, we take these object proposals as nodes and then formulate them into a fully-connected conditional random field (CRF) according to the contextual information. Specifically, the unary potentials are determined by the category scores from the Faster R-CNN. And the pairwise potentials are decided according to the relative layouts and categories of the object proposals. To efficiently inference in this CRF model, we utilize a fast mean field inference algorithm from [28,29] to yield semantically coherent detection results for the input image.

We have extensively evaluated our method on the PASCAL VOC2007 dataset [30]. The experimental results show that our method can achieve better performance than the baseline Faster R-CNN. And for some small objects like the bottles the average precisions can obtain a great improvement. Accuracies of very few categories are worsened by contextual information.

The rest of this paper is arranged as follows. First, we briefly review a few of recent work on object detection in Section 2. Then in Section 3 we describe the framework of our contextual model for object detection and the inference algorithm in detail. After

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048

that, we evaluate the performance of our method on the challenging databases PASCAL 2007 in Section 4. Finally, in Section 5, we present our conclusions and discuss the future work.

2. Related work

In this section, we briefly review the recent work on object detection. Object detection has been active research areas in recent years, which has led to a large amount of approaches to address the problems in it.

In the literature of object detection, the part-based model is one of the most powerful approaches in which deformable part-based model (DPM) [7] is an excellent example. This method takes the histogram of oriented gradients (HOG) features [24] as input and detects objects in an image by a sliding window approach. Specifically, it utilizes mixtures of multi-scale deformable part models to represent highly variable objects. Each part captures local appearance properties of an object while the deformable configuration is characterized by the spring-like connections between certain pairs of parts.

Recently, deep convolutional neural networks (CNN) have emerged as a powerful machine learning model on a number of image recognition benchmarks, including the most noticeably work by [27]. That aroused a significant body of methods [9,10,31–36] addressing this problem with the CNN model. Among these approaches, the regions-with-convolutional-neural-network (R-CNN) framework [9] has achieved excellent detection performance and became a commonly employed paradigm for object detection. Its essential steps include object proposal generation with selective search [2], CNN feature extraction, object candidates classification and regression based on the CNN feature. However, R-CNN brings excessive computation cost because it extracts CNN feature repeatedly for thousands of object proposals. Spatial pyramid pooling networks (SPPnets) [33] were proposed to accelerate the process of feature extraction in R-CNN by sharing the forward pass computation. The SPPnet approach computes a convolutional feature map for the entire input image once and then generates a fixed-length feature vector from the shared feature map for each object proposal. Fast Region-based Convolutional Network method (Fast R-CNN) [34] utilizes a multi-task loss, which leads to an end-to-end framework where the training is a single-stage and no disk storage is required for feature caching. The drawback of Fast R-CNN is that this method still use bottom-up proposal generation which is the bottleneck of efficiency. Instead, the authors proposed a Region Proposal Network (RPN) method [10] that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. These techniques, however, still mostly perform detection based on only local appearance features of the object proposals.

In the other hand, semantic contextual information also plays a very important role in the history of object detection methods [13,15,17,19–21]. The statistics of low-level features across the entire scene were used to predict the presence or absence of objects and their locations [13]. In [15], the authors demonstrated that contextual relations between objects' labels can help reduce ambiguity in objects' visual appearance. Specifically, they utilized image segmentation as a pre-processing step to generate object proposals. Then a conditional random field (CRF) formulation was exploited as post-processing to infer the optimal label configuration of this CRF model, which jointly labels all the object candidates. [17] extend this approach to combine two types of context co-occurrence and relative location with local appearance based features. And [20] introduced a unified model for multi-class object recognition that learns statistics that capture the spatial arrangements of various object classes in real images.

Besides, some work [37–39] which focus on detecting some specialized object class were proposed to use contextual information to support detection. However, these models mostly work on contextual information represented by traditional visual features such as HOG or GIST. Thus, we are motivated to move on to more powerful features provided by the deep CNN model.

3. A fully-connected CRF for object detection

In this section, we address the general object detection problem with a novel object detection system which considers both the local appearance and the contextual information. And the contextual information comprises the coherent constraints among objects and the global scene information. To process each input image, our approach includes three main stages. At first, we generate a pool of object proposals with the Faster R-CNN [10] method. Then we employ a conditional random field (CRF) framework to model the object detection problem with contextual information. Finally, we utilize an efficient mean field approximation method to inference and finally maximize object label agreement.

In Section 3.1, we introduce the process of the object proposal generation based on the Faster R-CNN. In Section 3.2, we give the formulation of the CRF model for the object detection problem. Then we will describe the unary potentials obtained from the results of Faster R-CNN, the pairwise potentials between object candidates and the global potentials determined by the global scene feature describing the entire image. After all, Section 3.3 will give the inference algorithm in detail.

3.1. Object proposals

Our approach will generate object proposals following Faster R-CNN [10], which is one of the state-of-the-art object detection methods. In contrast to R-CNN [9], the Faster R-CNN utilizes a Region Proposal Network (RPN) instead of other bottom-up approaches to output a set of object bounding boxes. Since RPN slides a small network over the last convolution feature map, which makes it can share forward pass computation with a Fast R-CNN object detection network [34]. And that leads to great advantages on efficiency.

The Faster R-CNN method can depend on different CNN architectures such as the Zeiler and Fergus model [40] (ZF), which has 5 shareable convolution layers, and the Simonyan and Zisserman model [41] (VGG), which has 13 shareable convolution layers. To verify our method is insensitive to different object proposal methods, we conduct experiments on the Faster R-CNN based on both the ZF and VGG CNN architectures. To train the network, we optimize parameters with the popular stochastic gradient descent (SGD) [42] with momentum. The Faster R-CNN model is pre-trained on the ImageNet dataset [43] and finetuned on the PASCAL VOC 2007 dataset [30]. More details on the training procedure can be found in [10].

3.2. CRF formulation

In this section, we describe how to employ a CRF framework to model the object detection problem. Given an image I, we first apply the Faster R-CNN to it and obtain n object proposals, denoted by $\mathbf{X} = \{X_1, \ldots, X_n\}$. Suppose there are K categories in the image dataset, we introduce a label variable y_i for each object proposal X_i and the domain of the variable y_i is a set of categories $L = \{0, 1, 2, \ldots, K\}$ in which 0 represents the background class. Taking contextual information into account, we model the joint

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048

distribution over the label variables $\mathbf{y} = \{y_1, \ldots, y_n\}$ as

$$\begin{cases} P(\mathbf{y}|\mathbf{X}, I) = \dfrac{Unary(\mathbf{y}) \times Global(\mathbf{y}) \times Pair(\mathbf{y})}{Z(X_1, \ldots, X_n, I)}, \\ Unary(\mathbf{y}) = \prod_{i=1}^{N} P(y_i|X_i), \\ Pair(\mathbf{y}) = \prod_{i,j=1}^{N} P(y_i|y_j, X_i, X_j, I)), \\ Global(\mathbf{y}) = \prod_{i=1}^{N} P(y_i|I) \end{cases} \tag{1}$$

where $Z(X_1, \ldots, X_n, I)$ is the partition function and $P(y_i|y_j, X_i, X_j, I)$ measures the probability that an object with label y_i appears when there is an object labeled with y_j in the same image. In addition, $P(y_i|I)$ represents the probability for an object with label y_i exists in image I.

More specifically, we view each object proposal as a node and consider a conditional random field $\mathbf{y}$ defined over the label variables $\{y_1, \ldots, y_n\}$ which is characterized by a Gibbs distribution as

$$P(\mathbf{y}|\mathbf{X}, I) = \frac{1}{Z(\mathbf{X}, I)} exp(-\sum_{i=1}^{N} \phi_u(y_i) - \omega_g \sum_{i=1}^{N} \phi_g(y_i) - \omega_p \sum_{i,j=1}^{N} \phi_p(y_i, y_j)) \tag{2}$$

where $Z(\mathbf{X}, I)$ is the partition function, ω_p and ω_g are the trade-off parameters of the pairwise potentials and the global potentials learned by cross-validation. The $\phi_u(y_i)$ is the unary term computed independently for each object proposal based on the local appearance. The $\phi_g(y_i|I)$ measures the probability that an object with label y_i appears given the scene information of image I. And the $\phi_p(y_i, y_j)$ measures the probability that an object with label y_i appears when there is an object labeled with y_j in the same image. Inferring in this CRF model is to adjust the scores of each node according to the global scene information and the contextual information based constraints among the nodes.

3.2.1. Unary potentials

In our conditional random field (CRF) model, the unary potential $\phi_u(y_i)$ measures the probability that the object proposal X_i belongs to the category y_i according to the local appearance evidence. Suppose Faster R-CNN generates some object proposals and their corresponding scores matrix $P \in \mathbb{R}^{N \times K}$ which $P_{i,k}$ represents the probability of the object proposal i belongs to category k. Then we use a rescaled score based on the results of the Faster R-CNN as unary potential. This lets us write our unary term as

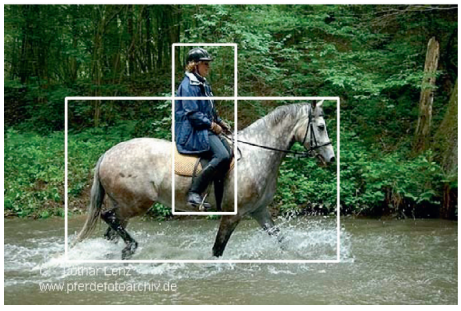
$$\phi_u(y_i) = -log(P(y_i|X_i)) \tag{3}$$

where $P(y_i|X_i)$ is the confidence that proposal X_i belongs to category y_i.

3.2.2. Pairwise potentials

Here we describe our pairwise potentials that purpose to capture the contextual information between multiple object candidates. Our pairwise model takes both the semantic and spatial relationships into account like [17,20]. In our method, we define a function $\phi_p()$ to estimate the pairwise potential of a proposal X_i based on a known object X_j in the same image. The input of $\phi_p()$ is the category information of the X_i and the X_j and their relative location. To leverage this spatial relevance better, we consider 11 different layouts for two object bounding boxes. If two candidates do not have intersection, then the spatial relationships of them can be classified into far, up, down, left and right. Otherwise, they can be defined by inside, outside, up, down, left and right. For example, some pottedplants are above another in Fig. 3(a) and do not intersect between each other. In Fig. 3(b), however, the human is on the horse and their bounding boxes have intersection. We define these situations as two different spatial relationships.

(a) Pottedplants



(b) Horse and Human

Fig. 3. Different spatial relationship.

Following [17,20], this function can be defined according to the probability $P(y_i, |y_j, X_i, X_j)$ which is learned from statistic summary of the training dataset.

$$\phi_p(y_i, y_j) = -log(P(y_i, |y_j, X_i, X_j, I)) \tag{4}$$

where $P(y_i, |y_j, X_i, X_j, I)$ measures the probability that an object with label y_i appears when there is an object labeled with y_j for a given relative location relationship between X_i and X_j.

3.2.3. Global potentials

In addition to object-level contextual information, we also introduce image-level signal to reason about the presence or the absence of some object in the image. The key point is to find global image features which can represent various scene well.

Scene categorization is also a challenge task in computer vision. Before, most work focuses on shallower hand-crafted features empirically and the databases that they used lack of abundance and variety. Recently, the Places2 dataset [44] is provided which contains more than 10 million images comprising 400+ unique scene categories. Moreover, the dataset features 5000–30,000 training images per class, consistent with real-world frequencies of occurrence.

With this large dataset, we can apply the powerful CNN model to extracting informative features representing the scene context information. Specifically, the CNN model takes the whole image as input and outputs a score for each category. Following [45], we train the VGG network [41] on the dataset using Caffe deep learning toolbox. And the last fully connected layer is used to represent the scene context. After obtaining the generic deep scene features for visual recognition, we use a logistic regression model to fit it and output $P(y_i|I)$ which measures the probability of existence of

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048

the category y_i in the input image.

$$\phi_g(y_i) = -log(P(y_i|I)) \tag{5}$$

where $P(y_i|I)$ measures the probability that an object with label y_i appears in the input image I.

3.3. Inference algorithm

To tackle this fully-connected CRF model, we use the mean field approximation method to minimize the objective function. Following [28], we adopt a fast mean field approximation algorithm to compute the marginals. Given the current mean field estimates $\{Q_i\}$ of the marginals, the update equation can be written as

$$Q_i(y_i) \propto exp(-\phi_u(y_i) - \omega_g\phi_g(y_i) - \omega_p \sum_{y_j \neq y_i} \sum_{j \neq i} Q_j(y_j)\phi_p(y_j, y_i)). \tag{6}$$

After convergence, we obtain an approximate posterior distribution of object labels for each node. To obtain the final results, we can employ the mean field approximate marginal probability $Q_i(y_i)$ as a detection score. Since the number of object proposals is mostly around 300, the time cost is almost free. The whole procedure for inference has been presented in Algorithm 1.

Algorithm 1 Mean field in fully connected CRFs.

1: Initialize Q

$$Q_i(y_i) = \frac{1}{Z_i} exp\{-\phi_u(y_i) - \omega_g\phi_g(y_i)\}$$

2: **while** not converged **do**

3: Message passing.

$$\tilde{Q_i}(l) = \sum_{j \neq i} Q_j(l)$$

4: Compatibility transform

$$\hat{Q_i}(y_i) = \sum_{l \in L} \tilde{Q_i}(l)\omega_p\phi_p(l, y_i)$$

5: Local update

$$Q_i(y_i) = \frac{1}{Z_i} exp\{-\phi_u(y_i) - \omega_g\phi_g(y_i) - \hat{Q_i}(y_i)\}$$

6: **end while**

4. Experiments

In this section, we conduct a series of experiments to evaluate the performance of our approach and compare it against the state-of-the-art object detection baseline. We evaluate our method on object detection benchmark datasets PASCAL VOC 2007 [30]. There are 20 different categories of objects and every dataset is divided into train, val and test subsets. In this dataset, object appearances vary greatly from changes in different illuminations, poses, locations, viewpoints and the presence of occlusions. We compare the performance in terms of mean average precision (mAP) which is the principal quantitative measure in VOC object detection task [30]. The results demonstrate that our method can boost the detection performance effectively.

4.1. Experiment and evaluation details

All of our experiments use the Faster R-CNN method [10] as our baseline. In our experiments, we use the python implementation of Faster R-CNN from [46]. Their code gives similar, but not exactly the same, mAP as the MATLAB version used in [10]. In order to show that our method is insensitive to the stage of object proposal, we utilize two CNN architecture including ZFNet [40] and VGGNet [41] to train the Faster R-CNN system. In what follows, we use ZF and VGG to denote Faster R-CNN with ZFNet and VGGNet respectively. Pp means the pairwise potentials, while Gp means the incorporation of the global potentials. The balance parameters for the pairwise potentials and the global potentials are selected via cross validation. We show the performance in terms of AP for each class on VOC 2007 test dataset.

Table 1
The result of VOC 2007 test dataset based on ZFnet trained with VOC2007 and VOC2012 trainval dataset.

Class	ZF	ZF+Pp	ZF+Gp	ZF+Gp+Pp
Aero	67.82	68.13	68.84	**69.27**
Bike	71.22	70.87	70.92	**71.25**
Bird	59.15	**60.35**	60.07	60.22
Boat	49.88	49.54	**50.22**	48.91
Bottle	33.89	35.68	37.60	**38.32**
Bus	71.75	**72.16**	71.74	71.74
Car	**75.21**	74.99	75.16	75.00
Cat	**79.94**	79.35	78.95	79.58
Chair	38.45	39.39	39.13	**39.79**
Cow	70.41	**72.38**	70.43	71.89
Table	58.93	59.86	60.36	**60.87**
Dog	**74.18**	74.03	73.54	73.99
Horse	79.83	80.49	79.87	**80.51**
m-bike	72.53	71.75	**72.69**	72.68
Person	65.06	**65.51**	65.22	65.49
Plant	30.36	**33.11**	32.19	32.77
Sheep	66.45	**67.68**	66.59	67.54
Sofa	60.61	59.19	**61.12**	60.10
Train	**72.47**	72.30	72.17	71.91
TV	58.51	59.23	59.65	**59.75**
mAP	62.83	63.30	63.32	**63.58**

4.2. Analysis

First, we examine the influence of the trade-off parameters in our model. Then we show the performance in terms of average precision on the VOC 2007 test.

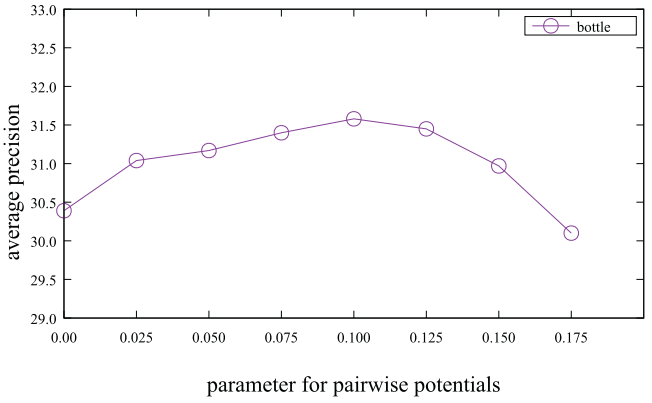
4.2.1. Parameter selection

In our CRF framework, the parameters ω_p and ω_g in (2) balance the influence of the local appearance and the contextual information. Here we focus on ω_p which represents the importance of the object-level contextual information and other hyper-parameters are fixed. The results of average precision for classes bottle, cow and pottedplant are shown in Fig. 4. As shown in Fig. 4, when the parameter ω_p becomes larger, our method achieves better AP at first and then the performance decreases. It demonstrates that there exists a balance for local appearance and contextual coherent constraints among object candidates in the same image.

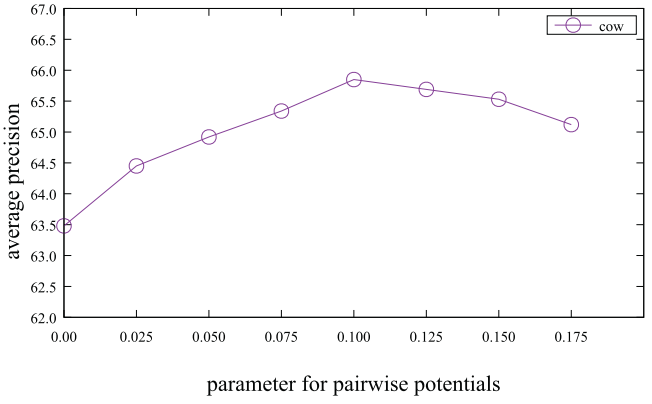
4.2.2. Performance on the ZFnet

In Table 1, we take ZFnet which is trained on the union of VOC 2007 and 2012 trainval datasets as the baseline and our approach obtains 0.75% improvement in terms of mAP. In this experiment, our method yields better performance over 13 classes than the baseline. Among these categories, the class bottle is enhanced by 4.43% and class plant achieves 2.41% improvement. Both pairwise potentials and global potentials support the object detection. Moreover, we can see that the pairwise potentials part plays a more important role because it achieves better improvement in 6 classes compared to 2 classes' enhancement caused by the global potentials.

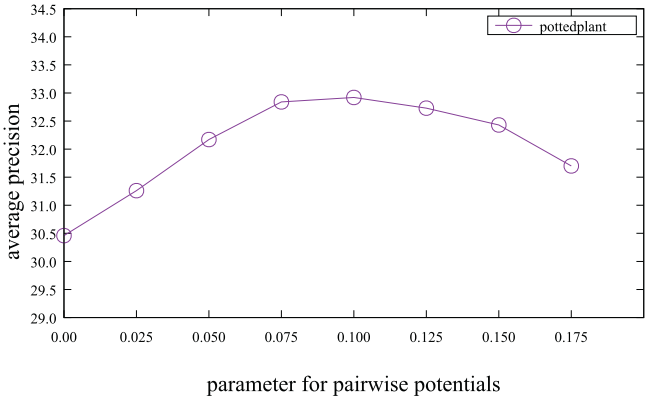
Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048



(a) AP for Bottle



(b) AP for Cow



(c) AP for Pottedplant

Fig. 4. AP for three classes with different ω_p.

4.2.3. Performance on the VGGnet

From Table 2, we can see that the results of our method based on VGGnet is 73.51% in terms of mAP which achieves 0.70% improvement than the Faster R-CNN method. And it outperforms baseline method on 14 classes out of a total of 20 of them. Furthermore, the results demonstrate that our method has the potential to work similarly well on different object detection methods since it only needs the object proposals and the corresponding scores. However, the improvement becomes lower than which it achieved with the model based on ZFnet. The reason maybe that the VGGnet CNN model which focuses on candidates themselves are powerful enough to overlook the help from contextual information in some extent.

Table 2
The result of VOC 2007 test dataset based on VGGnet trained with VOC2007 and VOC2012 trainval dataset.

Class	VGG	VGG+Pp	VGG+Gp	VGG+Gp+Pp
Aero	75.47	75.98	**76.33**	76.00
Bike	79.82	80.06	80.02	**80.50**
Bird	74.52	75.34	74.68	**75.38**
Boat	59.90	**59.97**	58.06	59.43
Bottle	52.70	53.97	**55.62**	55.36
Bus	82.93	**83.01**	82.39	82.81
Car	84.65	84.79	84.68	**84.81**
Cat	**88.33**	87.94	87.82	87.87
Chair	52.56	53.34	**53.76**	53.38
Cow	79.29	**83.49**	79.46	83.47
Table	66.04	65.72	**66.92**	66.08
Dog	84.81	85.45	84.98	**85.51**
Horse	85.00	85.55	84.00	**85.57**
m-bike	**76.82**	76.65	76.59	76.76
Person	76.64	**77.01**	76.52	76.95
Plant	36.86	40.28	40.36	**40.94**
Sheep	75.65	75.91	75.60	**75.98**
Sofa	**73.15**	71.58	72.78	72.25
Train	81.79	**82.22**	82.15	82.18
TV	**71.58**	71.55	71.11	71.41
mAP	72.93	73.49	73.19	**73.63**

Table 3
The results of VOC 2007 test dataset. The results of the other methods are taken from [35,47,48].

Method	mAP
DeepID-Net [47]	64.1
AC-CNN [48]	72.0
ION [35]	75.6
Ours	73.5

4.3. Comparison with other methods considering contextual information

To prove the effectiveness of our proposed method, we compare the following three existing algorithms which take the contextual information into account:

- DeepID-Net [47]: This method learns a deep model for the image classification task taking scene information into consideration. The image classification scores are used as contextual features, and concatenated with the object detection scores to form a feature vector, based on which a linear SVM is learned to refine the detection scores.
- ION [35]: This method exploits information both inside and outside the region of interest. Contextual information outside the region of interest is integrated using spatial recurrent neural networks. Inside, the proposed algorithm uses skip pooling to extract information at multiple scales and levels of abstraction.
- AC-CNN [48]: This method combines one attention-based global contextualized (AGC) sub-network with one multi-scale local contextualized (MLC) sub-network to capture global context and local context, respectively. Then the global and local context are fused together for making the final decision for detection.

In Table 3, we compare our method against many existing models, including DeepID-Net [47], ION [35], AC-CNN [48] and our proposed method on the detection mAP on VOC 2007 test dataset. The results in Table 3 show that our method can obtain competitive performance with those state-of-the-art approaches.

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048

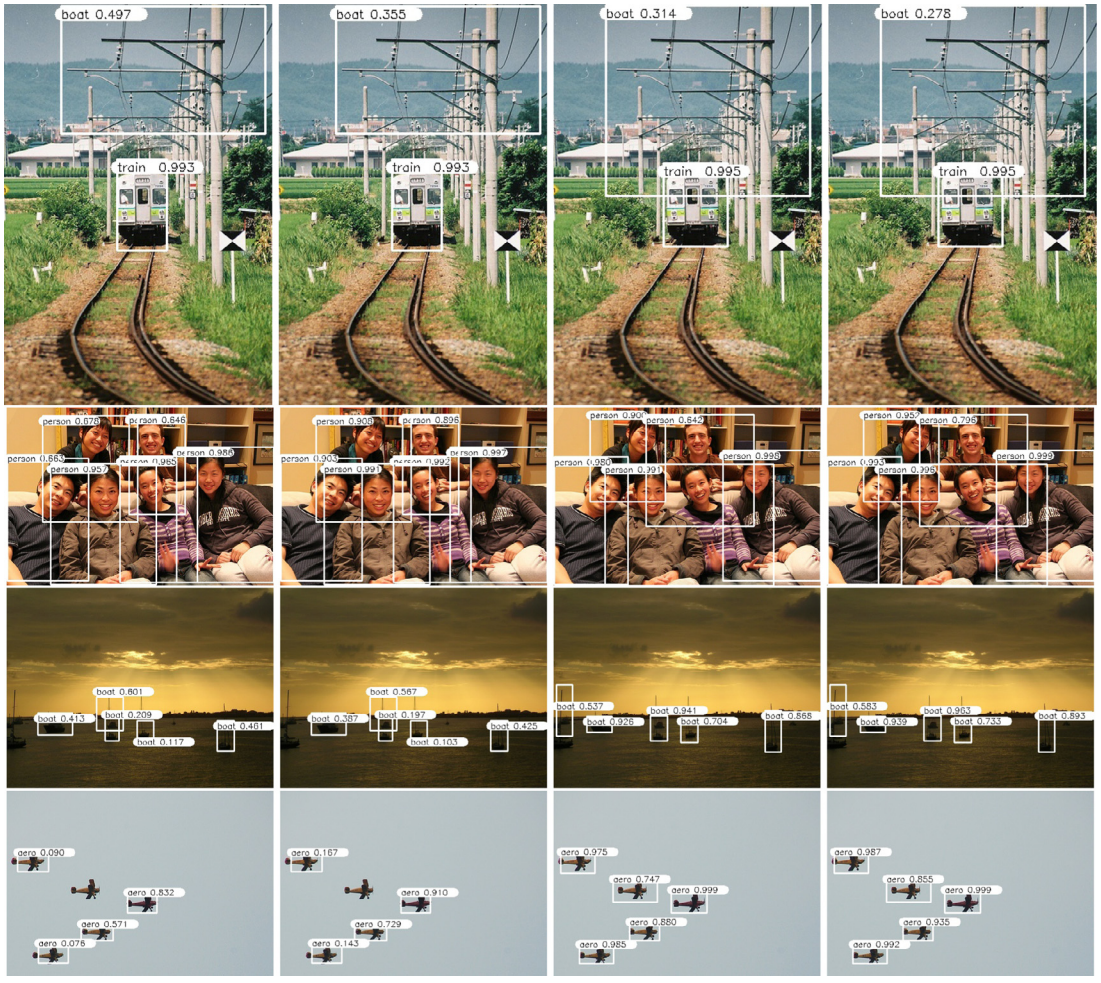


Fig. 5. Visualization of hits and misses on VOC 2007 images. In each image, the first and third columns show the results of ZFnet and VGGnet. The results of our contextual model based on ZFnet and VGGnet are displayed in second and fourth columns.

4.4. Visualization of more results

Fig. 5 visualizes the performance of our method against Faster R-CNN on some VOC2007 images. In most situations, our method can improve the detection performance. However, for the image in the third line which is full of tiny boats, the results are worse than that of ZFnet. Actually, we find that it is because our global contextual part could not recognize the scene as a lake well. And it reminds us that there still has very large development space. In addition, the results in the fourth line also show that our method cannot improve those situations when the objects were not marked by the object detector.

5. Conclusions

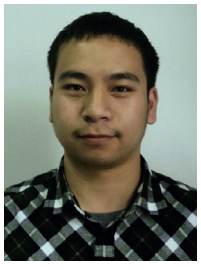
In this paper, we have proposed a novel object detection system which combines the local appearance of the object proposals and the contextual information around them. Here, we employ the powerful deep convolutional neural network to obtain unary potentials for the object proposals and extract the global features representing scene context. In addition, the pairwise potentials which take both semantic and spatial relevance into account for different object proposals are utilized to produce a semantically coherent interpretation of the input image. We formulate the whole problem in the form of a fully-connected CRF model which can be efficiently solved by a fast mean field inference method. Furthermore, our experimental evaluation has demonstrated that our approach could effectively leverage the contextual information to improve detection accuracy, thus outperforming existing detection techniques on benchmark datasets. In the future, we will devise a better pairwise model based on CNN and incorporate it into an end-to-end framework.

References

[1] J. Carreira, C. Sminchisescu, Cpmc: automatic object segmentation using constrained parametric min-cuts, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1312–1328.
[2] J.R. Uijlings, K.E. van de Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.
[3] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 328–335.
[4] P. Krähenbühl, V. Koltun, Geodesic object proposals, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 725–739.
[5] C.L. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, in: Computer Vision—ECCV 2014, Springer, 2014, pp. 391–405.
[6] M. Guo, Y. Zhao, C. Zhang, Z. Chen, Fast object detection based on selective visual attention, Neurocomputing 144 (2014) 184–197.
[7] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[8] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[9] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 580–587.
[10] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048

[11] I. Biederman, Perceiving real-world scenes, Science 177 (4043) (1972) 77–80.
[12] I. Biederman, R.J. Mezzanotte, J.C. Rabinowitz, Scene perception: detecting and judging objects undergoing relational violations, Cogn. Psychol. 14 (2) (1982) 143–177.
[13] A. Torralba, Contextual priming for object detection, Int. J. Comput. Vis. 53 (2) (2003) 169–191.
[14] D.J. Crandall, D.P. Huttenlocher, Composite models of objects and scenes for category recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, IEEE, 2007, pp. 1–8.
[15] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, IEEE, 2007, pp. 1–8.
[16] Z. Tu, Auto-context and its application to high-level vision tasks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE, 2008, pp. 1–8.
[17] C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE, 2008, pp. 1–8.
[18] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 1585–1592.
[19] M.J. Choi, J.J. Lim, A. Torralba, A.S. Willsky, Exploiting hierarchical context on a large database of object categories, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 129–136.
[20] C. Desai, D. Ramanan, C.C. Fowlkes, Discriminative models for multi-class object layout, Int. J. Comput. Vis. 95 (1) (2011) 1–12.
[21] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.
[22] M. Lin, C. Zhang, Z. Chen, Global feature integration based salient region detection, Neurocomputing 159 (2015) 1–8.
[23] C. Conde, D. Moctezuma, I.M. De Diego, E. Cabello, Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments, Neurocomputing 100 (2013) 19–30.
[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, IEEE, 2005, pp. 886–893.
[25] V.-D. Hoang, M.-H. Le, K.-H. Jo, Hybrid cascade boosting machine using variant scale blocks based hog features for pedestrian detection, Neurocomputing 135 (2014) 357–366.
[26] S. Yao, S. Pan, T. Wang, C. Zheng, W. Shen, Y. Chong, A new pedestrian detection method based on combined HOG and LSS features, Neurocomputing 151 (2015) 1006–1014.
[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[28] V. Koltun, Efficient inference in fully connected CRFs with Gaussian edge potentials, Adv. Neural Inf. Process. Syst 2 (3) (2011) 4.
[29] Z. Hayder, M. Salzmann, X. He, Object co-detection via efficient inference in a fully-connected CRF, in: Computer Vision—ECCV 2014, Springer, 2014, pp. 330–345.
[30] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes challenge 2007 (VOC 2007) results (2007), 2008.
[31] C. Szegedy, A. Toshev, D. Erhan, Deep neural networks for object detection, in: Advances in Neural Information Processing Systems, 2013, pp. 2553–2561.
[32] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, Scalable object detection using deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2147–2154.
[33] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[34] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[35] S. Bell, C.L. Zitnick, K. Bala, R. Girshick, Inside–outside net: detecting objects in context with skip pooling and recurrent neural networks, arXiv preprint arXiv:1512.04143. (2015).
[36] S. Gidaris, N. Komodakis, Locnet: improving localization accuracy for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 789–798.
[37] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human–object interaction activities, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 17–24.
[38] B. Li, T. Wu, S.-C. Zhu, Integrating context and occlusion for car detection by hierarchical and-or model, in: Computer Vision—ECCV 2014, Springer, 2014, pp. 652–667.
[39] T.-H. Vu, A. Osokin, I. Laptev, Context-aware CNNs for person head detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2893–2901.
[40] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision—ECCV 2014, Springer, 2014, pp. 818–833.
[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556. (2014).
[42] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE, 2009, pp. 248–255.
[44] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million Image Database for Scene Recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[45] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[46] R. Girshick, py-faster-rcnn, 2016, (https://github.com/rbgirshick/py-faster-rcnn) [Online; accessed 1 December 2016].
[47] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, et al., DeepID-Net: deformable deep convolutional neural networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2403–2412.
[48] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, S. Yan, Attentive contexts for object detection, IEEE Trans. Multimedia 19 (5) (2017) 944–954.



Wenqing Chu received his B.E. degree in Computer Science and Technology from Huazhong University of Science and Technology, in 2014. He is currently a Ph.D. candidate in Computer Science at Zhejiang University. His research interests include machine learning, computer vision and data mining.

Deng Cai is a professor in the State Key Lab of CAD&CG, College of Computer Science at Zhejiang University, China. He received the Ph.D. degree in computer science from University of Illinois at Urbana Champaign in 2009. His research interests include machine learning, data mining and information retrieval.

Please cite this article as: W. Chu, D. Cai, Deep feature based contextual model for object detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.048