Pattern Recognition Letters 86 (2017) 56–61

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Discovering overlooked objects: Context-based boosting of object detection in indoor scenes

Jongkwang Hong[a], Yongwon Hong[a], Youngjung Uh[a], Hyeran Byun[a,*]

Department of Computer Science, Yonsei University, Shinchon-Dong, Seodaemun-gu, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:
Received 7 June 2016
Available online 21 December 2016

MSC:
41A05
41A10
65D05
65D17

Keywords:
Object detection
Context modeling
Neural networks

ABSTRACT

Contextual detection not only uses visual features, but also leverages contextual information from the scene in the image. Most conventional context based methods have heavy training cost or large dependence on the original baseline detector. To overcome such shortcomings, we propose a new method based on co-occurrence context. It is built upon recent off-the-shelf baseline detector and achieves higher accuracy than existing works while detecting additional true positives which the baseline detector could not find. Furthermore we construct an indoor specific NYUv2-context dataset to investigate context-based detection of indoor objects. It is a subset of original NYU-depth-v2 dataset and to be published online to encourage context researches. In the experiment, the proposed method obtained 21.22% mAP which outperforms the baseline and compared context-based work by 0.91 and 0.36 percentage point mAP respectively.

1. Introduction

The goal of the object detection is to find the precise location of objects and to recognize their class in the scene. This field is one of the key technologies in scene understanding, which leads to many applications. While conventional studies focus on visual discriminance of objects, variational studies harness additional contextual information to improve the detection performance. Context is extra information that can be inferred from the image besides appearance of objects, e.g., relative location, co-occurrence, and usual size.

There are two approaches utilizing context information. First type uses context information for training feature representation or classifier [1–3], and second type attach post-processing on baseline detectors result using context information [4,5]. First type may get improved result, but they have critical weakness that they require large amount of training data and plenty of time to train. Especially, since the recent state-of-the-art detection methods [6–8] are based on deep learning technique which takes long time to train, these approaches are inapplicable due to necessity of training the entire model including baseline detector (which do not use context information) to reflect context. On the other hand, second type does not require much data and time since it leaves baseline detector untouched and generate context information.

In this paper we propose a method to impose contextual information by a new context model on the state-of-the-art baseline detector. Our method has two advantages: (1) It requires small training only for context, except for baseline detector. (2) The method improves detection accuracy by finding additional 'new' objects which were not found in the baseline detector (see Fig. 3).

Furthermore, we present an indoor specific NYUv2-context dataset from NYU-depth-v2 [9] to encourage indoor context research. We consider context information in the indoor scenes is more meaningful and cohesive than outdoor scenes because most indoor objects are arranged by humans with common principle. For instance, a keyboard is likely to be located below a monitor, pillows are on a bed. More details are described in Section 3. The experiment on this dataset shows that we increase the baseline's mean average precision 0.91 percentage point, *i.e.*, a 4.48% relative improvement.

The rest of our paper is organized as follows: Section 2 describes general object detection and context-based boosting approaches related with ours. Section 3 describes our context based detection method and explains the effect of our re-scoring method. In Section 4, a new detection benchmark dataset 'NYUv2-context' is introduced and described. Later, the proposed framework is evaluated in Section 5. Finally, conclusions are given in Section 6.

2. Related work

Object detection: Most of object detection researches studied how the visual information can be represented better, how loca-

* Corresponding author.
E-mail address: hrbyun@yonsei.ac.kr (H. Byun).

http://dx.doi.org/10.1016/j.patrec.2016.12.017
0167-8655/

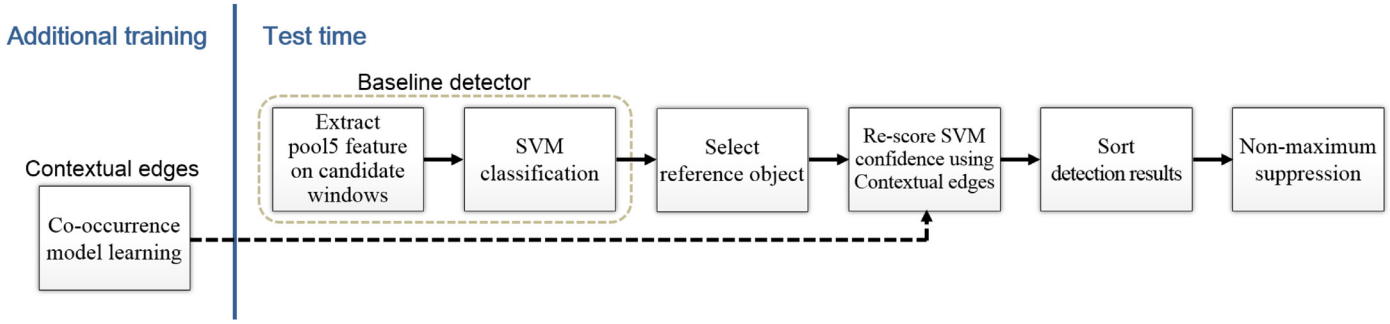

Fig. 1. Flowchart of our method. Dotted box contains baseline detector and dotted lines represent information for re-scoring.

tion of the objects can be found correctly. R-CNN [6] achieved big improvement by using deep convolutional neural network so called Alex-net [10], which is different with hand-crafted features like HOG [11] or SIFT [12]. Furthermore, they proposed fine-tuning technique that additionally trains with target domain data. It also improves detection accuracy because Alex-net is trained with classification data. SPP-net [7] improves R-CNNs result in speed, and detection accuracy. They proposed detection framework which is invariant to image aspect ratio and scale by applying spatial pyramid pooling [13] idea, which is used in bag-of-words approaches, in convolutional neural network feature. Feature reusing idea is also proposed, which forwards whole image at once and shared by object proposal boxes, to improve detection speed. Faster R-CNN [8] improves detection accuracy and speed with big margin, proposing region proposal network that generates object proposals which is deeply trained with bounding box annotations.

Context-based boosting: As deep learning based object detection researches are emerging, some other approaches that utilizing non-visual cues —relative location, co-occurrence, usual size— are consistently having studied. Guillaumin et al. [1] learn multimodal features to encode contextual information. They extract context from GIST descriptor [14] (capturing coarse texture) and tags. Zheng et al. [3] proposed Polar Geometric Descriptor, which carries Thing-Thing or Thing-Stuff contexts. In addition, Zheng et al. [3] proposed Maximum Margin Context(MMC) model which helps combining several types of context information and can be trained similar way to canonical Support Vector Machine(SVM). Instead of carrying context in features, Chen et al. [2] introduce a method to iteratively and mutually boost object classification and detection by taking the outputs from one task as the context of the other. However, these approaches require large amount of training data and time, because they need to train entire detection system including feature representation and classifier. Furthermore, whenever a state-of-the-art baseline detector is newly introduced, these approaches should be carefully modified accordingly.

On the other hand, there are researches to refine baseline detection results with post-processing. Choi et al. [5] train co-occurrence and relative location in a tree-based model and re-estimate detection scores using the trained model. Their advantage lies in that their re-estimation procedure can be used as a post-processing step on the result of existing off-the-shelf object detectors without adjusting the detector itself. However, their result largely depends on the baseline detectors output so that they cannot get 'new' labels other than the given labels by modifying the detection scores.

Our method can be regarded as a post-processing which improves baseline detector by context-boosting. As anticipated, it requires small training data, trains fast, and can be applied to any type of baseline detector, like other post-processing typed approaches. Furthermore, contrarily to existing post-processing methods, our proposed method can detect new true positive, which can provide richer correct detection results to users.

3. Context model and re-scoring framework

In this section we describe our contextual detection. The flowchart of our method is shown in Fig. 1. The contextual edge is trained for learning co-occurrence context. In test time, the result of the baseline detector (which does not use context model) is post-processed.

3.1. Co-occurrence model

Co-occurrence of multiple objects is fundamental yet powerful contextual information. To represent the co-occurrence as a measure, we construct a fully connected undirected graph. The graph has a node for each object category and weighted edges which indicate the relationship between the nodes. The weights of the edges are computed using the mutual information equation.

$$edge(i,j) = sign_{i,j} * mutual(i,j) \tag{1}$$

$$sign_{i,j} = \begin{cases} 1 & \text{if } joint(i,j) < p(i=1)p(j=1) \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

$$mutual(i,j) = (entropy_{i,i} + entropy_{j,j} - entropy_{i,j}) \tag{3}$$

$$entropy_{i,j} = -\sum joint(i,j)log(joint(i,j)) \tag{4}$$

Here $joint(i,j)$ and $mutual(i,j)$ are binary joint probability distribution and mutual information between object category i and j respectively. Eq. (2) determines the edge's sign to be negative if the joint probability $joint(i,j)$ is smaller than the product of the two marginal probabilities $p(i=1)p(j=1)$, which means the object i has an unfavorable relationship with the object j.

To train the edges from the data, we need to calculate binary joint probability distribution of every object-object relationship. We follow the implementation in tree-context [5] which trains a binary joint probability distribution by counting presence/presence, presence/non-presence, non-presence/presence, and non-presence/non-presence frequency of each pair of object categories on the images. Fig. 2 represents the edges trained on NYUv2-context in a scaled image.

3.2. Reference object selection

The co-occurrence model represents the relationship between two objects. To apply the co-occurrence information to original detection results in test-time, we propose to choose some object categories as the principal objects for each image. We term the categories as 'reference object(RO)'. We have to choose RO carefully because we apply context starting from it. To select the RO, we adopt 'presence prediction', which measure presence potential of each object category in certain image. The object categories that