Abnormal Object Detection by Canonical Scene-Based Contextual Model

Sangdon Park, Wonsik Kim, and Kyoung Mu Lee

Department of EECS, ASRI, Seoul National University, 151-742, Seoul, Korea
{sangdon,ultra16,kyoungmu}@snu.ac.kr
http://cv.snu.ac.kr

Abstract. Contextual modeling is a critical issue in scene understanding. Object detection accuracy can be improved by exploiting tendencies that are common among object configurations. However, conventional contextual models only exploit the tendencies of normal objects; *abnormal objects* that do not follow the same tendencies are hard to detect through contextual model. This paper proposes a novel generative model that detects abnormal objects by meeting four proposed criteria of success. This model generates normal as well as abnormal objects, each following their respective tendencies. Moreover, this generation is controlled by a latent scene variable. All latent variables of the proposed model are predicted through optimization via population-based Markov Chain Monte Carlo, which has a relatively short convergence time. We present a new abnormal dataset classified into three categories to thoroughly measure the accuracy of the proposed model for each category; the results demonstrate the superiority of our proposed approach over existing methods.

Keywords: abnormal object detection, generative model, sampling.

1 Introduction

Contextual modeling is a critical issue in scene understanding, particularly in object detection [2–4, 1, 5, 6]. Contextual models exploit the prior knowledge that in a specific scene, specific objects follow *common* or *normal* configurations, such as “cars” on the “road.” Thus, conventional contextual models weaken “car” bounding boxes floating over the “road,” or reinforce bounding boxes in correct positions. However, detecting “cars” actually floating on the “road” is difficult if contextual models only consider normal configurations. In this paper, we propose a novel generative model that can detect out-of-context objects, also referred to as *abnormal objects.*

Finding abnormal objects is an important and interesting task. With the advent of image-manipulation tools such as Adobe Photoshop, the number of artificially manipulated images continues to increase. Additionally, the ability to detect abnormal objects can be useful in surveillance systems. Therefore, models able to understand abnormal scenes are needed.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 651–664, 2012.

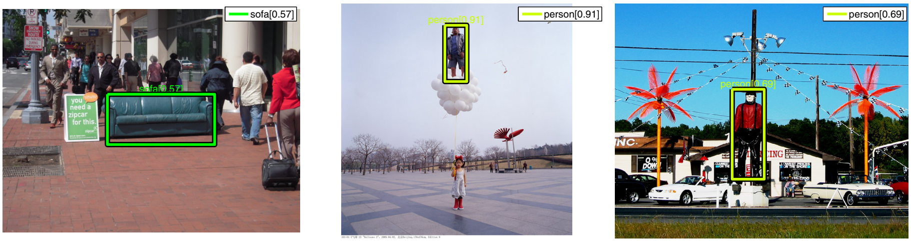


Fig. 1. Examples of abnormal objects detected using the proposed method. Each image contains the topmost abnormal object that violates co-occurrence, relative position and relative scale among objects.

This paper focuses on *abnormal object detection*, that is, finding abnormal objects in given scenes. We define abnormal objects as those that do not match expectations set by the surrounding context or by common knowledge. Specifically, abnormal objects (1) do not co-occur with surrounding objects (*co-occurrence-violating objects*), (2) violate positional relationships with other objects (*position-violating objects*) or (3) have relatively huge or small sizes (*scale-violating objects*) (Fig. 1). Exploiting the distributions of normal objects is necessary for abnormal object detection. This is because abnormality can be defined based on the extent to which an object is not normal.

Four necessary properties of abnormal object detection methods are defined as follows: (1) affluent object relations, (2) quantitative object relations, (3) affluent context types, and (4) prior-free object search. First, affluent relations among objects, such as a fully connected relation among objects, is critical because abnormal objects rarely occur. If an abnormal object, such as a floating "car", is related with only one object, "road," then identifying whether the target object is abnormal or not becomes difficult, because models can only identify the floating car's abnormality once the "road" has been detected correctly.

Second, object relations, particularly relative position/scale relation, should be defined quantitatively because qualitative representation is improper for determining the contextual violations. For example, if the relative position is qualitatively defined, such as *above*, then "person" and "road" can be related by the *above* relation. However, identifying the abnormality of a person hanging in the air becomes difficult because the "person" is still *above* the "road". Third, contextual models become more informative when the more context types, such as co-occurrence and relative position/scale among objects, are used [7]. Finally, the models should not restrict the interpretation of scenes to find abnormal object properly. If an abnormal object-detecting model has prior knowledge of searching objects, then finding abnormal objects is nearly impossible because abnormal objects do not exist in locations where they are expected.

We propose a novel generative model that generates both normal and abnormal objects, following a learned configuration among the objects. This learned configuration is determined by transforming a standard configuration, also called a *Canonical Scene*. This Canonical Scene, conditioned on a scene type, is a space wherein normal and abnormal objects can exist, following their physical positions and scale. In this sense, the Canonical Scene is similar to the real world. In the Canonical Scene, normal objects co-occur with each other without violating relative position and scale among objects, but abnormal objects do not. The proposed method can check whether the object configuration of the input scene is similar to that of the Canonical Scene by modeling the generative model of the Canonical Scene, thereby identifying abnormal and normal objects. The first three aforementioned necessary properties for abnormal object detection models are satisfied by the definition of the Canonical Scene. In addition, the proposed model does not dictate which or where objects exist, which satisfies the final property.

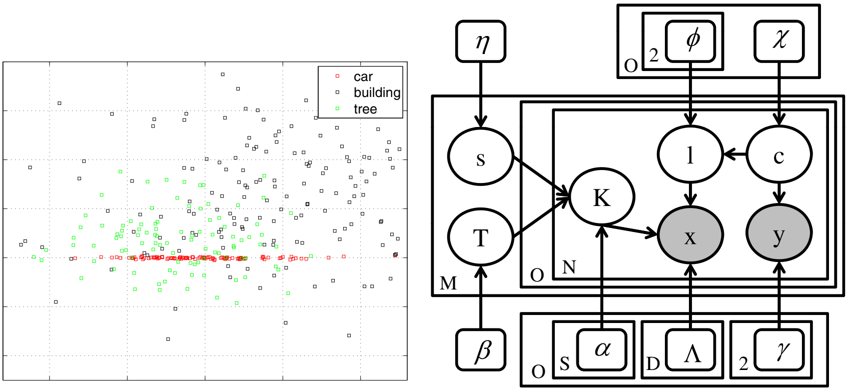
We use a population-based Markov Chain Monte Carlo (Pop-MCMC) technique [8] to predict the existence of abnormal objects via the proposed model. This technique generates samples from multiple, dependent chains, resulting in a high mixing rate. With regard to optimization, Pop-MCMC has more chances of escaping from the local optimum, making it a proper optimization tool for multimodal and/or high dimensional models.

We also propose a novel annotated dataset that contains one or more abnormal objects in each images to verify the accuracy of abnormal object detection tasks. This dataset consists of three separate sets of scenes where (1) fewer co-occurring objects and (2) location-violating objects exist, and (3) scale-violating objects exist. Whether the proposed abnormal object detection methods are adequate for finding each type of abnormal object must be verified. At present, only a few studies on abnormal object detections have been conducted [1, 5]. This paper is inspired by these studies.

2 Related Work

Although a variety of contextual models have been proposed over the past decade [2–4, 1, 5, 9, 7, 6, 10, 11], contextual models that meet all four criteria for find abnormal objects are difficult to find.

One group of such models incorporates object-object interaction [12] which is used to directly enforce contextual relation among objects. In [9], a conditional random field (CRF)-based contextual model provides only co-occurrence statistics among objects to refine miss-labeled image segments. This work is further expanded in [7] so that relative position context, such as “above/below” and “inside/around,” among objects is qualitatively encoded in the CRF model. Furthermore, [4] proposed a “far/near” relation between objects to reduce the ambiguity of qualitative context representation. In [1], Choi represents the relation among objects in a tree model for computational efficiency, taking only a few significant relationship among the objects into consideration. This tree model is more expanded in [5] to also encode “support” context, a geometric context for prohibiting floating objects. The strategy of [3] is to first find easy objects, and then search for difficult objects based on the locations of the easy



(a) (b)

Fig. 2. (a) Possible locations of objects in learned "outdoor" Canonical Scene. (b) The graphical model of the Canonical Scene model.

objects. This model combines boosting and CRF to efficiently search for unknown objects based on known ones, making it unnecessary to search the entire image to locate objects. [6] leverages the qualitative spatial contexts of "stuff," such as "sky" or "road," to improve object detection in satellite images.

Another group of contextual models indirectly embeds the contextual relationship among objects via a latent variable, such as a scene. [2] correlates objects through a common "cause" scene, thus including co-occurrence statistics in a hierarchical Bayesian model. The same contextual information is exploited by [10] to achieve scene classification, annotation, and segmentation altogether in a single framework. [13] solves the object detection problem by assuming that holistically similar images share the same object configuration. This method retrieves normal, annotated images with similar Gist feature [14] and then constructs a graphical contextual model to constrain which objects appear and where they can be found.

3 Transformed-Canonical Scene Generating Model

Our goal is to model a joint distribution over scenes and objects; Simultaneously inferring the scene type, which objects exist, where the objects are positioned, and what the objects' sizes are is made possible by maximizing distribution. From these inferred ones, we also infer which objects are abnormal ones.

The proposed graphical model is described in Fig. 2(b). The observed variables $\mathbf{x}$ and $\mathbf{y}$ can be considered as an input image of the whole system. We assume that the input image is represented as $\sum_{o=1}^{O} N_o$ number of image patches (*candidate objects*), where O is the number of object categories and N_o is the number of candidate objects in object category o. Normally or abnormally located candidate objects are called just normal or abnormal *objects*, respectively, when the candidate objects are correctly detected. Each candidate objects is located at $x_{o,n}$ and has the appearance score $y_{o,n}$, where o represents an object category and n is the index of an instance of the candidate object. $\mathbf{x}$ is a location matrix, where the element $x_{o,n}$ is the location vector of the corresponding candidate object. $\mathbf{y}$ is an appearance matrix, where the element $y_{o,n}$ represents a quantified similarity between the appearance of object category o and the corresponding candidate object. In this paper, $O = 10$ and $N_o = 10$ for all object category o.

The latent variables s, $\mathbf{c}$, and $\mathbf{l}$ can be considered as outputs of the system. The variable $s \in [1, 2, \ldots, S]$ is a scene category of the given image where S is the number of scene categories. The matrix $\mathbf{c}$ is a correctness matrix of appearance, where the element $c_{o,n}$ is a boolean-valued flag. When the flag is set to 1, the corresponding candidate object is similar to object o in appearance. The matrix $\mathbf{l}$ is a location type matrix, where each element $l_{o,n}$ is also a boolean-valued flag. $l_{o,n} = 1$ means that the corresponding candidate object is positioned at a normal location of object category o. Note that the "location" is used to represent context information because this paper assumes that the co-occurrence, position, and scale information of the objects are geometrically represented. We assume that two or more objects can exist in the same scene if and only if they co-occur with each other. Moreover, both the position information and scale information of the objects can be geometrically represented by using the "undo" projectivity technique, which is described in the following section. This paper distinguishes between normal and abnormal objects through geometric representation by assuming all normal objects, not "stuff," are located on the ground plane. In summary, all context information are compactly represented as objects' locations.

3.1 Image Representations and Assumptions

This paper adopts object-level image representation (such as [15, 1]), which differs from conventional low or mid-level representations of images, such as colors, SIFT or bag of visual words. Representing an image at object-level is necessary because our goal includes modeling the object's position/scale in the real world based on the objects in the image.

An image is described as a set of candidate objects, $\{(x_{o,n}, y_{o,n}) | 1 \leq o \leq O, 1 \leq n \leq N_o\}$. This set can be a set of bounding boxes generated by applying conventional object detectors on a given image, but only the top N_o scored candidate objects are used among the outputs of the object o detector. $x_{o,n} = (b_{o,n}, h_{o,n})$ is the location and height of the corresponding candidate object, where $b_{o,n}$ is a y-coordinate of the candidate object's center and $h_{o,n}$ is the height of the candidate object. We do not consider an x-coordinate and the width of a candidate object because these values are not informative [14]. Thus, the words "scale" and "height" are used interchangeably. $y_{o,n}$ is a score that represents the similarity between the appearance of the corresponding candidate object and that of the object category o. $y_{o,n}$ can be calculated by applying an object detector (such as [16]) to the corresponding candidate object.

We transform the location of the candidate object from the image coordinate into the camera coordinate to exploit the position relationships and the scale information between the objects. For this purpose, we apply Hoiem's "undo" projectivity technique [17], assuming that all instances of the same object category o have the same physical height. The object's location, that is, in an image coordinate is transformed into a camera coordinate by the following triangulation [1]: $G_{o,n} = (G^y_{o,n}, G^z_{o,n}) = (\frac{b_{o,n}}{h_{o,n}} H_o, \frac{f}{h_{o,n}} H_o)$. The coordinates are calculated based on the center location $b_{o,n}$ of the candidate object, the candidate object's height $h_{o,n} \in [h_1, h_2]$, the hand-crafted physical height of the objects $H_o \in [H_1, H_2]$, and the focal length f. This paper assumes that the height of the images is normalized to one, the principal axis passes through the center of the image, $f = 1$, $h_1 = 0.1$, $h_2 = 1$, $H_1 = 0.1$, and $H_2 = 30$. Furthermore, with a normalization of the location space to $\mathcal{S} = [-0.5, 0.5] \times [0, 1]$, we restate the location of the corresponding candidate object as $x_{o,n} = f_{norm}(G_{o,n})$, which represents both position and scale information.

This paper assumes that all instances of an object category o have the same physical height. Thus, the relatively small object instance in the image also has the same height as the normal objects in the camera coordinate. Even if the small object is difficult to identify based on its height, the value of the second component of the small object's $x_{o,n}$ is larger than that of the normal ones. Thus, the abnormal object can be identified using the second component of $x_{o,n}$.

This paper employs the exact sense of "objects" used by [18]. Therefore, we use "car object," "person object," or "sky stuff." However, "object" and "stuff" are called "object" if no confusion exists.

3.2 Canonical Scene for Location Model

This subsection presents one of the main idea of this paper: a novel approach for designing the location model $p(\mathbf{K}, \mathbf{x}|s, \mathbf{l}, T)$. This model measures whether the candidate objects' configuration in an input scene is similar to that of a predefined scene template, also called the *Canonical Scene*. This idea can be considered a template matching problem. Thus, the following processes are required: defining a scene template, matching transformation, and measuring the similarity between the input scene and the template.

Canonical Scene. Given a scene category s, a Canonical Scene $\mathcal{L}_s = \{L_{o,n,d}\}$ is a set of locational random vectors of nth candidate object of object category o, where $d \in \{0, 1\}$; $L_{o,n,0}$ and $L_{o,n,1}$ represent locations of abnormal and normal candidate objects, respectively; and $L_{o,n,d}$ follows truncated Gaussian distributions. For example, Fig. 2(a) represents a realized "outdoor" Canonical Scene in which possible normal locations for "car," "building," and "tree" are defined. Note that even though the range of $L_{o,n,1}$ is restricted in $\mathcal{S}$, for simplicity, we use Gaussian instead of truncated Gaussian distribution by assuming that all masses of distributions of candidate objects' location is inside of the restricted domain. Likewise, truncated Gaussian distributions are approximated as uniform distributions by letting locations of abnormal candidate objects follow uniform distributions. This paper represents the parameters of Gaussian and uniform distributions together as $\alpha_{s,o,d}$.

Matching Transformation. The Canonical Scene is transformed to match the input scene. This matching transformation can be arbitrary, depending on the applications. Only isometric transformation is considered in this paper because if changing the scale of the Canonical Scene would make identifying abnormal objects difficult. For example, if a scene has many "person" objects and one floating "person," the distance from the ground plane to the floating "person" is a critical clue for identifying the abnormal object. However, if the scale of the Canonical Scene is changed, the distance would becomes vague. Thus, the two-dimentional (2-D) isometric matching transformation T only consists of translation $[\tau\ 0]'$ and rotation R_θ, where $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. Moreover, isometry is only restricted on congruence mapping with no reflection, and is therefore invertible.

Similarity Measure. Matching measures can be defined in many ways, such as SAD, SSD, and maximum-likelihood measure [19]. The maximum-likelihood similarity measure between the Canonical Scene and the input scene is defined as

$$m^T_{s,\mathbf{l}}(\mathcal{L}_s, \mathbf{x}) \equiv \prod_{o,n} m^T_{s,l_{o,n}}(L_{o,n}, x_{o,n}), \tag{1}$$

where $m^T_{s,l_{o,n}}(L_{o,n}, x_{o,n})$ is an arbitrary similarity measure between a template feature $L_{o,n}$ and an image feature $x_{o,n}$. This paper measures the similarity between the locations of candidate objects in the transformed Canonical Scene $\mathcal{T}(\mathcal{L}_s)$ and the locations of candidate objects in the input scene $\mathbf{x}$, where $\mathcal{T}$ is the isometric transformation by T. Because $L_{o,n} = \{L_{o,n,0}, L_{o,n,1}\}$, matching one of them to $x_{o,n}$ is required. This paper defines the aforementioned matching as follows: If $l_{o,n} = d$, then $L_{o,n,d}$ is mapped to $x_{o,n}$. Thus, $\mathbf{l}$ control matches between $\mathcal{L}_s$ and $\mathbf{x}$.

Connection with the Location Model. If a Canonical Scene is properly learned, the configurations of candidate objects in the Canonical Scene are the same as those in the real world. For example, Fig. 2(a) is an "outdoor" Canonical Scene in which possible normal locations of candidate objects, such as the normal locations of "car" and "building", are defined. The possible location of "car" is on the ground plane on average and that of "building" is above the ground plane. Moreover, "sofa" does not exist in the "outdoor" Canonical Scene. Normal objects in the Canonical Scene co-occur with each others and exist without violating relative position/scale among objects.

Because the Canonical Scene embeds objects as they are located in the real world, the similarity between the Canonical Scene and an input scene can confirm whether or not the configuration of candidate objects in the input scene follows that of candidate objects in the real world. When letting $m^T_{s,l_{o,n}}(L_{o,n}, x_{o,n}) \equiv p(K_{o,n}, x_{o,n}|s, l_{o,n}, T)$, where $\mathcal{T}(L_{o,n}) = K_{o,n}$, the location model $p(\mathbf{K}, \mathbf{x}|s, \mathbf{l}, T)$ can be naturally interpreted as the maximum-likelihood measure $m^T_{s,\mathbf{l}}$ of template matching problems, thus also being possible to check normality.

3.3 Joint Location and Appearance Model

A joint distribution $p(s, \mathbf{l}, \mathbf{c}, T, \mathbf{K}, \mathbf{x}, \mathbf{y})$ is designed as a graphical model (Fig. 2(b)) to join the location model with the appearance model. The joint distribution is restated as

$$\begin{aligned} p(s, \mathbf{l}, \mathbf{c}, T, \mathbf{K}, \mathbf{x}, \mathbf{y}) &= p(\mathbf{K}, \mathbf{x}|s, \mathbf{l}, T)p(\mathbf{y}|\mathbf{c})p(s, \mathbf{l}, \mathbf{c}, T) \\ &= \left\{ \prod_{o,n} p(K_{o,n}, x_{o,n}|s, l_{o,n}, T) \right\} \\ &\quad \left\{ \prod_{o,n} p(y_{o,n}|c_{o,n}) \right\} \left\{ p(s)p(T) \prod_{o,n} p(c_{o,n}|l_{o,n})p(l_{o,n}) \right\}, \end{aligned} \tag{2}$$

where the first term is the location model, the second is the appearance model, and the last term is the prior distribution over latent variables.

The marginal location model $\int p(K_{o,n}, x_{o,n}|s, l_{o,n}, T)\mathrm{d}K_{o,n}$ can be analytically represented. Because $p(x_{o,n}|l_{o,n}, K_{o,n})$ is defined as $\mathcal{N}(x_{o,n}|K_{o,n}, \Lambda_o)$ when the candidate object is correctly detected as a normal object, $l_{o,n} = 1$, the marginal location model is also a Gaussian with mean of $R_\theta \mu_{s,o,1} + \tau$ and covariance of $R_\theta \Sigma_{s,o,1} R_\theta^{'} + \Lambda_o$. Moreover, if $l_{o,n} = 0$ and $x_{o,n}$ is independent on $K_{o,n}$ and follows a uniform distribution, then the marginal location model also follows a uniform distribution. The appearance model $p(y_{o,n}|c_{o,n})$ is adopted from the conventional models [2, 1]. The $p(y_{o,n}|c_{o,n})$ is indirectly defined by the Bayes theorem, $p(y_{o,n}|c_{o,n}) = \frac{p(c_{o,n}|y_{o,n})p(y_{o,n})}{p(c_{o,n})}$, where $p(c_{o,n}|y_{o,n})$ is defined as a logistic regression model $\sigma(\gamma_{o,c_{o,n}}^{'}[1\ y]^{'})$. The scene distribution is defined as $p(s) \sim Mult(\eta)$ and the transformation distributions $p(T = (\theta, \tau))$ is defined as a bivariate Gaussian distribution $\mathcal{N}((\theta, \tau)|\beta = (\mu, \Sigma))$ with no correlation between θ and τ. The distribution over correctness of appearance $p(c_{o,n}|l_{o,n})$ is modeled by the Bernoulli distribution with parameter $\chi_{o,l_{o,n}}$ and location distribution $p(l_{o,n}) \sim Bern(\phi_o)$.

4 Parameter Learning

This paper explicitly estimates T to convert all random variables into observed ones for learning Canonical Scenes. Given the scene/object-level annotated dataset [20, 21], separate dataset $D = \{D_s\}$ and use D_s to build the Canonical Scene for scene category s. The distributions of $L_{o,n,1}$ are estimated by maximum a posterior criterion by considering the relation $\mathcal{T}(L_{o,n,1}) = K_{o,n,1}$. Assuming that all objects on the images are located in the ground plane, then Algorithm 1 estimates T, which transforms objects on the ground plane to the slanted plane in the camera coordinate. In this estimate, only the locations of "objects," not those of "stuff," are used. Moreover, the distribution of object o's normal location $L_{o,n,1}$ which does not exist in a D_s is set to $\mathcal{N}(\mathbf{0}, \infty^{-1})$

Learning the logistic regression model $p(c_{o,n}|y_{o,n}) = \sigma(\gamma_{o,c_{o,n}}^{'}[1\ y_{o,n}]^{'})$ in the appearance model may seen trivial. However, when data classes are imbalanced,

Algorithm 1. Estimation of the isometry T of an image

```
Input: Locations of normal objects, $K_{o,n,1}$, for all $o$ and $n$
Output: $\hat{T}$: $L_{o,n,1} \mapsto K_{o,n,1}$
Estimate line $z = a_0 + a_1 y$ passing through points $K_{o,n,1}$ by minimizing least square errors
$K_{o,n,1} = R_{\hat{\theta}} L_{o,n,1} + [\hat{\tau}\ 0]'$, where $R_{\hat{\theta}}$ is a 2-D rotation matrix with $\hat{\theta} = -\arctan(\frac{1}{a_1})$ and $\hat{\tau} = -\frac{a_0}{a_1}$.
```

parameter learning with a conventional machine learning algorithm can reduce the accuracy of the classification problem [22]. The data used in learning the appearance model are also imbalanced. The data consist of a minor number of correctly detected candidate objects, thus resulting in low accuracy for classifying correctness. To handle this problem, we adopt conventional random oversampling: Uniform sampling on minor classes provides balanced data. However, in our case, we apply more weight to samples with high $y_{o,n}$. Other parameters, such as β, η, ϕ, χ, and Λ, are experimentally set.

5 Pop-MCMC for MAP Inference

Maximizing posterior distribution (eq. (3)) is required to detect abnormal objects.

$$\hat{s}, \hat{\mathbf{l}}, \hat{\mathbf{c}} = \arg\max_{s,\mathbf{l},\mathbf{c}} \int p(s, \mathbf{l}, \mathbf{c}, T | \mathbf{x}, \mathbf{y}) \mathrm{d}T. \tag{3}$$

Because the integral in eq. (3) cannot be analytically solved and is computationally expensive, we approximate its value by assuming that most mass of the distribution over T are concentrated at a single point. This assumption seems valid because pictures are commonly taken in the conventional way. Therefore, $\int p(s, \mathbf{l}, \mathbf{c}, T|\mathbf{x}, \mathbf{y})\mathrm{d}T$ is approximated as $p(s, \mathbf{l}, \mathbf{c}, \hat{T}|\mathbf{x}, \mathbf{y})$, where $\hat{T}$ is estimated by solving the non-linear optimization problem using the gradient ascent method: $\hat{T} = \arg\max_T p(T) \int p(\mathbf{K}, \mathbf{x}|s, \mathbf{l}, \mathbf{c}, T)\mathrm{d}\mathbf{K}$. Because T follows Bivariate Gaussian distribution, the mean of T is a proper initial for the gradient method.

This paper adopts a sampling method, which is called Pop-MCMC [8], to solve the optimization problem (3). Pop-MCMC generates multiple samples, also called *chromosomes*, from multiple Markov chains in parallel. This method can make global moves because it exchanges information between samples, thus leading to a higher mixing rate compared with conventional single chain MCMC samplers. When it comes to an optimization problem, it is possible to escape from local optimum via chromosomes of the Pop-MCMC. Therefore, optimization by means of Pop-MCMC is efficient when the objective function is multimodal and/or high-dimensional, such as the proposed model as well as MRF models for stereo problems [23].

Note that detections should be scored for ranking and drawing precision-recall curve to measure object detection results. The natural score for the detections becomes the posterior marginals [1] or log-odds ratio [4]. The log-odds ratio can quantify oddness of realizations of random variables. In this paper, the log-odds ratio represents how abnormal a candidate object is. The odd ratio of object category o's nth candidate object is calculated as follows: $m((l_{o,n}, c_{o,n}) = (0,1) = \log \frac{p((l_{o,n},c_{o,n})=(0,1)|\mathbf{x},\mathbf{y})}{p((l_{o,n},c_{o,n})\neq(0,1)|\mathbf{x},\mathbf{y})}$. This log-odd ratio can be approximated as [4] and calculated by modifying the MAP inference of eq. (3).

6 Evaluation

6.1 Abnormal Dataset and Experimental Setup

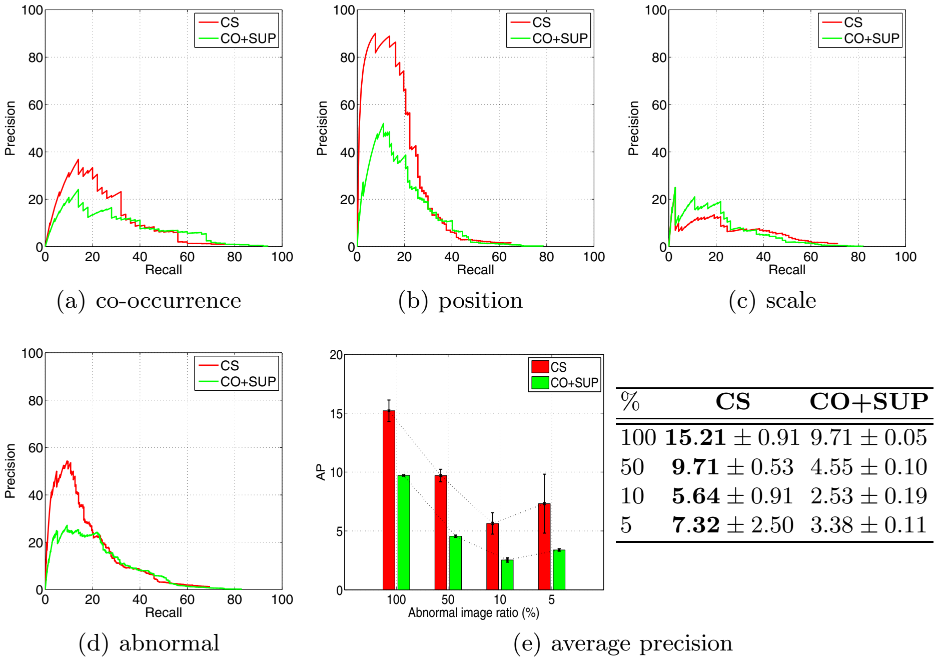
The dataset should be separated into each types of abnormal object to allow thorough evaluation of the accuracy of abnormal object detection models. The new dataset[1] consists of three different types of images (50 images each, for a total of 150 images containing annotated objects that violate co-occurrence, relative position, and relative scale) and additional extra images that contain two or more different types of abnormal objects. The abnormal objects are annotated relative to normal objects contained in the normal dataset. For example, a conventional normal dataset, such as LabelMe, consists of "cars" on the road, and so a flying "car" is annotated as an abnormal object. Likewise, a "person" who is taller than common adults is also annotated. Even though an abnormal dataset has already been established [1, 5], the set is unsuitable for a thorough evaluation of all three types of abnormal objects because the set has not been separated into each type.

Two datasets, one normal and one abnormal, are used for training and test. The SUN dataset [20, 1] is used for the first dataset which has scene-level annotations as well as many annotated objects in a single image. Therefore, the SUN is the proper dataset for training and testing the relationships among objects. The second dataset is the proposed abnormal dataset. Object detector models [16] and the parameters of the proposed model are trained on randomly separated images from the SUN dataset. Evaluation is conducted on the test set of the SUN dataset and the entire abnormal dataset. Two scene categories and ten object categories[2] are used to train and test the proposed model. The parameters for the proposed model are $\alpha = -\frac{\pi}{9}\, \beta = \frac{5\pi}{36}, t_1 = -0.02, t_2 = -0.004$ and $\Lambda_o = 10^{-128} \cdot I$. Note that optimizing eq. (3) takes about two minutes using an Intel Quad Core 3.3GHz PC platform with 8GB of memory.

We choose a hybrid model of co-occurrence contextual (CO) and support contextual (SUP) model [5] as the baseline method for abnormal object detection, because our proposed model simultaneously detects abnormal objects that violate co-occurrence context and position/scale context. The number of objects used in the baseline is 10+2 objects, including support objects such as "floor" and "road." These supporting objects, useless for the proposed model, are positively necessary for the baseline method because of their dependency on supporting objects. This baseline method is quantitatively (Fig. 3) and qualitatively (Fig. 4) compared with the proposed method.

[1] L. Wei [24] has copyright on several abnormal images.

[2] indoor, outdoor / bed, bottle, building, car, monitor, person, sky, sofa, toilet, tree.



(a) co-occurrence

(b) position

(c) scale

(d) abnormal

(e) average precision

%	CS	CO+SUP
100	**15.21** ± 0.91	9.71 ± 0.05
50	**9.71** ± 0.53	4.55 ± 0.10
10	**5.64** ± 0.91	2.53 ± 0.19
5	**7.32** ± 2.50	3.38 ± 0.11

Fig. 3. Quantitative comparisons. Each precision-recall curve ((a) to (d)) is generated using classified abnormal datasets and a combination of the three classified datasets and the extra abnormal images. Experiment (e) is conducted on mixtures of abnormal and normal dataset. Each mixtures consists of a set of 150 images composed of 100%, 50%, 10%, and 5% abnormal images.

6.2 Results

Precision-recall curves and average precision measures are used to evaluate abnormal object detection accuracy. These methods are conventional measures [25] for object detection tasks.

The average precision in Fig. 3(e) shows that the proposed method outperforms the baseline method on both the abnormal and mixed datasets. Our method is robust in all three types of dataset, particularly in the position dataset Fig. 3(b), as shown by the analysis of precision-recall curve (Figs. 3(a) to 3(c)). The reason is that the proposed Canonical Scene method is more robust on separating floating objects than the baseline method. Fig. 3(d) shows that our method also outperforms in the combination of each type of dataset and extra abnormal images. Fig. 3(e) shows average precision scores on 150 images with varying ratios of abnormal and normal images. If we assume that a general image set is composed of 5% of abnormal images, then this experiment verifies the robustness of the proposed method on a general image set. The proposed method may also be superior on the general image set based on the experiment

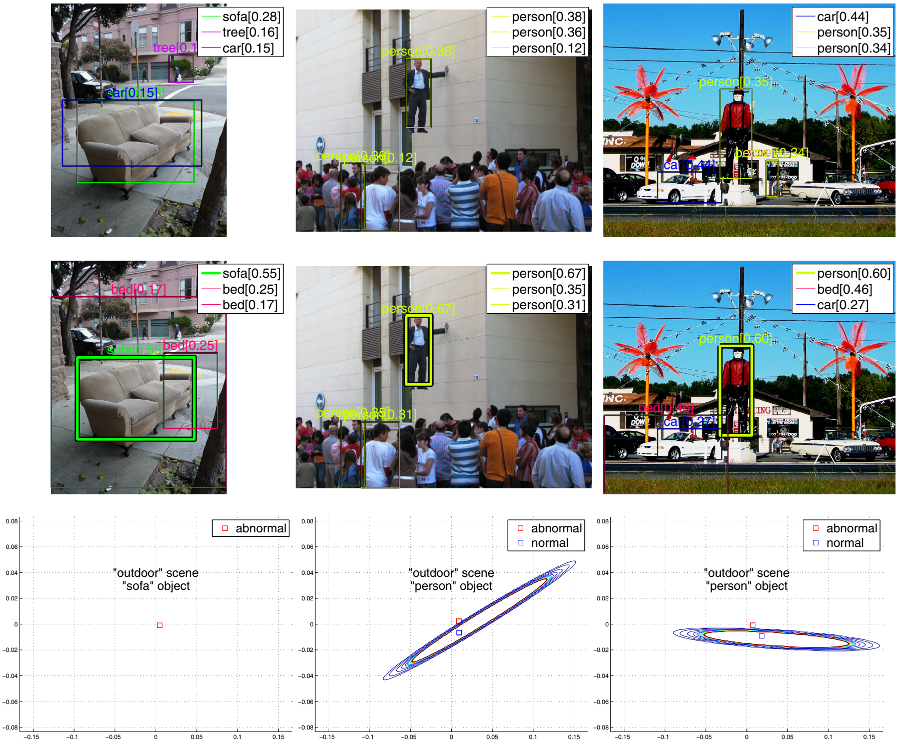


Fig. 4. Qualitative comparisons. The results in the top and the middle-rows are generated using the baseline method and the proposed method, respectively. In each images, the top-three abnormal objects are represented in bounding boxes. In the third row, the distribution of objects' locations in the Canonical Scene is represented with estimated abnormalities.

results. Note that in Fig. 3(c) shows that the proposed method is weak at classifying abnormal objects with abnormally huge or small sizes. Improperly learned Canonical Scenes may cause this weakness.

The qualitative comparisons and results are shown in Figs. 4 and 5. Based on the last row of Fig. 4, we can conclude that the proposed method exploits the rich relations among objects and instance of objects. For example, the second figure in the last row represents the distribution of a normal "person" in an "outdoor" scene. The learned normal distribution of "person" is transformed to cover "person" on the road. Therefore, the transformed distribution, or Canonical Scene, cannot cover the floating "person", thus making it possible to classify the floating "person" as an abnormal object. Fig. 5 illustrates the positive and negative results of the proposed method. The negative results are due to the improperly learned Canonical Scene, the experimentally set parameters, and the weakness of the base object detectors.

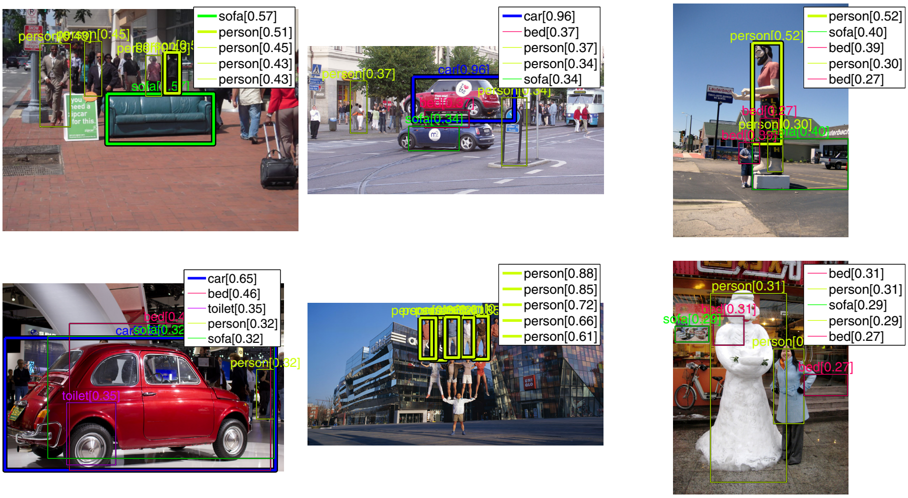


Fig. 5. Qualitative results. The images in the first, second and third columns consist of co-occurrence-violating, relative position-violating, and relative scale-violating images, respectively. In each image, the top-five abnormal objects are represented in bounding boxes.

7 Conclusion

We proposed a generative model for abnormal object detection. The model mainly exploited the rich and quantitative relations among objects and objects' instances via latent variables. Because of these considerations, our model outperformed the state-of-the-art method for abnormal object detection. In addition, we were able to thoroughly analyze the accuracy of the proposed method for different types of abnormality by classifying evaluation dataset into three types. The analysis revealed that our model is strong at detecting abnormal co-occurrence and position, but not as effective at detecting scale-violating objects. We expect that accuracy will be increased by fully learning of the proposed model.

References

1. Choi, M., Lim, J., Torralba, A., Willsky, A.: Exploiting hierarchical context on a large database of object categories. In: Proc. of IEEE Conf. on CVPR (2010)
2. Murphy, K., Torralba, A., Freeman, W.: Using the forest to see the trees: a graphical model relating features, objects and scenes. In: NIPS (2003)
3. Torralba, A., Murphy, K., Freeman, W.: Contextual models for object detection using boosted random fields. In: NIPS (2005)
4. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: Proc. of ICCV (2009)
5. Choi, M., Torralba, A., Willsky, A.: Context models and out-of-context objects. Pattern Recognition Letters 33, 853–862 (2012)

6. Heitz, G., Koller, D.: Learning Spatial Context: Using Stuff to Find Things. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 30–43. Springer, Heidelberg (2008)
7. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: Proc. of IEEE Conf. on CVPR (2008)
8. Jasra, A., Stephens, D., Holmes, C.: On population-based simulation for static inference. Statistics and Computing 17, 263–279 (2007)
9. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Proc. of ICCV (2007)
10. Li, L., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: Proc. of IEEE Conf. on CVPR (2009)
11. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph Cut Based Inference with Co-occurrence Statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
12. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Computer Vision and Image Understanding 114, 712–722 (2010)
13. Russell, B., Torralba, A., Liu, C., Fergus, R., Freeman, W.: Object recognition by scene alignment. In: NIPS (2007)
14. Torralba, A.: Contextual priming for object detection. IJCV 53, 169–191 (2003)
15. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: NIPS (2010)
16. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on PAMI 32, 1627–1645 (2010)
17. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. In: Proc. of IEEE Conf. on CVPR (2006)
18. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: Proc. of IEEE Conf. on CVPR (2010)
19. Olson, C.: Maximum-likelihood image matching. IEEE Trans. on PAMI 24, 853–857 (2002)
20. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proc. of IEEE Conf. on CVPR (2010)
21. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: a database and web-based tool for image annotation. IJCV 77, 157–173 (2008)
22. Van Hulse, J., Khoshgoftaar, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: ICML (2007)
23. Kim, W., Park, J., Lee, K.: Stereo matching using population-based mcmc. IJCV 83, 195–209 (2009)
24. Wei, L. (2012), `http://www.liweiart.com`
25. Everingham, M., Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV 88, 303–338 (2010)