What's Wrong with that Object? Identifying Images of Unusual Objects by Modelling the Detection Score Distribution*

Peng Wang†, Lingqiao Liu‡, Chunhua Shen‡, Zi Huang†, Anton van den Hengel‡, Heng Tao Shen†
† The University of Queensland, Australia ‡ The University of Adelaide, Australia

Abstract

This paper studies the challenging problem of identifying unusual instances of known objects in images within an "open world" setting. That is, we aim to find objects that are members of a known class, but which are not typical of that class. Thus the "unusual object" should be distinguished from both the "regular object" and the "other objects". Such unusual objects may be of interest in many applications such as surveillance or quality control. We propose to identify unusual objects by inspecting the distribution of object detection scores at multiple image regions. The key observation motivating our approach is that "regular object" images, "unusual object" images and "other objects" images exhibit different region-level scores in terms of both the score values and the spatial distributions. To model these distributions we propose to use Gaussian Processes (GP) to construct two separate generative models, one for the "regular object" and the other for the "other objects". More specifically, we design a new covariance function to simultaneously model the detection score at a single location and the score dependencies between multiple regions. We demonstrate that the proposed approach outperforms comparable methods on a new large dataset constructed for the purpose.

1. Introduction

Humans have an innate ability to detect an unusual object, even when they have no experience of the particular manner in which it is unusual. Mimicking this ability in computer vision has a range of applications such as surveillance or quality control. Existing studies towards this goal are usually conducted on small datasets and controlled scenarios i.e., with relatively simple backgrounds [2] or specific type of unusualness [4, 11]. To address this issue, in this work we present a large dataset which captures more general forms of unusualness, and has more complex backgrounds. Moreover, we adopt a more realistic "open world" evaluation protocol. That is, we need to distinguish the unusual version of an object-of-interest not only from typical examples from the same category but also from objects from other categories.

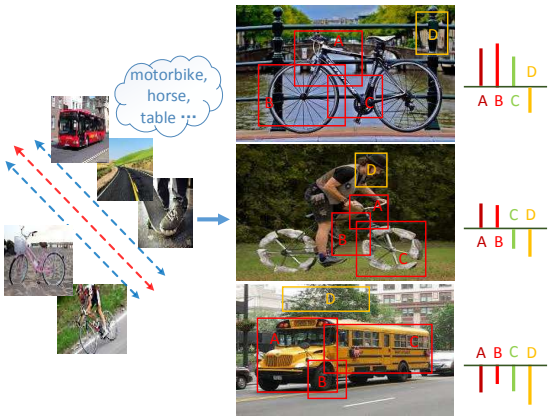

Figure 1. Illustration of the method applied to identifying the unusual "bicycle". By applying a detector trained on "regular bicycles" and "other objects", we are able to identify the "regular bicycle", the "unusual bicycle" and the "other object" (a bus in this case) through analysis of the distribution of the scores of multiple detectors. The discriminative information lies in both the values of the detection scores and the spatial dependencies between those scores, e.g. the score dependency between neighbouring proposals B and C.

Humans recognise unusual objects by identifying instances which share the key characteristics of the class, but not all of the typical incidental characteristics. Images of unusual objects are thus expected to be more similar to those of other instances of the same class of objects, than to those of other objects. Suppose we apply a detector trained using images of typical examples of a class of object as the positive data, and images of other objects as the negative data. The detection scores of the "unusual object" images are expected to be less than those of the typical objects of the same class, but greater than those of objects from other

*The first two authors contributed to this work equally. P. Wang's contribution was made when visiting The University of Adelaide. C. Shen is the corresponding author (e-mail: chunhua.shen@adelaide.edu.au). This work was partially supported by the Data 2 Decsions CRC.

classes. Empirically, however, we have seen that the detection score is insufficient to distinguish unusual objects from regular ones and other objects. We thus propose here not only to exploit the detection score values, but also the spatial distributions of the detection scores.

As is illustrated in Fig. 1, positive detection scores should densely overlap in images of regular object instances, while in unusual-object images the score distribution will be altered by the existence of unusual parts. To model these two factors, we propose to use Gaussian Processes (GP) [13] to construct two separate generative models for the detection scores of "regular object" image regions and "other objects" image regions. The mean function is defined to depict the prior information of the score values of either "regular object" images or "other objects" images. A new covariance function is designed to both non-parametrically model the detection score at a single region, and capture the inter-dependencies between scores over multiple regions. Note that unlike the conventional use of GP in computer vision, our model does not assume that the region scores of an image are i.i.d. This treatment allows our method to capture the spatial dependencies between detection scores, which turns out to be crucial for identifying unusual objects.

By comparing with several alternative solutions on the proposed dataset, we experimentally demonstrate the effectiveness of the proposed method. To summarize, the main contributions of this paper are:

- We propose a large dataset and present a more realistic "open world" evaluation protocol for the task of unusual-object identification from images.
- We propose a novel approach for unusual-object detection by looking into the detection score values as well as the spatial distributions of the detection scores of the image regions. We propose to use Gaussian Processes (GP) to simultaneously model the detection score at a single region and the score dependencies between multiple regions.

1.1. Related Work

Irregular Image/Video Detection. There exists a variety of work focusing on irregular image and/or video detection. While some approaches attempt to detect irregular image parts or video segments given a regular database [21, 2, 22, 7], other efforts are dedicated to addressing some specific types of irregularities [11, 4] such as out-of-context via building some corresponding models.

Standard approaches for irregularity detection are based on the idea of evaluating the dissimilarity from regular. The authors of [22, 7] formulate the problem of unusual activity detection in video into a clustering problem where unusual activities are identified as the clusters with low inter-cluster similarity. The work [2] detects the irregularities in image or video by checking whether the image regions or video segments can be composed using large continuous chunks of data from the regular database. Despite the good performance in irregularity detection, this method severely suffers from the scalability issue, because it requires to traverse the database given any new query data. Sparse coding [9] is employed in [21] for unusual events detection. This work is based on the assumption that unusual events cannot be well reconstructed by a set of bases learned from usual events.

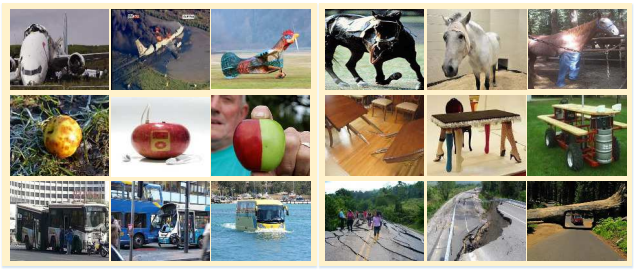

Figure 2. Examples of irregular images. Left column: aeroplane, apple, bus. Right column: horse, dining table, road.

Another stream of work focus on addressing specific types of irregularities. The work of [3, 4] focus on exploiting contextual information for object recognition or out-of-context detection, like "car floating in the sky". In [3], they use a tree model to learn dependencies among object categories and in [4] they extend it by integrating different sources of contextual information into a graph model. The work [11] focuses on finding abnormal objects in given scenes. They consider wider range of irregular objects like those violate co-occurrence with surrounding objects or violate expected scale. However, the applications of these methods are very limited since they rely on pre-learned object detector to accurately localize the object-of-interest. Recently the work in [14] delves into various types of atypicalities and makes a more comprehensive study.

Gaussian Processes in Computer Vision. Due to the advantage in nonparametric data fitting, GP has widely been used in the fields like classification [1], tracking [17], motion analysis [8] and object detection [19, 20]. The work [8] uses GP regression to build spatio-temporal flow to model the motion trajectories for trajectory matching. In [19, 20], object localization is done via using GP regression to predict the overlaps between image windows and the ground-truth objects from the window-level representations.

2. A New Dataset

2.1. Dataset Description

Here we propose a new dataset for the task of irregular image detection. The data is collected from *Google Images*

Table 1. Comparison of the proposed dataset with existing datasets. The work of [4] addresses the irregular type of *out of context*. The work of [11] deals with violations of *co-occurrence, positional relationship* and *scale.*

dataset	# images	irregular category	accurate detector
[11]	150	specific	yes
[4]	218	specific	yes
ours	20,420	general	no

and *Bing Images* which is composed of 20,420 images belonging to 20 classes. We choose the 20 classes referring to the PASCAL VOC dataset [5] but replace some classes that are not suitable for the task. The images of each class are composed of both regular images and irregular images. For regular images, we try different feasible queries to collect sufficient data. Taking "apple" for example, we try "fuji apple", "pink lady", "golden delicious", etc. To collect irregular images, we use keywords like "irregular", "unusual", "abnormal", "weird", "broken", "decayed", "rare", etc. After the images are returned, we manually remove the unrelated and low-quality data. Also, we perform near-duplicate detection to remove some duplicate images. In general, the number of irregular images per class is comparable to the sum of regular and "other class" images. Fig. 2 shows some examples of irregular images.

There exist some other datasets [4, 11] for irregular image detection. A comparison between our dataset and the existing datasets is summarized in Table 1. The main difference is twofold.

- Our dataset is large-scale comparing to the existing datasets, increasing the number of images from several hundred to more than twenty thousand.
- While the existing datasets are proposed for specific irregular category such as "out-of-context", "relative position violation" and "relative scale violation", our dataset is for general irregular cases.

Besides the above differences, we adopt a more practical evaluation protocol compared with [4, 11]. That is, we evaluate the irregular object detection with the presence of irrelevant objects. This is different from [2] where irregularity detection is performed in controlled environment with relatively simple background.

2.2. Problem Definition

For a given object category $\mathcal{C}$, we divide it into two disjoint subcategories, a regular sub-class $\mathcal{C}^r$ and an irregular sub-class $\mathcal{C}^u$, with $\mathcal{C} = \mathcal{C}^r \cup \mathcal{C}^u$ and $\mathcal{C}^r \cap \mathcal{C}^u = \emptyset$. We call an image I a regular image if $I \in \mathcal{C}^r$ and an irregular image if $I \in \mathcal{C}^u$. If an image I does not contain the given object, we label it as belonging to the "other class" set $\mathcal{C}^o$. The task is to determine if a test image $I \in \mathcal{C}^u$. Note that for $\mathcal{C}$, only the regular and "other class" images are available for training.

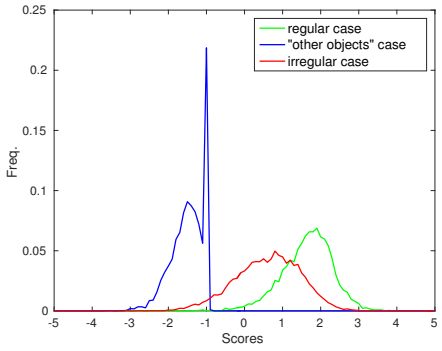

Figure 3. Histograms of decision scores for regular images, irregular images and "other class" images in the testing data. The decision scores are obtained by applying the classifiers learned from global images.

3. Key Motivation

Regular object images of the same class are alike; each irregular object image, however, is irregular in its own way. Thus, it is somehow impossible to collect a dataset to cover the space of the irregular images and one common idea to handle this difficulty is to build a "regular object" model to identify the "irregular objects" as outliers. While most traditional methods [21, 2] build this model based on the visual features extracted from images, our approach takes an alternative methodology by firstly training a detector from the "regular object" images and "other objects" images and then discovering the irregularity based on the detection score patterns. The merit of using detection scores for irregularity detection are as follows. (1) It is more computationally efficient since the appearance information has been compressed to a single scalar of detection values. This enables us to explore complex interaction of multiple regions within an image while maintaining reasonable computational cost. (2) It naturally handles the background and "other class" distraction since our detector is trained by using the "regular object" and "other objects". More specifically, our method is inspired by two intuitive postulates of how humans recognize an "irregular object", which are elaborated as follows.

Postulate I: discrimination in detection score values. From the perspective of human vision, an irregular object is something "looks like an object-of-interest, but is still different from its common appearance". If we view the object detection score as a measure of the likelihood of an image containing the object, then the above postulate could correspond to a relationship in detection scores $f(I^o) < f(I^u) < f(I^r)$, where $f(I^o)$, $f(I^u)$ and $f(I^r)$

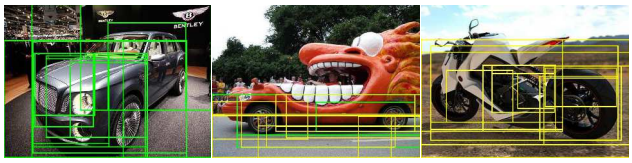

Figure 4. Visualization of spatial distribution of detection scores for test images of *car* class. Top-20 scored bounding boxes of an image are visualized. Positive proposals are visualized in green box and negative are visualized in yellow. From left to right: regular car, irregular car and other object (motorbike).

denote the detection score of the "other object", "irregular object" and "regular object" respectively. To verify this relationship, we train an image-level object classifier and plot the accumulated histograms of the scores of regular, irregular and other-class images of each class in Fig. 3. It can be seen from this figure that the distribution of the score values is generally consistent with our assumption. However, there are still overlaps especially between regular and irregular images, which means that using this criterion alone cannot perfectly distinguish the irregular images.

Postulate II: discrimination in the spatial dependency of detection scores. When exposed to part of the regular object, human can predict what the neighbouring parts of the object should look like without any difficulty. But irregular object may break this smoothness. This suggests that if we apply an object detector to the object proposals of an image, the region-level detection scores of the three different types of images may exhibit different dependency patterns. Fig. 4 shows the top 20 regions of some example images of *car* class according to the values of the detection scores. As seen, for regular car the positive bounding boxes are densely overlapped and images from other classes such as *motorbike* are supposed to have no positively scored proposals. Detection scores of irregular images may disobey both of these two distribution patterns. For example, two strongly overlapped regions may have opposite detection scores.

4. Proposed Approach

Motivated by the above analysis, we propose a two-step approach to the task of irregular image detection. We first apply a Multi-Instance Learning (MIL) approach to learn a region-level object detector and then design Gaussian Processes (GP) based generative models to model the detection score distributions of the "regular object" and the "other objects". Once the model parameters are learned, we can readily determine whether a test image is irregular by evaluating its fitting possibilities to these two generative models.

4.1. Object Detector Learning

Taking the region proposals of images as instances, we represent each image as a bag of instances. Since we only have the image-level label indicating the presence or absence of the object, the learning of region-level detector is essentially a weakly supervised object localization problem. Considering both the localization accuracy and the scalability, we follow the MIL method in [10] to learn an object detector for each class. For a class C, we have a set of regular images containing the object as positive training data and a set of images belonging to other classes where the object concerned do not appear as negative training data.

We use Selective Search [16] to extract a set of object proposals for each image and from the perspective of MIL, each proposal is regarded as an instance. Then each image I^i is represented by a $N_i \times D$ matrix $\mathbf{X}^i$ where N_i denotes the number of proposals and D represents the dimensionality of the proposal representations. Inspired by [10], we optimize the following objective function to learn the detector,

$$\mathcal{J} = \sum_i \log(1 + e^{-y^i \max_j\{\mathbf{w}^\mathbf{T}\mathbf{x}_j^i + b\}}), \qquad (1)$$

where $\mathbf{w} \in \mathbb{R}^{D\times 1}$ serves as an object detector, $\mathbf{x}_j^i$ indicates the jth instance of the ith image and $\mathbf{w}^T\mathbf{x}_j^i + b$ is its detection score. The single image-level score is aggregated via the max-pooling operator $\max\{\cdot\}$ and it should be consistent with the image-level class label $y^i \in \{1, -1\}$. The parameters $\mathbf{w}$ and b can be learned via back-propagation using stochastic gradient descent (SGD).

4.2. Gaussian Processes Based Generative Models

In this section, we elaborate how to use GP to model the distribution of the region-level detection scores. Unlike traditional GP based regression [20] which takes a single feature vector as input, we treat multiple proposals within an image as the input and our model will return a probability to indicate the fitting likelihood of the proposal set.

GP assumes that any finite number of random variables drawn from the GP follow a joint Gaussian distribution and this distribution is fully characterized by a mean function $m(x)$ and a covariance function $k(x, x')$ [13]. In our case, we treat the detection score of each proposal as a random variable. The mean function depicts the prior information of the score values, e.g. the value tends to be a positive scalar for the "regular object" images. The covariance function plays two roles. (1) As in standard GP regression, it serves as a non-parametric estimator of the score value. More specifically, if a proposal is similar (in terms of a defined proposal representation) to a proposal in the training set, it encourages them to share similar scores. (2) As one of our contributions, we also add a term in the covariance function to encourage the overlapped object proposals within the

same test image to share similar detection scores. In the following subsections, we introduce the details of the design of the mean function and covariance function.

4.2.1 GP Construction

For each class $\mathcal{C}$, we will construct two GP based generative models for regular images and "other objects" images separately. Without losing generality, we will focus on regular images in the following part.

Suppose that we have $N^{\mathcal{C}}$ positive training images for class $\mathcal{C}$. For each image I^i ($i \in \{1, 2, \cdots, N^{\mathcal{C}}\}$), we use the top-$n$ scored proposals s^i_j ($j \in \{1, 2, \cdots, n\}$) only in order to reduce the distraction impact of the background. Their associated detection scores can be obtained via the function $f(s^i_j)$. In our model we assume that f is distributed as a GP with a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$

$$f \sim \mathcal{GP}(m, k). \tag{2}$$

Mean function: We define the mean function $m(s) = \mu$, where μ is a scalar constant learned through parameter estimation. It can be intuitively understood as the bias of the detection score in the regular object or other object cases. For example, it tends to be a positive (negative) value for the "regular (other) object" case.

Covariance function: As aforementioned analysis, the covariance function is decomposed into two parts, an inter-image part and an inner-image part. While the inter-image part is employed to regress the proposal-level detection score in the light of the proposals in the training set, the inner-image part is used to model the dependencies of the scores within one test image. To define the inter-image covariance function for a proposal pair belonging to different images, it needs to design a representation for each proposal so that their similarity can be readily measured. We leverage the spatial relationship between a proposal and the proposal with the maximum detection score within the same image as this representation. More specifically, assuming the maximum-scored proposal in an image I^i is s^i_{max}, the representation of a proposal s in I^i is defined as,

$$\phi(s) = [\mathrm{IoU}(s, s^i_{max}), c(s, s^i_{max})], \tag{3}$$

where $\mathrm{IoU}(s, s^i_{max})$ denotes the intersection-over-union between s and s^i_{max} and $c(s, s^i_{max})$ denotes the normalized distances between the centers of s and s^i_{max}. Note that these two measurements reflect a proposal's overlapping degree, distance to the maximum-scored proposal and indirectly the size of the proposal. Intuitively, these factors could be used to predict the detection score value of a proposal.

With this representation, we can define the inter-image covariance function $k_{inter}(s, s')$ of s and s' as,

$$\exp\Big(-\frac{1}{2}\big(\phi(s) - \phi(s')\big)^T \mathrm{diag}(\gamma)\big(\phi(s) - \phi(s')\big)\Big), \tag{4}$$

where $\mathrm{diag}(\gamma)$ is a diagonal weighting matrix to be learned.

The inner-image covariance function serves as one of the key contributions of this work, which poses a smoothness constraint over the scores of the overlapped object proposals in an image. For a pair of inner-image proposals s and s', we define the inner-image covariance function as follows (if two proposals s and s' are from different images, $k_{inner}(s, s') = 0$),

$$k_{inner}(s, s') = \frac{2S(s \cap s')}{S(s \cap s') + S(s \cup s')}, \tag{5}$$

where S stands for the area. Note that the formula is variant to standard intersection-over-union [5] commonly used as detection metric. The reason why we define it like this is because it is exactly χ^2 kernel and can guarantee the covariance matrix to be positive definite [18].

With both the inter-image and inner-image covariance function, we can obtain the overall covariance function of any proposal pair s and s' as,

$$k(s, s') = a \cdot k_{inner}(s, s') + b \cdot k_{inter}(s, s'), \tag{6}$$

where a, b are hyper-parameters regulating the weights of these two kernel functions.

4.2.2 Hyper-parameter Estimation

In this part, we introduce the hyper-parameter learning for the GPs. Still, we use regular images for description. In the definition of the mean and covariance functions of the GP, we introduce the hyper-parameters $\theta = \{\mu, \gamma, a, b\}$. We estimate the hyper-parameters by minimizing the negative logarithm of the marginal likelihood of all the detection scores of the training proposals given the hyper-parameters,

$$-L = -\log p(f(\mathcal{S})|\mathcal{S}, \theta), \tag{7}$$

where $\mathcal{S}$ denotes the training proposals and $f(\mathcal{S})$ denotes their detection scores. We use the toolbox introduced in [12] for hyper-parameter optimization.

4.2.3 Test Image Evaluation

For class $\mathcal{C}$, let $\mathbf{s_r}$ be a set of proposals of regular training images and $\mathbf{f_r}$ be their detection scores. We can establish the covariance matrix K for the training data. Given a target set of proposals $\mathbf{s_t}$ from a test image and their detection scores $\mathbf{f_t}$, the joint distribution of $\mathbf{f_r}, \mathbf{f_t}$ can be written as,

$$\begin{bmatrix}\mathbf{f_r}\\ \mathbf{f_t}\end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}K & k(\mathbf{s_r}, \mathbf{s_t})\\ k(\mathbf{s_r}, \mathbf{s_t})^T & k(\mathbf{s_t}, \mathbf{s_t})\end{bmatrix}\Big), \tag{8}$$

where μ is the mean vector, $k(\mathbf{s_r}, \mathbf{s_t})$ calculates the inter-image covariance matrix between training set and testing

set and $k(\mathbf{s_t}, \mathbf{s_t})$ calculates the inner-image covariance of the test data. The fitting likelihood of the testing set to the generative model of the regular images can be expressed as,

$$\mathbf{f_t}|\mathbf{f_r} \sim \mathcal{N}\Big(\mu + k(\mathbf{s_r}, \mathbf{s_t})^T K^{-1}(\mathbf{f_r} - \mu), \\ k(\mathbf{s_t}, \mathbf{s_t}) - k(\mathbf{s_r}, \mathbf{s_t})^T K^{-1} k(\mathbf{s_r}, \mathbf{s_t})\Big). \tag{9}$$

Similarly, we can obtain the likelihood of the testing set given the "other class" training set. After obtaining the likelihood of the testing set given both regular training data and "other class" training data, we can compute the logarithm of the overall fitting likelihood of $\mathbf{f_t}$ as

$$\max\big(\log p(\mathbf{f_t}|\mathbf{f_r}), \log p(\mathbf{f_t}|\mathbf{f_o})\big), \tag{10}$$

where $\mathbf{f_o}$ represents the scores of "other class" training set. For either regular or "other class" test images, they could fit one of the generative models better than the irregular images. In other words, irregular images are supposed to obtain lower values in Eq. (10). Since the score obtained from Eq. (10) is negative (logarithm of a probability), we use the negative value of the score as the irregularity measurement.

5. Experiments

5.1. Experimental Settings

In this paper, we use the pre-trained CNN model [15] as feature extractors for object detector learning. Specifically, we use the activations of both the second fully-connected layer and the last convolutional layer as the representation of the object proposal or the whole image. Feeding an image into the CNN model, the activations of a convolutional layer are $n \times m \times d$ (e.g., $14 \times 14 \times 512$ for the last convolutional layer) with n, m corresponding to different spatial locations and d the number of feature maps. Given a proposal, we aggregate the convolutional features covered by it via max pooling to obtain the proposal-level convolutional features. We perform $L2$ normalization to these two types of features separately and concatenate them as the final representation. The dimensionality of the features is 4,608.

For each class, we construct GP based generative models for regular images and "other class" images separately. For regular images, we initialize the value of the mean function as 3 and for "other class" images we set the initial value to be -3. The hyper-parameters a, b in Eq. (6) are both initialized to be 0.5 and γ is initialized randomly. We use the top-20 scored proposals of each image for both generative model construction and test image evaluation. The test data of each class is divided into three parts including regular images, irregular images and images belonging to other classes. We label irregular images as 1 and label regular and "other class" images as -1. Mean Average Precision (mAP) is employed to evaluate the performances of the approaches.

5.2. Experimental Results

5.2.1 Alternative Solutions

We compare our method to the following methods.

Positive-negative Ratio If we apply an object detector to the image regions, considerable portion of the regions of a regular image should be positively scored. While on the contrary, images of other classes are supposed to have negatively-scored proposals only. Based on this intuitive assumption, we use the ratio of positive proposal number to the number of negative proposals within one image as its representation to construct two Gaussian models for regular images and "other class" images separately. Given a test image, we determine whether it is irregular via evaluating its fitting degree to these two Gaussians.

Global SVM According to the analysis in **Postulate I** in Section 3, the classification score of an image reflects the degree of containing the regular object-of-interest and the scores of the three types of images (regular, irregular, other class) should form the relationship of $f(I^o) < f(I^u) < f(I^r)$. For this method, we train a classifier for each class based on the global features of the images using linear SVM [6] where regular images are used as positive data and "other class" images are treated as negative data. Assuming the mean of the decision scores of irregular images is 0, we use negative absolute value of the decision score $-|f(I^t)|$ as the irregularity measurement for a test image I^t.

MIL + Max The global representation of an image is a mixture of the patterns of both the object-of-interest and the background. To avoid the distraction influence of the background, for the second solution we use the maximum proposal-level score $f_{max}(I^t)$ as the decision score of each image based on the object detector learned from MIL. Similarly we use $-|f_{max}(I^t)|$ as the irregularity measurement.

MIL + Max + Gaussian Different from above **MIL + Max** strategy, we take into consideration the uncertainty of the distribution of the maximum detection scores via modelling the maximum scores of regular images $\mathbf{I^r}$ and "other class" images $\mathbf{I^o}$ using two Gaussian distributions separately. We use maximum likelihood to estimate the parameters of these two Gaussians (means and variances). Given a test image I^t, we can calculate the likelihood of the image belonging to regular images as $p(I^t|\mathbf{I^r})$ and similarly the possibility of belonging to other classes as $p(I^t|\mathbf{I^u})$. Since an irregular image is expected to be able to fit neither of these two models, we set the final score of a test image as $-\max(p(I^t|\mathbf{I^r}), p(I^t|\mathbf{I^u}))$.

MIL + Top k Instead of using the maximum score only, for this method, we obtain the image-level score $f_{topk}(I^t)$ of a test image I^t by averaging the top k scores of its proposals. And the final score for an image is $-|f_{topk}(I^t)|$.

Table 2. Experimental results. Average precision for each class and mAP are reported.

Methods	aeroplane	apple	bicycle	boat	building	bus	car	chair	cow	dinging table	
Positive-negative Ratio	58.0	26.6	50.4	52.4	60.0	37.8	55.4	48.7	31.6	28.8	
Global SVM	88.8	70.8	81.3	82.9	85.5	76.4	87.6	69.7	61.7	79.8	
MIL + Max	86.9	70.0	85.0	78.8	81.7	77.6	87.8	70.5	63.9	76.4	
MIL + Max + Gaussian	86.0	72.1	83.1	78.5	74.5	76.3	83.2	59.3	56.7	68.4	
MIL + Top 20	86.7	78.3	86.6	86.9	79.6	75.2	86.5	64.0	63.8	56.8	
Sparse coding (200)	86.9	48.6	80.6	81.0	82.8	57.4	82.8	71.7	56.1	72.2	
Sparse coding (4,000)	93.6	74.5	89.8	86.7	94.5	86.1	92.8	78.7	76.8	**86.0**	
Ours	**95.4**	**82.2**	**91.2**	**93.0**	**94.6**	**92.8**	**95.1**	**92.8**	**92.0**	74.8	
Methods	horse	house	motorbike	road	shoes	sofa	street	table lamp	train	tree	**mAP**
Positive-negative Ratio	23.9	47.4	30.9	48.2	56.4	39.7	42.7	16.9	28.6	44.7	41.4
Global SVM	73.3	82.0	75.6	81.3	88.2	77.7	73.8	66.5	69.2	73.9	77.3
MIL + Max	70.3	80.0	74.8	78.1	87.7	76.4	69.1	65.1	67.3	77.0	76.3
MIL + Max + Gaussian	63.1	74.6	65.9	66.1	85.8	69.7	55.5	60.5	64.1	69.8	70.7
MIL + Top 20	63.7	76.4	76.9	73.6	90.3	69.7	63.7	52.3	67.2	75.2	73.7
Sparse coding (200)	61.5	71.3	61.0	80.1	82.3	80.2	84.1	52.3	65.5	57.6	70.8
Sparse coding (4,000)	80.0	89.3	75.5	89.9	87.2	87.7	91.1	67.9	81.9	78.9	84.4
Ours	**85.4**	**94.4**	**85.0**	**90.8**	**95.3**	**88.9**	**94.8**	**78.3**	**91.3**	**85.0**	**89.7**

Sparse coding Similar to [21], we use sparse coding based reconstruction error as the criterion for irregular image detection. The assumption is that both regular images and "other class" images can be well reconstructed by their corresponding dictionaries. For each class, we learn dictionaries for regular images and "other class" images separately. We try dictionary size 200, 4,000 and 5,000. Given a test image I^t, we infer the coding vectors of its proposals and calculate the reconstruction residues of the proposals. Let r_r^t be the mean residue for this image calculated based on the dictionary learned from regular images and r_o^t be the mean residue based on the dictionary learned from "other class" images. For an irregular image, the errors of both models will be large. Thus the irregularity measurement can be calculated as $\min(r_r^t, r_o^t)$.

5.2.2 Quantitative Results

Table 2 shows the quantitative results. As can be seen, our method outperforms other compared methods. Also we show the ROC performances of our method and two most competitive methods on some example categories in Fig. 5. Both these two measurements demonstrate the effectiveness of the proposed method.

The proposal ratio based method performs worst among these methods which indicates that the irregularity detection cannot be achieved by simply counting the number of positive and/or negative proposals. There are two reasons. The first is that the number of proposals varies between different images and the second reason is that for some irregular object images e.g., images of severely damaged cars, there may be no positively scored proposals detected.

The next four methods are classification-based methods. While the first three use single score per image from either the global image or the region with maximum detection score, **MIL+Top k** utilizes multiple region scores but treat them as i.i.d. **Global SVM** achieves a mAP of 77.3% (when using fully-connected features only, we obtain 75.4%) which to some extent justifies **Postulate I**. However, as illustrated in Fig. 3, this strategy fails to distinguish some irregular images that obtain extreme high or low decision scores. A drawback of using image-level representation is that the background can influence the decision score especially when the background dominates the image. Multi-instance learning is supposed to be a remedy because it makes it possible to focus on the object-of-interest via considering the proposal with maximum detection score. But using maximum detection score alone may risk missing the irregular part of the object. From Table 2, we can see **MIL+Max** obtains comparable results to **Global SVM**. To take into consideration the uncertainty of the detection scores, rather than directly using the maximum detection scores, we construct Guassian models for the maximum scores of regular images and "other class" images separately and determine whether an image is irregular via evaluating its fitting likelihood to these two Gaussian models. However, the performance degrades to 70.7%. The reason may be that the distribution of the maximum detection scores is not strictly Gaussian. Instead of using the maximum detection score of each image, in **MIL+Top20**, we aggregate the top 20 scores of each image via average pooling. Benefiting from this strategy, the performances on some classes like *apple*, *boat* are obviously boosted. However, on some other classes such as *horse*, *table lamp* it shows inferior performance to **Global SVM** and **MIL+Max**. As can be seen, our method significantly outperforms this strategy on all the classes. This big gap may to a large extent result from our capabilities of modelling the inter-dependencies of the proposal-level scores within one image.

For sparse coding, we first test the performance using

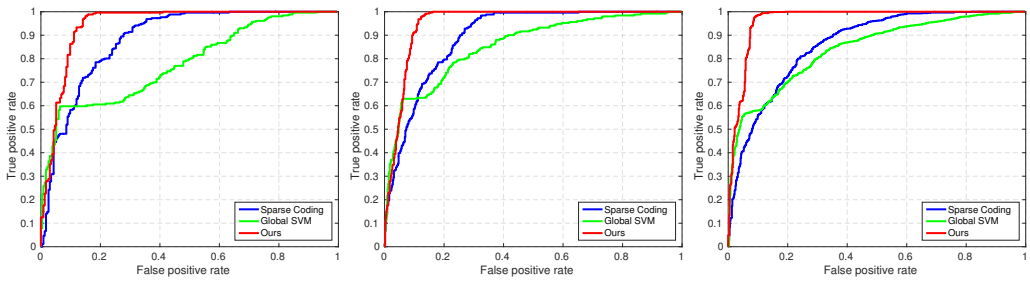

Figure 5. ROC curve for **Sparse coding, Global SVM** and **our method** on three categories. From left to right: boat, motorbike, shoes.

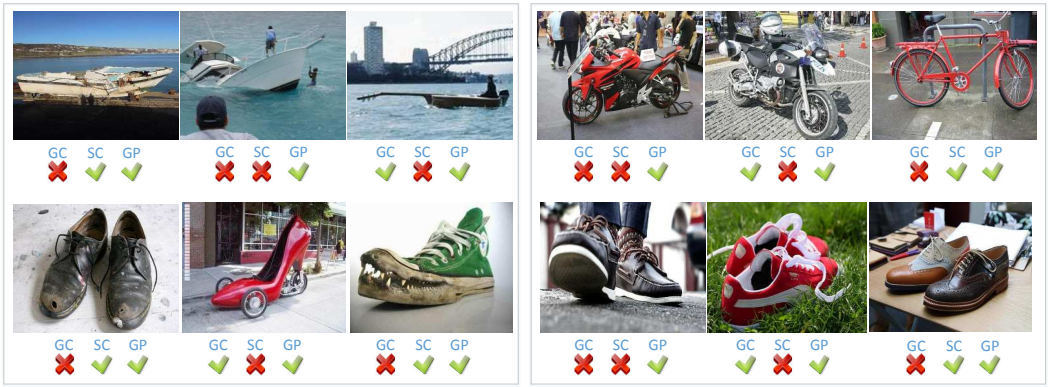

Figure 6. Qualitative performance comparison between our method (GP) and two alternative solutions, **Global SVM** (GC) and **Sparse coding** (SC). Left column displays the **false negative** examples when fixing the false positive rate to be 0.2 where cross mark indicates false negative and check mark indicates true positive. Right column displays the **false positive** examples when fixing the true positive rate to be 0.9 where cross mark denotes false positive and check mark denotes true negative. Three categories are boat, shoes and motorbike.

dictionaries of size 200 as [21] and the result is unsatisfactory which means 200 bases are not sufficient to cover the feature spaces of regular images or "other class" images. When the dictionary size is increased to 4,000, the performance is significantly improved. But after that continuing to increase the dictionary size (we test **5,000**) can lead to no improvement any more. Our method outperforms sparse coding by 5.3%. Apart from effectiveness, our method is also more efficient than sparse coding. Given a test image, while sparse coding needs to infer the coding vector for the high-dimensional appearance features our method works on quite low-dimensional space as defined in Eq. (3).

5.2.3 Qualitative Results

Fig. 6 demonstrates the qualitative comparison between our method and two compared methods **Global SVM (GC)** and **Sparse coding (SC)** on three object categories that are boat, motorbike and shoes. Comparing to our method, GC suffers from two drawbacks: 1) it subjects to the distraction influence of the background, and 2) it may ignore the fine details of the objects. Due to the influence of the background, GC may mistakenly classify the regular object within complex background into irregular object like the "shoes" on the right side of Fig. 6. Also, only looking at the global appearance makes it hard for GC to identify some irregular objects with fine irregularities such as the "broken boat" and "broken shoes" in Fig. 6. SC has similar deficiency that is it can be distracted or even dominated by the background. For example, the "capsized boat" is identified as "regular boat" while "regular motorbike" within complex background is regarded as "irregular motorbike". Comparing to these two methods our method is more robust. While using detection scores enables us to getting rid of the distraction influence of the background, modelling the inter-dependencies of the detection scores at multiple regions can help us to effectively discover the finer irregularities.

6. Conclusions

We have proposed a novel approach for the task of irregular object identification in an "open world" setting via inspecting the detection score patterns of an image. We propose to use Gaussian Processes to model the values as well the spatial distribution of the detection scores. It shows superior performance to some compared methods on a large dataset presented in this work.

References

[1] Y. Altun, T. Hofmann, and A. J. Smola. Gaussian process classification for segmenting and annotating sequences. In *ICML*, 2004.

[2] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 2007.

[3] M. J. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[4] M. J. Choi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012.

[5] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008.

[7] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: representing activities as bags of event n-grams. In *CVPR*, 2005.

[8] K. Kim, D. Lee, and I. Essa. Gaussian process regression flow for analysis of motion trajectories. In *ICCV*, 2011.

[9] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007.

[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[11] S. Park, W. Kim, and K. M. Lee. Abnormal object detection by canonical scene-based contextual model. In *ECCV*, 2012.

[12] C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (gpml) toolbox. *JMLR*, 2010.

[13] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. 2005.

[14] B. Saleh, A. Elgammal, J. Feldman, and A. Farhadi. Toward a taxonomy and computational models of abnormalities in images. In *AAAI*, 2016.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[16] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[17] R. Urtasun, D. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

[18] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *TPAMI*, 2011.

[19] A. Vezhnevets and V. Ferrari. Associative embeddings for large-scale knowledge transfer with self-assessment. In *CVPR*, 2014.

[20] A. Vezhnevets and V. Ferrari. Object localization in imagenet by looking out of the window. In *BMVC*, 2015.

[21] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 2011.

[22] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004.