Landmark Detection with Surprise Saliency Using Convolutional Neural Networks

Feng Tang  
*Fordham University*

Damian Lyons  
*Fordham University, dlyons@fordham.edu*

Daniel Leeds

Follow this and additional works at: https://fordham.bepress.com/frcv_facultypubs

Part of the [Artificial Intelligence and Robotics Commons](https://fordham.bepress.com/artificialintelligencerobotics_commons), [Graphics and Human Computer Interfaces Commons](https://fordham.bepress.com/graphics_and_human_computer_interfaces_commons), and the [Robotics Commons](https://fordham.bepress.com/robotics_commons)

Recommended Citation

Abstract—Landmarks can be used as reference to enable people or robots to localize themselves or to navigate in their environment. Automatic definition and extraction of appropriate landmarks from the environment has proven to be a challenging task when pre-defined landmarks are not present. We propose a novel computational model of automatic landmark detection from a single image without any pre-defined landmark database. The hypothesis is that if an object looks abnormal due to its atypical scene context (what we call surprise saliency), it then may be considered as a good landmark because it is unique and easy to spot by different viewers (or the same viewer at different times). We leverage state-of-the-art algorithms based on convolutional neural networks to recognize scenes and objects. For each detected object, a surprise saliency score, a fusion of scene and object information, is calculated to determine if it is a good landmark. In order to evaluate the performance of the proposed model, we collected a landmark image dataset which consists of landmark images, as we have defined them with surprise saliency above, and non-landmark images. The experimental results show that our model achieves good performance in automatic landmark detection and automatic landmark image classification.

I. INTRODUCTION

A landmark is an object in a scene which can be used as a navigational reference mark and which enables people or robots to orient themselves and to navigate in this environment. Landmarks play a key role in autonomous mobile robot navigation: Probabilistic mapping and localization [1], [2] allow a robot or team of robots to construct a map of the environment and to localize with respect to this map. Unique landmarks however are necessary for loop closure in such mapping. In topological mapping, unique landmarks are necessary to recognize when the robot has traversed from one place to an adjacent place within the topological map. A team of autonomous mobile robots [3] can use their recognition and measurement of a commonly agreed landmark to combine the maps they each generated. Finally unique landmarks can also play a role in human-robot communication by providing a common visual vocabulary for humans and robots. However, autonomous mobile robots need automatic discovery of these landmarks unless there is a set of pre-agreed upon landmarks (e.g., fiducial markings) in the environment.

Automatic definition and extraction of appropriate landmarks from the environment has proven to be a challenging task. Autonomous mobile robots using machine visual system can recognize pre-defined landmarks based on object detection algorithms developed by the computer vision community. However obvious landmarks (e.g. the Empire State Building in New York City) are not widespread in most operating environments. People and robots still need landmarks to navigate or localize themselves even if there is no obvious landmark in the current scene. Hence they need to find some objects in the scene that can be easily identified (and re-identified); these kinds of objects are ideal landmarks. For an autonomous mobile robot, many objects are visible in a given view of the environment, but only some of these objects are useful as landmarks. Our research addresses the automatic detection of landmarks from a single image problem by fusing information from two sources (scene and object) to obtain Surprise saliency: an object is salient because of its uniqueness and atypical context.

A. The Definition of Landmarks

Being a landmark is a relative property: An object being considered as a landmark depends not only on its own attributes but also on the distinction to the attributes of its environment. For example, a car is not a landmark if it is on the road. However a car appearing in the hall may be considered as a landmark because there is usually no car in the indoor environment.

Sorrows and Hirtle [4] stated the characteristics of a landmark including (1) singularity, or sharp contrast with its surroundings; (2) prominence of its spatial location; (3) its content, meaning, use, or cultural significance; and (4) its prototypicality that is, how typically it represents a category. According to these characteristics, they categorize landmarks into visual (certain visual characteristics), structural (prominence of their spatial location), and cognitive (semantically meaningful to personal interests or experiences) ones.

Raubal and Winter [5] followed the framework proposed by Sorrows and Hirtle [4] to present a formal model of landmark saliency including visual, semantic, and structural attraction which can be used to identify landmarks. Visual saliency is visual properties including facade area, shape, color, and visibility which make an object be easily noticed. Semantic saliency is the cultural and historical importance as well as function of a landmark. For example, in [6] a survey of the prominence of campus features has demonstrated that the single chapel on campus is more prominent than other buildings with multiple occurrences on campus, such as halls and bookstores. Structural saliency is that an object may be defined as a landmark because it plays an important role in the structure of the spatial environment.

*This research is supported by the Defense Threat Reduction Agency, Basic Research Award #HDTRA1-11-1-0038.

Feng Tang, Damian M. Lyons, and Daniel D. Leeds are with the Department of Computer and Information Science, Fordham University, Bronx, NY 10458, USA {ftang,dlyons,dleeds}@fordham.edu
For example, Lynch [7] stated that nodes, boundaries, and regions are the major elements that structure a city because of their individual structuring properties. As a result, these elements are considered as landmarks in a city.

B. Contribution

The main contribution of this work is a novel model of automatic landmark detection using surprise saliency based on convolutional neural networks (CNNs). We assume there are no pre-defined landmarks in the database. The hypothesis of this work is that if an object looks salient due to its atypical context, then it may be considered as a landmark because it is unique and easy to spot by different viewers (different robots or the same robot at a different time). Given an input image, our model will answer three questions: (1) Does the image contains landmarks? (2) Which categories do the landmarks belong to in the image? and (3) What are the positions of these landmarks in the image? To evaluate the performance of our model, we collected a landmark image dataset which consists of landmark images and non-landmark images.

II. RELATED WORK

A. Scene Classification

In our model, we need to classify a scene to know what objects are atypical for that scene. Scene classification deals with the problem of determining the scene category in which an image or photograph was taken. It is a challenging task because there are many factors (e.g., variability, ambiguity, illumination) which affect the classification accuracy. A typical approach has been to extract hand-crafted descriptors such as SIFT [8] and HOG [9], construct a global representation from these descriptors using bag-of-words [10] or sparse coding [11], then pool them into an image-level high-dimensional vector, employ some classification algorithm, and possibly use late fusion to combine the results of multiple features. An improved strategy is to use intermediate representations for classifying scenes (e.g., spatial envelope [12], patch codewords [13]).

Recently, the use of CNNs [14] for scene classification has gained tremendous attention because of its success in raising the accuracy of recognition. A CNN is a type of feed-forward artificial neural network which contains a number of layers. There are three main types of layers: convolutional layer, pooling layer, and fully-connected layer. Generally a CNN consists of several convolutional layers and pooling layers followed by at least one fully connected layer. Donahue et al. [15] trained a CNN with ImageNet data and applied it to scene recognition and object recognition. Koskela and Laaksonen [16] trained CNNs with external object-centric data and with different architectures, and used them as feature extractors in a standard visual recognition framework for scene recognition. Zhou et al. [17] built a large scale scene-centric database called MIT Places and trained AlexNet on it for scene recognition tasks.

B. Object Detection

Recognizing objects and localizing them in images is one of the most challenging problems in computer vision. A typical approach [10], [18] is to extract local image descriptors (e.g., SIFT, HOG), build a global representation of the descriptors using bag-of-words or sparse coding, and apply some standard classifiers. This approach uses the position of the visual words to locate the objects. However the performance of locating objects is not satisfied: (1) some visual words represents only parts of an entire object. (2) some visual words describes the background rather than the objects. An improved strategy is to detect the presence of objects by applying rich object models (e.g., deformable part models [19]). This strategy can obtain useful information (e.g., location, pose) of objects, but this approach is often trained on images with known locations of objects or even their parts.

CNNs have significantly improved object detection accuracy [20], [21]. The current state of the art for object detection is the Region-based Convolutional Neural Networks (R-CNN) method [22] and its successors. The general idea of R-CNN is: extract bottom-up region proposals from an input image using a proposal mechanism such as Selective Search [23]; compute features for each proposal using a CNN; classify each region using class-specific linear SVMs. Fast R-CNN [24] enabled end-to-end detector training on shared CNN features and has demonstrated better speed and accuracy than R-CNN. Faster R-CNN [25] is a Region Proposal Network method that shares full-image convolutional features with the CNN so that it achieves nearly cost-free region proposals.

C. Landmark and Abnormality Detection

Landmark detection and abnormality detection have the similarity that both of them are trying to find something that attracts the visual attention of observers. Landmark detection is related to objects and their surrounding context; detecting a landmark means finding something special in the scene in order to enable an agent (a person or a robot) to establish their location easily. Abnormality detection tries to find not only objects with abnormal positions in given scenes (e.g., a tiger in sky), but also objects with abnormal attributes (e.g., a cat has wings). In the latter case an object is considered an abnormality without consideration of its environment.

1) Landmark Detection: Ranganathan and Dellaert [26] proposed a hierarchical graphical model for the representation of a place and use this model and Bayesian surprise to perform landmark detection. Hayet et al. [27] used visual functions to extract and recognize quadrangular visual landmarks. Todt and Torras [28] presented an algorithm for natural landmark detection without an a-priori knowledge of the landmark characteristics. This leveraged the visual saliency concepts that a multi-scale opponent feature (color and orientation) method can be used to detect good landmark candidates in outdoor environments. Some of these works only consider visual saliency in an image which lacks in semantic information. Some works detect a landmark with
an a-priori knowledge of the landmark’s characteristics or the positions of objects. Our model will detect an entity as a landmark without any priori knowledge of the landmark’s characteristics or positions of objects.

2) Abnormality Detection: Saleh et al. [29] proposed a graphical model to detect abnormality by the attributes extracted from objects and they introduced a dataset of abnormal images for quantitative evaluation along with ground truth from human subjects. They are mainly focused on abnormalities of the object itself regardless of the scene context of the object. Park et al. [30] proposed a generative model to exploit the quantitative relations among objects by latent variables so as to detect an abnormal object. Saleh et al. [31] presented various types (object-centric, scene-centric, and contextual) of abnormality in images in a more comprehensive way and propose a computational model that can predict all different types of abnormality in images. They also introduce a new dataset of abnormal images showing a wide range of atypicalities. All this research is related to abnormality detection which is different from landmark detection (we are interested in context-dependence, as discussed previously).

III. COMPUTATIONAL MODEL

A. The Overview of Computational Model

Our landmark detection algorithm consists of four parts: scene classification, object detection, scene-object relation and landmark detection. Figure 1 illustrates this computational model. CNNs have been developed to benefit from large-scale data, hence we use two deep CNNs: one for scene classification and one for object detection. Please note, our computational model is capable of using other and different algorithms for scene classification or object detection as long as the two algorithms are not correlated closely. An entire image is input into the scene classification CNN and the object detection CNN. The results from the CNNs are probabilities that this input image belongs to certain scene categories and, separately, that the regions in the image belong to some object categories. Landmarks are identified according to a surprise score calculated based on the probabilities of scene and object, and pre-defined scene-object relation.

B. Scene Classification and Object Detection

For scene classification, we use Places-CNNs: the CNNs trained on the MIT Places [17] database and used for scene classification. For object detection, we use Fast R-CNN [24], a fast region-based CNN method. Actually, one of the advantages of our model is that we are able to leverage any state-of-the-art algorithms for scene classification and for object detection. However, they should satisfy the requirement that the algorithm for scene classification and the algorithm for object detection are not closely correlated.

1) Relation between Scene Classification and Object Detection: Human beings are experts at perceiving scenes and understanding their contents, and we generally recognize the scene based on understanding the objects and the visual attributes (e.g., color, shape) in this scene. For example, we generally perceive a scene as a living room because it’s an indoor environment and there is a sofa, a television, chairs and so forth in it. Similarly, some scene classification methods use objects as high-level semantic information, as opposed to low-level visual features, to classify scenes (e.g., [32], [33]). However we argue that this kind of approach is not very useful for landmark detection. Suppose there is an outdoor scene in which a sofa and a television and several chairs are in the sea. It is abnormal for a sofa, a television or a chair to appear in the sea. Therefore they may be considered as good landmarks in this scene. But an object-based scene classification method may consider this scene to belong to the indoor scene category just because it contains these objects. If we use this kind of scene classification approach, then as a result, landmark detection fails to identify what may be very useful landmarks. In order to ensure that the two methods are independent, we choose a scene classification method which is based on global features of an image, and an object detection method based on regions of an image.

2) Scene Classification: MIT Places [17] is a scene-centric database which is the largest publicly available image database of scenes and places at this time (over 7 million labeled pictures of scenes from 476 place categories). MIT Places is large enough to train algorithms that require huge amounts of data, such as CNNs. The results obtained by the deep features from CNNs trained with MIT Places significantly outperforms those obtained by the deep features from the same network architecture trained with object-centric database ImageNet [34]. In our model, we use deep CNN
to learn scene features from images for scene classification, and the CNN is trained on MIT Places database. The deep CNN that we use is VGG-16 [20], which has achieved high accuracy rates in several kinds of recognition tasks (the first and the second places in the localization and classification tasks respectively in ILSVRC 2014 competition).

3) Object Detection: Fast R-CNN [24] achieves near real-time speed and high detection accuracy using a deep CNN. Region proposals are first extracted from an input image. Then this input image and its region proposals are fed into a fully convolutional network. A region of interest (RoI) pooling layer extracts a fixed-dimensionality feature vector from the convolutional feature map for each region proposal. Finally each feature vector is input into fully connected (FC) layers and two output vectors are obtained: softmax probabilities and per-class bounding-box regression offsets. In our model, we use Fast R-CNN to detect objects from images. The CNN we use is also VGG-16 [20].

C. Relation between Scenes and Objects

The context contains information about how scenes and objects are related to each other, such as that refrigerators belong in the kitchen and sofas in a living room. An object may look abnormal due to its atypical context, and therefore this object may be a good landmark candidate because it is unique and easy to spot by different viewers or the same viewer at different times. For example, a car appearing in a lobby is a good landmark because it is unlikely that this will be a very common occurrence. Using it as a point of reference for navigation or for directions would be more successful than picking a landmark that is difficult to spot or is commonly duplicated. Hence we need prior knowledge about the probability of an object appearing in a scene. For instance, the probability that an elephant in the sky is almost 0%. Inspired by [31], we propose two methods to define this prior knowledge or relation: one is a human subject experiment, and the other one is learning from large scene datasets.

1) Human Subject Experiment: Humans are experts in determining typical object context. For this approach, we built a list of scene-object pairs, and the human subject was asked to determine whether the object should appear normally in the scene or not. The answer in each case is “Yes” or “No”. For each pair of object and scene the probability of the object appearing in the scene was calculated as the number of “Yes” answers divided by the total number of human subjects. For example, for the scene-object pair “car” and “street”: if 100 human subjects took the experiment and 98 of them answered “Yes” for the question “Should a car appear in the street?”, the probability of “cars” appearing in the “street” would be

\[
P(O = \text{car} | S = \text{street}) = \frac{98}{100} = 0.98.
\]

In our experiment, we polled one human subject to get the results. In the future, we will use Amazon Mechanical Turk to collect a large number of people responses for scene-object relations.

2) Learning from Datasets: The most common places where an object is found within the SUN database are listed in the annotations of the database [35]. For example, for the object “table”, the top 10 most common places are “dining room”, “kitchen”, “bedroom”, “living room”, “dineette vehicle”, “dineette home”, “art studio”, “conference room”, “restaurant”, and “bar”. We consider “table” showing in these scenes is normal, the relation probabilities are 1. The probabilities of “table” showing in other scenes are 0. For now we use the Human Subject Experiment method with one subject to obtain the scene-object relations because it generates accurate results by definition; we will use the dataset method in our future experiments.

D. Landmark Detection and Landmark Image Classification

Assume there are \(M\) scene categories and \(N\) object categories. For an input image \(IM\), we apply scene classifiers to obtain probabilities \(P(S_i | IM), (i = 1, 2, \ldots, M)\). \(P(S_i | IM)\) denotes the probability that given the image \(IM\) the scene category is \(S_i\). Then \(K\) objects are detected from the image \(IM\), and we get \(P(C_k = O_j | IM), (j = 1, 2, \ldots, N; k = 1, 2, \ldots, K)\) as the probability that each object \(C_k\) belongs to object category \(O_j\). An object is detected when its probability of being in a category is larger than a threshold. The conditional probability that object category \(O_j\) appears in the scene category \(S_i\) is \(P(O_j | S_i)\). Therefore the surprise score for object \(C_k\) in the image \(IM\) is computed as following:

\[
\text{Surprise}(C_k, IM) = \sum_{j=1}^{N} \left[ P(C_k = O_j | IM) \sum_{i=1}^{M} P(-O_j | S_i) P(S_i | IM) \right].
\]  

where the surprise score \(\text{Surprise}(C_k, IM)\) determines whether object \(C_k\) is a landmark or not in the image \(IM\). For a detected object \(C_k\), the product \(P(C_k = O_j | IM) P(S_i | IM)\) is the probability that the object \(C_k\) belongs to the category \(O_j\), and it is in the scene \(S_i\) given that these are independent (and we have chosen the scene and object detection methods with this in mind, as discussed). \(P(-O_j | S_i) = 1 - P(O_j | S_i)\) is the probability that the object category \(O_j\) is not in the scene category \(S_i\) normally. Therefore \(P(C_k = O_j | IM) P(-O_j | S_i) P(S_i | IM)\) defines how surprise score that the object \(C_k\) of category \(O_j\) is in the image \(IM\) which is of scene type \(S_i\). We sum all the scores of the surprise score to decide if \(C_k\) is a good landmark in the image \(IM\). It is a good landmark if \(\text{Surprise}(C_k, IM) \geq T_{os}\), where \(T_{os}\) is a threshold we establish for landmarks.

To detect landmarks in an image, the time computational complexity is \(O(KMN)\), where \(K\) is the number of objects that are detected, \(M\) is the number of scene categories, and \(N\) is the number of object categories.

If an object \(C_k\) is considered as a landmark in the image \(IM\), we use the following formula to determine its object category.

\[
\arg \max_j P(C_k = O_j | IM) \sum_{i=1}^{M} P(-O_j | S_i) P(S_i | IM).
\]  

A landmark image means that there are one or more landmark objects in this image. Landmark image classification is a useful application because it can work as a fast filter which
selects landmark images from massive datasets. For example, an application requires that the landmark detection accuracy should be more than 95% for a complicated image dataset. It is impossible for machines to achieve this goal nowadays, but humans can perceive landmarks precisely. We apply the model for this case using overall surprise score of an image IM which can be formulated as:

\[
\text{Surprise}(IM) = \sum_{k=1}^{K} \sum_{j=1}^{N} P(C_k = O_j|IM) \sum_{l=1}^{M} P(\neg O_j|S_l) P(S_l|IM). \tag{3}
\]

Surprise(IM) means the possibility that there is at least one landmark in the image IM. Similarly, we set a threshold \( T_{th} \) for images, and images whose overall surprise score smaller than \( T_{th} \) will be filtered.

IV. EXPERIMENTS AND RESULTS

A. Landmark Image Dataset

We collected an image dataset which consists of landmark images and non-landmark images. A landmark image refers to an image in which there are one or more landmark objects. Conversely, a non-landmark image refers to an image not containing a landmark object. To collect this dataset, we gathered landmark images from the 1001 Abnormal Image dataset [31] and other miscellaneous image websites, and non-landmark images from the test set of the Pascal VOC 2007 dataset [36]. This dataset contains 6 object categories: namely “Aeroplane”, “Boat”, “Car”, “TV/ Monitor”, “Motorbike”, and “Sofa”. There are more than 250 images in the current version of dataset, and half of them are landmark images. The overall dataset contains more than 450 objects. The collected images were annotated by marking bounding boxes around the objects in each image using the Training Image Labeler app in MATLAB. Figure 2 shows some landmark images of this dataset. Figure 3 shows an image and the associated XML annotation file. In each XML annotation file, there are the following fields: scene category of the whole image, the information of objects including their object categories, positions and whether they are landmarks or not. In the dataset, we also built a table of relations between scenes and objects using a limited human subject method in which a single subject labeled each relation between an object and a scene based on his common judgement. Table I shows examples of some items in the relation table.

**TABLE I**

<table>
<thead>
<tr>
<th></th>
<th>aeroplane</th>
<th>boat</th>
<th>car</th>
<th>TV/monitor</th>
<th>motorbike</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>bridge</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>highway</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>sky</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 3. Example of an image and its XML annotation file. An XML annotation file includes information: scene category of the image, and object category and bounding box, and whether it is a landmark or not for each object in the image.

B. Experimental Setup

Our landmark image dataset will only be used for testing. The MIT Places [17] database is the training set for scene classification, and 205 scene categories and 2.5 millions of images with a category label from this database are used. The training sets for object detection are the ImageNet [34] database and a training set of the Pascal VOC 2007 database [36], where ImageNet is an auxiliary dataset for supervised pre-training with image-level annotations. There are several reasons for this training and testing setup. The most important one is that we can benefit from big data. Both MIT Places and ImageNet are massive databases which capture large underlying patterns and features for categories. For example, we use 205 scene categories from MIT Places and it is enough to describe these scenes of the images in our landmark dataset. The objects in our landmark dataset are normal, and ImageNet has enough these kinds of images for training. The second reason is that it is useless to train CNNs on our dataset currently, because CNNs require huge amounts of training data, whereas our landmark dataset is relatively small. We use VGG-16 [20] CNN for the experiments. We implemented our model based on Caffe Deep Learning Framework [37]. The results, including recall, precision, and F1 score, and accuracy, of the experiments are reported.

C. Results

1) Landmark Detection: In the landmark detection experiment, an object is considered a good landmark if its surprise score is larger than a threshold of 0.4 (the optimal threshold obtained from a small sample test). The confusion matrix of landmark detection is shown in Figure 4.

The average accuracy of landmark detection is **65.29%**. Table II reports the precision, the recall, and the F1 score of landmark detection. Categories “aeroplane”, “car”, “motorbike”, and “sofa” have good detection results. Categories “boat” and “TV/monitor” have fair detection results. Actually this task is difficult in that we should not only detect whether there are one or more landmarks in an image but also know
Fig. 2. Examples of landmark images from our dataset. Columns: images from six categories (aeroplane, boats, cars, TV/monitors, motorbikes, sofas)

Fig. 4. The confusion matrix for landmark detection results.

To which category a landmark belongs. Figure 5 shows some examples of landmark detection results. The first row shows the landmarks which are detected correctly. The second row shows landmarks that are not detected. The third row shows objects which actually are not landmarks but are considered as landmarks.

![Confusion Matrix for Landmark Detection](image)

**TABLE II**

<table>
<thead>
<tr>
<th></th>
<th>aero-</th>
<th>boat</th>
<th>car</th>
<th>TV/monitor</th>
<th>motorbike</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>1.0</td>
<td>0.75</td>
<td>0.865</td>
<td>0.857</td>
<td>0.88</td>
<td>1</td>
</tr>
<tr>
<td>Recall</td>
<td>0.625</td>
<td>0.429</td>
<td>0.696</td>
<td>0.667</td>
<td>0.917</td>
<td>0.674</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.769</td>
<td>0.545</td>
<td>0.771</td>
<td>0.75</td>
<td>0.898</td>
<td>0.806</td>
</tr>
</tbody>
</table>

2) **Landmark Image Classification:** In the landmark image classification experiment, an image is considered as a good landmark image if its overall surprise score is larger than the threshold of 0.4 (the optimal threshold obtained from a small sample test). The confusion matrix of landmark detection is shown in Figure 6.

The average accuracy of landmark image classification is **87.42%**. Table III reports the precision, the recall, and the F1 score of landmark image classification. As Table III shows, both the accuracy and F1 score of landmark classification are above 80%, which is high, demonstrating that our model has ability to figure out which images potentially have landmarks.

![Confusion Matrix for Landmark Image Classification](image)

**TABLE III**

<table>
<thead>
<tr>
<th></th>
<th>landmark</th>
<th>non-landmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.927</td>
<td>0.81</td>
</tr>
<tr>
<td>Recall</td>
<td>0.770</td>
<td>0.95</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.845</td>
<td>0.874</td>
</tr>
</tbody>
</table>
Fig. 5. Examples of landmark detection results. First row: landmarks that are detected correctly. Second row: landmarks that are not detected. Third row: objects which are not landmarks are considered as landmarks.

D. Discussion

We use two deep CNNs – one for scene classification and one for object detection – because CNNs have achieved outstanding performance in these tasks. As mentioned above, one of advantages of our model is that it can leverage other different algorithms for scene classification or object detection as long as the two algorithms are independent of each other. These two topics are developing rapidly in the computer vision and machine learning community so that we can leverage them to improve the performance of our model. Landmark detection is different from object detection because it needs to not only recognize an object and locate it but also identify whether it is a landmark or not. Based on the landmark detection experiment Fast R-CNN [24] fails to detect landmarks.

In the landmark detection experiment, there are 6 landmark categories plus a “nothing” category. The probability of a guess is less than 14.29% because we should recognize the category and the location of an object as well as whether it is a landmark or not. Our model has achieved the average accuracy of 65.29% in landmark detection which is far more than the probability of a guess. The probability of a guess at classifying landmark image is 50%. Our model has achieved the average accuracy of 87.42% for landmark image classification. Hence our model is reliable to work as a filter which selects landmark images from a massive database.

V. Conclusions

Autonomous mobile robots need to discover landmarks automatically for loop closure and place recognition in navigation; We introduce a novel model for landmark detection, fusing scene classification and object detection. The process is that an image is input into the scene classification CNN and also the object detection CNN. The results are the probabilities that this image belongs to certain scene category, and the regions in the image belong to some object categories. Landmarks are identified according to a surprise score, which is calculated based on the probabilities of scene and object, and a pre-defined scene-object relation. We collected a landmark image dataset which consists of landmark images and non-landmark images to evaluate the performance of our model. The experimental results have demonstrated the good performance of landmark detection (the accuracy is 65.29%) and landmark image classification (the accuracy is 87.42%).

There are several improvements that can be considered for this research. The first one is that we will expand our landmark dataset by adding more landmark images. If we have a large landmark image dataset, we can train the models of different CNNs on it with the auxiliary databases and check whether it will improve the accuracy of landmark detection or not. The second is that we can use Amazon Mechanical Turk to collect a large number of people responses for scene-object relations. The third one is that we can consider the location of an object in a scene in the future. For example, if the object is a car and the scene is a street, our model won’t consider this car is a landmark. If this car in the air, it may be a good landmark candidate. Currently our model will fail to detect a landmark in this scenario.

Many applications rely on real-time, robust landmark detection. For example, in Counter-Weapons of Mass Destruction missions, mobile robots need to handle threats to human beings in a timely and safe fashion. Real-time landmark detection will be helpful to establish the location of the robots and threats such as biochemical bombs. In order to achieve real-time landmark detection, faster object detection algorithm is needed. Object detectors YOLO [38] and Faster R-CNN [25] can detect object in real-time while
achieving state-of-the-art object detection accuracy. We will incorporate these two algorithms in our landmark detection model in the future.

REFERENCES