NATURAL OBJECT DETECTION IN OUTDOOR SCENES BASED ON PROBABILISTIC SPATIAL CONTEXT MODELS

Jiebo Luo1, Amit Singhal1, and Weiyu Zhu2

1 Electronic Imaging Products, R & D
Eastman Kodak Company
{luo, singhal}@image.kodak.com

2 Dept. of Electrical Engineering & Computer Science
University of Illinois, Urbana-Champaign
weiyu.zhu@uiuc.edu

ABSTRACT

Natural object detection in outdoor scenes, i.e., identifying key object types such as sky, grass, foliage, water, and snow, can facilitate content-based applications, ranging from image enhancement to other multimedia applications. A major limitation of individual object detectors is the significant number of misclassifications that occur because of the similarities in color and texture characteristics of various object types and lack of context information. We have developed a spatial context-aware object-detection system that first combines the output of individual object detectors to produce a composite belief vector for the objects potentially present in an image. Spatial context constraints, in the form of probability density functions obtained by learning, are subsequently used to reduce misclassification by constraining the beliefs to conform to the spatial context models. Experimental results show that the spatial context models improve the accuracy of natural object detection by 13% over the individual object detectors themselves.

1. INTRODUCTION

Many image-related applications can benefit from object detection. In content-based image retrieval (CBIR), more and more researchers are looking beyond low-level color, texture, and shape features in pursuit of more effective searching methods. For example, Naphade and Huang [1] use a list of semantic objects, including sky, snow, rock, water, and forest, in a factor graph-based framework for semantic indexing and retrieval of video. Color has been the central feature of existing methods. However, one important trait of humans that they only examine isolated patches of pure object materials without context, which is extremely challenging even for human observers. Many object materials can have the same appearance in terms of color and texture, while the same object may have different appearance under different imaging conditions (e.g., lighting, magnification). However, one important trait of humans is that they examine all the objects in the scene before making a final decision on the identity of individual objects. The key in this holistic recognition process is the use of spatial context. Spatial context refers to the relationships among the location of different objects in the scene and is useful in many cases to reduce the ambiguity among conflicting detectors and eliminate improbable spatial configurations in object detection. Two types of spatial contextual relationships exist in natural images. First, relationships exist between co-occurrence of certain objects in natural images. For example, detection of grass with high probability would imply low snow probability. Second, relationships exist between spatial locations of certain objects in an image: sky tends to occur above grass, foliage above grass, sky above snow, etc.

2. OVERVIEW OF THE A HOLISTIC OBJECT DETECTION SYSTEM

Fig. 1 shows the overview of our holistic natural object detection system. The input image is first processed by a number of individual object detectors, each which outputs a belief map indicating the likelihood of a region being a certain object. Individual belief maps are passed to the object belief fusion module. This module integrates the individual belief maps and computes a combined object belief map for the image. Meanwhile, the individual belief maps are also used to segment the image into regions that are homogenous with respect to the outputs of the object detectors. The segmentation map, along with the combined belief maps for each region, is passed to the spatial context-based belief refinement module. This module revises the belief vectors of each region in the image by incorporating the object spatial context models and generates an updated object belief map that is coherent with the object spatial constraints and the initial object detection beliefs.

![Fig. 1. Architecture of the holistic object detection system.](image)
2.1. Baseline natural object detectors

Objects types most frequently present in consumer photos provide the highest payoff. Numerous studies at Kodak have shown that the most valuable object types or subject matters in consumer photos for photo-finishing are people, sky, and grass. After considering both frequency (see Table 1) and detectability by automatic algorithms, we have selected pure material types such as sky, grass, foliage, open water, and snowfield for our natural object detection system. These object types are also often related to the background of a scene and have well-defined spatial relationships among each other.

Table 1 Statistics of major object types in consumer photos.

<table>
<thead>
<tr>
<th>Object type</th>
<th>% of all pixels</th>
<th>% of all images</th>
<th>% of indoors</th>
<th>% of outdoors</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td>5.0</td>
<td>31.0</td>
<td>3.1</td>
<td>55.0</td>
</tr>
<tr>
<td>grass</td>
<td>4.1</td>
<td>29.0</td>
<td>1.8</td>
<td>52.0</td>
</tr>
<tr>
<td>foliage</td>
<td>3.2</td>
<td>35.0</td>
<td>9.5</td>
<td>65.0</td>
</tr>
<tr>
<td>water</td>
<td>2.1</td>
<td>10.0</td>
<td>1.0</td>
<td>19.0</td>
</tr>
<tr>
<td>Snow/ice</td>
<td>0.3</td>
<td>2.4</td>
<td>0.8</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Existing methods for detecting individual natural objects such as sky and grass [1–4] follow a common architecture similar to the one shown in Fig. 2, where low-level, feature-based classification is combined with some region-level analysis to generate individual object belief maps (many are simply binary). First, color and/or texture features are computed on the input image. Preferably, the computed features should be effective in differentiating the pixels of the concerned natural object from those of other object types. The features are fed to a classifier, such as a trained neural network, which produces a probability or belief value for each pixel in the image according to the color and/or texture characteristics of the natural object. The collection of pixel belief values forms a pixel belief map.

To generate the color and texture features, the images are first converted from the RGB color space to the LUV color space. We have observed that sky and snow regions tend to be the brightest regions in an image because sky is almost always the main source of illumination in outdoor scenes and snow reflects almost all of the light incident upon it. The LUV color space allows us to take advantage of this observation for the cloudy sky and snow detectors. A normalization step is used to define a pixel-level luminance feature \( I' \) on a per image basis

\[
l' = \frac{l}{\max_l}
\]  

where \( l \) is the raw luminance value for each pixel in the image. This physics-motivated feature extraction step leads to significant reduction in false positive detection of other grayish colored subject matter, such as fabric, road, rock, dry wall, and concrete surface as sky or snow. However, it can also lead to missed detection of very dark gray skies or snowfields with shadows on them. The normalized LUV triplet for each pixel provides three color features for the neural network classifier.

The texture features are computed using a wavelet transform based on classic multiresolution analysis [11]. A two-level wavelet decomposition was performed on the luminance channel. We ignored the lowpass coefficient, which is a measure of the local average of the signal and already accounted for by the use of the luminance color feature (the color and texture features are classified using a single neural network). Since we are only interested in the magnitude of the texture energy, absolute values of the highpass coefficients are used as texture features. This gives us a total of 6 texture features for each pixel.

A bootstrapping process using a separate validation set is used to retrain the network for each object type to reduce the number of false positives and false negatives (misses). However, we cannot include confusing materials that are only different semantically and nearly impossible to differentiate by color and texture features, e.g., white shirt when training a cloudy sky detector, or clouds when training a snow detector. The neural network generates a pixel-level map that associates each pixel in the image with a belief in that pixel being a certain object type.

After pixel classification, spatially contiguous regions are obtained from the raw pixel belief map after thresholding the belief values. Finally, each spatially contiguous region is further analyzed according to certain unique region-based characteristics of the concerned natural object type (e.g., color gradient in blue sky [5], if any, and the confirmed regions form the final region belief map where each region is associated with a uniform belief value. While some of these individual object detectors (e.g., for blue sky) have very good performance because of object-specific region analysis that removes false positives, other natural object detectors (such as those for open water and snowfields) suffer substantially high misclassification rates. Used independently, the accuracies of our detectors for blue sky, cloudy sky, grass, snow and open water range from 52\% (open water) to 96\% (blue sky).

2.2. Object belief fusion

The purpose of object belief fusion is to obtain an initial unified belief vector for each segmented image region according to the initial beliefs generated by the low-level individual object detectors and the image orientation information. This step may be unnecessary if one uses a single classifier to detect all the natural object types [1]. For each segmented region, a two-layer Bayesian network is trained for object belief fusion. The architecture of the fusion system is illustrated in Fig. 3.

![Fig. 3. Bayesian network for object belief fusion.](image-url)

The root node of the network represents the set of selected object types for each region in the image. These object types could be one of several natural objects of interest, such as sky, grass,
foliage, water, snow, and background. The root node is connected to a number of evidence nodes via conditional probability matrices (CPMs). Currently, the evidence nodes consist of five object detectors, one for each object type of interest, and one region location detector. Each object detector has two labels, activated and non-activated (denoted as "+'" and "-'"), respectively, where activated means that the detector has a non-zero output for that region and non-activated otherwise. The location evidence node has three labels, top, middle, and bottom, corresponding to three quantitative locations of a region in an image. Given a segmented image region, the beliefs generated by the individual detectors and the location of the region in an image are fed into the leaf nodes of the network and propagated to the root node, which outputs a composite object belief vector for that region.

2.3. Spatial context-based belief refinement

We further augment the object classification by incorporating global spatial constraint information through the use of spatial context models. This allows us to further refine the initial belief vectors by using spatial relationships, such as "sky occurs above grass," to reduce misclassifications that are not eliminated by the use of the orientation information.

The purpose of the spatial context-based object belief refinement module is to use the spatial relationship models described in the next section to evaluate and adjust the belief vectors obtained from the low-level object belief fusion module. The basic methodology here is to build graphical structures that capture the spatial constraint information of all the objects in an image; followed by an inference engine to generate new belief vectors that represent the best compromise between the spatial constraints and the original belief vectors. While it may be intuitive to encode the spatial relationships between all the regions in the image as a single graphical structure, this results in a loopy network for which a closed-form Bayesian inference solution does not exist. To make the inference engine practical, we use a series of non-loopy networks that approximate the global spatial constraints among all the regions in the image. These networks can be solved in an iterative manner, resulting in a spatially coherent, object belief map.

3. ON SPATIAL CONTEXT MODELING

There is existing work on using high-level scene models for spatial context-based object detection. Batte et al. [8] provide a comprehensive review of most of the work related to building scene models for specific image types. They describe techniques where spatial models can be constructed for scene types, such as a house scene, a road scene, and an urban scene. In each of these scene types, there is a strong expectation regarding the occurrence and location of various object types in the image. Even simple rule-based, spatial-context models can provide good performance in these scenarios. The main limitation of this technique, however, is that a different model must be built for each scene type, restricting its applicability to a general scene understanding application. Lipson et al. [9] present a spatial context modeling approach, called configuration-based scene modeling, for content-based indexing and retrieval applications. They model the qualitative and photometric relationships between various objects in a scene in a spatial sense and use these relationships to extract other scenes with semantically similar content. The scene models are extremely specific to the layout of a scene, e.g., ocean on top of sand is different from ocean beside sand.

Consumer photo applications have highly varied scene content for which multiple scene models would become computationally intractable. We propose the use of probabilistic spatial context models, defined as probability density functions over the universe of natural image content, for modeling the spatial relationships between various objects of interest in an image. Depending on the requirements of the application, the set of spatial relationships can be rich (many spatial relationships with minor differences between each) or sparse (fewer distinct relationships). The set of spatial relationships modeled in our system are \{above, far_above, below, far_below, beside, enclosed, and enclosing\}. A threshold on the distance between the nearest pixels of two regions is used to discriminate between above and far_above (and below and far_below).

In this study, we chose to quantify the spatial relationship of two regions in two ways. One is by checking the bounding boxes of the regions and the other is by using a lookup table of the directional weights of two regions computed via a statistical counting approach based on weighted walkthrough [10]. The bounding box method is easy to implement, but may encounter difficulties when the bounding boxes of the regions overlap. The lookup table method is robust to the size and location of regions, but is computationally more complex than the bounding box method. We use a hybrid scheme to determine the spatial relationship of two regions in an image. First, we check the bounding boxes of the regions. If the bounding boxes are not overlapping, we can use their coordinates to quickly derive the spatial relationship.

The spatial context models are built by learning probability density functions corresponding to the spatial relationships described above, as opposed to handcrafted rules. We have collected extensive ground truth for various object types for a set of 1800 consumer images. A simple frequency counting approach suffices to generate discrete probability density functions for all the spatial relationships in the system. An example of the probability density function generated for the relationship far_above is shown in Fig. 4. As an example, the first row in the matrix shows that any region far above a sky region is most likely to be sky, followed by foliage, and other (it cannot be grass, water, or snow).

$$CPM_{far\_above} = P_{far\_above}(A \mid B)$$

![Fig. 4. Probability density function for far_above.](image)

4. EXPERIMENTAL RESULTS

We conducted experiments by applying the holistic object classification system to a set of 780 consumer images containing at least two different object types, allowing for the use of spatial contextual information. The detection accuracy, defined as the number of regions correctly classified using the individual detectors, is approximately 70.6%. Addition of orientation information in the object belief fusion step increases the accuracy
Fig. 5 shows the object belief vectors for one test image at various stages of the system. The original image is shown on the top left and the ground-truth map is shown on the top right. A human observer generated the ground truth by selecting polygonal regions of homogenous content and appropriately labeling selected regions via visual inspection. The second row shows the belief vectors generated using the output of the individual object detectors and without the use of orientation information. At this stage, region 1 and region 2 have the highest belief for snow and background (“bg”) respectively, although they also have lower beliefs in sky and foliage (the correct classifications). Use of the orientation information changes the maximum belief in region 1 to be sky (because snow rarely occurs on the top of the image), but does not cause any major changes in the belief values of region 2. The last image shows the results obtained after applying the spatial context models. All four regions have the highest belief value for the correct object type. In Fig. 6, the region in the top of the image incorrectly classified as snow initially is corrected by incorporating orientation information. The two regions initially labeled as background regions were eventually reclassified to the same classes as the ground truth because their initial lower beliefs in foliage and grass were reinforced by the spatial context model.

5. CONCLUSIONS AND DISCUSSION

We have demonstrated a holistic approach to natural object classification that uses spatial context constraints to increase the accuracy of the initial classification by constraining the beliefs to conform to the spatial context models. Although the use of spatial context model may sometimes degrade the detection, experimental results show that the spatial context-aware models improve the accuracy of an orientation-aware, outdoor natural object detection system by 13%.

One obvious area of improvement in the future is in the individual natural object detectors. Our current sky and grass detectors perform well as individual object detectors with close to 90% accuracy, even without orientation. However, the cloudy sky, snowfield, open-water, and foliage detectors have accuracies of approximately 50-80% and can be substantially improved. It would be of great interest and a great challenge to extend our work to indoor scenes.

6. REFERENCES