Context-Dependent Object Proposal and Recognition

Ray-I Chang ¹, Chao-Lung Ting ¹, Syuan-Yi Wu ¹ and Peng-Yeng Yin ^{2,3,*}

¹ Department of Engineering Science and Ocean Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan; rayichang@ntu.edu.tw (R.-I.C.); chaoting@ntu.edu.tw (C.-L.T.); sandywu0603@hotmail.com.tw (S.-Y.W.)

² Department of Information Management, National Chi Nan University, No. 1, University Rd., Puli, Nantou 54561, Taiwan

³ Institute of Strategy and Development of Emerging Industry, National Chi Nan University, No. 1, University Rd., Puli, Nantou 54561, Taiwan

* Correspondence: pyyin@ncnu.edu.tw

Received: 19 August 2020; Accepted: 28 September 2020; Published: 30 September 2020

Abstract: Accurate and fast object recognition is crucial in applications such as automatic driving and unmanned aerial vehicles. Traditional object recognition methods relying on image-wise computations cannot afford such real-time applications. Object proposal methods appear to fit into this scenario by segmenting object-like regions to be further analyzed by sophisticated recognition models. Traditional object proposal methods have the drawback of generating many proposals in order to maintain a satisfactory recall of true objects. This paper presents two proposal refinement strategies based on low-level cues and context-dependent features, respectively. The low-level cues are used to enhance the edge image, while the context-dependent features are verified to rule out false objects that are irrelevant to our application. In particular, the context of the drink commodity is considered because the drink commodity has the largest sales in Taiwan's convenience store chains, and the analysis of its context has great value in marketing and management. We further developed a support vector machine (SVM) based on the Bag of Words (BoW) model with scale-invariant feature transform (SIFT) descriptors to recognize the proposals. The experimental results show that our object proposal method generates many fewer proposals than those generated by Selective Search and EdgeBoxes, with similar recall. For the performance of SVM, at least 82% of drink objects are correctly recognized for test datasets of various challenging difficulties.

Keywords: object proposal; low-level cue refinement; context-dependent refinement; SIFT; bag of words; support vector machine

1. Introduction

Traditional object recognition techniques require computations for extracting local features among pixels or salient points, such that high recognition accuracy can be obtained by conducting matching between the target object and an exhaustive set of candidate regions containing the local features. However, this sort of image-wise feature description technique, e.g., Hough transform [1] and scalable recognition [2], entails numerous computations. Object proposal generation is a process that intends to superimpose bounding boxes on the regions containing the objects of interest in the image or video. The aim of generating object proposals is to reduce the computations for pixel-by-pixel matching between the sought object and the entire image/video, but only focusing on the candidate object proposals that are deemed to contain the target object. The performance of an object proposal generator is thus justified by the number of generated object proposals and the recall of target objects. An ideal object proposal generator not only improves the recall but also produces fewer object proposals. Object proposal generation has become a prevailing preprocessing step for several image processing

and pattern recognition tasks, such as image segmentation, object recognition, target tracking, and video retrieval [3].

The output of proposal generation methods should contain two types of information: the localization of the candidate object proposals and the likelihood ranking of these candidate proposals [4]. The localization is usually represented by the minimum bounding box of the proposal, in which case the coordinates of the upper-left point and the width and height of the bounding box are tallied. To find the bounding box for the candidate proposals, two categories of approaches have been proposed. The bottom-up approaches start with pixel-level local feature descriptions such as edge gradients or color intensities and then hierarchically group the pixels with high proximity and homogeneity (such as similar edge orientations or color distributions) into super-pixels. The top-down approaches use several sliding windows with various scales to raster-scan the entire image and calculate at every sliding location how likely the region within the window is to contain an object. To estimate the likelihood for a proposal being a real object, several heuristics for measuring the objectness of a proposal have been proposed in the literature. These heuristics range from low-level cues such as edge orientation, texture distributions and color contrast [5,6], contours and symmetry at the region border [7,8], and the aspect ratio and size of the bounding box [5], to high-level semantics and context such as the idea that the road should appear below cars in the automatic driving context [9].

There exist some drawbacks with the traditional object proposal methods. To achieve a high recall rate, the bottom-up approaches usually apply multiple grouping criteria and produce many false-positive and redundant proposals. On the contrary, the top-down approaches rely on prudent choices of appropriate sliding windows and may suffer inferior recall and lack localization accuracy. Recently, two strategies have emerged to eliminate these drawbacks. The first strategy is to conduct a postprocessing step named proposal refinement, which removes false positives and refines the localization and size of the initial proposals produced by the generation methods [8–11]. The second strategy is to narrow down the proposal scope to a particular context. Most of the existing proposal generation approaches adopt the context-independent, i.e., class-agnostic, assumption for the sought objects. The curse of the context-free scenario poses a strong challenge for achieving a high recall rate with only a few hundred candidate proposals. Zhong et al. [9] proposed a class-specific re-ranking method in the automatic driving context. High-level features such as semantic segmentation, stereo information, and contextual information are employed to significantly reduce the number of proposals without deteriorating the recall as compared to the original object proposal method.

In this paper, we combine the two strategies. We focus our application on the context of drink commodity proposal and recognition. Therefore, both low-level cues and high-level semantics related to drink commodity objects can be used to refine the initial proposals produced by existing generation methods. In particular, we employ the fast edge detection method of EdgeBoxes [7] and refine the edge map by removing non-interesting edge groups. The context-dependent application and proposal refinement heuristics allow us to reach higher recall with fewer proposals as compared to with Selective Search [6] and EdgeBoxes [7]. The reason we focus on drink commodity proposal and recognition is three-fold. (1) The drink commodity has the largest sales (34.8%) [12] in convenience store chains in Taiwan. The context of drink commodity proposals benefits the deployment of advanced cashier-free stores such as Amazon Go. (2) Day et al. [13] disclosed a strong relationship between purchase decision strategies and eye-movement sequences. By locating the drink commodity proposals, one can analyze the decision strategy of customers for drink purchases in grocery stores. This information would be useful in planning the drinks' shelf arrangement. (3) In addition to tracking eye movement, a scene-drink association analysis from an enormous volume of images and videos is mandatory for knowing the best-selling drink during various scenes, such as reading, exercising, meeting, etc. The discovered knowledge is of great value in practicing precise marketing and drink promotions.

The remainder of this paper is organized as follows. Section 2 reviews relevant proposal generation and proposal refinement methods. Section 3 articulates the proposed method with its distinct features.

Section 4 describes the experimental results and discussion. Finally, conclusions and future research lines are given in Section 5.

2. Related Work

In this section, we review the related literature that covers main proposal generation and proposal refinement methods. The proposal generation methods produce the bounding boxes of object proposals from scratch, while the proposal refinement methods intend to refine the position and size of the initial proposals from generation methods or to reduce the number of the initial proposals via re-ranking techniques.

The proposal generation methods can be roughly divided into two categories: grouping-based and objectness-based. The grouping-based approaches usually start with gradient or edge maps and then hierarchically merge the regions with homogeneous properties such as edge orientations or color features. The objectness-based methods use a sliding window to scan multi-scale images and justify the objectness for each window according to some heuristic scoring criteria such as the edge distributions, aspect ratio, and size. Selective Search [6], which is one of the earliest grouping-based proposal generation methods, applies multi-scale color-space analytics to obtain a hierarchical merging of color homogeneous regions. The homogeneity is measured according to the diversity distributions in different color channels and the feature statistics such as texture, size, and compatibility. The bounding boxes that cover homogeneous object proposals are learned by an support vector machine (SVM). The two-stage Cascade SVM (CSVM) [14] is another grouping-based method that obtains high-quality proposals by using gradient features and learns all the quantized scales and aspect ratios of bounding boxes for calibration. The binarized normed gradients (BING) method [15] learns the objectness measure to produce object proposals. It slides a fix-sized window on multi-scale images and calculates normed gradients maps, which are then used to measure the objectness with the CSVM. BING is a computationally efficient proposal generator, training the CSVM on binary features. Objectness [5] uses a sliding window to find candidate proposals at salient locations and then measures the objectness of each proposal according to multiple cues of object properties, such as the saliency, color contrast, edge density, location, and size statistics. EdgeBoxes [7] considers the object contour as important information to discriminate objects from non-objects. The edges with similar orientations are grouped together, and then, each bounding box is scored by relevant edge groups without any object-parameter learning. Some recent approaches take the advantageous power of convolutional neural networks (CNN) to learn both the localization and the objectiveness of each proposal. The Faster Region-based CNN (Faster R-CNN) [16] introduces a region proposal network that shares full image convolutional features with the original Fast R-CNN [17] for object detection. The Faster R-CNN does not apply Selective Search as does Fast R-CNN. Instead, the region proposal network directly learns the proposal parameters from the same convolution neural networks for object detection. Thus, the additional cost for learning object proposals is reduced. The region proposal network simultaneously predicts object bounds and objectness scores at each position while the Fast R-CNN is learning the convolutional features for object detection. Redmon et al. [18] proposed a You Only Look Once (YOLO) method that considers object detection as a regression problem from the full image to proposal bounding boxes and associated class probabilities. YOLO is very fast because a single convolution neural network is used for learning the regression in one evaluation from the full image. However, YOLO sometimes loses precision in predicting proposal location due to the coarse 7×7 grid formation for the full image. Liu et al. [19] proposed a Single Shot Multibox Detector (SSD) method, which takes advantage of both YOLO and Faster R-CNN to reach a balance between proposal precision and computational efficiency. SSD also uses a single convolution neural network and adjusts the bounding box by considering local features to better match the object shape. Kwon and Lee [20] proposed an edge-based proposal generation method by developing an edge fields (EFs) technique that first binarizes the color channel images with various thresholds for keeping the salient regions and then blurs the image to extract high-quality proposals. Finally, several heuristics for measuring objectness are proposed to remove

duplicated bounding boxes and score the retained proposals. Kalampokas et al. [21] tested eleven CNN models by combining three different feature-learning sub-networks with five meta-architectures to segment vineyard images into semantic clusters of white grapes, red grapes, and leaves. They found out that the model combining the ResNet50 sub-network with the Full-Resolution Residual Network meta-architecture was the best among all the tested models for achieving semantic segmentation for the automation of harvest.

For the proposal refinement methods, DeepBox [10] proposes a four-layer CNN architecture for refining the proposals produced by existing object proposal approaches. DeepBox is a re-ranking method considering high-level structures to compute the objectness of candidate proposals and re-rank proposals using the computed objectness. The authors show with experiment results that DeepBox improves the bottom-up proposal ranking approach by achieving the same recall with 500 proposals as achieved by bottom-up ranking with 2000. Zhong et al. [9] proposed a re-ranking method for object proposals by reference to class-specific feature weights learned by Structured SVM. The re-ranking method can be combined with existing object proposal generators to enhance the recall rate. The considered re-ranking features are semantic segmentation, stereo information, contextual information, objectness measured by DeepBox, and the low-level cue produced by the original generator. It was shown with the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) benchmark that the re-ranking method can significantly improve the recall rate even with fewer proposals. Refinedbox [11] is a proposal refinement method that builds a computationally lightweight neural network and uses the initial object proposal boxes as input. The network has two branches, namely, re-ranking and box regression. The re-ranking branch learns the probability of each box being a true proposal or not by calculating its objectness. The box regression branch learns the offset of the position and size between the initial box and the true box. Ke et al. [8] first apply Selective Search to merge super-pixels hierarchically in order to produce redundant bounding boxes. Then, a uniform, fully CNN is used to extract the object properties, namely, the deep contour and symmetry. Finally, Bayesian scoring is used to score each bounding box by considering the corresponding contour and symmetry.

3. Proposed Methods

In this section, we present our system overview and elucidate the corresponding functional processes.

3.1. System Overview

The overview of our object proposal and recognition system is shown in Figure 1. The system consists of two main parts, proposal generation and refinement and context-dependent learning. The proposal generation and refinement part contains three processes. Firstly, the fast edge detection process originated from EdgeBoxes [7] is applied to produce the edge image. Secondly, the low-level cue refinement process is used to refine the edges by applying low-level image processing. Thirdly, the context-dependent refinement process taking advantage of our application context can effectively reduce the number of proposals while still preserving a high recall of true objects. The context-dependent learning part also consists of three processes. We extract the Scale-Invariant Feature Transform (SIFT) descriptors [22] of all the training images and use them to construct a codewords dictionary according to the Bag of Words (BoW) model [23]. Therefore, each training image can be represented by reference to the codewords dictionary. The support vector machine (SVM) is employed as our drink commodity classifier, and it is trained with the codewords representation of all the training images. All of the noted processes in our system are articulated in the following sections.

Figure 1. Overview of our object proposal and recognition system.

3.2. Fast Edge Detection

The Fast Edge Detection method [24] employed in EdgeBoxes [7] improves the edge map by grouping related edges together based on sketch tokens, which are mid-level representations for object contour. Various structures of 16×16 dimension edge patterns are used to define the sketch tokens. The image patches that really contain object contour are then classified as one of the sketch tokens. The trained classifier is a structured random decision forest that stochastically samples a set of image patches with replacement from the training set and uses the samples to train a decision tree. As there are multiple decision trees in the random forest, the sampling and training process is repeatedly practiced until the classifier is well trained. The random decision forest has a good capability for generalization because it trains multiple decision trees with random training samples and aggregates their results as the final classification. The Fast Edge Detection method of EdgeBoxes is suited to our application for drink commodity proposal and recognition since most texture-like edges are ignored by the Fast Edge Detection method and the texture-like edges are often features of the background such as wooden surfaces and bushes. As shown in Figure 2, we apply several edge detectors,

namely, Sobel, Canny, and Fast Edge Detection, to an image containing a drink commodity on a table with a texture surface and bush background; the Fast Edge Detection ignores most uninteresting texture edges as compared to Sobel and Canny.

Figure 2. The comparative results of various edge detectors. (a) Original input image; (b) The edge image obtained by the Sobel operator; (c) The edge image obtained by the Canny operator; (d) The edge image obtained by the Fast Edge Detection method.

3.3. Low-Level Cue Refinement

This step considers only low-level cues to refine the quality of potential object edges. Firstly, the edge image obtained by the Fast Edge Detection method is binarized to generate generic edge components. Secondly, a morphology dilation-and-erosion approach is used to fill the gap between very near edge components. In particular, the dilation operator is applied three times with a 3×3 square structure element to expand the borders of all the edge components such that two edge components with an inter-distance less than six pixels will be merged into one. At this stage, very small edge components in terms of the number of edge pixels can be removed because they cannot suffice for rendering a meaningful object. The erosion operator is applied after further context-dependent refinement heuristics are performed as noted in the next section. Figure 3 shows the intermediate results from the low-level cue refinement. Figure 3a is an input image, whose extracted edge image, the binarized edge image, and the edge image after dilation are shown in Figure 3b–d, respectively. Apparently, some texture edges were ignored and small gaps between segments were filled. However, the edge components obtained by the low-level cue refinement may not all represent meaningful candidates for rendering drink commodities such as the components enclosed in the red box in Figure 3c, which require further checking by the high-level context-dependent refinement heuristics.

Figure 3. The intermediate results obtained after various refining processes. (a) Original input image; (b) Extracted edge image; (c) Binarized edge image; (d) Expanded edge image obtained after dilation; (e) Retained edge components after symmetry checking; (f) Shrunk edge image obtained after erosion.

3.4. Context-Dependent Refinement

The context-dependent refinement heuristics take advantage of the application context in order to effectively remove false object proposals. For our application context, the drink commodity has several useful geometric properties that can be employed to evaluate the confidence of a proposal. We articulate the context-dependent refinement heuristics in the following sections.

3.4.1. Symmetry

From our collection of drink commodity images containing various categories of carton packages or PET bottles, we observe that the lengths of the two opposite sides of a drink object are nearly the same. Due to the quality of edge extraction, the lengths of the two opposite sides of the drink's edge component may be not the same, but they should be still comparable. Therefore, we design a symmetry-checking heuristic as follows. For each edge component, we first horizontally and then vertically scan the component and examine the number of crossings. If the component is formed by the edge of a drink commodity, more than one crossing is expected between the scanning line and the component. We define such a scanning line as a successful scanning; otherwise, it is a failed scanning. We count the number of successful scanings both horizontally and vertically. The edge component passes the symmetry-checking heuristic if the number of successful horizontal scanings or vertical scanings is greater than a threshold. On the other hand, if the component does not pass the symmetry-checking heuristic, it is considered as an absence of a drink commodity and should be removed from our candidate proposals. Figure 3e shows the retained edge components after applying the symmetry-checking heuristic. It is seen that the edge components enclosed in the red box in Figure 3c have been completely removed by the symmetry-checking heuristic. Before we introduce other context-dependent heuristics, the edge components are shrunk by applying erosion three times with a 3×3 square structure element to restore their correct sizes as shown in Figure 3f.

3.4.2. Size, Density, and Aspect Ratio

Next, the minimum bounding box for each retained edge component is located. We name the minimum bounding box the region of interest (RoI). The drink commodity usually has a complex label design pasted on the outside of the carton package or PET bottle in order to attract customers' attention. As a result, the density of the edge points within the RoI containing a drink commodity is higher than that in other RoIs. Moreover, the size of the RoI should be large enough to contain a drink commodity. In Figure 4, we plot the size and density of all the generated RoIs for our collection of images. It is seen that all true RoIs containing a drink commodity have a density greater than 0.4, and their size is greater than 2500, although a few false RoIs can also pass the two conditions. We thus deploy the context-dependent refinement heuristic by stipulating that the size and density of the RoIs should be larger than 2500 and 0.4, respectively. Finally, we calculate the range of the aspect ratio for true RoIs and remove the RoIs whose aspect ratios are outside of this range.

Figure 4. The size and density of all generated regions of interest (RoIs). The red dots indicate the true RoIs containing a drink commodity, while the blue dots represent the false RoIs without a drink commodity inside.

3.4.3. Merging

At this stage, the retained RoIs have a high likelihood of containing or overlapping with a drink commodity object. To further refine the location and size of the output object proposals, it is desired to merge the near RoIs that correspond to the same drink commodity object. The context-dependent semantics we refer to are that if the drink commodity object is broken into several RoIs during our previous processes, these RoIs should be near and they share some homogenous property. In our observations, every drink label is printed in similar colors with dark and light variations. Hence, we employ the hue intensity in the HSV color space to measure the similarity between near RoIs, because the hue intensity of a color is insensitive to dark and light variations. For the proximity measure, we consider two cases as shown in Figure 5. In the first case (see Figure 5a), one RoI overlaps or contains another RoI. In the other case (see Figure 5b), two RoIs are within proximity. We consider the RoIs in both cases as being near. Thus, our context-dependent merging heuristic will merge the RoIs if they have similar color hues and meet one of the two proximity cases.

Figure 5. The merging of near and homogeneous RoIs. (a) An RoI overlaps or contains another RoI. (b) Two RoIs are within a proximity.

An illustrative example for each of the two proximity cases is shown in Figure 6, where we focus on the blue RoI and we check its proximity relationship with the pink RoI. In Figure 6a, the blue RoI overlaps with the pink RoI, so they meet the first proximity criterion and should be merged. Meanwhile, in both cases of Figure 6b, the blue RoI does not overlap with the pink RoI. An extended line (as indicated as the broken yellow line) is drawn along each side of the blue RoI. If the extended line intersects with the pink RoI and the distance between the two RoIs is less than a threshold (see the white arrow in the left case of Figure 6b), then the two RoIs meet the second proximity criterion and are merged. If the extended line does not intersect with the pink RoIs, however, the four sides of the rectangle that is formed between the nearest corners of the two RoIs all have lengths less than a threshold (see the green rectangle in the right case of Figure 6b); the two RoIs are also considered to meet the second proximity criterion and should be merged.

Figure 6. The illustration of the two RoI-merging proximity criteria. (a) The first proximity criterion. (b) The second proximity criterion.

To obtain the best parameters of the proposal generation and refinement method, we intensively explored various values for the parameters such as the threshold for image binarization, size of the morphology structure element, threshold for successful symmetry crossings, threshold for size, density and aspect ratio of the proposals, and threshold for the RoI-merging proximity criteria. Another promising direction is the adoption of an ensemble approach [25] that is able to automatically determine the best parameters on the fly but, however, is computationally expensive. Our search results for parameter values can be used as initial settings for the ensemble approach to automatically determine the best parameter values.

3.5. Learning and Recognition

To proceed with drink commodity recognition, the adopted classifier should learn from the true RoIs containing a drink commodity. These true RoIs are segmented and labeled by an expert who has worked in the field of object proposal for two years. The context-dependent learning consists of feature extraction, image representation, and classifier training. For feature extraction, we adopt the local feature descriptor, namely, the scale-invariant feature transform (SIFT) [22] because SIFT is robust in the face of scale, rotation, and illumination variations. SIFT constructs several octaves to generate a Gaussian pyramid. For each octave, there are several convolution layers with the Gaussian kernel. Hence, the octave can generate a set of feature images with different smoothing degrees from fine to coarse. The next octave is a down-sampling from its immediately previous octave. All these octaves form a Gaussian pyramid containing many feature images under cascading smoothing and scaling processes. To locate the high-quality keypoints for object matching, another pyramid named Difference of Gaussian (DoG) is constructed from the Gaussian pyramid. The layer of an octave in the DoG is the difference between the corresponding pair of adjacent layers in the octave of the Gaussian pyramid. The keypoints are the maxima and minima values over neighboring points in the same and adjacent layers in the DoG. SIFT considers pixels around a radius of the keypoints for matching, and it is robust with local affine variation because the keypoints are detected in a varying smoothed and scaled pyramid. In this paper, we consider pixels within a 16×16 window for each keypoint and employ a Gaussian radius for gradient weighting. The mean gradients along eight orientations are computed for each 4×4 sub-window, generating, in total, 128 gradients as our local SIFT descriptor for each keypoint. These local SIFT descriptors are used to represent the true RoIs in the training set.

Although SIFT is robust with local affine and blurring variations, it may make the subsequent recognition process time-consuming if the number of keypoints extracted from the image is great. Many existing studies [23,26,27] have proposed implementations to expedite the recognition process based on SIFT local features. One of the prevailing implementations is combining SIFT and the Bag of Words (BoW) model [23]. We adopt this implementation as noted later.

SIFT has been combined with other heuristics to enhance its robustness. For example, Heikkilä et al. [28] presented a novel RoI descriptor by combining SIFT with the Local Binary Patterns (LBP) texture operator. The new descriptor is called the center-symmetric local binary pattern (CS-LBP) descriptor, which has been justified with a recent test protocol. The authors showed that the CS-LBP descriptor outperforms the SIFT descriptor in terms of tolerance to illumination changes and robustness on flat image areas. In this paper, we do not combine SIFT with texture operators. Instead, we employ the BoW model, whose early versions such as Leung and Malik [29] have taken texture recognition into account and achieved excellent performance. The BoW model recognizes texture patterns by using a histogram of visual words to record the multiplicity of salient texture.

Our image representation scheme follows the BoW model. The SIFT descriptor for each true RoI is considered as a visual word. As there are many visual words, the vector quantization technique needs to be applied to construct a codeword dictionary of which the number of codewords is significantly lower than the number of visual words produced from all the true RoIs in the entire training set. We apply K-means as our vector quantization technique to generate the cluster centroids as the codewords for the codeword dictionary formation. Each visual word (SIFT descriptor) contained

in a true RoI is assigned to the most relevant codeword according to the dictionary. The histogram, which counts the number of visual words assigned to each codeword, is used to represent the RoI for object recognition.

The support vector machine (SVM) is employed as our classifier for the classification of each drink commodity. The reason we chose SVM is because SVM has been coupled with histogram-based features and produced satisfactory results for learning object categories [23,30]. In this paper, the radial basis function (RBF) is used as the kernel function for SVM to map the sample vectors to a hyper-dimensional feature space. The RBF is defined as follows.

$$K(x_1, x_2) = \exp(-\gamma(x_1 - x_2) \cdot (x_1 - x_2)) \quad (1)$$

It is used to calculate the distance between two sample vectors x_1 and x_2 mapped to a hyper-feature space with the RBF parameter γ . The larger the RBF parameter γ , the fewer the retained support vectors in the SVM classifier. Our current image dataset is not large enough to conduct a meaningful cross-validation process to determine the best parameters for the SVM; therefore, we search for the best parameters in representative value combinations. It will be worth studying in our future research a comparison of our current implementation with other alternative classifiers and training the classifiers with a larger dataset.

The histogram of each true RoI is a sample vector, and the SVM is trained by all the sample vectors in the corresponding training set. Finally, in the recognition step, the segmented and refined proposals from a test image are sent to the trained SVM for recognition.

4. Experimental Results and Discussions

The experimental platform is a personal computer with a 1.7 GHz CPU and 4 GB RAM; the programs were codified in MS Visual Studio using the C/C++ language.

4.1. Performance of Proposal Generation and Refinement

We generate a test image dataset that contains 220 images. These images were obtained by taking pictures of various drink commodities in indoor, outdoor, or natural environments, with different scales, orientations, background complexities, and illumination. The test dataset poses different degrees of difficulty for proposal generation and refinement. In order to realize which object appearance criteria present the highest challenges to our method, we manually divide these images into three test subsets by reference to criteria such as the object size, viewing angle, occlusion and truncation ratio, illumination variation, and background complexity. These criteria are the most common challenges for humans and computers when performing object detection. Shown in Figure 7 are some typical test images for each subset. The number of images and drink objects in each subset is tabulated in Table 1.

Table 1. The number of images and drink objects in each test subset.

	Easy	Moderate	Hard
No. of images	92	107	21
No. of drink objects	110	134	27

Figure 7. The exemplar images in each test subset. (a) Easy; (b) Moderate; (c) Hard.

Figure 8 shows an example in each subset for the proposals generated by our method. It is seen that each drink commodity in the image is overlapped with some generated proposals, benefiting the subsequent recognition task. There are still some false object proposals such as the chair armrest, chair back, handbag, and wooden batten. Nevertheless, the number of proposals generated by our method is much lower than that obtained by the proposal generation method without adopting context-dependent proposal refinement heuristics, such as Selective Search [6] and EdgeBoxes [7]. Taking the test image in Figure 8c as an example, the numbers of proposals generated by the proposed method, Selective Search, and EdgeBoxes, are 24, 59, and 6824, respectively. Our method significantly reduces the number of false-positive proposals and achieves better recall.

Figure 8. Some examples of the proposals generated by our method. (a) Easy; (b) Moderate; (c) Hard.

A common criterion for measuring the comparative performance of object proposal methods is the Intersection over Union (IoU). The IoU is defined as the percentage of a true object covered by the proposals. The performance of object proposal methods is thus measured by comparing the recall of true objects at various IoU thresholds. Figure 9 shows the recall vs. IoU threshold for our approach, Selective Search, and EdgeBoxes, for the three subsets, Easy, Moderate, and Hard, respectively. It is observed that our approach performs much better than Selective Search and EdgeBoxes for the two subsets Easy and Moderate. For a commonly used criterion with the IoU threshold set to 0.5, our approach obtains 0.9 and 0.75 recall rates for Easy and Moderate subsets, while the Selective Search can only reach 0.35 and 0.45, respectively. In our experiments, EdgeBoxes often performs worse than the other two competitors when a low IoU threshold is specified, and it outperforms Selective Search as the IoU threshold increases. This demonstrates that the whole list of proposals generated by EdgeBoxes have lower likelihoods of being true objects as compared to those generated by our approach and Selective Search. When a higher IoU threshold is stipulated, the confidence for the top proposals in the list generated by EdgeBoxes surpasses that for that generated by Selective Search. For the Hard subset, our approach outperforms Selective Search, with a higher recall at every IoU threshold. However, EdgeBoxes outperforms our approach in recall when the IoU threshold is higher than 0.25. The reason is because our approach applies context-dependent refinement heuristics including symmetry, size, density, and aspect ratio to rule out non-drink-like objects. Although these heuristics successfully reduce the number of proposals for Easy and Moderate subsets, they may remove some obscure true proposals for very complex images in the Hard subset. It is worth further studying the design of context-dependent refinement heuristics that maintain a good balance between the number of proposals and the image complexity.

Figure 9. Recall vs. Intersection over Union (IoU) threshold for our approach, Selective Search, and EdgeBoxes, for the three subsets: (a) Easy; (b) Moderate; (c) Hard.

4.2. Performance of Object Learning and Recognition

As articulated in Section 3, our context-dependent learning trains an SVM classifier that recognizes seven categories of popular drink commodities in Taiwan. The training images for these drink categories were carefully generated with a white background and good illumination. As the test images could be generated under many real situations that result in various object scales and viewing angles, the preparation of training images should take these variations into account. Tai [31] generated the training images for drink carton packages and PET bottles by manually cutting off the packages and bottles and flattening them to generate panoramic images. Thus, the detection method can tolerate different viewing angle variations. However, the creation of the training dataset is very labor intensive. We decided to adopt another approach, which minimizes the human effort but still exhibits a high tolerance for scale and angle variations, as follows. Twenty-one pictures for each drink category were taken, with three different scales and seven distinctive angle views per scale. Hence, there were a total of 147 ($7 \times 3 \times 7$) training images. Preliminary experiments were conducted with different combinations of scales and angles. We empirically found that the following setting is best in the tradeoff between human effort and recognition results. The image scales were determined by setting the length of the longer side of the drink commodity to 200, 350, and 600 pixels, and the viewing angles were set to -90° , -60° , -30° , 0° , 30° , 60° , and 90° offsets to the front face of the drink commodity. The true drink RoI for each training image was segmented and labeled by an expert who had worked in the field of object proposal for two years. Figure 10 shows the seven angle views for each category of drink commodities at one of the scales.

Figure 10. Seven angle views for each category of drink commodities at one of the scales.

Following the BoW model, a codeword dictionary was constructed based on the SIFT descriptor for all true drink RoIs in the entire training set. As such, a histogram counting the number of visual words assigned to each codeword could be used to represent the true drink RoI for training the SVM for recognition. After the SVM was trained, its performance was evaluated with the entire set of proposals that were generated by our proposal generation and refinement method as noted in the previous section. We adopt three performance metrics widely used in the object recognition literature, namely, the Precision, Recall, and Accuracy, as defined as follows.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

where TP and FP denote true positives and false positives, and TN and FN indicate true negatives and false negatives. The results of the performance measures for the SVM are listed in Table 2. For each subclass of test images, the number of true objects and the number of proposals generated by our proposal generation and refinement method are displayed. It is seen that the number of generated proposals is lower than or approximately equal to double of the number of true objects, significantly reducing the effort for the subsequent recognition of the proposals. The precision for all the three subsets is above 0.5, Recall is above 0.82, and Accuracy is above 0.95. The high recall demonstrates that most of the true objects are correctly recognized by the SVM due to the good quality of the proposals generated by our method. Some such examples in the three subsets are shown in Figure 11, where the location and size of the true objects are correctly captured. Another interesting case is that the SVM with the BoW model can correctly recognize the drink commodity even when the IoU of the object proposal is low. Figure 12 shows such an example where the generated proposal is significantly larger than the true drink object, resulting in an IoU less than 0.4. Although the location and size of the proposal may be not very accurate, the SVM still correctly recognizes the object category because the BoW model extracts the frequency of visual words instead of their location and size.

Table 2. The recognition performance measures for the SVM with the three test subsets.

	No. of True Objects	No. of Proposals	Precision	Recall	Accuracy
Easy	110	156	0.80	0.88	0.97
Moderate	134	241	0.50	0.82	0.95
Hard	27	58	0.56	0.90	0.97
All	271	455	0.60	0.85	0.96

As we adopted the BoW model to transform the SIFT local features to a histogram of visual words, the computational time for recognition on the three different datasets was significantly shorter than that by direct recognition with the SIFT local features. The computational time for recognizing a test image ranges from 0.03 to 0.13 s depending on the remaining keypoints in the test image. Thus, the calculation for extracting the SIFT keypoints is the main consumer for our computation rather than the SVM recognition. In general, the computational time is longer if the test image belongs to a more challenging category in the order of Easy, Moderate, and Hard. However, there is no clear-cut difference in the computational time between different image categories because we did not create the image categories by reference to the numbers of extracted keypoints.

Figure 11. Some examples of the proposals correctly recognized by our method. (a) Easy; (b) Moderate; (c) Hard.

Figure 12. The SVM with the BoW model can correctly recognize the drink commodity even when the IoU of the object proposal is low.

The high accuracy shown in Table 2 reveals another fact that the SVM not only correctly recognizes true drink objects but also successfully identifies the true-negative proposals that actually contain no drink commodities. This can be realized by comparing Figure 11c to Figure 8c, where the latter contains 24 proposals but only two of them are actually drink objects and correctly recognized; the rest are identified by the SVM as non-drink objects (i.e., true negatives). The general performance statistics with all the images in the three subsets are provided at the bottom of Table 2. Although the recall and accuracy are satisfactory, the precision still has room for improvement.

The current challenge to our method is that the precision performance of the SVM drops from 0.8 with the Easy subclass to 0.5 with the Moderate subclass and 0.56 with the Hard subclass. The low precision with the Moderate and Hard subclasses is due to many false-positive RoIs, which contain color regions looking similar to the label design pasted on the outside of the drink carton packages or PET bottles. In particular, there are 169 false-positive RoIs in the recognition results for all the images. Figure 13 shows such examples for the Moderate and Hard subclasses. The left RoI contained in Figure 13a is a drink object; however, it is incorrectly recognized as another drink commodity by the

SVM. The right RoI in Figure 13a is not a drink object, but it is mistakenly recognized by the SVM as a drink object. For the more complex case in Figure 13b, it is worth noting that the four RoIs all contain color regions similar to different drink label designs. The lower-left three RoIs are recognized as labeled drink cartons, and the lower-right RoI is recognized as a drink PET bottle. Those false-positive RoIs as illustrated in Figure 13 show that our future research direction should enhance the proposal refinement heuristics and augment the size of the training set for SVM by including more images with complex scenes.

Figure 13. Some examples of the false-positive RoIs incorrectly recognized by the SVM. (a) Moderate; (b) Hard.

5. Conclusions

In this paper, a novel context-dependent object proposal and recognition method has been proposed. Object proposal generation is emerging as a mandatory form of preprocessing for efficient and accurate object recognition, and it has been adopted by many methods in object recognition competitions such as PASCAL VOC 2007 and ILSVRC. Most of existing object proposal methods are crippled by generating too many proposals in order to maintain a satisfactory recall of true objects. To remedy this drawback, we developed proposal refinement heuristics based on both low-level cues and context-dependent features. In particular, the context of the drink commodity is considered to extract useful features such as symmetry, size, density, and aspect ratio. The proposal refinement heuristics assist us to remove many false proposals. We further developed an SVM based on the BoW model with SIFT descriptors to recognize the proposals. A test dataset containing indoor and outdoor images with the presence or absence of various drink commodities was generated. The experimental results show that our context-dependent method generates many fewer proposals than Selective Search and EdgeBoxes, with the same recall level, implying our method can save more computational time spent in the recognition task. For the performance of the SVM, at least 82% of true objects are correctly recognized for test subsets of various challenging difficulties. The number of false positives may be still high for some test images with very complex scenes.

Our future research will aim at the following directions. (1) The improvement of our context-dependent refinement heuristics to help to effectively remove the false proposals in the first phase, and the use of a large training set to assist the SVM to reduce the number of false positives in the second phase. To obtain the best parameters for our method, the ensemble approach can be adopted to automatically determine the best parameters by using the parameters empirically found in this paper as initial settings. (2) As previously noted, SIFT has been combined with other heuristics such as texture measures. It is worth investigating whether these SIFT implementations increase the robustness of our method against illumination changes and texture variations. (3) The comparison for the use of the same strategy with different classifiers or deep-learning techniques could find the

best coupling methodology, and the use of a large dataset would benefit the training of classifiers based on SIFT. (4) For the recognition of PET bottles, the variations in plastic reflection, liquid quantity, and high transparency may cause misclassifications for the PET bottles. Our future work can consider only the label part of the PET bottles for training instead of the whole bottle. This could improve our recognition results.

Author Contributions: Conceptualization, R.-I.C. and C.-L.T.; methodology, R.-I.C.; software, S.-Y.W.; validation, R.-I.C. and P.-Y.Y.; writing—original draft preparation, P.-Y.Y.; writing—review and editing, P.-Y.Y.; visualization, S.-Y.W.; funding acquisition, R.-I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MOST Taiwan, grant numbers 109-3116-F-002-002-CC1 and 108-2410-H-002-230-MY2. The APC was funded by 109-3116-F-002-002-CC1 and 108-2410-H-002-230-MY2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hough, P.C.V. Method and Means for Recognizing Complex Patterns. U.S. Patent 3,069,654, 18 December 1962.
2. Romberg, S.; Pueyo, L.G.; Lienhart, R.; Zwol, R. Scalable Logo Recognition in Real-World Images. In Proceedings of the ACM Proceedings of the International Conference on Multimedia Retrieval, Trento, Italy, 18–20 April 2011.
3. Wang, Y.; Huang, L.; Ren, T.; Zhong, S.H.; Gu, H.; Liu, Y. Insights of object proposal evaluation. *Multimed. Tools Appl.* **2019**, *78*, 13111–13130. [CrossRef]
4. Hosang, J.; Benenson, R.; Schiele, B. How good are detection proposals, really? In Proceedings of the BMVC, Nottingham, UK, 1–5 September 2014.
5. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef] [PubMed]
6. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
7. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 391–405.
8. Ke, W.; Chen, J.; Ye, Q. Deep contour and symmetry scored object proposal. *Pattern Recognit. Lett.* **2019**, *119*, 172–179. [CrossRef]
9. Zhong, Z.; Lei, M.; Cao, D.; Fan, J.; Li, S. Class-specific object proposals re-ranking for object detection in automatic driving. *Neurocomputing* **2017**, *242*, 187–194. [CrossRef]
10. Kuo, W.; Hariharan, B.; Malik, J. Deepbox: Learning objectness with convolutional networks. In Proceedings of the International Conference on Computer Vision, ICCV, Santiago, Chile, 7–13 December 2015; pp. 2479–2487.
11. Liu, Y.; Li, S.; Cheng, M.M. Refinedbox: Refining for fewer and high-quality object proposals. *Neurocomputing* **2020**, *406*, 106–116. [CrossRef]
12. Department of Statistics, Ministry of Economic Affairs. Sales of Wholesale, Retail and Food Services in June 2020. Available online: https://www.moea.gov.tw/Mns/dos/content/ContentLink.aspx?menu_id=9431 (accessed on 1 August 2020).
13. Day, R.F.; Yin, P.Y.; Wang, Y.C.; Chao, C.H. A new hybrid multi-start tabu search for finding hidden purchase decision strategies in WWW based on eye-movement. *Appl. Soft Comput.* **2016**, *48*, 217–229. [CrossRef]
14. Zhang, Z.; Torr, P.H. Object proposal generation using two-stage cascade SVMs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 102–115. [CrossRef] [PubMed]
15. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. Bing: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA, 23–28 June 2014; pp. 3286–3293.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Neural Information Processing Systems, NIPS, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
17. Girshick, R. Fast_R-CNN. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1440–1448.

18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [\[CrossRef\]](#)
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [\[CrossRef\]](#)
20. Kwon, J.; Lee, H. Visual tracking based on edge field with object proposal association. *Image Vis. Comput.* **2018**, *69*, 22–32. [\[CrossRef\]](#)
21. Kalampokas, T.; Tziridis, K.; Nikolaou, A.; Vrochidou, E.; Papakostas, G.A.; Pachidis, T.; Kaburlasos, V.G. Semantic segmentation of vineyard images using convolutional neural networks. In Proceedings of the 21st International Conference on Engineering Applications of Neural Networks (EANN 2020), Halkidiki, Greece, 5–7 June 2020; pp. 292–303.
22. Lowe, D.G. Object recognition from local scale-invariant features. *Int. Conf. Comput. Vis.* **1999**, *2*, 1150–1157.
23. Lee, F.F.; Fergus, R.; Torralba, A. Recognizing and learning object categories: Year 2007. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007.
24. Dollár, P.; Zitnick, C.L. Structured forests for fast edge detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1841–1848.
25. Yin, P.Y.; Tsai, C.C.; Day, R.F.; Tung, C.Y.; Bhanu, B. Ensemble learning of model hyperparameters and spatiotemporal data for calibration of low-cost PM2.5 sensors. *Math. Biosci. Eng.* **2019**, *16*, 6858–6873. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Conf. Vis.* **2004**, *60*, 91–110. [\[CrossRef\]](#)
27. Lu, C.H. Improving Efficiency and Accuracy for SIFT Streetscape Recognition System. Master's Thesis, Institute of Engineering Science and Ocean Engineering, National Taiwan University, Taipei, Taiwan, 2012.
28. Heikkilä, M.; Pietikäinen, M.; Schmid, C. Description of Interest Regions with Center-Symmetric Local Binary Patterns. In *Lecture Notes in Computer Science*; Kalra, P.K., Peleg, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4338, pp. 58–69.
29. Leung, T.; Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.* **2001**, *43*, 29–44. [\[CrossRef\]](#)
30. Fergus, R.; Fei-Fei, L.; Perona, P.; Zisserman, A. Learning object categories from Google's image search. In Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005.
31. Tai, C.C. A Beverage Package Recognition System for Mobile Application. Master's Thesis, Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan, 2011.