Local Region Proposing for Frame-Based Vehicle Detection in Satellite Videos

Junpeng Zhang, Xiuping Jia * and Jiankun Hu

School of Engineering and Information Technology, University of New South Wales, Canberra 2612, Australia; junpeng.zhang@student.unsw.edu.au (J.Z.); J.Hu@adfa.edu.au (J.H.)

* Correspondence: x.jia@adfa.edu.au

Received: 31 August 2019; Accepted: 8 October 2019; Published: 12 October 2019

Abstract: Current new developments in remote sensing imagery enable satellites to capture videos from space. These satellite videos record the motion of vehicles over a vast territory, offering significant advantages in traffic monitoring systems over ground-based systems. However, detecting vehicles in satellite videos are challenged by the low spatial resolution and the low contrast in each video frame. The vehicles in these videos are small, and most of them are blurred into their background regions. While region proposals are often generated for efficient target detection, they have limited performance on satellite videos. To meet this challenge, we propose a Local Region Proposing approach (LRP) with three steps in this study. A video frame is segmented into semantic regions first and possible targets are then detected in these coarse scale regions. A discrete Histogram Mixture Model (HistMM) is proposed in the third step to narrow down the region proposals by quantifying their likelihoods towards the target category, where the training is conducted on positive samples only. Experiment results demonstrate that LRP generates region proposals with improved target recall rates. When a slim Fast-RCNN detector is applied, LRP achieves better detection performance over the state-of-the-art approaches tested.

Keywords: satellite videos; region proposals; convolutional neural networks; tiny and dim target detection; component mixture model

1. Introduction

As one of the most promising developments in remote sensing imagery, the satellite videos captured by Skybox and JL-1, have facilitated several emerging research and applications, including super resolution [1,2], video encoding [3,4] and target tracking [5,6]. They expand the earth observation capacity to rapid motion monitoring, such as vehicle and ship tracking [5,7,8]. To reveal these rapid motions, targets of interests need to be located throughout the satellite video first, and the extracted targets in each frame are then associated to construct the trajectories of targets of interest. Therefore, target detection in satellite videos is a fundamental and critical step for target tracking and motion pattern analysis.

Detecting objects of interest in a video can be achieved by the motion-based detectors, which search the changed pixels in a sequence of images by comparing with an estimated background model [9,10]. Various algorithms, such as Frame-Difference [5,11,12], Median Background [13], Gaussian Mixture Model (GMM) [14,15] and Visual Background Extractor (ViBe) [7,16,17], were developed for moving object detection. However, these approaches are prone to the inadequate background modelling and affected by the problem of parallax caused by the motion of the camera.

Alternatively, the image-based object detectors can extract objects of interest from a video frame by frame [18], whose performance is less affected by the parallax motion. By taking the advantage of the discriminative learning methods, these approaches employ a classifier to scan over possible
locations of targets in an image by sliding window [19–21]. To reduce the number of the candidate locations to examine, region proposals, which refer a sparse set of potential target locations, are introduced to replace sliding windows over the entire image. For common computer vision tasks, generating region proposals are commonly guided by the object saliency, such as the edges [22–24], or based on superpixels [25–29] or segmentation masks [30,31]. In aerial videos, the coherent regions extracted by Maximally Stable Extremal Regions (MSER) [32,33] or Top-hat-Otsu [34] are also adopted for region proposal generation. Due to the weak contrast between targets and background in satellite videos, saliency-based approaches result in degraded region proposal performance—either generating too many region proposals or producing a low target recall rate. These approaches also lack the mechanisms for quantifying the region proposals’ likelihood of being a target, and place the entire burden of handling a large number of region proposals in the target recognition stage. Convolutional Neural Networks were applied for searching region proposals in recent years. These approaches can provide the confidence score for each region proposal, and a significant portion of false alarms in the region proposals are removed before the recognition stage [35–38]. However, they heavily rely on the training of a reliable region proposal network using a large amount of training samples.

To improve the region proposal performance to handle dim and small target detection in satellite video, we propose a Local Region Proposing (LRP) approach with three steps in this study. Our observation is that vehicles in satellite videos appear small and dim globally. Therefore we propose to perform segmentation at a coarse scale to form semantic region first. Possible locations of small targets in each semantic region are then extracted. To reduce the false alarm further and alleviate the computation burden on further target recognition stage, a discrete Histogram Mixture Model (HistMM) is proposed to quantify their likelihoods towards the target category. HistMM presents little difficulty in cooperating with most detectors, as it is estimated separately and only positive samples are required for estimating the model.

The remaining part of this paper is structured as follows. Section 2 presents the proposed local region proposal approach, after which the experimental results are presented in Section 3. We conclude this paper in Section 4 with remarks on the promising direction for future study.

2. Local Region Proposing

Figure 1 shows the Local Region Proposing approach (LRP) developed in this study is composed of three steps. First semantic regions are extracted by coarse-scale segmentation, then possible target locations are searched in each extracted region. The Histogram Mixture Model is developed for removing obvious false alarms from the region proposals.

![Figure 1](image)

**Figure 1.** Overview of the proposed region proposal algorithm.

2.1. Semantic Region Extraction

Extracting semantic regions from a video frame can be by segmentation at a coarse scale, and the majority of pixels in each extracted region are more likely from a single land cover type.
The Felzenszwalb’s graph-based segmentation approach [39] is a typical method for extracting the semantic regions.

By this graph-based segmentation approach, the scale of the generated superpixels can be controlled by a parameter $k$. Increasing $k$ would lead to more coarse-scale superpixels, and these superpixels tend to present regions from different land cover types. The semantic regions are allowed to be larger than the target size on purpose. Decreasing $k$ would generate fine-scale superpixels. However, it is often difficult to make superpixels to associate with small targets in satellite videos, due to the low spatial resolution and the low contrast of targets, for example, vehicles, to the background in satellite videos.

2.2. Searching Possible Locations in Semantic Regions

Unlike most dominating saliency object-based approaches, such as Selective Search [26,40], which merge superpixels to form region proposals, the proposed LRP searches region proposals inside semantic regions, where an adaptive threshold is introduced to accommodate the statistics of individual regions.

Note the set of extracted semantic regions as $\mathcal{R}$, for a semantic region that contains $m$ pixels, the set of the pixels’ coordinates is noted as $r = \{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\} \in \mathcal{R}$. The intensity of a pixel at location $(x, y)$ is referred to $I(x, y)$. The blobs with high local saliency are constructed by the pixels with intensities over a threshold $thr_r$, $I(x, y) > thr_r(x, y) \in r$. The threshold $thr_r$ is defined by

$$thr_r = \mu_r + f \sigma_r,$$

where $\mu_r$ and $\sigma_r$ are the mean and standard deviation of pixel intensities in this local region $r$. The factor $f$ is the expected saliency against the backgrounds. For each extracted blob, a corresponding boundary box is extracted as a possible location.

In the complex scenarios of satellite videos, this searching strategy may be affected by the presence of crowded vehicles and the blurred boundaries of vehicles, which results in merged proposals or incomplete proposals within an original boundary box extracted. We handle these cases by generating multiple proposals. The large boxes should be divided into sub regions to match the target size approximately and the small boxes should be expanded by half of the target size in each direction as a conservative treatment. Figure 2a shows an example where 4 region proposals are generated. To address those incomplete proposals, as shown in Figure 2b, the given bounding box is expanded in each directions.

![Figure 2](image_url)

**Figure 2.** Generating multiple region proposals from a possible location. The red box refers to the groundtruth, green solid box refers to the extracted possible location, and green dash boxes refer to the generated region proposals. (a) and (b) illustrate two examples of generating region proposals by splitting and expanding original region proposals, respectively.
2.3. Histogram Mixture Model

2.3.1. Histogram Mixture Model for Removing Obvious False Alarms

The proposed Histogram Mixture Model (HistMM) measures the likelihoods of the generated region proposals towards their corresponding target category, so that obvious false alarms could be removed at an early stage. The HistMM is a mixture model built on a set of histograms, and training or estimating HistMM depends only on positive training samples.

Note the entire set of initial region proposals on a video frame as \( X_{rp} = \{ x_0, x_1, \ldots, x_{n_{rp}} \} \), and \( n_{rp} \) is the number of initial region proposal on a given frame. For a region proposal \( \forall x \in X_{rp} \), it is marked as either target or background. We decide if \( x \) belongs to the target category (\( T \)) or the background category (\( B \)) by a Bayesian decision function,

\[
R = \frac{p(T|x)}{p(B|x)} = \frac{p(x|T)p(T)}{p(x|B)p(B)}, \tag{2}
\]

in which \( R \) measures the membership rate of \( x \) belonging to the target category versus belonging to the background category. \( R \geq 1 \) implies \( x \) is a target. The corresponding decision function for \( x \) that belongs to \( T \) can be simplified as

\[
p(x|T) \geq c_t, \tag{3}
\]

where \( c_t \) is a threshold.

The \( p(x|T) \) refers to the likelihood of a region proposal \( x \) to the target category. We model it by a mixture model composed by a set of \( n_H \) histograms, \( H = \{ h_1, h_2, \ldots, h_{n_H} \} \). In this paper, we assume that each histogram contributes equally to the likelihood \( p(x|T) \), therefore, the possibility of a proposal \( r \) that belongs to \( T \) is defined as,

\[
p(x|T) = \frac{1}{n_H} \sum_{i=1}^{n_H} p(x|h_i). \tag{4}
\]

The decision function in Equation (3) can be then interpreted as

\[
p(x|T) = \frac{1}{n_H} \sum_{i=1}^{n_H} p(x|h_i) \geq c_t \Rightarrow \exists h \in H, p(x|h) \geq c_t, \tag{5}
\]

which means the likelihood to at least one histogram \( h_i \) in \( H \) is larger than \( c_t \). On the contrary, a region proposals is a background when all likelihoods toward histograms in \( H \) are less than the threshold \( c_t \), as

\[
p(x|h) < c_t, \forall h \in H. \tag{6}
\]

For a given pair of a region proposal \( x \) and a histogram in \( h \in H \), we appropriate \( p(x|h) \) by the Intersection of Histogram (IoH) between the histogram \( h \) and the histogram extracted from the region proposal \( x \). For simplicity, we employ the Histogram of Color (HoC) for calculating \( p(x|h) \), as

\[
p(x|h) = \text{IoH}(h, \text{HoC}(x)) = \sum \min(h, \text{HoC}(x)), \tag{7}
\]

which sums up the minimum values in all pairs of corresponding bins from \( h \) and \( \text{HoC}(x) \). As shown in Figure 3, the IoHs on HoCs are distinct for distinguishing targets and backgrounds, although less information is provided due to the dim appearance of the vehicles.

Our HistMM removes obvious false alarms by the threshold \( c_t \). A larger \( c_t \) tends to removal more possible false alarms, whereas it also risks abandoning some target instances. A smaller \( c_t \) may improve the coverage of targets in the region proposals, but the remaining number of proposals would be high. The detailed effects of different parameter settings are discussed in Section 3.2.
Figure 3. Histogram of Color can distinguish targets from backgrounds. Region proposal A and B are vehicles, whereas the region proposal C and D are obvious false alarms. For the four selected region proposals, their corresponding HoC are extracted, as shown in the right part of the figure. For A and B, the IoH is high, while both C and D have low IoH due to the extremely low similarities.

2.3.2. Estimating Histogram Mixture Model

For a set of \( n_{rp} \) possible region proposals \( X_{rp} \) on a video frame, we predict a region proposal \( x \in X_{rp} \) as a target or a background by Equation (6), as summarized in Algorithm 1. The complexity for predicting region proposals by HistMM grows linearly with the size of \( X_{rp} \), \( O(n_H \times n_{rp}) \). Therefore, our proposed HistMM is computationally feasible and scalable for the case with a large number of region proposals.

Algorithm 1 Removing Obvious False Alarms by Histogram Mixture Model (HistMM)

| Input: \( X_{rp} = \{x_0, x_1, \ldots, x_{n_{rp}}\} \), \( c_t > 0 \), and \( H = \{h_1, h_2, \ldots, h_{n_H}\} \) |
| Output: \( X_{rp}^1 \) |
| 1: for \( x \in X_{rp} \) do |
| 2: if \( \forall h \in H, p(x|h) \leq c_t \) then |
| 3: Remove \( x \) from \( X_{rp} \). |
| 4: end if |
| 5: end for |
| 6: return \( X_{rp}^1 \) |

HistMM is estimated by a recursive learning algorithm on the positive samples of groundtruths [14,41]. Note the estimated set of histograms by \( \hat{H} = \{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{n_H}\} \), and all the positive samples in the groundtruths is denoted by \( X_{gt} \). For a groundtruth \( x_{gt} \in X_{gt} \), a histogram \( \hat{h}_m \), \( m \in \{1, \ldots, n_H\} \), is updated by

\[
\hat{\pi}_m \leftarrow \hat{\pi}_m + o_m(x_{gt}) \\
\hat{h}_m \leftarrow \frac{\hat{h}_m \times \hat{\pi}_m + \text{HoC}(x_{gt}) \times o_m}{\hat{\pi}_m + o_m},
\]

where \( \hat{\pi}_m \) counts the updates of estimated histogram \( \hat{h}_m \), and, as \( \hat{\pi}_m \) increases, the lower fraction of the new samples are taken into \( \hat{h}_m \). \( o_m(x_{gt}) \) defines the \( x_{gt} \)'s ownership of an estimated histogram \( \hat{h}_m \) as

\[
o_m(x_{gt}) = \begin{cases} 
1, & p(x_{gt}|\hat{h}_m) \geq c_t \text{ and } m = \arg \max_{i \in \{0,1,\ldots,n_H-1\}} p(x_{gt}|\hat{h}_i) \\
0, & \text{otherwise} 
\end{cases},
\]

by which \( o_m(x_{gt}) = 1 \) indicts that the new sample \( x_{gt} \) updates the histogram \( \hat{h}_m \) by Equation (8). Otherwise, \( o_m(x_{gt}) = 0 \) means no nearby histogram component exists for this sample \( x_{gt} \), and a new histogram component \( \hat{h}_{n_H} \) is added to \( \hat{H} \). \( \hat{\pi}_{n_H} \) is then initialized as 1 and the added histogram component \( \hat{h}_{n_H} \) is initialized by \( \text{HoC}(x_{gt}) \). This update procedure continues until it finishes iterating over the groundtruth set \( X_{gt} \), as summarized in Algorithm 2.
Algorithm 2 Training procedure of Histogram Mixture Model (HistMM)

**Input:** \( \mathcal{X}_{gt} = \{x_1, \ldots, x_{n_{gt}}\}, c_t > 0 \)

**Output:** \( \hat{\mathcal{H}} \)

1. for \( x \in \mathcal{X}_{gt} \) do
2.   if \( \exists \hat{h} \in \hat{\mathcal{H}}, p(x|h) \geq c_t \) then
3.     Find the updating histogram \( \hat{h}_m \) and the ownership \( o_m(x) \) by Equation (9).
4.     Update \( \hat{h}_m \) by
5.       \[ \hat{\pi}_m \leftarrow \hat{\pi}_m + o_m(x) \]
6.       \[ \hat{h}_m \leftarrow \frac{\hat{h}_m \times \hat{\pi}_m + HoC(x) \times o_m}{\hat{\pi}_m + o_m} \]
7.   else
8.     Initialize a new component by \( HoC(x) \), and add it to \( \mathcal{H} \).
9.   end if
10. end for
11. return \( \hat{\mathcal{H}} \)

3. Experimental Results

3.1. Datasets

Two satellite video datasets, SkySat-Las Vegas dataset and SkySat-Burj Khalifa dataset, were used for experimental evaluation of the proposed method for efficient region proposal. For both datasets, the satellite videos were collected by SkySat, which recorded 1800 frames with 30 frames per second. The spatial resolution of each frame in this video is 1.5 m and the frame size is 1920 \( \times \) 1080 pixels.

The SkySat-Las Vegas dataset refers to the satellite video captured over Las Vegas, USA in March 2014. As illustrated in Figure 4a, two sub-regions were selected for training and one sub-region was selected for evaluation.

The SkySat-Burj Khalifa dataset refers to the satellite video, which is captured over Burj Khalifa, United Arab Emirates on April, 2014. This video is 60 seconds long, which counts up to 30 frames per second. As shown in Figure 4b, 3 sub-regions were selected from the original video, two of which were for training and the remaining one for evaluation.

![Figure 4](image-url)

**Figure 4.** Two typical frames from the two satellite video datasets used. (The regions surrounded by the rectangle in yellow color are for training, while the regions in green color are for testing.)

For both datasets, vehicles on five frames from each datasets were annotated, and their corresponding boundary boxes were provided as labelled samples. As we can see in Table 1, the average target sizes are very small.
Table 1. Detailed information for the datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Region</th>
<th>Size</th>
<th>Average Vehicle Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkySat-Las Vegas</td>
<td>Train. 1</td>
<td>360 × 360</td>
<td>7.09 × 5.12</td>
</tr>
<tr>
<td></td>
<td>Train. 2</td>
<td>580 × 1070</td>
<td>6.27 × 5.03</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>720 × 700</td>
<td>7.54 × 6.00</td>
</tr>
<tr>
<td>SkySat-Burj Khalifa</td>
<td>Train. 1</td>
<td>300 × 400</td>
<td>6.52 × 5.11</td>
</tr>
<tr>
<td></td>
<td>Train. 2</td>
<td>450 × 650</td>
<td>7.07 × 5.28</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>500 × 670</td>
<td>6.97 × 5.80</td>
</tr>
</tbody>
</table>

3.2. Parameter Discussion

The LRP approach is mainly controlled by 3 parameters: the local region scale $k$, the threshold factor $f$ and the threshold $c_t$ in HistMM. The effect of each of them is discussed below. Their performance were evaluated in terms of the coverage of targets (recall), where a target is recalled if there is at least 50% of IoU between any proposals and the ground-truth bounding box. These evaluations were conducted by the Leave-One-Out Cross Validation (LOOCV) strategy on training set of the SkySat-Las Vegas dataset.

- **Semantic region Scale** $k$ controls size of the semantic regions generated. A larger $k$ is preferred as it will generate a coarse segmentation as required. The semantic regions are allowed to be larger than the target size on purpose. As presented in Figure 5, reducing $k$ gives fine-scale segmentation and leads to an increased number of region proposals with lower recall rate, while with increasing $k$, LRP generates fewer region proposals with improved recall rate.

- **Threshold Factor** $f$ controls the segmentation threshold in each semantic region. Selecting a large $f$ would result in fragmented region proposals and decrease recall scores. As illustrated in Figure 5, increasing $f$ from 1.0 to 3.5, the recall scores experience a drop of over 40%.

- **HistMM Threshold** $c_t$ is the Bayesian decision threshold in the HistMM for removing obvious false alarms as presented Section 2.3. The HistMM model with a smaller $c_t$ tends to keep more obvious false alarms, which leads to unnecessarily more region proposals decreases. On the other hand, increasing $c_t$ would filter out more obvious false alarms from the searched region proposals. As shown in Figure 6, when $c_t$ increases to 0.5, the number of region proposals ($N_{rp}$) reduces significantly, while the recall scores holds nearly stable about 80%, which presents the most efficient case.

When $c_t$ was set to 0.5 based on the cross validation on using the training data, the number of region proposals are reduced by over 60% by HistMM with almost no decrease in recall rate, las presented in Table 2 and Figure 7, which demonstrates the effectiveness of the proposed HistoMM model.

Table 2. Evaluation on the effectiveness of HistMM.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recall Before</th>
<th>Recall After</th>
<th>Diff</th>
<th>$N_{rp}$ Before</th>
<th>$N_{rp}$ After</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkySat-Las Vegas</td>
<td>75.92%</td>
<td>75.10%</td>
<td>−0.82%</td>
<td>30,614</td>
<td>10,100</td>
<td>−67.01%</td>
</tr>
<tr>
<td>SkySat-Burj Khalifa</td>
<td>77.31%</td>
<td>76.83%</td>
<td>−0.48%</td>
<td>17,017</td>
<td>6525</td>
<td>−61.66%</td>
</tr>
</tbody>
</table>
3.3. Comparison of Region Proposal Approaches

The region proposal performance was compared with a set of existing region proposals approaches for both common object detection tasks as well as aerial object detection tasks. Inspired by the systematic region proposal evaluation research [42], the proposed region proposal scheme was evaluated against Superpixels (SP) [39,42], Selective Search (SS) [26] and Region Proposal Network.
Saliency Object-based Region Proposals (SS) [37] and Superpixel (SP) [38] generate region proposals for each extracted superpixel. For both SS and SP, the extraordinarily tiny or large region proposals are considered impossible for vehicles in satellite videos and removed by post-processing. In addition to these well-known region proposals techniques, two approaches for aerial object detection are also included for comparison, which are Maximally Stable Extremal Regions (MSER) [33] or Top-hat-Otsu [34].

Qualitatively, the region proposals generated by our LRP are more concentrated on possible targets, while those saliency object-based approaches, SS and SP, produce more evenly distributed region proposals, as shown in Figure 8. A similar phenomenon is observed on the results by RPN, as both RPN and our LRP remove those obvious false alarms from the background.

Then quantitative performance evaluation on different approaches was conducted in terms of recall scores. Benefiting from the adopted searching strategy and the HistMM, LRP generates a reasonable number of region proposals with good coverage of the possible targets. As presented in Table 3 and Figure 9, our LRP achieves the highest recall scores on both evaluation datasets. In term of the number of the generated region proposals, it seems like our LRP generates more region proposals than SP, but it should be noted that more than one region proposals are generated by LRP for most possible targets, as shown in Figure 8. Although RPN generates more region proposals with better recall rates, it takes advantage of the finetune scheme from our Fast R-CNN model.

### Table 3. Evaluation on region proposal performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>SkySat-Las Vegas</th>
<th>SkySat-Burj Khalifa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N_{rp})</td>
<td>Recall</td>
</tr>
<tr>
<td>SP</td>
<td>4092</td>
<td>37.95%</td>
</tr>
<tr>
<td>SS</td>
<td>18,222</td>
<td>20.00%</td>
</tr>
<tr>
<td>MSER</td>
<td>15,347</td>
<td>37.73%</td>
</tr>
<tr>
<td>Top-hat-Ostu</td>
<td>1329</td>
<td>2.01%</td>
</tr>
<tr>
<td>RPN (Finetuned from Fast-RCNN-LRP)</td>
<td>13,288</td>
<td>90.00%</td>
</tr>
<tr>
<td>lLRP</td>
<td>9874</td>
<td>80.00%</td>
</tr>
</tbody>
</table>

Besides, we also compare the detection performance by using a slim Fast-RCNN detector. This slim Fast-RCNN receives 128 × 128 video frame as input, and it includes two groups of convolutional layers and a branch of fully connected layers for classification, where the branch for boundary box regression are replaced with carefully selected anchor distribution. Each group of convolutional layers contains three layers with kernel in the same size of 3 × 3, and the number of output channels is 16 and 32 for the first and second convolutional layer group, respectively. After each convolutional layer, a non-linear transformation is conducted by a Rectifier Linear Unit (ReLU) [43,44], which is followed by a Batch Normalization (BN) layer [45]. The output size by Roi Pooling is 2 × 2, which is followed by two fully connected layers with 512 and 32 hidden neural units, respectively. A Faster R-CNN model is also included for comparison. Due to the limited number of training samples, directly training a Faster R-CNN model is challenging, therefore, this Faster R-CNN model is finetuned from our Fast R-CNN-LRP. The performance evaluation is based on the PASCAL VOC metrics, where we use Average Precision (AP) instead of Mean Average Precision (mAP), since only one target category is contained in both datasets.

Compared with detection results by SP and SS approaches, our approach recalls most of the targets with the highest AP scores, as presented in Table 4 and Figure 10. Compared with the state-of-the-art Faster-RCNN model, the developed LRP with Fast-RCNN model achieves slightly improved detection performance. As illustrated in Figure 11, fewer false alarms with higher detection scores are produced by the Fast R-CNN model using the proposed LRP approach.
In addition to aforementioned single-frame-based detection approach, we also compare our approach with three popular background subtraction-based approaches — Gaussian Mixture Model (GMM) [46], GMMv2 [14] and Visual Background Extractor (ViBe) [16] approaches (A post-processing is applied to all these background subtraction-based approaches for removing extremely small or
large blobs). Their performance are compared in terms of recall, precision and $F_1$ scores at IoU = 0.5. Compared with these background subtraction-based approaches, Fast-RCNN-LRP that uses our region proposals generates better $F_1$ scores, and the background subtraction-based approaches suffer from poor precision, as shown in Table 5.

Table 4. Detection performance evaluation.

<table>
<thead>
<tr>
<th>Method</th>
<th>SkySat-Las Vegas</th>
<th>SkySat-Burj Khalifa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rcll</td>
<td>Prcn</td>
</tr>
<tr>
<td>Fast R-CNN-SP</td>
<td>34.32%</td>
<td>35.53%</td>
</tr>
<tr>
<td>Fast R-CNN-SS</td>
<td>14.32%</td>
<td>19.57%</td>
</tr>
<tr>
<td>Fast R-CNN-MSER</td>
<td>30.45%</td>
<td>31.16%</td>
</tr>
<tr>
<td>Fast R-CNN-Top-hat-Ostu</td>
<td>1.82%</td>
<td>8.08%</td>
</tr>
<tr>
<td>Fast R-CNN-LRP</td>
<td>58.18%</td>
<td>43.91%</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>59.32%</td>
<td>55.53%</td>
</tr>
</tbody>
</table>

Faster R-CNN (Finetuned from Fast R-CNN-LRP)

Figure 10. Visualization on detection results by selected approaches on SkySat-Burj Khalifa dataset.

Table 5. Detection results comparisons.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Rcll</th>
<th>Prcn</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkySat-Las Vegas</td>
<td>GMM</td>
<td>45.8%</td>
<td>49.6%</td>
<td>47.6%</td>
</tr>
<tr>
<td></td>
<td>GMMv2</td>
<td>64.7%</td>
<td>26.7%</td>
<td>37.8%</td>
</tr>
<tr>
<td></td>
<td>ViBe</td>
<td>58.0%</td>
<td>16.7%</td>
<td>25.9%</td>
</tr>
<tr>
<td></td>
<td>Fast-RCNN-LRP</td>
<td>58.18%</td>
<td>43.91%</td>
<td>50.05%</td>
</tr>
<tr>
<td>SkySat-Burj Khalifa</td>
<td>GMM</td>
<td>33.5%</td>
<td>56.7%</td>
<td>42.1%</td>
</tr>
<tr>
<td></td>
<td>GMMv2</td>
<td>70.1%</td>
<td>37.7%</td>
<td>49.0%</td>
</tr>
<tr>
<td></td>
<td>ViBe</td>
<td>74.6%</td>
<td>22.0%</td>
<td>34.0%</td>
</tr>
<tr>
<td></td>
<td>Fast-RCNN-LRP</td>
<td>64.09%</td>
<td>42.49%</td>
<td>51.10%</td>
</tr>
</tbody>
</table>
4. Discussion and Conclusions

Region proposal extraction is a valuable step to make target detection efficient. However, it is challenging to generate a small number of region proposals without missing any targets. This is more difficult when the targets are small and dim, such as those presented in satellite videos, due to their limited spatial resolution.

To address the degraded performance of current region proposal extraction methods for satellite videos, we proposed a novel region proposal approach (LRP), in which possible locations of targets are searched in semantic regions by coarse-scale segmentation and a Histogram Mixture Model (HistMM) is proposed to select region proposals with high likelihood from them.

The proposed LRP achieves improved recall rates of the targets with an acceptable increase in time cost, when compared with saliency object-based region proposal approaches, such as Superpixels (SP), Selective Search (SS), Maximally Stable Extremal Regions (MSER) and Top-hat-Otsu. Although the Region Proposal Network (RPN) recalls more targets with less time cost, it requires sufficient training samples or finetuning from a pre-trained model, such as the one obtained from LRP. Another advantage of the proposed LRP is that its training procedure only relies on positive training samples, even when a limited number of training samples is available.

With the improved recall rates by LRP, the detection performance by it with a slim Fast R-CNN is also superior to other saliency object-based region proposal approaches. The detection results are comparable with those by a finetuned Faster R-CNN model from our Fast R-CNN model. Compared with those background subtraction techniques, the proposal LRP approach outperforms them in term of precision, as fewer false alarms are generated.

As more satellite video data are available, more extensive testing can be conducted in the future study. In addition, the approach proposed in this manuscript is developed and tested on a panchromatic video data without color information. It may be extended to multi-channel data in the future research and improved detection performance can be expected.

**Author Contributions:** conceptualization, J.Z.; methodology, J.Z. and X.J.; software, J.Z.; validation, J.Z.; writing—original draft preparation, J.Z. and X.J.; visualization, X.J.; supervision, X.J. and J.H.; project administration, X.J. and J.H.; funding acquisition, X.J.

**Funding:** This research received no external funding.

**Acknowledgments:** This work is partially supported by China Scholarship Council. The authors would like to thank Planet Team for providing the data in this research [47].

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


