APPLICABILITY EVALUATION OF OBJECT DETECTION METHOD TO SATELLITE AND AERIAL IMAGERIES

K. Kamiya [a], T. Fuse [a], M. Takahashi [a]

[a] Dept. of Civil Engineering, The University of Tokyo, 7-3-1 Hongo Bunkyo Tokyo 113-8656, Japan
- (kamiya, takahashi)@trip.t.u-tokyo.ac.jp, fuse@civil.t.u-tokyo.ac.jp

Commission VII, WG VII/4

KEY WORDS: Satellite and Aerial Imageries, Object Detection Method, BING Method, Multi-Resolution Images

ABSTRACT:

Since satellite and aerial imageries are recently widely spread and frequently observed, combination of them are expected to complement spatial and temporal resolution each other. One of the prospective applications is traffic monitoring, where objects of interest, or vehicles, need to be recognized automatically. Techniques that employ *object detection* before *object recognition* can save a computational time and cost, and thus take a significant role. However, there is not enough knowledge whether object detection method can perform well on satellite and aerial imageries. In addition, it also has to be studied how characteristics of satellite and aerial imageries affect the object detection performance. This study employ binarized normed gradients (BING) method that runs significantly fast and is robust to rotation and noise. For our experiments, 11-bits BGR-IR satellite imageries from WorldView-3, and BGR-color aerial imageries are used respectively, and we create thousands of ground truth samples. We conducted several experiments to compare the performances with different images, to verify whether combination of different resolution images improved the performance, and to analyze the applicability of mixing satellite and aerial imageries. The results showed that infrared band had little effect on the detection rate, that 11-bit images performed less than 8-bit images and that the better spatial resolution brought the better performance. Another result might imply that mixing higher and lower resolution images for training dataset could help detection performance. Furthermore, we found that aerial images improved the detection performance on satellite images.

1. INTRODUCTION

Satellite and aerial imageries have been extensively used in various fields because of their capability to simultaneously observe wide area in one image. These days, the widespread of UAV, the recent appearance of miniaturized satellites and the deregulation of commercial satellite use are enhancing the environment in which a lot of high resolution images are widely spread and frequently observed. While aerial imageries have higher spatial resolution, they need to have a flight to be taken. On the other hand, satellite imageries can be periodically sampled without an additional expense although the resolution is inferior. Accordingly, by combining these two types of images, it is expected to complement spatial and temporal resolution. One of the prospective applications with combination of satellite and aerial images is monitoring, e.g. traffic monitoring, maritime security, disaster investigation and so on.

Monitoring is required to recognize objects of interest in their images. To deal with vast amounts of information, the automatic object recognition is a significantly important task. For the sake of the advanced monitoring, *general object recognition* is desired, which is a most difficult but important task that recognizes objects of any category in arbitrary images. General object recognition consists of *object detection* and *object classification*. Object detection outputs a set of proposal regions that likely contain an object of any category, and object classification recognizes what category the proposal region belongs to. Though, object recognition means object classification in limited sense. Normally, general object recognition employs object detection in pre-processing fashion. Doing object detection before object recognition, especially on satellite and aerial imageries, takes a significant role because it can save a computational time and cost: Otherwise, object recognition will have to be performed on every region in a large image despite a region will not contain any object.

A few studies have tried to apply object detection method to vehicle detection task from satellite images. Elmikaty (2014) eliminated detection windows that unlikely contain vehicles using object detection scores. Qu (2016) introduced object detection method as pre-processing before convolutional neural network architecture. However, applicability of object detection method to satellite and aerial imageries has not been mentioned. To the best of our knowledge, its performance evaluation is typically conducted with normal images, or snapshots: There is not enough knowledge whether object detection method can perform on satellite and aerial imageries as well as on normal images. Moreover, such images have a number of characteristics that should be considered; such as image band, bit depth, and multi-resolution. It also needs to be studied how they affect the object detection performance. Hence, the purpose of this study is to evaluate the applicability of object detection method to satellite and aerial imageries.

The remainder of this paper is organized as follows: In Section 2, a framework of object detection method to satellite and aerial images is introduced, and then we decide to use BING method in this study. An explanation of BING method is given in Section 3. Section 4 shows our datasets. We conduct several experiments and applicability evaluations in Section 5. Finally we conclude in Section 6.

2. FRAMEWORK OF OBJECT DETECTION METHOD TO SATELLITE AND AERIAL IMAGES

2.1 Object detection method

Under the assumption that all objects of interest share common visual properties that distinguish them from the background, object detection method, once given an image, outputs a set of proposal regions that are likely to contain objects (Hosang, 2015). *Object detection* method is often used before *object recognition* method. This is because that object detection method can efficiently propose object regions regardless of object categories. Thus, this property can lead to save computational cost in employing the more consuming object recognition method. Using benchmark workshops such as PASCAL VOC (Everingham, 2010), ImageNet (Deng, 2009) and MS COCO (Microsoft, 2015), a number of top performing object detection methods has been proposed; for example, SelectiveSearch (Uijlings, 2013), Objectness (Alexe, 2012), EdgeBoxes (Zitnick, 2014), BING (Cheng, 2014), CPMC (Carreira, 2010), R-CNN (Girshick, 2014) and so on.

2.2 Requirements for satellite and aerial imageries

In order to use these object detection methods on satellite and aerial imageries, we first list the characteristics of these images compared with normal images or *snapshots*. Satellite and aerial imageries may:

- have limited object shapes from the sky;
- know the camera parameters such as distortion, focus and exposure;
- target wider area for analysis;
- contain arbitrary object orientation;
- include more noise due to the atmosphere;
- be able to utilize information rather than visible waves;
- be able to use deeper images than 8-bit.

The first two characteristics are considered to bring advantages of object detection performance. However, to deal with wider area, it is required fast operation on object detection method. Also, to handle with the problems of arbitrary orientation and noise, object detection method has to be robust to them. Moreover, it has to have capability of the extensions to multi-band and deeper depth images.

According to a review and analysis on object detection method performances by Hosang (2015), BING method runs significantly fast and is robust to rotation and noise. Thus, this study employs BING method (Cheng, 2014) for applicability evaluation of object detection method to satellite and aerial imageries. Also, we modify the original BING algorithm to apply on multi-band and deeper depth images.

3. BINARIZED NORMED GRADIENTS (BING)

In this section, we explain BING method. An overview of BING is introduced followed by some extensions for applying BING to satellite and aerial images.

3.1 Overview of BING

BING is a fairly simple but powerful feature to use for object detection method. The main motivation of BING is that generic object with well-defined closed boundaries share strong correlation when looking at the norm of the gradient, after resizing their corresponding image regions to small fixed size (e.g. 8 × 8) (Qu 2016). Accordingly, BING method uses a simple 64D norm of the gradients (NG) feature, as well as its binary approximation, i.e. binarized normed gradients (BING) feature, for accelerated computing.

To find object regions, an image is scanned over with multiscale and multi aspect ratio windows. Each window is scored with a linear model $\mathbf{w} \in \mathbb{R}^{64}$,

$$s_l = \langle \mathbf{w}, \mathbf{g}_l \rangle \tag{1}$$

$$l = (i, x, y) \tag{2}$$

where s_l = filter score
$\mathbf{g}_l$ = NG feature
l = location
i = window size
(x, y) = position of a window

Using non-maximal suppression, a small set of proposals is selected from each size i. Then calibrated filter score is defined with a coefficient $v_i \in \mathbb{R}$ and a bias term $t_i \in \mathbb{R}$ as

$$o_l = v_i \cdot s_l + t_i \tag{3}$$

To learn a linear models $\mathbf{w}$ in (1), v_i and t_i in (3), ground truth object regions are used for positive sample, and linear SVM is employed.

In sake of speeding up the feature extraction and testing process, linear model $\mathbf{w}$ gets approximated with a set of basis vectors $\mathbf{a}_j$ as follows:

$$\mathbf{w} \approx \sum_{j=1}^{N_w} \beta_j \mathbf{a}_j \tag{4}$$

where N_w = the number of basis vectors
$\mathbf{a}_j \in \{-1, 1\}^{64}$ = a basis vector
$\beta_j \in \mathbb{R}$ = a corresponding coefficient

Plus, each $\mathbf{a}_j$ is redefined with a binary vector and its complement: $\mathbf{a}_j = \mathbf{a}_j^+ - \overline{\mathbf{a}_j^+}$, where $\mathbf{a}_j^+ \in \{0, 1\}^{64}$. This redefinition makes the score of a binarized feature $\mathbf{b}$ calculable using fast BITWISE AND and BIT COUNT operations:

$$\langle \mathbf{w}, \mathbf{b} \rangle \approx \sum_{j=1}^{N_w} \beta_j \left(2 \langle \mathbf{a}_j^+, \mathbf{b} \rangle - |\mathbf{b}| \right) \tag{5}$$

If we approximate the NG values of the corresponding image regions using the top N_g binary bits of the BYTE value, then a 64D NG feature $\mathbf{g}_l$ can be approximated by N_g binarized NG (i.e. BING) features as

$$\mathbf{g}_l \approx \sum_{k=1}^{N_g} 2^{8-k} \mathbf{b}_{k,l} \tag{6}$$

Then the filter score of an image window corresponding to BING feature $\mathbf{b}_{k,l}$ can be efficiently tested using:

$$s_l \approx \sum_{j=1}^{N_w} \beta_j \sum_{k=1}^{N_g} 2^{8-k} \left(2 \langle \mathbf{a}_j^+, \mathbf{b}_{k,l} \rangle - |\mathbf{b}_{k,l}| \right) \tag{7}$$

3.2 Extensions to satellite and aerial images

Although the original BING method considers 3 bands (BGR) and 8 bits images, this study has to consider applying BING method to satellite and aerial images, which have more than 3

color bands and 8 bits per pixel. Thus, we implement extended BING method. The implementation is fairly simple: We modify the original code that calculates gradient values to deal with 4 bands and 11 bit depths.

Another extension is motivated by Cheng (2014). In their study, they found that the undetected objects of BING method have something in common. Therefore, they improved BING method by making a new training set with undetected objects, training an assistant filter on this new training set. They reported that the assistant filter supplements BING method by detecting objects which BING method misses successfully. Accordingly, we introduce their idea to our verification experiments.

4. DATASETS

4.1 Satellite and aerial datasets

We use high resolution satellite imageries from WorldView-3, a commercial Earth observation satellite owned by DigitalGlobe. Aerial imageries in this study are provided by Asia Air Survey Co., Ltd. The details of satellite and aerial imageries are shown in Table 1 and 2 respectively.

Table 1. Details of satellite imageries

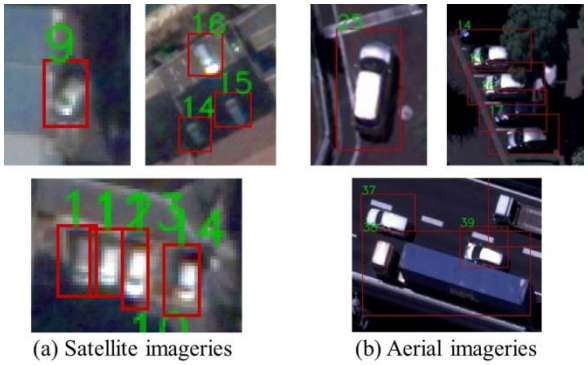
Region	Parts of Tokyo, Japan
Image date	14 Jan 2015, 13h25m30s-33s
Satellite	WorldView-3
Cloud coverage	0%
Solar altitude	30°
Off-nadir angle	7.04°
Type of image	Pan-sharpen image
Process	Ortho ready (OR2A)
GSD	0.31m
Bit depth	11-bit
Image bands	B: 450-510nm G:510-580nm R:630-690nm IR:770-895nm

Table 2. Details of aerial imageries

Region	A part of Tokyo, Japan
Image date	22 July 2015, 10h25m26s (local time)
Camera	Digital Mapping Camera (DMC)
Lens focus	120mm
Altitude	730m
GSD	0.073m

4.2 Training dataset

Using these satellite and aerial imageries, we create ground truth object regions for positive training samples and verification data. Following PASCAL VOC dataset (Everingham, 2010), we manually identify a location and a size of each *object* bounding box in every image. Here, we define an *object* as a *vehicle* in this study, because vehicles are main objects of interest in traffic monitoring with satellite and aerial images. We finally collect 3,009 vehicles from 49 satellite images and 1,166 vehicles from 35 aerial images. Figure 3 shows several examples of vehicle bounding boxes we made. Since the spatial resolution of aerial images (7.3cm) is much greater than that of satellite images (31cm), vehicle bounding boxes in aerial images is averagely 70 pixels in width while 20 pixels in satellite images.



(a) Satellite imageries (b) Aerial imageries

Figure 3. Examples of vehicle bounding boxes

5. EXPERIMENTAL RESULTS

In this section, we conduct several experiments to evaluate applicability of BING method to satellite and aerial imageries. First, experiments to compare the performances with both images are conducted. Secondly, verifications are done whether a training dataset with a combination of different resolution images improve the performance. Finally, we analyze the applicability of mixing satellite and aerial imageries for BING method as well as the effect of the assistant filter. The experimental details are shown in Table 4.

Our experimental code is based on the original C++ source code provided by Cheng (2014). Through our experiments, we keep some parameters that represent N_w and N_g fixed constant values. For having higher validity, we conduct 5-fold cross-validation for each experiment, and show an average result in following figures.

Table 4. Experimental details

Sec.	Analysis details		Train/Test
5.1.1	Type of image	Satellite, Aerial,	Sat/Sat Aer/Aer
5.1.2	Image band	4 bands(BGR+IR), 3 bands(BGR)	Sat/Sat
	Bit depth	11 bits, 8bits	Sat/Sat
5.1.3	Spatial Resolution	7cm, 15cm, 30cm	Aer/Aer
5.2.1	Combination of multi spatial resolution		Aer/Aer
5.2.2	Combination of satellite + aerial, Assistant filter		Mix/Sat

5.1 Comparison between different images

5.1.1 Type of image: First of all, we analyse whether the BING performance depends on type of image; satellite and aerial imageries. We train the BING model from two different types of image resources respectively, and test the BING method on each.

Figure 5 shows detection rate curves of each experiment. The vertical axis depicts what percentage of all *objects*, or vehicles, is detected. We follow the definition by VOC2007 that it is correct detection if the overlap of a ground truth object region and a proposal region is greater than 50%. The horizontal axis depicts the number of proposal regions in logarithmic scale. Since this detection rate curve monotonically increases, the higher graph shows the better performance.

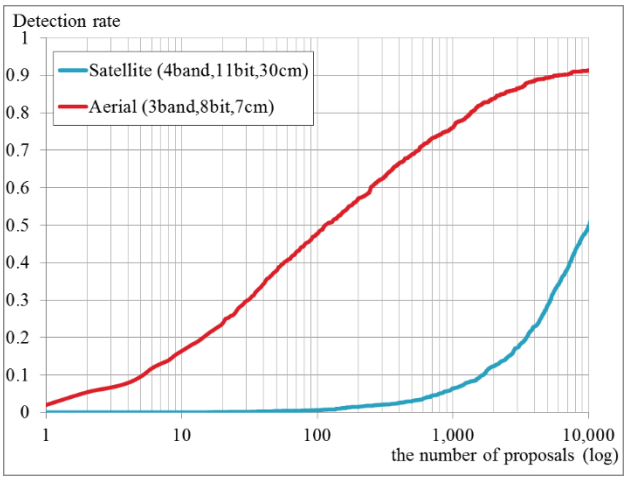


Figure 5. Comparison of detection rate between satellite and aerial images

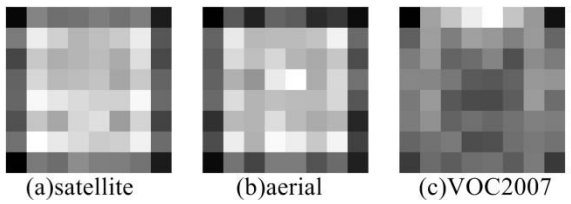


Figure 6. Learned linear model w

As we can see from Figure 5, BING performed well in order of aerial and satellite imageries. With a smaller number of proposal regions (e.g. 1,000), BING on aerial imageries detected about 75% of all vehicles, while only 5% on satellite imageries. With a larger number proposals (e.g.: 10,000), detection rate on aerial imageries converged to 90, although that on satellite imageries is still 50%. This shows there is an applicability of object detection method to aerial imageries to some extent, although that to satellite imageries is limited.

Figure 6 shows the linear model or NG filter **w** learned from each dataset. For purpose of comparison, we refer to the NG filter learned from VOC2007 dataset by Cheng (2014). Although there is little difference between models learned from satellite and aerial imageries, there is a great difference between VOC2007 and the others. Model learned from satellite and aerial images tend to be black (low value) in their edge and white (high value) in their center. This is probably due to the characteristic of the arbitrary orientation. In other words, in normal images, objects are almost fixed vertically, and are sometimes connected to the ground, which make the linear model asymmetry, top white and bottom black.

5.1.2 Image band and bit depth: In the next analysis, we compare differences of the BING performance between image bands and between bit depths respectively. We originally use the satellite images with 4 bands (Blue, Green, Red and Infrared) and 11-bit depth per pixel. To make 3-band satellite images, the infrared band is omitted using ArcGIS. Also, to make 8-bit satellite images, standard deviation stretching and gamma stretching implemented in ArcGIS are employed.

Figure 7 shows detection rate curves of four experimental patterns changing band size and bit depth. When comparing different band size (blue vs. navy in Figure 7), infrared band

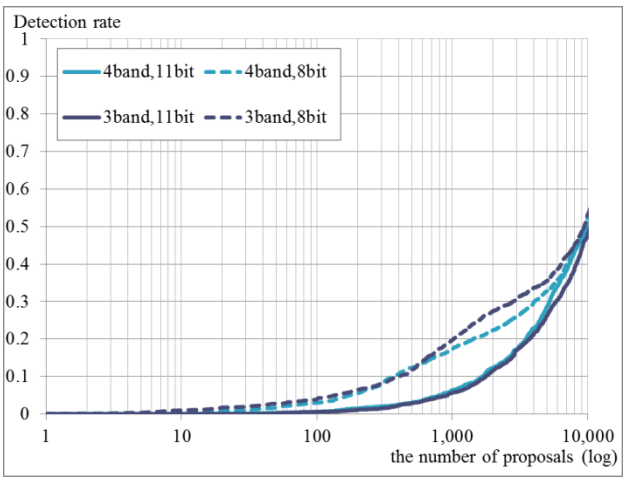


Figure 7. Comparison of detection rate between different image bands and bit depths

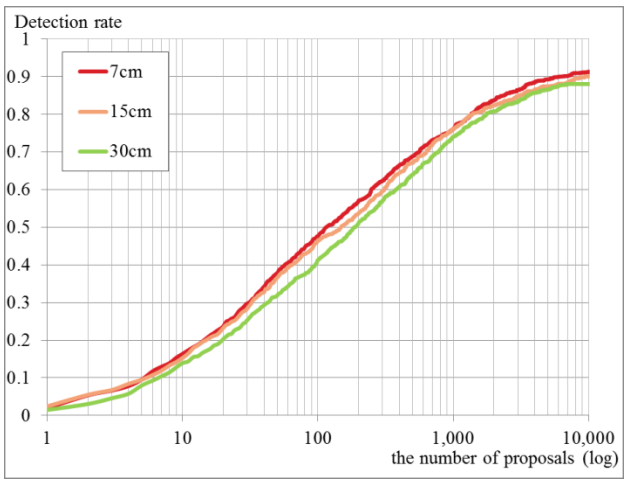


Figure 8. Comparison of detection rate between different spatial resolutions

had little effect on the detection rate of the BING method. This is probably because a red band histogram and an infrared band histogram are similar each other. Thus, the existence of an infrared band did not strongly affect processing gradients.

On the other hand, BING on 11-bit images performed less than 8-bit images (solid vs. dotted in Figure 7). This makes us slightly supersized because richer information in 11-bit images did not bring a better performance. It is considered that BING on 11-bit images is less robust to noises because gradients are calculated more sensitively.

5.1.3 Spatial resolution: Finally, we conduct sensibility analysis of BING performance according to spatial resolution. We originally use the aerial images with spatial resolution of about 7cm, and make modified images with resolution of 15cm and 30cm. A bicubic interpolation technique is employed in C++ development environment.

Figure 8 shows detection rate curves of 7cm, 15cm and 30cm spatial resolution images respectively. It tells that as spatial resolution is getting greater, the BING performance is getting better too, though the differences are not so serious. Because 8pixel × 8pixel filter corresponds to 2.4m × 2.4m region in 30cm resolution images, it is considered difficult for this

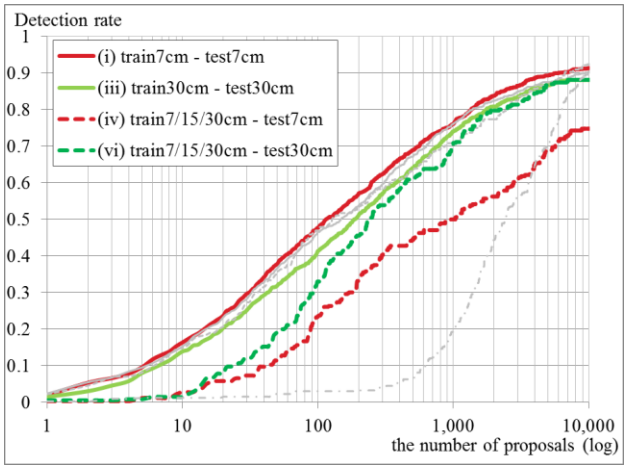


Figure 10. Comparison of detection rate between combinations of multiple spatial resolutions (i, iii, iv and vi)

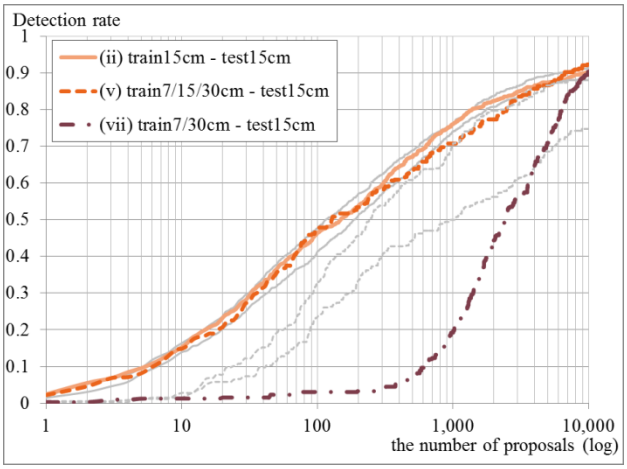


Figure 11. Comparison of detection rate between combinations of multiple spatial resolutions (ii, v and vii)

Table 9. Experiment patterns of multi spatial resolution

#	Resolution of train dataset	Resolution of test dataset
i	7 cm	7 cm
ii	15 cm	15 cm
iii	30 cm	30 cm
iv	7 cm , 15 cm and 30 cm	7 cm
v	7 cm , 15 cm and 30 cm	15 cm
vi	7 cm , 15 cm and 30 cm	30 cm
vii	7 cm and 30 cm	15 cm

small filter to cover all vehicles. Thus, in 30cm resolution images, even if uncountable proposal windows were produced, detection rate finally converged to less than 90%.

5.2 Comparison between combinations of multiple resources

When considering applying BING method to practical problems such as traffic monitoring in the future, it is thought that multiple resources are used to train BING model for conquering the insufficiency of training dataset. Next, we try to investigate whether multiple resources bring a good performance to BING method.

5.2.1 Multiple spatial resolutions: We suppose a situation where several aerial images with different spatial resolution are obtained. Thus, we test BING performance by using sets of different resolution images for training. In Table 9, experiment patterns we conducted are listed.

In Figure 10, experimental results of (i), (iii), (iv) and (vi) are shown. Through the comparison of (i) vs. (iv), and (iii) vs. (vi), it is shown the contamination of fairly different resolution images on training dataset leads poor detection performance. Specially, the performance shown in the result of (iv), where BING is trained with a set of 7, 15 30cm and tested on 7cm, is significantly dropped compared with that of (i). In Figure 11, the rest of experimental results, i.e. (ii), (v) and (vii) are shown. According to the result of (viii), although the lack of 15cm resolution images in its training dataset caused poorer performance with a small number of proposal windows, detection rate recorded 90% as well as (ii) and (v). These results may imply that mixing higher and lower resolution images for training dataset can help detection performance.

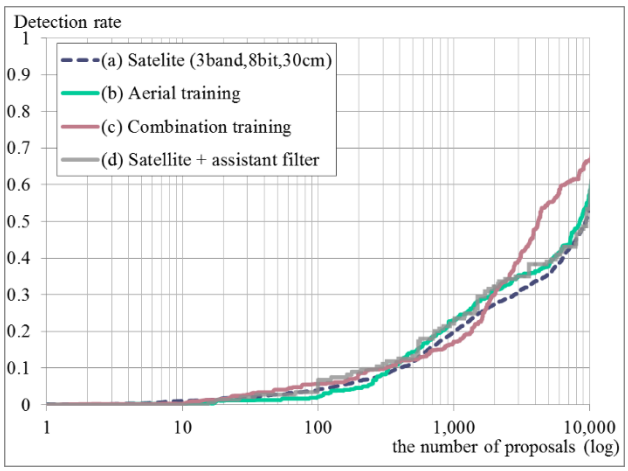


Figure 12. Comparison of detection rate between combinations of satellite and aerial imageries

5.2.2 Combination of satellite and aerial imageries: In the end of our experiments, we consider the combination of satellite and aerial images for training dataset. We train BING model (a) by the 30cm resolution satellite images, (b) by the 30cm resolution aerial images and (c) by a set of 30cm resolution satellite and aerial images. Then, each learned BING is tested on the 30cm resolution satellite images. Additionally, we introduce the idea of the assistant filter as experiment (d). We observe undetected vehicle regions in experiment (a), let them a new training dataset, learn an assistant filter and test on another test dataset.

Figure 12 shows experimental results of (a) to (d). Here, results of (b) and (c) performed better than (a). This may imply that BING learned from aerial images improve the detection performance even if it is tested on satellite images. Furthermore, combination with assistant filter brought a slight improvement of detection rate.

6. CONCLUSIONS

In this study, we first listed the characteristics of satellite and aerial imageries that should be considered in applying object detection method. Based on the result, this study employed binarized normed gradients (BING) method, which runs significantly fast and is robust to rotation and noise. Moreover, we extend the original BING method to deal with more than 3

color bands and 8 bit depth images. For our experiments, we created thousands of ground truth samples that represent regions of vehicles in images.

We conducted several experiments to compare the performances with different images, to verify whether combination of different resolution images improved the performance, and to analyze the applicability of mixing satellite and aerial imageries. The results showed that infrared band had little effect on the detection rate, that 11-bit images performed less than 8-bit images and that the better spatial resolution brought the better performance. Also, another result might imply that mixing higher and lower resolution images for training dataset could help detection performance. Furthermore, we found that aerial images in training dataset improved the detection performance on satellite images.

In our future works, it is desired to try object detection task than on vehicles. Since BING parameters are fixed to the default value through those experiments, the most suited values should be studied for satellite and aerial imageries. Our final goal is realization of general object recognition on satellite and aerial imageries. Thus, it is considered necessary to construct object classification architecture combining a state-of-the-art technique such as deep convolutional neural networks. Applicability evaluation throughout general object recognition; from object detection to object classification, is also required.

ACKNOWLEDGEMENTS

The authors would like to thank the provision of the aerial imageries provided by Asia Air Survey Co., Ltd.

REFERENCES

Alexe, B., Zurich, Zurich, ETH.Z., Deselaers, T. and Ferrari, V., 2012. Measuring the objectness of image windows. *IEEE PAMI*, 34(11), pp.2189-2202.

Carreira, J. and Sminchisescu, C., 2010. Constrained parametric min-cuts for automatic object segmentation. *IEEE CVPR*.

Cheng, M.M., Zhang, Z., Lin, W.Y. and Torr, P., 2014. BING: binarized normed gradients for objectness estimation at 300fps. *IEEE CVPR*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. *IEEE CVPR*.

Elmikaty, M. and Stathaki, T., 2014. Car detection in high-resolution urban scenes using multiple image descriptors. *IEEE ICPR*, pp. 4299-4304.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman A., 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), pp. 303-338.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE CVPR*, pp. 580-587.

Hosang, J., Benenson, R., Dollár, P. and Schiele, B., 2015. What makes for effective detection proposals? *IEEE PAMI*.

Microsoft, 2015. Microsoft Common Objects in Context. http://mscoco.org/home/

Qu, S., Wang, Y., Meng, G. amd Pan, C., 2016. Vehicle detection in satellite images by incorporating objectness and convolutional neural network. *Jounal of Industrial and Intelligent Information*, 4(2), pp.158-162.

Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeuldersm A.W.M., 2013. Selective search for object recognition. *International Journal of Computer Vision*, 104(2), pp. 154-171.

Zitnick, C.L. and Dollár, P., 2014. Edge Boxes: locating object proposals from edges. *Computer Vision ECCV*, 8693, pp. 391-405.