2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Leveraging Object Proposals for Object-Level Change Detection

Sugimoto Takuma Tanaka Kanji Yamaguchi Kousuke

Abstract— **Feature-based image differencing is an efficient approach to image change detection, which performs fast enough for self-driving car and robotic applications. Extant approaches typically take local keypoint features as input to the differencing stage. In this study, we aim to extend the differencing stage to consider object-level features. Our object-level approach is inspired by recent advances in two independent object-region proposal techniques: supervised object proposal (e.g., YOLO) and unsupervised object proposal (e.g., BING). A difficulty arises from the fact that even state-of-the-art object proposal techniques suffer from misdetections and false alarms. Our key concept is combining the supervised and unsupervised techniques into a common framework that evaluates the likelihood of change at the semantic object level. We address a challenging urban scenario using the publicly available Malaga dataset and experimentally verify that improved change detection performance can be obtained with our approach.**

I. Introduction

Image change detection is a fundamental problem for intelligent vehicles and other many important applications, including surveillance and city-model maintenance. In this study, we consider the problem of single-view change detection from a vehicle mounted front-faced camera. We do not assume the availability of a 3D model [1] nor any 3D reconstruction techniques, such as structure-from-motion or simultaneous-localization-and-mapping (SLAM) [2]. However, we focus on 2D-to-2D image comparison using an on-board monocular camera. This is a significantly challenging setting owing to a limited amount of perceptual information, large variety of object classes (e.g., cars, pedestrians, buildings, roads, vegetation, sky), intra-class variation, and inter-class confusion.

One of the most basic schemes for handling this problem is feature-based image differencing [3]. In this scheme, a given pair of query and reference images is first aligned to the same coordinate frame. Then, local keypoint features (e.g., scale invariant feature transform (SIFT) [4]) are matched between the aligned image pair. More formally, for each feature in the query image, an efficient nearest neighbor (NN) search over features in the reference image is performed. Then, its distance to the NN reference feature is interpreted as the likelihood of change (LOC) of that query feature.

This study aims to extend the differencing stage to consider object-level features. Compared to traditional local keypoint features, object-level features intuitively provide more rich semantic information that should be expected to be beneficial for change detection. Our object feature approach is inspired by recent advances in two independent object region proposal techniques: supervised object proposal (e.g., YOLO [5]) and unsupervised object proposal (e.g., BING [6]). The unsupervised proposal techniques provide category-independent object proposals, even for objects with unseen classes. The supervised proposal techniques provide more precise object regions supported by rich semantic information for objects with known classes. Both supervised and unsupervised proposals are beneficial as objects with both known and unknown classes can be change objects (Fig. 1).

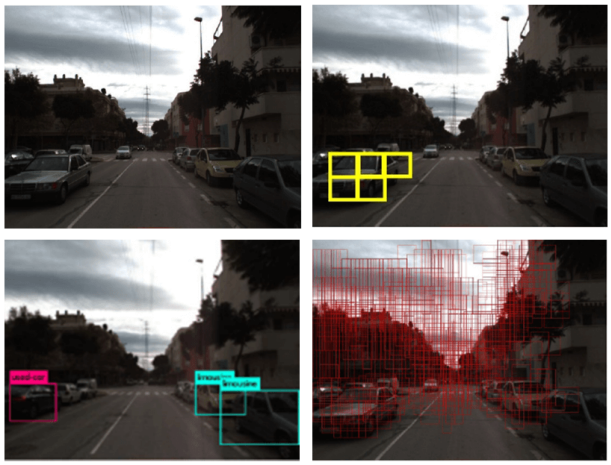

Fig. 1. Leveraging object region proposal techniques for object-level change detection. The change detection task takes as input a query image (top-left) and provides a ranked list of image grid cells in terms of LOC or the change mask (top-right). The supervised proposal techniques provide precise object regions supported by rich semantic information for objects with known classes (bottom-left). The unsupervised proposal techniques provide category-independent object proposals, even for objects with unseen classes (bottom-right).

A difficulty arises from the fact that even state-of-the-art object region proposal techniques suffer from misdetections and false alarms. The unsupervised approach often fails to capture the spatial context or semantic information. Thus, it suffers from false alarms. The supervised approach typically assumes a predefined set of object classes and works only on known class objects. Apparently, both unsupervised and the supervised approaches have advantages and disadvantages. Our objective is combining both approaches in a common framework to evaluate LOC at the semantic object level.

This paper makes several contributions. First, we extend the efficient scheme of feature-based change detection to consider object-level features. We also introduce an efficient

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 26330297, and (C) 17K00361.

The authors are with Graduate School of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

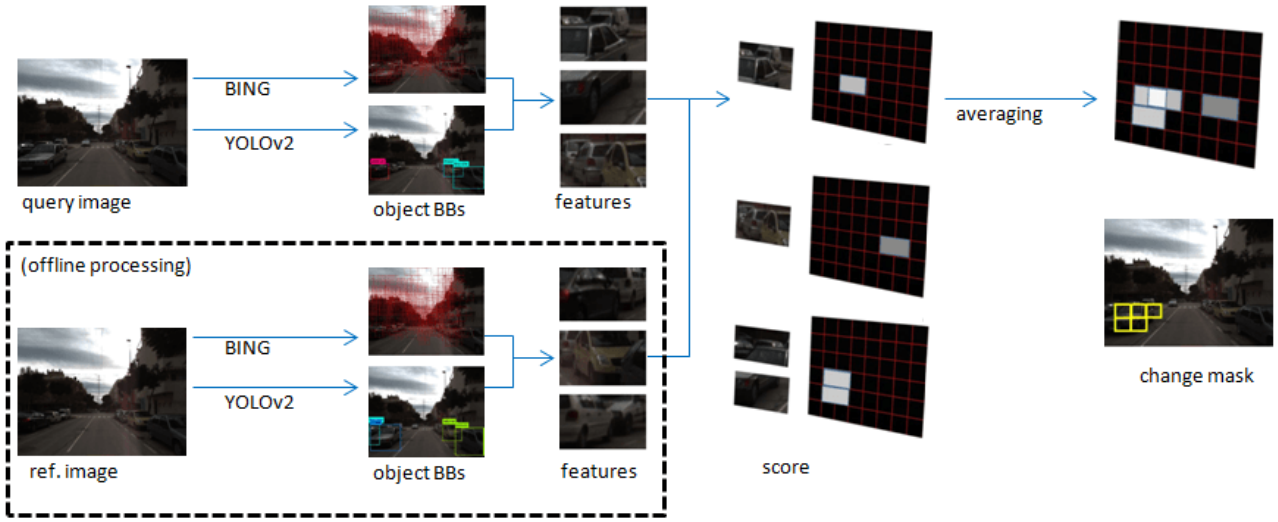

Fig. 2. Algorithm pipeline. In offline preprocessing, object region proposals are extracted from each reference image by both BING and YOLOv2 techniques. Then, object-level features are extracted for each object proposal by using LCFs and integral images. Online, object-level features are extracted from the given query image using the same procedure as the reference images, and feeds to an NN search over the corresponding reference image's features. Then, the NN distance values are interpreted as the LOC.

algorithm for object-level feature extraction that employs integral images that aggregate local keypoint features into object-level features with a small extra cost. Then, we implement several different combinations of both supervised and unsupervised object proposal techniques and investigate their effect on object-level change detection tasks. We also investigate to what extent the usage of the object proposals and features, both powered by deep convolutional neural network (DCN) techniques, can reduce misdetections by developing and comparing several change detection algorithms. We then address a challenging urban scenario from the publicly available Malaga dataset. We verify that improved change detection performance can be obtained by using the proposed approach.

II. Related Work

Change detection has been widely studied in various task scenarios. These scenarios are categorized into 2D-to-2D-matching [7], 3D-to-3D-matching [8], and 2D-to-3D-matching [1]. [7] performed 2D-to-2D-matching of 2D images from overhead imagery using an image registration and SIFT. [8] addressed 3D-to-3D-matching using point cloud data for novelty detection by patrol robots. [1] considered 2D-to-3D-matching between monocular image and cadastral 3D city models. Our approach uses 2D-to-2D matching, and our setting is far more complicated than the case of overhead imagery [5], because our vehicular applications require the capability to handle more general six-degrees-of-freedom freely-moving camera motions in a 3D space.

A pixel-wise comparison task, in which operations are carried out on a given pair of raw query and reference images, is the most common application. In [9], a scene registration method for image differencing was proposed, based on ground surface reconstruction, texture projection, image rendering, and registration refinement. In [9], a deep deconvolutional network for pixel-wise change detection was trained and used for differencing query and reference image patches. However, these required explicit memorization of every possible raw image and exhaustive many-to-many image comparisons, which severely limited its scalability in time and space.

Another line of research is the recent paradigm of efficient feature-based change detection [2]. In [2], SIFT features were employed in a change detection system that combined geometric, appearance, and semantic information. The work in [10] is considered one of the most relevant works to our study. In their work, change detection for overhead imagery was addressed. A bag-of-words model with tree-of-shape features was employed to achieve a more effective accuracy-efficiency trade-off. Based on these features, linear canonical correlation analysis was employed to learn a subspace to encode the notion of change between images. To reduce the cost of label acquisition by human photo-interpreters, a semi-supervised support vector machine framework was introduced. Our algorithm is inspired by these feature-based approaches and advances their models from the perspective of object-level features.

This work is a part of a study on scalable change detection and long-term vehicle navigation [11]. In [12], we focused on the image alignment stage in the change detection pipeline under large viewpoint uncertainty. In the area of field robotics, there is substantial work on change detection applications for patrolling [13], agriculture [14], tunnel inspection [15], and damage detection [10]. However, the above works did not focus on the use of object-level features from state-of-the-art object proposal techniques.

III. Approach

The proposed approach consists of a pipeline of four distinct steps (Fig. 2): (1) supervised object proposal; (2) unsupervised object proposal; (3) object feature extraction; and (4) LOC evaluation. These steps are detailed in the following subsections.

A. Supervised Object Proposal

Supervised object proposal techniques aim to detect bounding boxes (BBs) of possible object region proposals to classify each into one of several pre-learned object classes. We use a state-of-the-art object proposal technique called YOLOv2 [16], which is an extension of YOLO. Compared to faster region detection with convolutional neural networks (R-CNN) or YOLO, YOLOv2 improves localization errors and low recall. The network architecture is also modified by adding batch normalization to each convolution layer by introducing a new Darknet-19 structure and by adding a path-through layer to localize smaller objects. The approach also provides a method for dealing with larger numbers of unseen object classes, where no labeled training data are available, by exploiting ImageNet and WordNet to construct a hierarchical model of visual concepts. In experiments, YOLOv2 achieved higher mean average precision (mAP) performance than other state-of-the-art tools, including faster R-CNN, while still performing fast enough for self-driving car and robotic applications.

B. Unsupervised Object Proposal

Unsupervised object proposal techniques provide object region proposals from a given query/reference image without requiring prior knowledge of object appearance. We use the BING object proposal algorithm [6] to achieve this, because it is highly efficient with providing category-independent object proposals. The basic idea is to first train linear filters for each so-called quantized scale and aspect-ratio using simple binary gradient features. Then, it learns another global linear filter to rank BBs and output object proposals from the top ranked ones. For efficient computation, the quantization scheme maps every possible object scale and aspect ratio to predefined and fixed quantized ones, reducing the proposal searching space logarithmically, leading to very high computational efficiency. Moreover, BING adopts model binarization approximation to speed up feature extraction and testing. The learned linear model is approximated using a set of binary basis vectors. Each is further approximated using its top few binary bits. This enables very efficient cumulative computation, which is conceptually similar to the integral image.

The BING algorithm produces a large number of object proposals (e.g., 2.5×10^3 proposals per image). The proposals generally contain numerous false positives. Evaluating all proposals is computationally undesirable. Therefore, we select a marginal portion of object proposals. First, we evaluate area [pixel] of an object region and check if the area lies within the range of 5,000 to 20,000. Then, we eliminate near duplicate proposals by means of non-maximal suppression. With this process, a proposal is eliminated from the candidates when its intersection-over-union (IoU) with any other proposal having a higher LOC is larger than a pre-defined threshold of 0.5.

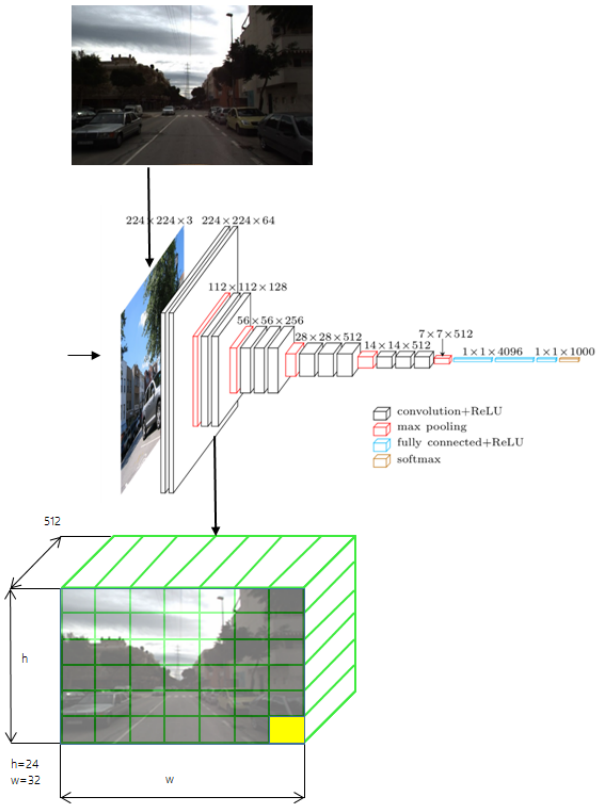

Fig. 3. Extracting local convolutional features. The feature extraction task takes as input a query/reference image and provides intermediate responses from convolutional layers of Vgg-16 as local convolutional features.

C. Object Feature Extraction

Object feature extraction aims to extract object-level features within a given BB, which is produced by supervised or unsupervised object proposal techniques. Our feature extraction approach is based on recent findings that reveal that the convolutional layer of a DCN can be viewed as a descriptor of patch-level image features (i.e., local convolutional features (LCFs) [4]). We observe that local features are more suited to changeable environments than global image descriptors (e.g., fully connected layer features), because local features are more robust against partial differences in appearance (i.e., changes) [17]. For the LCF feature, we extract a 32 $\times$ 24 array of 512-dim LCF vectors from the Conv5 layer of the Vgg-16 network [18] from either 1,024 $\times$ 768 image. Then, each feature is L2-normalized.

Our object-level feature is a summary of all the LCFs belonging to a given BB. LCF features are extracted on the regular $W \times H$ image grid, while a subset of LCF features located within a $w \times h$ BB is sized, wh ($w \leq W$, $h \leq H$). For efficient computation, we store all 512-dim LCF features on a $512 \times W \times H$ 3D grid, which can be viewed as an array of

512 2D grids of size $W \times H$. By applying the integral image technique to each 2D grid, we compute the sum or average of LCF vectors located within any BB. Our concept is to use a weighted average of LCF feature (See III-D) as the object-level feature of the BB.

D. Change Prediction

Change prediction aims to evaluate the LOC for each local image region. Given the above obtained object proposals and their respective object-level distance metrics, the NN distance between a query feature of interest and its NN reference feature is interpreted as the LOC value of the corresponding query object. However, with typical change detection applications, we are interested in pixel-level LOC rather than feature-level LOC [3]. To obtain pixel-level LOC values, we introduce a coarse $W' \times H'$ image grid of LOC values, where the width, W', and height, H', of the grid is set to $\lfloor W/32 \rfloor$ and $\lfloor H/32 \rfloor$ for $W \times H$ image. The LOC value of each grid cell is obtained by averaging the LOC values of all intersecting BBs. For averaging, we use a weighted average, considering the supervised and unsupervised methods have different levels of reliability. Then, we assign different weights for LOC values from supervised and unsupervised methods: $W_{supervised} = 1$ and $W_{unsupervised} = 1/20$.

IV. Experiments

We verified our method with a collection of urban images and compared the results obtained to the comparing methods. Our method, which combines the supervised and unsupervised object proposals, is denoted as "hybrid" (H). We adopted a simple "point-wise" method as the baseline approach. In it, individual LCFs are treated as independent point-wise features that are directly compared to query and reference images. This method can be implemented as a special case of the proposed method, by treating each cell in the image grid as each object BB. Additionally, we also compared classical keypoint feature methods, including huesift, opponentsift, and sift, with dense keypoint sampling. We also employed two alternative comparing methods: "supervised (S)" and "unsupervised (US)," which respectively employ the supervised or unsupervised object proposals alone.

Fig. 4 shows example outputs of the three methods. In the first example, both unsupervised and supervised method detected the false object in the left part. On the other hand, the proposed method, i.e., a combination of the supervised and unsupervised methods, successfully removed that parts. This is due to that the change mask of the proposed method is a weighted average of those of the supervised and unsupervised methods, and the weighted average acted as a kind of logical AND operation of the two change masks in this case.

In the experimental scenario, change detection from a vehicle mounted monocular camera is a challenge, owing to the large number of misdetections and false alarms produced by the supervised and unsupervised object proposal techniques. Moreover, even when object proposals and classifications are nearly perfect, the problem of discriminating changes

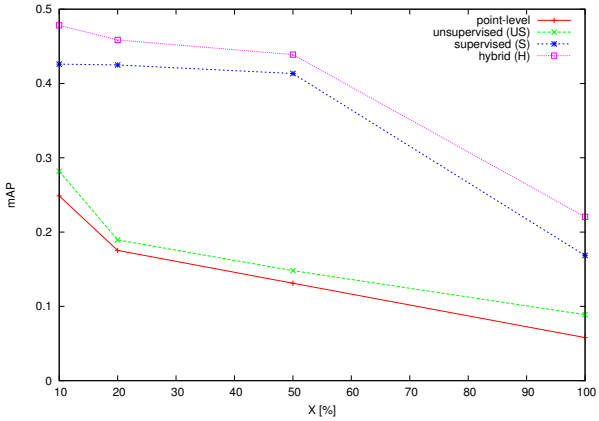

Fig. 6. mAP@X%recall performance of different approaches.

TABLE I
MAP PERFORMANCE [%].

Strategy	Method	mAP
Object feature -based	Proposed (H)	37.7
	Supervised (S)	34.5
	Unsupervised (US)	14.8
	Point-wise	12.4
Keypoint feature -based	huesift	3.6
	opponentsift	4.1
	sift	3.5

(e.g., stopping cars) from no-changes (e.g., parked cars), remains a difficult problem. Our application scenario has similar difficulty with alternative applications of parking lot-monitoring and car-counting [19].

Test sets were created from the Malaga dataset in the following procedure. First, images were paired by their ground-truth viewpoint locations, available in the dataset. To avoid making the change detection a trivial task, an image pair was accepted as a dataset element only if there exists change objects in the image and the vehicle's travel distance between the two viewpoints of the paired images was sufficiently large (e.g., $>$ 100 m). Note that such a viewpoint pair corresponds to the situation of "loop closing" in the literature of mobile robotics and SLAM, in which a vehicle traverses a loop-like trajectory and returns to a previously explored location. For each image pair, one with new object apperance was chosen as the query and the other was determined to be the reference image. Then, successive image pairs with near-duplicate query images were manually eliminated. The resulting collection of 325 image pairs were manually annotated with ground-truth BBs of change objects appearing in the query image, and was used as the test set.

Change detection performance is typically evaluated using precision and recall [3]. We adopted one of their variants, mAP, derived from the field of visual object detection [20]. In this model, a change detection algorithm outputs a prediction of the LOC for each cell on a $w \times h$ image grid of LOC values. For the evaluation, the LOC value of each cell in the image grid is thresholded into "change" or "no-change." Then, the binary change mask is compared to its ground-truth

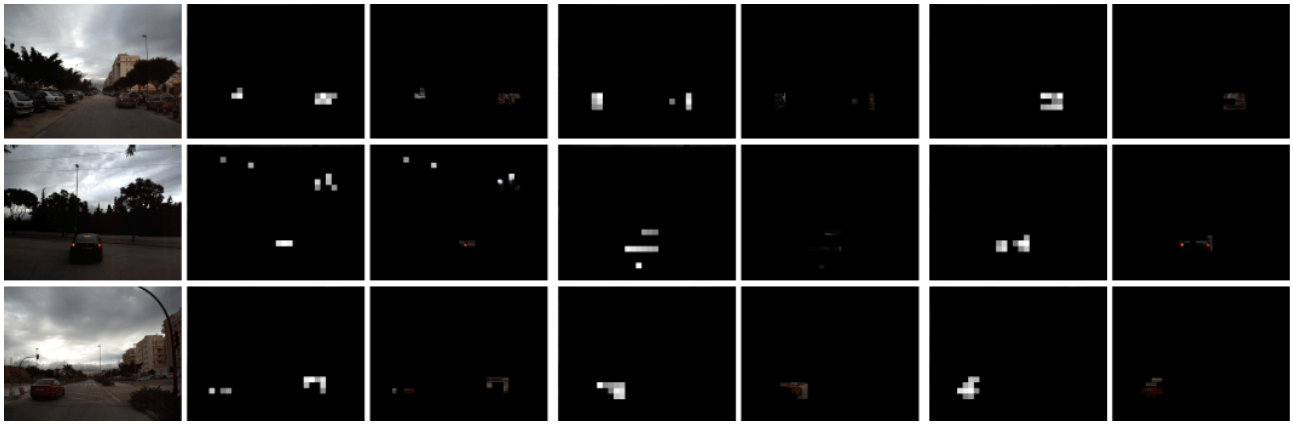

Fig. 4. Change detection results by object proposal techniques. From left-to-right, each panel corresponds to the input query image and change detection results using different change detection methods: "US," "S," and "H." Each panel shows change masks (left) and image patches (right) for cells having top 10 % LOC values.

Fig. 5. Example results.

change mask. Different threshold values provide different trade-off points between precision and recall. Following the literature [20], we approximated average precision (AP) using an average of "interpolated" precision values for 11 recall values: 0, 10, $\cdots$, 100 [%]. The mAP was obtained as the mean of the AP values over all the query images. The mAP can be viewed as a summary of these different trade-off points, approximating the value of the receiver operating characteristic. Considering the fact that achieving 100 % recall is not necessarily required in typical change detection applications, we also evaluate mAP@X%recall for different percentage points: X =10, 20, 50, and 100. Fig. 5 shows success and failure examples. We used the proposed approach of object-level change detection with the hybrid object proposal technique for the strategy. As shown in "success examples," the proposed approach was successful even for small-size change objects. However, failure often occurred from non-discriminative change objects, as shown in "failure examples."

Table I shows mAP performance for several change detection algorithms. It becomes apparent that the proposed object-level approach with the hybrid object proposal, powered by LCFs, outperforms the other LCF-based methods with supervised, unsupervised, and point-wise features and classical keypoint feature-based methods. There are several reasons for this. First, the method based on supervised object proposal alone suffers from misdetections. It often outputs sparse object proposals that do not contain any change object. Second, the method based on unsupervised object proposal suffers from false alarms. It usually outputs 5,000 to 20,000 object region proposals that result in very low precision, in terms of object detection. Third, even when object proposals are nearly perfect, it is still difficult to discriminate between change (e.g., stopping car) and no-change objects (e.g., parked car), especially when their visual appearance and locations are very similar. Fourth, the proposed method prioritizes supervised object proposals over unsupervised ones (i.e., $W_{supervised}/W_{unsupervised} = 20 \gg 1$). This can achieve high accuracy when supervised proposals contain change objects. Additionally, secondarily prioritized

unsupervised proposals can alleviate misdetections by using their spatially dense proposals as compensation.

It should be noted that we needed unsupervised object detector even when we have a better performing object detector, as the recall of supervised object detector was not good enough. Therefore, we needed object proposals from unsupervised method to increase the recall so that object detection covers enough region for change detection.

A lift from feature-based to object-based features has known in many fields and communities. A major concern of such object-based approach is how to handle an intra-class variation. It is not a trivial task to distinguish between a car and a different car parking at the same spot with different appearances. In the proposed LCF feature based approach, the issue of intra-class variation is addressed by explicitly describing appearance of individual objects. Thus, it enables to distinguish between same class objects with different appearances.

V. Conclusions

In this study, we extended the image-differencing stage of feature-based image change detection to consider object-level features. A difficulty arose from the fact that even the state-of-the-art object proposal techniques suffer from misdetections and false alarms. To address this issue, we implemented several different combinations of both supervised and unsupervised object proposal techniques, investigating their effects on object-level change detection. We addressed a challenging urban scenario using the publicly available Malaga dataset, experimentally verifying that improved change detection performance can be obtained by our proposed approach.

In this study, we assumed that the query and reference image pair are aligned to the same coordinate system in the pre-processing step. As mentioned in section I, the task of choosing the pair to ensure the alignment is a non trivial task. This alignment issue is focused in our recent paper in [12]. To integrate the alignment and comparison processes to realize a unified change detection framework is an immediate future work in our reasearch.

References

[1] A. Taneja, L. Ballan, and M. Pollefeys, "Geometric change detection in urban environments using images," *IEEE Trans. PAMI*, vol. 37, no. 11, pp. 2193–2206, 2015.

[2] J. Košecka, "Detecting changes in images of street scenes," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 590–601.

[3] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE transactions on image processing*, vol. 14, no. 3, pp. 294–307, 2005.

[4] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i Nieto, "Bags of local convolutional features for scalable instance search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16, 2016, pp. 327–331.

[5] van de Sande K. E. A., G. T., and S. C. G. M., "Empowering visual categorization with the gpu," *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 60–70, 2011.

[6] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3286–3293.

[7] W. Li, X. Li, Y. Wu, and Z. Hu, "A novel framework for urban change detection using vhr satellite images," in *Proc. ICPR*, 2006.

[8] P. Drews, P. Núñez, R. Rocha, M. Campos, and J. Dias, "Novelty detection and 3d shape retrieval using superquadrics and multi-scale sampling for autonomous mobile robots," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 3635–3640.

[9] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks." in *Robotics: Science and Systems*, 2016.

[10] L. Gueguen and R. Hamid, "Large-scale damage detection using satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1321–1328.

[11] K. Tanaka, Y. Kimuro, N. Okada, and E. Kondo, "Global localization with detection of changes in non-stationary environments," in *Proc. IEEE ICRA*, vol. 2, 2004, pp. 1487–1492.

[12] T. Murase, K. Tanaka, and A. Takayama, "Change detection with global viewpoint localization," in *Pattern Recognition (ACPR), 2017 4th IAPR Asian Conference on*. IEEE, 2017, pp. 31–36.

[13] H. Andreasson, M. Magnusson, and A. Lilienthal, "Has somethong changed here? autonomous difference detection for security patrol robots," in *Proc. IEEE/RSJ Int. Conf. IROS*, 2007, pp. 3429–3435.

[14] P. Ross, A. English, D. Ball, B. Upcroft, G. Wyeth, and P. Corke, "Novelty-based visual obstacle detection in agriculture," in *Proc. IEEE ICRA*, 2014, pp. 1699–1705.

[15] S. Stent, R. Gherardi, B. Stenger, K. Soga, and R. Cipolla, "An image-based system for change detection on tunnel linings." in *MVA*, 2013.

[16] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.