Training a robust UAV river patrol AI for different river and an analysis of the training dataset

Yuta Takahashi[*1]　　Junichiro Fujii [*1]　　Masazumi Amakata [*1]

[*1] Yachiyo Engineering Co., Ltd.

Japan has a lot of river. For the efficient river patrol, it is planned to use UAV & AI. The river patrol must detect the various target from the various back ground. This study focuses on the aerial image on the two river in urban or mountainous area. The target such as illegal dumping or people, bicycle, and cars can have the different distribution, and it is considered that the distribution of them and the variety of back grounds can affect the AI detection precision. The training uses the Faster R-CNN. The dataset is composed by the data on each river only or mixed with the ration changed. In this study, by the precision of the models are verified, the possibility of training models robustly for the different river and the dataset are verified.

1. Introduction

The global warming has a great possibility to makes water disasters intensified [UNESCO 2020]. In urban area, flood often occurs rather the huge economical loss than human damage [Pellicani 2018]. River embankment are also the infrastructure which has a part of the living space, and require regular patrols and maintenance. For example, in Japan, there are over 30,000 rivers, and a few great flood in every year. Area along with river has many objects (garbage, bench, small ships etc...), they can expand the damage of flood at overtopping. For keeping the state of embankment, the patrol stipulated by law are carried out however the efficiency is not high because of patrol by human. Since this patrols is also be done to check the breakage, there is dangerous for the confirmation works. If the UAV can patrol in this situation and the state of the embankment can be confirmed safely and quickly, the damage of flood can be reduced and the resilience after disaster become higher.

River space is open, it is suitable for UAV flight and also easy to apply image processing by AI technology. Recent year, although the UAV patrol under a water disaster is often disturbed by strong wind and heavy rain, some feasibility studies try the real time detection by UAV and analysis for embankment breakage, under an extreme good conditions [Brauneck 2016]. The progress is drastic, and the weather-proof UAV has begun to appear [PRODRONE 2020]. Promotion of UAV/AI application in river space will be a milestone for application to river infrastructure inspection such as bridges. This study focuses the detection of illegal dumping which is seem to be effective capturing image from UAV [Lega 2012] because it can have various contexts.

AI needs a learning with rich dataset. A lot of application case to civil engineering fields has be appeared, however, learning is often sensitive and difficult because of a less particular data, for instance, in the abnormal or damage states. As methodological improvements, there is unsupervised learning [Goodfellow 2014], however, the learning is often unstable. This results from the problem of which the boundary condition of data in civil engineering fields cannot be defined at ease. On the other hand, as learning dataset improvements, Data Augmentation is most popular, yet this method is not so almighty in less data. In previous study about self-attention which is SoTA technology in Natural Language Processing [Wang 2020], it is confirmed that the map in network can be recovered by the first few largest singular values. Thus, it is suggested that the selection of appropriate feature can improve the dataset quality.

At start of the UAV river patrol, dataset can be defined as what less aerial image taken by UAV and much ground image (taken by patrol staff and others). If ground image can be reused to learning AI for objects detection, dataset cover the features value which aerial image doesn’t have. Thus, this study tunes the extraction method of dataset which can contribute for improvement of learning. The best method is better to be able to embed the new data to pre-dataset with increase of dataset, however this study doesn’t limit the method and adopt the stochastic or geometrical ones, or both of them. Bounding Box Occupancy rate (BBO: our proposal criterion) are applied to the dataset which aerial and ground image are mixed. BBO is the rate of Bounding Box area in annotation and pixel image size [Takahashi 2021]. As object detection AI, Faster R-CNN [Ren 2015] is used.

This study focuses on the data taken in the two different river on urban or mountainous area. Urban area has a variety of back ground and target, and mountainous ones is not. In this study, the feasibility of training the robust model to each region is verified by mix of their dataset.

2. METHODOLOGY

2.1 Faster R-CNN

Faster R-CNN is an object detection deep learning model which is developed by Microsoft in 2015. At the ILSVRC in 2012, a team using deep learning left excellent results [Krizhevsky 2012], and the research has progressed rapidly as an image recognition technology. In 2015, some models exceeded human cognition of classification [Zhang 2016].

There are several networks for deep learning. Faster R-CNN uses CNN (Convolutional Neural Networks) for the extraction of feature maps. The output of full-connected MLP (Multi-Layer Perceptron) [Gardnera 1998] becomes one-dimensional simple vector, however by adding a convolutional layer, features which

Adress : Yuta Takahashi, Yachiyo Engineering Co., Ltd., CS Tower, 5-20-8, Asakusabashi, Taito-ku, Tokyo 111-8648, Japan, +81-3-5822-2903, yt-takahashi@yachiyo-eng.co.jp

maintain the input dimension are extracted enables more advanced learning. Faster R-CNN is performed in two stages: the object detection stage of specifying the object range by RPN (Region Proposal Network) and the object classification stage reuses same feature map which is used in RPN. The loss function of RPN is shown below from [Ren 2015].

$$L(\{p_i\},\{t_i\}) = {}^1/_{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + {}^\lambda/_{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (1)$$

Where, i is the index of Anchor point, p_i is the probability that Anchor point i is object. p_i^* is a compared label with Ground-Truth: when Anchor point i is object, p_i^* is 1, others is 0. t_i indicates the coordinates of predicted Bounding Box, and t_i^* indicates coordinates of Bounding Box in Ground-Truth. In [Ren 2015], N_{cls} is the mini batch size, and N_{reg} is the number of Anchor in feature map. λ is balanced parameter for the second term in right hands. In this paper, $\lambda = 10$ based on Reference. In the environment which can use GPU, N_{cls} often depends on the number of GPU. L_{cls}, the classification loss in whether object or not, is described by cross entropy. L_{reg}, the estimation loss of Bounding Box can be described with rectangle regression smooth_{L1} as below:

$$L_{reg}(t_i, t_i^*) = \mathrm{smooth}_{L1}(t_i - t_i^*), \quad (2)$$

$$\mathrm{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise.} \end{cases} \quad (3)$$

In the object detection stage, k Anchor Boxes with different aspect ratios are applied around each Anchor point of the CNN to classify whether or not they are objects. At this time, the IoU (Intersection over Union) of the Bounding Box in the Anchor Box and Ground-Truth images is calculated, and a threshold is set to distinguish the background from the object. In this experiment, IoU < 0.3 is the background and IoU > 0.6 is the object. Here, the second term on the right side of Eq. (1) is not considered if the object is not detected. That is, it is calculated only when IoU > 0.6. Through this algorism, L_{reg}, which is the deviation between each Anchor Box and the Bounding Box in the Ground-Truth image, is regressed. At the object classification stage, the RoI Pooling layer is applied to the feature map output by CNN and converted into a fixed-length feature vector. This is connected to two fully connected layers to obtain the classification of the presence or absence of an object and the output for rectangular regression. The model is learned by alternately updating the gradient of these two steps. As other classical detection models, You Only Look Once (YOLO) [Bochkovskiy 2020] or Single Shot multi-box Detector (SSD) [Liu 2015] are known. These models are composed only one step which is combined detection and classification, thus their inference speed is higher than Faster R-CNN. However, their accuracy of inference is inferior to Faster R-CNN. Additionally, because YOLO and SSD learns back ground of image by their own each architecture, it suggests that they don't suite for object detection in river patrol, for instance, illegal dumping which the back ground can change for every time to take image. Therefore, this study focuses Faster R-CNN.

3. Experiment

3.1 Training data taken by UAV on a different river

Two river is used as a test field. River A is in urban, and it has various back ground and target. Background has grass field, concrete or golf coat. River B is in mountainous area. Since there are often rocks or large wooded area, it has less artificial things. Added dummy dumping is similar on the rivers respectively.

Original images size are 3840 x 2160. The object is dummy dumping such as a box or plastic sheet or plastic bottle (red box in Figure.1). When images input to AI without cropping, the dumping size becomes small by resizing from 3840 x 2160 to 224 square and it may be vanished. In this study, learning image is cropped to 640 x 480 from original image also for data augmentation (orange dashed line box in Figure 1). This size is defined because 224 square cropping from original directly is too small to learn efficiently and 640 x 480.

In order to improve the efficiency of feature extraction, pre-trained ResNet50 by ImageNet is used, and the extraction layer is 40 ReLU layers. An RGB image input is resized to 224 x 224 and used for learning. The MATLAB 2021b environment is used for learning, the gradient optimization is SGDM (Stochastic Gradient Descent with Momentum) [Bottou 2010], the mini-batch size is 2, the learning rate is 0.0001, and epochs is fixed at 10.

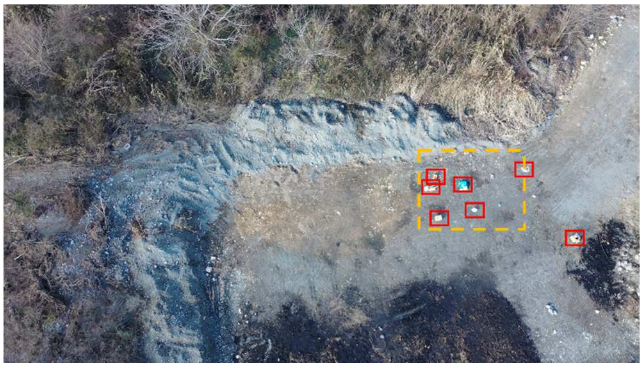


Figure .1 Aerial image (3840x2160 [pixel]) on River B

3.2 Making dataset

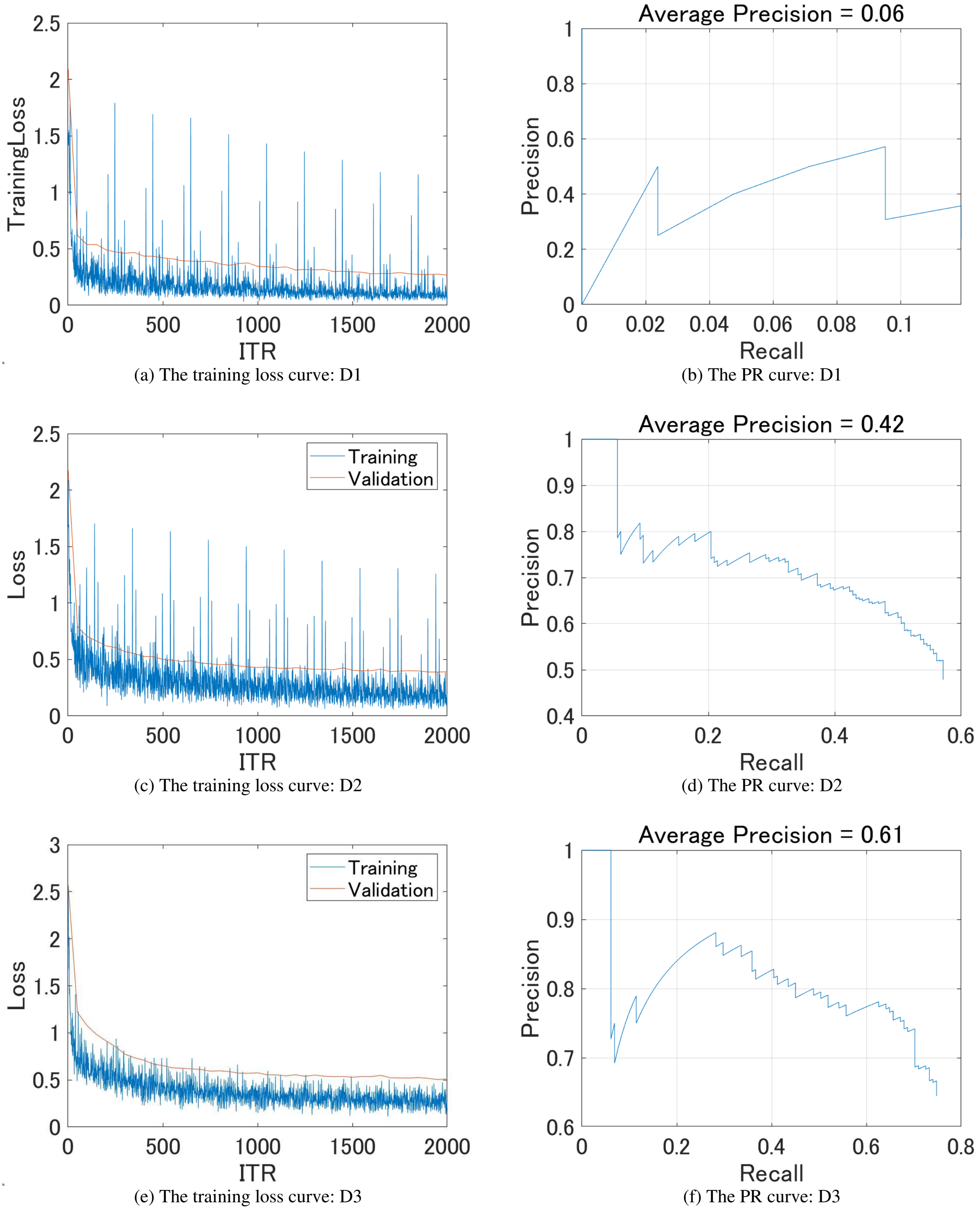
This study makes three dataset: 400 images on only River A, mixed 400 images(each 200 images on River A, B, respectively), 400 images on only River B. Validation and test data has 40 images from each river, and the maximum size become 80 images. Their dataset is called as D1,D2,D3. River A has a variety because of a lot of artificial things, and it is predicted that the model score becomes low. On the other hands, D3 has less them and less variety of background. Thus, it is expected that the score of D3 is top and D2 is middle of them. Less data in training can be assumed, however, this study try to train the robust model under less data such as start of UAV & AI patrol, thus this experiment is carried out by limit of amount data.

Table.1 The parameter in training

Learning rate	0.0001
Optimizer	SGDM
Epoch	10
Back bone	ResNet50
Extraction Layer	40_relu

Table.2 Dataset composetion

	Training		Validation		Test	
	River A	River B	River A	River B	River A	River B
D1	400	0	40	0	40	0
D2	200	200	40	40	40	40
D3	0	400	0	40	0	40



(a) The training loss curve: D1

(b) The PR curve: D1

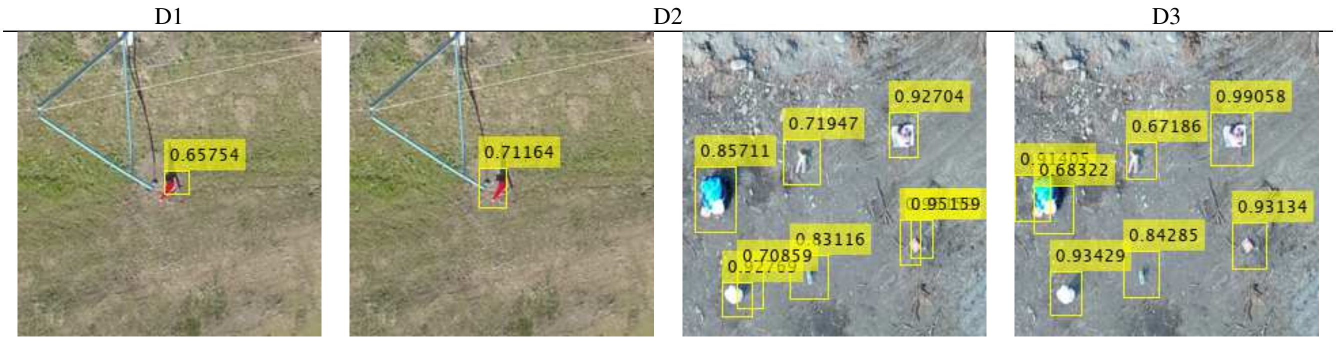
(c) The training loss curve: D2

(d) The PR curve: D2

(e) The training loss curve: D3

(f) The PR curve: D3

Figure.2 The training loss and PR curve of D1-3



(a) D1 (River A) (b) D2 (River A) (c) D2 (River B) (d) D3 (River B)

Figure.3 Sample result of D1-3

4. Result and discussion

4.1 Result

The loss curve of D1 (Figure.2(a)) suggests the convergence of training. The PR curve is in Figure.2(b). However, the AP is significant low. The sample result is Figure.3(a). There are a lot of artificial things and they disturbs the detection. The loss curve of D2 (Figure.2(c)) also suggests the training is converged, and the AP is not so high(Figure.2(d)). The sample result (Figure.3(b)(c)) shows AP is effected from the data of River B. However, the confidence in River A data grows from D1 result. The loss curve of D3 (Figure.2(e)) also suggests the convergence. The AP is top on them in Figure.2(f). The sample result (Figure.3(d)) shows the precision of D3 is higher than D2 result.

4.2 Discussion

The D1 result shows the large training data is necessary for AI on urban because of the variety of background and target. Thus, it is suggested that the addition of another river data can improve the score of AI for urban area. On the other hands, it is also suggested that AI on mountainous area which has often less variety of them can improve the model to robust with similar region data.

5. Conclusion and future works

This study verified the precision of Faster R-CNN trained by the aerial images taken from UAV on different rivers (urban or mountainous). The findings is as follows:

(1) The data on urban area can need a lot of data for training because of the variety of back ground and target. On mountainous area, the training can be carried out by less data.

(2) Addition of different river data to another can complement the training data on urban area. On the other hands, it is suggested that the model for mountainous area should be trained by similar region data.

This study uses the simple mix rate and random selection. Future works apply the pre-trained network such as ShuffleNet to dataset as prefilter for the selection of image data and it is verified whether their mix can improve the training.

Refference

[UNESCO 2020] World Water Assessment Programme, Disaster Risk Reduction, The United Nations world water development report 2020: water and climate change, Executive Summary, pp.3, 2020.

[Pellicani 2018] R. Pellicani, A. Parisi, G. Iemmolo and C. Apollonio, Economic Risk Evaluation in Urban Flooding and Instability-Prone Areas: The Case Study of San Giovanni Rotondo (Southern Italy), Geosciences, 8(4), 112, 2018.

[Brauneck 2016] J. Brauneck, R. Pohl and R. Juepner, Experiences of using UAVs for monitoring levee breaches, IOP Conf. Series: Earth and Environmental Science 46, 6th Digital Earth Summit, 2016.

[PRODRONE 2020] For example, PRODRONE: https://www.prodrone.com/

[Lega 2012] M. Lega, D. Ceglie, G. Persechino, C. Ferrara and Napoli R.M.A., Illegal dumping investigation: a new challenge for forensic environmental engineering, WIT Transactions on Ecology and The Environment, Vol 163, 2012.

[Goodfellow 2014] For example, GAN: I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, Generative adversarial nets, In Proceedings of NIPS, pages 2672– 2680, 2014.

[Wang 2020] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang and Hao Ma, Linformer: Self-attention with linear complexity., CoRR, abs/2006.04768, 2020. URL https://arxiv.org/abs/2006.04768.

[Takahashi 2021] Takahashi, Y., Fujii, J., Amakata M., Yamashita, T.: An application of AI technology to UAV's river patrol and the features value of datasets, SHMII-10, 2021.

[Ren 2015] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497, 2015.

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever, and G. Hinton, Imagenet classification with deep convolutional neural networks, In NIPS, 2012.

[Zhang 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, In Proc. of CVPR, 2016.

[Gardnera 1998] M.W. Gardnera and S.R. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmospheric Environment, Vol.32, Issues 14–15, pp. 2627-2636, 1998.

[Bochkovskiy 2020] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934, 2020.

[Liu 2015] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, SSD: Single shot multibox detector, arXiv preprint arXiv:1512.02325, 2015.

[Bottou 2010] L. Bottou, Large-Scale Machine Learning with Stochastic Gradient Descent, Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT’ 2010), pp. 177–187, 2010.