An application of AI technology to UAV's river patrol and the features value of datasets

Yuta Takahashi[1*], Junichiro Fujii[1], Masazumi Amakata[1], Takayoshi Yamashita[2]
[1] Artificial Intelligence Analysis Unit, Research Institute for Infrastructure Paradigm Shift, Yachiyo Engineering Co., Ltd.
[2]Department of Information Engineering, Chubu University
[*] Corresponding Author
email: yt-takahashi@yachiyo-eng.co.jp, jn-fujii@yachiyo-eng.co.jp, amakata@yachiyo-eng.co.jp, takayoshi@isc.chubu.ac.jp

ABSTRACT: River patrol has an aptitude to application of UAV and AI. Weather-proof UAV appear and the ordinary use case is increasing all over the world. On the other hand, training AI in civil engineering is still unstable because data often has a boundary condition which is defined difficultly. Large amount data can relax this constrain, however anomaly data is less and various. In this study, the applicability of UAV and AI to river patrol is verified using aerial image of dummy illegal dumping with data augmentation. Additionally, for stabilization and improvement of learning, whether the feature in aerial image is complemented by ground image which is selected based on several criterion or deep network is validated. The ground image selected by Bounding Box Occupancy rate and deep network (ShuffleNet and Inception v3) can improve the score, the validity of complement by data of different back ground was confirmed.

KEY WORDS: River Patrol, UAV, AI, Features of Data in Civil Engineering.

1 INTRODUCTION

The global warming has a great possibility to makes water disasters intensified [1]. In urban area, flood often occurs rather the huge economical loss than human damage [2]. 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 [3]. The progress is drastic, and the weather-proof UAV has begun to appear [4]. 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 [5] 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 [6], 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 [7], 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. t-SNE (T-distributed Stochastic Neighbor Embedding) [8] and k-NN (k-nearest neighbor) [9], k-means clustering [10], and 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. Similar criterion is used for improvement dataset [11]. Additionally, this study also verifies the effectiveness for selection of data by pre-trained network, such as ShuffleNet [12] or Inceptionv3 [13]. The ground image in RiMaDIS (River Management Data Intelligent System: data store in Ministry of Land, Infrastructure, Transport and Tourism [14]) and UAV aerial images in practice is used for verification. As object detection AI, Faster R-CNN [15] is used. The three contributions of this study are shown as below:

1. This study aims to apply image processing technology with UAV and AI to river patrol automation. The feasibility is validated by the

detection AI which learned the aerial image of dummy illegal dumping with data augmentation.

2. This study focuses on ground images, for complement of feature in less aerial images. The availability is verified by learning the supplemented dataset with on-ground image which selected based on several similarity criterions.

3. This study proposes that the data selection by above similarity criterions regards as prefilter for improvement of dataset. The difficulty of defining boundary condition for data in civil engineering cause often the instabilization for learning and over fitting on modeling. This study validate whether their criterion improve dataset and grow score by selecting better ground images. The comparison results suggest that Boundary Box Occupancy and pre-trained network such as ShuffleNet or Inceptionv3 are useful criterion or prefilter for improvement of datasets by complement of features.

This study tries proposing the methodology for improvement of dataset (especially for image data) in civil engineering based on the feature extraction. Finally, the improvement score by BBO and networks have been verified, and cover the feature value which is less in aerial images (e.g.: plastic bottle) has been confirmed.

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 [16], and the research has progressed rapidly as an image recognition technology. In 2015, some models exceeded human cognition of classification [17].

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) [18] becomes one-dimensional simple vector, however by adding a convolutional layer, features which 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 [15].

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

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 [15], 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 detection models, You Only Look Once (YOLO) [19] or Single Shot multi-box Detector (SSD) [20] 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.

2.2 Features for select of Dataset

In this study, in order to select the images used for training data, t-SNE (T-distributed Stochastic Neighbor Embedding: t-distributed stochastic neighbor embedding method) and k-nearest neighbor method, k-means method, BB occupancy (Bounding Box occupancy in the entire image: proposed in this paper), ShuffleNet and Inception v3, these five features and feature extraction method are used. The first two are the similarity for each pixel, and can be said to indicate the similarity of the background information of the image. The same applies to the third one, however it was used as an index to indicate the taking conditions for image: how much the target

object is captured in the image. Both of deep network has different characteristic respectively. ShuffleNet has less parameters than other models. Inception v3 has similar number of parameter with ResNet50 [17] which is used for feature extraction in Faster R-CNN in this study. However, input size of ResNet50 and ShuffleNet is 224 pixels square, on the other hands, Inception v3 is 299. In this study verifies whether the difference can affect the detection result.

a) t-SNE and k-NN

t-SNE is a nonlinear dimensionality reduction method. Find the distance $d(x_i, x_j)$ between the data at points x_i, x_j in the high-dimensional data X, and define the similarity $p_{j|i}$ as follows:

$$p_{j|i} = \frac{\exp\left(-\frac{d(x_i, x_j)^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{d(x_i, x_k)^2}{2\sigma_i^2}\right)} \qquad (4)$$
$$p_{i|i} = 0$$

When the above equation d is the Euclidean distance, if x_j is selected in proportion to the Gaussian distribution with respect to x_i, the similarity is expressed by a conditional probability. Next, the coupling probability p_{ij} is defined below by symmetry of the conditional probability. However, N indicates the number of data.

$$p_{ii} = \frac{p_{i|j} + p_{j|i}}{2N} \qquad (5)$$

Next, when the similarity is calculated for the points in the low-dimensional space data Y, and it is as follows:

$$q_{ij} = \frac{\left(1 + \left|\left|\mathrm{y_i} - \mathrm{y_j}\right|\right|^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1 + \left|\left|y_k - y_l\right|\right|^2\right)^{-1}} \qquad (6)$$
$$q_{ii} = 0$$

At this time, q_{ij} assumes a t distribution, which is the reason for naming t-SNE. The amount of Kullback-Leibler information between the joint distributions P and Q from Eq. (5) and Eq. (6) is expressed as follows:

$$KL(P||Q) = \sum_j \sum_{i=j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \qquad (7)$$

By minimizing this index, which indicates the difference between the two probability distributions, t-SNE reduces the dimension.

In this study, k-NN applies to the two-dimensional data of the image obtained by t-SNE, and the ground image around the aerial image are selected for learning. Note that t-SNE minimizes the amount of Kullback-Leibler divergence, thus it is considered that images with stochastically close feature are selected for the entire image. Additionally, note that t-SNE enables high accurate non-linear dimension reduction, however it requires the entire calculation again to classify the data obtained after. This cycle is not suitable for practical schema, however this is adopt at first in this study because the method is based on stochastic theory clearly.

b) k-means method

The k-means clustering method is a non-hierarchical clustering method. For each data $x_i (i = 1, \ldots, n)$, clusters are randomly assigned and the center of each cluster $(V_j j = 1, \ldots, k)$ is calculated. The operation of finding the distance between each x_i and V_j, and reallocating x_i to the nearest cluster is repeated until the threshold is satisfied (Eq. (8)).

$$\arg\min_{V_1, \ldots V_k} \sum_{i=1}^{n} \min_j \left|\left|x_i - V_j\right|\right|^2 \qquad (8)$$

As this algorism shows, k-means method depends on the initial value. Thus, the improved method k-means ++ [10] is used in this study. As the same as t-SNE, the distance between each data is used, however the dimensional reduction isn't involved. Therefore it is considered that a ground image similar to the aerial image can be selected retaining the feature.

c) BB occupancy rate

BB occupancy is proposed in this study. Since object detection involves separation of background, the smaller the area of the background, the less likely the tensor of features will be sparse. This idea is used for improvement of dataset in a previous study [11]. Here, when the Bounding Box area is about the same as the input image size, it is considered that the same amount of features can be extracted. In this study, the area of the pixel ratio of Bounding Box for image size were calculated, and ground images were selected due to match to the mean of aerial images.

d) ShuffleNet

ShuffleNet is developed for fast inference such as edge computing. This model has 1.4 million parameters. It is much smaller than ResNet50 ones which is 25.6 million. ShuffleNet enables high accurate feature extraction with 100 MFLOPS order due to Group Convolution and Channel Shuffle. Group Convolution carries out dividing the feature map to kernel channel and calculates convolution only in divided group. This process reduces FLOPS to 1/G: G is number of groups. Channel Shuffle is used for convolution between channels along with Group Convolution in ShuffleNet. This shuffle enables Group Convolution to include the feature of between Groups.

Layer of Group Convolution and Channel Shuffle is different with general other model ones, for example, VGG [21], ResNet ones. In this study, this difference can be expected as complement of features in ResNet50.

e) Inception v3

Inception v3 is developed from GoogLeNet which has Inception module. This module is micro network which is composed with multiple convolution and pooling layer, and it was inspired from Network in Network [22] which is composed convolution layer and MLP. In addition, Inception v3 has also batch normalization layer for balance of between calculation batches. The principle of effect by batch normalization has been clear yet, it suggests that this process can avoid the complexity of increasing weight in deep learning [23]. In civil engineering context, these process can be expected as avoidance method of overfitting to abnormal or singular data which are very few in practice.

3 EXPERIMENT

3.1 Feasibility confirmation

The feasibility of applying UAV and AI technology to river patrol is verified by AI detecting dummy illegal dumping in images taken from UAV. Original images size are 3840 x 2160 and the number of them is 1209. The object is box or plastic sheet or plastic bottle (red box in Figure 1(a)). 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(a)). This size is defined because 224 square cropping from original directly is too small to learn efficiently and 640 x 480 are popular size in RiMaDIS (Figure 1(b)).

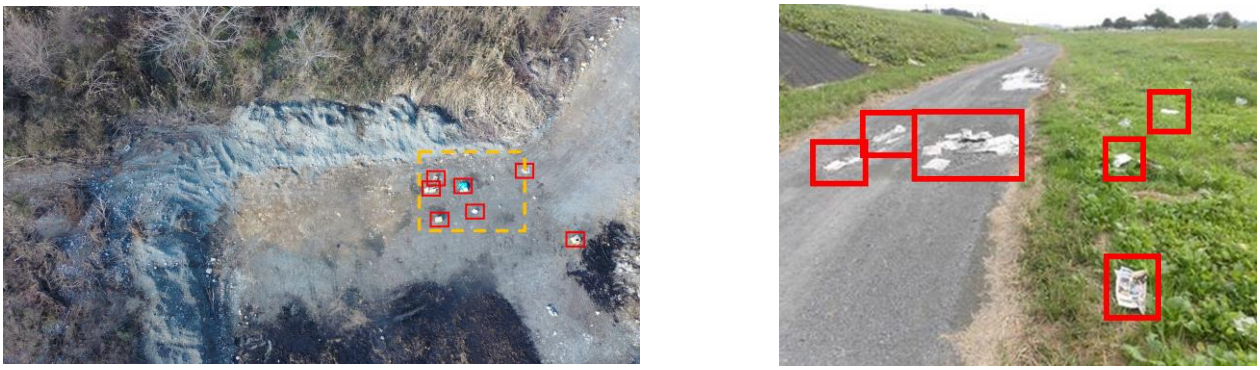
Cropped image are augmented to 17016 with shift. Training dataset is 10209, and test dataset is 6807. In order to improve the efficiency of feature extraction, pre-trained ResNet50 by ImageNet is used, and the extraction layer is 40 ReLU layers. An RGB image input is resized to 224 x 224 and used for learning. The MATLAB 2020a environment is used for learning, the gradient optimization is SGDM (Stochastic Gradient Descent with Momentum) [24], the mini-batch size is 2, the learning rate is 0.001, and epochs is fixed at 10.

The inference score and inference result sample are shown in Figure 2(a), (b) respectively. Training time is 77 hour with two TesraK80. In Figure 2(a), the score of average precision (AP) which is an area under the PR (Precision-Recall) curve is 0.89. In Figure 2(b), almost object can be detected. The yellow frame indicates the inferred Bounding Box, and the numbers indicate the confidence. This is high enough to be used in practice because an object can be captured in several images with continues through flight. This model is called FC model or later.

3.2 Benchmark experiment for feature complement

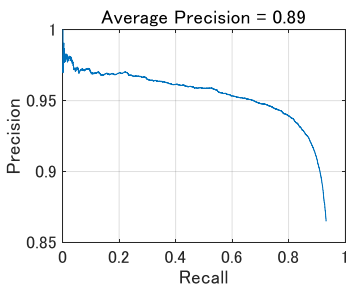
The score in FC model is high, however, this may result from large data augmentation which can cause overfitting to input, and thus the applicability to real various object which can be assumed as illegal dumping and there are not in dataset is not proven. In order to confirm whether the detection accuracy is improved by adding ground images, the benchmark is the case where only aerial images are used. This data is same as cropped ones used in FC model. 280 aerial image training data and 20 inference data are used. In this study, maximum of 800 ground images can be used. The aerial image is taken from the sky above the illegal dumping. Therefore, while the aerial image captures illegal dumping from almost directly above, the ground image is from the horizontal direction. The difference in this condition is expected to have an effect on how shadows are evaluated as features. The hyper parameter is same as FC model except the learning rate is 0.0001 because the amount of dataset becomes small than FC model one. These settings will not be changed for subsequent experiments.

Figure 3(a) shows the PR (Precision-Recall) curve and Figure 3(b) shows an example of the inference results for the inference results of the benchmark experiment. About half of the illegal dumping on the left side is caught in the Bounding Box, however it can be seen that the accuracy and reliability are not high. Plastic bottles (near the center) and magazines (upper right) have not been detected well. Since the blue object in the upper left is cut off, it is not set as illegal dumping in this training image, thus there is no problem even if it cannot be detected this time. On the other hand, since the center of Bounding Box is in the shadow of object, it can be supposed that the shadow is paying more attention. One of the causes is that the background is gray and bright, which is considered to



(a) Aerial image (3840x2160 [pixel]). (b) Ground image (640x480 [pixel]).

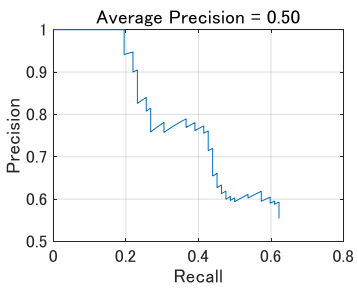
Figure 1. Aerial and Ground image for learning.



(a) The PR curve.

(b) Inference sample.

Figure 2. Feasibility Confirmation.



(a) The PR curve.

(b) Inference sample.

Figure 3. Benchmark.

have increased the attention to shadows. These results suggest that sufficient learning has not been carried out and the background and illegal dumping cannot be discriminated. Compared with this model, the ground image is added to the learning image and it is verified whether the accuracy is improved.

3.3 t-SNE and k-NN

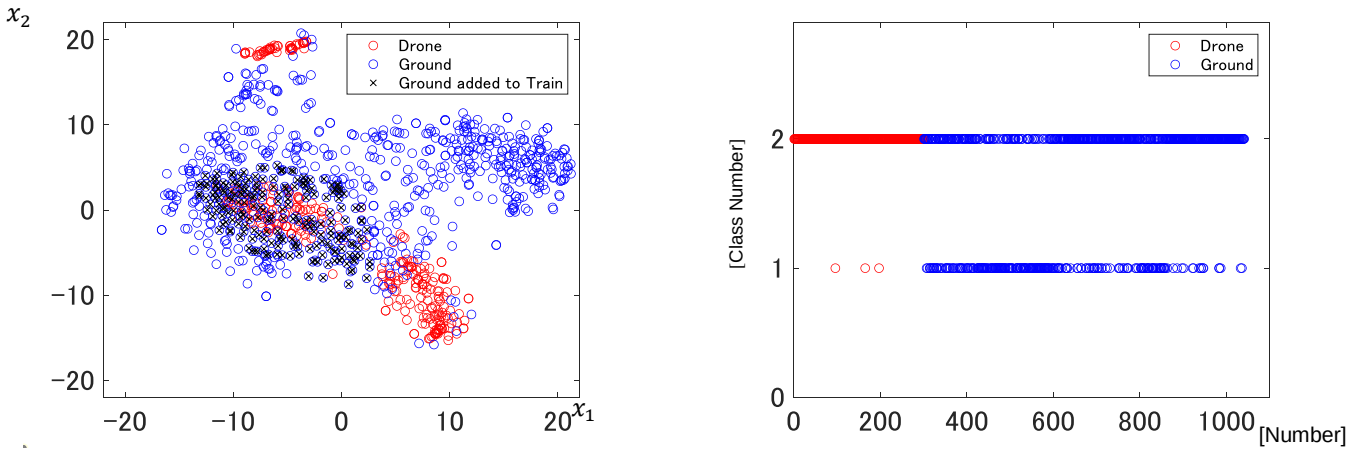
Figure 4(a) shows the two-dimensional data of each image data using t-SNE. The horizontal axis represents the first component x_1 and the vertical axis is the second component x_2. The red circle shows aerial images (legend: Drone), and the blue circle shows ground images (legend: Ground). In this study, the group of red circle on the left in the center (103 data) which has more blue circles around them are used for learning as aerial image. The k-NN selects 215 ground images around above group, and they are added to learning (black x: Ground added to Train in Figure. 4(a)). The ratio of the number of data is summarized in Table 1 in the discussion chapter. The PR curves of the inference results, and detection sample are shown in Figure. 5(a) and Figure. 5(b) respectively. Compared with PR curve of the benchmark, it is confirmed that score doesn't grow at all. In Figure. 5(b), the object, which was captured in case of the benchmark at least within about half of their area, was hardly detected. Focus on the background, it seems that the ground with light and dark like footprints is detected as object. From the results, it can be supposed that the feature amount of the aerial image is disturbed by ground image ones, and the detection accuracy becomes low.

3.4 k-means method

The k-means method classifies aerial and ground images to two class (Figure. 4(b)). The vertical axis expresses the class number and the horizontal axis is the data number. 277 aerial images (red circles) and 482 ground images (blue circles) classified into the same class (class 2) are used for learning. Compared with the t-SNE + k-NN experiment, the total training data has increased, and especially the ground image has more than doubled. The PR curve and an example of the inference result are shown in Figure. 6(a) and Figure. 6(b). The PR curve and mean Precision are better than the results of the t-SNE + k-NN, however not enough. At the example of the inference results, a plastic bottle near the center, which could not be detected even by the benchmark, was detected. It is assumed that this is because there was a lot of plastic bottles on the ground image, and it became possible to detect plastic bottles that were difficult to capture in aerial images. The idea about complement of the feature by the other data set isn't so rare (e.g. transfer learning), thus it suggest that this result has a validity.

3.5 BB occupancy rate

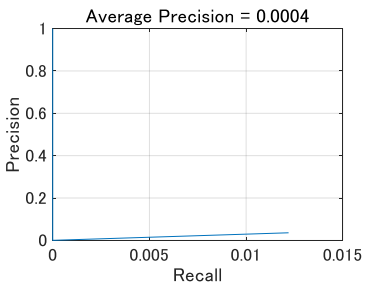
The BB occupancy rate (BBO) is the total value of the Bounding Box area of the image divided by the image size. In this study, the overlap of Bounding Box is allowed. Figure. 7 shows a histogram of the BBO in aerial images. The horizontal axis is the BBO, and the vertical one is frequency. The average was 8.88 [%]. Ground images (95 images) with a BBO which is range ±25% of the average value is selected for learning.



(a) Scatter plot of t-SNE and k-NN result

(b) Scatter plot of k-means++ result

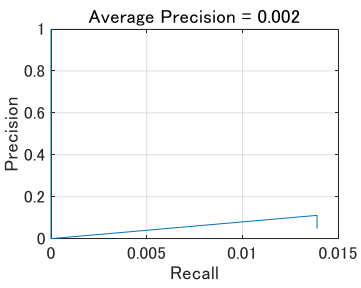
Figure 4. Classification result by t-SNE and k-NN, and k-means++



(a) The PR curve.

(b) Inference sample.

Figure 5. t-SNE and k-NN.



(a) The PR curve.

(b) Inference sample.

Figure 6. k-means++.

The PR curve is shown in Figure. 8(a). The AP is improved from the benchmark (0.50). An example of the inference result is shown in Figure. 8(b). It has higher confidence than before, and boxes are not duplicated or overlooked. In addition, as same as the case of k-means method, plastic bottles which could not be detected by the benchmark are detected with high confidence. Therefore, by increasing the number of ground images, complement of the features which is less in the aerial image data set and improvement of score can be confirmed.

3.6 ShuffleNet

For applying deep network to improvement of dataset as prefilter, it must be learned, for example, such as two-class classification. However, in the case of only aerial image are unsupervised learning, and can become unstable. On the other hand, classification learning of aerial image and some ground image can be affected from input ground image feature. This study uses back ground images which didn't capture an object for learning, as second class. Back ground image is not only easy to obtain, but also can be learned directly as adversarial data to aerial image which captures an object. Thus, it suggests directly that classified ground images as back ground is unsuitable for learning.

Training data uses 280 aerial image as same for detection model. Back ground image with 640 x 480 pixel size (309 images) is obtained from remains in cropping. The first class is aerial, the second one is back ground. Network is pre-trained by ImageNet. Validation rate is 0.3, learning rate is 0.0001, epoch is 8, mini batch size (MBS) is 128, final validation loss and accuracy (FVL and FVA) is 0.15, 0.93. These parameter is shown Table 3. On the inference, classified ground image as first class is 161 with threshold 0.85. The PR curve is shown in Figure 9(a). The AP becomes higher than BBO result. An example of the inference result is shown in Figure 9(b). The all confidence becomes higher and left object is detected at center in box. This result shows the validity of deep network as prefilter for improvement of dataset.

3.7 Inception v3

As same as ShuffleNet, Inception v3 is trained with back ground image. The changed parameter in Inception v3 is epoch and MBS, and they are 9 and 32 respectively. They are defined so that loss can become similar to ShuffleNet, and to make the condition match as possible and to avoid memory leak. FVL and FVA is 0.14, 0.98. These parameter is shown Table 3. On the inference, classified ground image to first class is 109 with threshold 0.85. The PR curve is shown in Figure 10(a) and the AP becomes higher than BBO result, however, it is inferior to ShuffleNet ones. An example of the inference result is shown in Figure 10(b). Detection is duplicate on right object. Others has highest confidence. From this result, the improvement is confirmed, however, it is suggested that merely to use deep network as prefilter without strategy has limit to improve.

4 DISCUSSION

Consider the six experiments carried out, including the benchmark. Tables 1 and 2 show the data distribution used in each experiment and the average Precision, which is the score of the obtained detection accuracy. The case numbers: Case 1 is benchmark, Case 2 is t-SNE and k-NN, Case 3 is k-means++,

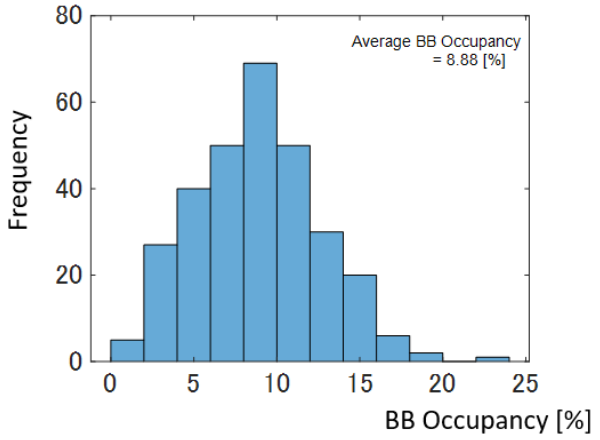
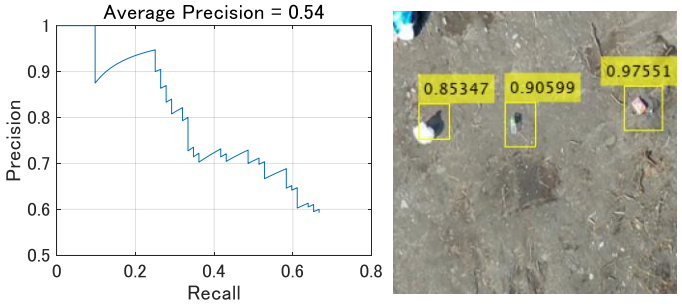
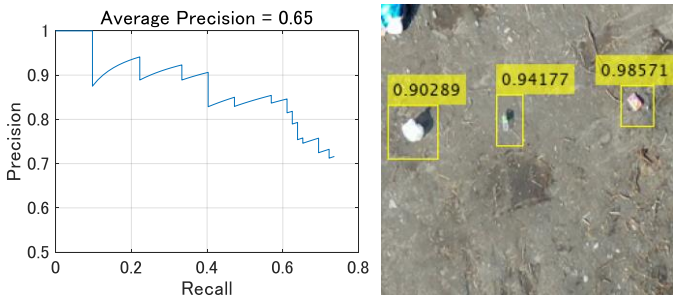


Figure 7. Histogram of BBO in aerial images.



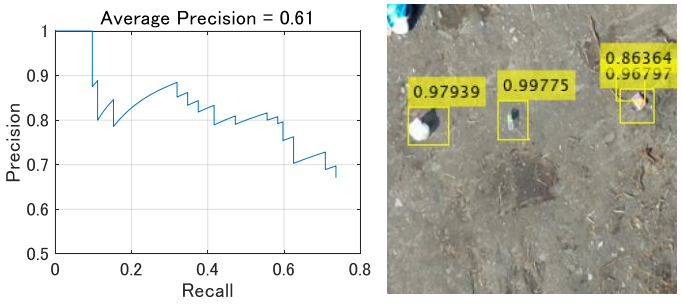
(a) The PR curve. (b) Inference sample.

Figure 8. BBO.



(a) The PR curve. (b) Inference sample.

Figure 9. ShuffleNet.



(a) The PR curve. (b) Inference sample.

Figure 10. Inception v3.

Table-1 The ratio of aerial and ground image in each experiment.

Case	1	2	3	4	5	6
Aerial image	280	83	277	280	280	280
For Case 1	0	-197	-3	0	0	0
Ground image	0	215	482	95	161	109
Total	280	298	779	375	441	389

Unit: image.

Table-2 Average Precision

Case	1	2	3	4	5	6
Average Precision	0.50	0.0004	0.002	0.54	0.65	0.61
For score of Case 1	0	-0.4996	-0.498	+0.04	+0.15	+0.11

Table-3 The parameter and score in training deep network for prefilter.

	ShuffleNet	Inception v3
Learning Rate	0.0001	
MBS	128	32
FVL [%]	0.15	0.14
FVA [%]	0.93	0.98

Case 4 is BBO, Case 5 is ShuffleNet, and Case 6 is Inception v3. Due to the selection by pre-filtering, it is possible that the number of aerial images will decrease, however, it can be considered that the data is also cleansed by narrowing down the features to be learned. In order to confirm the trade-off between data decreasing and cleansing, the increase / decrease of the aerial image used against the benchmark is compared. In Case 2, which has the lowest score, the number of ground images increased significantly, however the number of aerial ones decreased. Because many aerial images with lower similarity than ground images are excluded based on the features in Case 2, it suggests that the training dataset to learn becomes too less. In Case 3, where the number of aerial images used is almost the same as the benchmark, it seems that more ground images can be added to increase the training data, however the score deteriorated. It is suggested that learning became difficult by selecting ground images with features which are not common to aerial images. However, since plastic bottles that could not be detected by the benchmark were detected, thus it is suggested that addition of the ground image can complement the object which is difficult to detect only by the features of the aerial image.

Case 4-6 learning uses the same number of aerial images as the benchmark. The addition of ground images can improve the score and plastic bottles were successfully detected as in Case 3. According to the previous study, if the ratio of the background and the object is biased, AI assumes the correlation between the background and the object, and makes it difficult to learn properly [25]. This suggests that it is effective to utilize images with different backgrounds for robustness of AI. Based on this suggestion, in addition to complement of the training data as in the case of data augmentation, by selecting an appropriate image feature amount, different situations (e.g.: an image of a plastic bottle or an illegal bonfire itself) can be combined with back ground in aerial image, and they can be added to the training data. Therefore, it may detect even for object which cannot be obtained from aerial images. From these experiments and consideration, improvement of training data by adding images with different shooting conditions such as ground images to the aerial images taken is confirmed.

Finally, deep networks as prefilter are focused on. Inception v3 has more parameter and larger image size, however, ShuffleNet becomes the best model in this experiment. At first, in Case2 and 3, it can be assumed that the similarity is calculated based on each probability density function (PDF), and selected images have similar shape of PDF because of selection of similar stochastic values (e.g. standard deviation) corresponding to edge. On the other hand, in Case 4, BBO can be assumed as the feature about an area of probability density function between images, thus the texture. Previous study suggests deep networks prefer the texture [26], thus deep networks can be assumed as the prefilter which has the feature about texture at least. These consideration suggests that large parameter and dense image is not dominant for improvement of dataset and fitting prefilter input size to feature extraction layer input ones in detection model is more important. The other reason why ShuffleNet is better is seem to be robustness to less data by shuffle layer. Inception v3 has larger parameter and the learning needs larger time and data. Thus, larger input size cannot be over the effect of less data and it may not be suitable in assumed situation in this study. For pre-filtering, large network often take a longer time to learned or inference, thus more small network are preferable for use case in practice. Therefore, ShuffleNet can be assumed as more appropriate prefilter. On these analysis, the validity of deep network prefilter can be confirmed.

5 CONCLUSION

The aim of this study is the improvement of dataset with overfitting avoided by addition of ground images to the training data obtained only from aerial images taken by drones. This study explored and verified appropriate indicators and methods. Conclusion is summarized in the following four.

1) Feasibility study for application of UAV and AI to river patrol was carried out by dummy illegal dumping image with data augmentation and verified. Due to avoid the overfitting and improve the dataset, six features (t-SNE + k-nearest neighbor method, k-means method, Bounding Box occupancy (BBO), ShuffleNet and Inception v3) is proposed as data selection method and their improvement was compared. Based on the proposed features, ground images with high similarity to aerial images were selected as learning data, and the effect of difference in features to score was confirmed. (Correspond to Contribution 1.)

2) The BBO, ShuffleNet and Inception v3 can improve the average Precision over the benchmark which was learned only from the aerial images. On the other hand, t-SNE and k-nearest neighbor method and the k-means method deteriorated the score. These results suggest that the training data can be improved by selection of images in pre-filtering by appropriate features. (Correspond to Contribution 1.)

3) Addition of ground images enables AI to detect objects (plastic bottles, etc.) which cannot be sufficiently detected from aerial images alone. This result suggests that addition of an appropriate feature for learning can complement the dataset even if the image has different shooting conditions from the aerial image. (Correspond to Contribution 2.)

4) As prefilter, deep network, ShuffleNet and Inception v3, are proposed. Nevertheless Inception v3 has larger parameter and denser image than other ones, ShuffleNet becomes the best prefilter in six proposal criterion. This result suggests that large parameter and dense image is not dominant for improvement of dataset and fitting prefilter input size to feature extraction layer input ones in detection model is more important. Therefore, it can be confirmed that small network such as ShuffleNet is enough to apply as the prefilter to improve the dataset. (Correspond to Contribution 3.)

The images used in this learning had the almost minimum number for both of aerial and ground, thus it cannot conclude perfectly that it doesn't any data bias and any effect of the algorithm of Faster R-CNN. However, it is considered that the feature of texture such as BBO and deep networks causes this improvement at least. In process of drone river patrols with AI become more common, various ground images can be used for supplement of features which cannot be obtained with aerial images sufficiently. RiMaDIS has hundreds of thousands of images even if it is limited to illegal dumping, this study can be applied efficiently, because BBO is objective values and deep networks are applicable for mass image data at ease. As a future work, SSD (which isn't shown due to limitations of space in this study) is compared with Faster R-CNN. The input size of SSD is 299 as same as Inception v3, thus the score can become higher based on this study. In addition, this finding may be applied to the other context in civil, for example, corrosion of steel by FTIR (Fourier Transform Infrared Spectroscopy). Through these verifications, the strategy for improvement of AI in civil engineering context with the boundary condition defined difficultly can be revealed, and AI can recognize an anomaly place and degree after disaster accurately for suppression of damage expansion and increasing resilience.

ACKNOWLEDGMENTS

This research was carried out as part of the development of innovative river technology by the Ministry of Land, Infrastructure, Transport and Tourism. The Water and Disaster Management Bureau and the Chubu Regional Development Bureau have greatly cooperated in providing support and fields for the entire project. In conclusion section, the idea of applying this finding to corrosion of steel with FTIR, which is inspired from Rina Hasuike, Project Assistant Professor in Ryukyu University, is shown. We express to appreciation for them.

REFERENCES

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] L. van der Maaten and G. E. Hinton, *Visualizing Data using t-SN*, Journal of Machine Learning Research, Vol.9, pp. 2579–2605, 2008.

[9] B. V. Dasarathy, *Nearest-neighbor approaches*, Handbook of Data Mining and Knowledge Discovery, pp. 88–298, Oxford University Press, 2002.

[10] A. David and S. Vassilvitskii. *K-means++: The Advantages of Careful Seeding*., SODA 2007: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, 2007.

[11] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li and J. Sun, *Objects365: A large-scale, high-quality dataset for object detection*, Proceedings of the IEEE international conference on computer vision, pp. 8430–8439, 2019.

[12] X. Zhang, X. Zhou, M. Lin, and J. Sun, *ShuffleNet: An extremely efficient convolutional neural network for mobile device*, arXiv:1707.01083, 2017.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. *Rethinking the inception architecture for computer vision*, CoRR, abs/1512.00567, 2015.

[14] Ministry of Land, Infrastructure, Transport and Tourism, *River Improvement Measures Taken by the MLIT*, https://www.mlit.go.jp/river/basic_info/english/river.html

[15] 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.

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

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

[18] 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.

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

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

[21] K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*, In ICLR, 2015.

[22] M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013.

[23] I. Goodfellow, *Batch Normalization OpenAI*, Deep Learning Study Group, San Francisco, 2016. https://www.youtube.com/watch?v=Xogn6veSyxA&feature=emb_title

[24] 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.

[25] M. T. Ribeiro, S. Singh and C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.1135-1144, 2016.

[26] K. L. Hermann, T. Chen and S. Kornblith, *The Origins and Prevalence of Texture Bias in Convolutional Neural Networks*, In Pre-Proceedings of NeurIPS, 2020,