Satellite Image Semantic Segmentation Using Fully Convolutional Network

Atsushi Yoshihara, Tristan Hascoet
Graduate School of System Informatics,
Kobe University,
Nada, Kobe 657-8501,Japan

Tetsuya Takiguchi, Yasuo Ariki
Organization of Advanced Science and Technology,
Kobe University,
Nada, Kobe 657-8501, Japan

***Abstract*—In an event of large-scale disaster caused by an earthquake or tsunami, it is necessary to quickly grasp the damage situation of the area. In order to address this task, studies using deep learning have been done, in which CNN showed excellent performance in various fields in recent years. However, conventional classification methods using CNN have several problems. The first is that the input image size is fixed due to the fully connected layers at the final stage. Secondly, CNN extracts high-order features of images through repeating convolution, but the local information is lost as pooling is repeated. To solve this problem, a method using Fully Convolutional Network(FCN) has been proposed [1]. The purpose of our research is to perform the pixel-wise classification, that is, semantic segmentation using FCN in order to grasp the situation after the disaster.**

I. Introduction

Every year, large-scale disasters occur all over the world and cause great damage. Among them, the Japanese archipelago is located on multiple plates, so it is a region with many earthquakes. In particular, by the Great East Japan Earthquake that occurred in 2011, the Pacific coastal area of the Tohoku region suffered tremendously. In the event of such a large-scale disaster, securing safe evacuation and rescue routes, and considering reconstruction measures are very important tasks. For these tasks, it is necessayr for a person to actually visit and investigate the site in order to collect wide area information.

However, at the time of a large-scale disaster, it can't be done at once because it involves risks such as secondary disasters. Based on this background, methods using remote sensing technology have been studied. Remote sensing is a technique of observing the reflection of electromagnetic waves and measuring the objects remotely from sensors mounted on platforms such as artificial satellites and aircraft. This technique has advantages such as remoteness, wide area and periodicity, and it is utilized in various fields such as land use survey.

In order to acquire a wide area information at once, remote sensing technology and deep learning method are used. Especially, Convolutional Neural Network (CNN) shows excellent performance in various fields such as speech recognition, image classification and natural language processing. However, conventional classification methods using CNN have several problems. The first is that the input image size is fixed due to the fully connected layer at the final stage. Secondly, CNN extracts high-order features of images through repeating convolution, but local information is lost as pooling is repeated. As a method to solve this problem, Fully Convolutional Network (FCN) has been proposed [1] for semantic segmentation. The FCN can tolerate any input map size by using a convolution layer instead of the fully connected layer. They improve the segmentation accuracy by using a skip structure that combines the feature map at the lower layer with the feature map at the upper layer. In [2], [3], the skip architecture was further improved by employing the concatenation and CNN after upsampling .

The rest of this paper is organized as follows: In Section 2, related works are described and used data in this work is described in Section 3. In Section 4, our architecture is described and the experimental data is evaluated in Section5. The final section is devoted to our conclusions and future work.

II. Related Work

Classical method for land cover classification of multispectral satellite images includes supervised classifiers such as support vector machine (SVM) [5], [6], conditional random fields (CRF) [7], [8], and random forest (RF)[9], [10], [11].

The support vector machine(SVM) is supervised non-parametric statistical learning technique. Its training algorithm aims to find a hyperplane that separates the dataset into a discrete predefined number of classes in a fashion consistent with the training examples. In [5], [6], they shows that SVMs demonstrate good performance in the remote sensing field due to improvement of the classification accuracies.

The Conditional Random Fields are a probabilistic framework for labeling and contextual classification. The CRF is a form of undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. In [7], [8] (L.Albert et al.), a two-layer CRF model is proposed for simultaneous classification of land cover and land use. This results shows their approach yields good accuracies for the land use classes.

Random Forest (RF) is proposed by Breman in 2001 for classification and clustering. RF grows many decision tree in the forest. Each tree gives a classification, and the output of the classifier is determined by a majority vote of the trees. In [10] (Ozlem Aker et al.), the classification results of RF classifier are compared with the results obtained from other

TABLE I: Spectral bands used in the multispectral imagery

Band	Bandwidth [nm]
Red	$655 - 690$
Green	$510 - 580$
Blue	$450 - 510$
Near-infrared	$780 - 920$
Panchromatic	$450 - 800$

classification algorithms to evaluate RF performance. The experimental results indicate that RF algorithm gives higher classification accuracies than other methods.

While, in recent years, a Convolutional Neural Network (CNN) has shown excellent performance in various fields, such as speech recognition, image recognition and natural language processing[12], [13]. CNN consists of various combination of the convolutional layers, pooling layers and fully connected layers. They tightly couple feature extraction, model construction and classification.

However, the conventional method using CNN has the following problems: (1) the input map size is restricted, (2) local features are lost by passing through the pooling layer. In order to solve this problem, in [1], they construct a network that converts the fully connected layer of the final layer into a convolution layer. With this structure, it is possible to input a map with an arbitrary size. Also, in [1], [2], [3], they combined the information of the lower layer with the feature map of the upper layer, so that they learned the local information without losing it.

In this paper, inspired by [2], we constructed a network for producing the pixel-by-pixel multi-class segmentation map of the satellite image.

III. Data and study area

A. Satellite Image

The data used in this work are Geoeye-1 satellite images obtained from Geoeye-1 sensor with very high spatial resolution. Table. I shows the bands and their respective bandwidths. This sensor has 4 bands of R, G, B, and Near in-frared. The size of the orthographic images is 10,312 × 10,314 pixels with a spatial resolution of 0.5 meters per pixel.

The study area is located in Ishinomaki city damaged by the tsunami by the Great East Japan Earthquake that occurred in 2011. The satellite image is shown in Fig. 1.

B. Ground Truth

As ground truth, we used detailed map data collection provided by Esri Japan. The area such as "Facility area", "Road area", "Water area" and "Others" are automatically extracted from this data. The ground truth is shown in Fig. 2. In the figure, red, yellow, blue and black indicate "Facility area", "Road area", "Water area" and "Others" respectively. In this study, the ground truth was acquired from the detailed map, so that "Road area" may be included in "Facility area" in some cases.

Fig. 1: Study Area

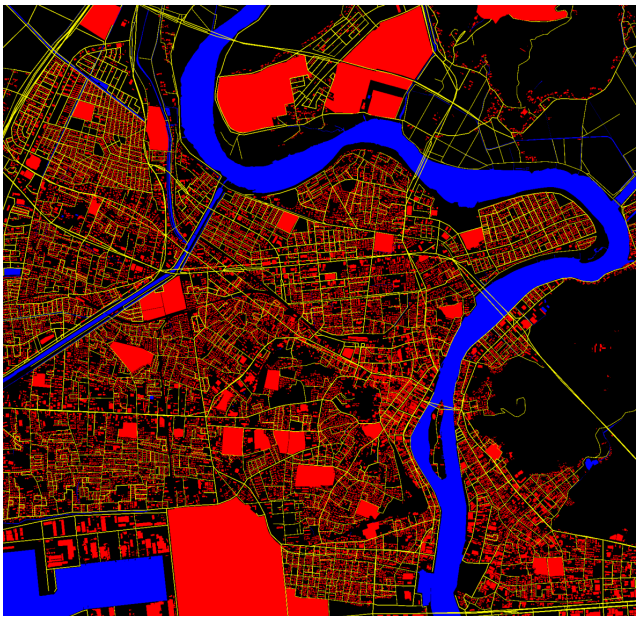

Fig. 2: Ground Truth

IV. Architecture

Fig. 3 shows the network structure for producing the pixel-by-pixel multi-class segmentation map for the satellite image by FCN used in this study. Our fully convolutional network architecture consists of encoder network and the corresponding decoder network like [2], [4]. Particulary, our architecture has a structure similar to the network of [2] where low-

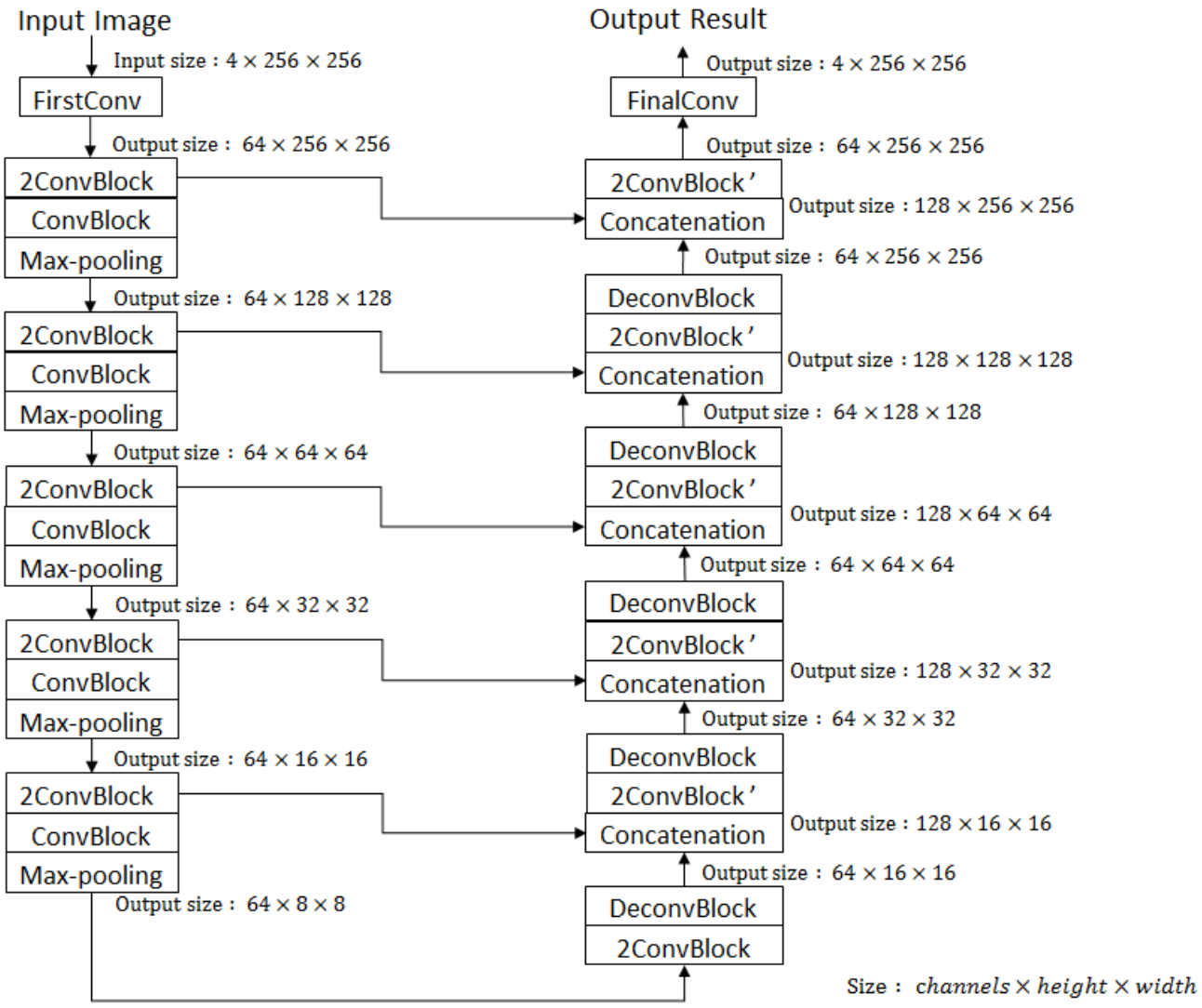

Fig. 3: Model Architecture

level feature maps are concatenated with high-level feature maps.The input size to the network has been changed from the employed value in [2] to 256 × 256.

The encoder network follows the typical architecture of a convolutional network. It consists of the repeated application of batch normalization, two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. In the convolution process, the value of padding is set to 1 so that the size of the feature map (height * width) becomes constant for the sake of concat processing simplification.

The decoder network consists of the repeated application of batch normalization, two 3x3 convolutions, 3x3 deconvolutions, a concatenation with the correspondingly cropped feature map from the encoder network and each followed by a rectified linear unit (ReLU) for upsampling. In the final layer, a 1x1 convolution is used to set the number of channels to the desired number of classes.

V. Experiments

Examples of the ground truth and the segmentation result are shown in Fig. 4 and 5 respectively when the test image is fed to the learned model. For the evaluation of the model, the confusion matrix of the whole batch was calculated. Table. II shows an example of confusion matrix of randomly selected batches. Table III shows calculation result of Recall and Precision for each class using this confusion matrix. Also, we calculate Overall accuracy, Mean Accuracy, Mean IoU, and IoU per class. Table IV shows these accuracy evaluation results.

From this result, it turns out that ”Fasility area” and ”Road area” are not well classified. Since these objects exist in a complicated and widly scattred way, it is thought that the boundaries can’t be well classified. On the other hand, since ”Water area” exists simply and densely in a specific area, it can be thought that it could be classified successfully.

TABLE II: Example of Confusion Matrix

		Ground Truth			
		Facility	Road	Water	Others
Predicted	Facility	1538870	2480	0	234090
	Road	6470	447230	0	535800
	Water	0	760	658510	4760
	Others	156600	50130	800	3370800

TABLE III: Recall and Precision of TABLE I （%）

	Facility area	Road area	Water area	Others
Recall	90.42	89.34	99.88	91.31
Precision	86.67	83.47	99.17	94.20

TABLE IV: Example of various evaluation indices （%）

Overall Accuracy	91.79
Mean Accuracy	90.88
Mean IoU	85.20

Class IoU	
Facility area	79.38
Road area	75.91
Water area	99.05
Others	86.45

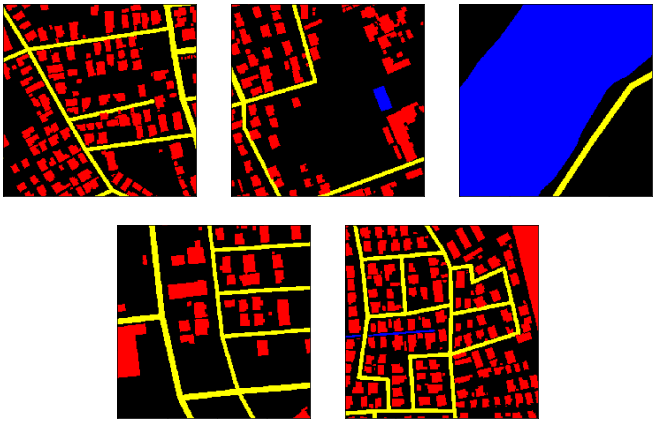

Fig. 4: Ground Truth

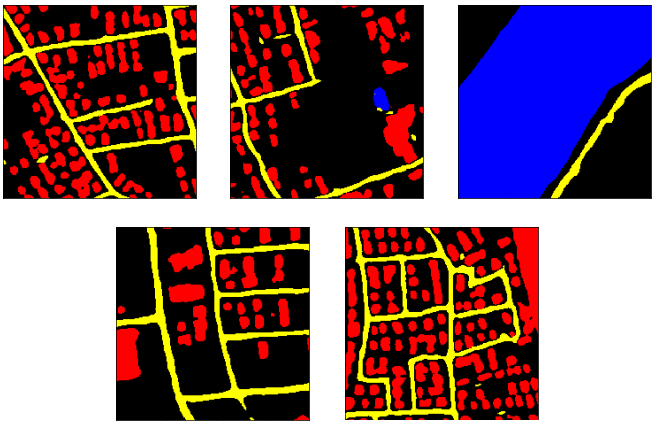

Fig. 5: Segmentation result

VI. Conclusion and Future Work

In this paper, we described semantic segmentation for pre-disaster satellite images using FCN as preliminary studies. In the experimental results, ”Facility area” and ”Road area” are classified slightly lower than ”Water area”. This is due to the fact that the boundaries can’t be well classified because these objects exist widly and complicated. Also, since ground truth was created from a detailed map, there were cases where even ”Road area” were labeled ”Facility area”. From this result, we found improvement will be further required for practical use.

In the future, we aim to produce geographical information after disaster.

Acknowledgement

A part of this study is subsidized by JSPS Grant-in=Ais for Scientific Research and Research granted JP 17K00236.

References

[1] Jonathan Long, et al. ”Fully Convolutional Networks for Semantic Segmentation.” CVPR (2015) : pp. 3431?3440.

[2] Olaf Ronneberger, et al. ”U-Net: Convolutional Networks for Biomedical Image Segmentation.” International Conference on Medical Image Computing adn Computer-Assisteed Intervention (2015)

[3] Simon Jegou, et al. ”The One Hundred Layers Tiramisu:Fully Convolutional DenceNets for Semantic Segmentation.”

[4] Vijay Badrinarayanan, et al. ”SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.” IEEE (2015)

[5] Giorgos Mountrakis, et al. ”Support vector machines in remote sensing.” ISPRS Journal of Photogrammetry and Remote Sensing 66 (2011): 247-259.

[6] Yuliya Tarabalka, et al. ”Spectral?Spatial Classification of Hyperspectral Imagery Based on Partitional Clustering Techniques.” IEEE Geoscience and Remote Sensing Society (2009): 2973-2987.

[7] L.Albert, et al. ”A TWO-LAYER CONDITIONAL RANDOM FIELD MODEL FOR SIMULTANEOUS CLASSIFICATION OF LAND COVER AND LAND USE.” Photogrammetry Remote Senssing and Spatial Information Sciences (2014): Volume XL-3, 17-24.

[8] L.Albert, et al. ”LAND USE CLASSIFICATION USUNG CONDITIONAL RANDOM FIELDS FOR THE VERIFICAION OF GEOSPATIAL DATABASES.” International Society for Photogrammetry and Remote Sensing (2014): Volume ll-4, 14-16.

[9] Jon Atli Nenediktsson, et al. ”Random Forests for land cover classification.” Pattern Recognition Letters 27 (2006): 294-300.

[10] Ozlem Akar, et al. ”Classification of multispectral images using Random Forest algorithm.” Journal of Geodesy and Geoinformation (2012): 105-112.

[11] Barrett Lowe, et al. ”MULTISPECTRAL IMAGE ANALYSIS USING RANDOM FOREST.” International Journal on Soft Computing (2015): Volume6 No.1

[12] Wei Hu, et al. ”Deep Convolutional Neural Networks for Hyperspectral Image Classification.” Journal of Sensors (2015): Article ID: 258619, 12 pages.

[13] Tomohiro Ishii, et al. ”Detection by Classification of Buildings in Multispectral Satellite Imagery.” International Conference on Pattern Recognition (2016): Paper ThAT2.1

[14] Martin Langkvist et al., ”Classification and Segmentation of Satellite Orthoimgery Using Convolutional Neural Networks.” Remote Sensing (2016): vol.8 ,issue 4, p.329.