14th IAPR International Conference on Machine Vision Applications (MVA)
May 18-22, 2015. Miraikan, Tokyo, Japan

Surface Object Recognition with CNN and SVM in Landsat 8 Images

Tomohiro Ishii[1,2,3], Ryosuke Nakamura[3], Hidemoto Nakada[3],
Yoshihiko Mochizuki[1,2], and Hiroshi Ishikawa[1,2]
[1]Waseda University, Japan
[2]JST CREST
[3]National Institute of Advanced Industrial Science and Technology, Japan

Abstract

There is a series of earth observation satellites called Landsat, which send a very large amount of image data every day such that it is hard to analyze manually. Thus an effective application of machine learning techniques to automatically analyze such data is called for. In surface object recognition, which is one of the important applications of such data, the distribution of a specific object on the surface is surveyed. In this paper, we propose and compare two methods for surface object recognition, one using the convolutional neural network (CNN) and the other support vector machine (SVM). In our experiments, CNN showed higher performance than SVM. In addition, we observed that the number of negative samples have a influence on the performance, and it is necessary to select the number of them for practical use.

1 Introduction

There is a series of satellites called Landsat, whose purpose is to observe the surface of the Earth. The Landsat series have been continuously operated for more than 40 years and have collected data over a long period of time. Landsat 8, which is the latest Landsat, was launched in 2013 and has a 16 day repeat cycle [1]. It sends 500GB of image data every day, which are all archived and made freely available. The amount of Landsat data is so large that it is hard to analyze all images manually. Thus, it is an important application area for machine learning and machine vision techniques.

Surface object recognition is one of the important applications of Landsat data. Its purpose is to survey the distribution of a specific object on the surface. One mode of surface object recognition is a binary classification of each pixel or localized region (cell) of a satellite image according to whether or not it represents the object of interest. There is a related work by Mnih et al. [2], in which they recognize road in aerial images using deep learning.

One issue in surface object recognition in satellite images is that the region representing the object of interest is extremely small compared with the region that represents everything else. This tendency is called the imbalanced data [3] and causes a low performance in recognition. Therefore, it is necessary to devise the learning and the evaluation methods in such a way that this issue is addressed.

In this paper, we analyze the surface object recognition in Landsat 8 images. We use and compare the convolutional neural network (CNN), which is one of the methods called deep learning, and more conventional support vector machine (SVM) for classification of features in the images. Deep learning is a machine learning method that recently gave a large impact in the field of image recognition. In addition, we use an undersampling of learning data and contrive the evaluation methods as a countermeasure against the issue of imbalanced data, in an attempt to improve the recognition performance.

2 Problem Formulation

In this paper, we propose two methods for surface object recognition in satellite images. We consider the satellite image classification tasks into binary classes, where the goal is to classify all localized regions (cells) of the image into those which include an object of interest (positive) and those that do not (negative). We use CNN and SVM, which are used in object recognition, as recognition methods, and analyze and compare the results of the two methods.

We use Landsat 8 satellite images [1]. Landsat 8 image is a multi-band image that shows a rectangle of approximately 185km $\times$ 180km with a 30m spatial resolution in the main band. In this paper, we use the three bands (4, 3, 2) that are near the wavelength region of RGB.

3 Methodology

We investigate the surface object recognition of a satellite image from two points of view. The first point is the recognition methods, and we use CNN and SVM. The second is a countermeasure against the issue of imbalanced data, and we investigate an undersampling of data and evaluation methods.

3.1 Convolutional neural network

CNN is a feed-forward network which repeat the two calculations of convolution and pooling alternately [4].

In image recognition, CNN is a classifier that gives the probability of each class given an image. In this paper, we use the CNN of Krizhevsky's method for learning and recognition.

3.2 Support vector machines

SVM is a learning method of a linear discriminant function for binary classification problem which realizes the maximum margin [4].

In image recognition, SVM is a classifier that outputs the class that the image belongs, given preselected features of the image. In this paper, we use the

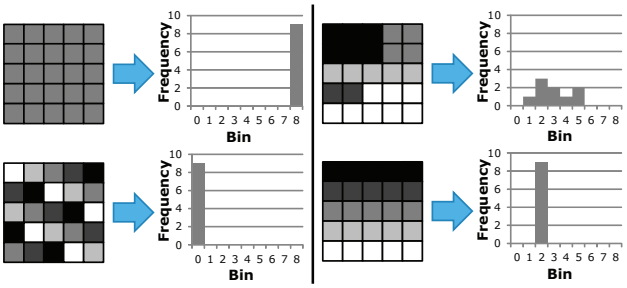


Figure 1. Examples of the neighboring-peer count histogram (NPCH) feature in cell. These are four examples of generating the NPCH in cell and they correspond to the process 1 and 2 in the text.

kernel method using Radial Basis Function (RBF) for learning and recognition. Parameters of the method, (C, γ), are searched by a grid search.

3.3 Peer Count Histogram

In the case of image classification, in addition to using a vector that is transformed from an image, there is a method that uses features of an image such as color histogram [5]. Here, we propose a histogram-based texture feature, which we call the neighboring-peer count histogram (NPCH), which improves upon the color histogram. The NPCH is invariant to illumination and viewpoint of satellite images.

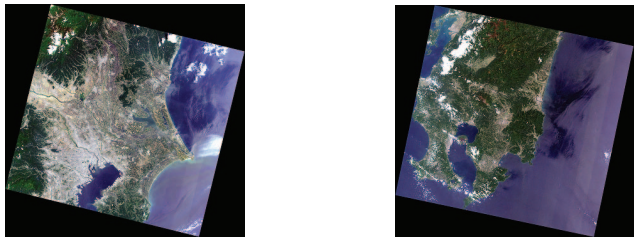
The feature NPCH represents the spatial spread of each color in the image. Examples of the NPCH in cell is shown in Figure 1, which is generated by the following procedure.

1. For each pixel p in a band image (except for the ones at the peripheral), count the number of neighboring pixels with the value within a predetermined tolerance from the value at p.
2. Make the histogram of the count for all pixels (except for the ones at the peripheral) with nine bins $(0, \ldots, 8)$.
3. Do the same for all three band images and combine the resulting histograms into one 27-dimensional vector.
4. Normalize the vector by dividing it by the number of pixels which is not in the peripheral.

3.4 Data generation and evaluation methods

In this paper, we divide a Landsat 8 image into cells, and determine whether or not each cell contains the object of interest. As Landsat has low spatial resolution, the object that can be recognized is limited to large ones. Here, we choose golf courses as the object of interest, since it has a chance of being recognized even with a resolution of Landsat. In addition, we make the cell size 16×16 pixels so that a golf course can be easily contained in one cell. The cell size should be altered depending on the size of the object of interest.

One of the problems of the object recognition in a satellite image is an availability of training data. In this paper, we have made the ground truth of the golf course manually because it can be recognized by visual inspection.



(a) The training image (KT)

(b) The evaluation image (KG)

Figure 2. Landsat 8 images we use in this experiment [7]. The region of the training image is different from that of the evaluation image. They are visualized with 3 bands (Red=Band 4, Green=Band 3, Blue=Band 2).

In addition, there is the issue of imbalanced data, which exhibits an unequal distribution between its classes [3]. Specifically, when we divide the satellite image into an object of interest and everything else, typically very few of the cells represent golf courses. This makes the evaluation difficult, since just determining all cells to be negative (not golf course) would be a fairly accurate classification in terms of the ratio of the correctly-determined cells (accuracy).

In this paper, we use the random undersampling method [3] in order to avoid imbalanced learning. We also calculate a confusion matrix, and use precision, recall and F-value[6] as the performance measure in order to evaluate the recognition results more accurately.

4 Experiment

To evaluate the recognition performances of the CNN and the SVM, we conduct the surface object recognition in the set of satellite images. In addition, since the data becomes imbalanced in case of surface object recognition in the satellite image, we also evaluate the influence of the undersampling for improving the recognition result.

4.1 Study area

We use two Landsat 8 images shown in Figure 2. Figure 2(a) is the image in Kanto region in Japan taken on May 31, 2014 (LO81070352014151KUJ00), and Figure 2(b) is the image around Kagoshima in Japan taken on May 2, 2014 (LC81120382014122KUJ00). In this experiment, we use the image in Kanto region as the training data and use the image around Kagoshima for evaluation in order to evaluate the generalization capability of the classifiers.

In this experiment, we divide a satellite image into cells as mentioned in chapter 2. The Table 1 shows the results of dividing the images of Figure 2. The result of dividing the training image is $\mathrm{KT}_{\mathrm{ALL}}$ and that of dividing the evaluation image is KG. We also do undersampling for $\mathrm{KT}_{\mathrm{ALL}}$ as the number of negative are 80000, 40000, 20000 and 10000, and the results of them are $\mathrm{KT}_{80\mathrm{K}}$, $\mathrm{KT}_{40\mathrm{K}}$, $\mathrm{KT}_{20\mathrm{K}}$ and $\mathrm{KT}_{10\mathrm{K}}$. In addition, we remove the cells which are located in the four corners of the image and where there are not any data in Figure 2.

Table 1. The number of samples in each dataset. The result of dividing the Figure 2(a) is KT_{ALL} and that of dividing the Figure 2(b) is KG. KT_{80K}, KT_{40K}, KT_{20K} and KT_{10K} are the undersampling version of KT_{ALL}.

	Label	positive	negative
training	KT_{ALL}	2155	149175
	KT_{80K}	2155	80000
	KT_{40K}	2155	40000
	KT_{20K}	2155	20000
	KT_{10K}	2155	10000
test	KG	259	151194

4.2 Training of CNN

We train CNN on each training dataset and evaluate it on KG. In this experiment, we use cuda-convnet [8] for the implementation of CNN. In addition, the architecture of CNN tuned for CIFAR-10 [9] are provided and it can recognize CIFAR-10 with error rates of 11%. In this experiment, we change the input image size to 14×14 and the number of units in the final output layers to 2 on the basis of the architecture.

We conduct data augmentation [8] which consists of generating image translations and horizontal reflections. We do this by extracting random 14×14 patches and the horizontal reflections from the 16×16 images and training the network on these extracted patches. We also subtract the pixel values of mean image of training images from that of each training image and input them to CNN.

In the learning phase, we initialize each weight of CNN by a random number that follows a normal distribution with mean 0 and variance σ^2, where variance σ^2 is different for all layers. We divide a training data set D into 5 batches D_i which are disjoint each other. First, CNN is trained on batches $D_1, \ldots, D_4$ for 500 epochs. Second, CNN is trained on batches $D_1, \ldots, D_5$ for 250 epochs. Then, repeat the training on $D_1, \ldots, D_5$ for 10 epochs with decreasing the learning rate.

4.3 Training of SVM

We use the SVM with a kernel of RBF. We train it on each training dataset and evaluate it on KG as with CNN. In this experiment, we use modified LIBSVM [10] for the implementation of SVM. It is important to tune the parameters, (C, γ), for the recognition task by SVM. We do a grid search and search a appropriate values. As a result of the grid search, the found parameters was $(1024, 1)$.

4.4 Results and discussions

Table 2 shows the experimental results of CNN and SVM. Figure 3 shows the result of undersampling about F-value [6]. Figure 5 shows an example of the results of recognizing golf courses in the evaluation image.

Comparing the recognition methods, it is clear that the performance of CNN is above that of SVM in any metric. Therefore, we conclude that CNN is more useful for surface object recognition in satellite images. This is also clear from Figure 5.

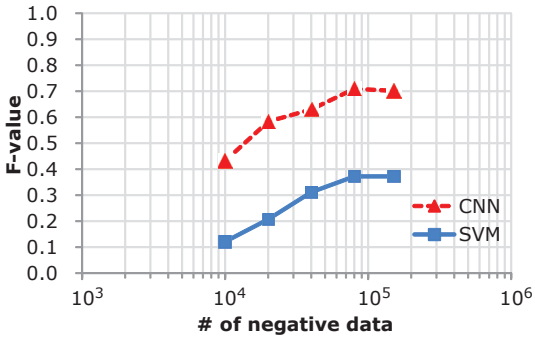


Figure 3. The evaluation result of F-value. This graph shows the changes of F-value in Table 2.

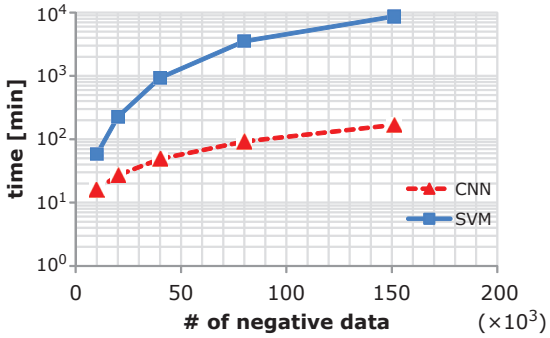


Figure 4. The evaluation results of the training time. This graph shows the changes of the training time of CNN and SVM

Comparing the results on the effect of undersampling, precision falls and recall rises as more undersampling is performed in both recognition method. From this, it can be seen that undersampling increases both the false positive and the true positive. This raises the question of how much increase in false negative we should tolerate. In terms of the F-value, the case of KT_{80K} is the best result. The reason for this may be that the cells corresponding to the boundaries of the golf course that are contained in the negative samples have been removed.

We show the training time of each recognition method in Figure 4. From this, it can be seen that the CNN also shows better results than the SVM in training time. However, it should be noted that, in this experiment, the CNN is computed by both CPU and GPU, whereas only the CPU is used for computing the SVM.

According to Figure 3 and 4, the more the number of negative samples are, the longer the training times are, but the performance is not the best when the number of negative samples is maximum. Therefore, it is necessary to select the number of them for practical use.

5 Conclusion

In this paper, we have presented an approach for object recognition in the Landsat 8 image using CNN and SVM. As a result, we observed that the CNN performs better than the SVM as a recognition method. We also confirmed that the result of the undersampling to reduce to half the negative samples is the best result of

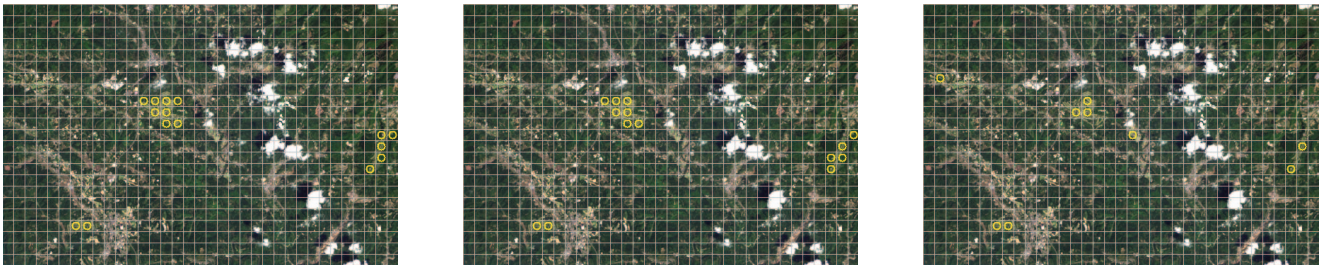
Table 2. The recognition results of CNN and SVM. These tables show the recognition results on the dataset KG by CNN and SVM which are trained on 5 datasets shown in Table 1. In this experiment, the weight β of F-value is 1.

CNN

Training data	TP	TN	FP	FN	precision	recall	F-value
KT_{ALL}	161	151155	39	98	0.805	0.622	0.702
KT_{80K}	180	151126	68	79	0.726	0.695	0.710
KT_{40K}	185	151051	143	74	0.564	0.714	0.630
KT_{20K}	196	150977	217	63	0.475	0.757	0.583
KT_{10K}	213	150680	514	46	0.293	0.822	0.432

SVM

Training data	TP	TN	FP	FN	precision	recall	F-value
KT_{ALL}	75	151125	69	184	0.521	0.290	0.372
KT_{80K}	105	150994	200	154	0.344	0.405	0.372
KT_{40K}	130	150749	445	129	0.226	0.502	0.312
KT_{20K}	160	150070	1124	99	0.125	0.618	0.207
KT_{10K}	181	148648	2546	78	0.066	0.699	0.121



(a) The ground truth (b) The result of CNN (c) The result of SVM

Figure 5. The recognition results of the golf courses [7]. This figure shows the recognition results for the part of Figure 2(b). (a) is the ground truth. (b) is the result of CNN and (c) is that of SVM. The cells including a yellow circle are recognized as the golf courses.

all undersampling.

In the future work, it is conceivable to improve the performance of classifiers by using the false positive data as new negative samples. In order to improve the performance of CNN, more bands of the satellite images could be used and more optimal network architecture could be constructed. In order to improve the performance of SVM, the use of a Bag-of-features with SIFT as the feature may be worth testing.

Acknowledgments

This work was partially supported by KAKENHI 24300075 and 25540075 from JSPS as well as CREST from JST.

The source data was downloaded from AIST's Landsat-8 Data Immediate Release Site, Japan. Landsat 8 data courtesy of the U.S. Geological Survey.

References

[1] DP Roy, MA Wulder, TR Loveland, CE Woodcock, RG Allen, MC Anderson, D Helder, JR Irons, DM Johnson, R Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

[2] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 567–574, 2012.

[3] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag New York, Inc., 2006.

[5] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[6] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer, 2010.

[7] The Landsat-8 data real-time release site, Japan. `http://landsat8.geogrid.org/`.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1106–1114, 2012.

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.