15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Damage Detection from Aerial Images via Convolutional Neural Networks

Aito Fujita*, Ken Sakurada†, Tomoyuki Imaizumi*,
Riho Ito*, Shuhei Hikosaka*, Ryosuke Nakamura*

*Advanced Industrial Science and Technology
Tokyo, Japan
fujita.713@aist.go.jp

†Nagoya University
Aichi, Japan

Abstract

This paper explores the effective use of Convolutional Neural Networks (CNNs) in the context of washed-away building detection from pre- and post-tsunami aerial images. To this end, we compile a dedicated, labeled aerial image dataset to construct models that classify whether a building is washed-away. Each datum in the set is a pair of pre- and post-tsunami image patches and encompasses a target building at the center of the patch. Using this dataset, we comprehensively evaluate CNNs from a practical-application viewpoint, e.g., input scenarios (pre-tsunami images are not always available), input scales (building size varies) and different configurations for CNNs. The experimental results show that our CNN-based washed-away detection system achieves 94–96% classification accuracy across all conditions, indicating the promising applicability of CNNs for washed-away building detection.

1 Introduction

In the event of catastrophic disasters such as an earthquake and subsequent tsunami, it is imperative to identify washed-away buildings quickly. To facilitate this process, pre- and post-tsunami aerial (or satellite) images are used along with pre-tsunami building maps.

However, currently, much of the identification is performed manually; i.e., by combing through pre- and post-tsunami images and checking an intimidating number of buildings in the images. This makes the process painstakingly slow and costly. In particular, for extensively devastated areas, it is no longer practical to obtain the results immediately after the disaster. Therefore, an automated damage detection system that is accurate and applicable at scale is indispensable.

Achieving an efficient damage detection algorithm has been difficult primarily because of the absence of both labeled datasets and effective feature extraction approaches for aerial or satellite images. Our contributions are as follows:

(1) For the scarcity of a labeled dataset, we construct a new benchmark dataset, the AIST Building Change Detection (ABCD) dataset. We intend to release this dataset to the public. The details are presented in Section 3.

(2) In terms of effective feature extraction, we harness Convolutional Neural Networks (CNNs). The efficacious use of CNNs for washed-away building detection is explored comprehensively (Sections 4 and 5).

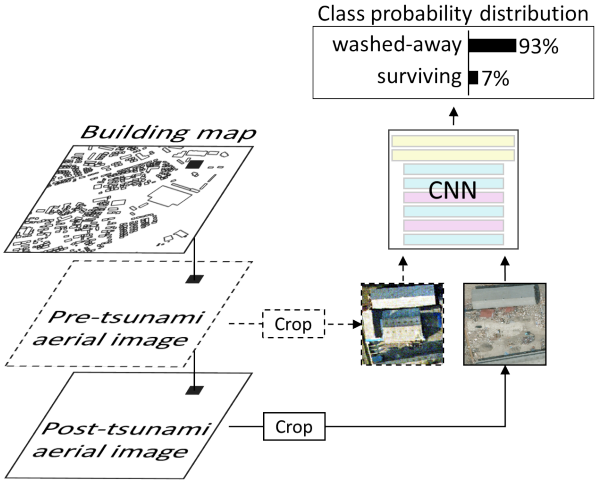

Figure 1. Overview of the proposed CNN-based washed-away building detection system. Based on the patches cropped from pre- and post-tsunami images, the CNN predicts whether a building at the center of the patch is washed-away or surviving.

Figure 1 shows a schematic of the proposed CNN-based washed-away building detection system. A brief description is as follows: (i) inputs to the system are a pre-tsunami building map and pre- and post-tsunami aerial images; (ii) for a given building in the map (black square), a patch corresponding to the building is cropped from the pre- and post-tsunami images; (iii) the two cropped patches are fed *in pairs* into a CNN; (iv) finally, the CNN predicts whether the building at the center of the patch is washed-away; and (v) (ii) to (iv) are repeated for each building in the building map.

Note that the system may be devoid of pre-tsunami images because such images are not necessarily available, and special attention should be paid to the crop size because there are various building sizes. Our goal is to explore CNNs that are suitable for such practical issues.

2 Related work

Several studies have examined damage detection using aerial or satellite images [1, 2]. These studies used pre- and post-event satellite images to detect the changes caused by natural or anthropogenic disasters. Gueguen and Hamid [2] showed that their hand-engineered features (dedicated to encoding object shapes) outperformed CNN-derived features. However, they did not explore the use of CNNs in details;

Figure 2. Six samples in the ABCD dataset. Each consists of pre- and post-tsunami patches. The target building is at the center for "washed-away" (left column) and "surviving" (right column).

i.e., they simply fine-tuned an ImageNet-pretrained CNN with ad-hoc object classes. Cooner et al. [1] evaluated a two-layer neural network but left the exploration of CNNs for future work. Differing from these studies, we explore several CNN configurations and train the networks from scratch using relevant data and classes.

The combined use of pre- and post-tsunami images in a single CNN is analogous to a comparison of natural image patch pairs [3, 4] and matching street- and aerial-view images [5]. Such pair comparisons have been addressed using one-branch [4] or two-branch (also called Siamese) [3, 4, 5] CNNs, and they achieved the best performance by a significant margin compared to hand-crafted features such as SIFT [6]. We evaluate these configurations in the context of damage detection from pairs of aerial images.

3 ABCD dataset

The ABCD dataset is a new benchmark geared toward constructing and evaluating damage detection systems to identify whether buildings have been washed-away. This dataset will be available at our web site.

3.1 Sources: a building map and aerial images

To establish a ground truth, we utilized the existing results of a post-tsunami survey. This survey result, created by MLIT (Ministry of Land, Infrastructure, Transport and Tourism) [7], is the outcome of an exhaustive investigation in the wake of the Great East Japan earthquake on March 11, 2011. The survey assessed over 220,000 buildings in the ravaged areas. Each building was assigned one of the seven designated damage levels (ranging from no-damage to washed-away). In the following experiments, these levels are reorganized into two levels, i.e. "washed-away" and "surviving", because there is no visual difference between the remaining levels (except for "washed-away") in aerial images.

In addition to the survey results, we obtained several pairs of pre- and post-tsunami aerial images covering 66 km^2 of tsunami-affected areas from the PASCO image archive [8]. The pre-tsunami images were acquired in August 2000 at a resolution of 40 cm, and the post-tsunami images were acquired within one month after

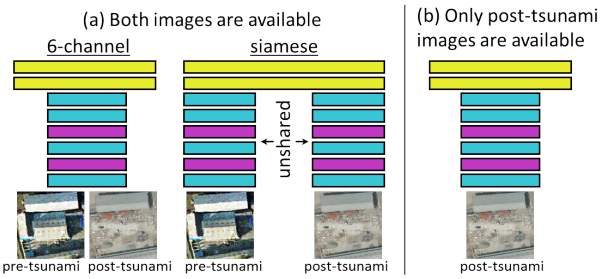

Figure 3. CNN configurations depending on the input scenarios. The two types in (a) are used when pre- and post-tsunami images are available. `siamese` has two distinct branches. Color blocks show Conv for cyan, max-pooling for purple and fully connected layer for yellow.

March 11, 2011 at a resolution of 12 cm. The post-tsunami images were resampled to 40 cm resolution and each pair of images was georegistered.

3.2 Derivation and details

Using the aforementioned sources, we cropped paired patches from pre- and post-tsunami images such that one patch enclosed a target building at its center. Note that a label assigned to each patch represents whether the target building is washed-away.

Three patch sizes were prepared: `fixed-scale`, `size-adaptive` and `resized`. For `fixed-scale`, the resolution of the patches was the same as the original images, i.e., 40 cm/pixel. The patch size was set as 160 $\times$ 160 pixels in order to crop the images such that the buildings were displayed in a reasonable context. For `size-adaptive`, patches were cropped depending on the size of the target building. Specifically, the crop size was three times larger than that of each target building because in the case of `fixed-scale`, a very small building is likely to be less conspicuous within a patch; `resized` is simply the `size-adaptive` patch resized to 120 $\times$ 120 pixels.

The resulting ABCD dataset comprised 10,777 pairs for `fixed-scale` (4,253 washed-away) and 11,394 pairs for `size-adaptive` and `resized` (4,546 washed-away). Some examples for `fixed-scale` are illustrated in Figure 2.

4 Proposed framework design

4.1 Problem formulation

As shown in Figure 1, the proposed system utilizes building maps to locate buildings in aerial images. We designed the CNN with regard to classification, i.e., it takes a patch (in pairs) as input and predicts whether a building at the center of the patch is washed-away. Formally, if $x \in \mathbb{R}^{hwc}$ denotes the input patch, where h and w are the patch height and width, respectively, c is the number of channels, and $y \in \{0, 1\}$ denotes the corresponding ground-truth binary label (washed-away or surviving), the CNN maps x to $p(y \mid x)$.

As in typical classification networks, we place a logistic sigmoid layer as the final output layer for all the configurations (explained in Section 4.2). During training, we compute cross-entropy loss and minimize

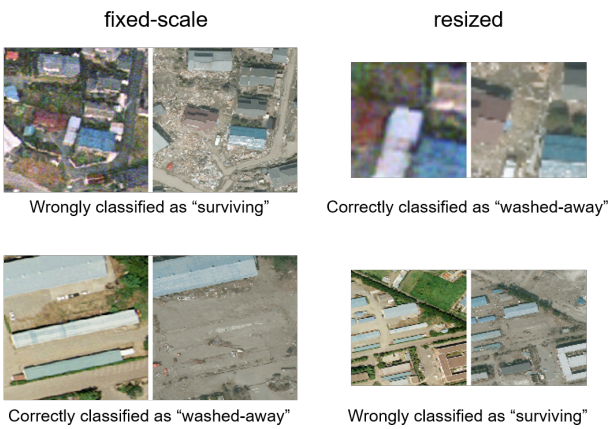

Figure 4. Fixed-scale and resized images: (top) a tiny building at the center of the patch is misclassified in the `fixed-scale` setting whereas correctly classified in the `resized` setting and (bottom) vice versa in the case of a large building. Note that the different scales are complementary with regard to washed-away classification.

the following objective:

$$\min_{w}\left[-\sum_{i=1}^{N}\{y_i \ln p_i + (1-y_i)\ln(1-p_i)\} + \frac{\lambda\|w\|_2^2}{2}\right],$$

where $y_i \in \{0,1\}$ is the ground-truth for the i-th training sample $x_i \in \mathbb{R}^{hwc}$, $p_i = p(y_i \mid x_i)$ is the corresponding network output, and w is the weight of the network.

4.2 Configuration design of CNN

As mentioned previously, in practice, pre-tsunami images may be unavailable or not readily available. Thus, we consider the following two scenarios: pre-tsunami images are accessible ($c = 6$) or not ($c = 3$). For the latter, our CNNs are reduced to familiar networks (Figure 3b) such as AlexNet [9] and VGG [10]; for the former, we evaluate two configurations on the basis of previous studies, i.e., `6-channel` [4] and `siamese` [3, 4, 5] (Figure 3a).

The `6-channel` configuration considers two patches of an input pair as a 6-channel image. Moreover, `siamese`, as seen in Figure 3, has two branches in the earlier layers, where each branch is responsible for each patch; the two branches join at a fully-connected layer in the later stage. In this study, the weights of the two branches are not shared because the appearance of pre- and post-tsunami patches is not similar (see Figure 2). Under such circumstances, having two unshared branches is more favorable because it provides flexibility and encourages each branch to learn a specific representation [5].

4.3 Patch scale

In a preliminary experiment, we observed that a CNN trained with the `fixed-scale` setting and one trained with the `resized` setting were complementary with respect to mistakes. Specifically, although these two networks had similar error rates, they tended to make different mistakes. For example, as shown in Figure 4 (top), tiny buildings were likely to be correctly classified in the `resized` setting whereas sometimes misclassified in the `fixed-scale` setting (e.g., because of surrounding distractions). The opposite was true for relatively large buildings (bottom). Consequently, we attempted to combine the two settings via an ensemble of predictions, i.e., by taking the mean of outputs from these two networks.

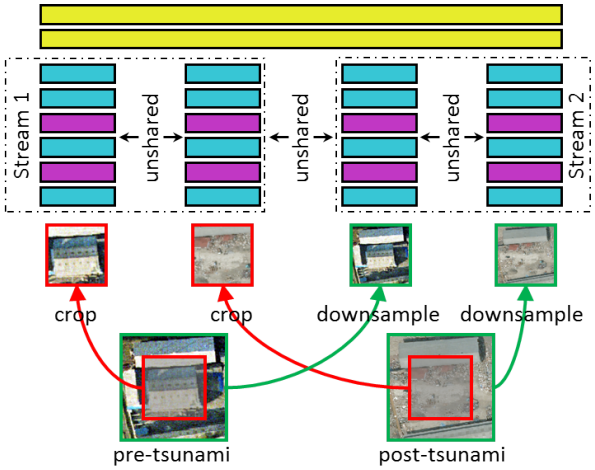

Figure 5. Schematic of a `central-surround` CNN. It can be incorporated into any configuration shown in Figure 3 (here, the `siamese` version is shown).

It is also possible for a CNN to have a multi-scale structure [4]. Such CNNs, e.g., `central-surround`, have two separate streams. Each of these streams *views* the center and the area surrounding of patches separately (Figure 5). Each stream can be replaced with any type of configuration shown in Figure 3.

5 Experiments

5.1 Setup

We performed a 5-fold cross validation to assess the accuracy of all models for the ABCD dataset. Initially, the "surviving" class was under-sampled (at random) to obtain a balanced dataset. The balanced dataset was then randomly split into five non-overlapping folds of nearly the same size, where each fold was balanced with respect to the number of patches per class. At each run, four folds were used as the training set, and the remaining fold was used as the test set. The results reported are the average of five runs followed by the standard deviation.

We employed a previously proposed architecture [11] as the base architecture for the CNNs. This network comprised four convolutional layers followed by two fully connected layers. A max pooling layer followed the first and second convolutional layers. Dropout was applied to the first fully connected layer. The nonlinear functions were all ReLU [12]. The hyperparameters such as kernel size, stride, number of units and dropout ratio were determined using Bayesian optimization [13].

In these experiments, Stochastic Gradient Descent with a constant learning rate of 0.001, momentum of 0.9, weight decay $\lambda = 0.005$ and mini-batch size of 100 was used to train the models. The weights were initialized randomly, and all models were trained from scratch. Each patch in the training set was normal-

Table 1. Accuracy for all applied CNN models

Config.	Scale	1-stream, w/o aug.	1-stream, w/ aug.	2-stream, w/o aug.
6-ch	fixed	93.8 ± 1.9	94.5 ± 0.5	95.2 ± 0.7
6-ch	resized	94.2 ± 0.7	94.7 ± 0.3	94.5 ± 0.4
6-ch	ensemble	95.2 ± 1.0	95.3 ± 0.2	95.4 ± 0.6
siam	fixed	94.6 ± 0.5	94.8 ± 0.3	94.8 ± 0.4
siam	resized	94.4 ± 0.6	94.9 ± 0.4	94.9 ± 0.5
siam	ensemble	95.5 ± 0.4	95.4 ± 0.4	95.6 ± 0.3
post	fixed	94.4 ± 0.1	94.7 ± 0.5	94.7 ± 0.7
post	resized	94.5 ± 0.6	94.5 ± 0.3	94.7 ± 0.3
post	ensemble	95.4 ± 0.4	95.1 ± 0.4	95.3 ± 0.6

ized such that they had zero mean and unit variance. We trained models for 12K iterations. Note that we did not encounter overfitting. To investigate the effect of data augmentation, all training data were augmented by four times with vertical and horizontal flipping. These settings were the same across all models. All models were implemented using Caffe [14].

5.2 Results and discussion

Table 1 summarizes the accuracy of different combinations of CNN configurations and input scales used in this study. In Table 1, "`post`" indicates "the case where only post-tsunami images are available"; "1-stream" and "2-stream" indicate CNNs without and with the `central-surround` structure, respectively. Overall, they all show reasonably good performance (roughly ranging between 94% and 96%), indicating the promising applicability of the proposed framework. Next, we discuss various aspects of the results.

Need for pre-tsunami images: First, using pre-tsunami images (cf. `post` vs. `6-ch` or `post` vs. `siam`) did not produce a clear benefit. A possible explanation for this is that in the context of washed-away detection, there is a correlation between labels and the visual appearance of post-tsunami patches. For example, in most cases, "washed-away" patches would present an earthy, relatively uniform texture, whereas "surviving" patches would have a more heterogeneous appearance. Since such a correlation will generally hold true, this result is significant from an application perspective.

Method of treating patch pairs: Two-branch CNNs (`siam`) outperform one-branch counterparts (`6-ch`), suggesting that unshared siamese networks can be more suitable for washed-away building detection.

Patch scale: A CNN trained with `fixed-scale` and that trained with `resized` are expected to work complementarily (Section 4.3). The ensemble of these scales yields consistent improvement, as expected. The same is true for the superiority of multi-scale CNNs (denoted 2-stream) over 1-stream CNNs.

Augmentation: Simple data augmentation by flipping slightly improves the performance for all configurations (cf. "1-stream, w/o aug." vs. "1-stream, w/ aug."); however, this does not occur in the case for `ensemble`.

6 Conclusion

This paper analyzed the use of CNNs in the context of washed-away detection. We constructed a new labeled dataset and conducted a comprehensive study that considered input scenarios (availability of pre-tsunami images), CNN configuration (`6-channel` or `siamese`), and input scales (`fixed-scale`, `resized`, their ensemble and `central-surround`). Overall, the proposed CNN-based washed-away detection system demonstrated 94–96% accuracy.

In this study, we used existing building location information to apply the CNN to buildings in aerial images. In future, we intend to investigate CNNs that can detect and classify object instances in an end-to-end manner, as reported in [15] and [16].

Acknowledgements: This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] A. J. Cooner, Y. Shao and J. B. Campbell: "Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake" in *Remote Sensing*, 2016.

[2] L. Gueguen and R. Hamid: "Large-scale damage detection using satellite imagery" in *CVPR*, 2015.

[3] E. Simo-serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer: "Discriminative learning of deep convolutional feature point descriptors" in *ICCV*, 2015.

[4] S. Zagoruyko and N. Komodakis: "Learning to compare image patches via convolutional neural networks" in *CVPR*, 2015.

[5] T.-Y. Lin, Y. Cui, S. Belongie and J. Hays: "Learning deep representations for ground-to-aerial geolocalization" in *CVPR*, 2015.

[6] D. G. Lowe: "Distinctive image features from scale-invariant keypoints", *IJCV*, 60(2):91–110, 2015.

[7] Ministry of Land, Infrastructure, Transport, and Tourism: "First report on an assessment of the damage caused by the Great East Japan earthquake", `http:\\www.mlit.go.jp\common\000162533.pdf` (published in Japanese)

[8] PASCO Corporation: `http://www.pasco.co.jp/eng/`

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton: "ImageNet classification with deep convolutional neural networks" in *NIPS*, 2012.

[10] K. Simonyan and A. Zisserman: "Very deep convolutional networks for large-scale image recognition" in *ICLR*, 2015.

[11] R. Ito, S. Iino, A. Fujita, T. Imaizumi and S. Hikosaka: "Assessing the land cover classification technique of the satellite imagery using deep learning" in *The 30th Annual Conference of the Japanese Society for Artificial Intelligence* (published in Japanese), 2016.

[12] V. Nair and G. E. Hinton: "Rectified linear units improve restricted boltzmann machines" in *ICML*, 2010.

[13] J. Snoek, H. Larochelle and R. P. Adams: "Practical bayesian optimization of machine learning algorithms" in *NIPS*, 2012.

[14] Y. Jia: "Caffe: An open source convolutional architecture for fast feature embedding", `http://caffe.berkeleyvision`, 2013.

[15] S. Ren, K. He, R. Girshick and J. Sun: "Faster R-CNN: Towards real-time object detection with region Proposal Networks" in *NIPS*, 2015.

[16] P. O. Pinheiro, T.-Y. Lin, R. Collobert and P. Dollár: "Learning to refine object segments" in *ECCV*, 2016.