Safety-level Estimation of Aerial Images based on Convolutional Neural Network for Emergency Landing of Unmanned Aerial Vehicle

Isana Funahashi* and Yo Umeki† and Taichi Yoshida‡ and Masahiro Iwahashi*
* Dept. Electrical Electronics Information Engineering, Nagaoka Univ. of Tech., Nagaoka, Niigata, 940-2137 Japan
E-mail: funahashi@stn.nagaokaut.ac.jp, iwahashi@vos.nagaokaut.ac.jp
† Dept. Information Science Control Engineering, Nagaoka Univ. of Tech., Nagaoka, Niigata, 940-2137 Japan
E-mail: umeki@stn.nagaokaut.ac.jp
‡ Dept. Communication Engineering Informatics, Univ. of Electro-Commun., Chofu, Tokyo, 182-8585 Japan
E-mail: t-yoshida@uec.ac.jp

***Abstract*—We propose an estimation method for the safety-level of local regions in aerial images for the emergency landing of Unmanned Aerial Vehicles (UAVs) based on Convolutional Neural Networks (CNNs), and introduce a new definition of safe areas and a new dataset. The estimation methods calculate scores of the safety-level for each region, and based on the results, the landing system detects safe areas where UAVs land without injuring humans, animals, buildings, artifacts, and themselves. Previous methods generally define natural flat regions, such as grass, lawn, soil and sand areas, as safe. However, if the flat regions are small and adjoin undesirable objects, the definition is dangerous and has the possibility of the injuring. Therefore, we introduce new definition to avoid the above complicated regions, and produce the dataset. Based on the dataset, we propose a CNN model to estimate scores of the safety-level. The proposed model can use various local and global features, and consider the environment of a target region. Hence, the proposed method estimates safe regions without the complicated ones, and then has better scores in the precision than the state-of-the-art method in experiments.**

I. Introduction

Unmanned Aerial Vehicles (UAVs) should be equipped with automatic landing for emergency, and the landing system frequently uses a technique estimating the safety-level of landing areas from aerial images [1]–[4]. The technique considers ground surfaces, humans, animals, buildings, and artifacts. Based on results of the technique, the landing system determines the area for the safety landing and UAVs move to there. UAVs should avoid injuring people and artifacts, and therefore the estimation methods require high precision.

Estimation methods of the safety-level are mainly classified into surface classification based methods and direct estimation methods. The surface classification based method is a technique which estimate types of ground surfaces using classification algorithms [1]–[3]. In the type, 'Safe' is defined as flat regions such as grass, soil, sand regions, and so on. On the other hand, direct methods estimate the pixel-wise safety-level by using the machine learning [4]. Since classification methods cannot judge unknown classes and are not practical. Hence, the direct estimation method has been proposed recently.

Unfortunately, the direct estimation method often produces the inaccurate safety-level in complicated regions. Complicated regions have the same visual feature of safe regions. For example, if a grass region adjoins an object such as human and buildings, conventional methods make the misdetection and cause an accident by the UAV landing. Although flat regions are defined safe areas in conventional methods, they are not always safety. Hence, the safety-level estimation methods need to consider the complicated regions for the accurate safety-level.

In Convolutional Neural Networks (CNNs), we presume that the receptive field is an important factor to estimate the accurate safety-level. The receptive field is the local region of input images that is used for one CNN outputs. The size of the receptive field is related to the size of objects that the CNN models can consider. In complicated regions, there are various objects with different sizes. Hence, CNNs should have multiple sizes of receptive fields for considering the object in complicated regions.

We propose a safety-level estimation method based on CNNs and introduce a dataset for UAVs landing systems. First, we propose a CNN architecture which considers surround features of a target area. The proposed CNN model considers both small objects and large regions by using residual blocks and dilated convolution layers [5], [6]. These architectures realize a CNN model with multiple sizes of receptive fields. A post-processing detects a safe landing area which has an enough size for UAV landing. Second, we propose a dataset which includes complicated regions and their safe-level. The dataset includes the different safety-level on same types of surface. By the learning with the dataset, the proposed CNN model can classify the complicated regions. In experiment, compared with the state-of-the-art method, the proposed method shows superior results in the precision.

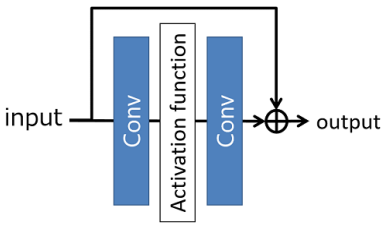


Fig. 1. An architecture of Residual block

II. Fundamentals of CNNs

A. Basic constructions of CNNs

Since the safety-level estimation is similar to image classification, we explain CNN architectures for classification algorithms in this section [5]–[7]. CNN models extract image features using pooling layers and convolution (Conv) layers. The models are construct by stacking those layers. An Conv layer has hyperparameters which are filter size, stride size, and zero-padding size. Zero-padding is the layer preprocess which adds the pixel with the value of zero around the image border. The padding is applied before the convolution operation. In Conv layer, rectified activation functions, such as Rectified Linear Unit (ReLU), Leaky Rectified Linear Unit (Leaky ReLU), and Parametric Rectified Linear Unit (PReLU), are used after Conv [8]–[10]. Then, the models apply the softmax operation to resultant values as

$$\hat{p}_{ik} = \frac{exp(\hat{x}_{ik})}{\sum_l exp(\hat{x}_{ij})}, \tag{1}$$

where l, $\hat{x}_{ik}$, and $\hat{p}_{ik}$ denote the number of pixels, resultant values of CNN models, and the class probability in the k-th class of the i-th pixel, respectively. Finally, pixels are classified according to p_{ik}. Through defining classes of the safety-level, we easily adjust the models to the safety-level estimation.

B. Residual block

Residual blocks are known as effective architectures for constructing deep CNN models [5], shown in Fig. 1. The structure of residual blocks consists of two Conv layer, an activation function, and a shortcut connection. The shortcut connection performs the element-wise summation of the input and the output. Several methods stack the block to construct deep CNN models [5], [11]. Hence, the model becomes ensemble of several CNNs with different sizes of receptive fields [12]. The shortcut connection well propagates errors between results and ground truths to previous blocks and usually avoids the gradient vanishing problem in the training of deep CNN models [13].

C. Dilated Conv layer

A dilated Conv layer performs the convolution with sparse filters which are constructed via up-sampling filters with arbitrary parameters [6]. The parameter indicates the size of space between filter elements, which called a dilation in this paper. The receptive field of the dilated Conv layer is wide without losing the spatial resolution of image features. Hence, the receptive field of a CNN model using dilated Conv layers is larger than one using Conv layers, and the CNN model can also extract features of local regions.

(a) Input image (b) Ground truth

Fig. 2. Examples of dataset images for training

III. Proposed method

A. Frame work

In this paper, we introduce a definition of the safety-level and a dataset for the UAV landing system, and propose a method of the safety-level estimation based on CNNs. Even if regions of the ground are classified into a same type of surfaces, they often have different labels of the safety-level because of surrounding environments. Hence, the proposed method considers to surrounds of landing areas using a new dataset. The proposed CNN model learned with the introduced dataset estimates accurate labels of the safety-level at complicated regions. The model estimates a map of pixel-wise safety-level which are classified into three classes, 'Other', 'Not recommend', and 'Safe'. We apply the morphological opening for 'Safe' areas as the post-processing, and a large area that is sufficient for the UAV landing is detected.

B. Dataset construction

For the introduced dataset, we randomly collected aerial images from Google Maps in a region bounded by (35.1°N, 138.5°W) and (36.4°N, 139.7°W). The dataset has 125 images and 50 images for training and testing, respectively. We assumed the flight altitude 140 m, and the size of aerial images is 576×576 px with approximately 8.2 px/m because of the regulation of Google Maps. We divided these image into patches with 16×16 px. UAVs are assumed to land at the center of patches, and the size of patches is approximately 2×2 m that is enough for safe landing.

We classified patches into three classes, 'Safe', 'Not recommended', and 'Other'. 'Safe' patches guarantee that UAVs can land there without damaging not only people and artifacts but also themselves. Conversely, at 'Not recommended' patches,

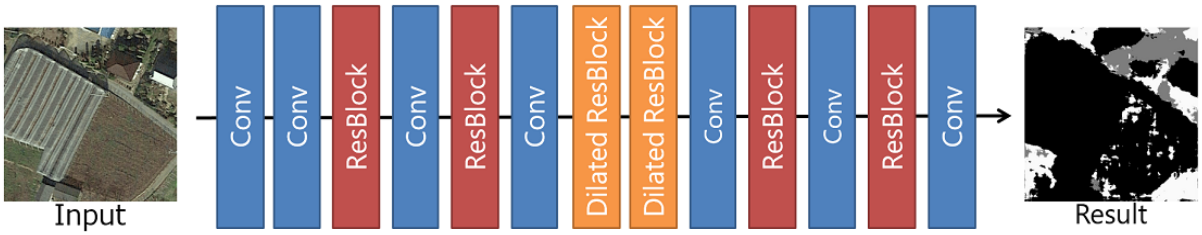


Fig. 3. Overview of proposed CNN model

TABLE I
PARAMETERS OF PROPOSED CNN MODEL

Layer type		Filter	Padding	Dilation	Ch.
Conv		3x3	1		32
ResBlock	Conv	3x3	1		32
	Conv	3x3	1		32
Conv		3x3	1		64
ResBlock	Conv	3x3	1		64
	Conv	3x3	1		64
Conv		3x3	1		64
Dilated ResBlock	DilatedConv	3x3	2	2	64
	DilatedConv	3x3	4	4	64
Dilated ResBlock	DilatedConv	3x3	4	4	64
	DilatedConv	3x3	8	8	64
Conv		3x3	1		32
ResBlock	Conv	3x3	1		32
	Conv	3x3	1		32
Conv		3x3	1		3
ResBlock	Conv	3x3	1		3
	Conv	3x3	1		3
Conv		3x3	1		3

UAVs land only without damaging people and artifacts. Residual patches are classified into 'Other', and are often danger areas.

Fig. 2 shows examples of the training set which are the input images and the ground truth. In the ground truth, white, gray, and black areas show the labels which are 'safe', 'Not recommend', and 'Other', respectively. Based on the above definition, ground truths are provided by various persons with the majority rule. Images of the dataset have variety of objects which are grass, trees, soil areas, artifacts, and so on. Patches of the 'safe' class are grass, soil, and sand areas which are not include water, trees, and artifacts. Patches of the 'Not recommend' class are almost forest, river, and sloping area. 'Other' class are not only car, buildings, roads, and other artifacts but also areas which are close to them and have similar visual features of 'safe' or 'Not recommend' patches, for example grass, trees, and gravel roads.

C. CNN model

The architecture of the proposed CNN model is shown in Fig. 3, where 'ResBlock' and 'Dilated ResBlock' denote residual blocks that uses normal Conv and dilated Conv layers, respectively. The model parameter is shown in Table I, where Filter, Padding and Ch. denote filter size of Conv layers, size of zero-padding and the number of output channels for Conv layers. We use the Leaky ReLU as the activation function in Conv layers and Residual Blocks [9]. The stride size of all Conv layers are 1. Let h and w be the height and the width of an input RGB image, and then output signals of the model are $h \times w \times N$, where N denotes the variation of the safety-level.

To use features in both small and large regions, we adopt residual blocks and the dilated Conv layers. Small receptive fields only include a part of large objects, and large receptive fields include not only object regions but also background regions for small objects. Small and normal sizes of receptive fields are realized by shallow and normal networks as mentioned in Sec. II-C. Since residual blocks make the model a union of shallow and normal CNNs, Residual blocks realize several sizes of reception fields. The dilated Conv layer realizes large reception fields without losing the spatial resolution. The spatial resolution is necessary for extracting features of small objects. Hence, the model has several receptive fields and considers small object and large areas.

D. Training

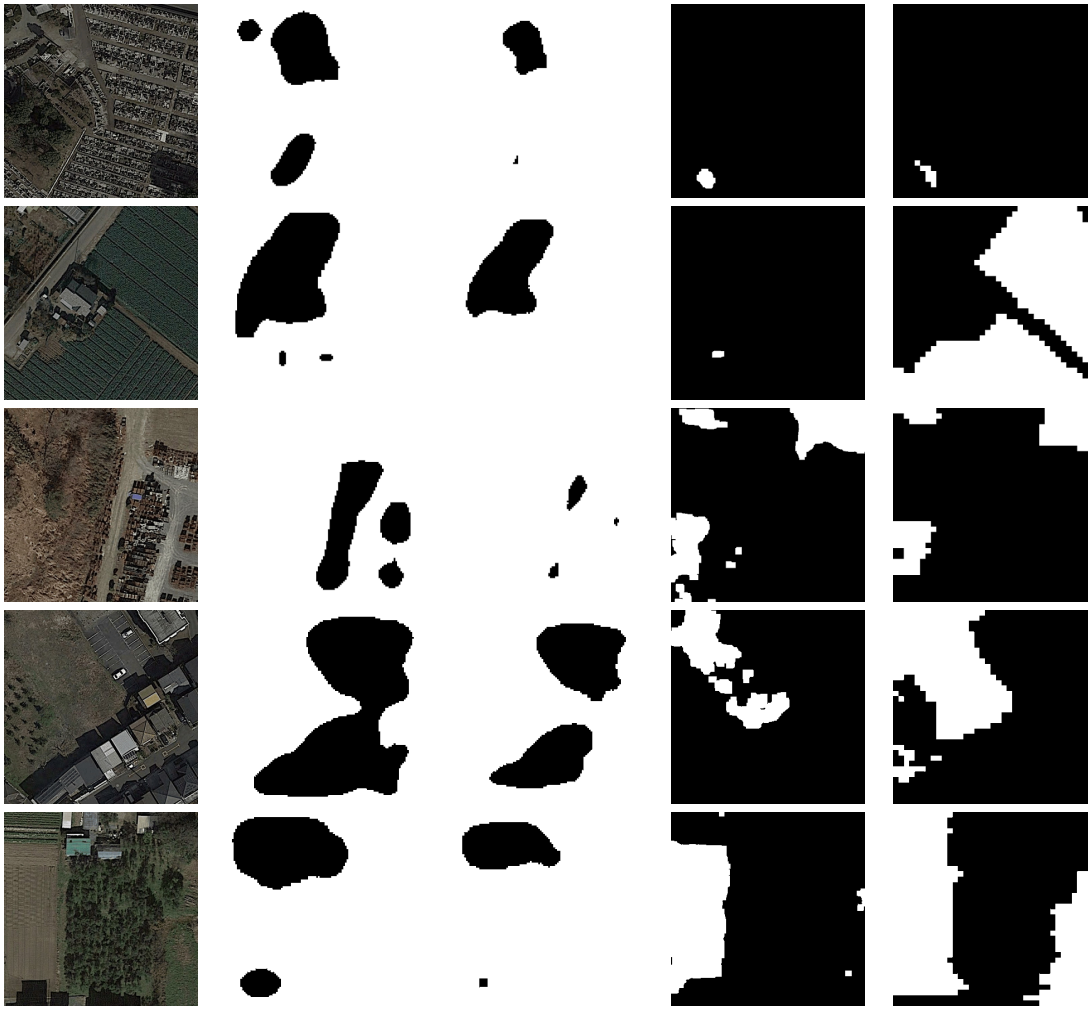
In the training of the proposed CNN model, we use the backpropagation algorithm and the mini-batch gradient descent for optimization [14]. We determine 6 images as the mini-batch size. We apply the batch normalization after all activation functions except the last layer of the proposed CNN model [15]. The model outputs the pixel-wise probability of each safety-level. Therefore, we use the pixel-wise cross entropy as the loss function, defined as

$$L = -\sum_{i=1}^{h \times w} \sum_{k=1}^{N} p_{ik} \ln \hat{p}_{ik}, \tag{2}$$

where p_{ik} is a one-hot vector of the ground truth class for the k-th element at i-th pixel and $\hat{p}_{ik}$ is defined in (1). For the loss function, we use the ADADELTA algorithm which automatically determines a learning rate [16].

E. Post processing

We apply the morphological opening to dump the small region which does not have the enough size for landing [17]. First, the proposed method makes a binary image from estimated safe areas. Then, we apply the morphological opening to the binary image. The filter size of the morphological opening is 16×16 px which is decided from the ground resolution of aerial images and the UAV sizes.



(a) Input (b) Comp. A (c) Comp. B (d) Prop. (e) GT

Fig. 4. Resultant safe area of test images

IV. Experiment

For experiments, we compare the proposed method with the state-of-the-art method of the safety-level estimation for the safe landing of UAVs [4], and use the proposed dataset as train and test sets. We applied several data augmentations to the train set, random clipping, rotation with (90, 180, 270), and Left-right flipping. Consequently, the size of train images is 480×480 px, and the number of their pairs is 2400. For the compared method, we divided train images according to its definition and trained it with parameters shown in the paper [4]. The proposed CNN model was trained with 12000 iterations.

We use the precision scores for quantitative evaluation [17]. In the case of the area detection for the safe landing of UAVs, the precision is important. The misdetection of safe areas induces dangerous accidents. UAVs requires only one area for landing. Therefore, the detection system for the safe landing requires a high precision score, and the recall and the F-measure are unsuitable for measuring.

We compare the proposed method with the state-of-the-art method by detection result of landing area [4]. The compared method estimates the map of safety-level which is in the range of $[0, 255]$, and detects landing areas using simple thresholding [4]. Thresholding parameters of the conventional method are 100 and 200 which are described in the paper [4]. The proposed method estimates three level of area which are 'safe', 'Not recommend' and 'Other'. Hence, we set the areas which estimated the 'safe' and 'Not recommend' class as detection results of the proposed method.

Table II, Table III show precision scores of only 'Safe' and both 'Safe' and 'Not recommended', and resultant images. Fig. 4 shows resultant images which are selected from results of 50 test images. In these tables and the figure, 'Comp. A', 'Comp. B', 'Prop.', 'GT', 'Average' denote the compared method with

TABLE II
PRECISION SCORES OF SAFE AREA

	Comp. A	Comp. B	Prop.
Image1	0.007	0.006	**0.556**
Image2	0.704	0.647	**1.000**
Image3	0.127	0.110	**0.645**
Image4	0.605	0.559	**0.997**
Image5	0.351	0.362	**0.993**
Average of 50 images	0.308	0.291	**0.638**

TABLE III
PRECISION SCORES OF SAFE AND NOT RECOMMENDED AREA

	Comp. A	Comp. B	Prop.
Image1	0.156	0.109	**0.740**
Image2	0.654	0.590	**0.816**
Image3	0.573	0.542	**0.904**
Image4	0.397	0.386	**0.983**
Image5	0.653	0.829	**0.979**
Average of 50 images	0.578	0.504	**0.871**

the thresholding parameter of 100 and 200, the proposed method, the ground truth, and average values of 50 images, respectively. These tables shows that the proposed method always outperforms the compared method. Fig. 4 shows that the proposed method often detects smaller areas than the ground truth, but its accuracy is usually high. The proposed method considers surroundings of the building and small boxes on the ground in comparison with conventional method. These results indicate that the proposed method reduces misdetection of the safety-level by considering surrounding of landing areas and undesirable objects.

V. CONCLUSION

In this paper, we introduce a definition of the safety-level and a dataset for the UAV landing system, and propose a method of the safety-level estimation based on CNNs. The proposed CNN model considers small and large region using Residual Blocks and Dilated convolution layers. Our dataset has images which have complicated regions. Thanks to the strategy, the proposed method provides better results than the state-of-the-art one.

REFERENCES

[1] L. Mejias, D. Fitzgerald, P. Eng, and X. Liu, "Forced landing technologies for unmanned aerial vehicles: towards safer operations," in *Aerial vehicles*. InTech, 2009.

[2] L. Mejias, "Classifying natural aerial scenery for autonomous aircraft emergency landing," in *Proc. Int. Conf. Unmanned Aircraft Syst.*, 2014.

[3] T. Patterson, S. McClean, P. Morrow, G. Parr, and C. Luo, "Timely autonomous identification of UAV safe landing zones," *Image and Vision Computing*, vol. 32, no. 9, pp. 568–578, 2014.

[4] X. Guo, S. Denman, C. Fookes, and S. Sridharan, "A robust uav landing site detection system using mid-level discriminative patches," in *Proc. Int. Conf. Patt. Recognit.*, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, 2016.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pat. Recognit.*, 2015.

[8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010.

[9] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[11] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit. Workshops*, 2017.

[12] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[13] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015.

[16] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv:1212.5701*, 2012.

[17] G. Rafael, C. and W. Richard, E., *Digital IMage Processing(3rd Edition)*. Prentice Hall, 2007.