A METHOD OF BUILDING DETECTION IN REMOTE SENSING IMAGES BASED ON DEEP LEARNING WITH MULTIPLE LIGHTNESS DETECTORS

Dongming Huang[1,2], Hongrui Zhao[1,2,*], Yi Yao[1,2]

[1] Department of Civil Engineering, Tsinghua University, Beijing 100084, China - hdm17@mails.tsinghua.edu.cn, zhr@tsinghua.edu.cn, yaoyi18@mails.tsinghua.edu.cn

[2] 3S Center, Tsinghua University, Beijing 100084, China

Commission IV, ICWG IV/III, WG IV/4

KEY WORDS: Remote sensing, Building detection, Deep learning, Multiple lightness detectors

ABSTRACT:

Buildings, where most human activities happen, are one of the most important crucial objects in remote sensing images. Extracting building information is of great significance importance for conducting sustainable development-related researches. The extracted building information is a fundamental data source for further researches, including evaluating the living conditions of people, monitoring building conditions, predicting disaster risks and so on. In recent years, convolutional neural networks have been widely employed in building detection, and have gained significant progresses. However, in these automatic detection procedures, the critical brightness information is often neglected, with all buildings simply classified into the same category. To make the building detection more efficient and precise, we propose a simple yet efficient multitask method employing several lightness detectors, each of which is dedicated to the building detection in a specific brightness interval. Experiment results show that the building detection accuracy could be improved by 8.1% with the assistance of the additional lightness information.

1. INTRODUCTION

Remote sensing is a fundamental technology that detects and analyses the natural resources and environment, and reveals the spatial distribution characteristics as well as the temporal and spatial evolutions of various elements on the Earth surface. Gong Peng (Gong, 2019) suggested that remote sensing could contribute to the 17 sustainable development goals of the 2030 Agenda. Among these goals, slum research, urban environment, and building facility monitoring, are all closely related to building extraction.

Automatic building detection from remote sensing images is one of the long-standing goals in remote sensing technologies. Remote sensing images contain basic characteristics or variations, such as shape, size, pattern, tone (or hue), texture, shadows, site, association, and spatial resolution, etc. Specifically, the tone (or hue) refers to the relative brightness or color of objects on an image(Lillesand et al., John Wiley & Sons, Inc, 2015). These features offer the basic paradigm for information extraction including building detection.

In recent years, the deep learning method has been widely used in building inspection due to its powerful feature extraction functionalities, and has achieved noticeable progresses. However, in previous automatic detection processes, the differences between buildings are artificially ignored, as a result, all buildings are simply classified into the same category. The huge information loss accompanying this indiscriminate building identification seriously hinders the automatic processing of building detection.

To overcome the shortcomings of the above brute-force building classification scheme, Hamaguchi (Hamaguchi et al., 2018) proposed a simple but effective multi-task model employing multiple detectors, each of which is dedicated to specific building size. This method actually utilizes the size information contained in the remote sensing image. Compared with the size, the brightness information is more obvious in different buildings, and can be straightforwardly calculated from the RGB value of the image. Therefore, building detection based on different lightness should be regarded as different classification tasks. In this work, we propose a multi-task model based on a combination of multi-lightness detectors, each of which is concentrated to building detection in a specific lightness range in remote sensing images.

2. METHODS

2.1 Overview

As illustrated in Figure 1, the proposed model is based on the U-Net model(Ronneberger et al., 2015), which consists of a shared feature extractor and three task branches (C_l, C_m, C_d) with identical structures. In the encoding stage, multiple detectors share the same feature extractor; while in the decoding stage, different detectors are responsible for different tasks. The model takes RGB images and its lightness information as inputs, and outputs three kinds of extraction results, corresponding to buildings with three lightness levels: light, medium, and dark. The final building extraction results are synthesized from these three results.

2.2 Lightness and its calculation

The distinction between lightness and brightness is usually vague but not completely untraceable. Actually, brightness corresponds to HSV/HSB color model, while lightness corresponds to the HSL color model. According to the existing literature (Gilchrist, 2007), lightness is the perceptual dimension that runs from black to white. The physical counterpart of lightness is the intrinsic property of a surface that determines what percentage of light it reflects. In short, lightness is perceived as reflectance. ON the

* Corresponding author

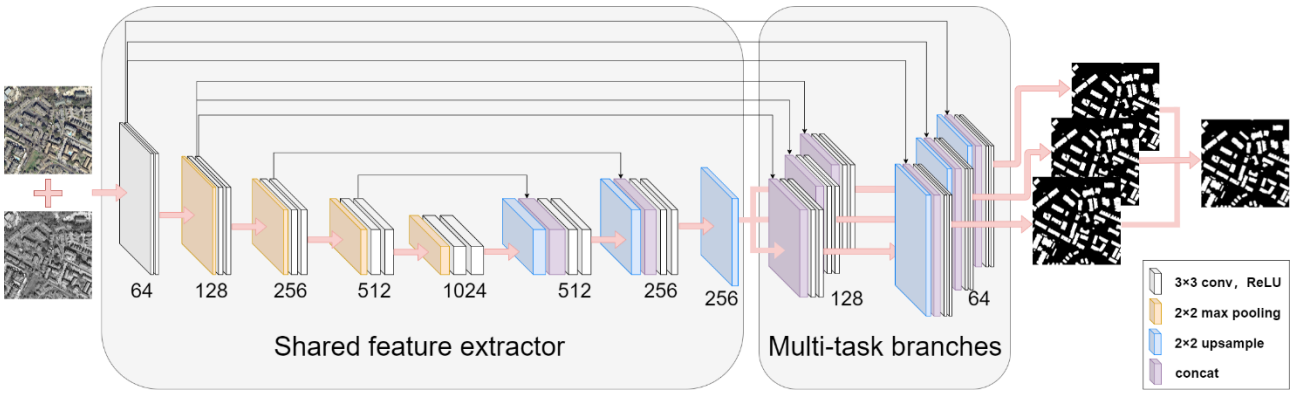


Figure 1. Overview of the proposed building detection method. The model is based on U-Net, and extends the last two decoder layers into multiple lightness detectors. It takes the RGB image and the lightness image as input, and outputs the "high", "medium", and "dark" building detection results respectively. The final extraction result is merged from the three branches.

other hand, brightness is the perceptual dimension that runs from dim to bright.. Lightness concerns the objective side of visual experience while brightness concerns the subjective side.

In general, remote sensing observes the reflectivity of a ground object in a particular band, which is similar to the physical meaning of lightness. Therefore, we choose the L value of the HSL color model as the lightness.

For any pixel of an RGB image, r, g, b corresponds to the values of three color channels, the lightness value l of the pixel is calculated as following.

$$l = \frac{1}{2}(max + min) \tag{1}$$

where $max = \max(r, g, b)$
$min = \min(r, g, b)$

Figure 2 shows the RGB image and its corresponding lightness image.

2.3 Lightness detectors

The proposed model has three lightness detectors (C_l, C_m, C_d), each of which is responsible for detecting "light", "medium", and "dark" buildings. For input $x \in \mathbf{X}$, the outputs of the detectors can be written as

$$p^k = C_k(F(x)), k = \{l, m, d\} \ . \tag{2}$$

To train the model, the multi-class labels $y_i = \{c_n, c_l, c_m, c_d\}$ which corresponding to the ground truth of "non-building", "light", "medium", and "dark" buildings are used.

The losses L_k of C_k is defined as follows.

$$H = -\frac{1}{n}\sum_{i=1}^{n}[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \tag{3}$$

$$J(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot \hat{y}_i}{y_i + y_i - \hat{y}_i \cdot \hat{y}_i} \tag{4}$$

$$L_k = H - \log(J), k = \{l, m, d\} \tag{5}$$

where n is the number of images in a batch
y is the ground truth
$\hat{y}$ is the prediction

The loss function of the model is defined as the sum of the losses from each detector.

$$L = L_l + L_m + L_d \tag{6}$$

where L_l, L_m, L_d is losses of each detector

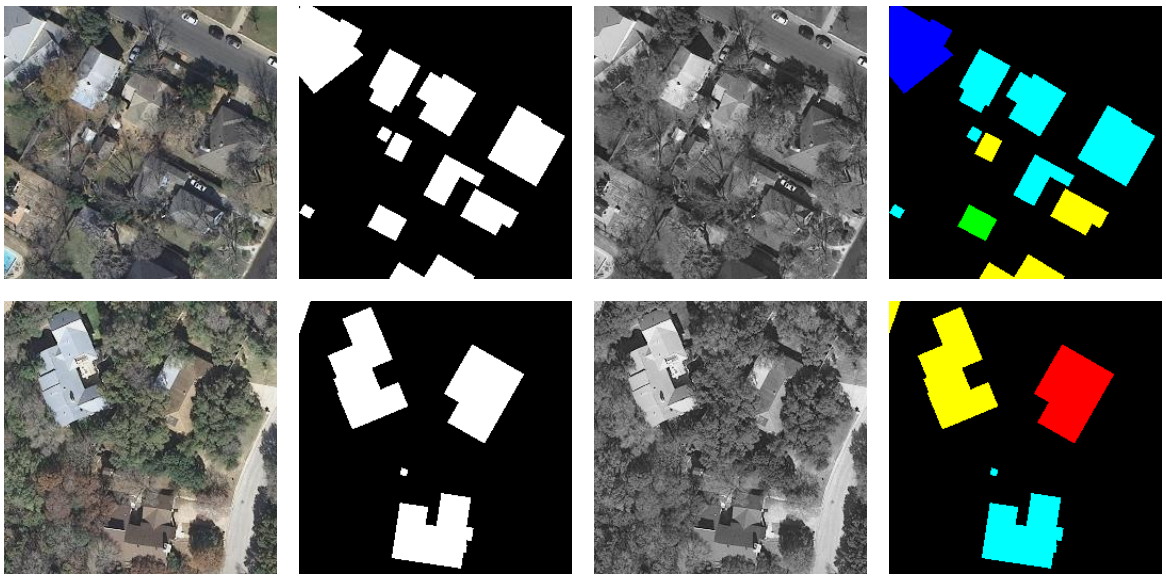
3. EXPERIMENTS AND RESULTS

3.1 Dataset

We used the building annotation information provided by Inria Aerial Image Labeling Dataset (Maggiori et al., 2017). The dataset consists of 360 high-resolution aerial images covering 9 different cities, but only 180 tiles in training set are provided with ground truths, and cover 5 cities, each of which has 36 tiles. The regions cover dissimilar urban settlements, ranging from densely populated areas to alpine towns. The size of each image is 5000 × 5000 pixels with the spatial resolution of 0.3 meters per pixel, and is composed of red, green and blue (RGB) channels. Only two semantic classes (non-buildings and buildings) were considered as the ground truth.

In order to train the proposed model, we need to change the annotation information to 4 classes: "non-building", "light", "medium", and "dark" buildings. Buildings are classified into three classes based on their lightness. We assume that the roof of a single building uses only one material, and the reflectance at each point is the same, with little difference in the lightness map. However, in the actual image, there is a certain difference in the brightness of each pixel of the building. To this end, we have considered three factors as following when making labels:

1. All pixels of a building should be divided into the same class to preserve their shape characteristics. Therefore, in order to avoid the different pixels of the same building being divided into different classes, the average brightness is used to replace the brightness of each pixel itself.



(a) Image (b) old label (c) lightness image (d) new label

Figure 2. Image and label examples. The first column is the remote sensing image, the second column is the old label, the third column is the generated lightness image, and the fourth column is the adjusted label. Note that there are more than three colors in the new label image because some buildings belong to more than one class. The new label is divided according to the average lightness of the building, where blue indicates "light", yellow indicates "light" and "medium", green indicates "medium", cyan indicates "medium" and "dark", and yellow indicates "dark". There are subtle differences between the class based on the average lightness and the brightness we perceive.

2. The balance between classes should be taken into account when setting thresholds.

3. There is actually a certain range of lightness at different points of the building, so a certain degree of overlap is maintained between the segmentation thresholds, which means that some buildings may be marked as one or more of "light", "medium", and "dark".

In the end, buildings with an average lightness greater than 150 represent for "light", between 80 and 180 represent for "medium", and less than 110 are defined as "dark".

3.2 Experimental details

We calculated the lightness images of the 180 scene images in the training set according to Equation (1). Following previous researches (Maggiori et al., 2017), we choose the number of 1-5 images of each city from the training set for the test. The model was strictly separated from the test data before the final test. We also choose the number of 6-7 images of each city for validation, and the remaining 140 tiles are used as training data. The labels of the training data and verification data are reset according to the above method. All the training images, validation images, and their new labels are divided into 400 small pictures of 256×256 pixels.

In the experiment, we used the same images and labels data to try three different training data: lightness images, RGB images, lightness images + RGB images. We implement our method based on Tensorflow and Keras. Adam was used for optimization with an initial learning rate of 0.001. Each training combination was trained 20 epochs, and the batch size is 4.

The predicted results of all models were not post-processed.

3.3 Results

To evaluate the quantitative performance of different training data, the overall $accuracy$, $precision$, $recall$, $F1-score$ and mean intersection over union ($mean-IoU$) are used as quality metrics. The five metrics are calculated as following.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \tag{7}$$

$$mean-IoU = \frac{tp}{tp+fp+fn} \tag{8}$$

$$precision = \frac{tp}{tp+fp} \tag{9}$$

$$recall = \frac{tp}{tp+fn} \tag{10}$$

$$F1-score = \frac{2\times precision\times recall}{precision+recall} \tag{11}$$

where tp is the number of true positives
tn is the number of true negatives
fp is the number of false positives
fn is the number of false negatives

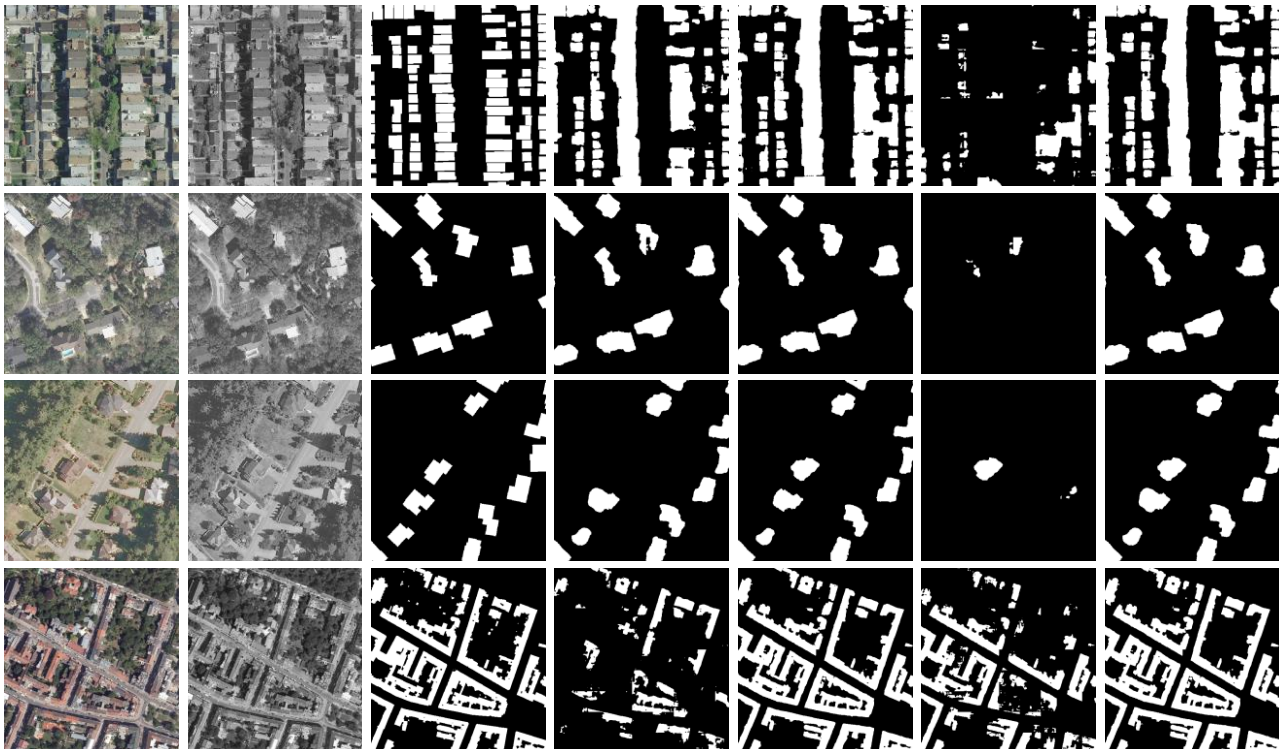


Figure 3. Experiment results of the model trained by lightness images + RGB images. The four images are from Chicago, Austin, Kitsap, and Vienna, covering different building density areas. The output of each detector shows that different detectors work as expected for detecting buildings with different lightness sections.

Table 1 shows the building detection results for different sets of training data on the selected 25 test images. Using only the lightness images as the training data, the highest *precision* is achieved, and the *accuracy* is comparable to other training data combinations. This means that lightness is important information in distinguishing building and non-building.

Training data	*accuracy*	*mean-IoU*	*precision*	*recall*	*F1-score*
lightness images	0.9427	0.7789	**0.8768**	0.6804	0.7662
RGB images	0.9557	0.8335	0.8588	0.8126	0.8350
lightness images + RGB images	**0.9579**	**0.8429**	0.8536	**0.8389**	**0.8461**

Table 1. building detection results for different training data. The highest values for the different metrics are highlighted in bold.

The lightness is calculated from RGB value according to Equation 1, it is strictly redundant data, but the experiment proves that all the metrics are improved to a certain extent, especially *recall* and *mean-IoU* metrics. This aspect proves the validity of our model. On the other hand, it is also shown that although deep learning has been proven to be able to extract features automatically, manually designing features such as lightness information is still an effective means to improve the metrics of automatic interpretation of remote sensing.

Figure 3 is an example of the typical results predicted by the model trained with lightness images + RGB images. The four images show the detector's ability to detect buildings in different building density areas. The results show that although there is a large overlap between the results of the different detector outputs, they can focus on the buildings within the target lightness range and form a good complement, especially in the third and fourth groups in the figure. The improvement after the combination of the bright detector and the dark detector of the image is obvious.

Method	*accuracy*	*mean-IoU*	*precision*	*recall*	*F1-score*
SegNet	0.865	0.737	0.867	0.767	0.849
FCN	0.889	0.773	**0.918**	0.831	0.872
U-Net	0.877	0.755	0.885	0.837	0.860
Tiramisu	0.869	0.743	0.911	0.801	0.853
FRRN	0.875	0.752	0.915	0.808	0.858
USPP	0.909	0.806	0.903	**0.882**	**0.893**
Our method	**0.958**	**0.843**	0.854	0.839	0.846

Table 2. Numerical Results of different methods on the INRIA testing dataset. In addition to our method, the remaining results are from (Liu et al., 2019). All models used the same test data. The highest values for the different metrics are highlighted in bold.

We further conducted a quantitative comparison with different models on the Inria Aerial Image Labeling Dataset. The literature(Liu et al., 2019) quantifies the performance of different methods including SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP on the dataset, using consistent evaluation metrics and the same test data with our method. The results are summarized in Table 2. Our method achieves the best results in metrics of *accuracy* and *mean-IoU*. Comparing to the U-Net model, the proposed method yields a higher accuracy by 8.1% (0.958 vs. 0.877) and a higher mean-IoU by 8.8% (0.843 vs. 0.755).

4. CONCLUSION

We proposed an automatic building detection method employing lightness as additional information channel. This method adopts a structure that uses three same shape detectors with different weights to detect buildings in different lightness ranges. By mining underlying lightness information, our method could improve the precision of building extraction without sacrificing accuracy. The detection improvement shows that even in the era of deep learning, extracting interpretation features from raw data as much as possible is still very important for automatic information extraction of remote sensing images.

ACKNOWLEDGEMENTS

This work was supported by the Fund of National Natural Science Foundation of China [Grant number 41571414].

REFERENCES

Gilchrist, A. L. 2007: Lightness and brightness. *Current Biology, 17*(8), R267-R269.

Gong, P. 2019: Towards more extensive and deeper application of remote sensing. *Journal of Remote Sensing, 23*(4), 567-569.

Hamaguchi, R., Hikosaka, S., Ieee. 2018. *Building Detection from Satellite Imagery using Ensemble of Size-specific Detectors.* Paper presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT (18-22 Jun 2018).

Lillesand, T., Kiefer, R. W., Chipman, J. 2015: *Remote Sensing and Image Interpretation, 7th Edition*: John Wiley & Sons, Inc.

Liu, Y., Gross, L., Li, Z., Li, X., Fan, X., Qi, W. 2019: Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder With Spatial Pyramid Pooling. *IEEE Access, 7*, 128774-128786. doi:10.1109/access.2019.2940527

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., Ieee. 2017: CAN SEMANTIC LABELING METHODS GENERALIZE TO ANY CITY? THE INRIA AERIAL IMAGE LABELING BENCHMARK. *2017 Ieee International Geoscience and Remote Sensing Symposium (Igarss)*, 3226-3229.

Ronneberger, O., Fischer, P., Brox, T. 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* Paper presented at the International Conference on Medical Image Computing & Computer-assisted Intervention (2015).