Building extraction from satellite imagery using a digital surface model

A V Dunaeva[1,2] and F A Kornilov[1]

[1]N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, S. Kovalevskaya Street 16, Yekaterinburg, Russia, 620990
[2]Ural Federal University named after the first President of Russia B.N. Yeltsin, Mira Street 19, Yekaterinburg, Russia, 620002

Abstract. In this paper, two approaches to building extraction from satellite imagery and height data obtained from stereo images or LIDAR are compared. The first approach consists of detecting high-rise objects in a digital surface model and then improving recognition accuracy using segmentation of spectral information. The second approach uses the U-Net convolutional neural network, which showed the best results for the extraction of objects from aerospace images on a number of large datasets. Extensive experiments were carried out to evaluate the dependence of the quality of U-Net-based building extraction on the different data types (including high-resolution satellite images and digital surface model data). Building extraction quality of the trained network was also evaluated on satellite images with different spatial resolutions.

1. Introduction

At the present time, monitoring the state of the Earth's surface based on aerospace images plays an important role in many fields such as agriculture, construction, emergency analysis and others. This approach has a number of undeniable advantages: timeliness, large coverage of areas and a relatively low cost. At the same time, many tasks are handled manually by an operator, which makes it important to develop tools for automatically retrieving required information from satellite images.

The extraction and classification of terrain objects are the most important tasks in processing satellite imagery. The solutions of such problems find their application in cartography (update of topographic maps), the detection of changes in the composition of the terrain objects (i.e. analysis of urban development and illegal deforestation), navigation of aircrafts and others. However, the methods existing at the moment extract and classify objects of a terrain from space images with insufficient accuracy. First of all, this is due to the complexity and variability of the scenes under consideration, which contain a huge number of objects of different nature significantly influenced by shooting conditions. All this contributes to the importance of research aimed at overcoming these difficulties and developing software systems for processing remote sensing data.

In this paper, in order to improve the quality of the building extraction, a digital surface model (DSM) containing heights of a terrain and objects on it was applied. Height information for the DSM can be obtained by shooting a terrain from two different angles. Two such images form a stereo pair, which is passed to the input of a stereo matching algorithm, for calculating a matrix of pixel shifts (a disparity map) belonging to the same objects between two images. After that, the digital surface model of the terrain is constructed by calculating the heights of the terrain and objects on the basis of

the disparity map and the parameters of shooting. In this paper, we used a digital surface model obtained as a result of the operation of the stereo matching algorithm proposed in [1]. Since the main interest was the urban development, panchromatic images with a resolution of 0.5 m were used for stereo matching, and the resolution of the final model was 1 m per pixel. Among the works devoted to the generation of DSMs, it is also worth mentioning work [2] which describes a fast algorithm for GPUs.

The paper is organized as follows. The second section describes an approach to building extraction, based on the detection of high-rise objects of a terrain from a DSM and the refinement of their boundaries on the basis of segmentation of spectral information. The third section is devoted to the U-Net [3] convolutional neural network and details of its training. The fourth section presents the results of the comparison of considered approaches using real data.

2. Image segmentation and detection of high-rise objects

As a starting point for building extraction, we consider a segmentation algorithm, which uses heights of objects in addition to the spectral information to provide a better segmentation of satellite images.

2.1. Satellite image segmentation

Segmentation is used to pre-process images in many computer vision tasks, since it allows analyzing homogeneous (by certain criteria) regions instead of separate pixels of an image. The characteristics used to split an image can be very different: the color, the texture, etc. To date, a large number of segmentation algorithms have been developed. In [4, 5], a description and comparison of some of them is given to address the problem of building extraction. The considered algorithms demonstrate high quality segmentation, but do not use height information that can permit a better detection of object boundaries in satellite images.

To test the influence of height information on the result of satellite image segmentation, an algorithm based on the idea of grouping pixels in regions on the basis of the similarity of their brightnesses and heights was presented in [5]. The algorithm receives a four-channel image and a digital surface model. The following sequence of steps is performed.

1. Median filtering of the image.

2. Preliminary selection of a region. For each pixel in the image we form a region by recursive search for neighboring pixels of similar height and color. After that, we calculate mode values of the height and each of the four image channels in the obtained region and find in it a pixel having a height exactly equal to the calculated mode of height and the closest brightness to the modes of each channel.

3. Selection of the region. Next, from the found pixel, we form a new region by repeating recursive search, and fill the new region with the color of the computed mode values.

4. Repeat steps 2-3. Due to the noisiness of the input data, the regions obtained in the previous steps will have a small area. The repetition of the segmentation procedure will allow merging the regions into larger ones.

5. Small region removal. As a result of the brightness or height discontinuity, regions with a negligible area can remain. Such regions are combined with neighboring ones in such a way that the region to which the pixels are added has a larger area, and its height and brightness are as close to the added pixels as possible. This procedure is performed for regions in order of increasing area.

2.2. Building extraction from a DSM

After segmenting the satellite image (using the algorithm mentioned above or another one), it is required to determine whether the obtained segments belong to buildings.

To do this, it is necessary to find objects in the DSM that rise above the terrain. Such objects include buildings and forest tracts. For this purpose, the following algorithm is proposed.

1. *Search for height differences*. For each pixel p in the DSM with the height value h, we consider its one-dimensional neighborhoods $O_d^{\rightarrow}(p)$ and $O_d^{\downarrow}(p)$ of the radius d in horizontal and vertical directions. For the pixels $p_i \in O_d(p)$, we calculate $\Delta h_i = h_i - h_{i+1}, i \in [-d, d-1]$, where the index 0 corresponds to the pixel p, which is the center of the neighborhood. And if $\Delta h_0 = \max_i \Delta h_i$ or

$\Delta h_0 = \min_i \Delta h_i$, and $|\Delta h_0| \geq T$, then the pixel p is marked as having a significant height difference. The threshold T specifies the minimum height of buildings to be detected. We do not consider the pixels at which the minimum and maximum are simultaneously reached, since they are the errors in the DSM.

2. *Obtaining high regions*. From the pixels that have a significant height difference, we construct line segments, which will form the target regions. For the construction of these line segments, we search horizontally and vertically in forward and backward directions in the DSM for pixels with a height difference upward. Then we begin to draw a line segment from each selected pixel that goes:

- to the pixel with the height difference upward, the value of which is added to the value of the difference at the starting pixel. Then the construction of the line segment continues;
- to the pixel at which the height value is less or close to the height value at the beginning of the line segment;
- until the line segment reaches the fixed length.

Performing this procedure in the forward and backward directions (from left to right and vice versa, from top to bottom and vice versa) allows us to find even those regions of buildings in which one of the sides can be blurred due to the influence of noise.

Next, the pixels of the segments selected only horizontally or vertically are discarded. The remaining pixels of the segments form regions. We select only those of them that contain at least one pixel with a significant height difference. This operation allows rejecting the false intersections of the vertical and horizontal segments.

3. *Refinement of the form of the found regions*. We consider simply connected regions in the DSM, all pixels of which have the same height. If more than half of the pixels of the region were marked as high in the previous step, the remaining pixels of the region are also marked as high. Thus, the boundaries of the detected regions are smoothed and false alarms associated with slopes of hills and other sharp differences of the terrain height are rejected. The regions with small area are removed from consideration.

4. *Extraction of buildings among the found high regions*. Not only buildings and structures, but also high vegetation (trees) is related to the detected high-rise objects. The normalized difference vegetation index (NDVI) [6], which is calculated from the red and infrared channels of a satellite image, effectively distinguishes vegetation from other objects. However, in addition to vegetation, this index also extracts buildings, which will lead to skipping objects when the building extraction algorithm is running. To solve this problem, it is necessary to perform additional color filtering of the vegetation extracted using the NDVI.

The segmentation procedure splits a satellite image into homogeneous regions; however, buildings can be covered by several regions. The detection of high objects in the DSM allows us to extract an object entirely, but with low accuracy of boundary localization. We suggested the combination of the outputs of high region extraction and segmentation algorithms as follows: if more than 75% of the pixels of the segmented region are marked as high, then the whole region is considered high (the threshold was chosen empirically). This enables to increase the accuracy of object localization and reduce the number of false alarms (for regions of trees closely adjacent to buildings). The result of the proposed approach is shown in figure 1 and in table 1.

3. Building extraction using convolutional neural networks

The use of convolutional neural networks made it possible to significantly improve the performance of computer vision algorithms in solving a variety of tasks, including the task of extracting and classifying objects in satellite images. Due to the increasing interest of various organizations in this task, at least four global online competitions [7-9] on the extraction and classification of objects in aerospace images were conducted in the last two years. Among the variety of solutions, the best performance for building extraction in [7,8] was shown by convolutional neural networks – modifications of the U-Net architecture [3]. It is worth noting that a better accuracy of object extraction is achieved by training a separate model of U-Net for each class of objects. In this work, we used the implementation of U-Net described in [10].

U-Net has a deep convolutional encoder-decoder architecture. At the input, the network receives a tensor containing satellite image channels and channels with additional data about the scene. At the output, U-Net returns a single-channel mask coinciding with the size of the original image and containing the found buildings. The network is trained on batches consisting of fragments with the size of 112 ×112 pixels randomly taken from the image, to which reflection over the x or y axes is randomly applied. Its training was performed on a four-channel satellite image of an urban settlement with the size of about 5 × 5 km and 1-m resolution that contains more than 3,000 objects such as multi-storey buildings, garages, hangars and cottages. In addition, we used a digital surface model obtained from the stereo matching algorithm proposed in [1]. The examples of the input data are presented in figure 1. To assess the influence degree of the data composition on the recognition quality, four U-Net models were trained on the following data sets:

- image, NDVI, DSM;
- image, DSM;
- image, NDVI;
- image.

The results of the U-Net network trained on different data sets are shown in figure 1 and in table 1. Building extraction quality of the trained network was also evaluated using satellite images with different spatial resolutions. The results of building extraction from the images with the resolution of 2 m and 3 m per pixel can be observed in the figure 2 and table 2.

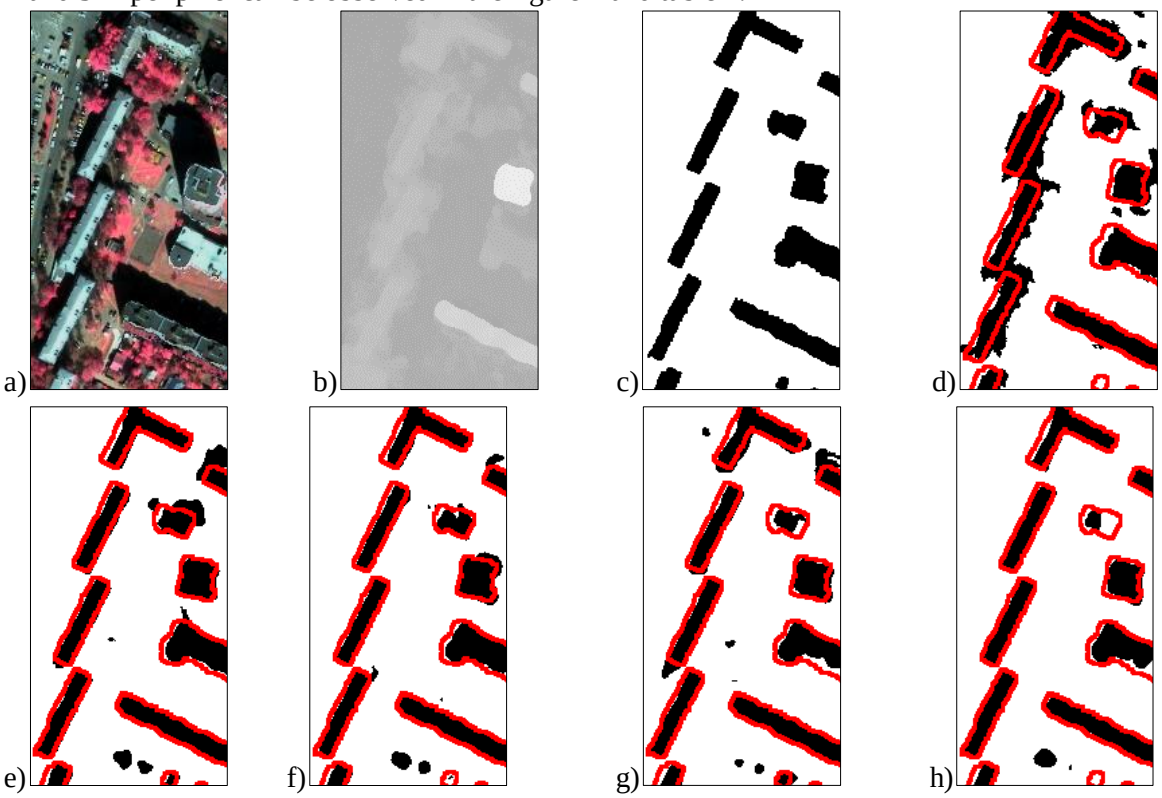

Figure 1. (a) Orthorectified four-channel satellite image (near-infrared, red, green channels) with the 1-m resolution, (b) digital surface model obtained from a stereo pair using a stereo matching algorithm, (c) the hand-drawn ground-truth building regions, (d) detection of buildings in the DSM and refinement of their boundaries using segmentation of spectral information. Building extraction by the U-Net network trained on: (e) image, DSM and NDVI, (f) image, DSM, (g) image and NDVI, (h) only on image. The extracted buildings and the ground-truth building boundaries are displayed in black and red respectively.

4. A comparison of considered approaches

The quality of the considered algorithms was evaluated using a fragment of a four-channel satellite image with the size of 900 × 500 pixels and 1-m resolution. The regions marked as buildings by the

algorithms were compared with the hand-drawn ground-truth building regions according to the following criteria:

1. Intersection over Union (also known as the Jaccard Index), which is sensitive to the accuracy of localization of the detected objects:

$$IoU = \frac{|S_{obj} \cap R|}{|S_{obj} \cup R|},$$

where S_{obj} – the regions of the extracted buildings, and R – the ground-truth building regions. Its value changes from 0 to 1, where 1 means full overlap of the found regions with the ground-truth ones.

2. The rate and amount of true positives, and false alarms in the output of the algorithms.

The results of the comparison can be observed in table 1 and in figure 1.

Of the two approaches to building extraction, the best performance was demonstrated by the U-Net network. Its advantage, as compared with the algorithm that extracts buildings only by using a DSM, is the ability to detect cottages and other low-rise buildings that are not visible in the DSM and these, at least, are one third of the objects in the initial 5 × 5 km image. Also, U-Net performs more accurate detection of the building boundaries.

According to table 1 and figure 1, U-Net shows the good result even without using additional information about the scene. However, the network was trained with height information better detect building boundaries, as was proved by the higher IoU value. The use of the NDVI lead to deterioration in building detection quality. A possible explanation for this might be that NDVI extracts not only vegetation but also buildings. The quality of the building extraction of U-Net was also evaluated on the image of another town with the size of about 5 × 5 km which can be seen in figure A1. Despite the fact that the image differs from the training set by shooting conditions and the composition and type of objects, U-Net (image, dsm) showed for it a similar rate of false alarms, correct and missed detections.

As regards building detection quality of U-Net for satellite images with different resolution, the network is also able to detect buildings in such cases. Figure 2 illustrates that the quality of building extraction depends on the building area and the network was unable to detect barely visible objects. Also, as the resolution is reduced, the quality of detection of building boundaries decreases. Perhaps, it is possible to improve the accuracy of boundary detection on images with different resolutions by adding such images to the training set of the network. It should be noted that to date several ways of improving the U-Net performance have been proposed [11-13], and it is most likely the conducted experiments will be valid for the proposed networks.

Table 1. A comparison of the building extraction algorithms. U-Net was trained on the 5 × 5 km image. The comparison was carried out on the 1 × 0.5 km image (116 buildings).

Algorithm	*IoU*	*True Positive*	*False Alarms*	*True Positive*	*False Alarms*
High-rise objects and segmentation	0.6	76	14	0.66	0.16
U-Net (image, ndvi)	0.74	95	10	0.82	0.09
U-Net (image, ndvi, dsm)	0.74	101	18	0.87	0.15
U-Net (image)	0.74	**102**	**7**	**0.88**	**0.06**
U-Net (image, dsm)	**0.76**	**102**	13	**0.88**	0.11

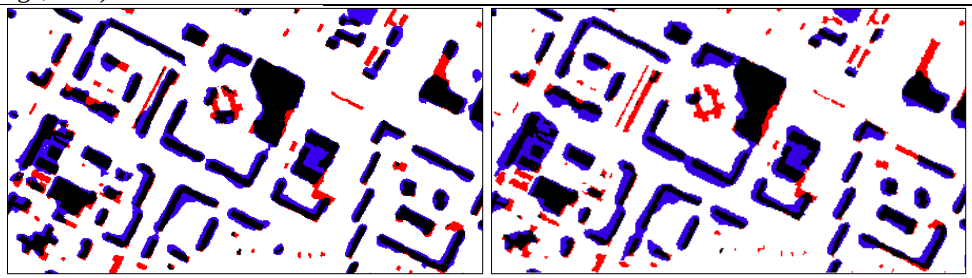

Figure 2. The results of building extraction using U-Net (image, DSM) for images with the 2-m resolution (left), and the 3-m resolution (right). True positive, false negative and false alarms regions are displayed in black, blue and red respectively.

Table 2. A comparison of the building extraction quality of U-Net using satellite images with different spatial resolutions. U-Net was trained on the satellite image with the 1-m spatial resolution (4 channels) and the digital surface model.

Satellite image resolution	*Total*	*True Positive*	*False Negative*	*False Alarms*	*True Positive*	*False Alarms*	*IoU*
1m	3141	3062	79	204	0.97	0.06	0.83
1m, only high buildings	735	728	7	–	0.99	–	–
2m	3285	1904	1381	109	0.60	0.05	0.50
2m, only high buildings	806	712	94	–	0.88	–	–
3m	3171	1282	1889	45	0.40	0.03	0.40
3m, only high buildings	754	645	109	–	0.86	–	–

5. Conclusion

This paper considered a comparison of two approaches to building extraction from satellite imagery and height data. The best performance was shown by the U-Net-based approach, which extracts a larger number of buildings with a better detection of their boundaries. Furthermore, U-Net shows the good result even without using additional information about scene. However, its application is possible only if there is a sufficiently large data set for training the network. The advantage of the approach based on extraction of buildings from a DSM is the absence of parameters that must be selected for each particular satellite image and it will work on images obtained from different satellites. In the future, it is planned to investigate the effect of the size and composition of the training set on the U-Net building extraction quality as well as the possibility of using trained models of neural networks for satellite images of different areas.

Appendix A

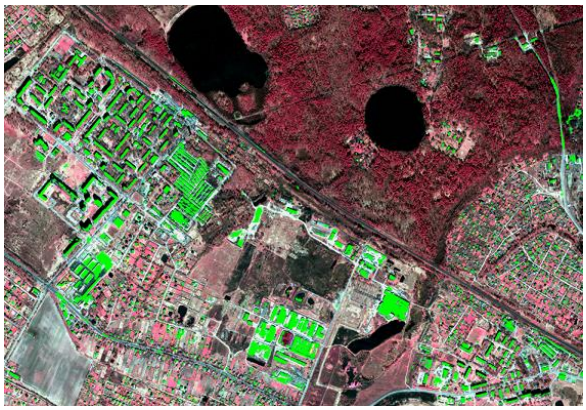

Figure A1. The result of building extraction using the trained U-Net model (image, dsm) for the image with the size of about 5 × 5 km and 1-m resolution. Detected buildings are displayed in green.

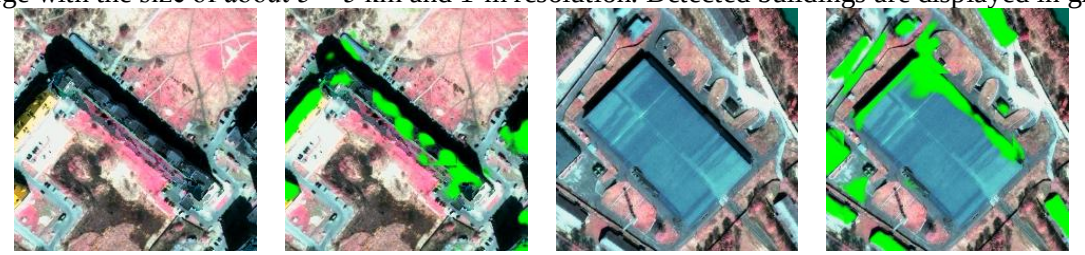

Figure A2. Examples of missed detections from figure A1. Left images in the pairs are fragments from the original image.

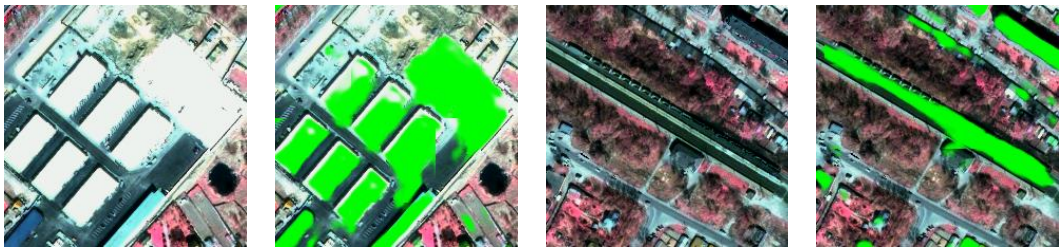

Figure A3. Examples of false alarms from figure A1. Left images in the pairs are fragments from the original image.

6. References

[1] Kostousov V B, Perevalov D S and Kornilov F A 2016 Digital terrain model generation from satellite stereoscopic data *Materials of the XXX conference of the memory of the outstanding designer of gyroscopic instruments N.N. Ostryakov* 382-388

[2] Fursov V A, Goshin Ye V and Kotov A P 2016 The hybrid CPU/GPU implementation of the computational procedure for digital terrain models generation from satellite images *Computer Optics* **40(5)** 721-728 DOI: 10.18287/2412-6179-2016-40-5-721-728

[3] Ronneberger O, Fischer P and Brox T 2015 U-Net: Convolutional networks for biomedical image segmentation *International Conference on Medical Image Computing and Computer-Assisted Intervention Lecture Notes in Computer Science* **9351** 234-241

[4] Maryanova A V 2015 The investigation of the quality of image segmentation algorithms depending on the size of objects in the image *Proceedings of the 46th International Youth School-Conference Actual Problems of Mathematics and its Applications* 129-134

[5] Dunaeva A V and Kornilov F A 2017 Building detection in remote sensing images using a digital surface model *Computational Mathematics and Information Technologies* **2(2)** 185-193

[6] Pettorelli N 2013 *The normalized difference vegetation index* (Oxford University Press) p 208

[7] *Dstl satellite imagery feature detection competition web page* (Access mode: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection) (01.05.2018)

[8] *The SpaceNet challenge web page* (Access mode: https://crowdsourcing.topcoder.com/spacenet) (01.05.2018)

[9] *Planet: understanding the Amazon from space competition web page* (Access mode: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space) (01.12.2017)

[10] Iglovikov V, Mushinskiy S and Osin V 2017 Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition *Preprint arXiv:1706.06169*

[11] Iglovikov V and Shvets A 2018 TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation *Preprint arXiv:1801.05746*

[12] Khalel A and El-Saban M 2018 Automatic Pixelwise Object Labeling for Aerial Imagery Using Stacked U-Nets *Preprint arXiv:1803.04953*

[13] Hamaguchi R, Fujita A, Nemoto K, Imaizumi T and Hikosaka S 2018 Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery *IEEE Winter Conference on Applications of Computer Vision* 1442-1450

Acknowledgments

This work was carried out within the program of the Ural Branch of the Russian Academy of Sciences № 18-1-1-14.