96
Professional paper
Doi 10.61892/STP202601095DJ
ISSN 2566-4484

STEPGRAD
СТЕПГРАД

Marko Đurđević, University of Banja Luka, marko.djurdjevic@student.aggf.unibl.org

BUILDING FOOTPRINT DETECTION USING ARTIFICIAL INTELLIGENCE

Abstract:

In the modern world, automatic classification of buildings from orthophoto images has become critical task in various domains, including urban planning, geographic mapping and infrastructure change monitoring. Traditional building detection methods rely on manual procedures or classical image processing algorithms, which often lead to limited accuracy, especially in complex urban environments where buildings vary in shape, size and spatial arrangement. The development of deep neural networks over the past decade has opened new opportunities for automating and improving this process. However, a significant challenge still remains in achieving an optimal balance between detection accuracy and model efficiency.

Keywords: Buildings, Deep neural networks, Orthophoto, Semantic segmentation

ДЕТЕКЦИЈА ОБЈЕКАТА ПРИМЈЕНОМ ВЈЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

Сажетак:

У савременом свијету, аутоматска класификација објеката из ортофото снимака постала је критичан задатак у различитим областима, укључујући урбано планирање, географско мапирање и праћење промјене инфраструктуре. Традиционалне методе детекције ослањају се на ручне процесе или класичне алгоритме обраде слика, што често резултира ограниченом тачношћу, нарочито у сложеним урбаним срединама гдје варијације у облику, величини и распореду објеката значајно отежавају прецизну сегментацију. Развој дубоких неуронских мрежа у посљедњој деценији отворио је нове могућности за аутоматизацију и унапређење овог процеса, али и даље постоји значајан изазов у постизању оптималног баланса између тачности детекције и ефикасности модела.

Кључне ријечи: Објекти, Дубоке неуронске мреже, Ортофото, Семантичка сегментација

1. INTRODUCTION

Buildings represent fundamental elements of urban environments. Accurate information about their spatial distribution is essential for a wide range of applications such as urban planning, disaster management, infrastructure monitoring, 3D city modeling, etc. Remote sensing data represents one of the most commonly used data sources for building footprint detection. A wide range of techniques exists for object detection in images. Before the advent of machine learning, approaches based on manual classification and classical image processing methods, such as edge detection filters, were widely applied. While these techniques yielded acceptable results for simple object structures, this assumption rarely holds in practical scenarios.

Currently, machine learning-based algorithms, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), U-Net and YOLO-based models, are predominantly used for object detection tasks. Each of these machine learning approaches offers distinct advantages depending on the characteristics of the data and the specific application scenario. In this work, a fully connected neural network was employed for object detection from orthophoto image. The network was designed to precisely distinguish and classify pixels into two classes: “building” and “non-building”. [1] utilized a convolutional neural network (CNN) and five distinct datasets Inria, AIRS, Tanzania, Česma A and Česma B (orthophoto + nDSM) for building detection form UAV imagery over the urban area of Banja Luka. A separate neural network was trained for each dataset, achieving precisions of 0.0966, 0.7204, 0.6565, 0.9151, and 0.9233 for the Inria, AIRS, Tanzania, Česma A, and Česma B models, respectively. Similarly, [2] applied a modified U-Net architecture for building detection from aerial and UAV imagery. This study demonstrates a transfer learning approach to investigate its impact on building detection using data from different sources. Training a deep learning model form scratch requires a large amount of training data and substantial computational time. When only a limited number of images is used for training, the model tends to suffer from overfitting. Transfer learning addresses this challenge by using the weights of a previously trained network on a large dataset as initial parameters for a new model. In this study, the Inria dataset was used for network training, while UAV imagery was applied for fine-tuning. The resulting model achieved a precision of 0.89 when evaluated on UAV imagery. The Inria dataset consists of aerial orthophoto images with spatial resolution of 30 cm. The YOLO framework was used by [3], consisting of four main components: Backbone, Neck, Head and Loss. The experimental data used in this study were obtained from a publicly available dataset complied by the 2016 Data Plus Energy Analytics Group. The dataset comprises high-resolution orthophoto images with spatial resolutions ranging from 0.15 m to 0.30 m, acquired from the United States Geological Survey (USGS), along with corresponding building footprints retrieved from OpenStreetMap (OSM). For the experiments, a total of 2500 ortho-images of size 640 x 640 pixels, covering five different cities, were selected as the experimental dataset. The dataset was randomly divided into training, validation and test sets in an 8:1:1 ratio. The YOLOv8n-seg, YOLOv8n-p2-seg, and the improved model achieved precision values of 0.839, 0.84, and 0.855, respectively. [4] proposed the joint use of S1 and S2 data together with OpenStreetMap (OSM) to generate deep learning models for building and road network detection. In this study, five S2 images were fused with two S1 images to form the input data, while OSM was used to generate the target labels. The proposed model is based on Fully Convolutional Networks (FCNs) and the U-Net architecture.

The main research problem addressed in this paper is the development of an efficient and accurate model for automatic building classification capable of overcoming challenges such as recognizing buildings under varying illumination conditions, identifying targets in high-density environments, and distinguishing buildings from other structures such as roads or vegetation. By leveraging advanced deep learning techniques, the goal is to design a system that achieves a high level of accuracy in building identification and segmentation while requiring minimal tuning and optimization. An effective solution to the problem of automatic object classification has a direct impact not only on academic research but also on practical applications across a wide range of industries. This study focuses on the investigation and development of a methodology that can address these issues, with the aim of providing a comprehensive and efficient approach to automatic object classification under diverse conditions.

2. METHODOLOGY

2.1. STUDY AREA

The area covered by the imagery represents a part of the western region of Banja Luka, including the settlements of Lauš, Petrićevac, Paprikovac and Nova Varoš. In the observed section of the city, residential buildings, infrastructure facilities and vegetation are interwoven. Due to its morphological diversity, this urban area poses a challenge for the application of artificial intelligence algorithms for building recognition. The algorithms can be directed toward the detection of various types of residential and commercial buildings, vehicles, vegetation and infrastructure elements such as roads and parking areas. The imagery exhibits variability in the shape and spatial distribution of buildings, which enables the testing and refinement of recognition models under diverse urban scenarios. The combination of traditional and modern architectural elements, together with the presence of vegetation and open spaces, provides a solid basis for comprehensive analysis and the development of advanced object recognition solutions. Such solutions can be applied in urban planning, property assessment and smart city applications aimed at improving the quality of life of residents.

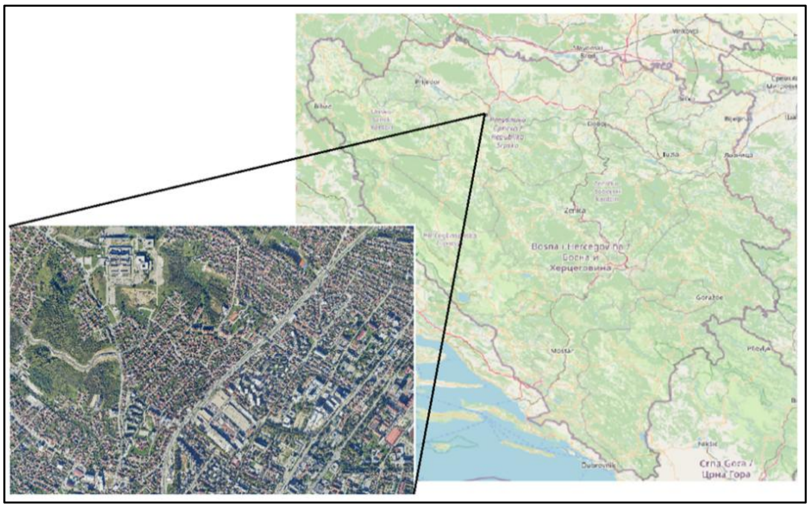

Figure 1. Study area (created by author).

2.2. DATASET AND APPROACH

In this study, data were acquired using a drone, specifically high-resolution orthophoto images with a spatial resolution of 20 cm. A digital orthophoto is a geometrically corrected photograph in digital form, obtained by transforming digital images from central projection into orthogonal projection. Although orthophoto maps have a photographic appearance, their geometric properties correspond to those of products derived from classical geodetic surveying.

Building classification from orthophoto images was performed using a fully connected neural network, specifically a multi-layer perceptron (MLP), implemented in the TensorFlow library [5]. A fully connected neural network is composed of basic computational units, referred to as neurons. The network is organized into multiple layers, each consisting of a set of neurons. Every neuron computes a linear combination of its input values and applies an activation function that introduces nonlinearity into the model. Input layer receives the input data, where each neuron corresponds to a single attribute of the input feature vector. Between the input and output layers, one or more hidden layers are present. In a hidden layer, neurons are connected to all neurons in the preceding layer, and their outputs are in turn connected to all neurons in the subsequent layer. The final layer of the network, which produces the output of the model, is known as the output layer. The number of neurons in the output layer depends on the number of classes or on the type of prediction task being performed [7].

The network used in this work was designed to precisely distinguish pixels into two classes: “building” and “non-building”. No additional processing was applied to the original orthophoto image. However, for the purpose of network training, object extraction was performed based on cadastral data by cropping the relevant objects from the image and generating a new raster containing only these data. This enabled the construction of appropriate training matrices from the corresponding RGB values. The training dataset was created by separately extracting image samples for buildings and for other objects such as vegetation and roads. After this step, the network was trained and the classification performance was evaluated in terms of accuracy.

Classification was performed based on the RGB pixel values. The neural network used in this study consists of an input layer, four hidden layers and an output layer. The first hidden layer contains 64 neurons, the second hidden layer 32 neurons, the third hidden layer 16 neurons, and the fourth hidden layer also 16 neurons. The input layer comprises three neurons corresponding to the RGB values, while the output layer consists of two neurons (0 or 1) representing the classification result, namely “building” and “non-building”.

2.3. IMPLEMENTATION AND TRAINING DETAILS

For model training, a drone-acquired orthophoto image was used. Based on cadastral object data, image subsets representing objects were cropped from selected areas of the image (Figure 2) and used for network training. In the same manner, subsets corresponding to areas representing non-object regions were extracted (Figure 3). These images were cropped using QGIS software [6]. In this way, two categories were created: pixels belonging to buildings and pixels belonging to other objects or the background. This strategy ensures class balance, reduces bias toward dominant classes, and enhances the model’s generalization capability by learning more discriminative features.

Figure 2. Training data for buildings (created by author).

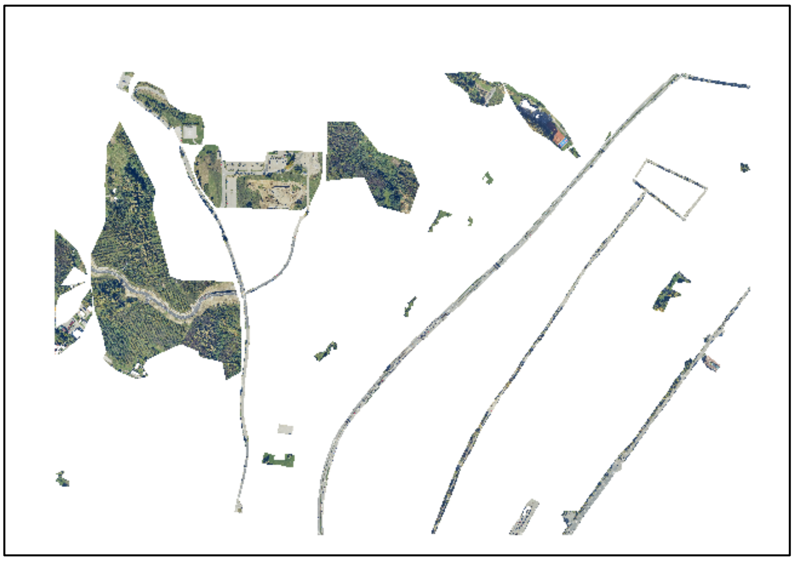

Figure 3. Training data for background and other objects (created by author).

The input data were reshaped so that each pixel was represented as a row containing three values (R, G, B), and pixels that were completely black or white were filtered out in order to remove irrelevant data. Using this approach, it was ensured that the model received a sufficiently diverse set of examples for training. The data were divided into training and validation sets in an 80:20 ratio. The training set was used for model learning, while the validation set served to monitor model performance during training. To prevent overfitting and improve model generalization, an EarlyStopping callback with a patience parameter of 10 was employed, meaning that training would be terminated if the model performance on the validation set did not improve for 10 consecutive epochs. The training process was initially set to 30 epochs but was stopped after the fifteenth epoch due to the implementation of early stopping. This approach ensures that the model is trained only up to the point at which performance is optimal, without unnecessary risk of overfitting.

2.4. ACCURACY ASSESSMENT

Model accuracy was evaluated using the accuracy metric, which measures the proportion of correctly classified pixels relative to the total number of pixels in the validation set. During training, the loss on both the training and validation sets was also monitored, and the results are presented in a graph (Figure 4) that allows a visual assessment of how the loss changes over the epochs.

Categorical Crossentropy Loss was used in this study as the loss function. This function measures the difference between the true and predicted probabilities for each class. A low loss value indicates that the predicted probabilities are close to the actual values.

In addition, the model's performance was evaluated using metrics such as precision, recall, F1 score, and Intersection over Union (IoU). These metrics are described by the confusion matrix [7]. Classification performance is optimal when the confusion matrix is diagonal, while non-diagonal elements represent classification errors. In the case of binary classification, one class is considered positive and the other negative. True positive instances are positive samples correctly identified by the model as positive, while true negative instances are negative samples correctly classified as negative. False positive instances are those incorrectly classified as positive, whereas false negative instances are positive samples incorrectly classified as negative. Precision represents the ratio between correctly identified positive instances and the total number of instances predicted as positive by the model. Recall represents the ratio between correctly identified positive instances and the total number of actual positive instances. The F1 score is defined as the harmonic mean of precision and recall, while Intersection over Union (IoU) describes the degree of overlap between the predicted positive area and the actual positive area.

3. RESULTS AND DISCUSSION

Following the training of the building classification model using UAV orthophoto imagery, an accuracy of 87.83% and a loss value of 0.2771 were achieved on the validation dataset. These results demonstrate that the proposed model effectively distinguishes building pixels from non-building pixels.

An accuracy of 87.83% indicates that, on average, nearly 9 out of 10 pixels were correctly classified as either "building" or "non-building." This result suggests a high level of reliability of the model in distinguishing buildings based on color and texture information extracted from the images.

A loss value of 0.2771 reflects the average discrepancy between the model predictions and the ground-truth labels on the validation set. A lower loss value is indicative of better model performance, as it implies that the model successfully minimized the difference between predicted and actual values during training.

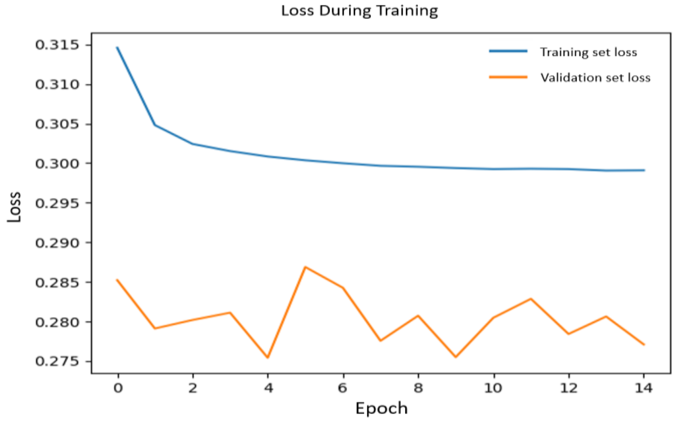

Figure 4. Figure 14. Training and validation loss (created by author).

The accuracy evaluation results (Table 1) show that the model is able to correctly identify pixels representing objects, with 82% of pixels classified as objects actually corresponding to real objects. However, recall shows that 29% of real objects are not correctly classified. The recall value is lower than the precision, which suggests that although the model is able to detect objects, some of them will be omitted. The IoU value of 0.62 indicates a moderate overlap between the detected and real objects, which leaves room for improvement in spatial accuracy.

Table 5. Model accuracy assessment

Parameter	**Precision**	**Recall**	**F1**	**IoU**
Value	0.8259	0.7147	0.7663	0.6211

As another measure of classification quality, a comparison of the areas of buildings from the cadastral data and from the classification results was made. The areas of buildings in the classification are significantly larger than the areas of buildings in the cadastre, and the reason for this is the different representation of buildings in the image and in the cadastre, because there are eaves in the image, while there are none in the cadastre. Another reason is that some buildings are missing in the cadastre, while they are shown in the image. In addition, the difference in areas is also influenced by the neural network model itself, due to the classification of incorrect parts of the image such as shadows, roads, etc.

Figure 5. Figure 15. Example of mismatch between objects from the image and objects in the cadastre (created by author).

Visually comparing the model classification results with the original orthophotos allows for an intuitive and easy-to-understand assessment of the model's performance. This comparison method provides insight into how well the model distinguishes buildings from other objects and backgrounds, and helps identify potential model weaknesses that could be improved.

Figure 6. Figure 16. Visual comparison of results (created by author).

The model successfully detects most of the buildings, preserving their structure and shape. Most of the buildings are correctly classified, which shows that the model has a high ability to distinguish buildings from other objects, which also confirms the accuracy achieved by the model during the evaluation. In several places, the model also classified non-building features, including wider areas of ground (road) that share similar characteristics with buildings (similar texture or color).

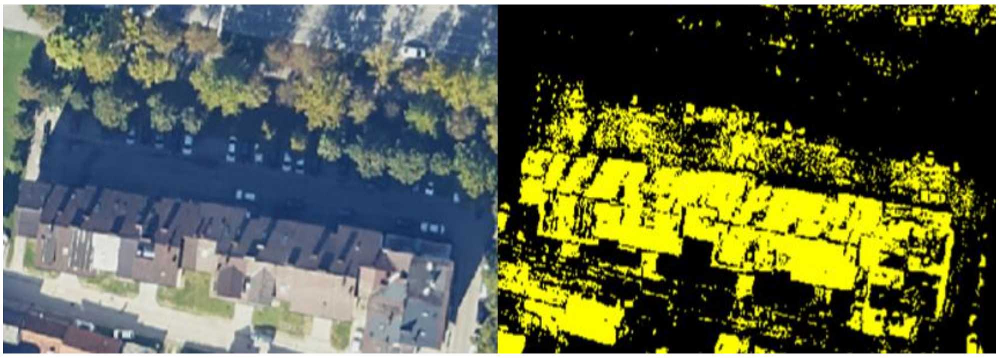

Figure 7. Figure 17. Shadow classification (created by author).

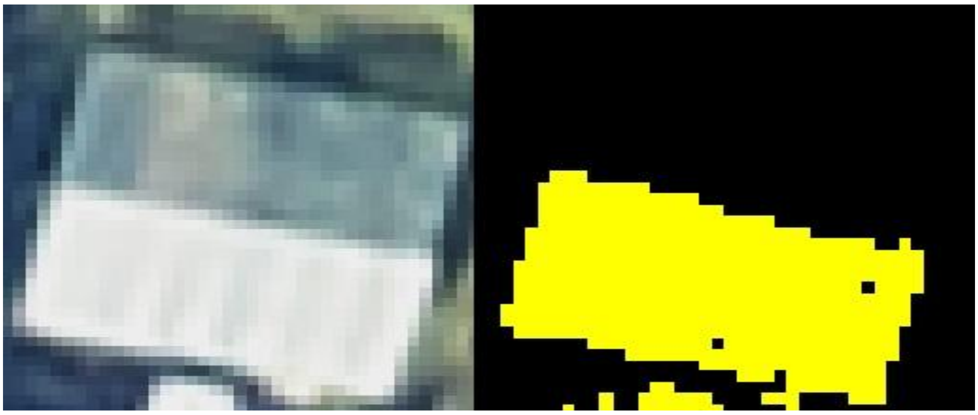

Figure 8. Figure 18. The darker parts of the roof are not classified as building (created by author).

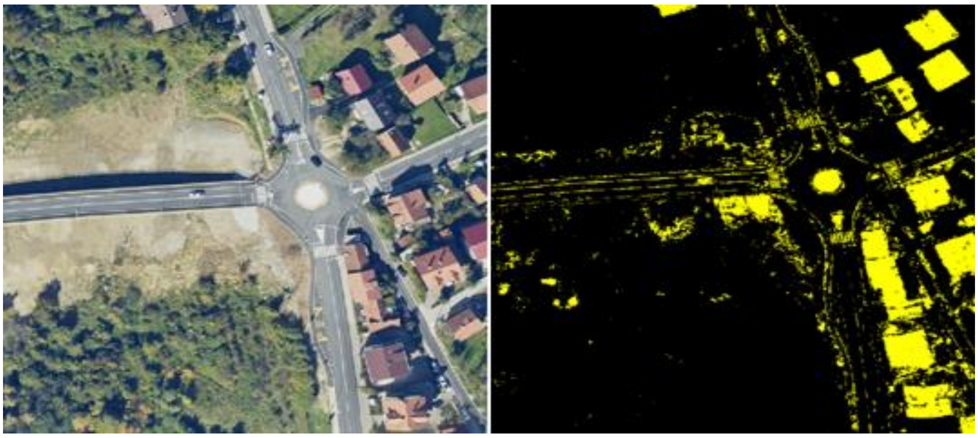

Figure 9. Figure 19. Road classification (created by author).

Finally, building footprint regularization methods could be incorporated to improve boundary accuracy and ensure more geometrically regular building representations.

4. CONCLUSION

This paper presents a methodology for automatic building classification from orthophoto images using machine learning techniques. The goal of the research was to develop a model that enables accurate detection and segmentation of buildings with minimal manual tuning and intervention requirements. Using a custom multi-layer neural network model and optimized architecture, satisfactory results were achieved on key metrics, including precision, recall, F1 score, and IoU. One of the key advantages of this approach is its ability to balance accuracy and efficiency, providing reliable results with relatively low computational resources. However, like any method, it has certain limitations, particularly in highly heterogeneous or extremely complex urban environments. The model's performance can be further improved by incorporating additional parameters, such as nDSM, increasing the number of neurons and hidden layers, and employing more extensive training, provided that sufficient computational resources are available. This can lead to enhanced accuracy and efficiency in building classification.

LITERATURE

[1] Z. Kokeza, M. Vujasinovic, M. Govedarica, B. Milojevic, and G. Jakovljević, "Automatic building footprint extraction from UAV images using neural networks," *Geodetski Vestnik*, vol. 64, no. 4, pp. 545–561, Dec. 2020, doi:10.15292/geodetski-vestnik.2020.04.545-561.

[2] S. Daranagama and A. Witayangkurn, "Automatic Building Detection with Polygonizing and Attribute Extraction from High-Resolution Images," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 9, p. 606, Sep. 2021.

[3] Z. Song, W. Wang, Z. Hao, X. Li, S. Tang, and L. Xie, "Building footprint extraction from aerial images using an edge-aware YOLO-v8 network," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-1-2024, pp. 605–611, May 2024.

[4] C. Ayala, R. Sesma, C. Aranda, and M. Galar, "A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in High-Resolution Satellite Imagery," *Remote Sens.*, vol. 13, no. 16, p. 3135, Aug. 2021.

[5] Google LLC, "TensorFlow," Accessed: Feb. 25, 2026. [Online]. Available: https://www.tensorflow.org/.

[6] QGIS Association, "QGIS Geographic Information System," Accessed: Feb. 25, 2026. [Online]. Available: https://qgis.org/

[7] M. Nikolić and A. Zečević, *Mašinsko učenje*. Belgrade, Serbia: University of Belgrade, Faculty of Mathematics, 2019.