A Semi-Automated Two-Step Building Stock Monitoring Methodology for Supporting Immediate Solutions in Urban Issues

Mehmet Isiler 1,*, Mustafa Yanalak 1, Muhammed Enes Atik 1, Saziye Ozge Atik 2 and Zaide Duran 1

1 Department of Geomatics Engineering, Faculty of Civil Engineering, Istanbul Technical University, 34469 Istanbul, Türkiye
2 Department of Geomatics Engineering, Faculty of Engineering, Gebze Technical University, 41400 Kocaeli, Türkiye
* Correspondence: isiler@itu.edu.tr

Abstract: The Sustainable Development Goals (SDGs) have addressed environmental and social issues in cities, such as insecure land tenure, climate change, and vulnerability to natural disasters. SDGs have motivated authorities to adopt urban land policies that support the quality and safety of urban life. Reliable, accurate, and up-to-date building information should be provided to develop effective land policies to solve the challenges of urbanization. Creating comprehensive and effective systems for land management in urban areas requires a significant long-term effort. However, some procedures should be undertaken immediately to mitigate the potential negative impacts of urban problems on human life. In developing countries, public records may not reflect the current status of buildings. Thus, implementing an automated and rapid building monitoring system using the potential of high-spatial-resolution satellite images and street views may be ideal for urban areas. This study proposed a two-step automated building stock monitoring mechanism. Our proposed method can identify critical building features, such as the building footprint and the number of floors. In the first step, buildings were automatically detected by using the object-based image analysis (OBIA) method on high-resolution spatial satellite images. In the second step, vertical images of the buildings were collected. Then, the number of the building floors was determined automatically using Google Street View Images (GSVI) via the YOLOv5 algorithm and the kernel density estimation method. The first step of the experiment was applied to the high-resolution images of the Pleiades satellite, which covers three different urban areas in Istanbul. The average accuracy metrics of the OBIA experiment for Area 1, Area 2, and Area 3 were 92.74%, 92.23%, and 92.92%, respectively. The second step of the experiment was applied to the image dataset containing the GSVIs of several buildings in different Istanbul streets. The perspective effect, the presence of more than one building in the photograph, some obstacles around the buildings, and different window sizes caused errors in the floor estimations. For this reason, the operator’s manual interpretation when obtaining SVIs increases the floor estimation accuracy. The proposed algorithm estimates the number of floors at a rate of 79.2% accuracy for the SVIs collected by operator interpretation. Consequently, our methodology can easily be used to monitor and document the critical features of the existing buildings. This approach can support an immediate emergency action plan to reduce the possible losses caused by urban problems. In addition, this method can be utilized to analyze the previous conditions after damage or losses occur.

Keywords: monitoring building stock; building floors estimation; OBIA method; YOLO algorithm; VHRS images; street-view images

1. Introduction

The Sustainable Development Goals (SDG) and the New Urban Agenda (NUA), which are guiding urban environment studies worldwide, address environmental and social
concerns in cities, such as resilience to natural disasters, insecure land tenure, inefficient energy consumption, and climate change [1]. Such international reports have also made urban problems more visible [1] and encouraged authorities to develop urban land policies to improve urban life. The availability of up-to-date, reliable, and accurate building information about cities from local to global scales plays a crucial role in solving urbanization issues. Establishing efficient land administration systems with a comprehensive approach takes considerable time. However, some practices should be implemented immediately to reduce the potential adverse effects of urban challenges on human life.

In developing countries, public records may not reflect the current status of buildings due to reasons including illegal housing and a lack of coordination between cadastral boards and municipalities.

Rapid urbanization always involves higher demand for land [2]. Accelerated urban growth frequently increases informal settlements and slum areas in less developed countries [3]. While slum areas have reduced considerably in developed countries, rural-to-urban migration that stimulates the formation of slums is still a significant problem in developing countries [4]. Informal housing and slums usually emerge due to various socio-economic and cultural reasons, including limited housing production, the economic priorities of countries, and legal gaps in regulations regarding city planning [5].

Nowadays, designing climate-change-sensitive and energy-efficient living spaces has become a primary objective of spatial planning. Developing countries have had to make considerable efforts to deal with informal housing and insecure land tenure issues, causing wasting time and financial losses, instead of adopting new urban designing approaches [6].

It should be noted that formal housing can be turned into illegal housing over time by adding stories and changing building characteristics without permission from the authorities [5]. Inadequate control methods and some legal gaps in the land registration system are the main reasons for changes in the legal status of buildings over time. In addition, the cadastral administration and local governments should be coordinated during the construction process. Otherwise, formal structures may not be recorded in the registry system.

Therefore, buildings should be continuously monitored to identify possible changes and prevent illegal construction, especially in developing countries. This study aims to develop a rapid and automated method for monitoring actual building stock using the potential of VHRS and GSV images. Our proposed approach can automatically identify essential building characteristics, such as the footprint of buildings (horizontal information) and the number of building floors (vertical information).

Today, remote sensing satellite images, street-view images, and open data movement [7] provide an opportunity to determine building characteristics in urban areas. Satellite images offer a wide-ranging overview of the area of interest. Thus, essential information about vast areas can be obtained all at once. Following the recent improvements in satellite sensor technology, satellite images with sub-meter spatial resolution are now commonly used data sources for extracting building footprint information with sufficient accuracy. However, satellite images are generally incapable of gathering complete information about building façades. In this case, street-view images (SVIs) can be used to extract the vertical details of urban features. SVIs provide human-centric views of urban streets. Street view creates a feeling of traveling in the city streets by combining images taken from different angles [8]. Street-view images can be accessed through online services such as Google, Microsoft, Baidu, and Tencent [9]. Due to their availability and free accessibility, the use of street-view images in urban studies has noticeably increased. In addition, advances in computer vision and deep learning techniques in image analysis enable automated and rapid solutions.

In our study, we first focused on automatically detecting the existing building stock in the study areas. Thus, the horizontal geometry of the extracted buildings can be obtained to control whether the identified buildings are registered in the related public documents. This information can be used to determine the horizontal occupancy of unregistered build-
ings. To this end, buildings in the study areas were extracted by applying the object-based image analysis (OBIA) method to a Pleiades satellite image with high spatial resolution, which covers two different urban streets in Istanbul and the Ayazağa Campus of Istanbul Technical University (ITU). The geographical coordinates of the central points of the extracted building footprints were obtained, and all extracted buildings were named with the geographical coordinates of their central points. These coordinates were used to obtain SV images of each building.

The second goal of our study is to document the vertical elements of building façades and automatically detect the number of building floors. Windows and doors are the main elements of buildings used to identify the building floors. We modified and enlarged a dataset prepared by Sezen et al. [10]. This dataset contained approximately 1000 Google Street-view images of the buildings in different streets of Istanbul and structures in ITU Ayazaga Campus. The YOLOv5 algorithm was used for window and door extraction. Then, the midpoints of the windows were calculated. The calculated midpoint coordinates were grouped along the y-axis of the image using the kernel density estimation method. It was accepted that the number of groups formed corresponds to the number of floors in the building. The general workflow of the methodology used in this study is shown in Figure 1.

**Figure 1.** General workflow of the proposed method.

This proposed method does not obtain 3D building information with the level of accuracy of a cadastral survey. However, the main aim of this study is to identify the critical elements of existing buildings in order to quickly produce immediate action plans to reduce the harmful effects of urban problems as soon as possible. The first stage was performed on high-resolution Pleiades satellite images to automatically detect existing buildings in the study areas. The accuracy of the building and nonbuilding classes was very good in all study regions. The second stage was carried out to determine the number of building floors based on the dataset obtained from the GSVIs of many buildings in Istanbul. The proposed
method calculated the number of floors for all buildings automatically, with 79.2% accuracy for the SVIs collected by operator interpretation. If the appropriate shooting conditions are provided, our method can produce satisfactory results and may be utilized to identify the vertical features of buildings.

1.1. Literature Review

1.1.1. OBIA for Building Extraction from VHRS Images

One of the main issues in satellite image analysis is obtaining accurate information about buildings' shape, size, and structure with satisfactory accuracy in short times [11]. High-resolution satellite (HRS) images provide an opportunity to extract building boundaries at a specific time and analyze the spatial distribution of buildings across a specific period. Studies of HRS images have recently increased for semi-automated or automated building extraction techniques. Building extraction from satellite images in urban applications is complicated [12,13] due to the different sizes and shapes of buildings, varying roof textures, and obstacles in surrounding buildings [13]. Classification is the focal point of building extraction techniques. Classification methods can be grouped into two main types: pixel-based image analysis and object-based image analysis (OBIA). Whereas pixel-based classification only uses the spectral information of every pixel within the boundaries of the study areas, object-based classification also considers its spatial, textual, or contextual information [13]. The object-based method can be considered more advanced in this regard.

Several previous studies have discussed the performance of object-based and pixel-based analyses. Myint et al. examined whether an object-based classification could accurately identify urban classes [14]. They used QuickBird images of central region in Phoenix, Arizona. The object-based classifier produced a high overall accuracy of 90.40%, whereas the maximum-likelihood classifier, a per-pixel classifier, provided a lower overall accuracy of 67.60% [14]. Rittle et al. investigated the performance of object-based and pixel-based techniques for producing land cover mapping of the Alto Riberia Tourist State Park, in the Brazilian Atlantic rainforest area [15]. This comparative study showed that object-based classification provided higher accuracy, with a kappa index value of 0.8687, than a hybrid per-pixel-based classification, with a kappa index value of 0.2224 [15]. Gao and Mas investigated the success of object-based image analysis in classifying satellite images with varying spatial resolutions, comparing the classification results from the pixel-based method [16]. This study showed that object-based image analysis in higher-spatial-resolution images achieved higher classification accuracy [16].

Many studies based on the OBIA technique have provided high-quality results for building extraction. Prathiba et al. proposed a method combining object-based nearest-neighbor classification and a rule-based classifier for extracting building footprints from VHRS images [17]. This method provided good results with an accuracy of over 82.5% [17]. Jamali et al. tested the two automated building extraction approaches on Istanbul’s aerial LiDAR point cloud and digital imaging datasets [18]. The object-based automated technique presented better results than the threshold-based building extraction technique in regard to visual interpretation [18]. Gavankar and Ghosh applied an object-based automatic building extraction technique to obtain building footprints using pan-sharpened IKONOS multispectral images [13]. This proposed method did not require training samples or a digital elevation model to extract highly accurate building footprints [13].

Some other studies also used OBIA on UAV and satellite images to identify the physical characteristics of urban slum settlements [19] and differentiate formal and informal housing in the built-up areas [20]. It should be noted that some alterations in formal buildings, which include increasing the number of floors without authorization, do not generally have common slum characteristics that differentiate them from the formal urban environment. Therefore, using common image analysis techniques that detect and classify formal and slum settlements is insufficient for these conditions. Our study aimed to develop an automated and rapid approach for extracting and documenting all existing buildings to
control and update public records, rather than classifying slums and formal housing based on HRS images.

1.1.2. YOLO Algorithm for the Extraction of Building Façade Elements from SV Images

Compared to the traditional remote sensing approach, SV images effectively contribute to vertical urban research, enabling us to understand street spatial structures, identify urban features, and extract building façade elements. Traditional aerial photogrammetry produces a two-dimensional image of the city using vertical overview (i.e., nadir) imaging [21]. Unlike overlook imaging, oblique imaging can accurately characterize the specific details of urban structures [21]. Oblique aerial images from UAVs can produce meaningful information for building façade studies. However, they also require the development of matching algorithms to cope with the unique characteristics of large-scale urban oblique aerial imagery, such as depth discontinuities, occlusions, shadows, low texture, and repetitive patterns [22]. One of the typical problems in UAV image analysis is that a transformation must be performed, since the building façade is not frontally viewed [23].

Street view images can be easily used to collect critical information about urban areas at the street scale. This new data tool has become popular in many kinds of research. These studies have mainly interested in evaluating visual perceptions of urban streets and street space qualities [9,24–28], detecting trees and plants in urban cities [29–31], exploring the climatic conditions and air pollution of cities [32–34], and determining walkability in the urban streets [35,36]. These efforts have proven that the analysis of SVIs produces comprehensive and reliable results for understanding human interactions with urban environments and outdoor dynamics.

In addition to cultural and social studies, it should be noted that building façade information can be used for other technical purposes such as building reconstruction [10], monitoring structural health [23], detecting façade faults [10], identifying urban heritage, building energy performance analysis [37] and low-energy building design [37,38]. Object detection is the central part of SVI analysis when fulfilling these technical purposes. Object detection is utilized to identify and locate one or more types of objects in images [39–41]. Detecting building façade elements, which is a sub-theme of object detection, is a critical issue in computer vision for image analysis [10]. Multi-source data, analyzed with machine learning and computer vision techniques, have provided a broad understanding of urban analysis. Deep learning algorithms have recently been widely employed in object recognition and detail extraction [10].

Traditional detection approaches usually encounter some challenges making them difficult to implement due to their high temporal complexity, low robustness, and strong scene dependency [42]. In recent years, target detection methods based on convolutional neural networks (CNNs) have provided satisfactory detection results [42]. Deep learning algorithms based on convolutional neural networks have significantly improved the efficiency and accuracy of target detection methods. Object detection methods based on deep neural architectures can be categorized as one-stage or two-stage detectors [43,44]. Classification and localization are the two significant issues for object detection, and two-stage algorithms have been developed to solve these difficulties [45]. R-CNN, Fast R-CNN, Faster R-CNN, SPP-Net, and Mask-RCNN are two-stage models that involve producing region proposals and classifying region proposals in the classifier and correcting positions [46]. They extract a set of candidates’ bounding boxes in the initial stage [45]. In the second step, the CNN detector collects objects from the selected candidate bounding boxes, and these features are then utilized for classification and bounding box regression [41]. The one-stage detectors, including the Single Shot Multibox Detector (SSD), RetinaNet, Fully Convolutional One-Stage (FCOS), DETection TRansformer (DETR), and the You Only Look Once (YOLO) family, predict the bounding boxes and compute the class probabilities of these boxes through a single network [39]. When developing a model, it is crucial to balance computation speed and detection accuracy [42]. Two-stage algorithms provide accurate
results but are slow and structurally complicated [45]. One-stage detection methods are simple and more suitable for quick object detection [41].

YOLO is one of the best known deep learning algorithms, and its single-stage detection architecture makes it particularly fast [47]. It addresses categorization and localization difficulties as a single regression problem [45]. YOLOv5, based on YOLOv4, is the most widely used YOLO series detection technique. Furthermore, YOLOv5 aims to improve small-target detection capabilities [37].

The YOLO algorithm is widely used for target detection in many different kinds of studies, such as estimating the positions of pedestrians from video recordings [48], identifying insect pests that influence agricultural production [49], iceberg and ship discrimination [50], ship detection and recognition in complex-scene SAR images [51], bridge detection [52], underwater object detection [53], automatic roadside feature detection [54], and object detection in automated driving [55].

Some studies have examined YOLO algorithms for door and window detections. Bayomi et al. described a deep learning approach to extract window and door elements that can be used in analyses of buildings’ energy performance [37]. Zhang et al. developed the DenseNet SPP–YOLO algorithm based on the YOLOv3 version, to enable autonomous mobile robots to recognize doors and windows in unfamiliar environments [43]. Sezen et al. compared the performance of YOLOv3, YOLOv4, YOLOv5, and Faster R-CNN in detecting doors and windows [10]. This study showed that YOLOv5 may be suitable for extracting the door and window elements of buildings by examining accuracy and speed together [10]. For this reason, YOLOv5 was chosen to perform window extraction in our study.

2. Materials and Methods

2.1. Study Area

For the study area, three different locations with dense settlements were selected from Istanbul in Türkiye (Figure 2). The selected regions include many different kinds of building structures. One of the selected regions is the Istanbul Technical University campus. Other areas were selected from the same city, one from Asia and one from the European continent. An experiment was conducted in the three selected areas for building extraction with OBIA. Additionally, the buildings in the regions were also included in the test data for building floor estimation.

Figure 2. Study area and high-resolution satellite image.
In the study, a very-high-resolution image of the Pleiades satellite with a resolution of 0.5 m was used for building extraction. The red–green–blue and near-infrared bands of the ortho-image product were used. Open Street Maps (OSM) building boundaries were utilized to show the difference between the building extraction boundaries (Figure 3). The missing building boundary vectors in the OSM data were added by the operator.

Using the ‘view street’ option of the Google Maps application, screenshots of randomly selected buildings in randomly selected urban cities and streets throughout Türkiye were obtained (Figure 4). We modified and enlarged a dataset prepared by Sezen et al. [10]. The street-view images were collected with the help of operator interpretation. There are 1006 building photos in the dataset, including detached houses, apartments, and residences. The doors and windows on the building façades were manually tagged after the photos were collected. The VGG JSON and COCO JSON formats were used to export the labels. The dataset is divided into 800 images for training, 100 images for validation, and 106 images for test data. Additionally, the images of the 39 buildings in the study areas were automatically obtained via the Google Street View API using the coordinates determined from the vector data and these buildings were added to our dataset to make extra analysis.

**Figure 3.** Open Street Maps (OSM) building boundaries. (a) Area 1; (b) Area 3; (c) Area 2.
2.2. Object-Based Image Analysis

OBIA includes two main stages: image segmentation and the classification of segmented image objects [16,17]. The segmentation process groups image pixels into meaningful image objects according to their spectral similarity and spatial characteristics, such as shape, area, and position [56]. In the classification process, the image objects obtained in the previous stage are assigned to different classes [17].

One of the main reasons for choosing OBIA as the classification method in this study is that it is aimed to obtain the building class with less of a salt and pepper effect. Unlike pixel-based classification, one of the main differences of the OBIA implementation is that the resulting classes consist of less noisy objects with a reduced salt and pepper effect. One of the main differences on which classification techniques are based is that, in OBIA, related neighboring pixels are grouped and classified into segments. Pixel-based classification is another classical method that is widely used in remote sensing applications. However, one of the main issues with pixel-based classification is that data from nearby pixels, which can help to precisely identify the target pixel class, are not used. Using OBIA for pixel grouping, similar pixel clusters can be examined by size, shape, texture, and spectral characteristics.

In our study, the multi-resolution segmentation (MRS) algorithm was used to extract buildings in the related study area. MRS is a widely used segmentation algorithm [57,58]. MRS methods can provide satisfactory classification results by selecting appropriate parameters for interested objects [58]. Multi-resolution segmentation uses an iterative method to reduce the average heterogeneity of image objects [18]. Image objects are arranged in groups until the maximum object variance is achieved as a threshold [18]. In the multi-resolution segmentation algorithm, the scale, shape, size and texture parameters are used [17,19]. The OBIA expert should determine several segmentation parameters according to the application.

2.3. You Only Look Once (YOLO) Algorithm

The YOLO algorithm considers object detection in images as a regression problem [45]. YOLOv5 consists of 3 main elements: the backbone, neck, and head [42]. The backbone collects input images and extracts image features [49] through convolutions at different scales.
The neck is used to collect and combine image features from the backbone, and these combined features are sent to the head to generate bounding boxes and class predictions [45].

The BottleneckCSP module extracts information from the image to create a feature map [59]. Unlike other large-scale convolutional neural networks, the BottleneckCSP structure reduces repetitive gradient information in the optimization of convolutional neural networks [49,59]. Depending on the width and depth of the BottleneckCSP module, different versions of YOLOv5 are created: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x [42,59]. YOLOv5 also includes a PANet as a neck using a new feature pyramid network FPN structure [49] with strong top-bottom semantic features and solid bottom-up positioning features [59]. Thus, this allows the model to process and detect targets at different scales [49] in the head elements. This architecture is illustrated in (Figure 5) as represented in Xu et al. [60].

![YOLO architecture](image)

**Figure 5.** YOLO architecture [60].

2.4. Kernel Density Estimation

Kernel density estimation (KDE) is a non-parametric and neighbor-based technique for predicting a random variable’s probability density function [61]. It essentially calculates the distribution in a set of points based on the selected kernel. A kernel is a $K(x, h)$ function that calculates the distribution of the points. Equation (1) for a Gaussian kernel is:

$$K(x) = \frac{1}{h\sqrt{2\pi}}e^{-0.5\left(\frac{x-x_i}{h}\right)^2}$$  \hspace{1cm} (1)

where $x_i$ is the observed data point, and $x$ is the point where the kernel function is calculated, and $h$ is known as the bandwidth [62]. The bandwidth is a smoothing parameter that controls the balance between bias and variation [62]. While a wide bandwidth causes high-bias density distribution, unrelated points can be assigned to the same cluster. If the bandwidth is low, a high-variance distribution occurs [62]. In simple 1d kernel density estimation, distribution is visualized as a histogram. In KDE, curved peak values appear in regions of point concentration. In this study, these peak regions were used to estimate the number of floors.
2.5. Evaluation Metrics

The standard accuracy assessment metrics used in building extraction are the F1 score and intersection over union (IoU), which are calculated using the metrics of true positive (TP), false positive (FP), and false negative (FN) [63]. Overall accuracy (OA), which is an estimation of the percentage of pixels that have been appropriately classified, is used as the evaluation metric for the classification [63]. TP is when the predicted label and ground truth are positive. FP is when the predicted label is positive, but the ground truth is negative. FN is when the predicted label is negative, but the ground truth is positive [64]. The IoU value is expressed as the ratio of correctly predicted pixels to the total number of correct and predicted pixels belonging to each class. [65]. The mean average precision (mAP) is the average value of the precision of all classes [64]:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (2)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3)
\]

\[
\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN} \quad (5)
\]

\[
\text{Overall accuracy} = \frac{1}{k} \sum_{i=1}^{k} \frac{N_{ii}}{N} \quad (6)
\]

\[
AP = \sum_{k=0}^{k=n-1} \left[ \text{Recall}(k) - \text{Recall}(k + 1) \right] \times \text{Precision}(k) \quad (7)
\]

\[
mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (8)
\]

3. Experiment

In the first step of the experiment, buildings were extracted from the high-resolution satellite image using OBIA (Figure 6). Buildings were extracted using the Pleadias satellite image of three areas in Istanbul province. First, the appropriate parameters were determined experimentally, and objects were generated with a multi-resolution segmentation algorithm in eCognition Developer 9 software. While the segmentation parameters were selected as 45 scale and 0.4 shape and 0.6 color for Areas 1 and 3, a shape of 0.3 was determined in Area 2. In the classification, the objects of the MRS algorithm results were utilized. RGB-NIR bands were used and the nearest neighborhood algorithm was applied. These objects were then assigned to the building and nonbuilding classes. According to the binary classification, the confusion matrix and evaluation metrics were calculated.

![Figure 6. Workflow of object-based image analysis for building extraction.](image)

Assuming that each floor has a window opening outwards, the windows need to be identified to estimate the number of building floors. An artificial-intelligence-supported approach was developed to estimate the number of floors (Figure 7). The YOLOv5 algorithm was used for window detection. Training and test data were created by labeling the windows and doors of the building images collected via Google Street View. The doors
were also labeled to prevent them from being predicted as windows by the algorithm due to their similarity with the windows. YOLOv5 creates a bounding box for object detection.

The midpoint of the bounding box of each window detected on a facade was calculated. Then, the pixel coordinates of the center point were obtained. The calculated midpoint coordinates were grouped along the $y$-axis of the image using the kernel density estimation method. It was accepted that the number of groups formed corresponds to the number of floors of the building. The PyTorch 1.10.0 library CUDA 11.1 was installed in the YOLO versions during training. When the batch size was entered as 16, the epoch parameter was set to 200. The training period lasted approximately 1 h and 21 min. For the experiments, i7-11800H, a 2.30 GHz processor, a GTX 3070 graphics card, and 32 GB RAM hardware were used.

4. Results

4.1. Results of Building Segmentation

As the reference data of the building footprints in the study area, the reference vector data area was prepared manually. Pixel-based accuracy analysis was performed between the ground truth and the prediction image. According to the classification results, two classes were evaluated. Except for the buildings, all other objects were labeled as the non-building class. Accuracy metrics were calculated based on the confusion matrix presented in Table 1. The table shows that the building class is more confused with the non-building class than the other way round. In Table 2, the accuracy metrics for the building and non-building classes were 85.77% and 99.70%, respectively. The IoU value is 81.86% for the building class and 98.8% for the non-building class. For the F1 score, a value of over 98% was achieved in both classes for Area 1. The visuals of the classification results are presented in Figure 8.

Table 1. Confusion matrix of the OBIA experiment.

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>Non-Building</th>
<th>Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Building</td>
<td>509,149</td>
<td>1514</td>
</tr>
<tr>
<td>Building</td>
<td>4509</td>
<td>27,185</td>
</tr>
</tbody>
</table>

Table 2. Performance metrics of the OBIA experiment in Area 1.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>IoU (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Building</td>
<td>99.70</td>
<td>98.80</td>
<td>98.65</td>
</tr>
<tr>
<td>Building</td>
<td>85.77</td>
<td>81.86</td>
<td>98.49</td>
</tr>
<tr>
<td>Average</td>
<td>92.74</td>
<td>90.33</td>
<td>98.57</td>
</tr>
</tbody>
</table>
Area 2 has a larger area than Area 1. The confusion matrix and evaluation metrics for Area 2 are presented in Tables 3 and 4, respectively. In Area 2, an average accuracy of 92.23% was obtained, according to Table 4. While the average IoU value is 84.99%, the average F1 score is 94.60%. The non-building class has higher metrics than the building class. In this region, it has been determined that some, albeit few, non-building regions with a similar texture to the buildings are assigned as buildings. However, ignoring the details of the building edges in the ground truth also reduced the accuracy. The IoU of the non-building class reached 97.05%, while the IoU of building class was 72.92%. Visuals of the classification results are also shown in Figure 9.

Table 3. Confusion matrix of the OBIA experiment in Area 2.

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>Non-Building</th>
<th>Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Building</td>
<td>1,550,896</td>
<td>26,721</td>
</tr>
<tr>
<td>Building</td>
<td>20,406</td>
<td>126,917</td>
</tr>
</tbody>
</table>

Table 4. Performance metrics of the OBIA experiment in Area 2.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>IoU (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Building</td>
<td>98.31</td>
<td>97.05</td>
<td>95.42</td>
</tr>
<tr>
<td>Building</td>
<td>86.15</td>
<td>72.92</td>
<td>93.77</td>
</tr>
<tr>
<td>Average</td>
<td>92.23</td>
<td>84.99</td>
<td>94.60</td>
</tr>
</tbody>
</table>

Figure 8. Segmentation results of the OBIA. (a) Ground truth; (b) classification result with OBIA; (c) comparing with ground truth.

Figure 9. Segmentation results of the OBIA in Area 2. (a) Ground truth; (b) classification result with OBIA; (c) comparison with ground truth.
Area 3 is a zone with a larger number and greater variety of buildings than the other zones. As in the other regions, the non-building class has higher metrics. The accuracy for the non-building and building classes is 96.64% and 89.20%, respectively. The footprints of almost all buildings, except a few, were determined in detail. In Area 3, the average IoU metric is 84.68%, and the average F1 score is 94.71%. Assigning the mosque courtyard to the building class was a factor that reduced accuracy because the ground reflectance value and the building reflectance value are similar. The evaluation metrics of Area 3 are shown in Tables 5 and 6. Additionally, the classification results are presented in Figure 10.

Table 5. Confusion matrix of the OBIA experiment in Area 3.

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>Non-Building</th>
<th>Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonbuilding</td>
<td>536,702</td>
<td>18,658</td>
</tr>
<tr>
<td>Building</td>
<td>10,180</td>
<td>84,081</td>
</tr>
</tbody>
</table>

Table 6. Performance metrics of the OBIA experiment in Area 3.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>IoU (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Building</td>
<td>96.64</td>
<td>94.90</td>
<td>95.34</td>
</tr>
<tr>
<td>Building</td>
<td>89.20</td>
<td>74.46</td>
<td>94.08</td>
</tr>
<tr>
<td>Average</td>
<td>92.92</td>
<td>84.68</td>
<td>94.71</td>
</tr>
</tbody>
</table>

Figure 10. Segmentation results of the OBIA in Area 3. (a) Ground truth; (b) classification result with OBIA; (c) comparison with ground truth.

4.2. Floor Estimation

Sezen et al. [10] calculated the precision, recall, and mAP values obtained with YOLOv5 were 85.0%, 72.0%, and 79.0%, respectively, for their dataset (Table 7). As seen the results in Table 7, YOLOv5 successfully detects doors and windows. However, the false detection rate is higher than YOLOv4. In our study, the precision, recall, and mAP values of YOLOv5 for door and window detection are 86.4%, 71.8% and 78.7%, respectively. Detections with a confidence value of 0.5 were accepted as correct. YOLOv5 detects the windows with 90.5% mAP. The predictions of YOLOv5 were preferable for floor calculations because it produced successful results in window detection. The visuals of the window detection results are presented in Figure 11.
Table 7. Comparison of different object detection algorithms in the study of Sezen et al. [10].

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv3</td>
<td>68.0</td>
<td>82.0</td>
<td>77.0</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>75.0</td>
<td>88.0</td>
<td>84.0</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>54.0</td>
<td>63.0</td>
<td>54.0</td>
</tr>
<tr>
<td>YOLOv5</td>
<td>85.0</td>
<td>72.0</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Figure 11. Window detection results of YOLOv5.

The middle points of the bounding boxes of the windows are grouped with KDE depending on the y coordinates. Thus, the windows in the horizontal direction were counted as a group. The total number of window groups is equal to the number of floors. For this, the maximum points of the histograms produced with 1D KDE are considered. The regions where the histogram is at its maximum represent the window group. As shown in Table 8, the number of building floors was estimated correctly in 84 of 106 test images and incorrectly in 25 of them. The proposed algorithm correctly estimates the number of floors at a rate of 79.2%. The floor estimation results for the SVIs obtained with API are also shown separately in Table 9.

Table 8. Floor estimation results of the proposed algorithm for SVIs obtained with operator interpretation.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Number of Predictions</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurate prediction</td>
<td>84</td>
<td>79.2</td>
</tr>
<tr>
<td>Incorrect prediction</td>
<td>22</td>
<td>20.8</td>
</tr>
</tbody>
</table>

Table 9. Floor estimation results of the proposed algorithm for SVIs obtained with API.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Number of Predictions</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurate prediction</td>
<td>22</td>
<td>56.4</td>
</tr>
<tr>
<td>Incorrect prediction</td>
<td>17</td>
<td>43.6</td>
</tr>
</tbody>
</table>

5. Discussion

In the first step, we aimed to determine the existing building stock quickly using high-resolution satellite images. This study used VHR remote sensing images from Pleiades, with 0.5 m spatial resolution, to detect buildings. The availability of sub-meter resolution data
from HRS images has directly promoted urban feature detection [13]. Additionally, HRS images can obtain data from across large areas containing various types of urban objects with minimum effort. It enables us to analyze large areas quickly without undertaking long-term fieldwork.

Recent developments in automated image analysis from satellite images can considerably improve and accelerate the building extraction process [66]. Object-based image analysis is one of the most common techniques, which enables the use of the spectral, contextual, textual, and geometric components of image objects when performing analysis on high-resolution satellite images [67]. Our proposed OBIA-based method used the MRS method in the segmentation stage. MRS parameters are selected according to the study region. Therefore, specialist knowledge is required for selecting proper segmentation parameters.

This automatic building extraction approach produced speedy and efficient results compared to traditional methods, such as land surveying and manually digitizing building footprints. The OBIA process is divided into three subprocess: MRS, sample selection, and classification. The total processing times for Area 1, Area 2, and Area 3 were 122.3 s, 185.4 s, and 203.1 s, respectively. It should be noted that our proposed OBIA-based method does not need any extra information, such as digital elevation models. Therefore, our approach is simple and uncomplicated.

Table 1 shows that the building class is more confused than the non-building class. Confusion increases at the building edges due to the spatial resolution of the satellite images.

The importance of monitoring the buildings continues to increase day by day. Land-use changes are usually caused by expanding built-up regions and urbanization [68]. However, high-resolution images are usually not freely accessible. This situation can be considered a limitation of this study. However, the most important priority is to prevent the possible loss of life and economic damage by making immediate and efficient decisions. Compared to the damage that can be avoided, the cost of the satellite images used in the process is insignificant. In addition, with the determination of the existing building stock, updated land registry information will contribute to the economic rights arising from property rights. Therefore, an effective and efficient real estate market can be developed.

In the second stage of this study, we focused on documenting the vertical details of buildings and determining the number of building floors. This information can be used for monitoring vertical changes in building inventory and for detecting the addition of unauthorized stories. Such additions place significant loads on the existing structure and reduce its durability. This situation adversely affects the purposes of zoning plans and reduces resistance against disasters in urban areas.

There are 1045 building images in the dataset, which includes residential building types such as detached houses, apartments, and residences. We modified and enlarged a dataset prepared by Sezen et al. [10]. The training and test data were created by labeling the windows and doors of the building images collected via Google Street View. YOLOv5 was selected to extract the windows and door elements, as shown in Figure 8.

The dataset used in this study was divided into two main parts. As shown in Tables 8 and 9, the floor estimation result of the proposed algorithm for SVIs obtained by operator interpretation has higher accuracy than the dataset containing SVIs obtained automatically through the API. The perspective effect, the presence of more than one building in the photograph, obstacles around the buildings, and different window sizes cause errors in the building floor estimation stage. Manual interpretation should be implemented during the SVI collection and documentation phases to reduce possible adverse effects on the analysis results.

Accordingly, the training set needs to be developed. Considering the lack of available data in the literature, creating an appropriate training set is of great importance for the success of the experiments.

The determined midpoint coordinates were clustered along the y-axis of the image using kernel density estimation. The number of groups was accepted as the number of
floors. The proposed semi-automated approach successfully calculates the number of floors with 79.2% accuracy. This building floor estimation method is more suitable for residential buildings that have regular façade elements.

Street-view images are freely and easily accessible, especially in big cities. SVIs provide an efficient tool for obtaining vertical details of street environments. It is clear that updating the street views periodically in cities struggling with urbanization problems will contribute significantly to studies about the urban environment and analyses of building stock changes.

The SVI analysis proposed in our study can document building façades and detect critical changes that can be identified from outside. Some critical indoor alterations that weaken structures against disasters cannot be detected using our method. However, information that is missing from the public documents can be completed by comparing the output results of each step with the existing records, and urgent measures can be taken for the renewal of building stock. This proposed approach can be integrated into a GIS-based system to continuously monitor the building stock.

6. Conclusions

In the early period of the urbanization process in developing countries, increasing rural-to-urban migration caused several problems in terms of meeting people's basic needs in urban areas, such as transportation and housing. Accordingly, unplanned and uncontrolled construction processes have continued for a long time [6]. This situation poses significant risks for urban life. Developing countries should first focus on solving these problems that have been carried from the past to the present. In order to establish effective and efficient land policies to reduce the negative impacts of rapid urbanization, existing building stock should be analyzed. Additionally, the United Nations (UN) Sustainable Development Goals (SDGs) can be related to building monitoring applications, and the importance of related studies increases day by day [69]. First, there is a need for up-to-date and accurate data that reflects the actual situation. Unfortunately, public records in developing countries may not present all existing buildings. Analyses of these data yield poor results.

Establishing a comprehensive land administration system in cities to improve the quality of urban life is a complicated and lengthy process. Therefore, a two-step semi-automated building stock monitoring approach that enables quick decision making was proposed in this study. Our method automatically detects the essential elements of buildings, including the footprint (horizontal information) and the number of floors (vertical information), using remote sensing technologies and the opportunities provided by street-view web services such as Google Street View.

Accuracy analyses of each step are presented to demonstrate the potential of the method proposed in this study. Object-based image analysis techniques and the deep learning approach used in our study provided automated, fast, and accurate results. The accuracy assessment results support the applicability of our approach.

In future studies, different procedures can be investigated to collect SVIs of buildings fully automatically through the existing web mapping services by providing the image conditions necessary for detecting and analyzing essential elements of the building façade. Thus, the effectiveness and speed of the proposed approach can be further increased.


Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would thank the Implementation and Research Center for Satellite Communications and Remote Sensing (CRCS) of Istanbul Technical University for the high-resolution satellite images of Pleiades used in our study.

Conflicts of Interest: The authors declare no conflict of interest.

References


5. Iban, M.C. Lessons from approaches to informal housing and non-compliant development in Turkey: An in-depth policy analysis with a historical framework. Land Use Policy 2020, 99, 105104. [CrossRef]


11. Shukla, A.; Jain, K. Automatic extraction of urban land information from unmanned aerial vehicle (UAV) data. Earth Sci. Informatics 2020, 13, 1225–1236. [CrossRef]


17. Prathiba, A.P.; Rastogi, K.; Jain, G.V.; Govind Kumar, V.V. Building Footprint Extraction from Very-High-Resolution Satellite Image Using Object-Based Image Analysis (OBIA) Technique. In Applications of Geomatics in Civil Engineering. Lecture Notes in Civil Engineering; Springer: Singapore, 2020; Volume 33. [CrossRef]


34. Wu, D.; Gong, J.; Liang, J.; Sun, J.; Zhang, G. Analyzing the Influence of Urban Street Greening and Street Buildings on Summertime Air Pollution Based on Street View Image Data. ISPRS Int. J. Geo-Inf. 2020, 9, 500. [CrossRef]


56. Wojtaszek, M.V.; Ronczyk, L.; Mamatkulov, Z.; Reimov, M. Object-based approach for urban land cover mapping using high spatial resolution data. *E3S Web Conf.* **2021**, *227*, 01001. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.