PFG (2018) 86:235–248
https://doi.org/10.1007/s41064-018-0060-5

ORIGINAL ARTICLE

A CNN-Based Approach for Automatic Building Detection and Recognition of Roof Types Using a Single Aerial Image

Fatemeh Alidoost[1] · **Hossein Arefi**[1]

Received: 31 May 2017 / Accepted: 14 December 2018 / Published online: 15 January 2019

Abstract
Automatic detection and reconstruction of buildings have become essential in many remote sensing and computer vision applications. In this paper, the capability of Convolutional Neural Networks (CNNs) is investigated for building detection as well as recognition of roof shapes using a single image. The major steps are including training dataset generation, model training, image segmentation, building detection and roof shape recognition. First, a CNN is trained for extracting urban objects such as trees, roads and buildings. Next, classification of different roof types into flat, gable and hip shapes is performed using the second trained CNN. The assessment results prove effectiveness of the proposed method with approximately 97% and 92% of quality rates in detection and recognition steps, respectively.

Keywords Convolutional neural network (CNN) · Deep learning (DL) · 3D modelling · Fine-tuning · Pattern recognition · Selective search

Zusammenfassung
Ein CNN-basierter Ansatz zur automatischen Erkennung von Gebäuden und Dachtypen in einem einzelnen Luftbild. Die automatische Erkennung und Rekonstruktion von Gebäuden ist bei vielen Anwendungen in Fernerkundung und Computer-Vision unerlässlich geworden. In diesem Beitrag wird die Fähigkeit von Convolutional Neural Networks (CNNs) zur Erkennung von Gebäuden und Dachformen in einem einzelnen Bild untersucht. Die wichtigsten Schritte sind die Erstellung von Trainingsdatensätzen, das Modelltraining, die Bildsegmentierung sowie die Gebäude- und Dachformerkennung. Zunächst wird ein CNN für das Extrahieren von städtischen Objekten wie Bäumen, Straßen und Gebäuden trainiert und der Datensatz klassifiziert. Anschließend erfolgt die Klassifizierung der Dächer in Flach-, Giebel- und Satteldach mit dem zweiten trainierten CNN. Die Ergebnisse belegen den Erfolg der vorgeschlagenen Methode mit ca. 97% bzw. 92% Klassifizierungsgenauigkeit bei Gebäudedetektion und Klassifizierung der Dachformen.

1 Introduction

Automatic extraction and accurate localization of man-made structures from remotely sensed data such as aerial and satellite images or LiDAR data is one of the most notable challenges for a number of applications such as environmental 3D modelling, 3D visualization, building change detection, urban growing analysis, transportation, disaster recovery, resource management, tourism, and others. In such applications, a considerable number of researches are dedicated to the automatic detection and localization of buildings, as the most prominent objects in urban environments. For knowledge based 3D building reconstruction, the information about footprints or locations of the buildings as well as the type of the roof shapes are very important and the complexity of 3D modelling can be reduced if these information are provided as prior knowledge with acceptable accuracies. Additionally, the final precision of the 3D building model is directly dependent on the accuracy of detection and recognition steps. Therefore, significant attempts have been made in development of building detection algorithms through which they are robust against shadow, relief displacements, viewing direction, noise and errors in dataset. Additionally,

✉ Hossein Arefi
hossein.arefi@ut.ac.ir

Fatemeh Alidoost
falidoost@ut.ac.ir

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran

the algorithms should be capable to separate buildings from non-ground objects like vegetation covers even in low resolution data. Moreover, due to the high variety of buildings and roof shapes particularly in urban areas, and furthermore the difficulties in providing high resolution 3D data such as LiDAR data and DSMs to provide the geometrical features of buildings directly, there is a significant effort for developing automatic building detection algorithms based on aerial and satellite images as a cost effective solution.

Nowadays, Convolutional Neural Networks (CNNs) and other deep learning algorithms have shown remarkable performances in automatic object detection and recognition using close range images for various computer vision applications. The goal of the present paper is to introduce the abilities of CNNs in automatic building detection—besides the recognition of roof shapes—using a single aerial image in a sequential framework. This approach is based on CNN training and the knowledge transferring concept using training datasets including urban objects such as buildings, roads, trees and different classes of roof shapes. The final results of our study are the locations of buildings surrounded by bounding boxes and the type of the roof shapes which are extracted from a single image automatically, unlike semi-automatic methods (Nalani 2014). These information are extremely important for many applications, particularly for 3D building reconstruction using model-based techniques. In summary, this paper contributes to the literature in three major aspects:

- In this paper, one of the most common challenges in the remote sensing community, namely building detection, is addressed based on an object-level detection algorithm [unlike pixel-level classification approaches (Saito and Aoki 2015)].
- A CNN-based method is proposed for automatic roof shape recognition of buildings from a single monocular remote sensing image which is an essential knowledge for 3D CAD model generation.
- A training dataset[1] including different urban objects are generated which could be employed in further CNNs-based object detection applications.

2 Related Work

There is extensive literature on building detection and extraction. In the present section a part of the related work is considered in short, and divided into the three general groups: segmentation-based, classification-based, and hybrid methods, i.e., classification following the segmentation-based methods.

There are numerous segmentation-based methods and algorithms for building detection and extraction. They are achieved by the application of thresholding, mathematical morphology and connected component labelling techniques (Maas and Vosselman 1999; Vu et al. 2009; Yu et al. 2010), the geometric criteria for grouping the planar and non-planar points (Gamba and Houshmand 2000), the combination of eigen analysis and the fuzzy *k*-means algorithm (Sampath and Shan 2010), and different geometrical and spectral features such as normal and roughness values of point cloud and the normalized differential vegetation index (NDVI) of images (Maltezos and Ioannidis 2015). In some approaches, a region-growing algorithm (Ballard and Brown 1982) is used to detect building segments (Zhang et al. 2006; Dorninger and Pfeifer 2008; Cheng et al. 2011). Instead of regions, the image edges are extracted by a gradient-based Line Segment Detection (LSD) algorithm (Von Gioi et al. 2010) and building edges are separated from other objects using characteristics of the building shadows (Singh et al. 2015). Furthermore, energy optimization methods are employed for building using the normal vectors (Kim and Shan 2011), an improved snake model (Kabolizade et al. 2010) or the graph-based algorithm named "Grab cut" (Khurana and Wadhwa 2015). Other methods and algorithms for building detection and extraction can be grouped into the classification-based methods such as the ISODATA algorithm (Haala and Brenner 1999), the Dempster Shafer theory (Rottensteiner et al. 2007), the Random Forests (RF) algorithm (Guo et al. 2011), and the object-based classification methods (Hermosilla et al. 2011). One of the state-of-the-art supervised classification algorithms is based on Deep Convolutional Neural Networks (DCNNs), which are widely employed in many computer vision and machine learning applications, for instance pixel level classification, automatic object detection and object recognition (Bengio 2009; Deng et al. 2009; Yu et al. 2010; Zeiler and Fergus 2014; Liu et al. 2015). Modern object detection techniques based on deep learning algorithms can be categorized as classification and regression-based methods. The classification-based methods include an additional step such as sliding window or selective search operations to generate the region candidates while in the regression-based methods the proposal generation stage is eliminated and the network takes an input image and learns the class probabilities and bounding box coordinates, directly. There is a trade-off between accuracy and processing speed for comparison of two categories. Currently, the region-based detectors including R-CNNs (Girshick et al. 2014) as well as different variations of it, such as Fast R-CNNs (Girshick 2015), Faster R-CNNs (Ren et al. 2016) and Mask R-CNNs (He et al. 2017) are the most accurate detectors which are less sensitive to the overfitting

[1] https://github.com/loosgagnet/Building-detection-and-roof-type-recognition.

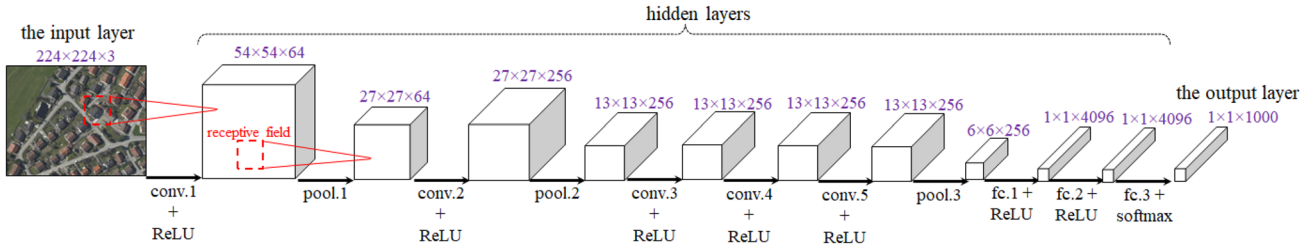


Fig. 1 Common architecture of CNN, namely VGG-F

problem. Moreover, these detectors are slower than regression-based solutions such as You Only Look Once (YOLO) (Redmon et al. 2016) and the Single Shot MultiBox Detector (SSD) (Liu et al. 2016) which can be utilized in real-time object detection. R-CNNs employ the selective search method to localize the region candidates. These regions are then fed into a CNN to extract corresponding features for each regions. Next, a binary SVM is trained on these features and finally, a regression model is also trained to reduce the localization errors of regions. Because of separate computation for CNN, SVM and regression models, R-CNNs are expensive and slow. Fast R-CNNs improve the speed of R-CNNs by sharing computation. In Fast R-CNNs, region proposals are extracted using the selective search method, similar to R-CNNs. Next, unlike R-CNNs, all regions of an image are aggregated as one region and then, this region is just fed into a CNN including a ROI (Region of Interest) pooling layer. Consequently, the feature vectors of regions are extracted by the ROI pooling layer. Finally, the softmax classifier and regression models are employed in a similar procedure as in R-CNNs. The ROI pooling layer improves the speed of R-CNNs. The Faster R-CNNs are composed of two networks as a Fast R-CNN and a Region Proposal Network (RPN) to improve the effect and the efficiency of detection. Instead of selective search, the RPN is used based on sliding windows to generate region candidates with multiple scales and aspect ratios and to achieve better results. Mask R-CNN is composed of two networks as a Fully Connected Network (FCN) and a Region Proposal Network (RPN). It is an extension of Faster R-CNNs for both object detection and pixel-level image segmentation. The FCN is a segmentation model to predict the segmented mask for each region. Instead of ROI pooling layer in Faster R-CNNs, a RoIAlign layer is used in Mask R-CNNs to improve the accuracy of aligning the regions to objects. YOLO detector is a regression-based method for object detection by looking at the entire image once during network training and assigning image pixels to coordinates of ROIs, directly. Since there is no need to extract region candidates as a primary step, the speed and efficiency of YOLO are higher than R-CNN-based methods. On the other hand, the detection accuracy of YOLO is less than R-CNNs especially for irregularly shaped objects or a group of small objects. The SSD uses the VGG-16 (Visual Geometry Group) network and produces a set of bounding boxes and scores them based on the presence of object class instances in those boxes. Additionally, the proposal generation and feature resampling stages are eliminated from processing steps of SSD. Since SSD is faster than YOLO and as accurate as Faster RCNN (Liu et al. 2016), it achieves a good balance between speed and accuracy. A comparison between classification and regression-based methods can be found in the Google research paper (Huang et al. 2017). In remote sensing application, the CNNs can be utilized to extract the building and non-building regions automatically (Makantasis et al. 2015; Saito and Aoki 2015; Vakalopoulou et al. 2015; Alidoost and Arefi 2016; Yuan 2016; Zhang et al. 2016). Lastly, in the hybrid methods for building detection, one can apply classification algorithms on segmentation results to label each segment based on different land cover classes. For instance, classification of non-ground points using image edge segments (Awrangjeb et al. 2013), the application of the region growing segmentation and an object-based classification algorithms (Höfle et al. 2009), the multi-region graph cut image segmentation and a rule-based classification algorithms (Karantzalos et al. 2015), as well as Markov Random Field (MRF) segmentation and fuzzy *k*-NN classification algorithms (Oztimur Karadag et al. 2015).

Several methods are available for extracting buildings from a single image. Some of these recent algorithms include an energy-based optimization algorithm using the Local Gradient Orientation Density (LGOD) (Benedek et al. 2012), a graph-based algorithm using shadow information of buildings (Izadi and Saeedi 2012; Ok et al. 2013), a combination of the *k*-means clustering algorithm and a Purposive FastICA (Fast Independent Component Analysis) model (Ghaffarian and Ghaffarian 2014), the multi label graph partitioning strategy (Manno-Kovacs and Ok 2015), a combination of Gaussian Mixture Model (GMM) clustering and Conditional Random Field (CRF) classification algorithms (Li et al. 2015), a self-supervised decision fusion framework (Senaras and Vural 2016), and a supervised segmentation algorithm based on the image descriptors (Dornaika et al. 2016).

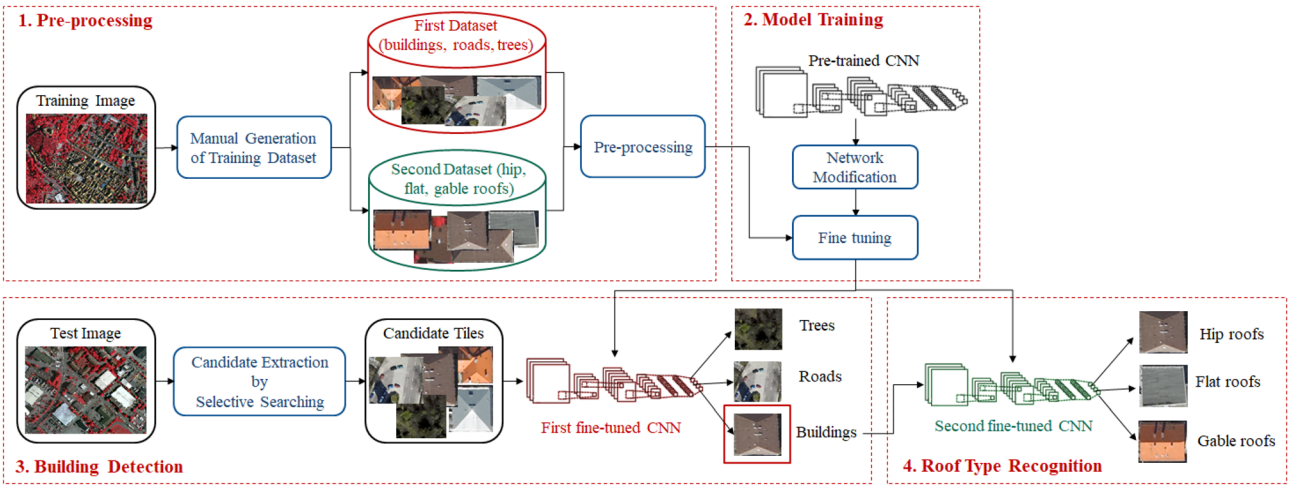


Fig. 2 The flowchart of proposed method

3 Convolutional Neural Network

Deep Learning (DL) is one of the subfields of machine learning and Artificial Neural Networks (ANNs). It is based on hierarchical learning representations of data using a deep structure composed of different hidden layers (Bengio 2009; Deng and Yu 2014; Schmidhuber 2015). DL is currently used in a variety of applications for instance object recognition (Zuo and Wang 2014), object detection (Girshick et al. 2015; Zhang et al. 2015), image classification (Chen et al. 2015; Tuia et al. 2015), and others. The Convolutional Neural Network (CNN) is a type of deep learning approach, which was inspired by biological processes (LeCun and Bengio 1995). As seen in Fig. 1, a common structure of CNN, which is employed in this study, includes an input layer (e.g., an image matrix), hidden layers, and an output layer (e.g., a fixed length vector of class scores). The hidden layers include the convolutional layers, the sampling layers (or in other words the pooling layers) and the fully connected layers (Phung and Bouzerdoum 2009; Liu et al. 2015).

There are two general strategies for training a CNN (Girshick et al. 2015; Maitra et al. 2015). As a first strategy, a CNN is trained based on the random initialization of weight parameters. This training method requires enough training data as well as a large amount of memory and its convergence takes a vast amount of time to provide an acceptable accuracy, i.e., 2–3 weeks for modern CNNs. In many classification applications, it is relatively rare to have a large training dataset. Therefore, to prevent the overfitting problem an existing pre-trained network is employed and its parameters are fine-tuned for a small dataset based on the concept of transfer learning (Bengio 2012; Donahue et al. 2014). A pre-trained network has been trained on a large dataset with approximately several million images using different toolboxes over several weeks using multiple GPUs. As shown in Fig. 1, the VGG-F network (Chatfield et al. 2014; Simonyan and Zisserman 2015) is a pre-trained network which is trained on the ImageNet ILSVRC benchmark dataset (Deng et al. 2009) of 1000 different categories using MatConvNet toolbox (Vedaldi and Lenc 2015).

There are two common transfer learning techniques for fine-tuning a pre-trained network. In the first technique, the last fully connected layer of the pre-trained model is replaced with a fully connected layer which is relevant to the new classification problem (the object classes of the new problem are considered in the new fully connected layer). Next, the rest of the network is used as a fixed feature extractor to extract features using the new dataset and a classifier such as the linear Support Vector Machine (SVM) or a logistic regression is trained on top of the network on the new dataset. In the second technique, in addition to replacing the last fully connected layer and retraining the classifier, the weights of all layers or just higher-level layers are also fine-tuned using a back propagation algorithm such as the Stochastic Gradient Descent (SGD) (Girshick et al. 2015).

4 Proposed Method

In the present paper, a sequential approach is proposed for automatic building detection and for recognition of roof types based on a CNN classification framework using intensity information as the spectral image channels exclusively. Due to high intra-class variability (e.g., the flat, gable and hip roofs from the building class) and high inter-class similarity (e.g., the flat roofs and roads classes or the roads with clear centreline and the gable roofs classes), two networks

are trained and used sequentially to decrease the complexity of and to improve the performance of the CNN. The first network is employed to classify objects such as buildings, roads, and trees in the detection step and the second network is utilized to classify the flat, gable and hip roofs in the recognition step. The main steps of the proposed approach are as follows (cf. Fig. 2):

1. Pre-processing;
2. Model training;
3. Building detection;
4. Roof type recognition.

The summary of each step and their main components are given in the following sub-sections.

4.1 Pre-Processing

As shown in Fig. 2, in the pre-processing step, two training datasets are generated by manually cropping the training image. The first dataset includes image tiles per class of buildings, roads, and trees for the detection step; and the second dataset includes image tiles per class of the flat, gable and hip roofs for the recognition step. Each class contains several image tiles centered on top view of the objects. Before training the network, additional processing steps should be applied to the cropped image tiles as resizing, normalization, mean subtraction, and data augmentation. Since the input layer of the proposed CNN is an image with the size of $224 \times 224 \times 3$, all image tiles are resized. Another important pre-processing step is normalization, which refers to applying a common radiometric scale (or a unit) to the input data with different scales in pixel values. There are various normalization methods in statistics but in the case of optical images, the relative scales of pixels should be equal for different input layers. Therefore, all resized image tiles are normalized to the range of 0–255 in the present study. Mean subtraction is the most common form of pre-processing, which has the geometric interpretation of centering the cloud of data around the center along every dimension. By subtracting the mean per channel from all scaled image tiles the data becomes zero-centered in each dimension. On the other hand, the CNN has several million parameters and the size of the generated training datasets is insufficient to learn all of these parameters without a considerable overfitting error. Therefore, the data augmentation (Krizhevsky et al. 2012) is applied as the easiest method to reduce the overfitting problem and carried out by enlarging two training datasets artificially.

4.2 Model Training

In the model training step, a CNN-based on the VGG-F architecture (Fig. 1) is used as a pre-trained network. Unlike objects in computer vision with high inter-class variability, the buildings with different roof shapes in aerial images (top view images) have high inter-class similarity (for example, the differences between gable and hip roofs or between gable and flat roofs in aerial images are just a few linear features). Therefore, the extracted features of different buildings need to be distinguishable enough. Among many choices for a CNN architecture, the VGG-F architecture is a simple one because only 3×3 convolution and 2×2 pooling layers are used throughout the whole network. Moreover, the VGG-F-based feature extractor is deep enough to extract complex features and the combination of such features from aerial images. In this paper, the given CNN is fine-tuned using two training datasets to adapt the parameters for building detection and roof shape recognition application. For fine-tuning the network, the last fully connected layer is removed and a layer with random weights is replaced based on the classes of each training dataset. At the first step of fine-tuning all layers except the last fully connected layer are frozen (i.e., the weights are not updated) and the network is trained for a while. Then, all layers except all fully connected layers are frozen and the network is trained for another few epochs. Finally, the network is trained with all layers, jointly. During fine-tuning, the weights are adjusted subsequently by continuing the back propagation based on the minibatch SGD algorithm with momentum which only takes into account the first derivative when performing the updates on the weights. Additionally, the softmax log function is applied as the loss function. Therefore, through this step two kinds of fine-tuned networks are generated: the first fine-tuned network including three classes of objects (i.e., buildings, roads, and trees) for the detection step and the second fine-tuned network including three classes of roof shapes (flat, gable and hip) for the recognition step.

4.3 Building Detection

Through the building detection step, a test image is selected outside the training area. As illustrated in Fig. 2, the selective search algorithm is employed for automatically extracting candidate regions from the test image at all scales. This algorithm is based on a graph-based segmentation method (Felzenszwalb and Huttenlocher 2004) and combines the strength of both exhaustive search and segmentation (Uijlings et al. 2013). After applying the selective search algorithm to the test image, the extracted candidate regions are fed into the first fine-tuned network and subsequently the

Table 1 The accuracy of networks on validation sets

Step	Image no.	Class no.	Time (h)	Obj.	Err.	Acc. (%)
Detection	2400	3	5.6	0.021	0.004	99.6
Recognition	2400	3	5.6	0.343	0.022	97.8

candidate regions' score matrix is calculated. The dimension of the score matrix is $S \times C$, where S and C stand for the number of regions and the number of classes, respectively. Since the extracted features are outstandingly distinguishable, by applying a maximum operation on each row of the score matrix, the class label for each region corresponding to maximum score values can be defined and all of the regions are classified into buildings, roads and trees classes.

4.4 Roof Type Recognition

After separating all building regions from other objects in the previous step, the optimum bounding box of each building can be generated based on the minimum bounding rectangle. In the recognition step, the building regions are fed to the second fine-tuned network to define the roof shape of each building as flat, gable, and hip classes (Fig. 2). The output of the network is a $S \times C$ score matrix, where S is the number of building regions and C is the number of classes. If the region belongs to a specific class of roofs, the score of the class is maximum value and the optimal label of the region can be defined.

5 Experiments and Results

In this study, two non-overlapping scenes of a single aerial mosaic ortho-photo of Vaihingen, Germany (ISPRS 2012) with three channels are employed as the training and testing images for evaluation of the proposed method performance. The aerial images are a subset of the data used for testing digital aerial cameras carried out by the German Association of Photogrammetry, Remote Sensing, and Geoinformation (DGPF) (Cramer 2010). This dataset contains 20 pan-sharpened colour infrared (CIR) images with a ground sampling distance (GSD) of 8 cm. In addition, the second dataset from Potsdam, Germany (ISPRS 2012) consisting of a true ortho-photo with a GSD of 5 cm is employed to assess the transferability of the trained networks. The Vaihingen and Potsdam test images are composed of about 63 and 93 buildings, respectively. The training image is cropped manually into 200 tiles per class of buildings, roads, and trees for the detection step; and 200 tiles per class of the flat, gable and hip roofs for the recognition step. In this study, the data augmentation is conducted by flipping the image horizontally and vertically, and by rotating by $-90°$ and $90°$.

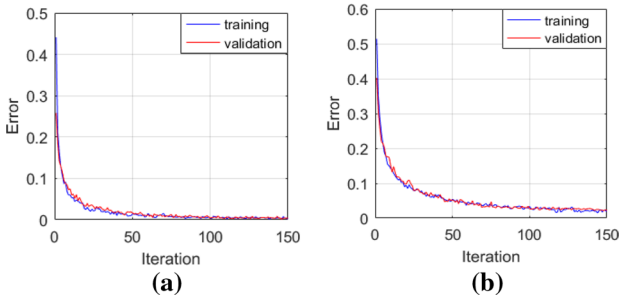


Fig. 3 The convergence graphs of training for the detection step (**a**), and recognition step (**b**)

Fig. 4 Errors in the training data for building detection

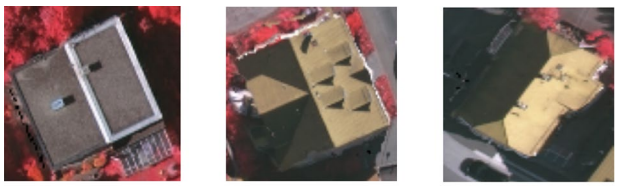


Fig. 5 Errors in the training data for roof type recognition

After the data augmentation process, each generated training dataset is containing 800 tiles per class. The network is trained using MatConvNet on a single NVIDIA GeForce 840 M with 2 GB of GPU memory and with a batch size of 100 for 150 epochs for the detection and recognition tasks, separately. The learning rate and the momentum are about 0.01 and 0.9, respectively.

In this study, two training datasets are generated manually. Therefore, the performance of the trained networks is first evaluated on the training datasets by reporting the accuracy on validation set (Table 1). Table 1 presents the quantitative parameters for the training step such as the learning

time, the error and accuracy for the last iteration and for the detection and recognition steps. In Table 1 the error on the validation set represents the frequency when the highest scoring estimation is wrong. The accuracy, in percent, would then be: 100 × (1 − error on the validation set) and used to describe how good the hyper parameters, the network structure as well as the training datasets are selected for this application. Moreover, Fig. 3 shows the convergence graphs of trained models after 150 iterations.

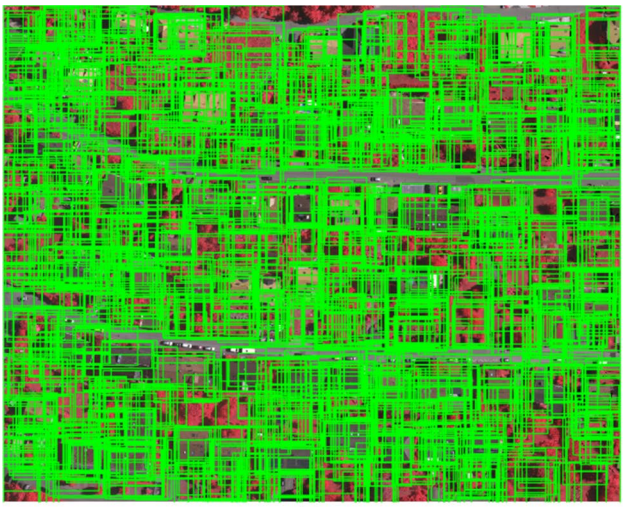
According to Table 1, the error on the validation set appears to be very low in both detection and recognition steps. In the detection step, due to the different nature of objects (buildings, roads and trees) in the first training dataset, the extracted features are more discriminative, and therefore, the accuracy of the trained model on the validation set is about 99.6%. It is also higher than this value for the recognition step. The sources of these errors are shown in Figs. 4 and 5. In the left image (Fig. 4a), the edge lines between vegetation, road and shadow areas act as the edge lines in a gable roof, so there is a similarity between the extracted features of this image and other building image tiles and the score of this image is high in the building class instead of the road class. In the right image (Fig. 4b), there are building structures on both sides of the image tile and there is a road between them, thus this image tile appears as a flat building. Therefore, this image is incorrectly labelled as a building.

There are similar errors in the training dataset for the recognition step (Fig. 5). For example, two neighbouring flat buildings in the left image (Fig. 5a) appear as gable buildings, which the network classifies it as a gable roof by mistake. Moreover, the small structures on the roof can make learning of the gable roof difficult (as seen in Fig. 5b). Since the variety of roof shape in a specific class is not high in the second training dataset, the network cannot be trained correctly and therefore, some buildings like the sample shown in the right image (Fig. 5c) are classified into the wrong class.

After evaluating the performance of trained networks on validation datasets, their performances are assessed on testing datasets as follows.

In the building detection step, the number of extracted regions by the selective search algorithm are about 26,731 and 23,822 for Vaihingen and Potsdam test images, respectively. The results of applying the selective search algorithm on the test images and the final results of the detected buildings with modified bounding boxes are shown in Figs. 6 and 7.

The standard quality measures of Completeness (or Recall), Correctness (or Precision), and Quality (McGlone and Shufelt 1994; McKeown et al. 2000) have been calculated as given in Eq. (1).



(a)

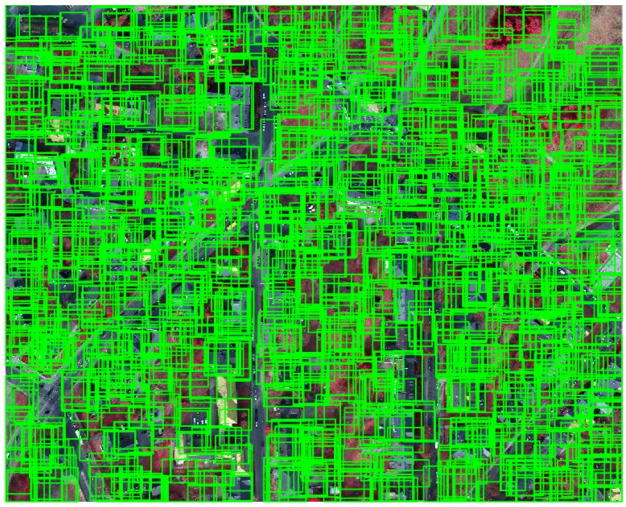
(b)

Fig. 6 The extracted bounding boxes by selective search algorithm (**a**) and detected buildings (**b**) for the Vaihingen test image

$$\text{Comp.} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \text{Corr.} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \text{Qual.} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \quad (1)$$

where *TP* is True Positive, *FP* is False Positive, and *FN* is False Negative.

In this study, the number of correctly detected buildings (*TP*), the number of non-building objects detected as buildings (*FP*) and the number of undetected buildings or detected buildings as non-building objects (*FN*) as well as completeness, correctness, and quality values of building detection are presented in Table 2 for both test areas.



(a)

(b)

Fig. 7 The extracted bounding boxes by selective search algorithm (**a**) and the detected buildings (**b**) for the Potsdam test image

For the Potsdam dataset, the completely detected buildings are just considered as True Positive samples in Table 2. Therefore, the number of undetected buildings is 4, as shown in Fig. 8. Moreover, the *TP*, *FP* and *FN* as well as Completeness, Correctness, and Quality values for roof shape recognition are given in Tables 3 and 4 for Vaihingen and Potsdam test images, respectively. In Table 4, the hip and the half-hip buildings are classified into the hip class. In addition, buildings with complex or non-clear roof shapes as well as roofs, which are not included in the second training dataset (Fig. 9), are not considered in Table 4.

Based on the results of the Vaihingen test image as shown in Figs. 10 and 11, the objects that are incorrectly detected as the buildings in the detection step have a direct effect on the recognition results (the case I in Fig. 10). Furthermore, the buildings with small structures on their roofs should be considered in the gable class in the training step, so that the trained network is able to recognize such buildings as gable roofs and separate them from hip buildings (the case II in Fig. 10). For detecting the complex buildings like cases III and IV, the training dataset should be composed of different classes of buildings, otherwise, the class scores are calculated for these complex buildings based on their maximum similarity to other available classes. The Potsdam test image consists of many complex roof shapes (the case I in Fig. 12) or roof types which are not defined in the roofs' training dataset such as the mansard roofs (the case IV in Fig. 12). As shown in Fig. 8, if the complete region of a building is not extracted through the segmentation step, the accuracy of building detection could be decreased. Additionally, the bounding boxes which they are larger or smaller than the building (e.g., the case II in Fig. 10 and the case III in Fig. 12) have negative effects on the results. On the other hand, the focus of this paper is to recognize three classes of gable, hip and flat roofs. Therefore, all of complex buildings are classified into one of these classes. Moreover, some non-building structures are classified into the building class (the case II in Fig. 12). Additionally, the large buildings are not detected completely (the case III in Fig. 12). In Figs. 11 and 13 the results of the segmentation, detection and recognition steps are zoomed for some example areas. It could also be concluded that there are some factors having direct effects on the final quality of detection and recognition results as follows:

- The generated training dataset should ideally include different types of objects and roofs from different views and with different sizes and shapes as positive and negative training samples as much as possible. It is crucial to train a flexible and robust network which has a high sensitivity for distinguishing between similar but different objects, e.g., between flat buildings and roads.
- Since the segmentation results are shown to have a direct effect on the final detection performance as well as the

Table 2 Quantitative evaluation results for building detection

Dataset	TP	FP	FN	Comp. (%)	Corr. (%)	Qual. (%)
Vaihingen	63	1	0	100	98.4	98.4
Potsdam	89	0	4	95.7	100	95.7

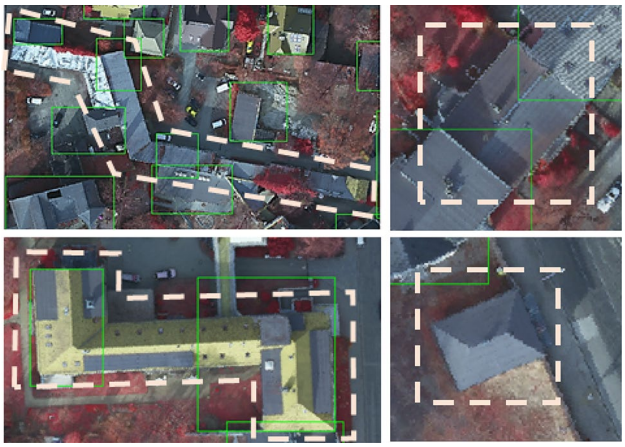


Fig. 8 The False Negative samples for the Potsdam test image

Fig. 9 Buildings with complex roof shapes

final recognition results, usage of additional data such as digital surface models (DSMs) could improve the results significantly, and could remove the unwanted areas in each building bounding box. On the other hand, the exact segments of the combined buildings, e.g., the half gable and half-hip buildings, should be extracted during the segmentation step to increase the accuracy of recognition.

6 Conclusion and Discussion

In this paper, a supervised, automatic method for building detection and roof shapes recognition from a single aerial image is proposed using a CNN-based approach. Based on the qualitative and quantitative assessments, the proposed method is shown to be capable of detecting buildings and identifying the types of roof shapes from a single aerial image based on CNNs with a quality rate of approximately 97% and 92% in the detection and recognition steps, respectively. Despite training the CNN from scratch (Yuan 2016; Zhang et al. 2016) with a large training time [i.e., about 2 days (Yuan 2016)], the proposed method has the benefit of

Table 3 Quantitative evaluation results for roof shape recognition in the Vaihingen test image

Roof class	TP	FP	FN	Comp. (%)	Corr. (%)	Qual. (%)
Flat	3	0	0	100	100	100
Gable	56	2	1	98.2	96.5	94.9
Hip	2	1	1	66.7	66.7	50.0
All classes	61	3	2	96.8	95.3	92.4

Table 4 Quantitative evaluation results for roof shape recognition in the Potsdam test image

Roof class	TP	FP	FN	Comp. (%)	Corr. (%)	Qual. (%)
Flat	5	1	0	100	83.3	83.3
Gable	59	1	2	98.3	96.7	93.6
Hip	21	2	1	95.5	91.3	87.5
All classes	85	4	3	96.6	95.5	92.4

Fig. 10 The final result of rooftop model recognition for the flat (blue boxes), gable (yellow boxes) and hip (green boxes) rooftop models for the Vaihingen test image, the False Positive and the False Negative samples are shown with pink crosses

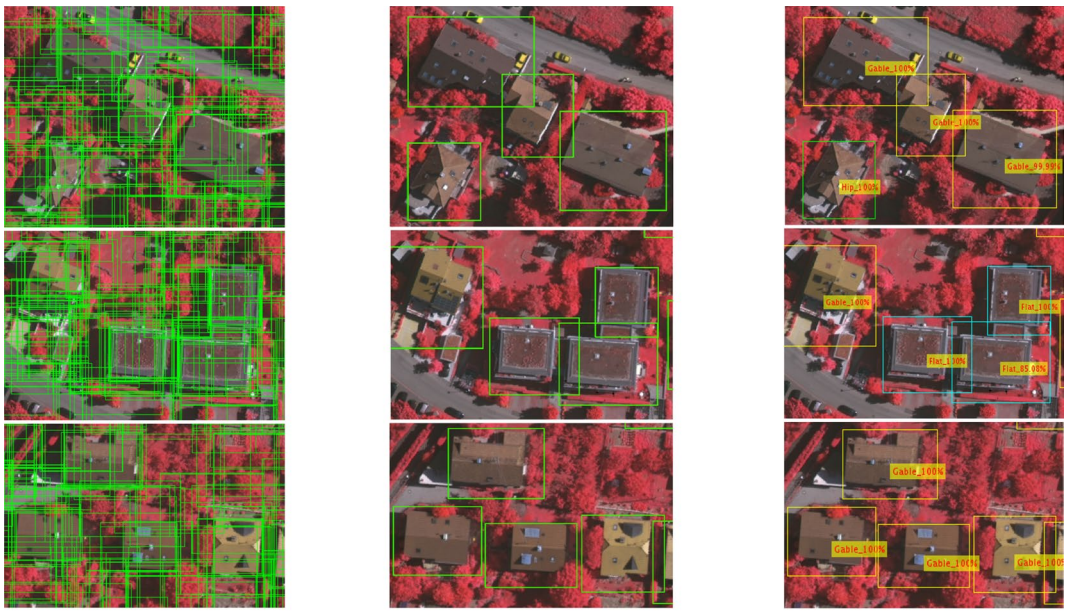


Fig. 11 The result of selective search algorithm (**a**), building detection (**b**) and recognition of rooftop models (**c**) for the Vaihingen test image

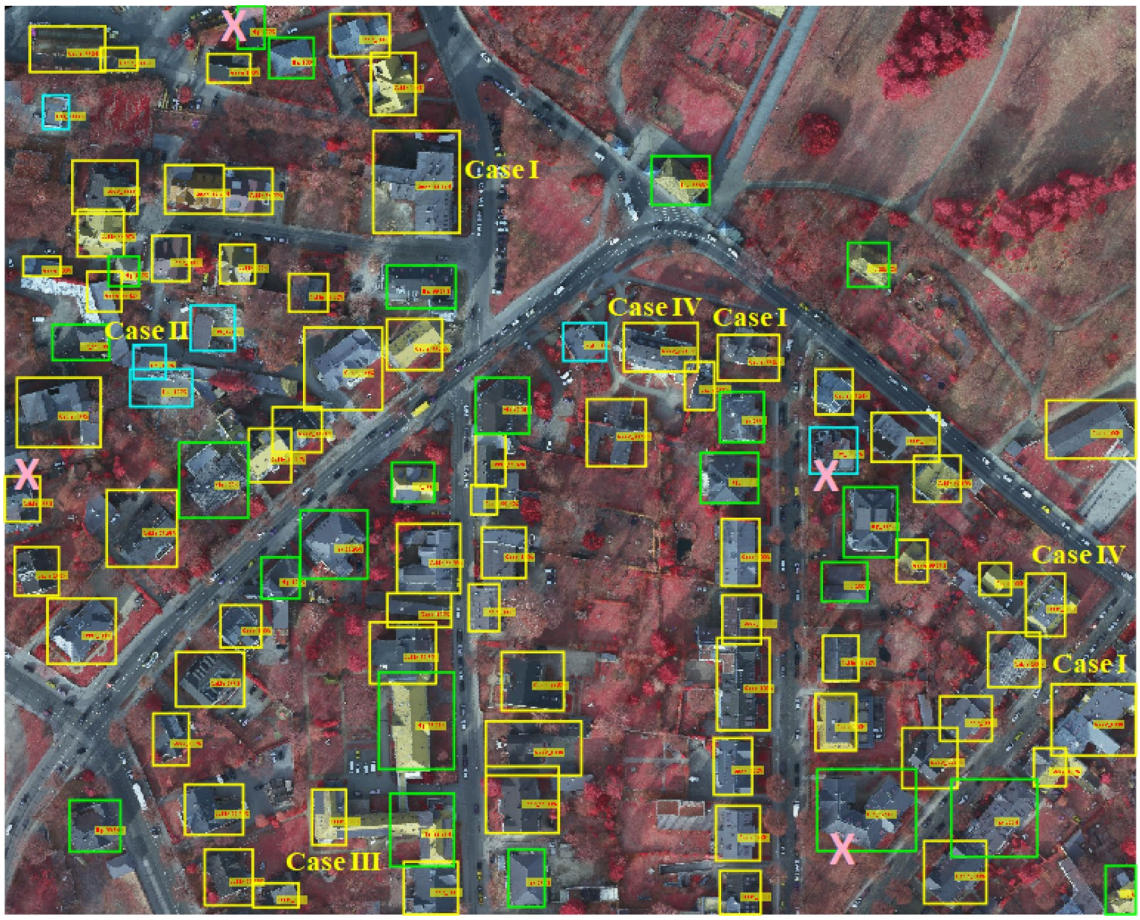


Fig. 12 The final result of rooftop model recognition for the flat (blue boxes), the gable (yellow boxes) and the hip (green boxes) rooftop models for the Potsdam test image, the False Positive and the False Negative samples are shown with pink crosses

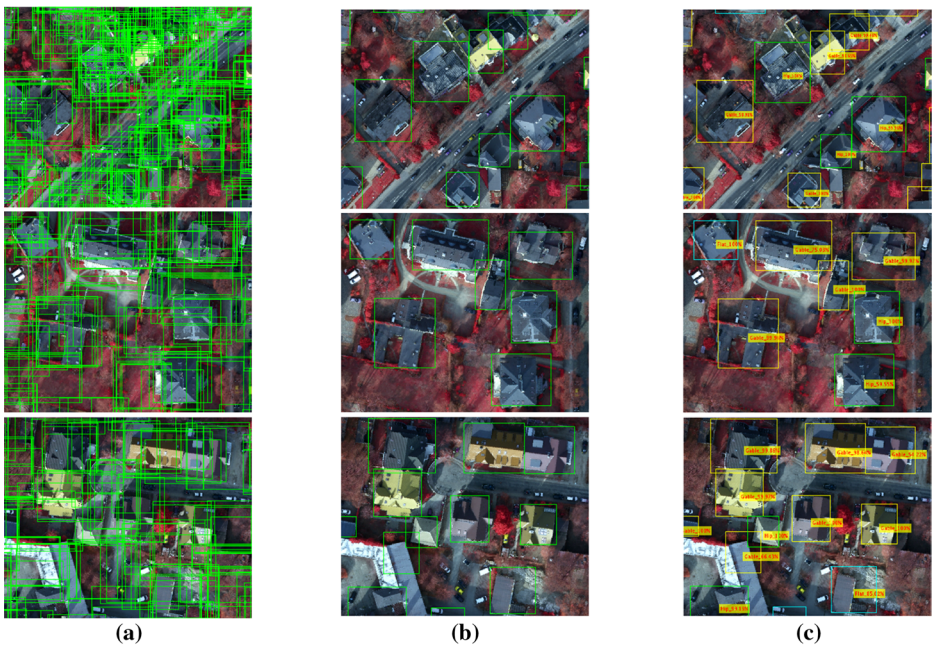


Fig. 13 The result of selective search algorithm (**a**), building detection (**b**) and recognition of rooftop models (**c**) for the Potsdam test image

the transfer learning techniques with a low computational cost. Moreover, in comparison with Zhang's study results (Zhang et al. 2016), the results of the building detection achieved by the proposed method do not depend on the complexity level and distribution of buildings. Based on the detection results shown in Figs. 6 and 7, the proposed method is also robust against areas occluded by trees and has good performance with a small training dataset (about 2400 image tiles), in contrast to the large or very large training datasets employed in previous studies (Vakalopoulou et al. 2015; Yuan 2016; Zhang et al. 2016). In the proposed algorithm by Yuan (2016), it is assumed that the building footprints are available from a GIS database, and in the algorithm reported by Zhang et al. 2016, the fixed size candidate images are extracted using the sliding windows algorithm with large numbers of non-optimal bounding boxes (Uijlings et al. 2013). The other advantages of the proposed method are the capacity to extract the bounding boxes of buildings automatically without additional information, to extract bounding boxes with different sizes related to building sizes, and to use the selective search algorithm to decrease the number of unnecessary extracted bounding boxes of objects. However, for a comprehensive evaluation and efficiency analysis of the proposed algorithm, a wider area containing buildings with more complex roofs should be considered for future experiments. Furthermore, the important focus in the future work could be on improving the automatic segmentation results as well as the automatic training dataset generation using of unsupervised learning techniques.

Acknowledgements The Vaihingen and Potsdam data sets are provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (ISPRS 2012; Cramer 2010) which is acknowledged by authors.

References

Alidoost F, Arefi H (2016) Knowledge based 3D building model recognition using convolutional neural networks from lidar and aerial imageries. Int Arch Photogramm Remote Sens Spat Inf Sci XLI-B3:833–840. https://doi.org/10.5194/isprsarchives-xli-b3-833-2016

Awrangjeb M, Zhang C, Fraser CS (2013) Automatic extraction of building roofs using lidar data and multispectral imagery. ISPRS J Photogramm Remote Sens 83:1–18. https://doi.org/10.1016/j.isprsjprs.2013.05.006

Ballard DH, Brown CM (1982) Computer vision. Prentice-Hall Inc, New Jersey

Benedek C, Descombes X, Zerubia J (2012) Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Trans Pattern Anal Mach Intell 34(1):33–50. https://doi.org/10.1109/TPAMI.2011.94

Bengio Y (2009) Learning deep architectures for AI. Found Trends® Mach Learn 2(1):1–127. https://doi.org/10.1561/2200000006.

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. JMLR 27:17–37

Chatfield K, Simoyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. Proc B Mach Vision Conf. arXiv:1405.3531

Chen Y, Zhao X, Jia X et al (2015) Spectral-spatial classification of hyperspectral data based on deep belief network. IEEE J Sel Top Appl Earth Obs Remote Sens 8(6):2381–2392. https://doi.org/10.1109/JSTARS.2015.2388577

Cheng L, Gong J, Li M, Liu Y (2011) 3D building model reconstruction from multi-view aerial imagery and lidar data. Photogramm Eng Remote Sens 77(2):125–139. https://doi.org/10.14358/PERS.77.2.125

Cramer M (2010) The DGPF-test on digital airborne camera evaluation—overview and test design. Photogramm Fernerkundung Geoinf 2010:73–82. https://doi.org/10.1127/1432-8364/2010/0041

Deng L, Yu D (2014) Deep learning: methods and applications. Found Trends® Signal Process 7(3–4):197–387. https://doi.org/10.1136/bmj.319.7209.0a

Deng J, Dong W, Socher R et al (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR2009). IEEE, Miami, FL, USA, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Donahue J, Jia Y, Vinyals O et al (2014) DeCAF: a deep convolutional activation feature for generic visual recognition. Proc 31st Int Conf Mach Learn, PMLR 32(1):647–655

Dornaika F, Moujahid A, El Merabet Y, Ruichek Y (2016) Building detection from orthophotos using a machine learning approach: an empirical study on image segmentation and descriptors. Expert Syst Appl 58:130–142. https://doi.org/10.1016/j.eswa.2016.03.024

Dorninger P, Pfeifer N (2008) A comprehensive automated 3D approach for building extraction, reconstruction, and regularization from airborne laser scanning point clouds. Sensors 8:7323–7343. https://doi.org/10.3390/s8117323

Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comp Vision 59(2):167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77

Gamba P, Houshmand B (2000) Digital surface models and building extraction: a comparison of IFSAR and LIDAR data. IEEE Trans Geosci Remote Sens 38(4):1959–1968. https://doi.org/10.1109/36.851777

Ghaffarian S, Ghaffarian S (2014) Automatic building detection based on purposive FastICA (PFICA) algorithm using monocular high resolution google earth images. ISPRS J Photogramm Remote Sens 97:152–159. https://doi.org/10.1016/j.isprsjprs.2014.08.017

Girshick R (2015) Fast R-CNN. In: Proceeding of IEEE conference on computer vision and pattern recognition (CVPR2014). IEEE, Santiago, Chile, pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR2014). IEEE, Columbus, Ohio, pp 580–587. https://doi.org/10.1109/CVPR.2014.81

Girshick R, Donahue J, Darrell T, Malik J (2015) Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 38(1):142–158. https://doi.org/10.1109/TPAMI.2015.2437384

Guo L, Chehata N, Mallet C, Boukir S (2011) Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS J Photogramm Remote Sens 66:56–66. https://doi.org/10.1016/j.isprsjprs.2010.08.007

Haala N, Brenner C (1999) Extraction of buildings and trees in urban environments. ISPRS J Photogramm Remote Sens 54:130–137. https://doi.org/10.1016/S0924-2716(99)00010-6

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of IEEE international conference on computer vision (ICCV2017). IEEE, Venice, Italy, pp 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hermosilla T, Ruiz LA, Recio JA, Estornell J (2011) Evaluation of automatic building detection approaches combining high resolution images and lidar data. Remote Sens 3:1188–1210. https://doi.org/10.3390/rs3061188

Höfle B, Mücke W, Dutter M, Rutzinger M (2009) Detection of building regions using airborne lidar: a new combination of raster and point cloud based GIS methods. Proc Geoinformatics Forum Salzburg. pp 66–75. https://ezproxy2.utwente.nl/login?url=https://webapps.itc.utwente.nl/library/2009/chap/rutzinger_det.pdf. Accessed 15 Jan 2017

Huang J, Rathod V, Sun C et al (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR2017). IEEE, Honolulu, HI, USA, pp 3296–3297. https://doi.org/10.1109/CVPR.2017.351

ISPRS (2012) Web site of the ISPRS test project on urban classification and 3D building reconstruction. Available at http://www2.isprs.org/commissions/comm3/wg4/detection-and-reconstruction.html. Accessed 17 Sep. 2016

Izadi M, Saeedi P (2012) Three-dimensional polygonal building model estimation from single satellite images. Geosci Remote Sens IEEE Trans 50(6):2254–2272. https://doi.org/10.1109/TGRS.2011.2172995

Kabolizade M, Ebadi H, Ahmadi S (2010) An improved snake model for automatic extraction of buildings from urban aerial images and lidar data. Comput Environ Urban Syst 34:435–441. https://doi.org/10.1016/j.compenvurbsys.2010.04.006

Karantzalos K, Koutsourakis P, Kalisperakis I, Grammatikopoulos L (2015) Model-based building detection from low-cost optical sensors onboard unmanned aerial vehicles. Int Arch Photogramm Remote Sens Spat Inf Sci 40:293–297. https://doi.org/10.5194/isprsarchives-xl-1-w4-293-2015

Khurana M, Wadhwa V (2015) Automatic building detection using modified grab cut algorithm from high resolution satellite image. Int J Adv Res Comput Commun Eng 4(8):158–164. https://doi.org/10.17148/IJARCCE.2015.4833

Kim K, Shan J (2011) Building roof modeling from airborne laser scanning data based on level set approach. ISPRS J Photogramm Remote Sens 66:484–497. https://doi.org/10.1016/j.isprsjprs.2011.02.007

Krizhevsky A, Sutskever I, Geoffrey EH (2012) ImageNet classification with deep convolutional neural networks. Proc 25th Int Conf Neural Infor Proc Syst, NIPS'12 1:1097–1105. https://doi.org/10.1109/5.726791

LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time-series.In: Arbib MA (ed) The handbook of brain theory and neural networks. MIT Press

Li E, Femiani J, Xu S et al (2015) Robust rooftop extraction from visible band images using higher order CRF. IEEE Trans Geosci Remote Sens 53(8):4483–4495. https://doi.org/10.1109/TGRS.2015.2400462

Liu T, Fang S, Zhao Y et al (2015) Implementation of training convolutional neural networks. arXiv:1506.01195

Liu W, Anguelov D, Erhan D et al (2016) SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016. ECCV 2016. Lecture notes in computer science. Springer, Cham

Maas HG, Vosselman G (1999) Two algorithms for extracting building models from raw laser altimetry data. ISPRS J Photogramm Remote Sens 54:153–163. https://doi.org/10.1016/S0924-2716(99)00004-0

Maitra DS, Bhattacharya U, Parui SK (2015) CNN based common approach to handwritten character recognition of multiple scripts. In: Proceedings of international conference on document analysis recognition (ICDAR2015). IEEE, Tunis, Tunisia, pp 1021–1025. https://doi.org/10.1109/icdar.2015.7333916

Makantasis K, Karantzalos K, Doulamis A, Doulamis N (2015) Deep supervised learning for hyperspectral data classification through convolutional neural networks. IEEE Int Geosci Remote Sens Symp 2015:4959–4962. https://doi.org/10.1109/IGARSS.2015.7326945

Maltezos E, Ioannidis C (2015) Automatic detection of building points from lidar and dense image matching point clouds. ISPRS Ann Photogramm Remote Sens Spat Inf Sci II-3/W5:33–40. https://doi.org/10.5194/isprsannals-ii-3-w5-33-2015

Manno-Kovacs A, Ok AO (2015) Building detection from monocular vhr images by integrated urban area knowledge. IEEE Geosci Remote Sens Lett 12(10):2140–2144. https://doi.org/10.1109/LGRS.2015.2452962

McGlone JC, Shufelt JA (1994) Projective and object space geometry for monocular building extraction. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR94). IEEE, Seattle, WA, USA, pp 54–61. https://doi.org/10.1109/CVPR.1994.323810

McKeown DM, Bulwinkle T, Cochran S, Harvey W, McGlone C, Shufelt JA (2000) Performance evaluation for automatic feature extraction. Int Arch Photogramm Remote Sens Spat Inf Sci XXXII I(B2):379–394

Nalani HA (2014) Automatic reconstruction of urban objects from mobile laser scanner data. Dissertation for awarding the academic degree Doktor-Ingenieur. Dresden, Germany

Ok AO, Senaras C, Yuksel B (2013) Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. IEEE Trans Geosci Remote Sens 51(3):1701–1717. https://doi.org/10.1109/TGRS.2012.2207123

Oztimur Karadag O, Senaras C, Yarman Vural FT (2015) Segmentation fusion for building detection using domain-specific information. IEEE J Sel Top Appl Earth Obs Remote Sens 8(7):3305–3315. https://doi.org/10.1109/JSTARS.2015.2403617

Phung SL, Bouzerdoum A (2009) Matlab library for convolutional neural networks. Technical report, ICT research institute, visual and audio signal processing lab, university of Wollongong. https://www.uow.edu.au/~phung. Accessed 15 Aug 2016

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR2016). IEEE, Las Vegas, NV, USA, pp 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rottensteiner F, Trinder J, Clode S, Kubik K (2007) Building detection by fusion of airborne laser scanner data and multi-spectral images: performance evaluation and sensitivity analysis. ISPRS J Photogramm Remote Sens 62:135–149. https://doi.org/10.1016/j.isprsjprs.2007.03.001

Saito S, Aoki Y (2015) Building and road detection from large aerial imagery. Proc. SPIE 9405, Image processing: machine vision applications VIII:94050K. https://doi.org/10.1117/12.2083273

Sampath A, Shan J (2010) Segmentation and reconstruction of polyhedral building roofs from aerial lidar point clouds. IEEE Trans Geosci Remote Sens 48(3):1554–1567. https://doi.org/10.1109/TGRS.2009.2030180

Schmidhuber J (2015) Deep Learning in neural networks: an overview. Neural Networks 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Senaras C, Vural FTY (2016) A self-supervised decision fusion framework for building detection. IEEE J Sel Top Appl Earth Obs Remote Sens 9(5):1780–1791. https://doi.org/10.1109/JSTARS.2015.2463118

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

Singh G, Jouppi M, Zhang Z, Zakhor A (2015) Shadow based building extraction from single satellite image. Comput Imaging XIII:94010F. https://doi.org/10.1117/12.2083500

Tuia D, Flamary R, Courty N (2015) Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions. ISPRS J Photogramm Remote Sens 105:272–285. https://doi.org/10.1016/j.isprsjprs.2015.01.006

Uijlings JRR, Van De Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171. https://doi.org/10.1007/s11263-013-0620-5

Vakalopoulou M, Karantzalos K, Komodakis N, Paragios N (2015) Building detection in very high resolution multispectral data with deep learning features. In: Proceedings of IEEE international geoscience remote sensing symposium (IGARSS2015). IEEE, Milan, Italy, pp 1873–1876. https://doi.org/10.1109/igarss.2015.7326158

Vedaldi A, Lenc K (2015) MatConvNet-Convolutional neural networks for MATLAB. In: Proceedings of the ACM international conference on multimedia. ACM, Brisbane, Australia, pp 689–692. https://doi.org/10.1145/2733373.2807412

Von Gioi RG, Jakubowicz J, Morel J-M, Randall G (2010) LSD: a fast line segment detector with a false detection control. IEEE Trans Pattern Anal Mach Intell 32(4):722–732. https://doi.org/10.1109/TPAMI.2008.300

Vu TT, Yamazaki F, Matsuoka M (2009) Multi-scale solution for building extraction from lidar and image data. Int J Appl Earth Obs Geoinf 11(4):281–289. https://doi.org/10.1016/j.jag.2009.03.005

Yu B, Liu H, Wu J et al (2010) Automated derivation of urban building density information using airborne lidar data and object-based method. Landsc Urban Plan 98(3–4):210–219. https://doi.org/10.1016/j.landurbplan.2010.08.004

Yuan J (2016) Automatic building extraction in aerial scenes using convolutional networks. http://jiangyeyuan.com/bldgExt.html. arXiv:1602.06564. Accessed 15 Jan 2017

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Comput vision–ECCV 2014 8689:818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang K, Yan J, Chen SC (2006) Automatic construction of building footpoints from airborne lidar data. IEEE Trans Geosci Remote Sens 44(9):2523–2533. https://doi.org/10.1109/TGRS.2006.874137

Zhang Y, Sohn K, Villegas R et al (2015) Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR2015), Boston, MA, USA, pp 249–258. https://doi.org/10.1109/cvpr.2015.7298621

Zhang Q, Wang Y, Liu Q et al (2016) CNN based suburban building detection using monocular high resolution google earth images. In: Proceedings of IEEE international geoscience remote sensing symposium (IGARSS2016). IEEE, Beijing, China, pp 661–664. https://doi.org/10.1109/IGARSS.2016.7729166

Zuo Z, Wang G (2014) Learning discriminative hierarchical features for object recognition. IEEE Signal Process Lett 21(9):1159–1163. https://doi.org/10.1109/LSP.2014.2298888