Transferable instance segmentation of dwellings in a refugee camp - integrating CNN and OBIA

Omid Ghorbanzadeh, Dirk Tiede, Lorenz Wendt, Martin Sudmanns & Stefan Lang

To cite this article: Omid Ghorbanzadeh, Dirk Tiede, Lorenz Wendt, Martin Sudmanns & Stefan Lang (2020): Transferable instance segmentation of dwellings in a refugee camp - integrating CNN and OBIA, European Journal of Remote Sensing, DOI: 10.1080/22797254.2020.1759456

To link to this article: https://doi.org/10.1080/22797254.2020.1759456

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 06 May 2020.

Submit your article to this journal

Article views: 160

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tejr20
Transferable instance segmentation of dwellings in a refugee camp - integrating CNN and OBIA

Omid Ghorbanzadeh, Dirk Tiede, Lorenz Wendt, Martin Sudmanns and Stefan Lang

Interfaculty Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria

ABSTRACT
The availability and usage of optical very high spatial resolution (VHR) satellite images for efficient support of refugee/IDP (internally displaced people) camp planning and humanitarian aid are growing. In this research, an integrated approach was used for dwelling classification from VHR satellite images, which applied the preliminary results of a convolutional neural network (CNN) model as input data for an object-based image analysis (OBIA) knowledge-based semantic classification method. Unlike standard pixel-based classification methods that usually are applied for the CNN model, our integrated approach aggregates CNN results on separately delineated objects as the basic units of a rule-based classification, to include additional prior-knowledge and spatial concepts in the final instance segmentation. An object-based accuracy assessment methodology was used to assess the accuracy of the classified dwelling categories on a single object-level. Our findings reveal accuracies of more than 90% for each applied parameter of precision, recall and F1-score. We conclude that integrating the CNN models with the OBIA capabilities can be considered an efficient approach for dwelling extraction and classification, integrating not only sample derived knowledge but also prior-knowledge about refugee/IDP camp situations, like dwellings size constraints and additional context.

KEYWORDS
Convolutional neural network (CNN); object-based image analysis (OBIA); knowledge-based semantic classification; refugee camps; humanitarian operations

Introduction

Both human-made and natural disasters are the main reasons of population displacement. According to (UNHCR, 2019), almost 70.8 million individuals were forcibly displaced worldwide as a result of persecution, conflict, violence, or human rights violations. Refugee and IDP (internally displaced people) camps are often the first accommodation for people who have been forced to flee their home. The availability of very high resolution (VHR) remote sensing (RS) images is growing because of the continuous development of advanced RS related technologies (G. Zhang et al., 2018). Nowadays, VHR satellite images are widely used for efficient support of camp planning and delivery of humanitarian aid (Witmer, 2015). Firstly, these satellite images have significant potential to provide humanitarian organisations with spatial information of difficult to access areas. Secondly, for large and complex refugee/IDP camps, VHR RS images make it possible to have a deeper and better understanding of the camp situation and camp dynamic at lower costs compared to traditional surveys and often as the only possibility due to security and accessibility reasons (Lang et al., 2019a). Therefore, VHR RS imagery is considered as the primary source of information for identifying the number, type and size of dwellings which can serve – amongst others as the input for an estimation about the number of people in such camps (using the dwellings as proxies).

Automated processing of the satellite images to produce applicable high-level information is still a challenging task. To extract and categorise dwelling types from RS images, previous studies have used various approaches as an alternative to visual interpretation, such as template-matching (Tiede et al., 2017), (semi-)automated workflows (Spröhnle et al., 2014), object-based image analysis (OBIA) (Lüthje et al., 2015; Tiede et al., 2010), mathematical morphology-based algorithms (Laneve et al., 2006), convolutional neural networks (CNNs) (Ghorbanzadeh et al., 2018a, 2018b; Quinn et al., 2018). Visual interpretation approaches have some major drawbacks: they depend on the expert’s experience to annotate different types of dwellings, and are expensive, as they are labour-intensive and time-consuming. Pixel-based approaches face severe problems in single dwellings extractions from VHR RS image, since they denote just a small fragment of a dwelling without taking into account spatial properties (like size/form, context), and usually fail in correctly extract dwellings and categorise them into different types.

OBIA deals with the shortcomings of pixel-based classification methods by grouping pixels into spectrally similar non-overlapping segments utilising abundant
features (Blaschke, 2010; Jozdani et al., 2019). A key element of OBIA is image segmentation whose aim is to generate image objects suitable for further classifications of spatial properties and context (Blaschke & Pirililou, 2018; Lang et al., 2019b). OBIA has been successfully used in multiple applications (Blaschke et al., 2014; Tavakkoli Pirililou et al., 2019) during the past few decades.

Whenever the segmentation process produces the objects that are used for the classification, the result may be influenced by the quality of this process (Pan et al., 2019). Especially, in the case of camps, some materials that are used to construct dwellings may produce higher colour contrast than other making it difficult to extract and categorise the different type of dwellings without any prior context knowledge or complex rulesets.

During the past decade, deep-learning methods, and in particular CNNs, have achieved cutting edge success in the computer vision (Krizhevsky et al., 2012), and are also used for RS image classification (Zhu et al., 2017). Compared to traditional pixel-based algorithms, they integrate spatial context and texture of images in the analysis process by design to some degree. Recently, the DeepGlobe 2018 datasets were introduced by (Demir et al., 2018) as datasets that can be considered as valuable benchmarks in satellite image processing to show the priority of any novel approach in this domain. CNNs have showed some improvements in classification and semantic segmentation of RS images (Du et al., 2019; Qayyum et al., 2019), object detection (Ghorbanzadeh & Blaschke, 2019; Guirado et al., 2019; Sameen & Pradhan, 2019) scene classification (Han et al., 2017), instance-aware semantic segmentation and instance segmentation (Dai et al., 2016; Igliovikov et al., 2018). In the case of semantic segmentation, each single pixel is labelled, while in the case of instance segmentation, instead of connecting each pixel to a label, target objects will be classified, and only pixels of those objects are labelled (Panboonyuen et al., 2019). Multiple hierarchical stacking and trainable layers of a CNN can learn characteristic features and abstractions from raw RS images (Fu et al., 2019).

Although CNNs have shown some of the state-of-the-art baselines in the mentioned domains, they still face some challenges. Generally, CNN-based models implement classification on the pixel level (Jin et al., 2019). Those cannot easily identify object borders and transitional zones of different dwelling types or other existing objects. The problem is even more apparent when we encounter non-regular camp structures and aim to also extract small types of dwellings (Ghorbanzadeh et al., 2018b). Solutions to overcome these difficulties were suggested, such as training data augmentation (Radovic et al., 2017), CNN model structure modification or using deeper CNN models (Sameen & Pradhan, 2019) with more hidden layers and nodes, which led to marginal improvements (Lin et al., 2016), using pre-trained CNN models and fine-tuning them to the new applications (Castelluccio et al., 2015). However, such solutions have some limitations such as overfitting problems (Hu et al., 2015). Even if it was found that training data augmentation helps enhance the classification performance, it could increase the dependency of the CNNs on the training sample size (Jin et al., 2019). Although OBIA methods have several advantages and are considered as advantageous compared to pixel-based ones especially in HR/VHR imagery, most of the CNN applications for the RS images rely on the pixel-based method (Liu & Abd-Elrahman, 2018). Any image understanding process (i.e., image classification and object annotation) using CNN models require the input training data to be fixed-size square window sample patches of pixels. (Ghorbanzadeh et al., 2019) applied CNN models with different layer depths to extract slope failures (as a spatial feature) within different window sizes ranging from 12 to 48 pixels. (Sameen & Pradhan, 2019) trained different CNN models using the same window size patches of 15 pixels from RS imagery. However, the target objects in RS images usually exhibit a wide range of different sizes. Focusing on the larger objects requires large window size patches. At the same time, using a large fixed-size window can make it challenging to annotate small target features (Liu et al., 2019). Distances between closely located dwellings, for instance, are very short, and using a large fixed-size window has the effect that a large part of the window is filled with dwelling bodies, which might mislead a CNN to classify all area as only one huge dwelling.

So far, a limited number of studies tried to use capabilities of both CNN models and OBIA methods. Lv et al. (Lv et al., 2018) set multi-scale samples with six different window sizes from 15 to 65 using CNN model integrated with a segmentation process and a majority voting strategy to solve the boundary problem in urban image classification. Zhao et al. (Zhao et al., 2017) used a fixed-size square window of 18 pixels applied in a five-layer CNN to classify complex urban objects in high-resolution RS images. An object-based classification method was then applied to classify the resulting features from the CNN model. A similar approach was used to integrate OpenStreetMap data in an object-based CNN approach within the urban area (Zhao et al., 2019). Liu et al. (S. Liu et al., 2019) applied OBIA as a post-classification to refine the LULC mapping based on a CNN model using Sentinel optical and SAR data. Their proposed approach improved the overall accuracy of LULC mapping for different datasets including the Sentinel Guangzhou, the Zhuhai-Macau LCZ, and University of Pavia dataset. The use of image patches as input data form for training CNN models provides a natural opportunity to use them in the OBIA approach, and have shown higher
performance in the object-based RS applications (S. Liu et al., 2019).

The studies mentioned in our literature review tried to deal with the challenges of intra-class heterogeneity, inter-class homogeneity, and delineate more accurate object boundaries in the semantic classifications. Some instances of intra-class heterogeneity and inter-class homogeneity in our case of dwelling classification of refugee/IDP camps are represented in Figure 1. In our study, an integration approach of a CNN model with an OBIA classification method is presented for classification of bright dwellings of an of IDP camp. Our main purpose is evaluating our proposed solution of physically incorporating the concept of the object to a CNN model. Therefore, our focus is not to evaluate the performance of different existing CNN structures or designs for state-of-the-art dwelling classification. Our contribution is to show possibilities for integrating expert knowledge through the OBIA rulesets to the decision making step of a CNN model. Thus, any CNN model can be applied for this aim and a non-complex designed one is much more desired to represent the integration capabilities rather than a complex one. The significant contributions are: (1) training a CNN model with a time series of images and testing the model on an image of a new date where no samples were seen by the algorithm at all; (2) implementing on top of the CNN results a straight-forward rule-based OBIA method for dwelling classification including the probabilities of the CNN model and encoding expert knowledge about dwelling size and distribution; (3) applying an object-based accuracy assessment methodology to evaluate the results of our integrated approach.

Workflow

Overall methodology

In this study, we integrate a CNN model with an OBIA classification method for the extraction of different dwelling types of the Minawao refugee camp in Cameroon. The workflow of this study is as follows:

1. Prepare a manually labelled dataset based on different VHR images from different sensors acquired in two different years (2015 and 2016) of a very dynamic refugee camp.
2. Establish a CNN model concerning the considered input window size patches of the applied RS imageries.
3. Apply OBIA classification (rule-set integrating expert knowledge) to the probability values of different dwellings and non-target classes based on the CNN results.
4. Apply an object-based accuracy assessment to validate the model performance.

The description of the applied integrated approach, the experimental results, further explanations and discussions are organised in the following sections of this paper.

Data collection

The test site for this study is the camp Minawao in the Far North Region of Cameroon, at 10°33′30″ N 13° 51′30″E (see figure 2). This refugee camp is run by UNHCR, and mainly houses refugees from the neighbouring Nigerian Borno State, where activities of the Boko Haram militia forced over 240,000 people to flee.
their homes (UNHCR, 2019). The camp has existed since 2012. Population numbers started to increase rapidly from early 2014 onwards, and by the time of the first image, acquisition on 12 April 2015 the camp housed approximately 34,000 people. Over the period considered in this study, the population grew to approximately 61,000 inhabitants (Wendt et al., 2017). During this time, the camp extent grew from ~230 ha to 625 ha. The camp still hosts nearly 60,000 people (UNHCR, 2019). An operational service produced the training and test data used in this study for humanitarian mapping at the University of Salzburg, Department of Geoinformatics (Z_GIS) in the context of mapping requests from Doctors without Borders (MSF) within the last years. The existing classifications are based on the semi-automated method, i.e. the output is subsequently refined manually and validated by a trained operator. Information products based on dwelling extractions are used operationally by MSF to obtain population numbers independently from the inhabitant registration system run by UNHCR, and for the planning of healthcare, water and sanitation services and campaigns.

The training data set was prepared based on RS images from different sources and times. The input data are taken from WorldView-3 images captured on 12th April and 13 October 2015, with four spectral bands, blue (450–510 nm), green (510–580 nm), red (630–690 nm), and near-infrared (770–895 nm). In addition, we used a GeoEye-1 image acquired on 1 April 2016 with blue (450–510 nm), green (510–580 nm), red (655–690 nm), and near-infrared (780–920 nm) spectral bands. The last image we used to produce the training data set was a WorldView-2 image captured on 3 June 2016.

A WorldView-2 image, acquired on 17 February 2017 from the same IDP camp was used for testing the integrated approach. The spectral bands of this image were applied for the test process. All dwellings were extracted and labelled manually in the context of an operational humanitarian mapping task by experts of the University of Salzburg. For all labelled objects, the centroids were taken to create convolution input sample patches (see Figure 3). A data augmentation method of shifting window was used for increasing the number of sample patches of those dwelling classes, which were less in number than the other classes. Also, the numbers of each dwelling type differ at different times. For example, tunnel shape dwellings decreased strongly from 2015 to 2016; they were replaced with other types of dwellings. Drop shape dwellings emerged in 2016 with no occurrence of this type before. Therefore, data augmentation helped keep the balance between different types of classes. Table 1 shows the number of convolution input sample patches of each class. The table also indicates the number of produced sample patches and not the number of objects in each class. Moreover, the 2017 image was used for testing our approach; no image patches were created from this image. Figure 4 shows examples of
selected convolution input sample patches for each class from each image.

**Convolution neural network (CNN)**

CNNs have led in the state-of-the-art feature extraction results and are considered a hot topic in the image processing and computer vision fields, which gradually overcome traditional methods (Pena et al., 2019). CNNs as kind of mature network of deep learning model is inspired from the biological multi-layer neural networks architectures, which enable them to form high-level semantic features from the existing low-level features in an image (Jin et al., 2019). Multi-layer neural networks are interconnected to each other by a set of learnable weights and biases. The input of layers is small patches of the image that move over the entire image to obtain different feature characteristics (C. Zhang et al., 2018). These image patches are generalised through two main blocks of any CNN model including convolutional and pooling layers and generate a group of feature maps. The resulting feature maps from each layer feed-forward to the next layers until the high-level semantic features are captured (Ghorbanzadeh et al., 2019). Feed forwarding of the feature maps enables the learnable filters of convolution layer to learn by different feature extractors and large amount of image patches required to learn (Maggiori et al., 2017) appropriately. Since the local features are translation-invariant, the extracted features are more important than their location in the training data (Yang et al., 2017). The pooling layer is applied in a local window on the resulting feature maps from each convolution layer. The condensed feature maps by the pooling layer are usually insensitive to spatial translations (Yang et al., 2017). Although pooling layers have performed using the average or summing functions, the pooling layer using the max value, the so-called max-pooling, is the most common, and it keeps only the maximal

---

**Figure 3.** An illustration of convolution input sample patches selecting based on the position of the centroid of any labelled object.

**Table 1.** The figure in the table indicates the number of training sample patches per different target dwelling class (bright dwellings) and non-target classes (small dark structures, bare soil etc.). Numbers of samples are different from the real mapped dwellings, due to the applied augmentation routines to improve the balance between the different classes.

<table>
<thead>
<tr>
<th>Image</th>
<th>Date</th>
<th>Facility</th>
<th>Target classes</th>
<th>Non-target classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Buildings</td>
<td>Rectangular Shape</td>
<td>Tunnel Shape</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1226</td>
<td>1383</td>
</tr>
<tr>
<td>Testing data</td>
<td>WV3 12 April 2015</td>
<td>1821</td>
<td>2584</td>
<td>2833</td>
</tr>
<tr>
<td></td>
<td>GE1 1 April 2016</td>
<td>2739</td>
<td>3356</td>
<td>1810</td>
</tr>
<tr>
<td></td>
<td>WV3 3 June 2016</td>
<td>2823</td>
<td>3564</td>
<td>1389</td>
</tr>
<tr>
<td>Test data</td>
<td>WV2 17 February 2017</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

---
values of the feature maps. Furthermore, an element-wise non-linear activation function (e.g., ReLU, sigmoid, and hyperbolic tangent) is taken for non-linearity amplification of the convolutional layers.

In the present study, we considered different sizes of dwellings and different distances between pairwise dwellings, and multiple sample patch window sizes of $16 \times 16$, $20 \times 20$, $32 \times 32$ pixels. Using cross-validation, the sample patch window size of $20 \times 20$ pixels was selected as the optimal one for further processes. Although some current studies attempt to find an automatic framework of the optimal CNN architecture for their datasets (e.g., Pavia University Scene Data) (Hang et al., 2019), their transferability to other cases such as dwelling detection is not clear. Since designing the optimal architecture for each specific training dataset remains under-explored, some researchers like (Ghorbanzadeh et al., 2019) and (Sameen & Pradhan, 2019) applied different CNN architectures and compared the results. The patch window sizes are considered as one of the most critical parameters in optimal CNN architecture and consequently the number of convolution/pooling layers and the size of applied kernels (Ghorbanzadeh et al., 2019). In this study, considering the size of our optimal sample patch window, the number of network layers was tuned to five (see Figure 5). The CNN was using a convolution layer with a kernel size of $5 \times 5$ and stride of 1 as the first convolution layer and continuing with additional convolution layers with the kernel size of $3 \times 3$ after the only max-pooling layer. As our sample patch window size is relatively small ($20 \times 20$ pixels), only one max-pooling layer (with a kernel size of $2 \times 2$ and stride of 2) was used immediately after the first convolution layer. Therefore, the information loss to the next convolution layers is minimised. In this case, our designed CNN was fed by the sample patches of $20 \times 20 \times 4$ units, where $20 \times 20$ is the window size of the input sample patches, and 4 is the number of spectral bands (RGB and NIR). These sample patches were selected out of $6301 \times 6201$ pixels of the labelled original images. Different numbers of feature maps were used in each layer. The CNN model was designed and trained in Trimble’s eCognition software, based on the Google TensorFlow software library. In the model, batch normalisation (BN) layers were applied after all convolution layers to have higher learning rates. Using the BN helps reduce overfitting by improving the capacity of generalisation of the CNN architecture and increasing the speed of its learning process by accelerating the convergence (Pena et al., 2019). The learning rate of this work is featured a gradual reduction starting from 0.001 dropping down to learning rate of 0.0006, which resulted in an acceptable performance within cross-validation. Lower learning rates increase the time of the learning process and the chance of being stuck in local minima and subsequently end up with incorrect weights. Although higher learning rates like 0.001 would increase the speed of the learning process, the network may not reach the minimum and again get wrong weights. A batch size of 50 and 5000 training steps were used in this work to obtain the best performance.

**Object-based image analysis (OBIA)**

For each pixel, the predictions of the CNN model are 10-dimensional vectors $P = (p_1, p_2, \ldots, p_{10})$, where 10 is the number of classes including both our target and non-target classes, and each dimension $I \in [1, 2, \ldots, 10]$ indicates the CNN model-predicted probability of the $i$th semantic class. In an ideal case, the probability should be 1 for a specific class and 0 for the other ones, which is usually not the case in real-world situations. The resulting probability for each semantic class can be easily shown as $(x) = (p_1[x \in [1, 2, \ldots, 10]),$ where $p_1 \in [0,1]$ and $\sum_{i=1}^{10} p_i = 1$ (Li et al., 2019). After the image segmentation, the corresponding probability of each resulting object is different (object-mean). For our case, the resulting probability for each semantic class within an object...
can be presented as \( f(y) = (p_y | y \in \{1, 2, \ldots, 10\}) \), where \( p_y \in [0,1] \) and \( \sum_{y} p_y \geq 1 \) (see Figure 6 and Table 2). The segmentation applied was a multi-resolution segmentation (Baatz & Schaepe, 2000) based on 4 pan-sharpened spectral bands (R-G-B-NIR), the variance-based homogeneity parameter (the so-called scale parameter) has been intentionally set to a low value of 10 to achieve over-segmentation and to avoid any under segmentation of target objects. Shape and compactness parameters of the algorithm were equally weighted with 0.5 aiming for quite compact objects. Target objects were then merged and classified based on a combination of context/neighborhood, spatial properties and similarity in the CNN probability results for the different target and non-target classes, which improved the segmentation results in contrast to an unsupervised segmentation (see Figure 6). This second supervised segmentation/classification step is implemented as a rule-set in CNL (Cognition Network Language within eCognition) able to address and manipulate single objects based on expert knowledge and included:

- Initial classification of objects, based on the CNN probability values using fuzzy membership values using two features, namely the probability mean values per object and the maximum probability value per image object
- Some wrongly classified vegetation objects were eliminated from the list of target objects using an NDVI threshold of 0.3.
- The membership values of the different objects are also used in the supervised merging process based on similarity and shared object borders, i.e. an object growing into neighbouring pixel took place if the membership value of that pixel in the specific class was not less than 90% of the growing object. This has been implemented as a loop & merge function.
- Post-processing:

![Figure 5. The structure of the applied CNN-OBIA approach and the applied parameter settings in this study.](image)

![Figure 6. A local result of the multiresolution segmentation (MRS) in conjunction with a second supervised segmentation (object-merging) step based on CNN probability values and context/neighborhood parameters. Four randomly selected objects (a, b, c, and d) are compared in detail (refer to table 2).](image)

<table>
<thead>
<tr>
<th>Probability of Object (a)</th>
<th>Probability of Object (b)</th>
<th>Probability of Object (c)</th>
<th>Probability of Object (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectangular Shape</td>
<td>0.87</td>
<td>0.06</td>
<td>0.58</td>
</tr>
<tr>
<td>Tunnel Shape</td>
<td>0.03</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>Facility Buildings</td>
<td>0.12</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Drop Shape</td>
<td>0.26</td>
<td>0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>Tukuls</td>
<td>0.2</td>
<td>0.47</td>
<td>0.09</td>
</tr>
<tr>
<td>Small Bright</td>
<td>0.49</td>
<td>0.36</td>
<td>0.23</td>
</tr>
<tr>
<td>Brown</td>
<td>0.53</td>
<td>0.54</td>
<td>0.46</td>
</tr>
<tr>
<td>Small Dark</td>
<td>0.17</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>Bare Soil</td>
<td>0.03</td>
<td>0.11</td>
<td>0.21</td>
</tr>
<tr>
<td>Vegetation</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

![Table 2. Corresponding resulting probability values for randomly selected objects (a, b, c, and d, see Figure 6) based on each semantic class.](image)
- Objects smaller than 3 m² were not further considered as dwellings.
- Objects classified based on the CNN probability values as Facility Buildings below a size of 50 m² were reclassified as rectangular dwellings.
- Neighbouring objects classified as Facility Buildings were merged without any size constraint.
- Objects classified as the tunnel and drop shape objects were evaluated using a “rectangular fit” option. The calculation is based on a rectangle with the same area as the image object. The proportions of the rectangle are equal to the proportions of the length to width of the image object. If the rectangular fit was higher than 0.9 (1 indicates a perfect rectangle) the objects were also reclassified as rectangular dwellings.

**Object-based accuracy assessment**

For the target classes a pixel-based accuracy assessment, resulting ratios in the pixel overlap between the reference and the classification were not considered as appropriate. The critical information to be delivered in the present context – supporting humanitarian action – is on number and size per dwelling and aggregated per dwelling type, in order to estimate population figures. Therefore, an object-based accuracy assessment has been conducted. For this, each extracted dwelling objects is spatially compared with the reference objects, similar to the approaches described in Radoux and Bogaert (2017) (Radoux & Bogaert, 2017) for spatial entity detection. A tool for assessing a spatially explicit accuracy has been developed in Python and implemented in ArcGIS. The comparison of reference and target dwellings is based on the following schema (see also Figure 7):

1. count how many extracted dwelling-polygons have an associated reference object and vice versa;
2. prevent double-counting (if, e.g., a reference polygon is covered by more than one dwelling/object), which would be equal to false positive;
3. for reference objects with more than one matching extracted object, the most appropriate object is counted (i.e. the object with the largest overlap);
4. each reference-object is buffered according to a certain user-defined distance, e.g., a minimum distance between dwellings (here: 2 m);

**Figure 7.** Object-specific accuracy assessment evaluating different potential cases, resulting in different accuracy aspect: a) Overlap of reference (black) and extracted objects (red) (TP = True positive), b) avoidance of double counts, only the best match is counted, additional overlaps are counted as false positives (FP), c) Overlap calculations includes a user-defined buffer range (here: 2m). Only the best match (i.e. largest overlap) is counted, d) if an extracted object is overlapping more than one reference object, also only the best match is counted as TP, other overlapping (or non-overlapping) reference objects are considered as FN (false negatives).
(5) Provision of per dwelling area compared to the matching reference (without double counts)
(6) Provision of direct overlap (intersect) between best matches and all matches

The mentioned buffer size ensures to count associated dwellings, which are not directly overlapping the reference due to errors in the classification procedure. In this case, the buffer size is set to 2 m, which reflects the absolute minimum distance between dwellings in refugee camps as recommended by UNHCR camp planning standards.

Using metrics of TP, FP, and FN, standard accuracy assessment parameters of Precision, Recall, and F1 can be calculated for the results. Precision indicates the proportion of target dwellings, which correctly identified by the proposed approach. The recall is the proportion of target dwellings in the labelled data that were correctly detected by the approach. F1 is used to balance Precision and Recall parameters (see equations 1–3).

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \tag{2}
\]
\[
F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}
\]

The accuracy assessment was conducted for the bright dwelling classes, which are used for housing (target classes). Larger buildings (> 50 m²) were excluded.
from the reference and the analysed data set since they are in such camps usually not used for population estimations (Facility Buildings etc.).

**Experimental results**

This paper integrates the probabilities resulting from a CNN model with an OBIA classification method to observe the accuracy of the classification of bright dwellings including Rectangular Shape, Tunnel Shape, Facility Buildings, and Drop Shape in a refugee camp. As there were many changes in this camp over three years (from 2015 to 2017) the transferability of the trained CNN model was a significant issue in our study. Some of the dwellings such as tunnel shape ones gradually disappeared or were replaced with other types, and in most of the cases, the shape of dwellings did change a lot through time.

Moreover, some new dwelling types such as drop shape dwellings appeared in 2016 and increased a lot until 2017, when the integrated approach was tested on that. To keep the balance between the numbers of sample patches in each class of training data a data augmentation was applied for those types of dwellings, which were less than the other ones. After testing the CNN model on the test image, the multi-resolution segmentation (MRS) was used to create the objects for further classification using OBIA. The segmentation process was based on the resulting probabilities of our CNN model and the spectral information of the test image. Therefore, the resulting objects for classification were created based on the feature extraction capabilities of a CNN model. For further classification, the membership values of an initial fuzzy dwelling classification were also used in a supervised

**Figure 9.** Enlarged maps of four challenging areas (mentioned in figure 1) for comparing the results of the OBIA-CNN approach with those of manually extracted labels.

**Figure 10.** Histogram showing the distribution of the matching area between reference objects and extracted objects along with basic descriptive statistics according to the object-specific accuracy assessment. Only best matching objects (as defined in the object-specific accuracy assessment) are taken into consideration (no-double counts etc.). The graph shows the overall well matching of the area on a single dwelling level, which is an important pre-requisite for any population estimation based on population occupation rate per dwellings of different size.
Table 3. Results of the object-specific accuracy assessment focusing on the target classes mainly relevant for population estimation in refugee camps (number and size of relevant classes). The second column shows the results without larger dwellings (e.g., Facility Buildings, here: > 50 m²).

<table>
<thead>
<tr>
<th></th>
<th>Acc Ass for all target classes</th>
<th>Without Facility Buildings (dwellings larger than 50 m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number CNN-OBIA</td>
<td>12,304</td>
<td>12,079</td>
</tr>
<tr>
<td>Number Reference</td>
<td>13,085</td>
<td>12,902</td>
</tr>
<tr>
<td>TP (overlap incl. 2 m Buffer, no double count)</td>
<td>11,898</td>
<td>11,667</td>
</tr>
<tr>
<td>FP (no overlap or not best match)</td>
<td>406</td>
<td>412</td>
</tr>
<tr>
<td>FN (reference not overlapping CNN-OBIA, or not best match)</td>
<td>1187</td>
<td>1235</td>
</tr>
<tr>
<td>Total Area overlap only best matches (intersect Reference + CNN/OBIA, m²)</td>
<td>152,187.58</td>
<td>136,869.11</td>
</tr>
<tr>
<td>Total Area overlap (intersect Reference + CNN-OBIA, m²)</td>
<td>164,280.75</td>
<td>137,411.91</td>
</tr>
<tr>
<td>Total Area Reference m²</td>
<td>203,900.09</td>
<td>178,710.56</td>
</tr>
<tr>
<td>Total Area CNN-OBIA m²</td>
<td>230,496.36</td>
<td>196,141.21</td>
</tr>
<tr>
<td>Precision</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Recall</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>F1</td>
<td>0.94</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Emerging process based on similarity (of CNN probabilities) and shared object borders. The fuzzy classification method is considered a reliable method for such complex situations. However, the transferability of the model depends on using appropriate membership values.

The instance classification results of the target classes on the Minawao refugee camp are shown in Figure 8. The proposed approach could classify different bright dwellings, and the accuracy assessment was applied to the resulting classification map.

**Accuracy assessment**

As described in the workflow section, the focus of our classification is set on the bright dwelling structures as a proxy for population estimations, other classes (dark structures, vegetation bare soil etc.) are combined as non-target classes in the approach. Such population estimations are usually conducted based on a single dwelling basis, taking into consideration the dwelling types (inhabited or not), dwellings size and the occupancy rates usually estimated from people working in the field (humanitarian organisations, see also (Grundy et al., 2012; Lang et al., 2010).

The following table depicts the values for the object-specific accuracy assessment for the target classes (bright dwellings) with the larger structures (Facility Buildings) and without. The accuracy assessment was conducted for the whole camp in the 2017 image. Total values for with and without Facility Buildings reveal a high amount of agreement with the total dwelling numbers. The object-specific accuracy assessment (counting double counts (several overlapping structures) as FPs) also reveal high F1 scores (0.94 incl. Facility Buildings, 0.93 without); precision rates are higher (0.97 both) than the recall values (0.91 and 0.93) (see Table 3). Figure 9 shows enlarged parts of some challenging areas (mentioned in Figure 1), which make it possible to visually compare the results of the OBIA-CNN approach with ground truth labels.

Figure 10 reveals some objects specifics statistics regarding the area of extracted objects (object delineations) and the object sizes of the reference (without Facility Buildings). Most of the objects match well in terms of the area with the reference objects, mean and median are around 1 (slight overestimation) and a standard deviation of around 7 m² could be reached. These results are quite promising concerning population estimations, where not only the number of dwellings is significant but also the size of the single dwellings, which are determining the occupancy rates. Most of the larger outliers were situations where extracted objects are overlapping more than one reference object, and vice versa, which was then considered as false positive and the corresponding area is set to zero.

**Discussion**

In this study, we proposed an integration approach that equips a simple CNN model with the OBIA classification capabilities for a refugee camp classification. Our approach was tested on a new image of the camp that was not applied for the training process of the CNN model. It obtained some remarkable classification results with more than 90% accuracies. Such results showed the potentials in the transferability of the integration of the CNN models with knowledge-based OBIA classification methodologies. This study demonstrated the transferability of the approach in terms of using different sensors, different conditions/dates, and partly different dwelling classes and types. Although the approach was applied for different training datasets and images, it was still tested on the same refugee camp and the same geographical conditions.

In contrast to recently published studies of CNN and OBIA integrations that successfully enhanced the overall accuracy by mainly focusing on some post-classification refinements (e.g., adding some auxiliary information) (Li et al., 2019; S. Liu et al., 2019), our approach is able to
incorporate the resulting probabilities of a CNN model alongside with the spectral information of the test image in the OBIA part in an effective manner. Therefore, this study is not a mere refinement of a CNN model resulting classification by an OBIA approach. In addition, as the performance of the segmentation and the classification of the applied OBIA method were based on the spectral information of the test image and the resulting probabilities of the CNN model respectively, expert rules in the OBIA part were as far as possible relying on camp planning standards, size and form differencing as parameters to ensure transferability. The addition of the knowledge-based OBIA part helped to sharpen the delineation of objects compared to the blurred results from the CNN approach only.

Moreover, the OBIA classification fairly avoided intra-class heterogeneity and inter-class homogeneity in dwelling classification, by integrating additional prior-knowledge in the approach. Based on the resulting classification and the corresponding accuracies, this study concludes that a combination of the probability results of the CNN models and an additional OBIA step can play an important role in further classification and mapping. Especially for population estimation based on dwelling type and size, such an approach has advantages e.g., to object-detectors only. Future studies will explore the transferability of the proposed approach between different areas by training and testing on different refugee/IDP camp sites. As already mentioned, a simple CNN model was used in the present study. The switch from merely pixel-based accuracy assessments to an object-based one is in our understanding a step forward: in applications cases, where the delineation of objects is important for further analysis (here: estimation of population figures based on dwelling occupancy rates), such an object-based accuracy approach contributes to accuracy values that are more realistic.

Conclusions

Within the present study, it could be shown, that for the case of refugee/IDP camps and their specific challenges (small size of dwellings, spectral heterogeneity and changing image conditions) a CNN network can be trained on different sensors and acquisition dates, and the successful transferability to a not-sampled image of the same camp – with different camp extents (number of dwellings) – is possible. The object-based approach built on top of the analysis helped to improve the dwelling delineation, which is important for specific dwellings area calculations and subsequent population estimations. We focused on bright dwellings only, the very small scales of darker structures and their problematic separability (also for human interpreters) is still a problem which needs further research. Besides, the transferability of the approach to other camps of similar structure but in different areas and under different (climate) conditions is still an open question. Nevertheless, we think that the work will contribute to further improve the quality of automated solutions and the speed & scale of expert interpretations of such analyses for the humanitarian domain.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is partly funded by the Open Access Austrian Science Fund (FWF) through the GIScience Doctoral College [DK W 1237-N23].

ORCID

Omid Ghorbanzadeh http://orcid.org/0000-0002-9664-8770
Dirk Tiede http://orcid.org/0000-0002-5473-3344
Martin Sudmanns http://orcid.org/0000-0002-0473-1260
Stefan Lang http://orcid.org/0000-0003-0619-0098

References


