Fully convolutional networks for land cover classification from historical panchromatic aerial photographs.

Nicholus Mboga¹, Tais Grippa¹, Stefanos Georganos¹, Sabine Vanhuysse¹, Benoît Smets², Olivier Dewitte², Eléonore Wolff¹, Moritz Lennert¹.

Correspondence: nmboga@ulb.ac.be*, tgrippa@ulb.ac.be, s.georgan@ulb.ac.be, svhuysse@ulb.ac.be, benoit.smets@africamuseum.be, Olivier.Dewitte@africamuseum.be, ewolff@ulb.ac.be, mlennert@ulb.ac.be

Corresponding author*

¹Department of Geosciences, Environment & Society, Université Libre de Bruxelles, Brussels Av Franklin Roosevelt 50-1050, Belgium.

²Department of Earth Sciences, Royal Museum for Central Africa, Leuvensesteenweg 13, 3080 Tervuren, Belgium

ABSTRACT

Historical aerial photographs provide salient information on the historical state of the landscape. The exploitation of these archives is often limited by accessibility and the time-consuming process of digitizing the analogue copies at a high resolution and processing them with a proper photogrammetric workflow. Furthermore, these data are characterised by limited spectral information since it occurs very often in a single band. Our work presents a first application of deep learning for the extraction of land cover from historical aerial panchromatic photographs of the African cities of Goma, Bukavu and Bujumbura. We evaluate the suitability of deep learning for land cover generation from a challenging dataset of photographs from the 1940s and 1950s that covers large geographical extents and is characterised by radiometric variations between dates and locations. A fully convolutional approach is investigated by considering two network architectures with different strategies of exploiting contextual information: one used atrous convolutional layers without downsampling, whereas the second network has both downsampling and learned upsampling convolutional layers (U-NET). The networks are trained to detect three main classes namely, buildings, high vegetation and a mixed class of bare land and low vegetation class. High overall accuracies of > 90% in Goma-Gisenyi and Bukavu, and >85 % in Bujumbura are obtained. This work provides a novel methodology that outperforms a baseline standard machine learning classifier for the exploitation of the vast archives of historical aerial photographs that can aid long-term environmental baseline studies. Future work will entail developing domain adaptation strategies in order to make the trained network robust for different image mosaics.

Keywords: fully convolutional networks, deep learning, panchromatic historical aerial imagery, land cover classification
1 Introduction

Historical datasets provide an avenue to study events that happen in the environment over an extended time period. A key issue affecting utilization of these datasets is often the availability of readily usable digital formats. In the remote sensing domain, historical aerial photographs were captured for various parts of the world starting in the early 1930s (Luman et al., 1997). Historical aerial black and white photographs have served as a key input for topics as different as for instance, landslide inventory and distribution analysis (Martha et al., 2012; Shu et al., 2019), vegetation dynamic studies (Carter et al., 2011; Hudak and Wessman, 1998; Jeter Jr and Carter, 2016; Kadmon and Harari-kremer, 1999; Lucas et al., 2010; Miller, 1999), landcover mapping (Caridade et al., 2008; Pinto et al., 2019; Robertson, 2012) and the evolution of river morphology (Abate et al., 2015).

Since majority of the historical aerial photographs were captured in panchromatic films, any study of the environment, its landcover and its changes over a multidecadal time scale can only be done if we grapple with the underlying challenges to their processing. First, analogue copies of the photographs suffer shrinkage and deterioration in brightness from ageing. In other instances, operators might write on the photographs using permanent markers that often defaces them. In addition, during the reproduction and scanning processes to create digital copies of the analogue photos, exposure differences in brightness leads to inconsistent distribution of brightness across the image (Luman et al., 1997).

The monochrome nature of the photographs reduces their discriminative character for land cover classification. While visual image interpretation has been the main mode of information extraction of landcover classes for remote sensing problems, it has not been without challenges. It is a labor-intensive task whose success depends on skilled operators. Furthermore, it is bound to consume a lot of time especially if the study area covers a large geographical extent (Forster, 2006). Traditional machine learning-based approaches have been used to automate the process of interpreting these images. Machine learning approaches may be applied directly on the panchromatic image, or on a combination of the raw image and extracted texture features. Textures encode salient information on the spatial distribution of tonal variations within a specific image channel (Haralick et al., 1973) and examples include the Grey Level Co-occurrence Matrix (GLCM) and the Local Binary Patterns (Ojala et al., 2002). However, the extraction of such texture features is characterized by the high number of free parameters that need to be optimized, the need for domain specific knowledge and the amount of human effort and computer resources needed.

In recent years, deep learning models have become useful in the field of remote sensing (Ma et al., 2019). While they are also considered as part of machine learning methods, they differ from traditional machine learning algorithm since they have a key property of automatically extracting spatial-contextual features directly from the input data, thus reducing the need for feature creation and selection by the operator. Convolutional neural networks are a class of deep learning algorithms that use a network of nonlinear layers to learn/extract low level, mid-level to high level features from the input data (LeCun et al., 2015). Several architectures of convolutional neural networks have been designed as for example, patch-based.
convolutional neural networks (Bergado et al., 2016; Krizhevsky and Hinton, 2012), and fully convolutional networks (Long et al., 2015; Sherrah, 2016). Availability of large-scale benchmark datasets and improved parallel processing through Graphical Processing Units (GPUs) have encouraged development in the field. While a lot of work is ongoing in the development and application of deep learning in the remote sensing field, most of the datasets being used are multi-spectral and have a sub-metric resolution for example the UC MECER Dataset, INRIA Aerial image labeling dataset, ISPRS Vaihingen (Cheng et al., 2017; Maggiori et al., 2017), and the DeepGlobe challenge (Demir et al., 2018). Furthermore, the images are usually recent and the process of creating reference data comparatively less challenging. In most of the existing works, both multispectral and panchromatic channels usually exist, and can be fed into the deep learning network by way of a pansharpened multispectral image, or separately feeding the channels and combining the feature maps at different stages of the network (Bergado et al., 2018; Gaetano et al., 2018; Kolokoussis et al., 2011; Liu et al., 2018; Tan et al., 2018). Historical photographs, however, have lower spatial and spectral resolution and lower quality because of degradation. External reference data is also either completely absent, incomplete, or not adapted to modern needs.

In this paper, our main objective is thus to develop a methodology for land cover extraction from historical aerial black and white photographs using deep learning. Our hypothesis is that (1) the deep learning networks will be robust to the various types of image distortions characteristic of historical aerial black and white photographs and (2) that it would outperform traditional machine leaning methods. We utilize two existing networks that have fully convolutional architectures as well as an object-oriented approach using Random forest classification (Breiman, 2001) and compare their performance. We test the robustness of the networks with datasets that cover large geographical areas for three cities in equatorial Africa. The uniqueness of our work stems from the fact that only panchromatic images are available for the study sites. Moreover, the three cities under study have diverse appearance of the landscape, which poses more challenges as opposed to the use of existing well-curated benchmark datasets. To the best of our knowledge, only one other work has focused on evaluating deep learning models using historical aerial photographs (Ratajczak et al., 2019). Their work carries out an extensive evaluation of texture features and makes a comparison to patch-based deep learning models in generation of land use classes from a historical aerial photographs dataset of France between 1970 and 1990.

The contributions of this work are:

i) Development of a methodology based on deep learning using fully convolutional networks for land cover classification from historical black and white photographs

ii) Generation of a historical land cover map for three cities in Central Africa that could be used to support long baseline studies in the region.
2 Materials and Methods

2.1 Study area
The study focuses on the cities of Bujumbura, Bukavu and Goma-Gisenyi (Figure 1). These cities, sited in the mountainous environments of the western branch of the East African Rift, had different urban structures in the 1950s (see section 2.3.2) and contrasting topographic landscapes. Goma-Gisenyi are sited in a plain along Lake Kivu. Bukavu develops on a hilly landscape with steep slopes and very few flat areas. Bujumbura extends mostly over a plain but its outskirts extend over very steep slopes. These characteristics make them interesting candidates for the methodological developments for land cover classification from historical photographs.

![Figure 1: Location map of the cities imaged with historical aerial photographs (yellow circles), i.e. Goma 1947, Bukavu 1959 (Democratic Republic of Congo), Gisenyi (Rwanda) and Bujumbura 1957/58/59 (Burundi). The green zones indicate the extent of the available historical aerial photographs. Goma and Gisenyi are adjacent and referred to as Goma-Gisenyi. The processing extent of Bukavu and Bujumbura is indicated by the black frames. The triangles indicate the Virunga volcanoes, with the two active ones, i.e. Nyiragongo and Nyamulagira in red.](image-url)
At the continental level, the cities are in a region that is one of the most exposed to seismic hazard (Delvaux et al., 2017; Smets et al., 2016). In addition, Bukavu and Bujumbura are exposed to landslides and flood hazards (Depicker et al., 2020; Dille et al., 2019; Monsieurs et al., 2018; Nibigira et al., 2018; Nobile et al., 2018) while Goma and Gisenyi are exposed to volcanic hazards (Chirico et al., 2009; Poppe et al., 2016; Smets et al., 2017, 2010). The region has also one of the highest population densities of Africa (frequently > 300 inhab/km²), and urban expansion is very fast and often uncontrolled, anthropogenic pressure high, transportation accessibility low and population vulnerability high (Linard et al., 2012; Michellier et al., 2016). The three cities have a population of circa 800,000 to 900,000 each. In the 1950s the cities had a population of a few thousands and their structure was the result of well-followed planning measures (Michellier et al., 2020, 2018).

This uncontrolled expansion, often due to weak governance, political instability, socioeconomic issues and insecurity, partly explains why the cities are poorly adapted to the environmental constraints associated with the natural hazards in the region (Michellier et al., 2020, 2018; Trefon, 2016). This stresses that, in addition to being relevant for methodological achievements, producing land cover datasets for the region is also of societal relevance.

2.2 Historical black and white photographs and orthomosaics

The aerial orthomosaics used in this work have been generated from the collections of aerial photographs of the Royal Museum for Central Africa (RMCA) and according to the photogrammetric processing workflow described in (Smets et al., 2020). The historical aerial photographs archived at the RMCA correspond to paper reproductions of aerial surveys performed during the 1940's and 1950's by local colonial, Belgian and French geographical institutes. The exploited orthomosaics are of limited quality for three main reasons. First, the quality of the paper photos is usually relatively poor, depending on the dataset and the photo band. This is due to low quality imaging (e.g., blurring effects, noisy chemical emulsion, under- and overexposure), aging effects (e.g., cracks, chemical alteration), non-optimal long-term archiving conditions (e.g., humidity spots and paper deformation), and/or man-made damages (e.g., permanent marking, deformed photo surface due to pencil marking). Second, the orthomosaics are based on scanned photos that already are photo reproductions. Consequently, additional blurring, vignetting effects, bad exposure and optical distortions are observed. Third, the quality of orthomosaic production is limited due to the absence of camera calibration reports associated with the aerial datasets and the unavailability or insufficient number of precise ground control points for proper georeferencing and orthorectification of the data. In addition of all these limitations, aerial surveys from three different years (1955, 1958, 1959) were merged as a single orthomosaic for the city of Bujumbura, in order to ensure a complete geographical coverage of the city. This merging led to brightness difference between the different band paths and thus to significantly lower image quality

It is evident that these datasets are characterised by a high level of complexity, motivating the need for a robust deep learning paradigm that detects the predominant classes at that epoch, namely buildings, high vegetation, bare land low vegetation
2.3 Method

2.3.1 Convolutional neural networks

In this work, we make use of the fully convolutional network architectures shown in in Figure 2. We first evaluate a fully convolutional architecture that is made of atrous convolutional layers developed in (Mboga et al., 2019) and referred to as FCN-ATR-SKIP. The network was designed with an aim of improving classification of buildings from very high-resolution aerial imagery. The network comprises convolutional layers only but does not have downsampling layers. Instead, atrous convolutions (Persello and Stein, 2017; Wu et al., 2018; Yu and Koltun, 2016), which involve interspersing the kernel values with zeros, are used to increase the field of view without raising the number of parameters in the network. The network comprises six convolutional layers, whose outputs are concatenated (through skip connections that reintroduce high spatial features that are degraded due to repeated convolution operations from lower layers to higher convolutional layers) and fed through a softmax layer to generate the pixel-wise class distribution scores.

For comparison we also use an adaptation of the U-NET architecture developed in (Ronneberger et al., 2015) for the semantic segmentation of biomedical images. The network has a characteristic U-shape, comprising encoding and decoding branches. In the encoding branch, downsampling is realised through maxpooling layers with a pooling size of $2 \times 2$, that halve the length and width dimensions of the feature maps. The expansion branch is made possible through use of transpose convolutional layers that allow for learned upsampling. At each resolution level, feature maps are copied from the contracting layers and concatenated to their corresponding expanding layers through skip connections. The Rectified Linear Unit activation (RELU) is applied after each convolution operation. Further, in the encoding layers, Batch Normalisation (BN) layers are applied before the maxpooling (MP) layers.
Model learning

We use visual image interpretation to prepare the reference data. The networks are different in that in order to increase the field of view, FCN-ATR-SKIP uses dilated convolutions while the U-NET uses downsampling layers. The impact of these differences will be visible in the generation of the final land cover maps and useful for understanding the nature of the respective architectures. For the deep learning experiments, we used Python 3.7 with Keras API and TensorFlow 1.13.1 deep learning framework, an NVIDIA GPU 1080 GTX with 8 GB of GPU-RAM.

2.3.2 Experimental setup

The reference data for training and testing is prepared through visual interpretation. First, three main classes, namely buildings, high vegetation and mixed bare land and low vegetation are identified and delineated. Next, using the unsupervised k-means clustering in GRASS GIS software (Neteler et al., 2012), pixels in the mixed bare land and low vegetation are grouped into four subclasses with varying grey levels. We selected this strategy to avoid having a very heterogeneous class that would have been unnecessarily difficult to classify. The generated homogeneous sub-classes would be used to train the network, and later, the subclasses would...
be merged as shown in Figure 3. We use fully labelled patches of 128 × 128 pixels as inputs to the networks. The training of the networks is done by minimizing the cross-entropy loss using the stochastic gradient descent optimizer with a momentum of 0.8, a learning rate of 0.1, and a learning rate decay of 0.001, using a batchsize of 16 patches over 150 epochs. The size of the training set is 20800, 21400 and 16000 patches respectively for Goma-Gisenyi, Bukavu and Bujumbura split in the ratio of 80/20 for training and validation sets.

<table>
<thead>
<tr>
<th></th>
<th>Panchromatic image</th>
<th>Reference (6 classes)</th>
<th>Reference (3 classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goma-Gisenyi</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td>Bukavu</td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
</tbody>
</table>
Figure 3 Raw and reference image for scenes in Goma-Gisenyi, Bukavu and Bujumbura. The second and third column show the mixed bare land and low vegetation class and its subclasses and when merged respectively. In the 6-class legend the classes are 1- building, 2-high vegetation, (3-6) mixed bare land and low vegetation and in the 3-class legend1- building, 2 high vegetation, 3 mixed bare land and low vegetation. 0 represents the unclassified pixels.

Figure 4, shows the locations of independent training and testing tiles are shown for Goma-Gisenyi, Bukavu and Bujumbura.
Figure 4: Locations of the training and testing tiles in the images of (A) Goma-Gisenyi, (B) Bujumbura and (C) Bukavu used in the experiments covering a geographical area of 80km$^2$, 265km$^2$ and 258km$^2$ respectively.

The total number of pixels per class present in each training set for each city is presented in Table 1. The city of Goma-Gisenyi was largely undeveloped, with few, small-sized buildings
and mostly covered with high vegetation and mixed bare land and low vegetation which would explain the low number of pixels from the building class.

Table 1: Number of pixels per class ×10^6 present in the training data for the cities of Goma-Gisenyi, Bukavu and Bujumbura. the classes are 1) building, 2) high vegetation and (3-6) represent subclasses of the mixed bare land and low vegetation

<table>
<thead>
<tr>
<th>City</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goma-Gisenyi</td>
<td>6.1×10^6</td>
<td>96×10^6</td>
<td>52.4×10^6</td>
<td>70.9×10^6</td>
<td>60.7×10^6</td>
<td>23.1×10^6</td>
</tr>
<tr>
<td>Bukavu</td>
<td>18.8×10^6</td>
<td>46×10^6</td>
<td>51.8×10^6</td>
<td>103×10^6</td>
<td>57.8×10^6</td>
<td>10.1×10^6</td>
</tr>
<tr>
<td>Bujumbura</td>
<td>17.8×10^6</td>
<td>11×10^6</td>
<td>33.7×10^6</td>
<td>62.6×10^6</td>
<td>57.2×10^6</td>
<td>47.8×10^6</td>
</tr>
</tbody>
</table>

2.3.3 Object-based image analysis (OBIA) with Random Forest Classifier

For the baseline experiment, we adopt an object-based image analysis approach with a random forest classifier to generate the classes. OBIA entails generation of homogeneous and disjoint regions (segments) based on the input image which can then be labelled manually or automatically using a classification algorithm (Blaschke et al., 2014). We use a complete OBIA toolchain developed using the GRASS GIS software (Grippa et al., 2017b). For the choice of segmentation parameters, we make use of the GRASS module “i.segment.uspo” that optimizes the search of segmentation parameters within locally defined zones (Grippa et al., 2017a). Here, each image was split into 1km² grid cells, and in each of the cells the “threshold” parameter that controls over- or under- segmentation was varied in the range [0.001,0.05]. Values of this parameter close to zero produce an over-segmented output while values close to 1 produce an under-segmented output. The smallest size of segments was set to 8 pixels, as preliminary tests revealed that a lower value would result into an over-segmented output.

Further, using the GRASS GIS add-on “r.texture”, texture features based on GLCM namely Variance, Entropy, Angular Singular Moments and Contrast were computed on the panchromatic image using a moving window of size 11×11 pixels, averaging the results in four directions [0°,45°,90° and 135°] (Haralick et al., 1973). For each segment, a total of 51 statistics were calculated using the “i.segment.stats” add-on: aggregate statistics, for both the original and texture bands, such as minimum, maximum, range, mean, standard deviation, 1st, 3rd quartiles, 90th percentile, as well as geometric features characterizing the segments such as area, perimeter, circle and square compactness and fractal. For each of the three cities, 14000 points randomly distributed over the labelled training tiles were used to extract segments and used to train a random forest classifier setting the number of trees to 500.

2.3.4 Accuracy Assessment

An independent reference set is used to test the models. All labelled pixels in each of the test tile are used as test sets. A confusion matrix is generated from which other metrics such as the producer accuracy (PA), the user accuracy (UA), the overall accuracy (OA) are calculated (Congalton, 1991). Also, F1 score of the three classes is computed according to the formula:

\[
F1\text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]
In addition to these quantitative metrics, a visual quality assessment of the classified maps is done and compared to the reference dataset, in order to detect issues not easily measured with the above indicators.

3 Results and Discussion

After training, the network is used to perform prediction on the entire image mosaics of the cities. Table 2 presents the accuracy measures of the three-class legend. Accuracy metrics are computed on a set of labelled tiles that were excluded from the training set. The accuracy metrics computed over the six training classes is presented in Table A1 in the Appendix.

Table 2: Producer accuracy (PA), user accuracy (UA), F1 score and overall accuracy (OA) for FCN-ATR-Skip, U-NET and OBIA-RF for the cities of Goma-Gisenyi, Bukavu and Bujumbura. The classes are represented as 1-buildings, 2-High vegetation and 3-Mixed bare and Low vegetation. The number of pixels for each class instance are displayed.

<table>
<thead>
<tr>
<th>Class</th>
<th>FCN-ATR-Skip</th>
<th>U-NET</th>
<th>OBIA-RF</th>
<th>Test Pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA</td>
<td>UA</td>
<td>F1</td>
<td>PA</td>
</tr>
<tr>
<td>Goma-Gisenyi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td>69.32</td>
<td>73.27</td>
<td>0.71</td>
<td>85.16</td>
</tr>
<tr>
<td>High veg</td>
<td>95.06</td>
<td>97.05</td>
<td>0.96</td>
<td>98.95</td>
</tr>
<tr>
<td>Mixed bare/low veg</td>
<td>96.25</td>
<td>93.79</td>
<td>0.95</td>
<td>98.74</td>
</tr>
<tr>
<td>OA</td>
<td>95.54</td>
<td>98.83</td>
<td>78.02</td>
<td>4599088</td>
</tr>
<tr>
<td>Bukavu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td>75.92</td>
<td>72.57</td>
<td>0.74</td>
<td>88.61</td>
</tr>
<tr>
<td>High veg</td>
<td>92.4</td>
<td>88.29</td>
<td>0.90</td>
<td>97.83</td>
</tr>
<tr>
<td>Mixed bare/low veg</td>
<td>94.26</td>
<td>95.91</td>
<td>0.95</td>
<td>94</td>
</tr>
<tr>
<td>OA</td>
<td>92.77</td>
<td>94.48</td>
<td>72.28</td>
<td>4572214</td>
</tr>
<tr>
<td>Bujumbura</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td>59.82</td>
<td>74.34</td>
<td>0.66</td>
<td>65.81</td>
</tr>
<tr>
<td>High veg</td>
<td>77.26</td>
<td>49.06</td>
<td>0.60</td>
<td>76.94</td>
</tr>
<tr>
<td>Mixed bare/low veg</td>
<td>89.77</td>
<td>96.11</td>
<td>0.93</td>
<td>89.06</td>
</tr>
<tr>
<td>OA</td>
<td>87.87</td>
<td>87.33</td>
<td>85.22</td>
<td>3077810</td>
</tr>
</tbody>
</table>

For the city of Goma-Gisenyi, high values for OA is obtained by both U-NET and FCN-ATR-Skip at 98.83% and 95.54% respectively. U-NET performs better for the three classes considering the PA, UA and F1 score. The PA and UA of FCN-ATR-Skip for the building class is much lower than U-NET. OBIA-RF has a low UA for buildings but high PA for mixed bare and low vegetation.

In the city of Bukavu, both deep learning networks report high OA values at 92.77 and 94.48 for FCN-ATR-Skip and U-NET respectively. However, FCN-ATR-Skip outperforms the U-NET in the building class by having a higher UA of 72.57 against 58.54 of U-NET. On looking at the PA, the results indicate that U-NET has a higher PA than FCN-ATR-Skip, indicating that it tends to overpredict the building class. However, for the Mixed bare land and low vegetation class, both networks produce high classification metrics. The accuracy metrics of
the building class in OBIA-RF are comparatively lower with a PA and UA of 62.69 and 31.6 respectively, suggesting poor generalization on a complex dataset.

In the city of Bujumbura, with its lower image quality, both deep learning networks have lower overall accuracy values as compared to the Bukavu and the Goma-Gisenyi images. The OA values are 87.87 and 87.33 for FCN-ATR-Skip and U-NET respectively. For FCN-ATR-Skip, the PA of the building class is lower than the PA from U-NET. The high vegetation class has lower user accuracy in the three methods. However, the Mixed Bare and Low vegetation class has high values for the UA and F1 scores. As for the OBIA-RF, high PA, UA and F1 are obtained for the mixed bare and low vegetation class comparable to the deep learning networks as opposed to the building and high vegetation classes. In order to further assess the classification performance of the networks, a visual inspection of the classified maps was conducted. Sample classified scenes for the deep learning networks are displayed Figure 5.
The quality of the classified maps produced by the deep learning networks is high as shown in Figure 5. There is a good distinction between the three classes namely high vegetation, building and mixed bare land and low vegetation. The image of Goma-Gisenyi had small sized buildings, but these are well captured by both networks. We observe that the classified maps are much smoother in the case of U-NET as opposed to the FCN-ATR-SKIP. This can be seen by inspecting the high vegetation class in the Figure 5.

Whereas the mixed bare land and low vegetation class was heterogeneous in the orthomosaic, it was consistently well classified by the networks as observed from the classification metrics in Table 2 and the maps in Figure 5. This good performance can be attributed to the strategy of using unsupervised k-means to generate four sub-classes, that are used in the training and predictions stage. High classification metrics were attained for the subclasses as seen in Table A1.

The baseline method overpredicts the building class in Bukavu while having challenges classifying the high vegetation class in both Bujumbura and Goma-Gisenyi. The quality of the OBIA-RF classification is dependent on the quality of the segmentation step necessary for generating distinct homogeneous zones. In each of these images, the similarity in spectral signature for different classes owing to the nature of the old aerial photographs could result in a less optimal segmentation output. An intensive feature extraction scheme might be needed, although likely to be limited by the fact that the starting point is a single channel from old aerial panchromatic photographs.

Despite good quality of the predictions, we observed some examples where the performance of the networks was challenged. Examples of such scenes are presented in Figure 6. For both cases, U-NET tends to predict instances of the building class as a blob, whereas the FCN-ATR-SKIP tends to differentiate between the building class and the Mixed bare and low vegetation classes.
One of the reasons for this is the poor radiometric quality of the scenes, that makes the task of differentiating between the building class and mixed bare and low vegetation class a challenge. Another reason for the performance of the U-NET lies in the downsampling property of the U-NET. With each downsampling, the contextual size of the input doubles. Due to poor contrast between the neighbouring classes, the classification output turns out as a blob. On the other hand, FCN-ATR-Skip does not have the downsampling operation, since it uses atrous convolutions. This could explain why the FCN-ATR-Skip seems more robust in scenes with poor radiometric quality as opposed to the U-NET.

The better performance of the deep learning networks can be attributed to the discriminative features learned directly from the panchromatic images as opposed to the standard baseline method where manual feature extraction had to be conducted. Due to the challenging nature of the dataset, it is evident that complex discriminative features were needed for the discrimination of the land cover classes. The comparatively higher UA for the deep learning networks as opposed to the standard baseline method suggest better generalisation capability. That notwithstanding, the OBIA-RF approach had a comparable performance for the mixed bare and low vegetation class.

The fully convolutional networks were conducted on a GPU and the training time for U-NET ranged between 8 and 10 hours across the three datasets. In all cases, U-NET was faster than the FCN-ATR-Skip in the training stage on average by 90 minutes. This is due to U-NET having downsampling layers that reduce the dimension of the height and width of the feature
maps, whereas the dimensions remain the same in the FCN-ATR-Skip at each convolutional layer. However, both networks predicted the entire images of the cities, corresponding to ≈15,000 patches of 128 × 128 pixels on average of 2 minutes.

4 Conclusion

In this work, we have proposed a deep learning approach that is based on fully convolutional networks to generate landcover maps from historical aerial black and white photographs. High classification accuracies of > 90% for the city of Goma-Gisenyi and Bukavu and > 85% for the city of Bujumbura have been obtained. The deep learning approach has clearly outperformed a standard baseline approach in comparative tests.

While deep learning algorithms are robust, data quality of the raw input image and the reference data can affect the quality of the outputs. This is more evidenced in the classification results of Bujumbura where the image comprised a mosaic of three years 1957, 1958, and 1959. We observed that the high vegetation class appeared variously on different image strips, resulting in remarkable radiometric differences, hence lower user accuracy. Furthermore, the vegetation signature can change according to the seasonality pattern of the rain (if the photos are taken at different time during the same year). Nonetheless, the high overall accuracy and F1 scores indicate strong generalisation ability of the deep networks.

Numerical and qualitative results indicate the applicability of the deep learning methodology based on fully convolutional networks for landcover classification from historical black and white photographs. While FCN-ATR-Skip appears robust to poor radiometric contrast, the U-NET architecture generates smoother predictions which makes it suitable for the extraction of high vegetation and the mixed bare and low vegetation classes.

During the 1940s and 1950s, most cities in sub-Saharan Africa were sparsely settled, and were characterised by extensive areas of forests, bare land and low vegetation with some being agricultural areas. Thus, developing a methodology that can discriminate these classes with a high confidence is desirable, a goal that our methodology will contribute to.

A bottleneck in the proposed methodology is that the reference data had to be generated through visual image interpretation, which was not a trivial task. This would imply that for each new city, new reference data would have to be generated reducing the scalability of the approach. However, there have been ongoing works exploring the use of crowdsourced geospatial data to train deep learning networks on recent multispectral aerial imagery (Kaiser et al., 2017). While the use of crowdsourced data such as open street maps may not be directly applicable for historical imagery acquired in the mid-20th century over cities in Central Africa, the use and development of domain adaptation techniques capable of exploiting the available reference data for one city at one date and providing high classification accuracies for a land cover mapping task at another date (or another city) is an avenue for future works.
The generated maps are a first result and provide an insight into the landscape of 1940s and 1950s in Central Africa. They will be post-processed and used as inputs for the generation of land use maps. They will be utilised for characterising the links between humans and land degradation processes in the study area. Lastly, the methodology developed in this work is a first step to proving the utility of digitizing analogue archives for the support of modern scientific research.

5 Acknowledgements

This research is funded by the Belgian Science Policy Office (BELSPO) in the frame of the Historical Aerial photographs and Archives to Assess Environmental Changes in Central Africa (PAS\text{ECA}) project (BR/165/A3/PASTECA; http://pasteca.africamuseum.be/).

6 Appendix

Table A1: The Producer accuracy (PA), user accuracy (UA), F1 score and overall accuracy (OA) for FCN-ATR-SKIP, U-NET and OBIA_RF for the cities of Goma, Bukavu and Bujumbura. The classes are represented as 1-buildings, 2-high vegetation and (3-6) mixed bare land and low vegetation.

<table>
<thead>
<tr>
<th>City</th>
<th>Class</th>
<th>FCN-ATR-SKIP</th>
<th>U-NET</th>
<th>OBIA-RF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PA</td>
<td>UA</td>
<td>F1</td>
</tr>
<tr>
<td>Goma-Gisenyi</td>
<td>1</td>
<td>69.32</td>
<td>73.27</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>95.06</td>
<td>97.05</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>90.82</td>
<td>85.93</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>89.47</td>
<td>86.89</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>88.57</td>
<td>88.87</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>94.71</td>
<td>93.89</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>OA</td>
<td>92.94</td>
<td>97.61</td>
<td>97.68</td>
</tr>
<tr>
<td>Bukavu</td>
<td>1</td>
<td>75.92</td>
<td>72.57</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>92.4</td>
<td>88.29</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>90.08</td>
<td>87.45</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>91.38</td>
<td>92.77</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>88.97</td>
<td>91.61</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>78.76</td>
<td>86.57</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>OA</td>
<td>89.22</td>
<td>92.72</td>
<td>96.05</td>
</tr>
<tr>
<td>Bujumbura</td>
<td>1</td>
<td>59.82</td>
<td>74.34</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>77.26</td>
<td>49.06</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>68.57</td>
<td>87.35</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>89.14</td>
<td>92.63</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>94.21</td>
<td>91.79</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>90.69</td>
<td>92.42</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>OA</td>
<td>83.88</td>
<td>84.87</td>
<td>73.68</td>
</tr>
</tbody>
</table>

7 References

https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.03.044


