Revised: 11 October 2018 | Accepted: 26 October 2018

DOI: 10.1002/arp.1731

RESEARCH ARTICLE

WILEY

Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland

Øivind Due Trier[1] | David C. Cowley[2] | Anders Ueland Waldeland[1]

[1] Section for Earth Observation, Norwegian Computing Center, Oslo, Norway

[2] Historic Environment Scotland, Edinburgh, UK

Correspondence
Øivind Due Trier, Norwegian Computing Center, Section for Earth Observation, Gaustadalléen 23A, P.O. Box 114 Blindern, NO-0314 Oslo, Norway.
Email: trier@nr.no

Funding information
Royal Society of Edinburgh, Arts & Humanities Awards Small Grant

Abstract

This article presents results of a case study within a project that seeks to develop heavily automated analysis of digital topographic data to extract archaeological information and to expedite large area mapping. Drawing on developments in computer vision and machine learning, this has the potential to fundamentally recast the capacity of archaeological prospection to cover large areas and deal with mass data, breaking a dependency on human resource. Without such developments, the potential of the vast amount of archaeological information embedded in large topographic and image-based datasets cannot be realized. The purpose of the case study reported on here is to assess existing developments in a Norwegian study against digital topographic data for the island of Arran, Scotland, examining the transferability of the approach and providing a proof of concept in a Scottish context. For Arran, three monument classes were assessed – prehistoric roundhouses, shieling huts of medieval or post-medieval date, and small clearance cairns. These present different challenges to detection, with preliminary results ranging from a manageable mix of false positives and true identifications to the chaotic. The influence of variable morphology and the occurrence of other, largely natural, objects of confusion in the landscape is discussed, highlighting the potential improvements in automated detection routines offered by adding anthropogenic and natural false positives to additional confusion classes.

KEYWORDS

airborne laser scanning, archaeological survey, computer vision, convolutional neural network, deep learning, transfer learning

1 | INTRODUCTION

Archaeological prospection and surveys have long relied on human observation, whether in the field or through desk-based work, for the identification of objects of interest (e.g. Bowden, 1999; Opitz & Cowley, 2013). In this approach, rates of coverage are inherently limited by the availability of human resource. This means that achieving a systematic large-area or national mapping of Scotland's archaeological remains, or indeed remains over a large region anywhere else, is a distant prospect, unachievable even over many decades. At the same time, extensive high-resolution topographic data are becoming more widely available, presenting a challenge to the capacity of human observer-based approaches to explore it. Developments in heavily automated data processing (e.g. Hesse, 2013) and computer vision offer a way forward to efficiently and rapidly explore these data and identify archaeological information (e.g. Bennett, Cowley, & De Laet, 2014; Sevara, Pregesbauer, Doneus, Verhoeven, & Trinks, 2016). For archaeological survey, computer vision offers the potential for a step-change in rates of coverage, and a mechanism to exploit the vast amount of archaeological information embedded in large topographic and image-based datasets.

A key driver for the exploration of such approaches is the recognition that existing heritage datasets are unrepresentative (Banaszek, Cowley, & Middleton, 2018; Cowley, 2016) and are a biased basis for management and research. An additional motivation is the ongoing acquisition by the Scottish Government of airborne laser scanning (ALS, aka airborne LiDAR) data, and the proliferation of aerial and satellite imagery, which all have enormous potential for archaeological mapping. However, to exploit this potential requires development of analytical methods that can efficiently deal with mass data, and computational approaches drawing on convolutional neural networks (CNNs) offer a way forward for archaeological prospection (Trier, Salberg, & Pilø, 2018). These draw on learning sets to 'look' at data in a manner inspired by the organization of the animal/human visual cortex (LeCun, Bengio, & Hinton, 2015), and have the potential to be more successful than past approaches to automated detection.

The work presented in this article was commissioned by Historic Environment Scotland (HES), the lead public body for Scotland's historic environment, and was designed to contribute to a range of HES strategic priorities. These include the need to build better knowledge of where the material remains of past activities survive, without which understanding of Scotland's history will be limited. This recognizes that the National Record of the Historic Environment (NRHE) is a partial record built up piecemeal over more than a century, and its contents are neither systematic nor representative of the archaeology that survives in the landscape (Banaszek et al., 2018; Cowley, 2016). Indeed, in Scotland there are extensive relict landscapes of preserved archaeological micro-topographic remains, and systematic survey in most parts of the country generates large increases (e.g. up to tenfold) in the numbers of known monuments. However, only about 10% of the country has been covered in this way, and this leads to an expectation that there are hundreds of thousands of unrecorded monuments preserved in the micro-topography of the landscape. This knowledge gap is a significant limitation on understanding, management and protection of Scotland's archaeological assets. Developing approaches to mass data analysis provides a mechanism to explore topographic data and imagery to provide systematic large-area coverage within short periods of time that can form the basis for a better understanding of the past, and for more effective management strategies, especially in the face of challenges such as climate change and human-induced landscape-scale changes. This work also addresses an ongoing paradigm-shift in archaeological prospection in response to the proliferation of digital data. If archaeological survey is to engage effectively with the complexity and scale of datasets such as ALS, this demands the development of analytical processes that are less dependent on human observer-based approaches.

2 | BACKGROUND

Automatic and semi-automatic methods for detection and mapping of archaeological structures from remote sensing data (Table 1) have emerged over the last 12 years (see Traviglia, Cowley, & Lambers, 2016, for a discussion of this trajectory). In an early work, Bescoby (2006) used the Radon transform to detect Roman land boundaries from aerial photographs. Template matching has been used to map burials from optical satellite data (Trier et al., 2009), and to identify a range of objects including pitfall traps, charcoal burning platforms, and grave mounds in a digital terrain model (DTM) derived from ALS (Schneider, Takla, Nicolay, Raab, & Raab, 2015; Trier & Pilø, 2012; Trier & Pilø, 2015; Trier, Pilø, & Johansen, 2015; Trier, Zortea, & Tonning, 2015). Also based on a DTM is an automatic pit filling method based on an inverted DTM to locate mound structures (Freeland et al., 2016); a combination of curvature estimates, topographic position index, and circular Hough transform to detect prehistoric barrows (Cerrillo-Cuena, 2017); a combination of segmentation and template matching to detect grazing structures (Toumazet et al., 2017); and local contrast in the DTM at three different scales and a random forest classifier to detect burial mounds (Guyot et al., 2018). A study to detect rectangular enclosures in panchromatic satellite images (Zingman et al., 2016) concluded that bespoke methods in some cases perform better than using a pretrained deep CNN, but at the cost of much longer development time. However, the use of a deep CNN for charcoal burning platforms showed considerable improvement on an earlier template matching approach (Trier et al., 2018). The term 'deep' is used to emphasize that the neural network has several layers between the input layer and output layer, whereas early neural networks only had three layers. Most of the existing works (Table 1) report their success in terms of rates of true positives and false positives, as indicating how good each method was in solving the particular automation problem at hand. However, for a number of reasons, these figures do not support systematic assessment of the effectiveness of the methods. Firstly, the quality of the data (i.e. how well archaeological structures are visible in the data, and how well they stand out from modern structures and natural terrain features) varies considerably between studies. Secondly, the number of training and test examples also vary between the studies, and thirdly, each method has a different set of parameters that are estimated during training. These issues highlight the challenges of creating transferable automated approaches across differing archaeological remains and terrains, with the added difficulty that large numbers of parameters require large numbers of training examples.

Notwithstanding these issues, the increasing number of case studies that demonstrate the utility of such automated approaches, and the rapid rate of development, has prompted HES to explore this methodology. Recognizing the challenges of transferability in existing studies, the development of a proof of concept for Scotland looked to capitalize on work undertaken elsewhere in areas with similar archaeological morphology and topography. The broadly similar forms of topographic expression amongst archaeological remains in parts of Norway and Scotland made the partnership between HES and the Norwegian Computing Center (NCC) an attractive proposition. Moreover, the broad framework of research by the NCC on semi-automatic methods for the detection and mapping of cultural heritage remains, with an overall aim to develop methods that can be used in national cultural heritage infrastructure contexts (Kermit, Hamar, & Trier, 2018), fitted well with HES' aspiration to improve the coverage of the NRHE in Scotland.

TABLE 1 Previous studies in rapid and/or automated archaeological mapping

Authors	Objects to detect	Remote sensing data	Method	True positive rate	False positive rate
Hesse, 2013	Potential archaeological features	Airborne laser scanning (ALS), 1/m^2	Manual interpretation of digital terrain model (DTM) visualization		
Bescoby, 2006	Roman land boundaries	Historic aerial photos	Radon transform		
Sevara et al., 2016	Burial mounds in grave field	ALS, 6/m^2	DTM openness + segmentation	91%	6%
Sevara et al., 2016	Various archaeological features	ALS, 5/m^2	DTM openness, slope, roundness + segmentation	100%	35%
Zingman, Saupe, Penatti, & Lambers, 2016	Fragmented rectangular enclosures	Satellite, optical 0.5 m	Rectangle detector	100%	34%
Zingman et al., 2016	Fragmented rectangular enclosures	Satellite, optical 0.5 m	Pre-trained deep convolutional neural network (CNN)	100%	124%
Trier et al., 2018	Charcoal burning platforms	ALS, 5/m^2	Template matching	70%	72%
Trier et al., 2018	Charcoal burning platforms	ALS, 5/m^2	Pre-trained deep CNN + support vector machine classifier	86%	37%
Trier, Larsen, & Solberg, 2009	Cropmarks of levelled grave mounds	Satellite, optical 0.5 m	Template matching		
Trier & Pilø, 2012	Pitfall traps	ALS, 7/m^2	Template matching + if-tests	86%	92%
Trier, Pilø, & Johansen, 2015	Burial mounds in grave field	ALS, 7/m^2	Template matching	65%	
Trier, Zortea, & Tonning, 2015	Grave mounds in forest	ALS, 1–22/m^2	Template matching + if-tests	50%	375%
Freeland, Heung, Burley, Clark, & Knudby, 2016	Earthworks mounds	ALS, 1/m^2	DTM local relief, ratios + segmentation	71%	14%
Freeland et al., 2016	Earthworks mounds	ALS, 1/m^2	Inverted pit filling	85%	18%
Cerrillo-Cuena, 2017	Prehistoric barrows	ALS, 0.5/m^2	Curvature, topograpic position index, circular Hough transform	46%	
Toumazet, Vautier, Roussel, & Dousteyssier, 2017	Grazing structures	ALS, 11/m^2	DTM local relief, segmentation, template matching	91%	34%
Guyot, Hubert-Moy, & Lorho, 2018	Burial mounds	ALS, 14/m^2	DTM local contrast at three scales, random forest classifier	98%	1%

2.1 | Towards neural networks

In developing semi-automated methods, the NCC research framework has taken two different approaches, as outlined earlier, and discussed in a little more detail here. Template matching proved successful for the detection of pitfall traps in Oppland County, Norway (Trier & Pilø, 2012). In this landscape, the pitfall traps stood out as clear anthropogenic structures in the ALS-derived DTM. Although the template matching gave some false positives, many of these were removed by including additional tests on the shape of each detected pit. Template matching was then used to detect pits and mounds on iron extraction sites, grave mounds and charcoal burning platforms. These results were less convincing than for the pitfall traps. In the case of the iron extraction sites and grave mounds this is probably because they were less distinct than the pitfall traps, and did not stand out from the natural terrain to the same degree. However, these reasons seemed less applicable for the charcoal burning platforms, which are very distinctive to a human observer. This introduced an additional problem, that the charcoal burning platforms had many different appearances in the DTM, and that it was difficult to construct a suitable template or a collection of templates to deal with the variability in form. However, applying both mound and pit detection, using small templates for pits and large templates for mounds produced some improvement in performance. In this case many charcoal burning platforms were detected as a range of features, including a central mound with pits around the circumference, a central mound, and pits in a circular arrangement. Never the less, many charcoal burning platforms were still missed.

Recent advances in computer vision, using deep CNNs, suggested an alternative approach (Trier et al., 2018). Using a network pre-trained on a million natural images, they discarded the last layer of the neural network and replaced it with a support vector machine classifier. By training the support vector machine classifier on 400 examples of charcoal burning platforms and 10 000 random terrain locations, 86% of the verified charcoal burning platforms were correctly detected, compared to 70% for the template matching approach. In an approach implemented in the Caffe library, the false positive rate is 37%, against 72% for template matching, also a significant improvement. The main limitation of the implementation is speed as 1 km by 1 km of DTM data requires several hours of processing. Using a step size of 1 m (5 pixels), the method extracts a 224 × 224 pixels image from the DTM, and scales the floating point elevation values to integer values in the range 0–255. This individual re-scaling is the main bottleneck, and could be avoided in the future by, for example, using local relief visualization (Hesse, 2010). Recently, the PyTorch library has emerged as a better alternative to Caffe, offering more flexibility in training and classification. Accepting the published reservations that deep CNNs might not always be the best solution (Zingman et al., 2016), this approach has been used on Arran

because transferability between datasets and the development of 'general purpose' archaeological CNNs is desirable if the discipline as a whole is to make better use of the methodology. For HES, the similarities in landscape forms, datasets and the basic morphology of archaeological monuments between some of the Norwegian case studies and the Scottish context meant that it was an attractive first step in establishing a proof of concept project on Arran.

3 | ARRAN CASE STUDY

The island of Arran is being used by HES to develop approaches to rapid large area mapping using remote sensing datasets (Banaszek et al., 2018; Cowley & López-López, 2017), and this includes a proof of concept for the use of automated object detection to expedite rates of coverage and to explore how archaeological objects are identified. Arran lies in the west of Scotland, extending to about 432 km^2, and is colloquially known as 'Scotland in miniature' because it has a range of landscapes from highlands to lowlands that are generally representative of the rest of Scotland (Figure 1).

3.1 | Data

The ALS data used for this project were collected between November 2012 and April 2014, commissioned by the Scottish Government, Scottish Environmental Protection Agency, sportscotland and 13 local authorities collectively (DTM and DSM are available under Open Government licence v3.0, the point cloud is available under a Non-Commercial Government Licence). The ALS data for Arran comprise 1 km by 1 km tiles of point cloud data (las file v1.2, compressed to laz), each point classified as one of eight classes, including 'ground', 'building' and 'vegetation'. The average 'ground' point density per square metre was 2.75, but this varies considerably from 0.43 to 7.44 across the 489 tiles depending on vegetation density and the presence of buildings. However, large areas of Arran are open land with low vegetation (Figure 2). From previous fieldwork and visual inspection of ALS-derived visualizations, learning sets comprising several hundred locations of historical structures and some modern structures were created (Table 2) for selected discrete areas of approximately 1 km by 1 km (Figures 3, 4).

4 | METHODS

4.1 | Pre-processing of ALS data

In order to allow for detection of cultural heritage structures at tile boundaries, each tile included a 50 m buffer of data from neighbouring tiles. Then, for each extended tile, all ALS ground points were used to create a DTM at 0.25 m grid spacing. At this resolution, no detail in the ALS data is lost, and the archaeological structures are clearly visible. The DTM was created using the IDL (https://www.harrisgeospatial.com) functions TRIANGULATE and TRIGRID. Since the archaeological remains of interest survive as local elevation differences, a smoothed version of the DTM was subtracted from the DTM, producing a simplified version of a local relief model (Hesse, 2010). The smoothing was

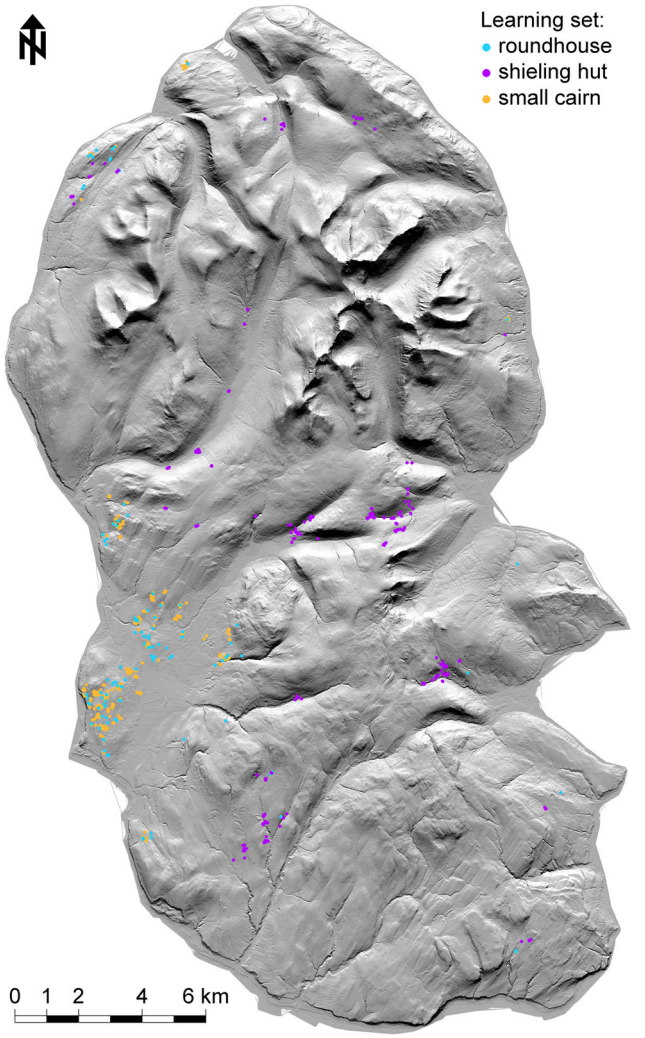

FIGURE 1 Hillshade relief visualization of the digital terrain model (DTM) of Arran, illustrating the complex topography of the island. The learning set locations are superimposed. DTM derived from airborne laser scanning (ALS). The DTM contains public sector information licensed under the Open Government Licence v3.0. [Colour figure can be viewed at wileyonlinelibrary.com]

done for each pixel by taking the mean value within a 30 × 30 pixels sliding window (i.e. 7.5 m × 7.5 m). The resulting image then contained local elevation deviations from the general smoothed terrain surface, the values for which were truncated to the range −2 m to +2 m.

While there are other DTM visualizations (Kokalj & Hesse, 2017) that could have been used instead of, or combined with, the local relief model, for the present study the simplified local relief model (SLRM), replicated for red, green and blue channels, is used. The SLRM, as a normalization of the topographic data that removes the influence of absolute elevation, seems to have some advantages for the present study. It is closer to the 'raw' terrain data than some other visualizations, such as hillshade or sky view factor, which build in more abstraction and modelling of interactions between terrain and light or sightlines. It is rotation invariant and, with the exception of larger, abrupt terrain features, that might distort the local terrain model, generally renders the archaeological remains effectively. However, the influence of different treatments of the raw terrain data on detection outputs is a potentially important issue that requires exploration.

FIGURE 2 This aerial view of the northwest of Arran shows the predominance of low to ground vegetation such as grass and heather, but also the large areas of 20th century coniferous plantations. DP252900 © Historic Environment Scotland [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Archaeological and modern structures in the learning set

Type of structure	Count
Roundhouse	121
Shieling hut	267
Small cairn	384
Burial cairn	6
Burnt mound	24
Cattle feed stance	24
Enclosure	11
Horse platform	1
Possible kiln	5
Rectangular building	15

4.2 | Neural network design

The ResNet18 implementation in PyTorch (https://pytorch.org/) was used as a starting point, an approach that includes a ready-to-use implementation that allows the user to take a network pre-trained as a starting point, and to refine it with their own data. This process is known as 'transfer learning' (e.g. see Liu et al., 2018; Pan & Yang, 2010). The neural network is pre-trained on the ImageNet (http://www.image-net.org/) database of 1.2 million images of natural scenes, each image tagged with one or more of about 1000 unique labels denoting image content. While the ImageNet database will not be representative of archaeological landscapes and sites, for the present there is a hope that the basic elements of low-level image detection are transferable from one type of image to another. That is, that there are common processes in the basic detection of edges, arcs, lines, contrast, texture, and so on, that are learned on one set of images, and may be applicable to other types of image. This too, is an issue requiring further exploration especially as archaeological learning sets may be small in comparison to other training sets, such as ImageNet.

ResNet is designed to work on images 224 × 224 pixels in size. The input layer of ResNet is organized as a 7 × 7 array of image feature detectors, each 32 × 32 pixels in size. By instead using 2 × 2 or 3 × 3 feature detectors, input image sizes of 64 × 64 or 96 × 96 pixels could be used. In addition, the final classification layer that maps to the 1000 ImageNet classes is replaced with a layer that maps to the actual classes of interest. In the present study, this could have been the four classes of roundhouse, shieling hut, small cairn, and background. However, with that design, we found it difficult to optimize performance on all classes simultaneously. So, we instead used three networks: roundhouse *versus* background; shieling *versus* background; and small cairn *versus* background.

4.3 | Training of the neural network

It is recommended to run the training phase on a computer with a graphics processing unit (GPU), otherwise training may be very slow. The workflow for this phase included the following distinct processes – extraction of training images, image augmentation, image cropping, training and validation iterations and storing the results of training.

The training of the neural network was done on image extracts 101 × 101 pixels in size centred on known locations of roundhouses, shieling huts and small cairns. Samples of the 'background' terrain, as a single class and excluding any 'foreground' locations, were also extracted on a random basis (also 101 × 101 pixels), taking care to avoid foreground structures in the learning set. To exclude the foreground structures, buffer zones around these locations were created

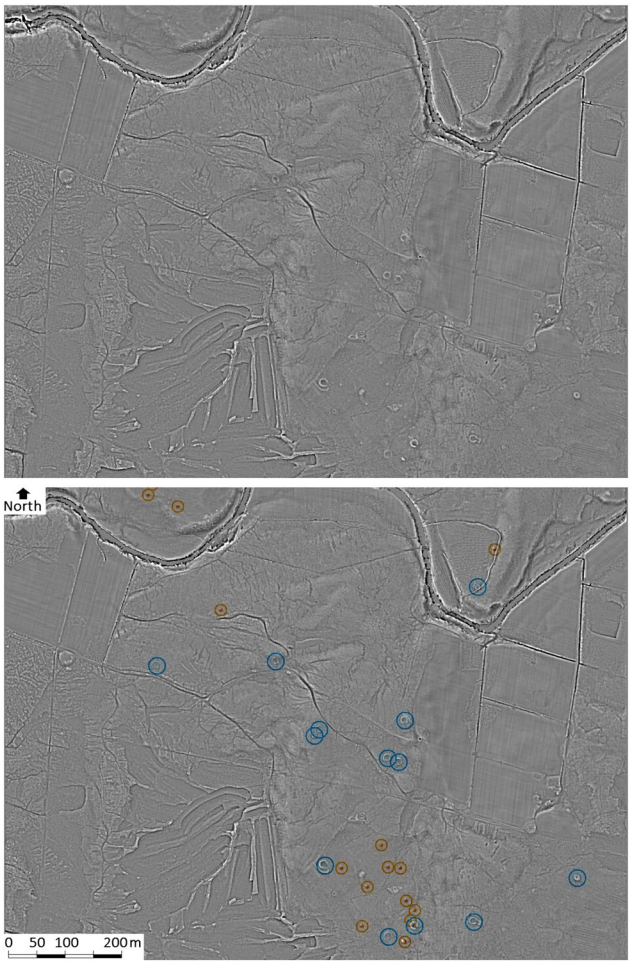

FIGURE 3 (Top) Detail of the digital terrain model (DTM) visualized as simplified local relief model (SLRM) for Machrie Moor on the west of Arran, with the roundhouses, small cairns and other archaeological monuments clearly visible. The 1100 m × 835 m image is centred on 190 600 east, 632 450 north (OSGB national grid). (Bottom) The learning sets of verified roundhouses (blue) and small cairns (brown) for the same area. The DTM contains public sector information licensed under the Open Government Licence v3.0 [Colour figure can be viewed at wileyonlinelibrary.com]

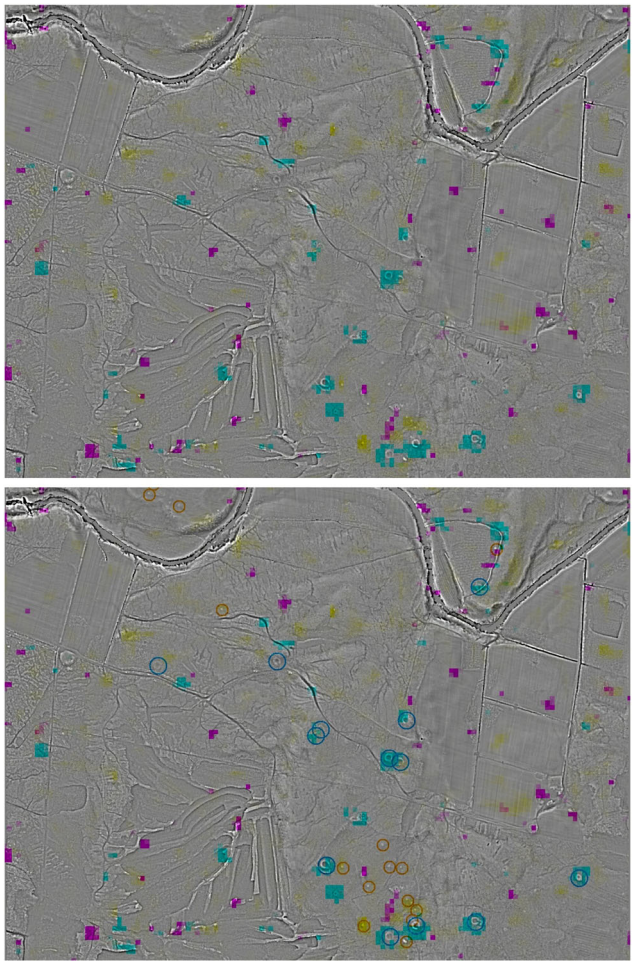

FIGURE 4 (Top) Machrie Moor with automatic detections of roundhouses (cyan), shielings (magenta) and small cairns (yellow) for the same area as Figure 3. (Bottom) Verified roundhouses (blue) and small cairns (brown) superimposed, for the same area. The digital terrain model (DTM) contains public sector information licensed under the Open Government Licence v3.0 [Colour figure can be viewed at wileyonlinelibrary.com]

to form a mask layer. The locations used for buffer zoning included roundhouses, shieling huts and small cairns, but also remains of other monuments such as burial cairns, burnt mounds, modern cattle feed stances, enclosures, and horse-engine platforms which could be confused with roundhouses, for example (see Figure 5(h, i)). A random number generator was used to select *x* and *y* coordinates of background locations from the 1 km × 1 km tiles that contained known locations of roundhouses, shieling huts and small cairns. Background locations within the buffer zone mask were discarded.

The image extracts were divided in two groups of 'training' and 'validation' (Table 3), so that the neural network learned its internal parameters from the training data, and evaluated detection performance on the validation data. This is designed to prevent overfitting on the training data, whereby the classifier recognizes the training data but may perform badly on data that it has not encountered during training.

The number of background examples (Table 3) was large for three reasons. Firstly, because 'background' is the most frequent situation in the landscape, and secondly because there are many different natural terrain structures that we did not want the neural network to classify as an archaeological object type. Thirdly, a large range of background examples may reduce the number of false positives (i.e. a location that the neural network predicts as being one of the cultural heritage types, but is in fact not). However, if the number of background examples is too large, the training phase becomes very slow, precluding experimentation with various parameter settings by running several training phases and comparing the results.

The next step of the workflow used image augmentation to address the issue that the neural network contains a large number of parameters that need to be trained, but that the initial set of learning examples is small. While the images in the validation set were kept unchanged, those in the training set were augmented by allowing the following changes to images to increase the size of the training set:

1. Horizontal flip (yes/no)
2. Rotation by 0, 90°, 180° or 270°
3. Random scaling 0.95–1.00 while keeping the aspect ratio

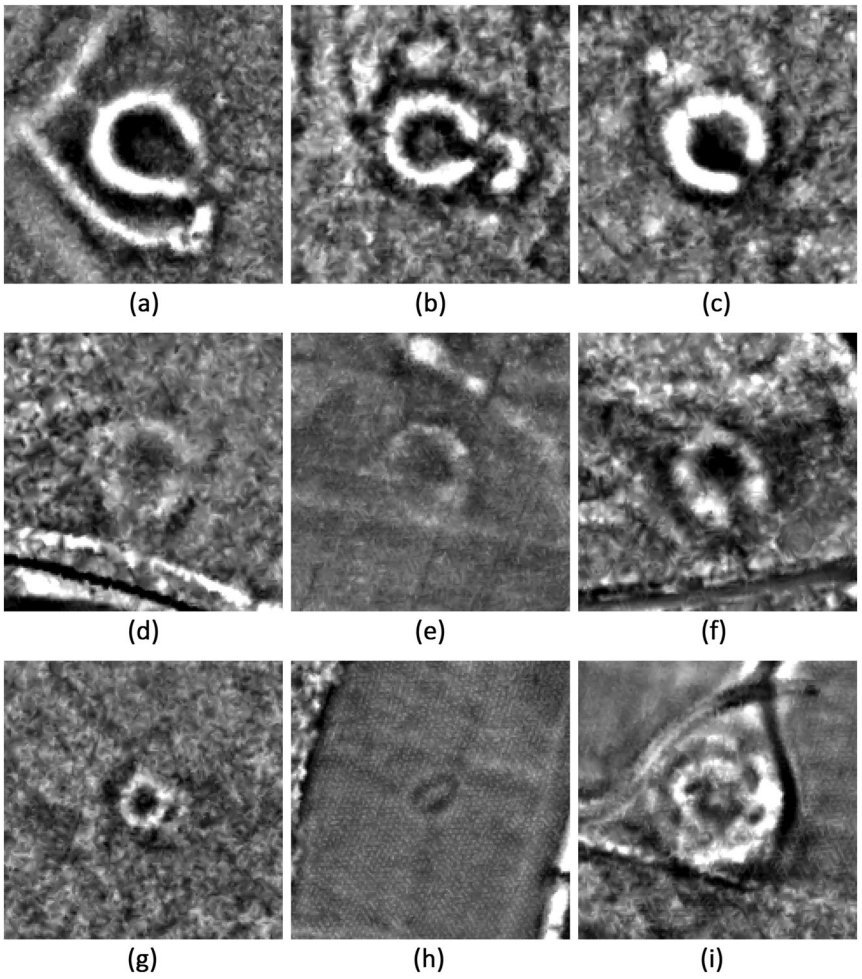

FIGURE 5 Roundhouse predictions and objects of confusion in the Machrie Moor area. Roundhouses (a)–(c) were correctly identified, despite a range of morphology and the potential influence of 'noise' from adjacent features such as field banks (a). However, three roundhouses in the learning set (d)–(f) were not identified, and this may be because of their slightly less bold expression in the visualization amongst other potential factors. One automatic identification (g) is probably too small to be a prehistoric roundhouse, but shares its basic morphology with them. The cattle feed stance is modern (h), but also shares some basic aspects of roundhouse morphology and is one of a group added to the study as a confusion learning set. A circular burial cairn (i) was identified as a shieling hut, while some others were identified as roundhouses. The locations are at OSGB coordinates: (a) 190 513 east, 632 288 north; (b) 190 778 east, 632 188 north; (c) 190 674 east, 632 180 north; (d) 190 217 east, 632 640 north; (e) 190 300 east, 633 013 north; (f) 190 428 east, 632 648 north; (g) 190 532 east, 632 237 north; (h) 190 065 east, 632 842 north; (i) 190 880 east, 632 353 north. The image portions are 40 m × 40 m. The digital terrain model (DTM) contains public sector information licensed under the Open Government Licence v3.0

TABLE 3 Training and validation sets used in training of neural networks

Class	Training	Validation	Total
Roundhouse	80	26	106
Shieling hut	177	88	265
Small cairn	259	125	384
Background	7355	1842	9197

4. Random scaling 0.95–1.00 without keeping the aspect ratio
5. Random translation by 0, 1 or 2 pixels either to the north or to the south.
6. Random translation by 0, 1 or 2 pixels either to the west or to the east

The first two of the mentioned changes, i.e. flip and rotation, produce eight times as many training images as listed in Table 3 under the 'training' column. The other four types of changes produce minor variations in the training images. The augmented training images were centre cropped to a size of 96 × 96 pixels for roundhouse detection (i.e. 24 m × 24 m as most roundhouses are in a range from 8 m to 15 m in diameter) or 64 × 64 pixels for shieling hut and small cairn detection (as these remains tend to be smaller at 2–6 m across) – noting that these images may contain other objects.

The deep CNN has many parameters that must be estimated from the training data, requiring training and validation iterations. The training was done by iteratively updating the parameters to minimize the error in the training set, and evaluating the network by classification performance on the validation data. One full iteration through the training set is called an epoch. After each epoch, the neural network internal parameters and classification performance (Table 4) are stored. At the end of the training phase, the neural network internal parameters, which were obtained from the epoch that produced the best validation accuracy, were saved to a file. This file could then later be read into the ResNet to restore the exact state of the classifier.

4.4 | Running the neural network on entire tiles

The ResNet may be run on different image sizes. When the network is applied to the full tile, the averaging operation is turned off, which makes the network produce a classification result for each 32 × 32 pixels image block. The maximum input image size was 2048 × 2048 pixels (512 m × 512 m) on the particular computer used at the NCC, though this may be smaller or larger on other computers. In practice, the 1.1 km × 1.1 km extended tiles were divided into overlapping sub-tiles, each of which was fed into the neural network classifier. The lower resolution classification results were then expanded to the original image resolution. The overlaps were removed and the

TABLE 4 Correct classification rates after each epoch of the training of the deep neural network for roundhouse detection

Epoch	Training accuracy	Validation accuracy
1	0.9847	0.9936
2	0.9876	0.9925
3	0.9890	0.9946
4	0.9856	0.9946
5	0.9895	0.9834
6	0.9872	0.9845
7	0.9894	0.9791
8	0.9883	0.9914
9	0.9903	0.9925
10	0.9907	0.9888

Note: On the validation data, the best classification rate, 99.46%, was obtained after epoch numbers 3 and 4

classification results merged into a raster file of the same size as the input extended tile. The classification results comprise one raster image for each archaeological object type, with values between 0.0 and 1.0, for each input tile. For the three archaeological object types of roundhouse, shieling huts and small cairns, the three classification results were combined into a single red–green–blue (RGB) image, colour coded with white as background, cyan as roundhouse, magenta for shieling huts, and yellow for small cairns.

By running the automatic method on the SLRM visualizations, predicted locations of roundhouses, shielings and small cairns were obtained. These outputs of the neural network are by default not normalized. The final layer is a softmax-normalization function that normalizes the output for each class between 0 and 1, and the sum over all the classes to 1. The output of the softmax function is interpreted as the probability that a given sample belongs to each of the given classes. Since this case study uses three classifiers (i.e. roundhouse *vs* background, shieling hut *vs* background, and small cairn *vs* background), another normalization would be needed to make sure they sum to one in each pixel, in order to treat them as probabilities. Such a normalization is not used at the moment.

5 | RESULTS

Two exemplar tiles from the Arran results, one with encouraging results, and another with a high level of noise, are discussed here, comparing the automated predictions with the desk and field-based observations of an experienced archaeological field worker. The example with the most encouraging results is on Machrie Moor, an area rich in known archaeological monuments. The landscape has generally low natural relief, and comprises a mosaic of improved pasture and rough grazing (grass and heather). Many of the known archaeological monuments are clearly visible in the DTM (Figure 3, top), which also shows features such as hollow trackways and peat cuttings. For this block 15 roundhouse footings and 20 small cairns were included in the Arran learning set (Figure 3, bottom). The roundhouse predictions are quite meaningful, in the sense that many true locations are identified, and that the number of false positives is not too large (Table 5). Thus, 11 of the 15 roundhouses (73%) were located by the automated method,

TABLE 5 Classification results for two areas, Machrie Moor and Glen Shurig

Monuments	Correct identifications		False identifications		Known monuments
Machrie Moor					
Roundhouse	11	73%	13	87%	15
Shieling hut	0		23		0
Small cairn	4	20%	18	90%	20
Glen Shurig					
Roundhouse	0		27		0
Shieling hut	5	26%	36	189%	19
Small cairn	0		2		0

with four false negatives where some known roundhouse locations were missed. The performance of the automated predictions for the small cairns is rather more uneven, with less than 50% matching verified locations, and many false positives (Figure 4, Table 5). Finally, the automated outputs for the shieling huts are mainly false positives, though in some cases these identify monuments of other forms, including a chambered cairn. Looking in more detail at the roundhouse results, a range of morphological forms have been correctly detected (Figure 5(a–c)). These include two examples with an adjacent field bank (e.g. Figure 5(a)) that has not precluded correct identification. For the three roundhouse locations that the automatic method missed (Figure 5(d–f)) their expression in the DTM visualizations are somewhat weaker, but not all so markedly as to convincingly explain why they were missed. The automatic detections also include an example of a round structure that is probably too small to be a roundhouse, and may be a shieling hut (Figure 5(g)).

The second example is in Glen Shurig, a steep sided valley on the east of the island, and here the automated predictions appear chaotic. Of the 19 shieling huts in this block that were included in the Arran learning set (Figure 1), only five (26%) were correctly identified, and the number of false predictions is high (Figure 6, Table 5). Moreover, there are large numbers of false positives for roundhouses and a

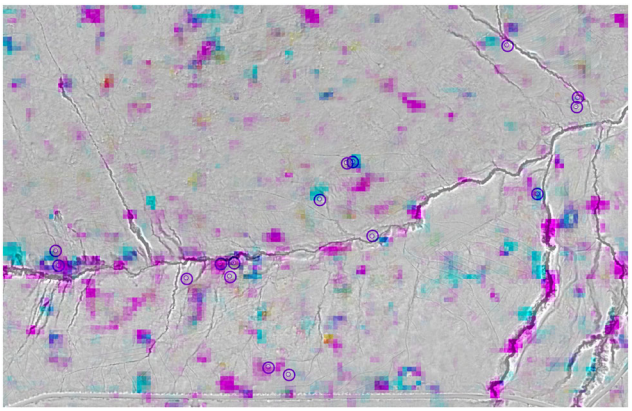

FIGURE 6 Detection results (coloured overlay) for Glen Shurig, an area with verified shielings (purple circles). The 1100 m × 705 m image is centred on 198 500 east, 636 355 north (OSGB national grid). This chaotic set of returns includes large numbers of false positives for shieling huts and for roundhouses. The digital terrain model (DTM) contains public sector information licensed under the Open Government Licence v3.0 [Colour figure can be viewed at wileyonlinelibrary.com]

scatter of false positives for shieling huts. Many of the false positives occur in areas of woodland with lower ground point densities, but are also scattered across open ground that is characterized by its unevenness, with many natural undulations and lumps of a similar scale to the archaeological remains. It is worth noting that many of the known huts in this area are relatively small, measuring less than about 3 m by 5 m across, and that the size of the training images (16 m by 16 m) may well be a factor in the chaotic outputs, as higher proportions of background are included.

These two examples provide highly contrasting results, and challenge us to understand why this might be the case. In exploring where improvements in the present version of the classifier outputs might be found, some parameters were varied to see the effect on the classification results, though the limited scope of the pilot study did not allow for a comprehensive exploration of these and their potential influence on classification accuracy. However, some useful observations can be made. Firstly, for the pixel size of the input DTM, 0.25 m pixel size seems to be appropriate for the archaeological structures in this study. A pixel size of 0.1 m appears to highlight archaeologically unimportant detail in the ALS data, while a 0.5 m pixel size makes the archaeological structures visually too small in the learning set. Secondly, although 224 × 224 pixels is the image size of the original ResNet code, this produces larger regions of each class in the detection results and therefore less precise location prediction. Smaller images also include less background terrain, but it is unclear if the character of the background terrain within each sub-image is an issue. Smaller images have the added advantage of making the training phase faster, thus allowing for more experiments. Thirdly, while the present study incorporated objects of potential confusion such as modern cattle feed stances which might easily be mistaken for roundhouses, natural terrain objects like rock outcrops and mounds of glacial origin were not included. In landscapes where anthropogenic and natural topographic features share some basic aspects in their morphology this may be an important factor.

6 | DISCUSSION AND CONCLUSIONS

The preliminary results of this automatic detection project on Arran are mixed. For the Machrie Moor area the method found many true roundhouse locations, and the number of false positives was not overwhelming. In contrast, all results for the Glen Shurig area were chaotic and overall the results for shieling huts and small cairns are less convincing. Indeed, these were included in this study because it was expected that they would be more difficult to detect than roundhouses because of their variability in form and the ease with which they might be mistaken for other features, including natural landforms. These results demonstrate the potential for further development, both for the Arran study, and to other case study areas in Scotland with similar remains. They also highlight a number of general issues.

6.1 | Artificial intelligence is being applied without proper understanding

Hutson (2018) notes an extensive critique of the widespread use of artificial intelligence without proper understanding of when and why it works well, and when and why it works less well. This legitimate concern stems from a major strength of deep neural networks – that they are easier to apply than handcrafted 'traditional' pattern recognition methods, and appear to work 'better'. And therein lies a fundamental problem, that a perception of working 'better' can preclude the exploration of how/why it works as 'only' of academic interest. However, as the study presented here demonstrates, the reasons for differing performance of deep neural networks are complex, and there is a pressing need to explore the reasons for this variability in output. These include that the structure of the deep neural network may not be fit for the pattern recognition problem at hand, so a better neural network structure must be found. In addition, the number of training examples is inadequate – a significant issue in archaeological applications. In attempting to address these issues the exploration of how transfer learning might be applied, using a large and, if possible, closely related data set for the initial training seems a useful approach. It is also worth highlighting that while it is only proper to demand better understanding of how and why applications of artificial intelligence work, this brings with it a need for the basis on which archaeologists identify and classify archaeological monuments to be made more explicit. Black boxes in processes of identification and classification are undesirable, whether they are computational or expressed through the archaeologist. The synergistic exploration of both how neural networks perform and how archaeologists identify and classify remains seems a useful common concern, escaping the oppositional expression of these processes in the past (e.g. Parcak, 2009, pp. 110–111).

6.2 | Training sets and the design of neural networks

The broader research framework that the Arran case study forms part of aspires to the creation of 'general purpose' archaeological CNNs, recognizing that this is desirable if applications in archaeology across a range of datasets and landforms are to be realized in a cost-effective way (i.e. each does not have to be engineered individually). Hence the use in this study of a ready-to-use implementation, pre-trained on a non-archaeological image database. While this may be expedient, it does require consideration of the influence of generic training sets and CNN design on the archaeological applications. Ideally, in the future, large tagged training sets of archaeological images can be used for pre-training, rather than hoping that common processes of detection learned on generic images (i.e. ImageNet) may be applicable to specifically archaeological images. The design of the neural network is dependent on the size of the cultural heritage structures subjected to mapping, and the number of training examples of each type of structure. Thus, if only a few hundred examples of each type exist (as with Arran), then a pre-trained network is likely to be the most effective solution – keeping in mind the caveats mentioned earlier. However, if there are a few thousand training examples, then the network may be trained from scratch, and may subsequently be used as a pre-trained network for types of remains with far fewer training examples. The training will also require several thousand examples of 'background' terrain without any cultural heritage remains. Multiscale approaches may also be required if the archaeological structures are of different scales, requiring DTM of various resolutions (e.g. 0.25 m

and 1.0 m), and/or neural networks with different input sizes (e.g. 96 × 96 pixels and 256 × 256 pixels).

Moreover, further consideration of the character of learning sets in specific implementations will be crucial. In the Arran case study, for example, the identification of objects of confusion, such as remains that may be mistaken for roundhouses, appears to be an important consideration in improving outputs. Adding such false positives to the learning set in confusion classes has the added benefit of taking a more holistic approach to the topography of the landscape, and creating more explicit understandings of how the archaeological elements of the landscape are defined and identified. For example, the landscape of Arran has many glacial landforms, some of which are similar in basic shape to the small cairns. In some cases, shieling huts are situated on glacially derived mounds adding further potential confusion, highlighting the importance of adding such natural landforms to a confusion class. This issue also highlights the potential influence of the size of training images on outcomes, as larger training images may contain more background noise. The need to identify such 'objects of confusion', anthropogenic or natural in origin, to CNN training sets has a direct analogue to the need for archaeological fieldworkers to be aware of natural and anthropogenic features that may appear similar to archaeological remains (Cowley, 2015). The main purpose of adding confusion classes is to assign non-archaeological structures to non-archaeological classes, thus reducing the number of false positive identifications of the archaeological structures being sought. However, there is also a risk that true archaeological structures may, in some cases, be mistaken as one of the confusion classes.

6.3 | Identification and classification

The purpose of the survey at hand is also a crucial consideration in the design of projects that may make use of automated detection, as the balance between the basic identification of archaeological remains and their correct classification may vary. For example, the automatic identifications on Figure 4 include a mound and a burial cairn (Figure 5(i)), which were predicted as being a shieling hut and a roundhouse, respectively. These examples illustrate what may happen when some classes are not present in the training data. A structure may have an appearance that is closer to one of the classes in the training data than to the general background terrain. If one wishes to address this confusion, it will be necessary to include many examples of, in this case, mounds and chambered cairns. Another alternative is to accept that rare classes may be confused with classes that are more numerous. The latter approach may be used when the purpose of automatic classification is to identify previously unknown archaeological structures, with less immediate concern that the classification to monument type is entirely accurate. The various classes of archaeological structure may be viewed as a means to design the classifier to better discriminate between archaeological structures on the one hand, and natural terrain features and modern anthropogenic structures on the other hand. Thus, depending on the purpose at hand, confusion between classes of archaeological structure may be regarded as a minor issue, when the identification of a broad range of archaeological site types is the main objective. Here, rather than being concerned with the balance of right and wrong classifications in the outputs, the extent of false positives and false negatives in identifications of objects of archaeological interest might be a key issue (see Opitz & Cowley, 2013, p. 7, for further discussion of the certainty of identifications).

6.4 | Next steps

This case study on Arran using CNN based methods of automated detection for archaeological purposes illustrates the potential (promising results of roundhouse detection in one area) and challenges (chaotic results for huts, small cairns, and roundhouses in a second area) of such an approach. The implementation of this approach within a programme of large area survey (e.g. Banaszek et al., 2018) clearly requires further development before it can be operationalized. The creation of large tagged archaeological training sets drawn from Scotland and beyond may be a prerequisite to the creation of flexible transferable archaeological CNNs, an endeavour that will require transnational cooperation. This will have the added benefit of obliging archaeologists to be more explicit in how they define the archaeological elements in the landscape, and how these are expressed in image-based and topographic datasets.

7 | CONCLUSIONS

We have demonstrated that deep CNNs have great potential for automating the archaeological mapping of Scotland. However, further improvements are needed, and we have discussed several possibilities. The most promising seem to be (1) adding false positives to the learning set as confusion classes, and (2) using a large set of labelled (archaeological and/or modern) DTM structures to pre-train the network. In these issues there is the potential to explore and refine the basis on which archaeological remains are identified, differentiating them from the natural background and modern confusion classes. There is also an important wider context to this work – that is the synergies between automated and observer-based processes, and the challenges presented to the ability of archaeologists themselves to explicitly identify and differentiate between archaeological object types.

ACKNOWLEDGEMENTS

Dave Cowley is grateful to the Royal Society of Edinburgh for an Arts & Humanities Awards Small Grant that funded a study trip to Norway and other activities in 2016. These played a key role in developing the Arran proof of concept project reported on here. He also thanks Lars Holger Pilø for a fascinating field trip in Oppland to examine automated detection in the field. The authors thank Łukasz Banaszek, George Geddes, Rachel Opitz, Rog Palmer and Robin Turner for their input to various iterations of this article, and the anonymous reviewers for their valuable comments. The authors have no conflicts of interest to declare.

ORCID

Øivind Due Trier https://orcid.org/0000-0002-3817-9777

REFERENCES

Banaszek, Ł., Cowley, D. C., & Middleton, M. (2018). Towards national archaeological mapping. Assessing source data and methodology—a case study from Scotland. *Geosciences*, *8*(272), 1–17. https://doi.org/10.3390/geosciences8080272

Bennett, R., Cowley, D., & De Laet, V. (2014). The data explosion: tackling the taboo of automatic feature recognition in the use of airborne survey data for historic environment applications. *Antiquity*, *88*(341), 896–905. https://doi.org/10.1017/S0003598X00050766

Bescoby, D. J. (2006). Detecting Roman land boundaries in aerial photographs using Radon transforms. *Journal of Archaeological Science*, *33*(5), 735–743. https://doi.org/10.1016/j.jas.2005.10.012

Bowden, M. (1999). *Unravelling the landscape: An inquisitive approach to archaeology*. Stroud: Tempus.

Cerrillo-Cuena, E. (2017). An approach to the automatic surveying of prehistoric barrows through LiDAR. *Quaternary International*, *435*(B), 135–145. https://doi.org/10.1016/j.quaint.2015.12.099

Cowley, D. (2015). Aerial photographs and aerial reconnaissance for landscape studies. In A. Chavarria Arnau, & A. Reynolds (Eds.), *Detecting and Understanding Historic Landscapes*. *PCA studies 2*. (pp. 37–66). Mantua: SAP.

Cowley, D. (2016). What do the patterns mean? Archaeological distributions and bias in survey data. In S. Campana, & M. Forte (Eds.), *Digital methods and remote sensing in archaeology - archaeology in the age of sensing* (pp. 147–170). New York: Springer. https://doi.org/10.1007/978-3-319-40658-9_7

Cowley, D. C., & López-López, A. (2017). Developing an approach to national mapping – preliminary work on Scotland in miniature. *AARGNews – The newsletter of the Aerial Archaeology Research Group*, *55*, 19–25.

Freeland, T., Heung, B., Burley, D. V., Clark, G., & Knudby, A. (2016). Automated feature extraction for prospection and analysis of monumental earthworks from aerial LiDAR in the Kingdom of Tonga. *Journal of Archaeological Science*, *69*, 64–74. https://doi.org/10.1016/j.jas.2016.04.011

Guyot, A., Hubert-Moy, L., & Lorho, T. (2018). Detecting Neolithic burial mounds from LiDAR-derived elevation data using a multi-scale approach and machine learning techniques. *Remote Sensing*, *10*(2), 225. 1–19. DOI: https://doi.org/10.3390/rs10020225

Hesse, R. (2010). Lidar-derived local relief models – a new tool for archaeological prospection. *Archaeological Prospection*, *17*(2), 67–72. https://doi.org/10.1002/arp.374

Hesse, R. (2013). The changing picture of archaeological landscapes: lidar prospection over very large areas as part of a cultural heritage strategy. In R. S. Opitz, & D. C. Cowley (Eds.), *Interpreting archaeological topography: Airborne laser scanning, 3D data and interpretation* (pp. 171–183). Oxford: Oxbow.

Hutson, M. (2018). Has artificial intelligence become alchemy? *Science*, *360*(6388), 478. https://doi.org/10.1126/science.360.6388.478

Kermit, M. A., Hamar, J. B., & Trier, Ø. D. (2018). Towards a national infrastructure for semi-automatic mapping of cultural heritage in Norway. In E. Uleberg, & M. Matsumoto (Eds.), *Oceans of data, proceedings of the 44th annual conference on computer applications and quantitative methods in archaeology, Oslo, Norway, 29 March–2 April 2016* (pp. 161–174). Oxford: Archaeopress.

Kokalj, Ž., & Hesse, R. (2017). *Airborne Laser Scanning Raster Data Visualization. A Guide to Good Practice*. Ljubljana: Založba ZRC.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Liu, X., Wang, C., Hu, Y., Zeng, Z., Bai, J., & Liao, G. (2018). Transfer learning with convolutional neural network for early gastric cancer classification on magnifying narrow-band imaging images. In *Proceedings of the 2018 IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018* (pp. 1388–1392). Piscataway: IEEE. https://doi.org/10.1109/ICIP.2018.8451067

Opitz, R. S., & Cowley, D. C. (2013). Interpreting archaeological topography: lasers, 3D data, observation, visualisation and applications. In R. S. Opitz, & D. C. Cowley (Eds.), *Interpreting archaeological topography: Airborne laser scanning, 3D data and interpretation* (pp. 1–12). Oxford: Oxbow.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(19), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Parcak, S. (2009). *Satellite remote sensing for archaeology*. London: Routledge. https://doi.org/10.4324/9780203881460

Schneider, A., Takla, M., Nicolay, A., Raab, A., & Raab, T. (2015). A template-matching approach combining morphometric variables for automated mapping of charcoal kiln sites. *Archaeological Prospection*, *22*(1), 45–62. https://doi.org/10.1002/arp.1497

Sevara, C., Pregesbauer, M., Doneus, M., Verhoeven, G., & Trinks, I. (2016). Pixel versus object – a comparison of strategies for the semi-automated mapping of archaeological features using airborne laser scanning data. *Journal of Archaeological Science: Reports*, *5*, 485–498. https://doi.org/10.1016/j.jasrep.2015.12.023

Toumazet, J.-P., Vautier, F., Roussel, E., & Dousteyssier, B. (2017). Automatic detection of complex archaeological grazing structures using airborne laser scanning data. *Journal of Archaeological Science: Reports*, *12*, 569–579. https://doi.org/10.1016/j.jasrep.2017.03.012

Traviglia, A., Cowley, D., & Lambers, K. (2016). Finding common ground: human and computer vision in archaeological prospection. *AARGnews – The newsletter of the Aerial Archaeology Research Group*, *53*, 11–24.

Trier, Ø. D., Larsen, S. Ø., & Solberg, R. (2009). Automatic detection of circular structures in high-resolution satellite images of agricultural land. *Archaeological Prospection*, *16*(1), 1–15. https://doi.org/10.1002/arp.339

Trier, Ø. D., & Pilø, L. H. (2012). Automatic detection of pit structures in airborne laser scanning data. *Archaeological Prospection*, *19*(2), 103–121. https://doi.org/10.1002/arp.1421

Trier, Ø. D., & Pilø, L. H. (2015). Archaeological mapping of large forested areas, using semi-automatic detection and visual interpretation of high-resolution lidar data. In F. Giligny, F. Djindjian, L. Costa, P. Moscati, & S. Robert (Eds.), *CAA2014. 21st Century Archaeology. Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Paris, France, 22–25 April 2014* (pp. 81–86). Oxford: Archaeopress.

Trier, Ø. D., Pilø, L. H., & Johansen, H. M. (2015). Semi-automatic mapping of cultural heritage from airborne laser scanning data. *Sémata*, *27*, 159–186.

Trier, Ø. D., Salberg, A.-B., & Pilø, L. H. (2018). Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In E. Uleberg, & M. Matsumoto (Eds.), *Oceans of data, proceedings of the 44th annual conference on computer applications and quantitative methods in archaeology, Oslo, Norway, 29 March–2 April 2016* (pp. 221–231). Oxford: Archaeopress.

Trier, Ø. D., Zortea, M., & Tonning, C. (2015). Automatic detection of mound structures in airborne laser scanning data. *Journal of Archaeological Science: Reports*, *2*(1), 69–79. https://doi.org/10.1016/j.jasrep.2015.01.005

Zingman, I., Saupe, D., Penatti, O. A. B., & Lambers, K. (2016). Detection of fragmented rectangular enclosures in very high resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4580–4593. https://doi.org/10.1109/TGRS.2016.2545919

How to cite this article: Trier ØD, Cowley DC, Waldeland AU. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeological Prospection*. 2019;26:165–175. https://doi.org/10.1002/arp.1731