Journal of
open archaeology data

Object Detection Model, Image Data and Results from the "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and Around State Game Lands in Pennsylvania" Paper

DATA PAPER

]u[ubiquity press

JEFF BLACKADAR
BENJAMIN CARTER
WESTON CONNER

*Author affiliations can be found in the back matter of this article

ABSTRACT

These data were used to build an object detection model to locate Relict Charcoal Hearths (RCH) as described in the paper "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania" [1]. This is the second grouping of data for the paper above. The first grouping is also available in this journal, see "Geospatial and image data from the "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania" paper" [2].

THESE FILES CONSIST OF:

- JPEGs representing tiles of larger Slope TIFF files derived from LiDAR for the State Game Lands (SGL) of Pennsylvania, United States [3, 4, 5, 6]. A subset of these tiles was used to train the model.
- A Shapefile of points of known relict charcoal hearths (RCH).
- XML files representing the pixel points of known RCHs on JPEG files used for training.
- Jupyter notebooks of programs used to prepare data and train a Mask R-CNN model.
- The Mask R-CNN model H5 file.
- Shapefile and GeoJSON of object detection results from the model showing locations of possible RCH in all SGLs.
- XML files representing the pixel points of predicted RCH on JPEG files used for predictions.
- GeoJSON of results using cluster analysis.

These data are stored on Zenodo.org. The programs are stored on Github.com.

CORRESPONDING AUTHOR:
Jeff Blackadar
Carleton University, CA
jeffblackadar@cmail.carleton.ca

KEYWORDS:
Relict Charcoal Hearth; RCH; Mask R-CNN; Pennsylvania; Digital Elevation Model; Iron Production; Deep Learning

TO CITE THIS ARTICLE:
Blackadar J, Carter B, Conner W. 2021 Object Detection Model, Image Data and Results from the "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and Around State Game Lands in Pennsylvania" Paper. *Journal of Open Archaeology Data*, 9: 12, pp. 1–5. DOI: *https://doi.org/10.5334/joad.81*

(1) OVERVIEW

CONTEXT

Spatial coverage

Description: State Game Lands within the State of Pennsylvania, United States.
(WGS84):
North_Bounding_Coordinate: 42.269479
South_Bounding_Coordinate: 39.719860
East_Bounding_Coordinate: –74.689583
West_Bounding_Coordinate: –80.519349

Temporal coverage

AD1700-AD1945

(2) METHODS

INTRODUCTION

These data were used to build an object detection model to locate Relict Charcoal Hearths (RCH) as described in the paper "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania" [1].

STEPS

A Python program was used to split the Slope TIFF files into smaller 1024x768 pixel JPEG tiles. If a tile contained at least one location point of a known RCH it was also retained to train the model. For each training image the program generated a corresponding XML file with the pixel coordinates of the boundaries of known RCHs in the image. Mask R-CNN is used to train the model using the images and XML files [7].

The trained model H5 file was used to detect RCHs in all of the JPEG tiles. This produces a Shapefile of predicted RCHs. The initial predictions from the model were processed using cluster analysis to produce a final list of detected RCHs, stored in GeoJSON. (***Figure 1***) lists the detailed steps.

SAMPLING STRATEGY

For training the model, 20% of the images across all SGLs were set aside for automated testing.

QUALITY CONTROL

The quality control for model training and selection had these steps:

- Each model was assessed for its Average Precision. A score above 50% meant the model was a possible candidate for object detection.
- Candidate models were used to detect RCHs in 20 images. The results were manually inspected for accuracy.
- Models that passed visual inspection were formally scored using a set of 100 images selected at random.
- Models passing formal scoring were reviewed in depth using large sample size, spanning multiple SGLs.

CONSTRAINTS

Predictions for State Game Lands 264 and 258 could not be processed due to problems in the source files.

(3) DATASET DESCRIPTION

OBJECT NAME

Since the dataset contains multiple objects, filenames are displayed in Zenodo.

DATA TYPE

Primary data: Locations of known RCHs. Secondary data: JPEG tiles, XML files, Object Recognition File, ShapeFile and GeoJSON of results.

FORMAT NAMES AND VERSIONS

TIFF, Shapefile, GeoJSON, JPEG, XML, H5, Jupyter Notebook

CREATION DATES

01/01/2020 – 31/08/2020

DATASET CREATORS

Jeff Blackadar, object recognition programmer, author, Carleton University. *https://orcid.org/0000-0002-8160-0942*

Ben Carter, archaeologist, author, data collection and verification, GIS expert, Muhlenberg College. *https://orcid.org/0000-0002-7464-0989*

Weston Conner, archaeologist, author, data collection and verification, GIS expert, Lehigh University. *https://orcid.org/0000-0001-9906-3762*

LANGUAGE

English

LICENSE

CC-BY.

REPOSITORY LOCATION

Documentation: *https://zenodo.org/record/4766351*

Data for

1. Vector files resulting from manual identification of relict charcoal hearths (RCHs). These were used to train the Mask R-CNN model.
   a. Shapefile - *https://zenodo.org/record/4593605*
   b. GeoJSON - *https://zenodo.org/record/4593622*
2. Program for splitting slope analysis (in TIFF format) into smaller tiles (in JPEG format).
   a. *https://github.com/jeffblackadar/charcoalhearths/blob/master/0_split_tifs_refactored.ipynb*
3. Data files for Training:

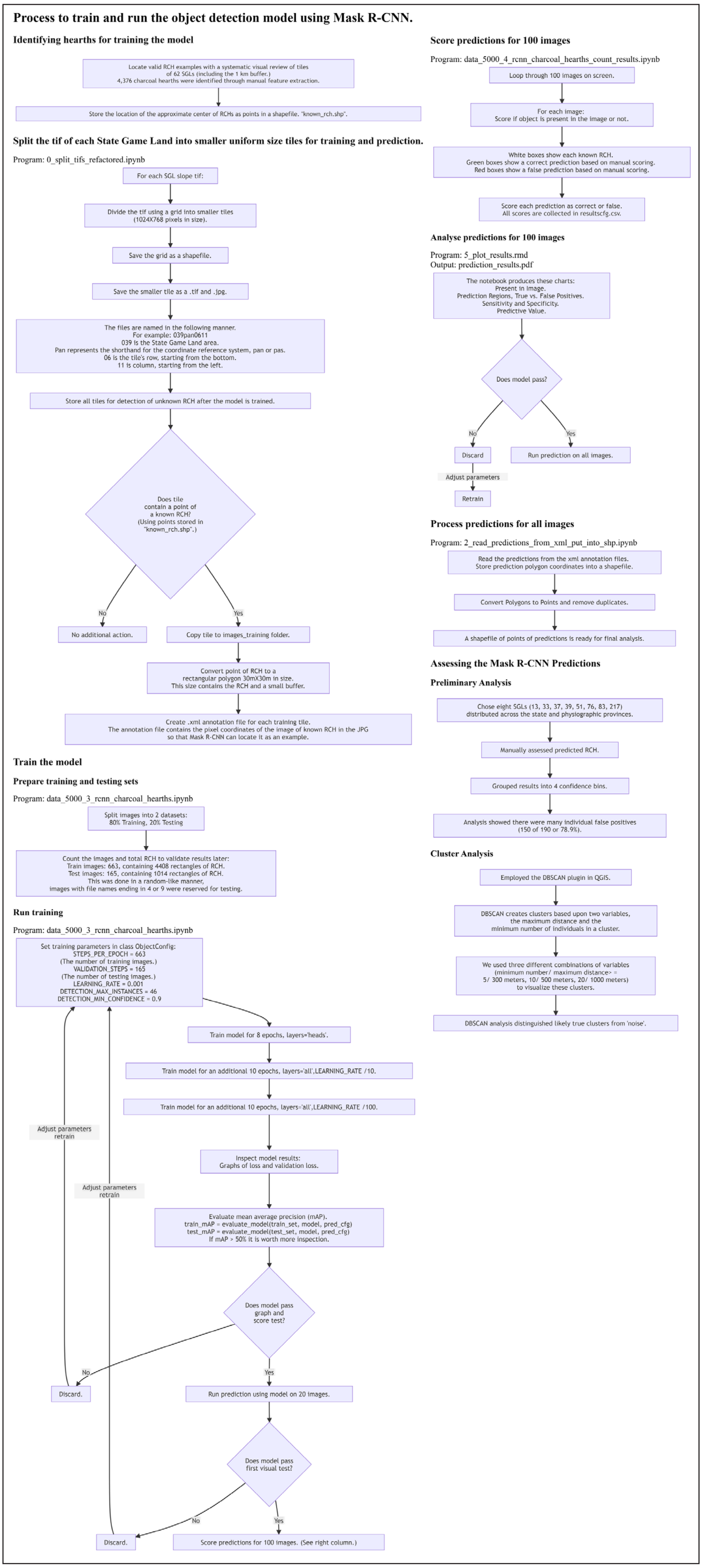
Process to train and run the object detection model using Mask R-CNN.
Identifying hearths for training the model
Locate valid RCH examples with a systematic visual review of tiles of 62 SGLs (including the 1 km buffer.) 4,376 charcoal hearths were identified through manual feature extraction.
Store the location of the approximate center of RCHs as points in a shapefile. "known_rch.shp".
Split the tif of each State Game Land into smaller uniform size tiles for training and prediction.
Program: 0_split_tifs_refactored.ipynb
For each SGL slope tif:
Divide the tif using a grid into smaller tiles (1024X768 pixels in size).
Save the grid as a shapefile.
Save the smaller tile as a .tif and .jpg.
The files are named in the following manner. For example: 039pan0611 039 is the State Game Land area. Pan represents the shorthand for the coordinate reference system, pan or pas. 06 is the tile's row, starting from the bottom. 11 is column, starting from the left.
Store all tiles for detection of unknown RCH after the model is trained.
Does tile contain a point of a known RCH? (Using points stored in "known_rch.shp".)
No
Yes
No additional action.
Copy tile to images_training folder.
Convert point of RCH to a rectangular polygon 30mX30m in size. This size contains the RCH and a small buffer.
Create .xml annotation file for each training tile. The annotation file contains the pixel coordinates of the image of known RCH in the JPG so that Mask R-CNN can locate it as an example.
Train the model
Prepare training and testing sets
Program: data_5000_3_rcnn_charcoal_hearths.ipynb
Split images into 2 datasets: 80% Training, 20% Testing
Count the images and total RCH to validate results later: Train images: 663, containing 4408 rectangles of RCH. Test images: 165, containing 1014 rectangles of RCH. This was done in a random-like manner, images with file names ending in 4 or 9 were reserved for testing.
Run training
Program: data_5000_3_rcnn_charcoal_hearths.ipynb
Set training parameters in class ObjectConfig: STEPS_PER_EPOCH = 663 (The number of training images.) VALIDATION_STEPS = 165 (The number of testing images.) LEARNING_RATE = 0.001 DETECTION_MAX_INSTANCES = 46 DETECTION_MIN_CONFIDENCE = 0.9
Train model for 8 epochs, layers='heads'.
Train model for an additional 10 epochs, layers='all',LEARNING_RATE /10.
Train model for an additional 10 epochs, layers='all',LEARNING_RATE /100.
Adjust parameters retrain
Adjust parameters retrain
Inspect model results: Graphs of loss and validation loss.
Evaluate mean average precision (mAP). train_mAP = evaluate_model(train_set, model, pred_cfg) test_mAP = evaluate_model(test_set, model, pred_cfg) If mAP > 50% it is worth more inspection.
Does model pass graph and score test?
No
Yes
Discard.
Run prediction using model on 20 images.
Does model pass first visual test?
No
Yes
Discard.
Score predictions for 100 images. (See right column.)
Score predictions for 100 images
Program: data_5000_4_rcnn_charcoal_hearths_count_results.ipynb
Loop through 100 images on screen.
For each image: Score if object is present in the image or not.
White boxes show each known RCH. Green boxes show a correct prediction based on manual scoring. Red boxes show a false prediction based on manual scoring.
Score each prediction as correct or false. All scores are collected in resultscfg.csv.
Analyse predictions for 100 images
Program: 5_plot_results.rmd
Output: prediction_results.pdf
The notebook produces these charts: Present in Image. Prediction Regions, True vs. False Positives. Sensitivity and Specificity. Predictive Value.
Does model pass?
No
Yes
Discard
Run prediction on all images.
Adjust parameters
Retrain
Process predictions for all images
Program: 2_read_predictions_from_xml_put_into_shp.ipynb
Read the predictions from the xml annotation files. Store prediction polygon coordinates into a shapefile.
Convert Polygons to Points and remove duplicates.
A shapefile of points of predictions is ready for final analysis.
Assessing the Mask R-CNN Predictions
Preliminary Analysis
Chose eight SGLs (13, 33, 37, 39, 51, 76, 83, 217) distributed across the state and physiographic provinces.
Manually assessed predicted RCH.
Grouped results into 4 confidence bins.
Analysis showed there were many individual false positives (150 of 190 or 78.9%).
Cluster Analysis
Employed the DBSCAN plugin in QGIS.
DBSCAN creates clusters based upon two variables, the maximum distance and the minimum number of individuals in a cluster.
We used three different combinations of variables (minimum number/ maximum distance> = 5/ 300 meters, 10/ 500 meters, 20/ 1000 meters) to visualize these clusters.
DBSCAN analysis distinguished likely true clusters from 'noise'.

Figure 1 Process to train and run the object detection model using Mask R-CNN.

   a. RCH Detection with Mask R-CNN Image Annotations. *https://zenodo.org/record/4575582*
      This file contains a collection of xml files that contain coordinates of the locations of known RCHs on images (from above). These files are known as annotations and are used by Mask R-CNN to identify objects to detect during training of a model.
   b. RCH Detection with Mask R-CNN Training Images. *https://zenodo.org/record/4579935*
      This file contains all of the images used for training the Mask R-CNN model. Each image contains at least one known RCH.
   C. Polygons for tiles of LiDAR data. *https://zenodo.org/record/4580726*
4. Program for Mask R-CNN training and prediction. Known as "data_5000_3_rcnn_charcoal_hearths.ipynb" [7] *https://github.com/jeffblackadar/charcoalhearths/blob/master/data_5000_3_rcnn_charcoal_hearths.ipynb*
5. Data files of the model and predictions
   a. Resultant trained Mask R-CNN model. *https://zenodo.org/record/4579946*
   b. Predictions from the model, in x, y coordinates (not geolocated) in XML format. *https://zenodo.org/record/4581281*
      The format of these files is similar to the training annotations.
   c. RCH Detection with Mask R-CNN Images. *https://zenodo.org/record/4583945*
      This file contains all of the images representing tiles of LiDAR images of State Game Lands. These images are used for predictions to locate RCHs. (A subset was used for training. See 3b RCH Detection with Mask R-CNN Training Images above.)
6. Program that produces confidence scores for the predictions above. Known as "data_5000_4_rcnn_charcoal_hearths_count_results.ipynb" *https://github.com/jeffblackadar/charcoalhearths/blob/master/data_5000_4_rcnn_charcoal_hearths_count_results.ipynb*
7. Program to remove duplicates (because some tiles included multiple SGLs and some tiles overlapped both the north and south LiDAR tile indices of the state), converts squares to their centroid for comparison and saves unique squares in geolocated vector files. Known as "2_read_predictions_from_xml_put_into_shp.ipynb" *https://github.com/jeffblackadar/charcoalhearths/blob/master/2_read_predictions_from_xml_put_into_shp.ipynb*
8. Vector file of prediction results.
   a. Shapefile- *https://zenodo.org/record/4593734*
   b. GeoJSON- *https://zenodo.org/record/4593747*
9. Prediction results with additional variables (bins for assessment, ID of training data, cluster analysis and visual confirmation)
   a. GeoJSON (no shapefile)- *https://zenodo.org/record/4593767*
   b. Variables:
      i. **id** = unique identifier starting with 3-digit SGL number, PAN or PAS (projections) and, within those a unique four-digit identifier
      ii. **score** = confidence score
      iii. **SGL** = State Game Land number
      iv. **SGLImage** = name of TIFF file of merged LiDAR tiles
      V. **Confirm** = Whether the predicted hearth was determined, through visual inspection, to be a likely true positive (Y) or a false positive (N)
      vi. **Bin#**- in assessing these predictions we "binned" the results based upon the confidence score.
      vii. **Bin_select** = 1 if this record (predicted RCH) was selected for assessment within that bin
      viii. **TrainID** = Original ID of the training data (only training data that matched with a prediction are included).
      ix. **Clusters5_300** = resultant clusters from DBSCAN where minimum cluster size = 5 and maximum distance = 300 meters
      x. **Clusters10_500** = resultant clusters from DBSCAN where minimum cluster size = 10 and maximum distance = 500 meters
      xi. **Clusters20_1000** = resultant clusters from DBSCAN where minimum cluster size = 20 and maximum distance = 1000 meters
      xii. **CLUSTERCT** = How many of the above clusters included the predicted RCH (0–3). Derived from the previous three variables.
      xiii. **3Cluster** = whether or not this predicted RCH was included in all three clusters.
10. False Negatives for the tiles around SGL 43 after close visual inspection (at 1:1000 scale).
   a. GeoJSON = *https://zenodo.org/record/4758647*

PUBLICATION DATE

11/03/2021

(4) REUSE POTENTIAL

The model may be reused to detect similar looking objects in other landscapes. The shapefile and images may be used to train an improved prediction model. Also, the images and programs may be re-used to train a different object detection model to locate other types of objects in SGLs.

FUNDING STATEMENT

We would like to acknowledge the support of the Social Sciences and Humanities Research Council of Canada. We thank the Provost's Office and the Department of Sociology and Anthropology of Muhlenberg College for financial and logistical support.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Jeff Blackadar *orcid.org/0000-0002-8160-0942*
Carleton University, CA

Benjamin Carter *orcid.org/0000-0002-7464-0989*
Data collection and verification, GIS expert, Muhlenberg College, US

Weston Conner *orcid.org/0000-0001-9906-3762*
Data collection and verification, GIS expert, Lehigh University, US

REFERENCES

1. **Carter BP, Blackadar JH, Conner WLA.** "When Computers Dream of Charcoal: Using Deep Learning, Open Tools, and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania." *Advances in Archaeological Practice*. 2021; 9(4): 257–71. Cambridge University Press. DOI: *https://doi.org/10.1017/aap.2021.17*
2. **Conner W, Carter B, Blackadar J.** "Geospatial and Image Data from the "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and Around State Game Lands in Pennsylvania" Paper." *Journal of Open Archaeology Data*. Ubiquity Press. 2021; 7. DOI: *https://doi.org/10.5334/joad.80*
3. **PAMAP Program, PA Department of Conservation and Natural Resources, Bureau of Topographic and Geologic Survey.** PAMAP Program LAS Files (LiDAR Data of Pennsylvania). n.d. [Online]. Available: *https://www.pasda.psu.edu/uci/FullMetadataDisplay.aspx?file=pamap_lidar_LAS.xml* [Accessed 30 April 2020].
4. **PAMAP Program, Bureau of Topographic and Geologic Survey, PA Department of Conservation and Natural Resources.** PAMAP Tile Index - North 2006; 2006. [Online]. Available: *https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=266* [Accessed 30 April 2020].
5. **PAMAP Program, Bureau of Topographic and Geologic Survey, PA Department of Conservation and Natural Resources.** PAMAP Tile Index – South 2006. 2006. [Online]. Available: *https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=267* [Accessed 30 April 2020].
6. **Pennsylvania Game Commission., PGC State Game Lands.** 2018. [Online]. Available: *http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=86* [Accessed 30 April 2020].
7. **Abdulla W.** Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow. n.d. [Online]. Available: *https://github.com/matterport/Mask_RCNN* [Accessed 30 April 2020].

TO CITE THIS ARTICLE:
Blackadar J, Carter B, Conner W 2021 Object Detection Model, Image Data and Results from the "When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and Around State Game Lands in Pennsylvania" Paper. *Journal of Open Archaeology Data,* 9: 12, pp. 1–5. DOI: *https://doi.org/10.5334/joad.81*

Published: 27 December 2021

Journal of Open Archaeology Data is a peer-reviewed open access journal published by Ubiquity Press.