Boise State University ScholarWorks

Research Computing Days 2021

Research Computing Days

3-31-2021

Digitizing Historical Forest Service Data

Floriana Ciaglia Boise State University

Chinwendum Njoku Boise State University

Issac Bard Boise State University

Digitizing Historical Forest Service Data

Abstract

When ecologists are working in the field, they often record their data on datasheets by hand. This hardwon information then tends to remain trapped in physical copies of datasheets which then get stored into filing cabinets, preventing further analysis. We are collaborating with the Sawtooth National Forest Service, which has collected decades of data on historical vegetation and soil conditions in the Sun Valley, Idaho area to digitize their historical data. The goal of this project is to create an Optical Character Recognition (OCR) model able to process the collected handwritten datasheets and generate a digitized version of them. By making nearly a century of environmental data ready for statistical analysis, this project will allow Forest Service and BSU scientists to answer important questions about how some of Idaho's most spectacular landscapes have been affected by climate change, sheep grazing, and natural resource management decisions across areas and timeframes that were previously impractical to tackle.

Digitizing Historical Forest Service Data

Floriana Ciaglia, Chinwendum Njoku, Isaac Bard Advisor: Dr. Catherine Olschanowsky, Dr. Kelly Hopping

1. Problem Statement

- Ecologists record vegetation data by hand onto physical paper-sheets.
- Historical Forest Data is inaccessible for further analysis and research.

2. Motivation

Vegetation and soil condition data from the Sun Valley, Idaho area has been collected by hand and is laying into dusty filing cabinets.

The goal of this project is to **digitize** the data forms to make them available for future scientific research.



3. Optical Character Recognition (OCR)

- Processes image.
- Recognizes ASCII characters in the provided image.
- Extracts the character and saves it into a machine-encoded text.

Forest Service

ForestService

4. Process

the man in)			0				1	
e	A-13	SITE ANALYSIS SUMMARY									
	Sawtooth	tooth D-5			ALLOTMENT Blue Ridge TYPE DESIGNATION KIND OF SS SS SS			lett	DATE 7/6/68		
	TRANSECT NO. PLOT SIZE PLOT INTERVAL			TYPE DES				SLOPE	ASPECT 5-W-W		
	Eastand west of Sprout crock, 6800-100										
		ТОТА	L PRODUCTIO	N - GREEN W	GREEN WEIGHT		* PROD. DRY WT.	PERCENT	CENT DESIRA- OSI- BILITY		
	SPECIES	TRANS 1	TRANS 2	TRANS 3	TOTAL	DRY WT.	PER ACRE	TION	RATING		
	35 AGSPI	- 219	75	45	1690	592	164	34	30 4		
	40 BRTE	3	State States	T	30.	12.	4	1		1	
	35 ELCI	[]	T		10	4	1		1		
in the	35. BRMA	<u>}</u>	6	11	170	. 60	20	H	4		
SES	Total grass	and the second					189	39	The second		
ASS	the second second									1.00	
GR .					The second of			and the			

Original data format

Ste 1. Identifying sub-fields in the form

- Extract sub-fields from the form using the OpenCV library.
- Each (x, y)coordinate is stored into a JSON file.



Field

Step 2. Bounding box around single characters

- Crop the image around each single character to feed to the model.

WRITEUP NO. A-13	4	1	3			
Field	Individual Cells					

Individual Cells

Step 3. Character Classification

- Feed the pre-processed images to a neural network (NN) to classify them.





5. Models

ResNet

The model used categorical cross entropy loss function, 50 epochs, the SGD optimizer.

The ResNet was trained on the MNIST and the Kaggle datasets.



EMNIST 0-9, A-Z and a-z 2BCOCFOLIKIMNOPARCTUVWXYZOO Abcdefg11JLLMNOPQrsfurwx1211 abCDEFGHLJKLMNOPGYSLUVWXY122 abcdefghisk/mnopgrstuvwx4233 ABCdefGLIJKIMNOPQFSTUVWXYZYY aBCdefghIJKLMNODENSTUVWXYZSS ABCdefGhJ KLMNOPQrStUUN14266 abodefbhask | Mnopgestuuwxyz11 2bCdeF9113K/MN019855TUVWXY288 ABCOER9BIJRIMNOPErSTUVWXYZ98

Image courtesy of: https://www.pyimagesearch.com/

6. Future Development and Challenges



Pre-process harder words



The letters "S" and "a" are connected by the same hand stroke

- Letters in hand-written text are often connected by the same hand stroke.
- Learn how to preprocess images where words have connected letters.

7. Acknowledgements

Boise State's Research Computing Department. 2017. R2: Dell HPC Intel E5v4 (High Performance Computing Cluster). Boise, ID: Boise State University. DOI: <u>10.18122/B2S41H</u>



Permutations of Convolutional Neural Networks

- Five permutations of different models: different amount dense layers, convolutional layers, neurons per layers and dropout, 10 epochs.
- The CNN was trained on the EMNIST dataset.

image courtesy of: https://www.researchgate.net/

Load data into database

- Previous work has been done to create a database where to store the collected digitized data from the Forest Service forms.
- Automate process to load the digitized data into database.