Rescuing Historical Climate Observations to Support Hydrological Research: A Case Study of Solar Radiation Data

Ogundepo Odunayo, Naveela N. Sookoo, Gautam Bathla, Anthony Cavallin
Bhaleka D. Persaud, Kathy Szigeti, Philippe Van Cappellen, Jimmy Lin

University of Waterloo
Ontario, Waterloo, Canada

ABSTRACT

The acceleration of climate change and its impact highlight the need for long-term reliable climate data at high spatiotemporal resolution to answer key science questions in cold regions hydrology. Prior to the digital age, climate records were archived on paper. For example, from the 1950s to the 1990s, solar radiation data from recording stations worldwide were published in booklets by the former Union of Soviet Socialist Republics (USSR) Hydrometeorological Service. As a result, the data are not easily accessible by most researchers. The overarching aim of this research is to develop techniques to convert paper-based climate records into a machine-readable format to support environmental research in cold regions. This study compares the performance of a proprietary optical character recognition (OCR) service with an open-source OCR tool for digitizing hydrometeorological data. We built a digitization pipeline combining different image preprocessing techniques, semantic segmentation, and an open-source OCR engine for extracting data and metadata recorded in the scanned documents. Each page contains blocks of text with station names and tables containing the climate data. The process begins with image preprocessing to reduce noise and to improve quality before the page content is segmented to detect tables and finally run through an OCR engine for text extraction. We outline the digitization process and report on initial results, including different segmentation approaches, preprocessing image algorithms, and OCR techniques to ensure accurate extraction and organization of relevant metadata from thousands of scanned climate records. We evaluated the performance of Tesseract OCR and ABBYY FineReader on text extraction. We find that although ABBY FineReader has better accuracy on the sample data, our custom extraction pipeline using Tesseract is efficient and scalable because it is flexible and allows for more customization.

CCS CONCEPTS

• **Applied computing** → **Environmental sciences**; **Optical character recognition**.

DocEng '21, August 24–27, 2021, Limerick, Ireland

ACM ISBN 978-1-4503-8596-1/21/08…$15.00
https://doi.org/10.1145/3469096.3474929

KEYWORDS

data rescue, data digitization, optical character recognition, OCR, page segmentation, table detection, climate, solar radiation

ACM Reference Format:
Ogundepo Odunayo, Naveela N. Sookoo, Gautam Bathla, Anthony Cavallin and Bhaleka D. Persaud, Kathy Szigeti, Philippe Van Cappellen, Jimmy Lin. 2021. Rescuing Historical Climate Observations to Support Hydrological Research: A Case Study of Solar Radiation Data. In *ACM Symposium on Document Engineering 2021 (DocEng '21), August 24–27, 2021, Limerick, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469096.3474929

1 INTRODUCTION

Prior to introducing electronic data storage and distribution in the workplace, hydrometeorological data were handwritten onto maps or inputted into tables, printed, and distributed. In time, developments in computing extended the capability of data input in hydrometeorological organizations, with information now stored in servers with massive storage capabilities and tailor-made software, making data analysis easy. However, there is a need to convert, i.e. rescue, historical data from print to machine-readable text; it offers a low-cost opportunity for climate risk reduction and disaster warning, especially in cold regions and developing countries. This can lead to transformative growth as climate data are needed to support almost every infrastructure development, from engineering to agriculture. Access to climate data such as solar radiation data in digital format can help answer research questions across various disciplines from water science to engineering. These include understanding energy-balance changes in the atmosphere and to what extent it controls biomass productivity in lakes, lake ice melting, and understanding the feasibility of developing solar-powered homes in a specific location. Current crowdsourcing approaches[1] used to digitize climate data help to alleviate the gap but can be time and labour intensive, not being able to keep up with the growth of climate change research. The World Meteorological Organization (WMO) alluded that machine learning technologies offer alternative solutions to make climate data rescue projects more efficient [1].

This paper reports on preliminary results to digitize data contained in *Solar Radiation and Radiation Balance Data* (SSR) booklets published by the Hydrometeorological Service of the USSR. The booklets are held by the Canadian Cryospheric Information Network/Polar Data Catalogue and contain station names and tabular climate data. To ensure the preservation of the data structure, it is essential to have segmentation in each document, making it possible to isolate the tables in each document and identify the table structure using the text detection feature of the Tesseract OCR engine.

[1] https://www.zooniverse.org/projects/edh/rainfall-rescue

We show that leveraging deep learning for automatic document layout analysis and table data extraction can unlock opportunities for digitizing such documents data at scale. We also report on the results of using different image preprocessing techniques and tools on optical character recognition (OCR) output and performance, tools we are currently testing to convert 20th-century paper-based climate records into a machine-editable open access format, such as comma-separated-values, for further climate analysis.

2 RELATED WORK

Rescuing paper-based climate data is currently done in many ways, which can be time consuming, labour intensive, and expensive [1]. There is ongoing work in the climate community which focuses on digitizing historical climate records, layout parsing, and table understanding, including table detection and table structure recognition. Sustaining these climate data rescue activities in an efficient manner by using OCR technology will expedite ongoing efforts in climate data rescue projects [2].

Following the launch of the ICDAR (International Conference on Document Analysis and Recognition) table competition [3] in 2013, there have been several deep learning approaches to help solve the problem of table detection and table structure recognition. TableNet [4], a deep learning solution that detects the table regions in an image and subsequently detects the columns in the table, uses a rule-based approach to extract the text contained in the cells of the detected table. CascadeTabNet [5] is an end-to-end deep learning approach to table detection and table structure recognition using the Cascade Mask R-CNN HRNet model. LayoutParser [6] is a document image analysis library useful for applying deep learning models to layout analysis, OCR, and parsing text in tables. They use the Tesseract OCR engine and the Google Cloud Vision API for text detection and localization.

3 DATA ACQUISITION

The data are held in 116 unique titles in booklet form, using both the English and Russian languages (in Cyrillic script). For this study we focused on fonts that are in English. The data are published in monthly issues, in tabular and map format. There are five series with a minimum run of 104 and a maximum run of 806 issues, respectively, which represent the biggest blocks of consecutive data. The data were collected by at least 298 stations across global WMO regions with observations occurring from the 1950s to 1990s. The number of reporting stations varied over the years and sometimes stations ceased to report.

In this study we focused on the SSR series composed of 332 booklets containing observations collected between January 1964 and March 1992. Each booklet contains at least 50 pages with reports from a varied number of stations. The first booklet published data from 302 stations and a spot check for 1984 revealed 326 stations. The values are reported in monthly summaries contained in three tables: (1) daily and monthly values of global solar radiation; monthly values of sunshine duration; (2) hourly, daily, and monthly values of radiation balance and monthly means of global radiation; (3) monthly means of global radiation for hourly intervals. The SSR booklet pages were scanned and stored in TIFF, totalling 927 GB and 33,742 files, each file an image representing one page of a booklet. Each image has a resolution of 600 DPI and an average size of 24 MB. Using November 1984 as a representative example, one booklet has a total of 96 pages with 135 data tables; station names, in table format, comprise 13 pages.

4 EXPERIMENTAL APPROACH

Our task is to extract the observed data from images captured in tabular format in scanned booklets using Tesseract OCR and ABBYY FineReader. This section provides details on our approach. We experimented with different techniques and document the challenges and proposed solutions in detail.

4.1 Tesseract OCR

We adapted Tesseract OCR [7] as one of the OCR engines for extracting information from documents. Tesseract is an open-source text recognition engine that supports the extraction of text from documents in a variety of languages.

4.1.1 Image Preprocessing. The better the quality of an image, the higher the accuracy of the OCR output. Image quality in this context refers to low pixel noise, properly defined character edges, and well-aligned characters. We use the OpenCV library [8] to preprocess the scanned images before text extraction. Below we describe in detail the approach taken using a tested combination of preprocessing techniques.

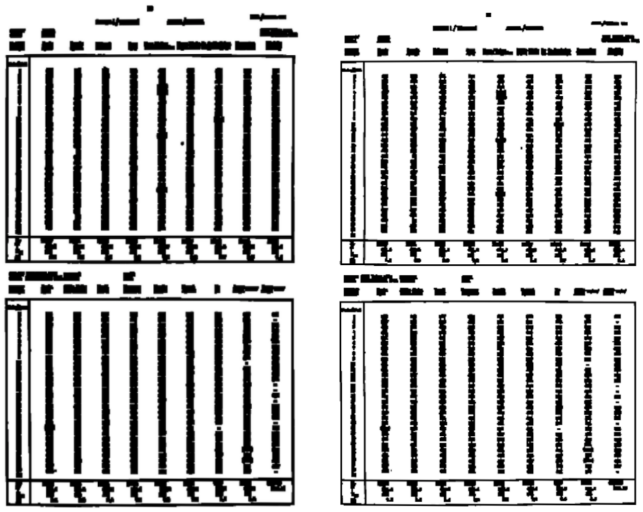
- Grayscaling, which is a simple and basic image processing technique that involves converting an image from any color space to grayscale. Grayscaling is the first processing operation performed on the image using the "`cvtColor`" function provided by the OpenCV library.
- Skew Correction. Some of the images were slightly skewed during scanning. OCR accuracy reduces significantly when the image is too skewed. Skew correction was performed to detect the skew angle and correct the skewness to ensure that the text lines are horizontal.

4.1.2 Table Detection/Page Segmentation. We tested a couple of techniques to segment the images and extracted the regions with tables and other necessary metadata such as the station names.

- Image Processing Segmentation. Here, the preprocessed image is run through a series of morphological operations such as erosion, dilation before thresholding to generate masks. Edge detection operation is then performed on the image using contours. It is assumed that the largest contours based on area represent the tables because they make up a significant portion of the page. The area of the contour depends on the size of the table.

  We perform the `Opening` operation to remove the noise in the image. In the dilation transform, we transform the original image to thicken the black pixel regions. We convert the original images into binary images before applying the dilation transform. Figure 1a shows the dilation transform while Figure 1b is the transformed dilated image. A kernel size of 15×15 for two iterations was used for the erosion transform, while we used the same kernel size for dilation for one iteration.

  The `Thresholding` transform was used to create masks using a pixel threshold of 150. After threshold comes the flood-fill



(a) erosion (b) dilation

Figure 1: Erosion and Dilation Operations — results of the erosion and dilation operations showing the thickening of the black pixel regions.

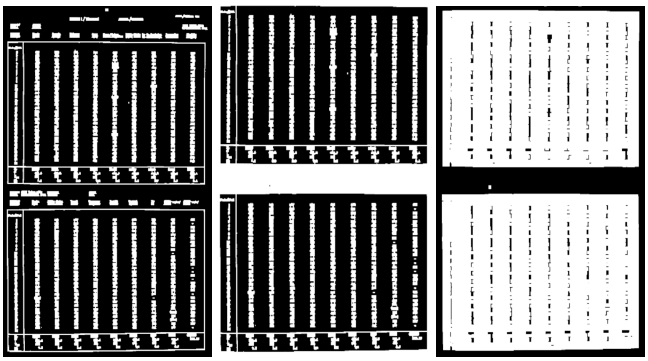


Figure 2: Threshold and Flood-fill Operations — result of the threshold and flood-fill operation showing the segmented table region

operation, this helps to fill up the crevices and further remove noise in the image. After performing the described operations, we apply a bitwise operation and use contour areas to extract the tables from the image, which are the regions of interest.

- Semantic Segmentation. For our dataset, the necessary table metadata (such as station names where the data were recorded) were captured outside the table boundaries. It was easier to extract tables and station names with a pixel-level instance segmentation model. We start with a pretrained CascadeTabNet model [5], fine-tuned it on our dataset, and used it to predict and draw bounding boxes around the table and station name regions.

  The model combines Cascade R-CNN [9] with HR-Net backbone network [10]. The training set consisted of 150 scanned images from the SSR containing 280 tables, each table with a corresponding station name block. The training set comprised 100 images, while the validation and test sets comprised 25 images each. The data was annotated using the COCO [11] JSON format. We used the MM-detection implementation format to train the model using the default configurations. Experiments show that these models yield good results for table detection.

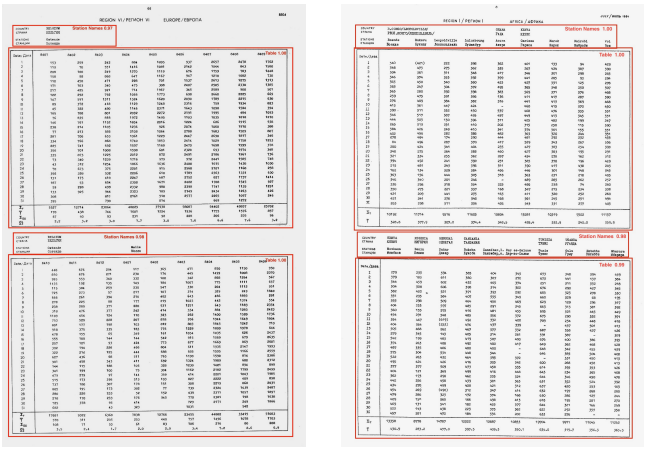


Figure 3: Model Output — running inference on an image showing bounding boxes around the table and station names.

4.1.3 Text Extraction. Tesseract comes with different configuration settings that can be adjusted depending on the use case. Recent versions come with pretrained neural networks (LSTM) in multiple languages to improve OCR accuracy. Tesseract works by analyzing and storing the outlines into blobs. Blobs are organized into text lines, analyzed for fixed pitch or proportional text and broken into words. The recognition step is a two-pass process. During the first pass, Tesseract tries to recognize all the words. The satisfactory words are passed into an adaptive classifier, which helps it recognize text down the page more accurately. The second pass attempts to recognize the words that were not appropriately recognized in the first pass. The final phase is used to resolve fuzzy spaces and locate small-cap text. To improve Tesseract OCR accuracy, we trained a custom model for extraction using the default Tesseract LSTM model for English language ("`eng.traineddata`") as the starting point. Training data consists of line images and ground-truth transcription. A shell script generates the line images alongside initial transcriptions, which are manually corrected before training. Running inference using the retrained model shows a significant improvement in the accuracy of the OCR output.

4.2 ABBYY FineReader

ABBYY FineReader OCR Editor is a popular and easy-to-use but proprietary software that produces digitized data from scanned images. It was designed and developed by an international company, namely "ABBYY" to provide OCR services.[2] To extract text from an image using ABBYY FineReader, we uploaded the image onto the ABBYY console. After uploading, the software allows one to draw bounding boxes around certain regions in the image to

[2] https://www.abbyy.com/

	% Accuracy	
	Station Names	Tables
Tesseract	84.6	91.71
ABBY Fine Reader	85.1	95.20

Table 1: Comparing OCR output accuracy for Tesseract and ABBYY FineReader

extract the text in that region of the image. This flexibility is helpful in instances where only a segment of an image needs to be digitized. This was particularly useful for the SSR series, where station names were outside the boundaries of the tables. ABBYY FineReader has automatic table structure recognition built into the software. It automatically detects columns and rows in the input image, maintaining this structure in the output text. If the table was not digitized correctly, we noticed that some rows and columns went missing and sometimes overlapped. This was one of the most common issues we encountered during extraction. When this happens, the table can be easily and quickly modified to ensure that the formatting is correct when OCR is performed. ABBYY FineReader had some difficulty analyzing large data tables in the SSR series.

ABBYY FineReader offers built-in patterns to recognize characters on an image without doing any further work. However, their built-in patterns appear to be very limited, generalized, and error-prone. The most significant error that we noticed using the built-in patterns was "." misidentified as "-" which may be due to the age and print quality of the booklets. To avoid this type of problem when extracting using the built-in patterns, we trained new patterns to recognize characters individually. For the Solar Radiation and Radiation Balance Data series, the document type selected on the ABBYY console was "Typewriter". The new pattern must first be created and then saved. It can then be selected, and the pattern can be trained to recognize characters. The process of training a new pattern involves manual observation of each individual character on an image, typing the correct character that matches the observed character, and continuing to the next. After training new patterns to improve OCR accuracy, we then run OCR on the scanned images and export the digitized data to a CSV file.

5 EVALUATION RESULTS

Here we present preliminary results on the performance and optimization of two different OCR engines using solar radiation data as a case study. We evaluated the performance of both Tesseract and ABBYY FineReader by measuring accuracy based on the number of cells and station names that were correctly transcribed. We ran OCR on two randomly selected booklets from different years, containing a total of 90 images. Each image contained two tables and an equal number of station name blocks containing English text.

The performance in Table 1 shows that ABBYY FineReader has slightly better accuracy than Tesseract on the font type present in the document. Both OCR tools offer promising results with high accuracy in extracting climate data within acceptable margins of error. However, with more training the accuracy of both OCR engines can be improved. Tesseract OCR offers added advantages as it is open-source and can be easily modify/trained to adapt to hydrometeorological data in different formats.

Improving the accuracy of ABBYY FineReader requires training individual characters alongside text extraction, and this involves uploading the files into the ABBYY console. Uploading a single booklet in one batch, approximately 150 to 200 scanned images, took, on average, two hours. Issues of font face variation and quality of text print impacted the time required to train the characters

6 CONCLUSION AND NEXT STEPS

Currently there are countless hydrometeorological observations in print, making the data difficult to use. Digitizing the data would aid researchers, but there are problems that include: image quality before and after scanning; font type; quality of the print; and, how the data are captured (bordered or borderless tables).

This study applied semantic segmentation, text recognition extraction techniques to convert solar radiation data in scanned images into machine readable texts using ABBYY FineReader and Tesseract. Future work will extend this data project to extract hydrometeorological data points on print maps, handwritten data and Cyrillic characters into machine readable formats. We believe that these OCR techniques could also be applied to reduce manual effort and costs to digitize data. In addition, this dataset would be timely for the academic community who require climate data to support their research. For instance, the Global Water Futures program, one of the largest university-led water research initiatives in Canada, would use these rescued data to support flood forecasting models development to help adapt and mitigate risks in vulnerable communities across cold regions during extreme climate events.

7 ACKNOWLEDGMENTS

This work was partially funded by the Canada First Research Excellence Fund's Global Water Futures Programme. We would like to thank the Canadian Cryospheric Information Network/Polar Data Catalogue for making booklets available to support this research.

REFERENCES

[1] World Meterological Organization. Guidelines on best practices for climate data rescue. pages 1–39, 2016.
[2] S. Brönnimann, Yuri Brugnara, Rob Allan, Manola Brunet, Gilbert Compo, Richard Crouthamel, P. Jones, Sylvie Jourdain, Jürg Luterbacher, Peter Siegmund, Maria Valente, and Clive Wilkinson. A roadmap to climate data rescue services. *Geoscience Data Journal*, 5:28–39, 06 2018.
[3] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. ICDAR 2013 table competition. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pages 1449–1453, 2013.
[4] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, and Lovekesh Vig. TableNet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. *CoRR*, abs/2001.01469, 2020.
[5] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. *CoRR*, abs/2004.12629, 2020.
[6] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. LayoutParser: A unified toolkit for deep learning based document image analysis. *CoRR*, abs/2103.15348, 2021.
[7] R. Smith. An overview of the Tesseract OCR engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 629–633, 2007.
[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
[9] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. *CoRR*, abs/1712.00726, 2017.
[10] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019.
[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.