Master Thesis Proposal: Whitewashde text detection in PDFs with deep learning

Student: Ammar Alhashimi (13836676)
Email: ammar.alhashmi@student.uva.nl
Supervisor UvA: Maarten Marx (M.J.Marx@uva.nl)

ABSTRACT

This research will investigate the use of deep learning and machine learning models in order to detect whitewashed text in large PDF documents. To be more clear, an attempt will be made to detect the regions in PDF documents that hide/whitewash text or information. Several approaches will be used to address this problem because different ways were used to whitewash text in PDF documents.

1 INTRODUCTION

In literature many approaches have been used to detect or extract features and information from PDFs such as region detection, image segmentation, colors detection and shape detection. For example, [1] presents a representative local region detector based on color-contrastpixel ranking and [2] proposes extracting more information and features from PDFs by converting PDFs to HTML form.

Furthermore, different research papers proposed different ways to extract information from PDFs after turning them into images. Such as using a Convolutional Neural Network model to extract image features. What most of the research papers have in common is that they used the extracted image features such as shape or color to detect certain objects or to classify types of documents. What was insufficient from the literature is conducting features extraction and / or segmentation on textual documents to detect useful information such as whitewash detection.

In this paper, an attempt was made to successfully detect whitewashed text in pdfs documents by applying deeplearning and machine learningtechniques. In other words, different approaches will be applied to detect the deleted or hidden text (whitewashed text) in PDF documents and the best approach will be chosen. The models performance will be checked by calculating the accuracy and precision. The running time will be considered since the models must go through large PDF documents.

1.1 RESEARCH QUESTION

What machine learning or deep learning approaches can be used to identify whitewashed text (hidden information) in large pdfs and which approach has the best performance?

1.2 SUB QUESTIONS

1. What is the detection performance of the models?

- The overall detection rate
- Detection rate per large pdf / per pdf type

2. What features or factors can be used to improve the detection rate?
    - Combination of existing approaches
3. What is a suitable threshold for errors to determine if a detected redaction is correct or not.

2 EXPLORATORY DATA ANALYSIS

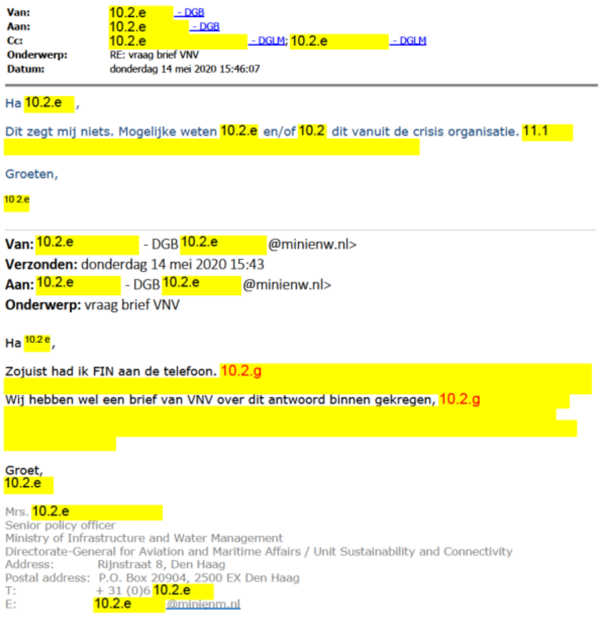
During this research different datasets will be discussed. First dataset consists of 365 PDF files which contain in total 32 thousand textual pages. The documents can be downloaded from the dutch government website webcovid19 (https://wobcovid19.rijksoverheid.nl/). This dataset will be further analysed and split into training and validation dataset.

Furthermore, since different approaches with different models will be used to detect or extract whitewashed text in PDF documents, models that are pre-trained will be used to make sure the chosen models have a relativly sufficient accuracy when extracting features. For the pre-trained models, datasets such the MS-COCO dataset will be used because it contains 328,000 images with of 80 object classes. These classes contains objects of different shapes and colors. Pre-trained models could be useful when detecting whitewashed text because then the models will have more range to identify irregular shapes or colors in PDF documents

3 RELATED LITERATURE

Whitewashing documents refers to hiding sensitive information by covering the regions of the documents that include sensitive information. Hiding the sensitive information could be done in different ways, such as covering the text with colored shapes, painting over the text or simply covering the text with square or rectangle shaped figures. E.g. figure 1 shows an example of a document with whitewashed content.

1.



Van: 10.2.e - DGB
Aan: 10.2.e - DGB
Cc: 10.2.e - DGLM; 10.2.e - DGLM
Onderwerp: RE: vraag brief VNV
Datum: donderdag 14 mei 2020 15:46:07

Ha 10.2.e ,

Dit zegt mij niets. Mogelijke weten 10.2.e en/of 10.2 dit vanuit de crisis organisatie. 11.1

Groeten,

10.2.e

Van: 10.2.e - DGB 10.2.e @minienw.nl>
Verzonden: donderdag 14 mei 2020 15:43
Aan: 10.2.e - DGB 10.2.e @minienw.nl>
Onderwerp: vraag brief VNV

Ha 10.2.e,

Zojuist had ik FIN aan de telefoon. 10.2.g

Wij hebben wel een brief van VNV over dit antwoord binnen gekregen, 10.2.g

Groet,
10.2.e

Mrs. 10.2.e
Senior policy officer
Ministry of Infrastructure and Water Management
Directorate-General for Aviation and Maritime Affairs / Unit Sustainability and Connectivity
Address: Rijnstraat 8, Den Haag
Postal address: P.O. Box 20904, 2500 EX Den Haag
T: + 31 (0)6 10.2.e
E: 10.2.e @minienm.nl

Figure 1

In order to address this problem, several feature detection and extraction techniques should be applied. Several research papers researched color-based methods to extract features.[1][4]. To detect colors [1] proposes a Color-Contrast-MSER (CCM) detector with color-contrast pixel ranking that can reduce the number of nonsensical background patches removed. This paper proposes a color contrast analysis to achieve a homogeneous contrast region detection that detects high contrast parts of a document. This approach could help detect whitewashed pages in PDF documents, especially the pages that have been whitewashed with color or colored shapes as shown in Figure 1.

Since whitewashing documents could also be done by placing objects in front of the text that is intended to be covered, an interesting approach [4] seemed suitable, which is called salient region detection. It is achieved by maximizing a submodular objective function.

4 METHODOLOGY

The main focus of this research will be on model validation. The following steps will be taken in this research:

1. Prepare different approaches by building different models.
2. Pre-process the large PDF data to split it into a test and validation dataset
3. Calculate the performance per approach
4. Look into possible factors in the models to improve the performance
5. Perform feature engineering to modify the data in order to improve the results
6. Compare and evaluate finings.

5 RISK ASSESSMENT

Several risks might be faced when conducting this research. The first risk is the lack of needed data. In order to train models, labeled data will be needed and there are no labeled whitewash documents. Thus, it is possible that the creation of labeled data should be done during this research. The second risk is that all the work that was done in the related literature was not done on PDf documents. Thus, using their approach might deliver some unexpected results. Another riks comes to mind the computational time since the models will go through thousands of PDF documents.

6 PROJECT PLAN

The table below shows an overview of the week by week planning of the thesis project.

Week	**Achievement**
Week 1	Researching and rebuilding different models and preparing and preprocessing the needed data
Week 2	Researching and rebuilding different models and preparing and preprocessing the needed data
Week 3	Researching and rebuilding different models and preparing and preprocessing the needed data
Week 4	Preparing the validation process
Week 5	Performing the validation process
Week 6	Working on the thesis
Week 7	Performing feature engineering to improve the data
Week 8	Improving the models
Week 9	Reevaluate the results
Week 10	Completing the thesis
Week 11	Completing the thesis
Week 12	Finished thesis & thesis defense

REFERENCES

[1] Yang Cao1,2, Ke Gao1 , Sheng Tang1 , Yongdong Zhang1. (2014). A Representative Local Region Detector Based On Color-Contrast-MSER

[2] Xiaohu YANG, Deliang JIANG. (2009). Converting PDF to HTML approach based on Text Detection

[3] Zhuolin Jiang, Larry S. Davis. (2013). Submodular Salient Region Detection

[4] Jinping Li, Xiangru Yu, Kusuma Wardhani Mega. (2018). Comparative Analysis of Color Edge Detection for Image Segmentation