ShabbyPages: A Recipe for Repeatable, Synthetic, Modern Document Images

Alexander Groleau[1,2], Kok Wei Chee[1], Jonathan Boarman[1], and Samay Maini[1]

[1] Sparkfish LLC (augraphy@sparkfish.com)
[2] Left Associates LLC (research@left.associates)

Abstract. The ShabbyPages document image dataset is produced using the Augraphy document image augmentation tool. The development of the computation pipeline used to generate the corpus is discussed, and the results presented. The final corpus contains over 6000 "born-digital" ground truth document images, sourced on the Web, with synthetically-noised counterparts ("shabby pages") that appear to have been printed and faxed, photocopied, or otherwise altered through physical processes. The release of this dataset and the recipe for its production attempts to address a growing need for labeled training document images for supervised learning tasks. The results of several experiments are discussed, in which the corpus trains performant convolutional denoisers, which remove real noise features with a high degree of human-perceptible fidelity, establishing baseline performance for a new ShabbyPages benchmark.

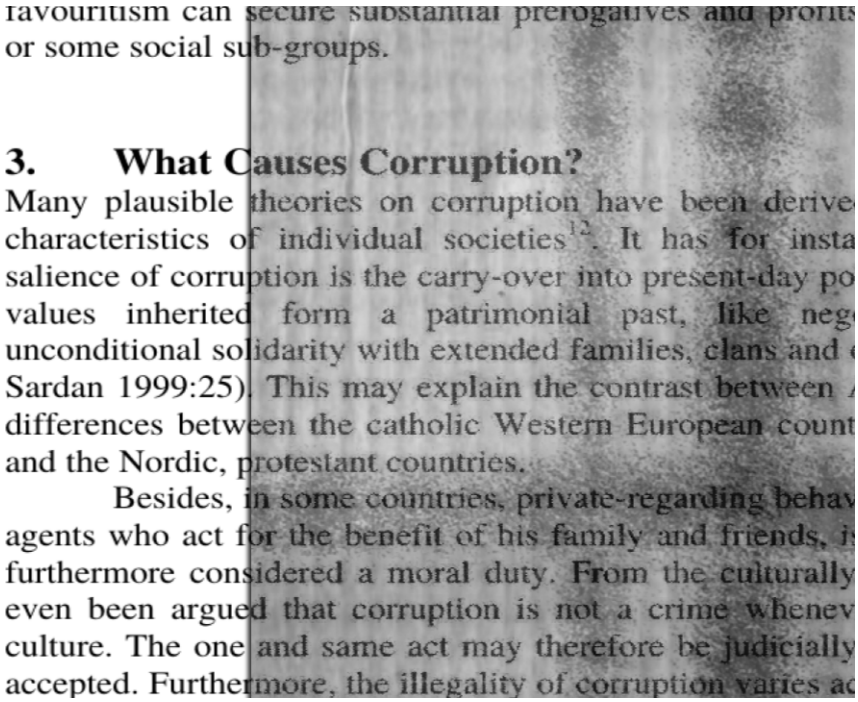
1 Introduction

Supervised learning requires a collection of training data and accompanying "labels", which in the graphical modality are clean images, ree of noise or other degradations. ShabbyPages is a large dataset composed of 6153 clean/noisy pairs of images taken from 600 digital documents. Information about the set's contents and its relation to other datasets are explored in the next section. The noisy images were created from their clean origins by the application of a pipeline developed with the Augraphy library; the production of ShabbyPages is detailed in the third section, with all source code and input materials made openly available. The final dataset is used to train several denoising models which are compared directly and whose prediction results are compared together in the final section.

2 The Dataset

ShabbyPages joins a rapidly growing number of open-source datasets freely available online. Human activity today generates a tremendous amount of data, with many people choosing to publish theirs on websites like Kaggle [1], Hugging Face Hub [2], and many more. All of this data is arranged into categories: sets exist which contain text, images, audio, and data from other modalities, with each set often collected and designed for a purpose. Of the many image datasets

available, the authors understand few to have been designed for groundtruthed supervised learning of document-specific tasks like binarization or denoising. In fact, we could find only one other set compiled for doing so with images of *modern* documents: NoisyOffice [6].


favouritism can secure substantial prerogatives and profits
or some social sub-groups.

3. What Causes Corruption?

Many plausible theories on corruption have been derive
characteristics of individual societies[12]. It has for insta
salience of corruption is the carry-over into present-day po
values inherited form a patrimonial past, like neg
unconditional solidarity with extended families, clans and
Sardan 1999:25). This may explain the contrast between
differences between the catholic Western European count
and the Nordic, protestant countries.

Besides, in some countries, private-regarding behav
agents who act for the benefit of his family and friends, i
furthermore considered a moral duty. From the culturally
even been argued that corruption is not a crime whenev
culture. The one and same act may therefore be judicially
accepted. Furthermore, the illegality of corruption varies ac

Fig. 1. Example ShabbyPages image, with the groundtruth partially augmented on the right.

2.1 Features

Modern artificial neural networks produce a latent representation of their training data, a space from which outputs are sampled. In deep networks, these representations live within multiple layers, each of which corresponds to a feature of the input data. Document images are a particularly interesting category, as the data within the documents each picture represents is itself frequently multi-modal; the document may contain an image and its description, a table with statistics, multiple regions of structured and formatted text, and even other documents.

Table 1 below contains a comparison of some of the key hyperproperties of the ShabbyPages set compared with other popular document image datasets.

Table 1. Document Image Dataset Feature Comparison

Feature	ShabbyPages	NoisyOffice	RVL-CDIP	Tobacco800
Images	6153	72	400000	1290
Document-specific	Yes	Yes	Yes	Yes
Synthetic	Yes	Yes	No	No
Groundtruthed?	Yes	Yes	No	No
Font size	Multiple	16pt	Multiple	Multiple
Paper styles	Multiple	1	Multiple	Multiple
Multilingual	Yes	No	No	No
Documents contain images?	Yes	No	Yes	Yes
Pre-categorized	No	No	Yes	Yes

NoisyOffice The NoisyOffice database [6] contains a very low 4 base images, 2 types of font, 3 font sizes, and 4 types of noise, for a total of 72 unique images. The groundtruth images were produced in a standard word processor, with the simulated noisy images produced by first adding noise to a blank sheet of paper, storing that as a background image after digital scanning, then overlaying the foreground text onto the noisy background texture. An additional set of real NoisyOffice images was produced, where the foreground text was printed on blank white sheets, the results of which were physically augmented by the introduction of footprints, coffee stains, wrinkles, and folds.

ShabbyPages began with 600 documents sourced from the Internet from culturally-important sources. A manifest is included in the dataset which contains an ordered table of the filenames, the languages present in the document, the download URL for the file, and an English name or acronym for the source organization or person. These documents were split into their constituent pages, which were then passed through an Augraphy pipeline we tuned over a month or two. This tuning amounted to tweaking numerical parameters for each of the 24 different augmentations from the Augraphy library. All or none of these may have applied to each input image, and only one pipeline run was completed per image. Each augmentation has a high degree of internal variability due to several calls to a pseudorandom number generator on which several transforms depend. The conjunction of all these factors implies ShabbyPages is a sample from a massive space. Indeed, the cosine similarity computed over OpenAI's CLIP representations of each output image averages to 0.49, much higher than NoisyOffice's 0.31. Augraphy does not yet have augmentations for footprints or coffee stains, so the dark regions and edges of the stains and prints were simulated with features introduced by other augmentations, but Augraphy does contain the PaperFactory transformation which selects and crops a texture from a local directory of paper images. The result of a PaperFactory application is the foreground text "printed" onto the background texture, similar to NoisyOffice's construction.

2.2 RVL-CDIP

The RVL-CDIP dataset [7] consists of 400,000 grayscale images divided into 16 document classes, each with a similar number of elements. The dataset is pre-segmented into 320,000 training, 40,000 test, and 40,000 validation images, facilitating immediate use in machine learning applications. The classes represented are the following:

letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo.

The images exhibit poor fidelity with what we can imagine their originals to have been, contain substantial noise, and many were scanned at around 100 pixels per inch, quite a low resolution for documents. None of the images in the CDIP set come with groundtruths.

The collection of RVL-CDIP is an impressive feat in its own right: nearly half a million scanned documents with real noise, already categorized. Augraphy makes synthetic documents, intended to simulate real noise without the need to physically produce it. One of the augmentations used in the ShabbyPages set is BadPhotoCopy, which mimics low-fidelity noisy or damaged scans.

2.3 Tobacco800

Tobacco800, like RVL-CDIP, is a subset of the Complex Document Image Processing collection, containing images of documents originally internal to tobacco companies, and released under the Tobacco Master Settlement Agreement. The corpus is composed of 1290 images, gathered and scanned at different times and by different devices. Several of the images in this set are consecutively-numbered pages, increasing the set's utility for certain image analysis tasks [4]. Document images in this set vary from from 150 to 300 DPI with dimensions between 1200 pixels wide and 2500 pixels tall on the low end and 1600 by 3200 on the high.

2.4 Statistics

ShabbyPages contains documents from multiple classes, which contain many types of information.

Table 2 below contains some statistics for the dataset, in grayscale, at a standard resolution of 150 ppi.

2.5 Limitations

The ShabbyPages collection is a comprehensive entrant to the public datasphere, but with a relatively low starting document volume, makes some important omissions.

Table 2. ShabbyPages Statistics

Statistic	Value
Images	6153
DPI	150
Max image size	-

First, we arbitrarily selected 600 as the number of source documents. The number of final images of pages depends on this, and so is similarly arbitrary.

There are only a few languages represented, which reflects a number of biases, including the following:

- The native languages of the document-searchers
- The locations from which the document-searchers searched
- The search techniques used by each document-searcher

This work does not account for them, but these biases are significant determiners on the set of input documents. We hoped to overcome this somewhat by picking a suitably large number of input documents, and settled at 600.

3 Motivation for Denoising

TODO

4 Construction

This section describes the dataset generation methodology; code for all of this is available on GitHub.

4.1 Data Gathering

A team of workers searched the public internet for PDFs of many different kinds. 600 unique documents were retrieved, totaling 6153 pages. A similar process was completed to collect paper textures on which to "print" the documents: 300 unique textures were gathered, all of which are either in the public domain or carry CC-0, CC-BY, or CC-BY-SA licenses. Care was taken to retrieve freely-available and attributable documents; a CSV file containing document and paper texture links (as of time of download), with their sources, is available and distributed with the dataset. We also manually reviewed the gathered documents to verify that no personal identity information was present within the corpus.

4.2 PDF to PNG

The `pdftoppm` tool from the `poppler-utils` package was used to split the PDF documents into individual pages. Each page was converted to a PNG image at

150dpi, a common printing resolution. Because the majority of these documents were created with the standard US Letter dimensions, this resulted in the most common image dimension being 1275 pixels wide by 1650 pixels tall.

```
pdftoppm document_name.pdf document_name -png -r 150
```

4.3 Developing the Pipeline

The Augraphy library presents an easy-to-use API for constructing feature pipelines, which has been designed for interoperability with other augmentation tools, and within the broader data ecosystem. While Augraphy's default pipeline has what we believe to be quite realistic defaults, we wanted a broader range of features from this dataset than those the default pipeline could produce. To address this, we broke all of the parameters for every augmentation constructor out into separate variables, tweaking these and committing the new pipeline to GitHub. We created an automated daily build in GitHub Actions to render a random selection of ground-truth images with the updated pipeline, then our team met frequently to discuss the output and make adjustments to the pipeline.

4.4 Processing Augraphy on a multicore system

Execution time is dependent on which augmentations are executed at runtime; an Augraphy pipeline can take several seconds to process large images. The library is under active development with performance enhancements underway, but the time cost to generate large datasets sequentially is prohibitive when dealing with thousands of files, so we use a multi-process pooling technique to distribute the workload across many processor execution threads. For each process, we generate a new pipeline, run the pipeline object on an image, and save the output, using the code below:

```
import os
import cv2
from multiprocessing import Pool
from pathlib import Path

input_path = Path("/path/to/input/images")
filenames = [(input_path / name) for name in os.listdir(input_path)]

pool = Pool(os.cpu_count())

def run_pipeline(filename):
    image = cv2.imread(filename)
    pipeline = get_pipeline() # returns the current Shabby Pages pipeline
    data = pipeline.augment(image)
    shabby_image = data["output"]
    cv2.imwrite(filename.parent / f"{filename.stem}-shabby{filename.suffix}"
```

```
pool.map(run_pipeline, filenames)
```

Processing all 6153 images took TODO:TIME on the 64 cores of a Graviton3 c7g.16xlarge instance on AWS.

5 Experimentation with the ShabbyPages Set

Binarization and denoising are two important techniques for the removal of unwanted data from images. Both approaches can be achieved by supervised learning of relationships between features of the input and output data. In this section, we train and compare several denoising models and their predictions on both the NoisyOffice and ShabbyPages datasets.

References

1. https://www.kaggle.com/datasets
2. https://huggingface.co/datasets
3. Christian Clausner, RDCL2019 Competition Dataset (Recognition of Documents with Complex Layouts) (RDCL2019) ,1,ID:RDCL2019_1,URL:https://tc11.cvc.uab.es/datasets/RDCL2019$_1$
4. David Doermann, Tobacco 800 Dataset (Tobacco800) ,1,ID:Tobacco800$_1$,URL : $https://tc11.cvc.uab.es/datasets/Tobacco800_1$
5. Harold Mouchère, ICDAR 2019 Competition on Recognition of Handwritten Mathematical Expressions and Typeset Formula Detection (ICDAR2019-CROHME-TDF) ,1,ID:ICDAR2019-CROHME-TDF_1,URL:https://tc11.cvc.uab.es/datasets/ICDAR2019-CROHME-TDF$_1$
6. Castro-Bleda, MJ.; España Boquera, S.; Pastor Pellicer, J.; Zamora Martínez, FJ. (2020). The NoisyOffice Database: A Corpus To Train Supervised Machine Learning Filters For Image Processing. The Computer Journal. 63(11):1658-1667. https://doi.org/10.1093/comjnl/bxz098
7. @articleDBLP:journals/corr/HarleyUD15, author = Adam W. Harley and Alex Ufkes and Konstantinos G. Derpanis, title = Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval, journal = CoRR, volume = abs/1502.07058, year = 2015, url = http://arxiv.org/abs/1502.07058, eprinttype = arXiv, eprint = 1502.07058, timestamp = Mon, 13 Aug 2018 16:48:54 +0200, biburl = https://dblp.org/rec/journals/corr/HarleyUD15.bib, bibsource = dblp computer science bibliography, https://dblp.org
8. F. Zamora-Martinez, S. España-Boquera and M. J. Castro-Bleda, Behaviour-based Clustering of Neural Networks applied to Document Enhancement, in: Computational and Ambient Intelligence, pages 144-151, Springer, 2007.
9. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
10. Journet Nicholas, Visani Muriel, Mansencal Boris, Van-Cuong Kieu, Billy Antoine, DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. J. Imaging 2017, 3, 62.
11.