Decision Engines Data Science Position Challenge

Simon Olsen

Approach

Thank you for the exciting challenge. Please see the attached script for my solution.
The three invoices differed quite a bit in their imperfections, which required different approaches for obtaining a best possible result for OCR.

I started by doing a manual analysis of the three invoices. My major concerns were related to the grayscale backgrounds in the forms for Gibson and Merck's invoices. My initial thought was to use opening (erode, then dilate) to try and remove the specks of the grayscaled backgrounds, while still retaining the important textual information. I knew OpenCV had such functionality ready, therefore I spent some time reading through their documentation on image processing, focusing on Smoothing and Morphological Transformations.

For Merck's invoice, opening seemed to be a good choice due to the high density of dots in the grayscale background and the small, but quite bold text.
For Gibson's invoice, however, an initial erosion would cause drastic information loss, especially in the contact information on the top left, due to its weakness and "white holes" within the text. Performing a closing before an opening would also not yield beneficial results, due to the grayscale background interfering with the text clarity. Smoothing, however, could efficiently remove the grayscale, without disrupting the text too much.

Upon experimenting I found a need to handle these two invoices quite differently. While an initial smoothing of Gibson's interface gave good results, it was too hard on the small text in Merck's invoice. Thus, I decided to implement a script with some options as to how the files should be processed.

Now, a simple opening on Merck's invoice did not give a perfect result. Some specks from the grayscale was still retained. An increase in the erosion's intensity was efficient in removal of the grayscale, but this also removed too much information from the foreground text. Upon closer inspection I found that after the initial erosion, much of these specks were single pixels that stood alone. In an attempt to further clear the grayscale background while still retaining the clarity of the text, I implemented a method for removing the pixels that stood alone, and ran this before the following dilation. This process is run by passing the argument --hard-open (or -ho) when running the script. The pixel removal method can be found in the script defined as remove_isolated_pixels.

The original GE Healthcare invoice was, in my opinion, quite good. Though, the top right box had a quite noisy black background, and some of the text was a bit thin. A simple smoothing reduced the noise drastically, while closing and dilating enhanced the text's clarity.

The Script

Please see the email attachments for the script. (ocr_preprocessor.py)

The script contains the following optional arguments:

--input, -i: Input Destination
Path to file, or contasining folder for process. If given path is a folder, all containing TIF's will be processed.

--out, -o: Output Destination
Path to folder where the optimized files should be stored.

--blur, -b: Blur
If provided performs median blur on the input.This helps remove unwanted grayscale at the cost of text clarity.

--hard-open, -ho: Hard Open
If given performs a hard open on the input. This helps remove grayscale but at a high cost on weak and unclear text with much "holes” within.

--open, -op: Open
If given performs a normal open on the input.

--close, -c: Close
If given performs a normal close on the input.

--erode, -e: Erode
If given performs a normal erosion on the input.

--dilate, -di: Dilate
If given performs a normal dilation on the input.

--debug, -d: Debug
If given prints debugging information, and shows one large and one small preview of the processed image.

Obtaining the Results from the Script

Below follows the commands to use with the script to obtain the attached results for each of the single invoices. If no output argument is passed, the result wl be stored in a subfolder named “results” with the same filename.
The instructions uses relative paths and assumes that the current directory contains both the script and each of the original invoices.
The scripts also assumes the following filenames for the invoices:

- *merck.tif*
- *gibson.tif*
- *ge.tif*

Merck
Only the extended opening (*hard_open*) is run.

>> python .\ocr_preprocessor.py -i merck.tif -ho

Gibson
First runs median blurring to clear background noise followed by a dilation and closing in an attempt to fill in gaps in the weakly visible text.

>> python .\ocr_preprocessor.py -i gibson.tif -b -di -c

GE Healthcare
A similar procedure to the one used on the Gibson invoice gave a good result. Swapping the order of the closing and dilation also further increased the clarity of the text.

>> python .\ocr_preprocessor.py -d -i ge.tif -b -c -di

Results

Invoice

ORIGINAL

MERCK

MERCK SHARP & DOHME CORP.
ONE MERCK DR.
WHITEHOUSE STATION NJ 08889-3400

Ship To:
MCKESSON MEDICAL SURGICAL INC
1005 SATELLITE BLVD
SUWANEE GA 30024-2882

Bill To: MCKESSON MEDICAL SURGICAL INC
DOCUMENT PROCESSING
P.O. Box 4059
DANVILLE IL 61834-4059

Invoice Number
7011376429
Invoice Date
03/07/2018
Delivery Number
814256021
Plant
1899

Customer Number	Purchase Order Number	Order Reference	Order Number
50040679	22065294	AUTOSYS	3016582746

Description	Item Number	Batch No	Batch Expiry	Quantity	Unit Price	Extension
STERILE DILUENT 0.7ML 10X1DOSE VIAL	0000006430900	N031941	08/03/2020	1		0.00
Thank You For Your Order						
The net effective discount price of each discounted Merck product must be reported accurately to all public and private third-party payers (including state Medicaid agencies and the Medicare program), patients and other appropriate individuals or entities to the extent required under applicable law						

Payment Terms	Total Quantity	Total Amount
2% 35 Days Net 36 Days	1	0.00

Remarks

Bill To: MCKESSON MEDICAL — REMITTANCE STUB - Please detach and send with payment

Invoice Number	Invoice Date	Net Due Date	Net Amount Due
7011376429	03/07/2018	04/12/2018	0.00
Customer Number	Discount Amount	Discount Due Date	Discount Amount Due
50040679	0.00	04/11/2018	0.00

Please remit payment to:

Merck Sharp & Dohme Corp.
P.O. Box 5254
Carol Stream, IL 60197-5254

Customer Service:
800-MERCK-RX (800-637-2579)

Amount Paid

00500406797011376429000000000000000000000000000000000000000000000000000000000000

Gibson BIOSCIENCE

Invoice

LX-94747

1040 Manchester Street
Lexington, KY 40508
Toll Free: 800.477.4763
Phone: 859.254.9557
Web: www.gibsonbioscience.com

Acct. No	Date	Invoice #
109526.5	1/15/2018	LX-94747

Packing Slip # / Proforma #
LX-54576

COUNTRY OF ORIGIN: USA

Bill To
ATTN: ACCOUNTS PAYABLE
MCKESSON MEDICAL SURGICAL INC
PO BOX 4059
DOCUMENT PROCESSING
DANVILLE IL 61834

Ship To
ATTN: RECEIVING / LAB
MCKESSON MEDICAL SURGICAL INC KANSAS
1405 NORTH CHOUTEAU
KANSAS CITY PC 003
KANSAS CITY MO 64120

PO #	Ship Date	Shipping Method	Tracking #
21615548	1/15/2018	LEX-FedEx 2Day®	789346692527
Ship Note	**Ship Note Ref Numbers**	**Comments**	
		SHIPPING 01/15/2018	

Catalog #	Quantity	Description	Unit Price	Amount
60060	2	ID-AE KIT	107.90	215.80

Thank you for your business!

Subtotal	215.80
Other (Royalty)	0.00
Shipping Cost (LEX-FedEx 2Day®)	50.14
Total	**265.94**

***** EACH LOT OF THE PRODUCTS IN THIS SHIPMENT MEETS APPLICABLE NCCLS (M22) STANDARDS. *****
VISIT OUR WEB PAGE AT http://gibsonlabs.com.

Remittance Advice

Invoice # LX-94747

Customer 109526.5 MCKESSON MEDICAL SURGICAL INC-LEX : MCKESSON MEDICAL SURGICAL INC KANSAS-LEX

Amount Due 265.94

Amount Remitted _____

GE HEALTHCARE

DBA: GE Medical Systems Information Technologies, Inc

FEDERAL ID# 39-1046671

AMOUNT DUE:	**204.00** (US DOLLARS)
DUE DATE:	**08-FEB-18**
Remit to:	GE Medical Systems Information Technologies, Inc
US MAIL:	Attn: Accounts Receivable * 5517 Collections Center Drive * CHICAGO IL 60693
Wire/EFT Information:	ABA 021000021 ACCOUNT 304183911
If Wire/EFT, please email remittance advice to: Remit.Healthcare@ge.com	

REMIT INVOICE NUMBER: 2861892

INVOICE DATE: 09-JAN-18

CUSTOMER ACCT: 128009

GE REFERENCE#: 80797369

CUSTOMER PO#: 21732386

SOLD TO:

MCKESSON MEDICAL SURGICAL INC
ACCOUNTS PAYABLE
PO BOX 4059
DANVILLE,IL 61834-4059

SHIP TO:

MCKESSON MEDICAL SURGICAL INC
7343 S HARDY DR STE 101
TEMPE AZ 85283-4479

Payment Terms: NET 30	**Contract #:**	**GE BA Number:**	
Cost Center:	**FE Badge:**	**FE Name:**	**ProdLine:**
GE Sales Rep or FE: MS SM Central Admin cc		**SERV MANAGER :**	

Inquiries regarding this **Invoice** should be directed to: 1-800-581-5600

QUANTITY	ITEM NUMBER	DESCRIPTION	NET UNIT PRICE	EXTENDED AMOUNT OF BILLING
		THIS INVOICE AMOUNT IS DUE AND PAYABLE NET 30 FROM INVOICE DATE.		
2	002203	DURA-CUF, ADULT, 2 TB SUBMIN, 23 - 33 CM, 5/ BOX	102.00	204.00
		Tax USAZ_EX_O2 @ 7.40		0.00

Please include the Invoice / Credit Memo number for proper credit: 2861892	TOTAL	204.00
	Tax	0.00
	SHIPPING/HANDLING	0.00
	Total Amount	204.00

PAST DUE INVOICES ARE SUBJECT TO A SERVICE CHARGE OF 1.5% PER MONTH, NOT TO EXCEED THE MAXIMUM RATE ALLOWED BY LAW. ALL ORDERS SUBJECT TO GE HEALTHCARE TERMS AND CONDITIONS.

Goods and services or reimbursements associated with the ordered products or services and provided under contract without separately identified charges constitute discounts or other reductions in price under applicable federal law. It is the customer's responsibility to disclose such discounts or other reductions in price in the manner required under state or federal program which provides reimbursement to the customer for or related to the products or services under the contract.