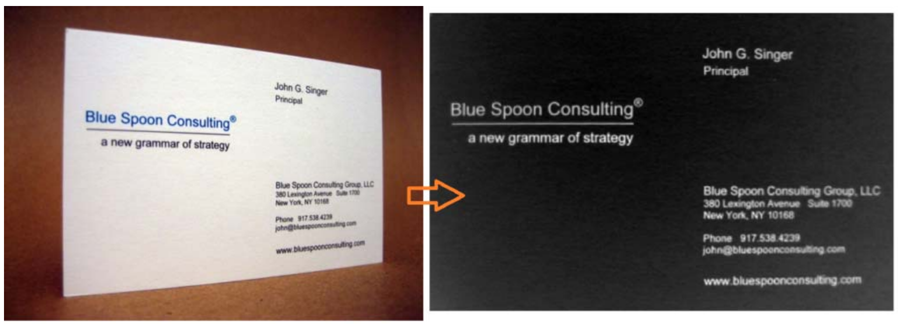
Reliancy

Textify: Extracting Structured Text from Images

July 17, 2017 Amer Agovic (http://reliancy.com/author/amer/)

In our previous post we showed how to apply computer vision methods to prepare an image of a document for text extraction. To make things interesting we picked a document, a business card, that was taken at an angle. Normally the projective distortion would make character recognition impossible. As a solution we showed how projective distortion can be reversed to obtain nice or rectified document devoid of any background clutter.

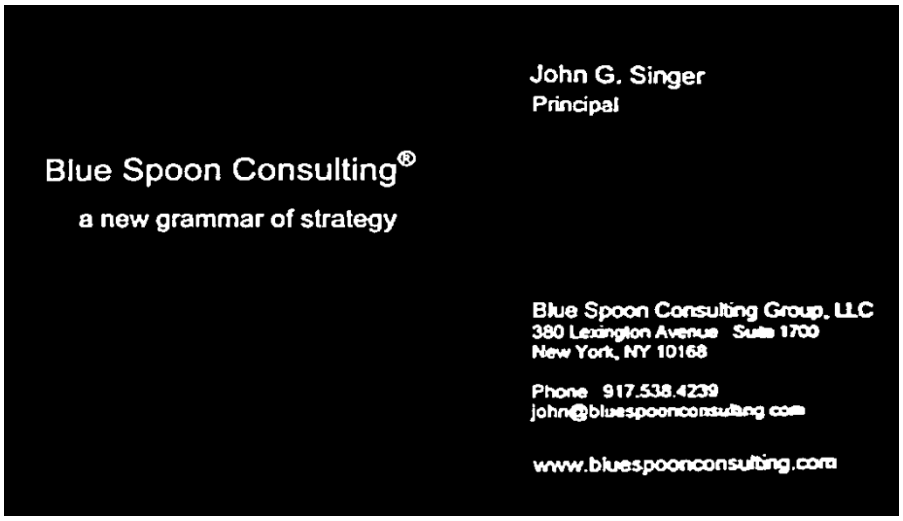

John G. Singer
Principal
Blue Spoon Consulting®
a new grammar of strategy
Blue Spoon Consulting Group, LLC
380 Lexington Avenue Suite 1700
New York, NY 10168
Phone 917.538.4239
john@bluespoonconsulting.com
www.bluespoonconsulting.com

The OCR engine performs its last step on a black/white binary image. If we do not binarize the input then OCR might do it but it might not do a good job. So we perform “thresholding” as one more step before OCR.

```
Mat binary= Mat.zeros(rectified.size(),rectified.type());
Imgproc.threshold(rectified,rectified,0,255,Imgproc.THRESH_BINARY+Imgproc.THRES
H_OTSU);
```

For thresholding we picked Otzu’s method (https://en.wikipedia.org/wiki/Otsu%27s_method). Unlike plain thresholding which requires us ,rightly or wrongly, to choose a thresholding value in Otzu’s method the value is automatically chosen for us. This method assumes an image with two groups of pixels and analyzes the histogram to find the value that best separates the two classes. Our text document is a perfect example of such image type because white text is one category and black background is the other. If we added colored or textured business cards our binary image would likely have more than two categories and Otzu’s method would fail.

The end result of thresholding or binarization should look similar to this:



John G. Singer
Principal

Blue Spoon Consulting®
a new grammar of strategy

Blue Spoon Consulting Group, LLC
380 Lexington Avenue Suite 1700
New York, NY 10168

Phone 917.538.4239
john@bluespoonconsulting.com

www.bluespoonconsulting.com

We emphasize “should” because we applied some additional denoising and clutter removal. As you can see with good results in some areas and not as good in others.

At last we come to the OCR stage. We feed the binary image to Tesseract and enumerate the results at word level:

```
Scalar CONTOUR_COLOR = new Scalar(0,255,0,255);
Mat rgb = new Mat(binary.size(),CvType.CV_8UC3);
ArrayList<Word> goodWords=new ArrayList<Word>();
BufferedImage img2=toBufferedImage(binary,null); // convert OpenCV image to Jav
a BufferedImage
ITesseract ocr=new Tesseract();
ocr.setDatapath(getTesseractFolder());
int pageIteratorLevel = TessPageIteratorLevel.RIL_WORD;
List<Word> result = ocr.getWords(img2, pageIteratorLevel);
//print the complete result
for (Word word : result) {
    if(word.getConfidence()>33.0f) goodWords.add(word);
    System.out.println("W:"+word.toString());
    Rectangle bb=word.getBoundingBox();
    Rect rect=new Rect(bb.x,bb.y,bb.width,bb.height);
    Imgproc.rectangle(rgb,new Point(rect.x,rect.y), new Point(rect.x+rect.widt
h,rect.y+rect.height),CONTOUR_COLOR);
}
```

The preceding code section will print out detected words and will also annotate an RGB image so we can show this:

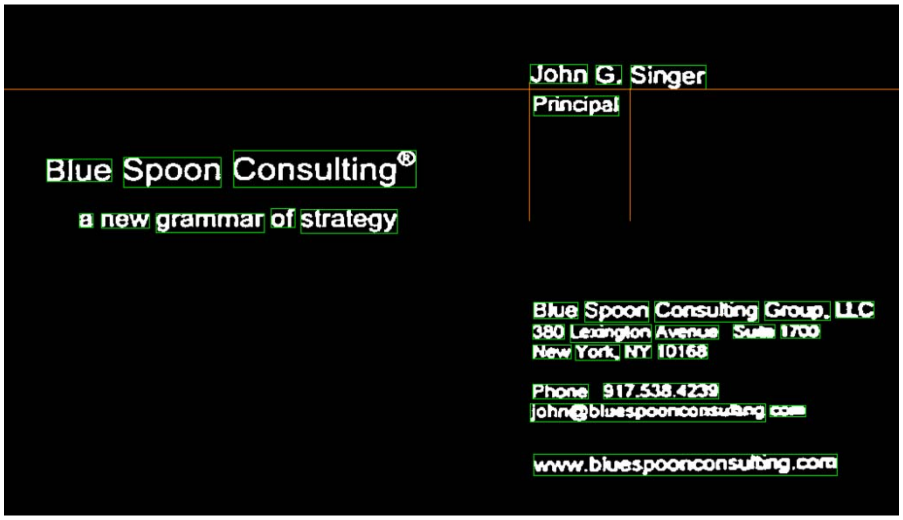
After we group all words that seem to lie on the same line we obtain our first textual result:

```
     0:                          John G. Singer
     1:                           Principal
     2: Blue Spoon Consulting®
     3:    a new grammar of strategy
     4:                             Blue Spoon (insulting Gnu). LLC
     5:                                    380 W Avenu- &- Im
     6:                                      New York NY 10168
     7:                                   Phone Sin-331m
     8:                                    johngbluupoormmm can
     9:                         www.bh:espooneonsd&lg.m
```

The results are good and bad. The left side of the document was clean and Tesseract did a good job. The right side was not binarized precisely enough. Part of it is due to binarization, another part is aliasing due to projective distortion. The image was low detail to begin with, after warping the right is blurred as you can see from the rectified image in the Computer Vision section. Amazingly the human eye can still distinguish and fill in the missing bits to recover the actual content. Tesseract not so much.

We undertake one extra step. We call it “Tabulating” and it is an attempt to recover tabular structure from the content. To that end we apply a ray casting algorithm that collects all possible tables. It returns back one dominant table. The dominant table is the one with the most columns and rows.

this case the selected table or its rays are depicted below:


John G. Singer
Principal

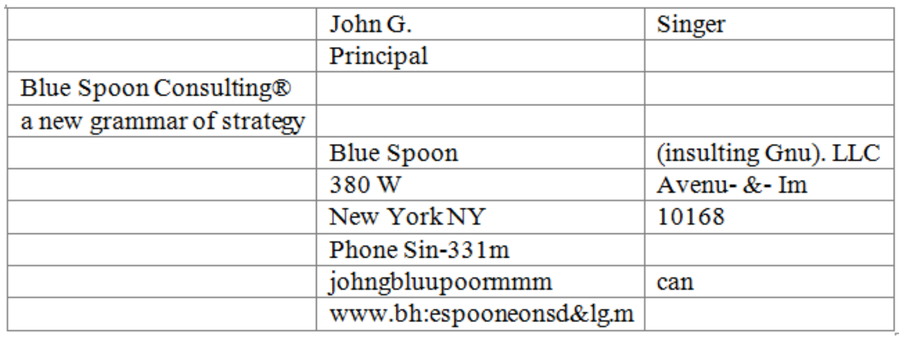
Blue Spoon Consulting®
a new grammar of strategy

Blue Spoon Consulting Group, LLC
380 Lexington Avenue Suite 1700
New York, NY 10168

Phone 917.538.4239
john@bluespoonconsulting.com

www.bluespoonconsulting.com

The resulting tabular structure and our final result are then:



	John G.	Singer
	Principal	
Blue Spoon Consulting®		
a new grammar of strategy		
	Blue Spoon	(insulting Gnu). LLC
	380 W	Avenu- &- Im
	New York NY	10168
	Phone Sin-331m	
	johngbluupoormmm	can
	www.bh:espooneonsd&lg.m	

The tabulation algorithm could clearly use some improving. For one it only considers left aligned columns. Moreover it does not deal with tables inside tables. So we will stop here since tabulation is a subject in its own right.

In this second part of our series we showed how images are prepared for character recognition, how recognized characters are enumerated and then formatted for further processing. The Computer Vision pre-processing and OCR stages have some limitations but perform with high reliability otherwise. Assuming that we have fine-tuned these stages the most interesting problem then becomes: Tabulating. That is recovering tabular structure so that we can locate content and properly channel it later on.

If you like to see the entire process live in action check it out at Reliancy Cloud (http://apps.reliancy.com/input/menu/main/DataCentral/Textify).

(http://reliancy.com/author/amer)

About the Author

Amer Agovic, PhD is the President and Founder at Reliancy. He has been the Chief Architect of the Reliancy's own software platform and has considerable experience in structuring and leading Software Development projects. His fields of expertise include Software Architecture, Robotics, Computer Vision and Natural Language Processing.

RELIANCY, LLC
1660 S Hwy 100, Suite 322
Saint Louis Park, MN 55416
(952)-406-8591

info@reliancy.com

DockML
Deployment Framework for Machine Learning Applications. Exposes Models as Services. Provides Workflow Orchestration, Exploration of Data and Effective Path To Production

DATA SCIENCE
Predictive Modeling
Data Exploration
Strategic Consulting
Big Data Platforms
Training

BLOG
Our Company Blog. Posts on Data Science, Data Virtualization, Application Development and related technology