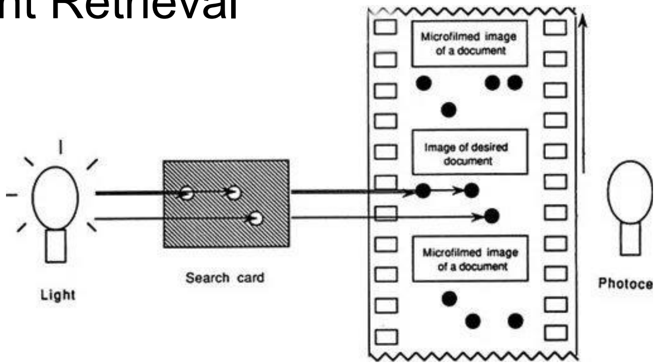
# Improvements to Offline Versus Online OCR

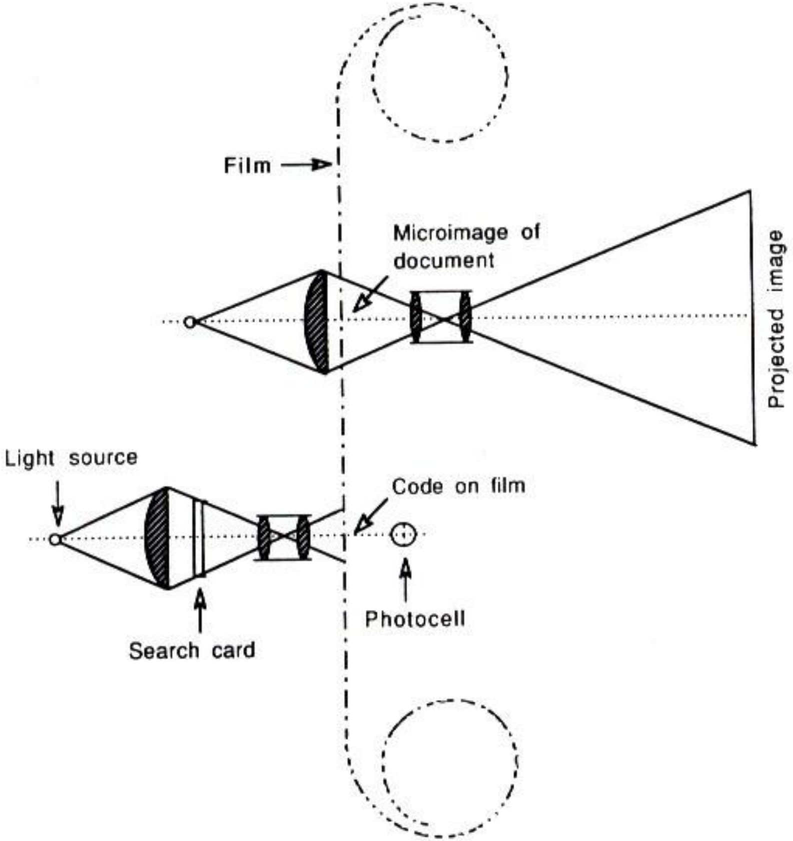
By Martin Cupic

# History

Origins:

- Emanuel Goldberg
  - Photoelectric cell and movie projector
  - Telegraphy
  - Document Retrieval


Microfilmed image of a document
Image of desired document
Microfilmed image of a document
Light
Search card
Photoce



# Types

- Intelligent Word Recognition
- Intelligent Character Recognition
- Optical Word Recognition
- Optical Character Recognition
- Optical Mark Recognition

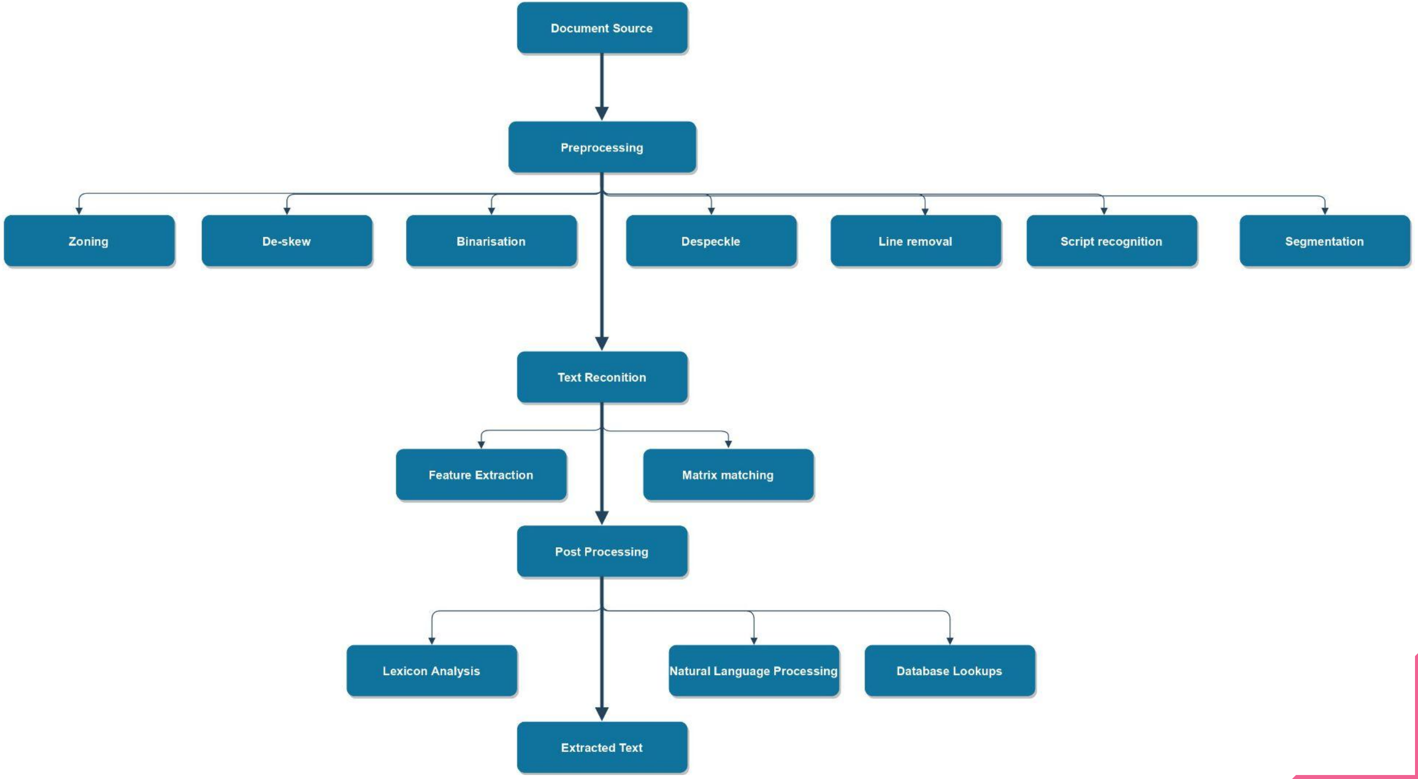
# Software Used to Execute OCR

- Tesseract
- Adobe Scan

Hypothesis:

Adobe scan will outperform

# How it works - Pipeline



Document Source

Preprocessing

Zoning

De-skew

Binarisation

Despeckle

Line removal

Script recognition

Segmentation

Text Reconition

Feature Extraction

Matrix matching

Post Processing

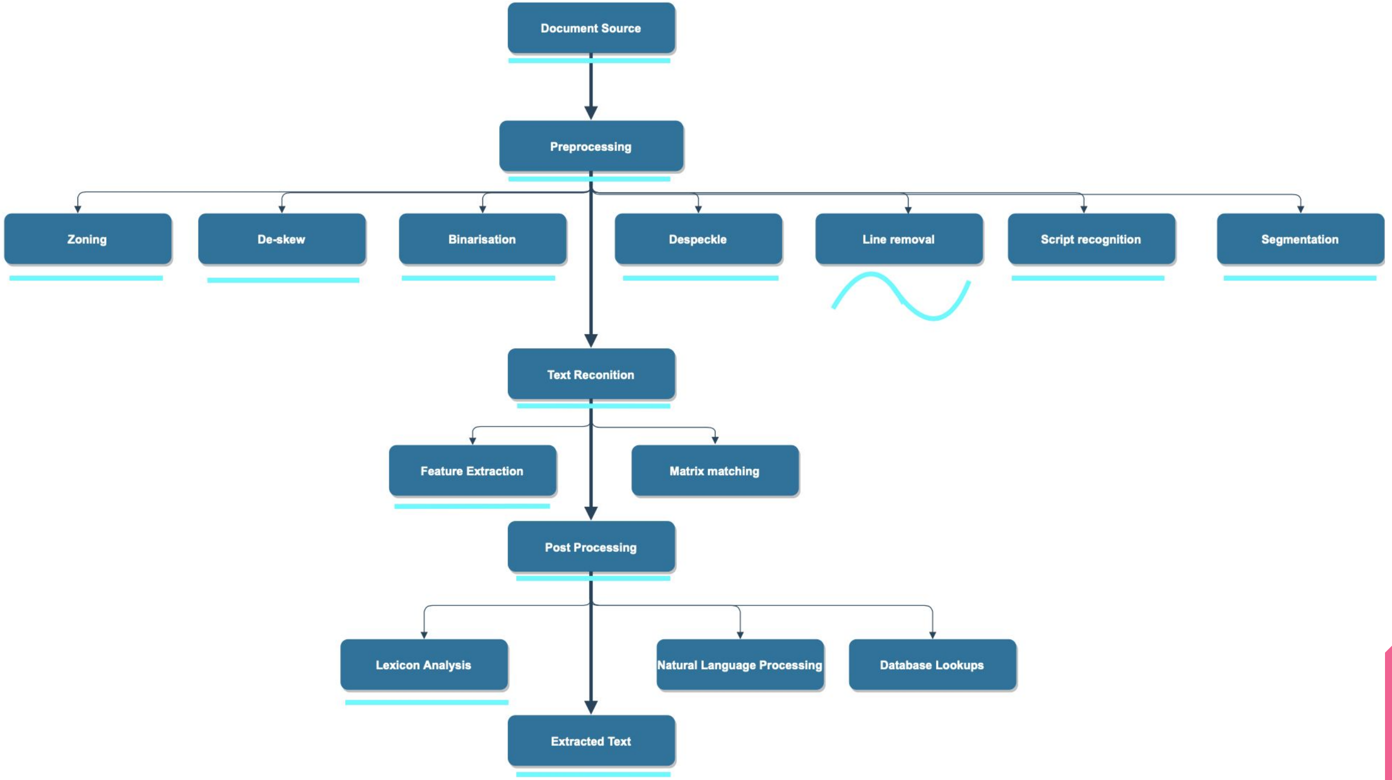
Lexicon Analysis

Natural Language Processing

Database Lookups

Extracted Text

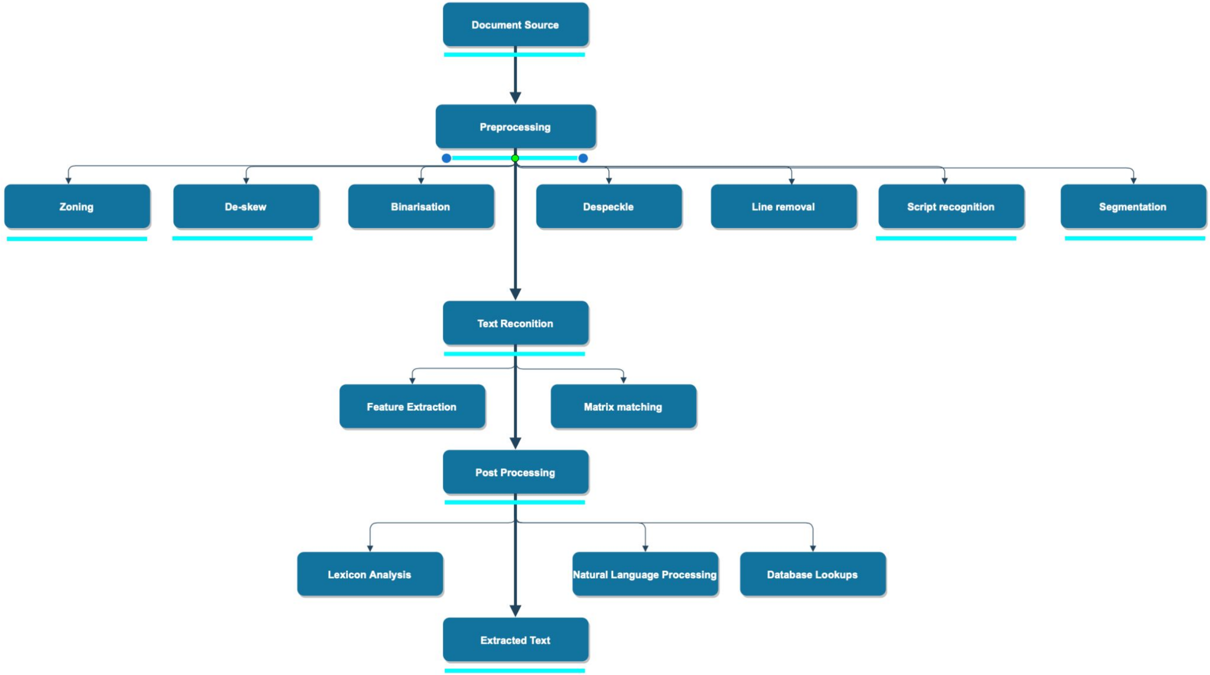
# Pipeline for Tesseract



- Tesseract is a offline and open source OCR software.
- Uses LSTM

# Pipeline for Adobe

- Adobe is an online OCR
  - runs OCR on the cloud



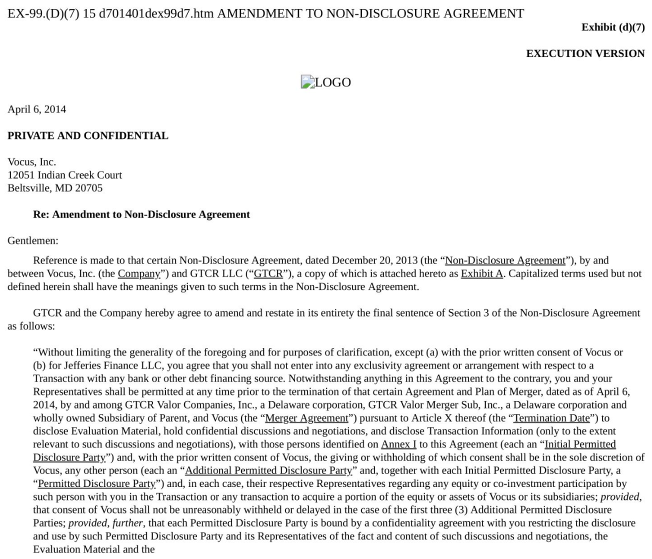
# Benchmark Datasets which were used

- For high resolution scanned documents:
  - https://github.com/applicaai/kleister-nda
- For low resolution scanned documents:
  - https://www.cs.cmu.edu/~aharley/rvl-cdip/
  - State-of-the-art: 92.21%

# Testing

High resolution scanned documents:

- Photos of scanned documents(100)



EX-99.(D)(7) 15 d701401dex99d7.htm AMENDMENT TO NON-DISCLOSURE AGREEMENT

**Exhibit (d)(7)**

**EXECUTION VERSION**

LOGO

April 6, 2014

**PRIVATE AND CONFIDENTIAL**

Vocus, Inc.
12051 Indian Creek Court
Beltsville, MD 20705

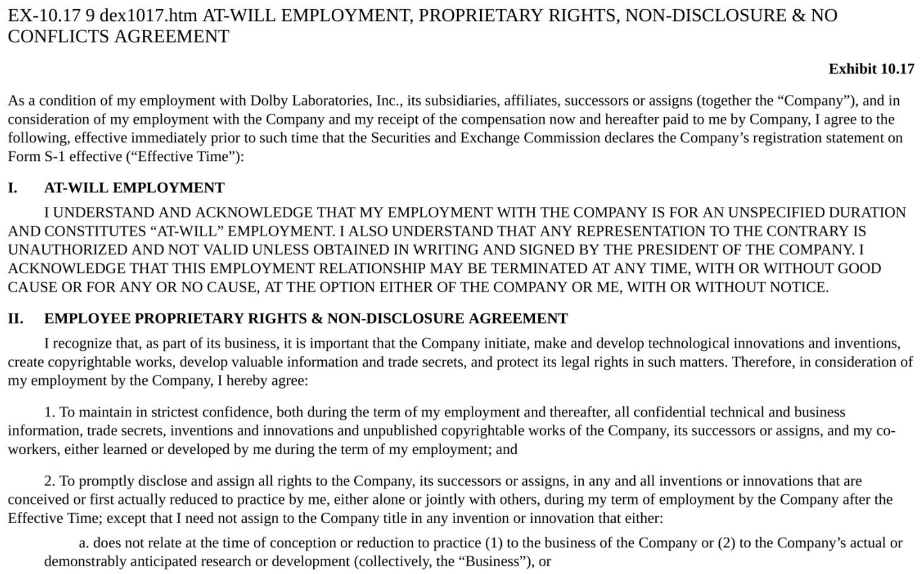
**Re: Amendment to Non-Disclosure Agreement**

Gentlemen:

Reference is made to that certain Non-Disclosure Agreement, dated December 20, 2013 (the “Non-Disclosure Agreement”), by and between Vocus, Inc. (the Company”) and GTCR LLC (“GTCR”), a copy of which is attached hereto as Exhibit A. Capitalized terms used but not defined herein shall have the meanings given to such terms in the Non-Disclosure Agreement.

GTCR and the Company hereby agree to amend and restate in its entirety the final sentence of Section 3 of the Non-Disclosure Agreement as follows:

> “Without limiting the generality of the foregoing and for purposes of clarification, except (a) with the prior written consent of Vocus or (b) for Jefferies Finance LLC, you agree that you shall not enter into any exclusivity agreement or arrangement with respect to a Transaction with any bank or other debt financing source. Notwithstanding anything in this Agreement to the contrary, you and your Representatives shall be permitted at any time prior to the termination of that certain Agreement and Plan of Merger, dated as of April 6, 2014, by and among GTCR Valor Companies, Inc., a Delaware corporation, GTCR Valor Merger Sub, Inc., a Delaware corporation and wholly owned Subsidiary of Parent, and Vocus (the “Merger Agreement”) pursuant to Article X thereof (the “Termination Date”) to disclose Evaluation Material, hold confidential discussions and negotiations, and disclose Transaction Information (only to the extent relevant to such discussions and negotiations), with those persons identified on Annex I to this Agreement (each an “Initial Permitted Disclosure Party”) and, with the prior written consent of Vocus, the giving or withholding of which consent shall be in the sole discretion of Vocus, any other person (each an “Additional Permitted Disclosure Party” and, together with each Initial Permitted Disclosure Party, a “Permitted Disclosure Party”) and, in each case, their respective Representatives regarding any equity or co-investment participation by such person with you in the Transaction or any transaction to acquire a portion of the equity or assets of Vocus or its subsidiaries; *provided*, that consent of Vocus shall not be unreasonably withheld or delayed in the case of the first three (3) Additional Permitted Disclosure Parties; *provided*, *further*, that each Permitted Disclosure Party is bound by a confidentiality agreement with you restricting the disclosure and use by such Permitted Disclosure Party and its Representatives of the fact and content of such discussions and negotiations, the Evaluation Material and the



EX-10.17 9 dex1017.htm AT-WILL EMPLOYMENT, PROPRIETARY RIGHTS, NON-DISCLOSURE & NO CONFLICTS AGREEMENT

**Exhibit 10.17**

As a condition of my employment with Dolby Laboratories, Inc., its subsidiaries, affiliates, successors or assigns (together the “Company”), and in consideration of my employment with the Company and my receipt of the compensation now and hereafter paid to me by Company, I agree to the following, effective immediately prior to such time that the Securities and Exchange Commission declares the Company’s registration statement on Form S-1 effective (“Effective Time”):

**I. AT-WILL EMPLOYMENT**

I UNDERSTAND AND ACKNOWLEDGE THAT MY EMPLOYMENT WITH THE COMPANY IS FOR AN UNSPECIFIED DURATION AND CONSTITUTES “AT-WILL” EMPLOYMENT. I ALSO UNDERSTAND THAT ANY REPRESENTATION TO THE CONTRARY IS UNAUTHORIZED AND NOT VALID UNLESS OBTAINED IN WRITING AND SIGNED BY THE PRESIDENT OF THE COMPANY. I ACKNOWLEDGE THAT THIS EMPLOYMENT RELATIONSHIP MAY BE TERMINATED AT ANY TIME, WITH OR WITHOUT GOOD CAUSE OR FOR ANY OR NO CAUSE, AT THE OPTION EITHER OF THE COMPANY OR ME, WITH OR WITHOUT NOTICE.

**II. EMPLOYEE PROPRIETARY RIGHTS & NON-DISCLOSURE AGREEMENT**

I recognize that, as part of its business, it is important that the Company initiate, make and develop technological innovations and inventions, create copyrightable works, develop valuable information and trade secrets, and protect its legal rights in such matters. Therefore, in consideration of my employment by the Company, I hereby agree:

1. To maintain in strictest confidence, both during the term of my employment and thereafter, all confidential technical and business information, trade secrets, inventions and innovations and unpublished copyrightable works of the Company, its successors or assigns, and my co-workers, either learned or developed by me during the term of my employment; and

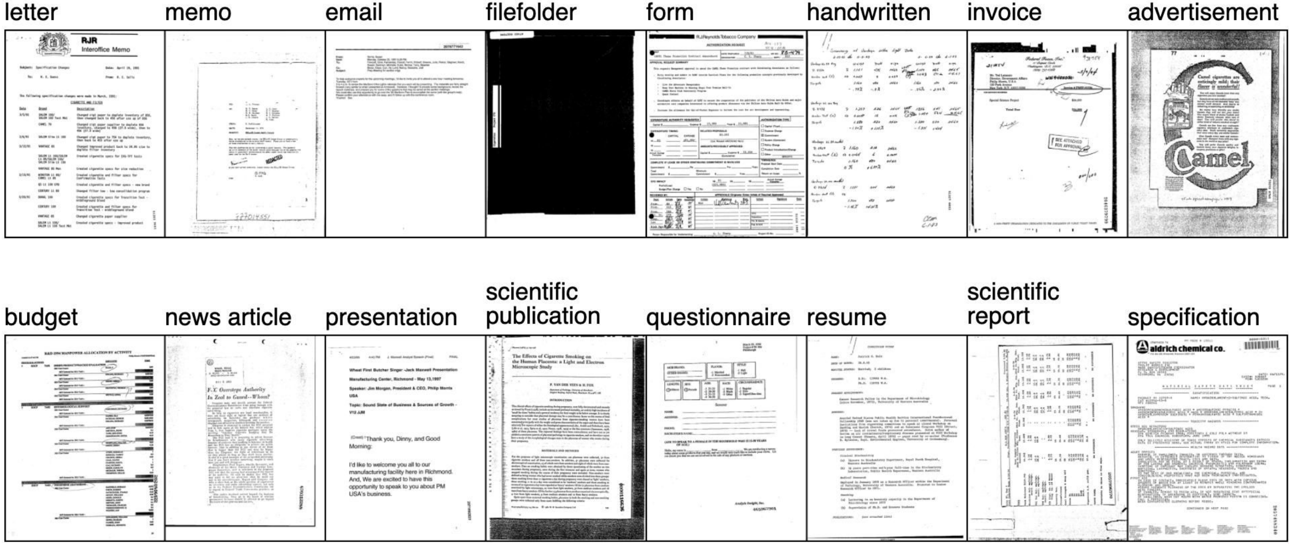
2. To promptly disclose and assign all rights to the Company, its successors or assigns, in any and all inventions or innovations that are conceived or first actually reduced to practice by me, either alone or jointly with others, during my term of employment by the Company after the Effective Time; except that I need not assign to the Company title in any invention or innovation that either:

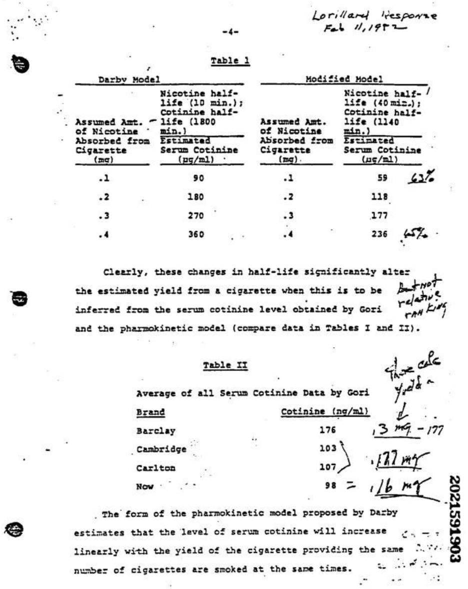
   a. does not relate at the time of conception or reduction to practice (1) to the business of the Company or (2) to the Company’s actual or demonstrably anticipated research or development (collectively, the “Business”), or

# Testing

Low resolution scanned documents:

- Multiple types of scanned documents(excl. handwritten)





Lorillard Response
Feb 11, 1982

-4-

Table 1

| Darby Model | | Modified Model | | |
|---|---|---|---|---|
| | Nicotine half-life (10 min.); Cotinine half-life (1800 min.) | | Nicotine half-life (40 min.); Cotinine half-life (1140 min.) | |
| Assumed Amt. of Nicotine Absorbed from Cigarette (mg) | Estimated Serum Cotinine (μg/ml) | Assumed Amt. of Nicotine Absorbed from Cigarette (mg) | Estimated Serum Cotinine (μg/ml) | |
| .1 | 90 | .1 | 59 | 63% |
| .2 | 180 | .2 | 118 | |
| .3 | 270 | .3 | 177 | |
| .4 | 360 | .4 | 236 | 65% |

Clearly, these changes in half-life significantly alter the estimated yield from a cigarette when this is to be inferred from the serum cotinine level obtained by Gori and the pharmokinetic model (compare data in Tables I and II).

But not relative ranking

Table II

Average of all Serum Cotinine Data by Gori

These calc yield

| Brand | Cotinine (ng/ml) | |
|---|---|---|
| Barclay | 176 | .3 mg - 177 |
| Cambridge | 103 | .177 mg |
| Carlton | 107 | |
| Now | 98 | = .16 mg |

The form of the pharmokinetic model proposed by Darby estimates that the level of serum cotinine will increase linearly with the yield of the cigarette providing the same number of cigarettes are smoked at the same times.

2021591903

# Results

**Tesseract**

High quality scanned documents:

- 100% accuracy. Confusion shapes/icons with characters.

Low quality scanned documents:

- Extracts text accurately rarely and is unable to extract full texts
- 33% accuracy

Overall:

- It is the state of the art OCR software when it comes to high quality images
- Very poor performance when it receives low quality images

# Results

**Adobe Scan**

High quality scanned documents:

- 100% accuracy

Low quality scanned documents:

- Extracts some of the text
- Small mistakes
- 54% accuracy

Overall:

- Adobe is clearly a superior OCR but still performs poorly with very low resolution images

# Results

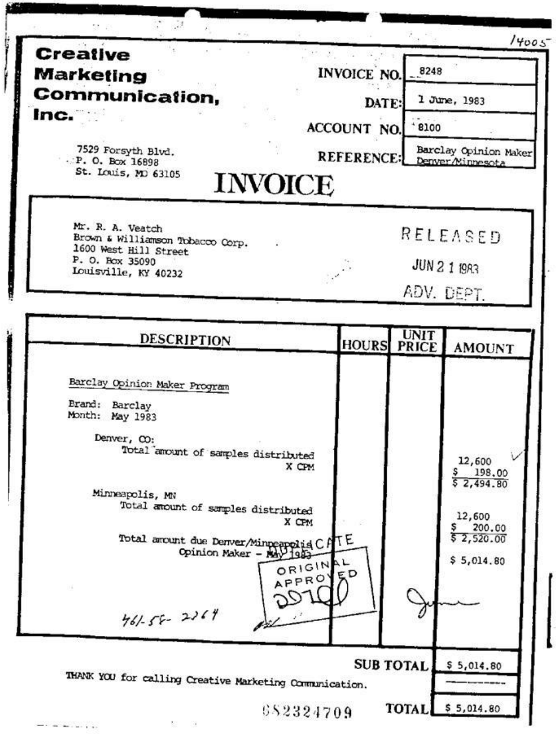
Example of results from low resolution scanned document

**Tesseract**:

- Unable to read stamps, misplaces characters, and unable to extract some characters
  - Overall character accuracy: 62%

**Adobe**

- Able to extract everything other than the bottom stamp. Some character mistakes and combination of words.
  - Overall character accuracy: 84%



14005

**Creative Marketing Communication, Inc.**

7529 Forsyth Blvd.
P. O. Box 16898
St. Louis, MO 63105

| | |
|---|---|
| INVOICE NO. | 8248 |
| DATE: | 1 June, 1983 |
| ACCOUNT NO. | 8100 |
| REFERENCE: | Barclay Opinion Maker<br>Denver/Minnesota |

INVOICE

Mr. R. A. Veatch
Brown & Williamson Tobacco Corp.
1600 West Hill Street
P. O. Box 35090
Louisville, KY 40232

RELEASED
JUN 21 1983
ADV. DEPT.

| DESCRIPTION | HOURS | UNIT PRICE | AMOUNT |
|---|---|---|---|
| Barclay Opinion Maker Program | | | |
| Brand: Barclay<br>Month: May 1983 | | | |
| Denver, CO:<br>Total amount of samples distributed<br>X CPM | | | 12,600<br>$ 198.00<br>$ 2,494.80 |
| Minneapolis, MN<br>Total amount of samples distributed<br>X CPM | | | 12,600<br>$ 200.00<br>$ 2,520.00 |
| Total amount due Denver/Minneapolis<br>Opinion Maker - May 1983 | | | $ 5,014.80 |

DUPLICATE
ORIGINAL APPROVED
20700

461-58-2264

| | |
|---|---|
| SUB TOTAL | $ 5,014.80 |
| | ---------- |
| TOTAL | $ 5,014.80 |

THANK YOU for calling Creative Marketing Communication.

682324709

# Results

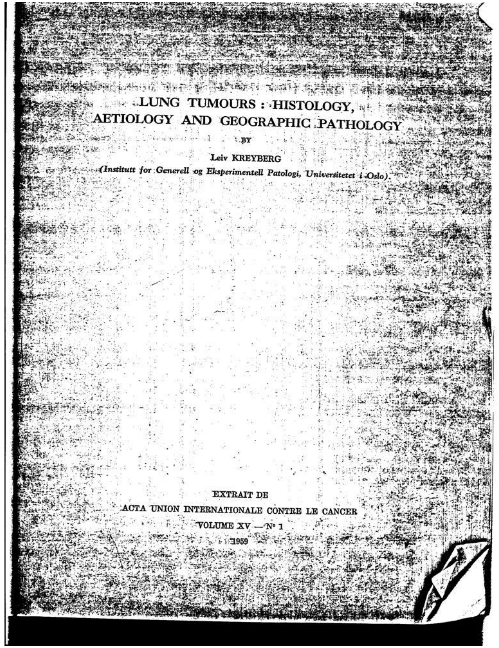
Example of results from low resolution scanned document

**Tesseract**:

- unable to extract most of the characters
  - Overall accuracy: 27%

**Adobe**:

- Manages to extract most of the text but the majority of it has mistakes(i.e. B getting extracted instead of E in word ”ETRAIT”)
  - Overall accuracy: 53%



LUNG TUMOURS : HISTOLOGY,
AETIOLOGY AND GEOGRAPHIC PATHOLOGY

BY

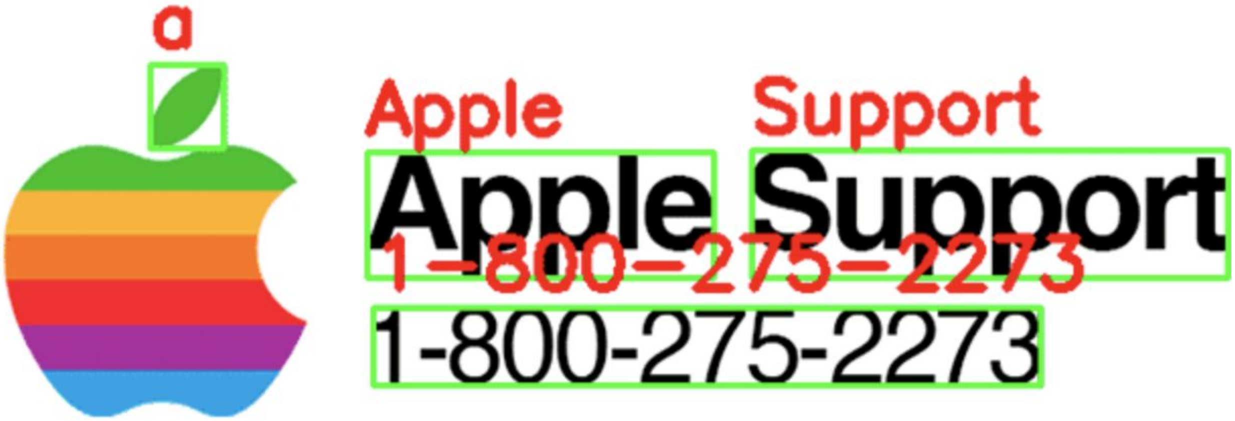
Leiv KREYBERG

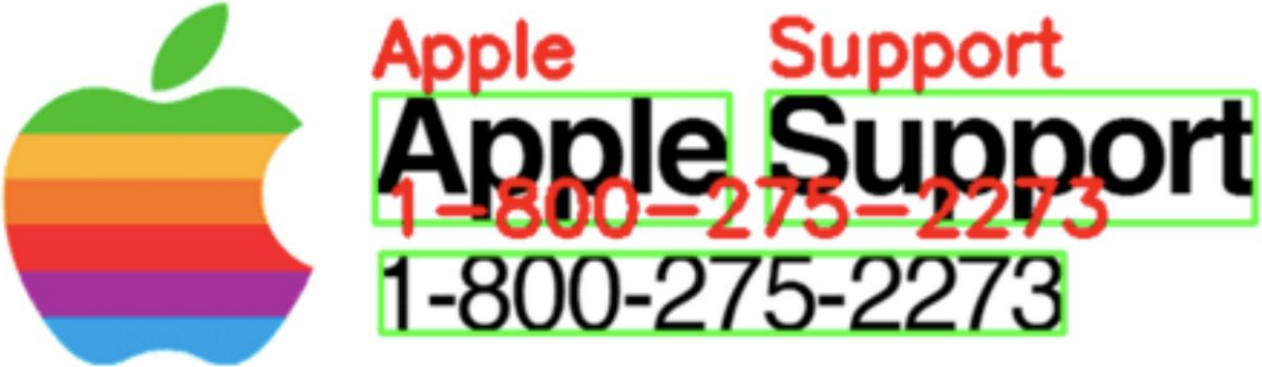
(Institutt for Generell og Eksperimentell Patologi, Universitetet i Oslo).

EXTRAIT DE
ACTA UNION INTERNATIONALE CONTRE LE CANCER
VOLUME XV — N° 1
1959

# Problem 1

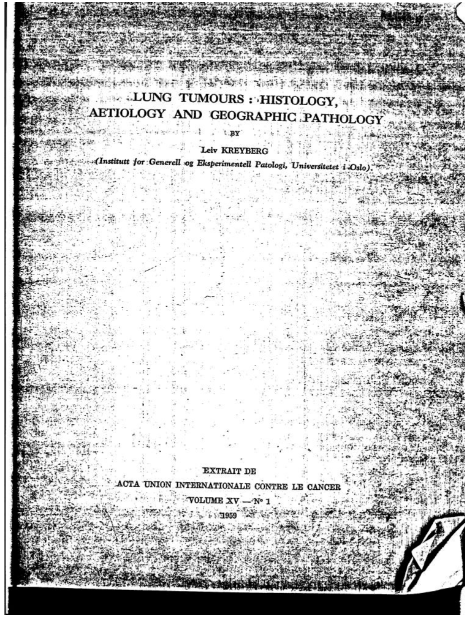
Min. Confidence:





# Problem 2

- Low resolution
  - Blur around characters
  - Specks from scanning



LUNG TUMOURS : HISTOLOGY,
AETIOLOGY AND GEOGRAPHIC PATHOLOGY

BY

Leiv KREYBERG

*(Institutt for Generell og Eksperimentell Patologi, Universitetet i Oslo).*

EXTRAIT DE
ACTA UNION INTERNATIONALE CONTRE LE CANCER
VOLUME XV — N° 1
1959

# Improvements

2 improvements:

- **Improvement** of confusion of shapes with letters:
    - Setting a minimum value
    - "--min-conf #", --min-conf 50

- **Improvement** of executing OCR on image with low resolution:
    - Increasing the contrast and sharpness

# Increasing Sharpness and Contrast

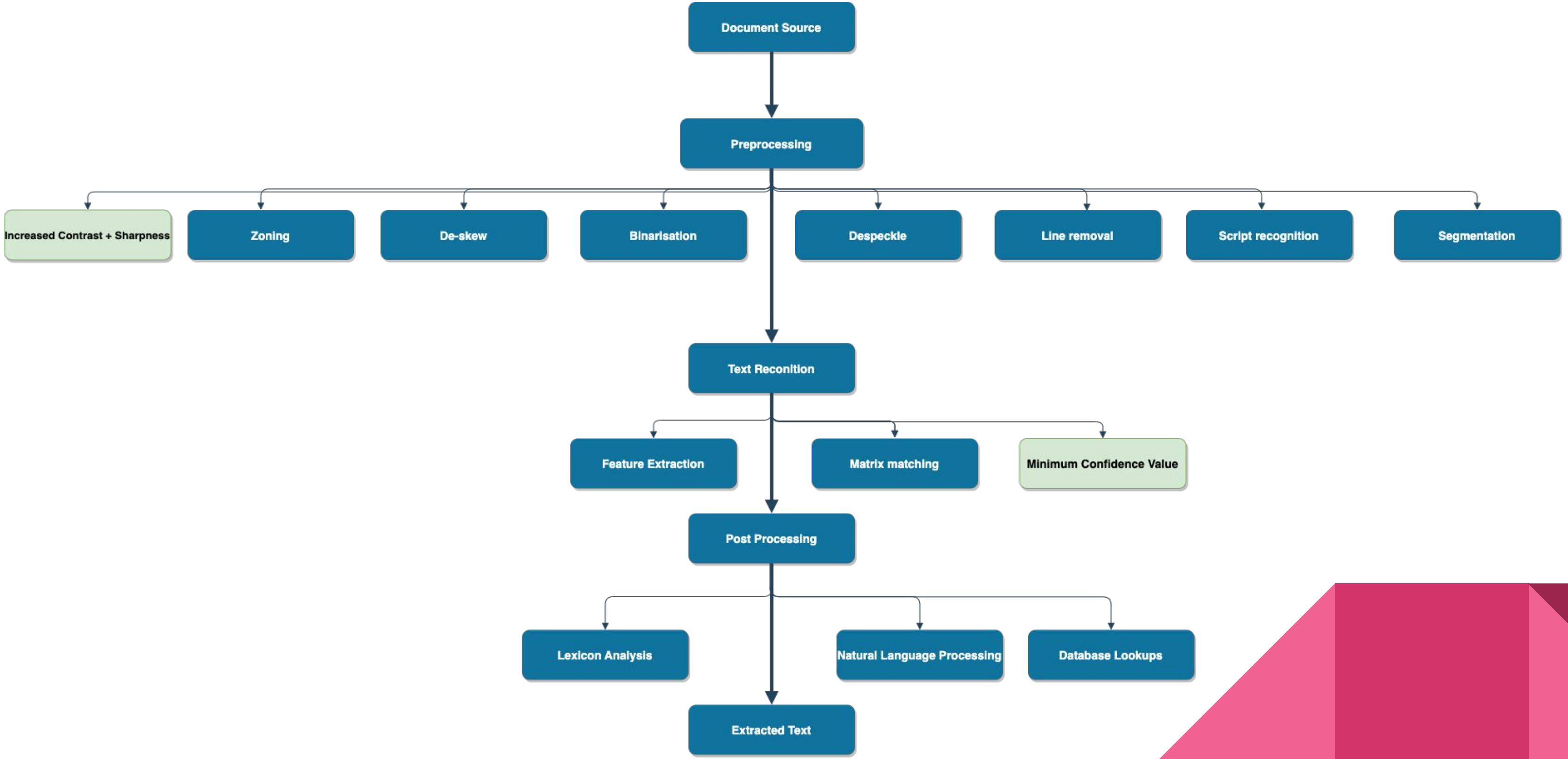
Pillow (PIL Fork)

```
filter = ImageEnhance.Contrast(img)

img.filter(float)


filter = ImageEnhance.Sharpness(img)

img.filter(float)
```

# Increasing Sharpness and Contrast

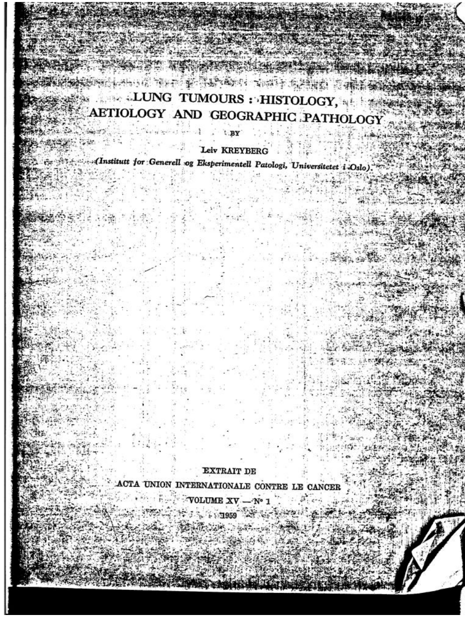
- Contrast 10%
- Sharpness 50%

# Improvement Pipeline



# Result after improvement

Example



LUNG TUMOURS : HISTOLOGY,
AETIOLOGY AND GEOGRAPHIC PATHOLOGY

BY

Leiv KREYBERG

*(Institutt for Generell og Eksperimentell Patologi, Universitetet i Oslo).*

EXTRAIT DE
ACTA UNION INTERNATIONALE CONTRE LE CANCER
VOLUME XV — N° 1
1959

**Tesseract**:

- Extracts slightly more words(extracted one more word in this example)
    - Overall accuracy: from 27% to 31%

**Adobe**:

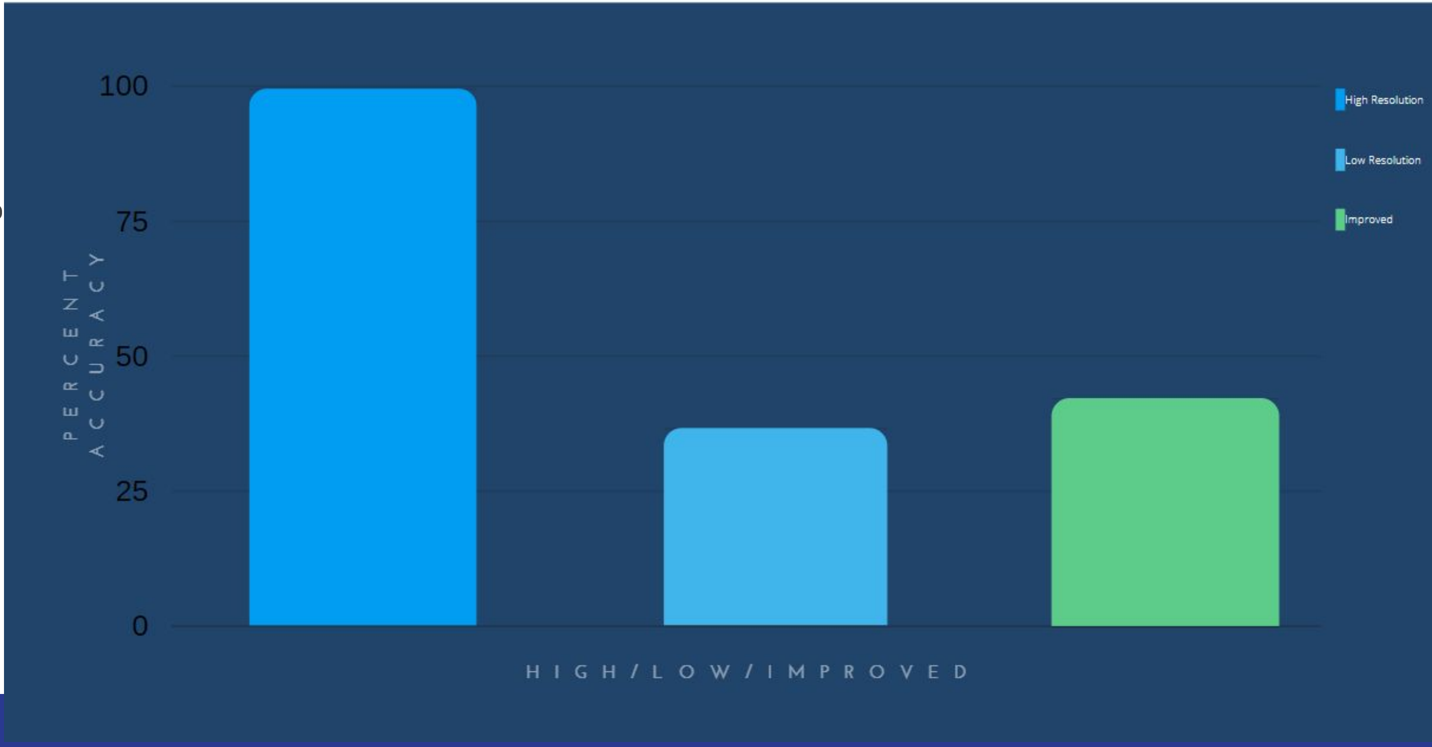
- Extracts slightly more text, does not combine words as much and extracts words more accurately
    - Overall accuracy: from 53% to 63%

# Final results:

## HIGH RESOLUTION VS LOW RESOLUTION VS IMPROVED LOW RESOLUTION

Tesseract:

- High: 100%
- Low: 33%
- Improved: 40%

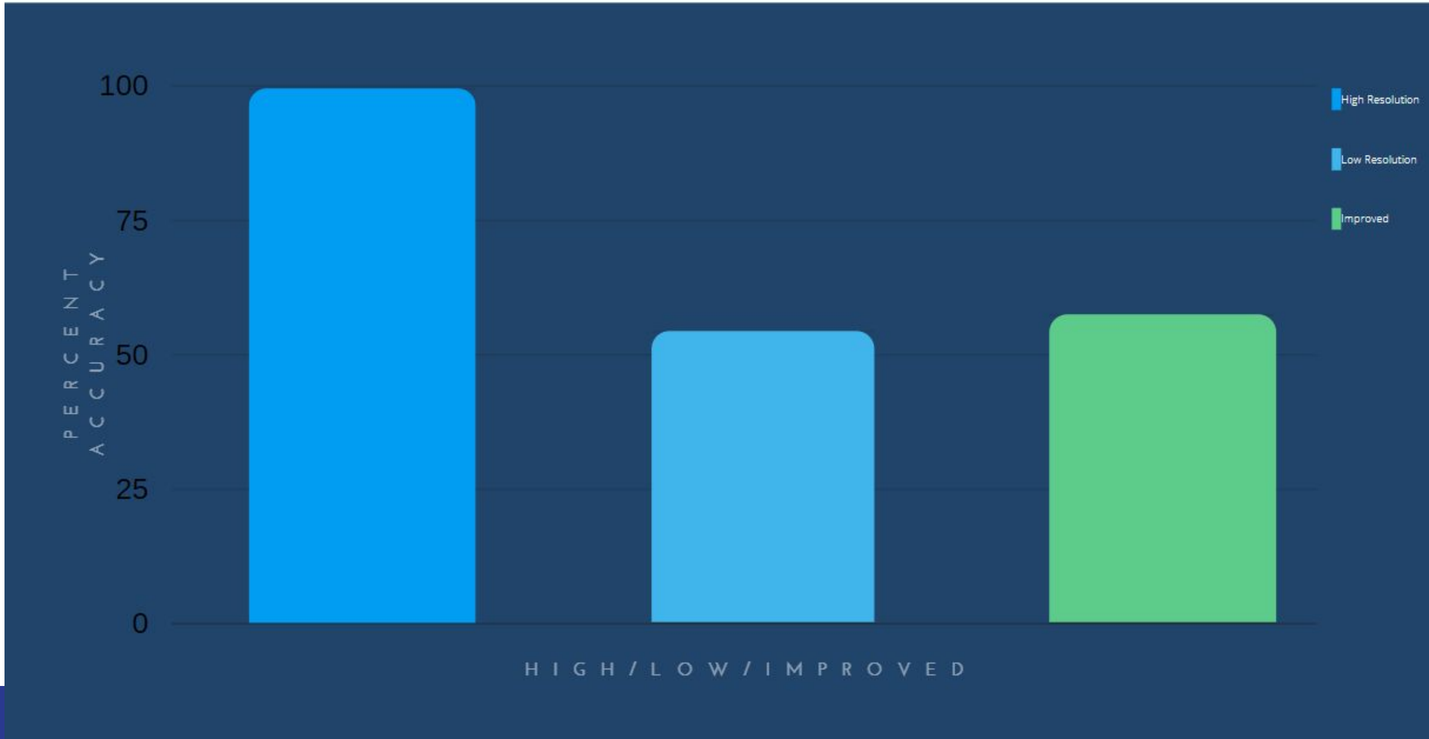


# Final results:

Adobe:

- High: 100%
- Low: 54%
- Improved: 58%

## HIGH RESOLUTION VS LOW RESOLUTION VS IMPROVED LOW RESOLUTION



# Conclusion after improvements

--min-conf 50:

- Confussion between symbols/icons and characters disappeared

Sharpness +50% AND Contrast +10%:

- Removes some blur around characters due to sharpness and makes the text more visible due to contrast

# Conclusion

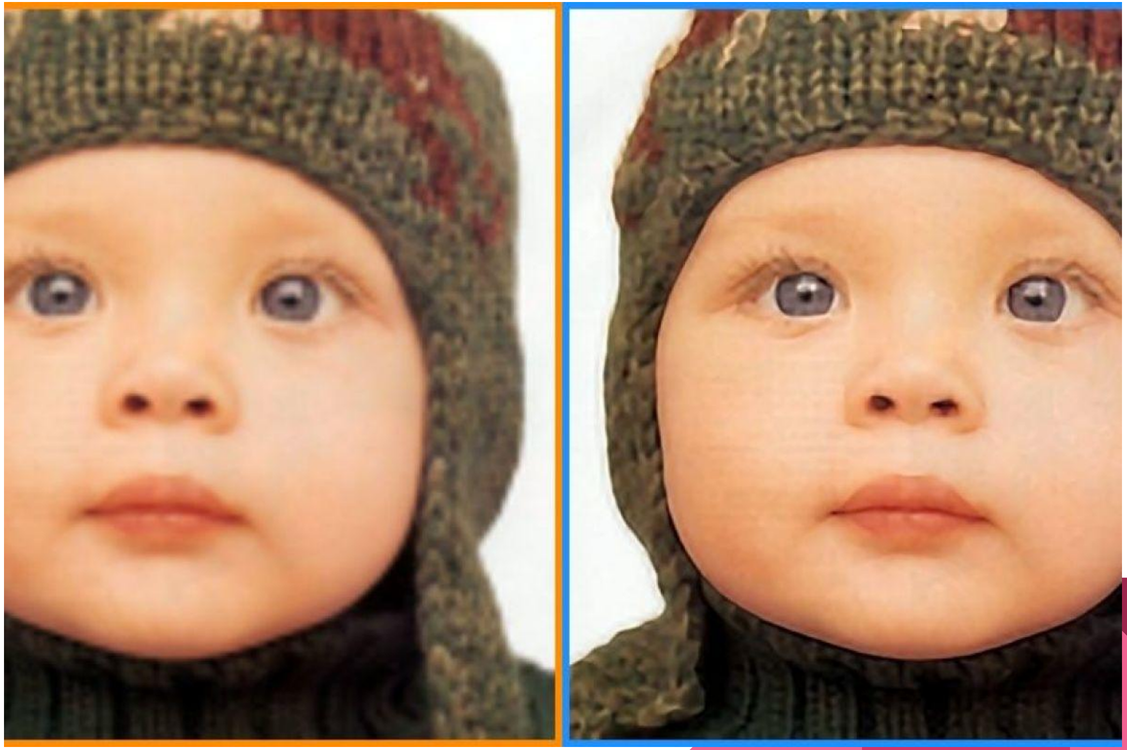
- Improvements not good enough to let it compete against state-of-the art online OCR software
  - 92.21% is the state-of the-art on this dataset

- Increase in accuracy
  - Tesseract: 33%-40%
  - Adobe: 54%-58%

- Biggest problem for OCR is image resolution

# Future Improvement: Super-resolution from OpenCV

- Super-resolution



# Bibliography

G. Chiron, A. Doucet, M. Coustaty and J. -P. Moreux, "ICDAR2017 Competition on Post-OCR Text Correction," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1423-1428, doi: 10.1109/ICDAR.2017.232.

Kae, Andrew & Huang, Gary & Doersch, Carl & Learned-Miller, Erik. (2010). Improving state-of-the-art OCR through high-precision document-specific modeling. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1935-1942. 10.1109/CVPR.2010.5539867.

Reul, Christian & Springmann, Uwe & Wick, Christoph & Puppe, Frank. (2018). State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines.

Netzer, Yuval & Wang, Tao & Coates, Adam & Bissacco, Alessandro & Wu, Bo & Ng, Andrew. (2011). Reading Digits in Natural Images with Unsupervised Feature Learning. NIPS.

*RVL-CDIP-I Dataset*. (2019, August 18). [Dataset]. https://www.kaggle.com/datasets/nbhativp/first-half-training

A. (2021). *GitHub - applicaai/kleister-nda* [Dataset]. https://github.com/applicaai/kleister-nda

Images:

Dodson, B. (2021, July 8). *It's all in the detail: Impressive new approach to super-resolution processing developed*. New Atlas.

https://newatlas.com/super-resolution-weizmann-institute/23486/

Rosebrock, A. (2021, October 25). *Tesseract OCR: Text localization and detection*. PyImageSearch. https://pyimagesearch.com/2020/05/25/tesseract-ocr-text-localization-and-detection/

*Goldberg Statistical machine*. (2006). School of Information Management & Systems.

https://people.ischool.berkeley.edu/%7Ebuckland/statistical.html