

# Preprocessing and Layout Analysis for Offline Handwriting Recognition

---

東北大学大学院工学研究科  
電子工学専攻川又研究室  
Perttu Pitkänen

2016年02月26日

# Overview

---

- Handwriting recognition
- Preprocessing
  - Stroke Width Analysis
- Layout Analysis
  - Bounding Box Expansion
  - RLSA
- Tests
- Test Results
- Conclusions
- Remaining Problems
- Future work

# Handwriting recognition 1

---

- Offline handwriting recognition (HWR) is the process of extracting text in digital form from handwritten images.
- Offline recognition process can be divided into three main phases:
  - Preprocessing
  - Feature extraction
  - Classification
- Implementation done with MATLAB and its image processing toolbox

# Handwriting recognition 2

---

## ■ Preprocessing

- Image is enhanced for feature extraction phase and the detected characters are segmented from the original image.
- Layout analysis can be considered to be a part of preprocessing.

## ■ Feature Extraction

- Shape describing features are extracted from previously acquired objects (words).

## ■ Classification

- Extracted features are used in machine learning algorithms to create the feature vector and to classify the inputs into word classes.

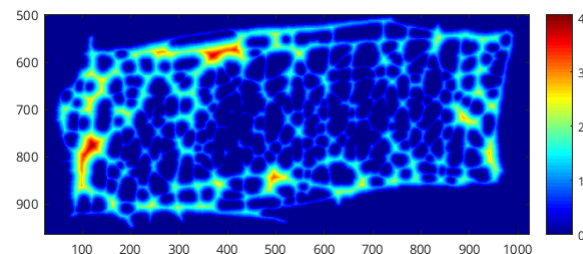
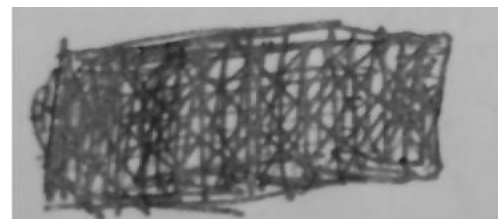
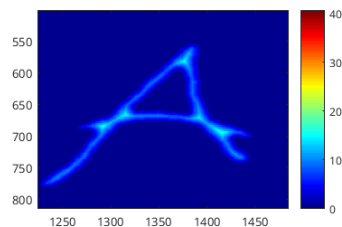
# Preprocessing

---

- Most of the preprocessing is same than previously
  - Image aquisition
  - Noise removal
    - Adaptive Wiener filter
  - Binarization
    - Sauvola algorithm
  - Object property analysis
    - ~~■ Features such as holes in object, size, area or aspect ratio~~
    - Stroke width variation
- All methods need pre-defined parameters!
- Object property analysis now uses sroke width instead of other features.
- Majority of methods proved to be useful in preprocessing.

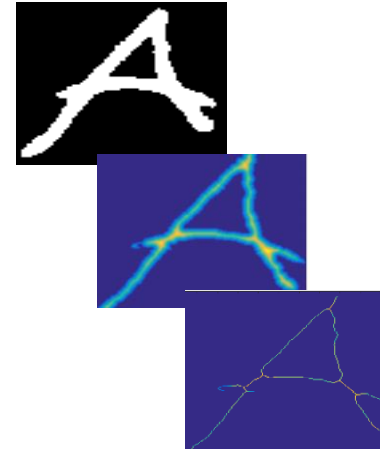
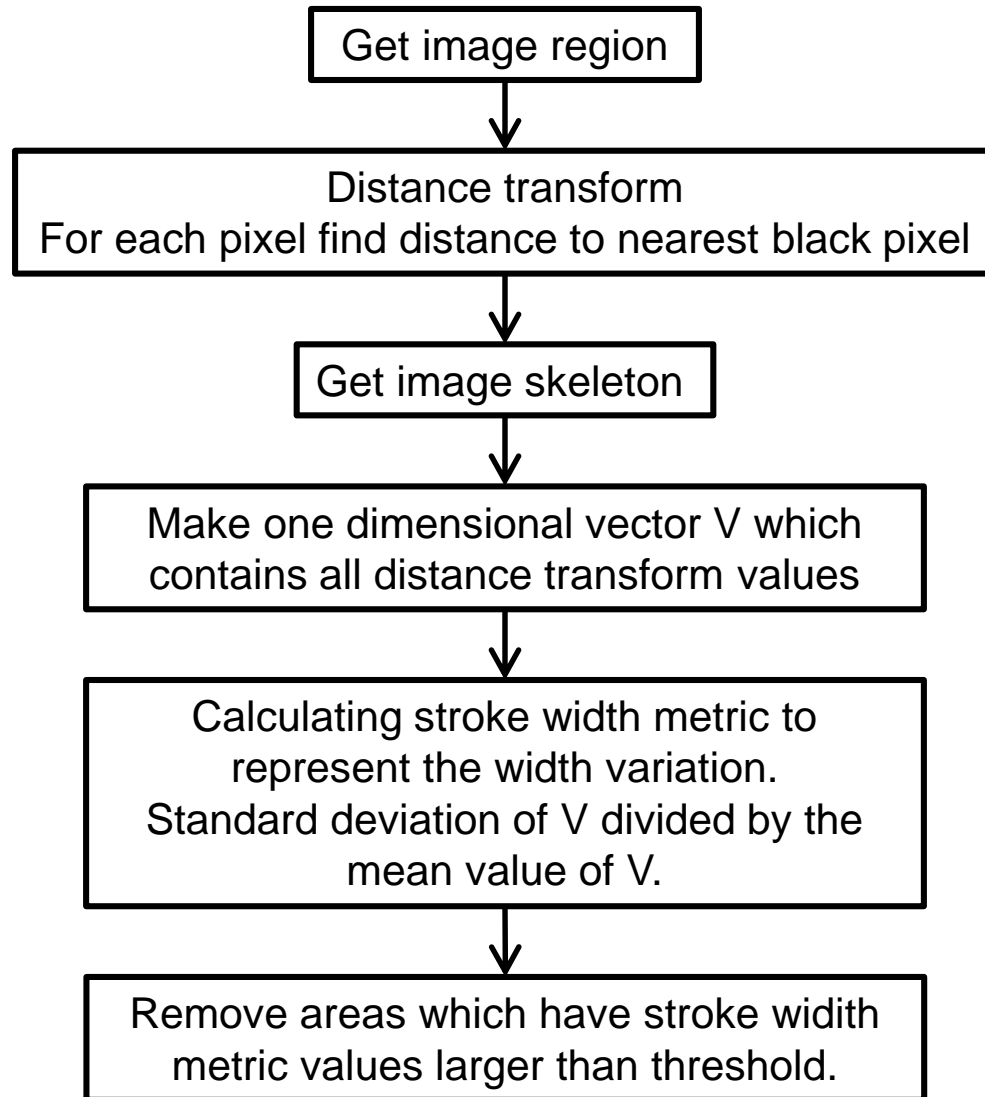
# Stroke Width Analysis 1

- One distinctive feature of text is that it consists of "strokes".
- Strokes have only a little variation in thickness.
- Other objects such as images can have lot of variation in thickness.
- The amount of variation can be used to distinguish text from other objects.



Dark blue represents thin stroke width and dark red thick strokes.

# Stroke Width Analysis 2

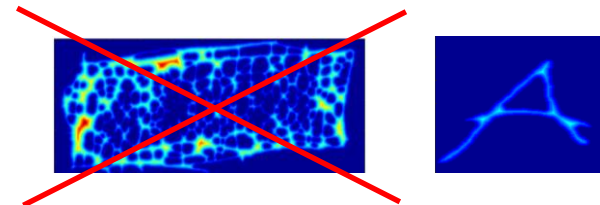


$$V = [6.403, 8.5440, 6.4031, 8.0623, 5.6569, 7.2801, \dots]$$

$$\frac{\sigma}{\bar{V}} = 0.2360$$

$\sigma$  = standard deviation  
of vector  $V$

$\bar{V}$  = mean of vector  $V$



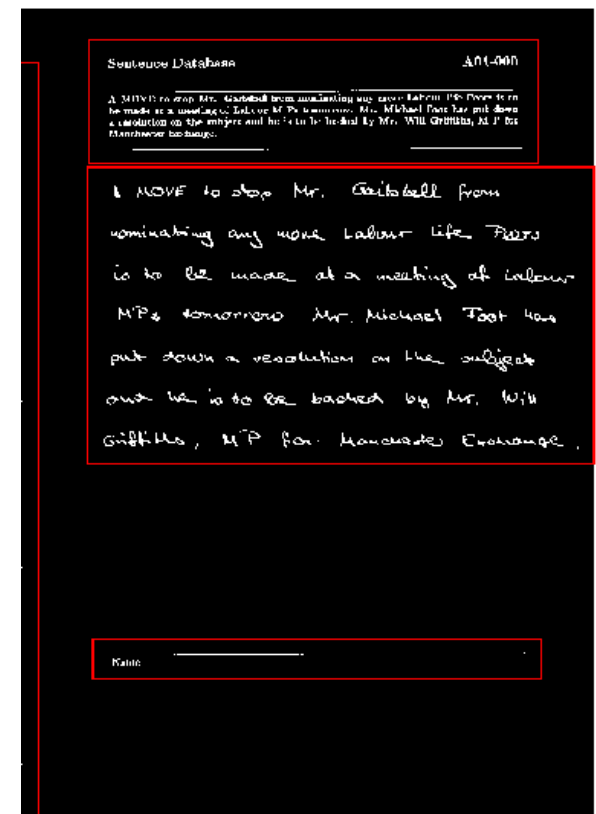
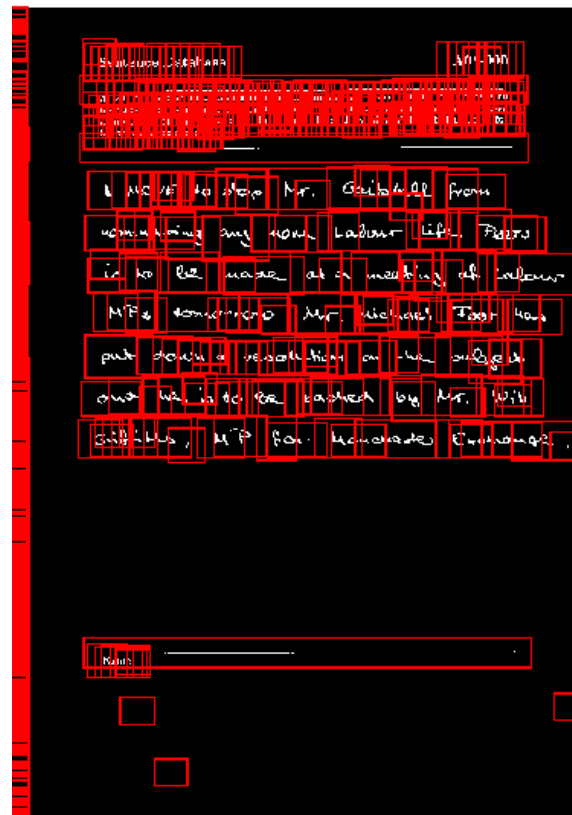
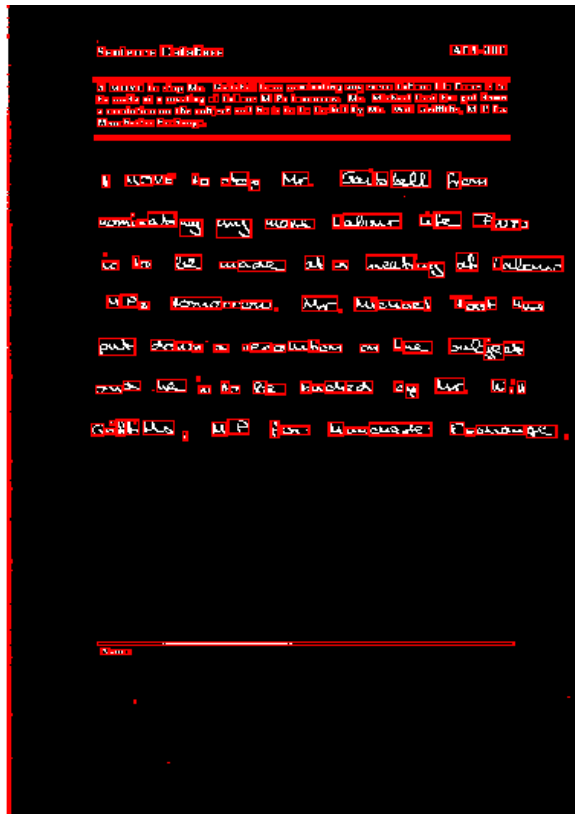
# Layout Analysis

---


- Analysing the image regions to find where the text is located and what kind of bodies of text it contains.
- Columns, rows, words.
- Proposed method to find the areas of interest is to draw bounding boxes over the text objects, expand them in all directions and combine overlapping boxes.
- Layout is saved hierarchically into areas of interest, rows and words.



# Bounding box expansion



In the last phase boxes that take only a small fraction of the total area are removed.

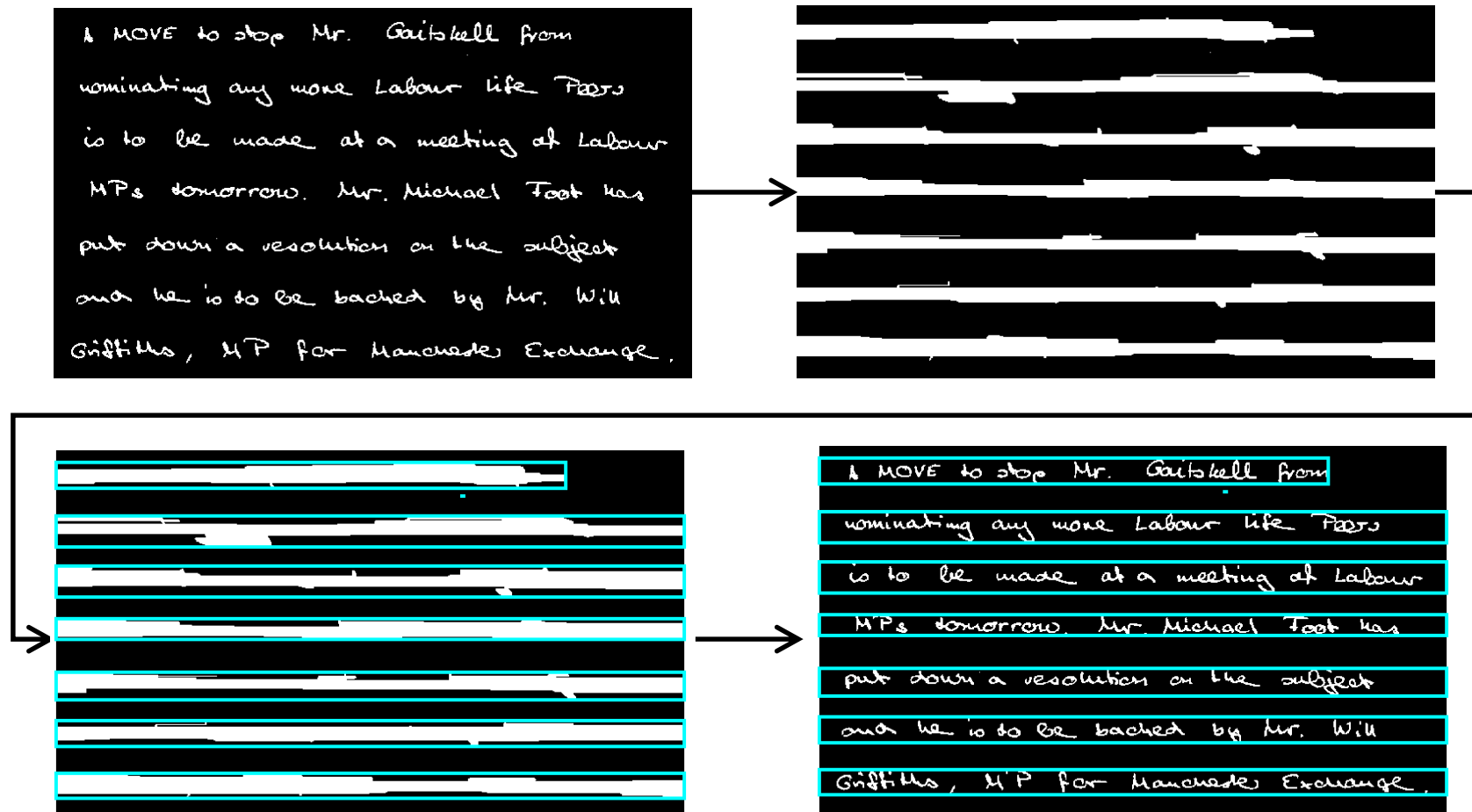
 = Area of interest box

# Layout Analysis 2

---

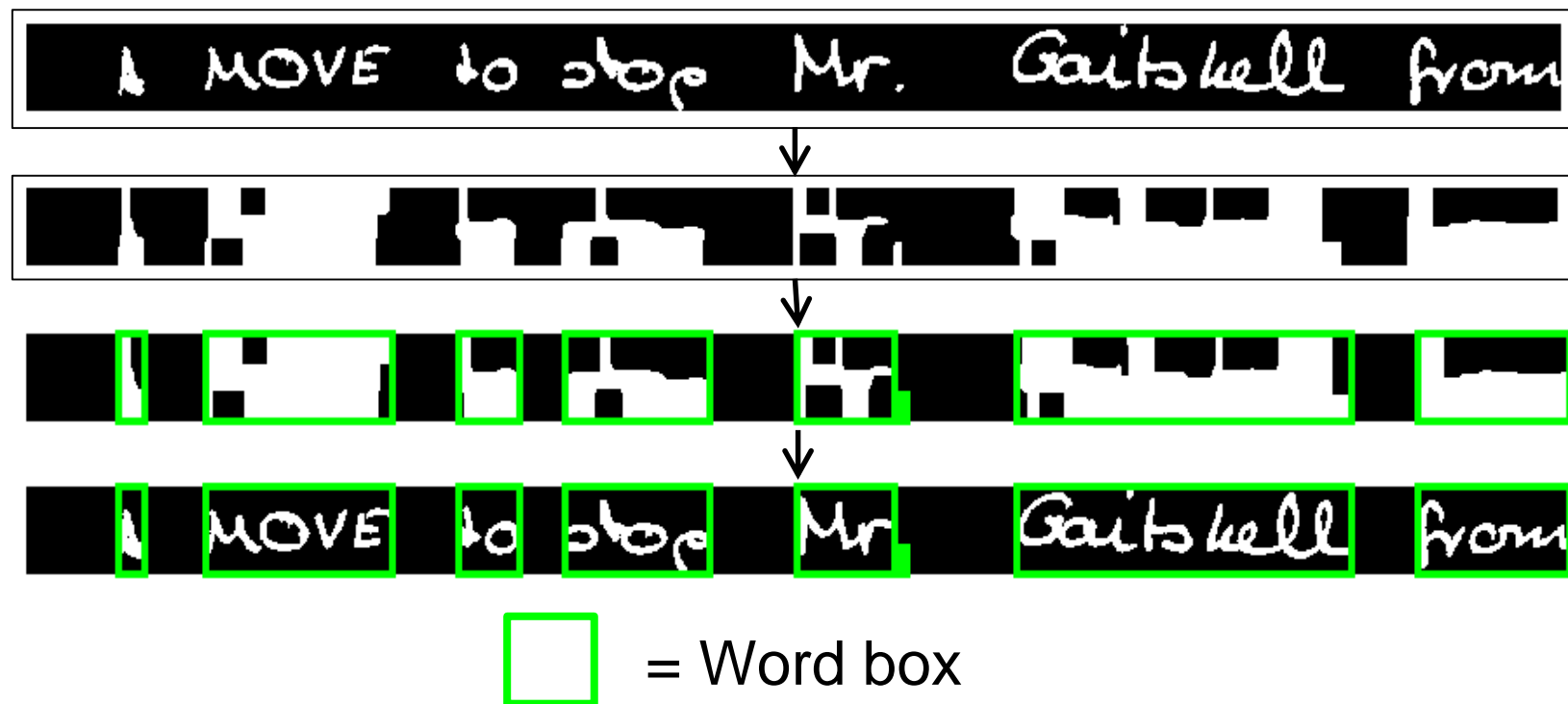
- To find rows of text the run length smearing algorithm is used.
- The RLSA finds rows of black pixels and changes them to white if their lengths are under given threshold.
- The bounding box is acquired for each object generated by RLSA. These bounding boxes represent the rows.
- The same method is used to find individual words within rows. For words the RLSA is executed also vertically. Smaller threshold values are used.

# RLSA for rows



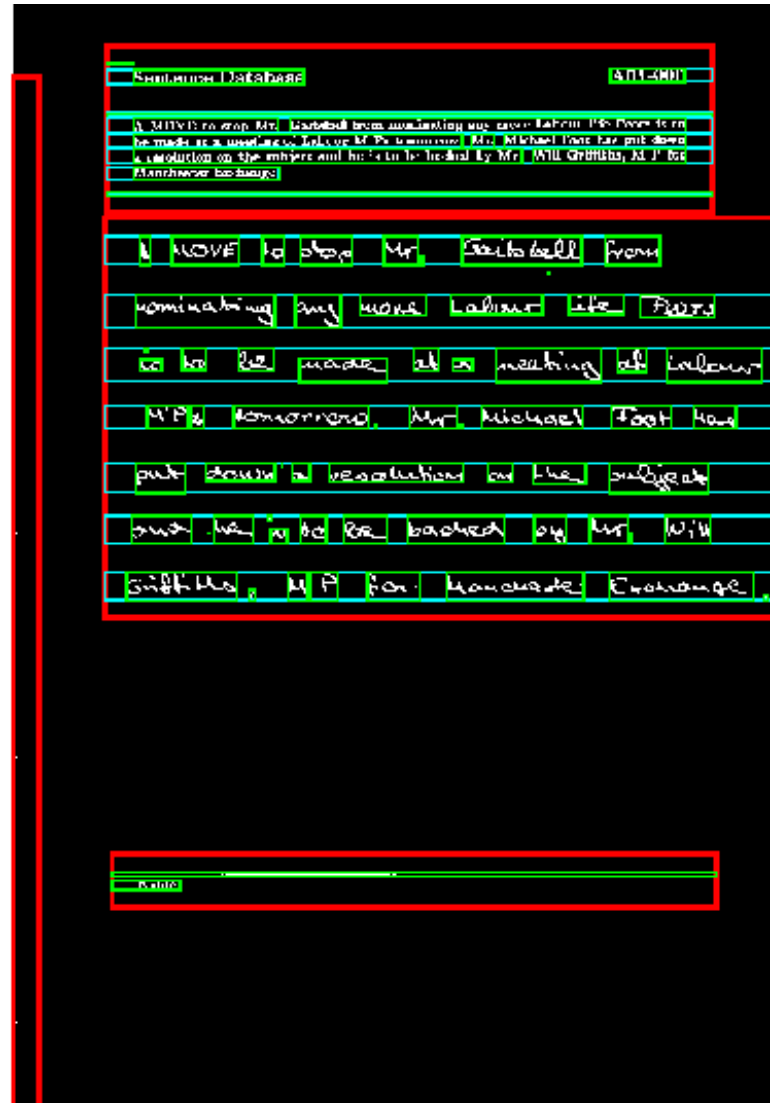
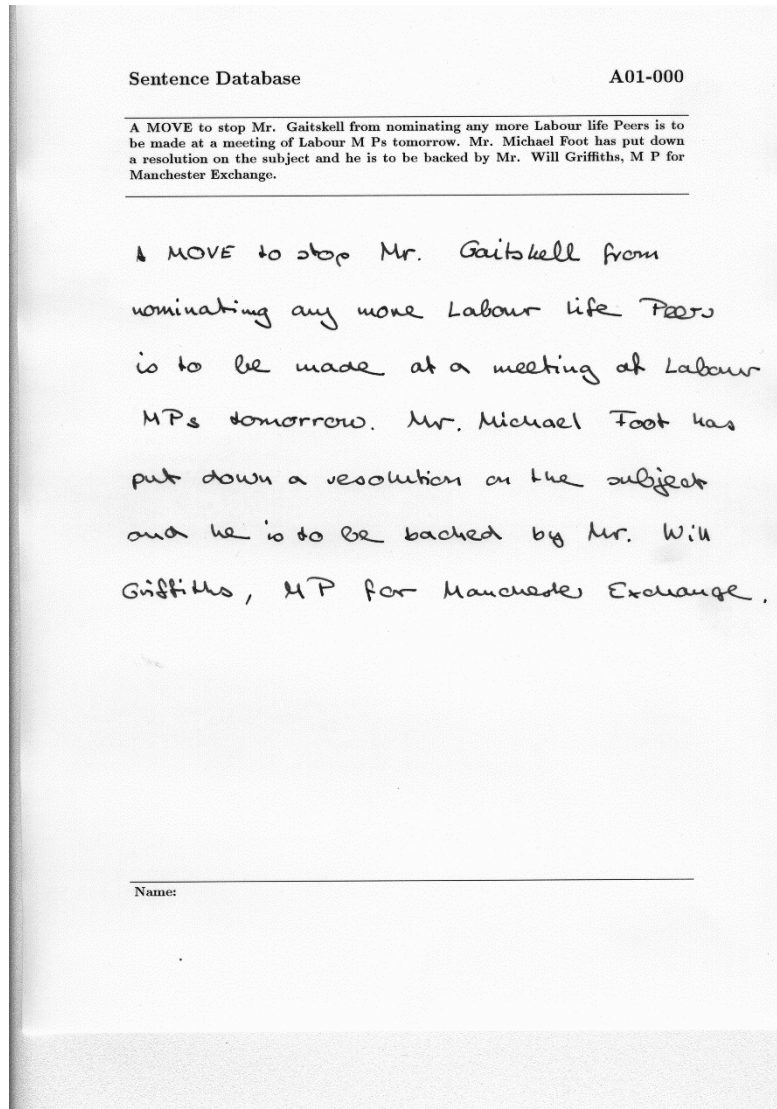
 = Row box




# RLSA for words



Same procedure but this time the algorithm is also run vertically to get dots or other broken characters into word.

# Full layout visualized



-  = Area of interest box
-  = Row box
-  = Word box

# Tests

---

- The system is still dependant on the parameter values.
- The tests goal is to find the optimal threshold values for IAM database images.
- First the optimal parameter values are found and then the accuracy is tested with these parameters.
- How to evaluate the accuracy of the system mathematically?
  - IAM database contains metadata about the number of rows and words.
  - Compare the real values to those detected by the system.

# Example entries in IAM handwriting database.

A MOVE to stop Mr. Gaitskell from nominating any more Labour life Peers is to be made at a meeting of Labour MPs tomorrow. Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, MP for Manchester Exchange.

Delegates from Mr. Kenneth Kaunda's United National Independence Party (280'000 members) and Mr. Harry Nkumbula's African National Congress (400'000) will meet in London today to discuss a common course of action. Sir Roy is violently opposed to Africans getting an elected majority in Northern Rhodesia, but the Colonial Secretary, Mr. Iain Macleod, is insisting on a policy of change.

Though they may gather some Left-wing support, a large majority of Labour MPs are likely to turn down the Foot-Griffiths resolution. Mr. Foot's line will be that as Labour MPs opposed the Government Bill which brought life peers into existence, they should not now put forward nominees. He believes that the House of Lords should be abolished and that Labour should not take any steps which would appear to "prop up" an out-dated institution.

Mr. Macleod thought the two Rhodesian parties had refused to attend the talks because Sir Roy had found messages sent from the Government were "unsatisfactory". African delegates to the talks yesterday called on Mr. Macmillan to cease his negotiations with Sir Roy's representative, Mr. Julius Greenfield. He was at Chequers last weekend. They said they regarded with "growing anger" the gross and unconstitutional interference by Sir Roy's Federal Government in the talks.

Only the handwritten part was used during the tests.

# Test procedure to find optimal parameters

5 different handwriting samples and their metadata



Iterate through a list of tested values



With each iteration run the preprocessing and layout analysis for each image



Get the number of detected rows and words



Get the percentage of difference compared to real values



Get the mean difference among the images for each tested parameter value



Visualize results and find the optimal value

Apply optimal value and choose next tested parameter



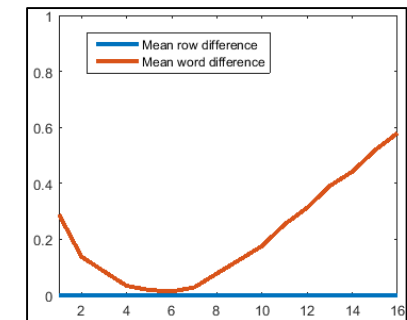
numberOfRows = 7  
numberOfWords = 52

```
testValues = [0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1];
```

```
for testedValue in testValues:  
    for image in images:  
        result = preprocess(image,testedValue);  
    end  
end
```

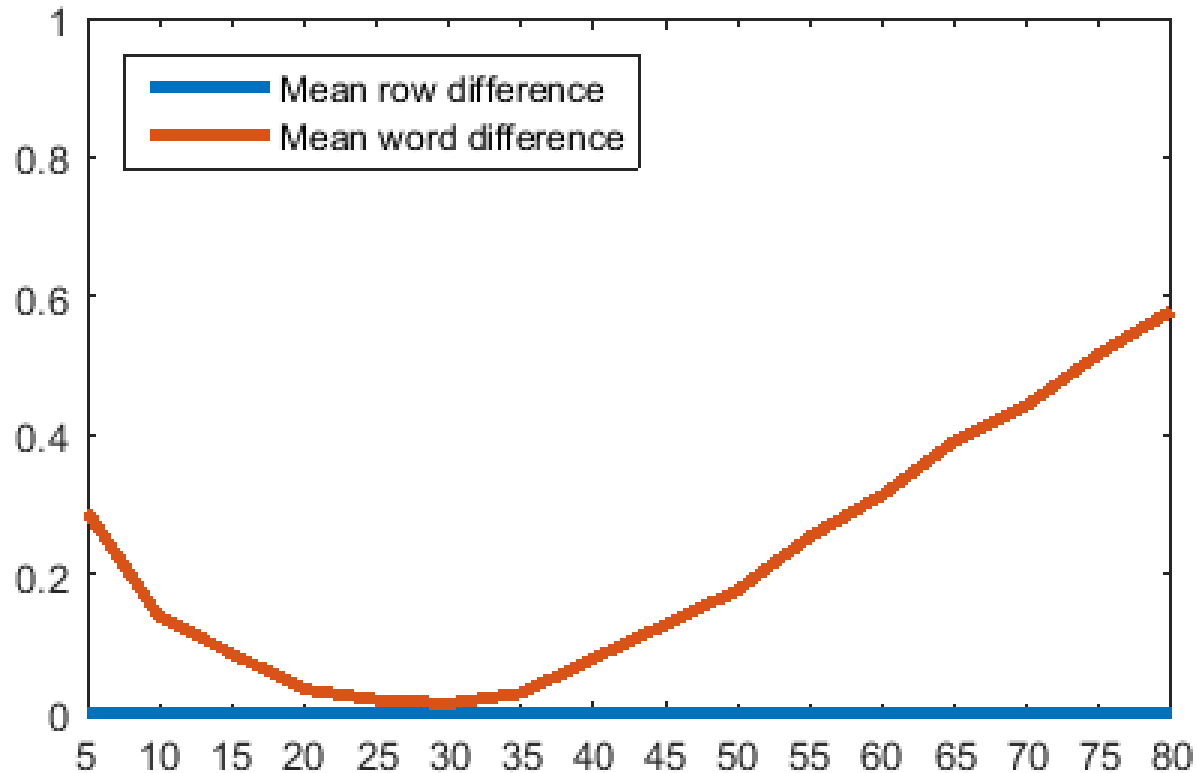
ImageName	RealRows	FoundRows	RealWords	FoundWords
'a01-000u'	7	8	52	61

RowDiffPercMean	WordDiffPercMean
1.5861	0.4841
0.9242	0.2555
0.7061	0.1922
0.5505	0.1650
0.4838	0.1475
0.5576	0.1591
0.4131	0.1300





# Example of test data



Result data for RLSA horizontal threshold test.  
Lowest difference is aquired with threshold 30.

# Test results

## Preprocessing parameters:

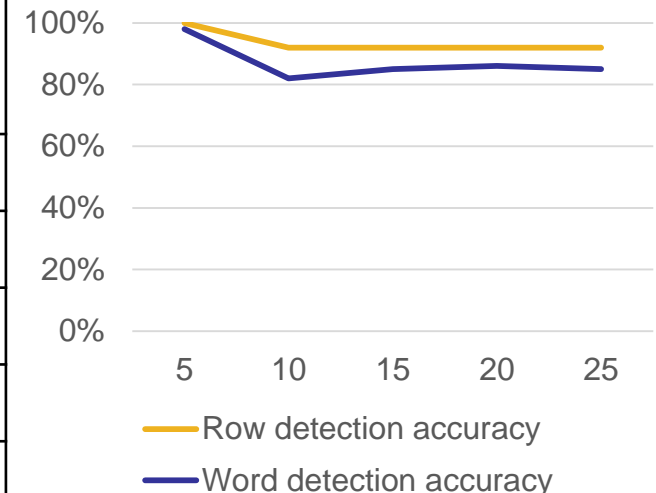
- wienerFilterSize = 15
- sauvolaNeighbourhoodSize = 180
- sauvolaThreshold = 0.3
- strokeWidthThreshold = 0.6

## Layout analysis parameters:

- aoiXExpansionAmount = 40
- aoiYExpansionAmount = 60
- rlsaRowThreshold = 300
- rlsaWordHorizontalThreshold = 30
- rlsaWordVerticalThreshold = 30

With above parameters the system achieved following performance:

Number of images	Row detection accuracy	Word detection accuracy
5	100%	98%
10	92%	82%
15	92%	85%
20	92%	86%
25	92%	85%



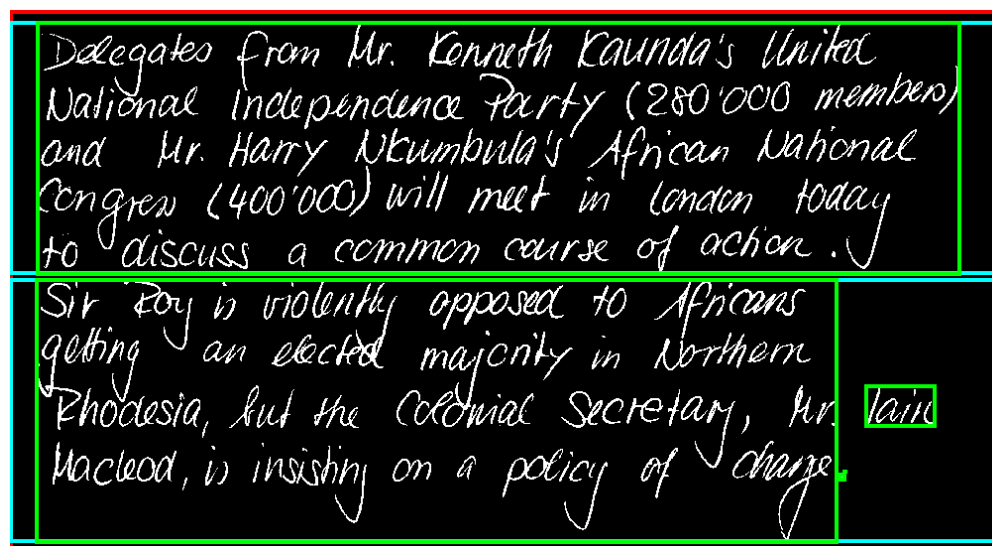
# Conclusions



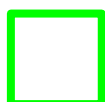
---

- The chosen methods proved to be useful in preprocessing and layout analysis.
- The tests proved that the chosen parameters work well for IAM database images.
- Word detection is slightly more sensitive about parameter changes than the row detection.

# Remaining problems

- The row and word detection isn't perfect
  - If two rows contain overlapping characters those rows are combined as one.
  - For the same reason multiple words can also be combined as one.



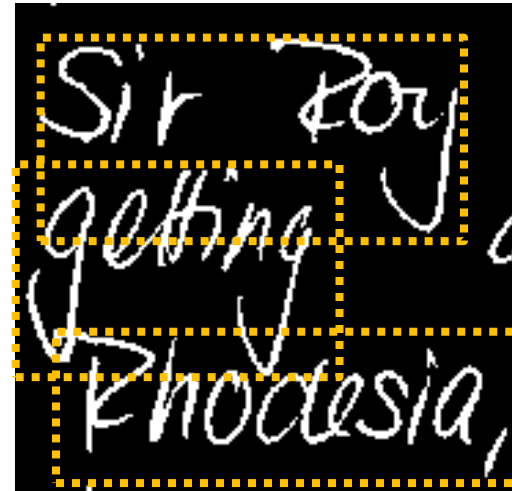
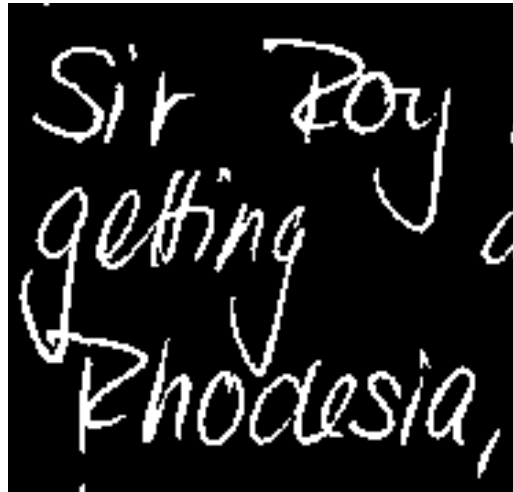
-  = Area of interest box
-  = Row box
-  = Word box


Result: 2 rows and 5 words

Real values: 9 rows and 68 words

# Remaining problems: Overlapping characters

---



 = Real word box (unrealized)

# Future work

---

- Solve the overlapping rows/words problem.
  - Some papers have been published regarding this problem. More research is needed.
- Automatic parameter choosing.
  - Use some mathematical properties such as object area or size to choose appropriate parameter value for some functions.
  - Average object area  $\rightarrow$  RLSA threshold?
- Continue to feature extraction phase.