Preprocessing 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

2015年12月25日

Overview

- Introduction
- Handwriting recognition
- Preprocessing
  - Binarization
  - Sauvola algorithm
  - Property analysis
- Results
- Conclusions
- Future work

Introduction

- Matlab was used for implementation with image processing toolbox.
- This research concentractes on writing done with English alphabet.

Handwriting recognition 1

- Handwriting recognition (HWR) is the process of extracting text in digital form from handwritten images or input devices.
- Divided into offline and online handwriting recognition
  - Offline HWR recognition takes static image and runs the recognition process to it.
  - Online HWR analyzes characters on input. For example with tablet pc.
- Used methods for handwriting recognition include Optical Character Recognition (OCR) or Intelligent Word Recognition (IWR).

Handwriting recognition 2

- OCR is used to recognize individual characters while IWR recognizes whole words.
- Offline recognition process can be divided into three main phases:
  - Preprocessing
  - Feature extraction
  - Classification
- The preprocessing and feature extraction phases are similar in both OCR and IWR.
- Each phase reduces the amount of data to be processed.

Handwriting recognition 3

- Preprocessing
  - Includes noise removal, binarization and segmentation.
  - Image is enhanced for feature extraction phase and the detected characters are segmented from the original image.
- Feature Extraction
  - Shape describing features are extracted from previously acquired objects.
  - Histogram of ordered gradients, horizontal and vertical histograms, topological features (loops, junctions), etc.

Handwriting recognition 4

- Classification
  - Extracted features are used in machine learning algorithms to create the feature vector.
  - The inputs are classified according to this vector.
  - For example simple k-nearest neighbors algorithm can be applied to find the correct category for the input.
- For now the research has concentrated on preprocessing.

Preprocessing 1

- During this research following preprocessing methods were considered for offline HWR:
  - Image aquisition
  - Noise removal
  - Binarization
  - Object property analysis
  - Object extraction

Preprocessing 2

- Image is converted to grayscale.
- Noise is removed using 6x6 adaptive wiener filter
  - Adapts to the variation. More smoothing if variance is large.
  - 6x6 neighbourhood optimal for most cases. Larger filter sizes resulted in exessive blur.
- Histogram equalization is not applied because in most cases it increased the visibility of irrelevant objects such as paper texture and noise.

Binarization

- Binarization is one of the most important parts of the preprocessing.
- Image may have uneven lighting resulting in visible shadows or gloss.
- Sauvola algorithm was designed for document binarization purpose.
- Sauvola algorithm resulted in best results with quick comparison with other algorithms.

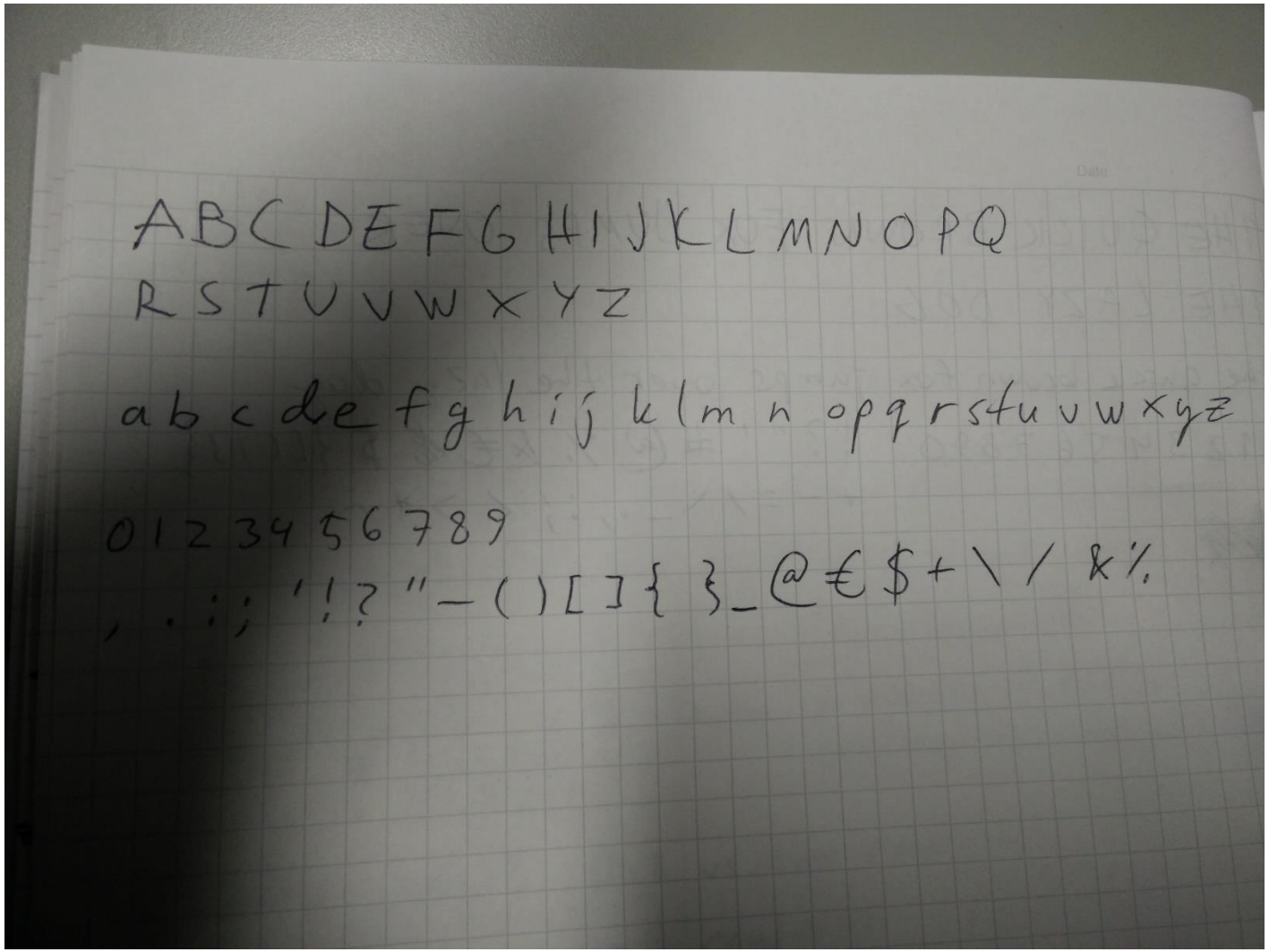
Sauvola algorithm

- Uses adaptive thresholding to binarize document images with uneven lightning or texture.
- Enhanced version of Nilback binarization algorithm.
- Can apply different algorithms to textual and non-textual areas of the image. (Nilback can only detect varying lightning)

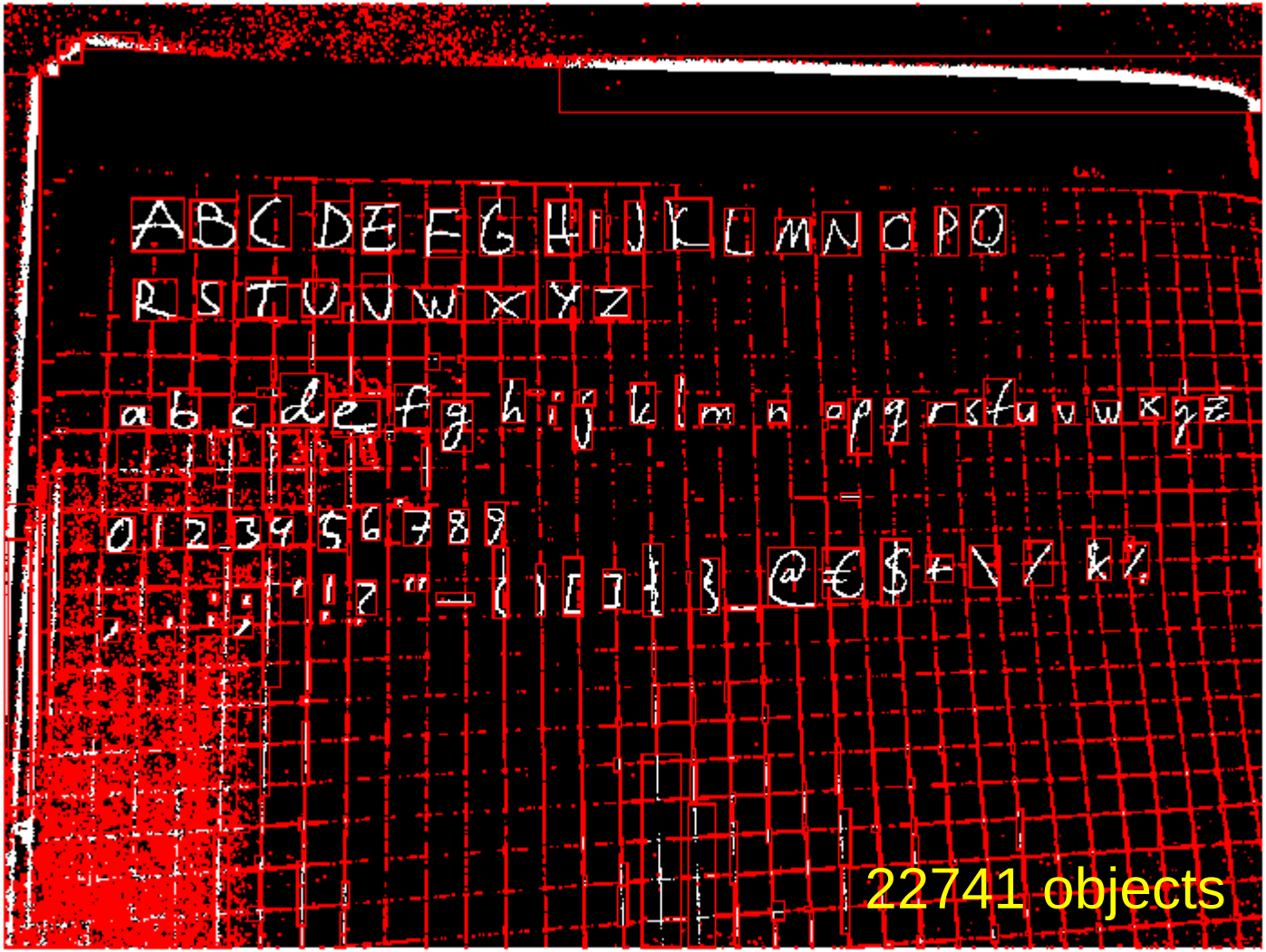
Sauvola algorithm

- The algorithm takes two arguments:
  - Window size w and user defined parameter k
- $T = m \times \left[1 + k \times \left(\frac{s}{R} - 1\right)\right]$
  - T new threshold
  - m mean of window size w
  - k user defined parameter ”sensitivity”
  - s local standard deviation in window w
  - R dynamic range of standard deviation (128 with 8-bit gray level images)

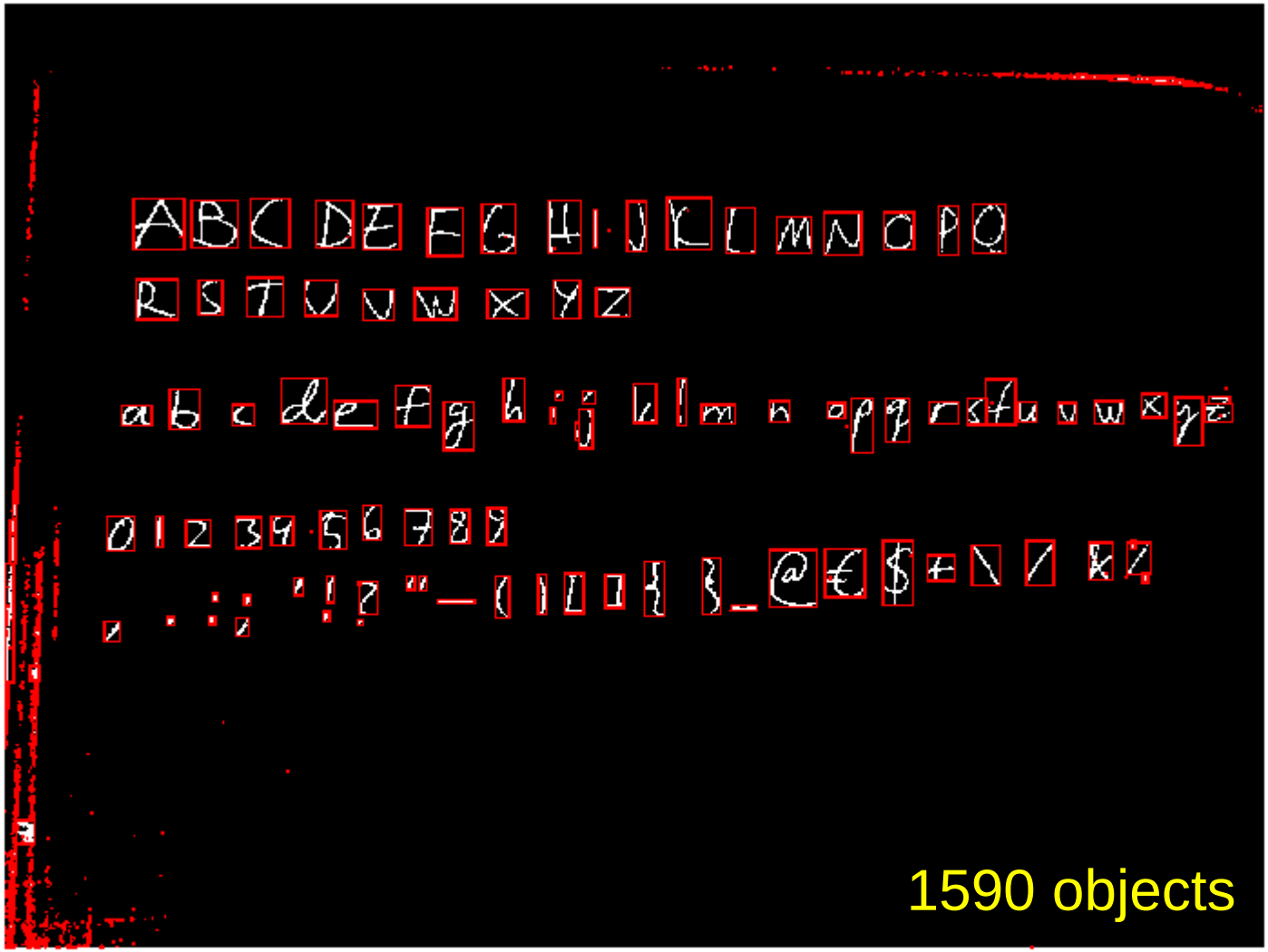
Original image with shadow. 95 individual objects.



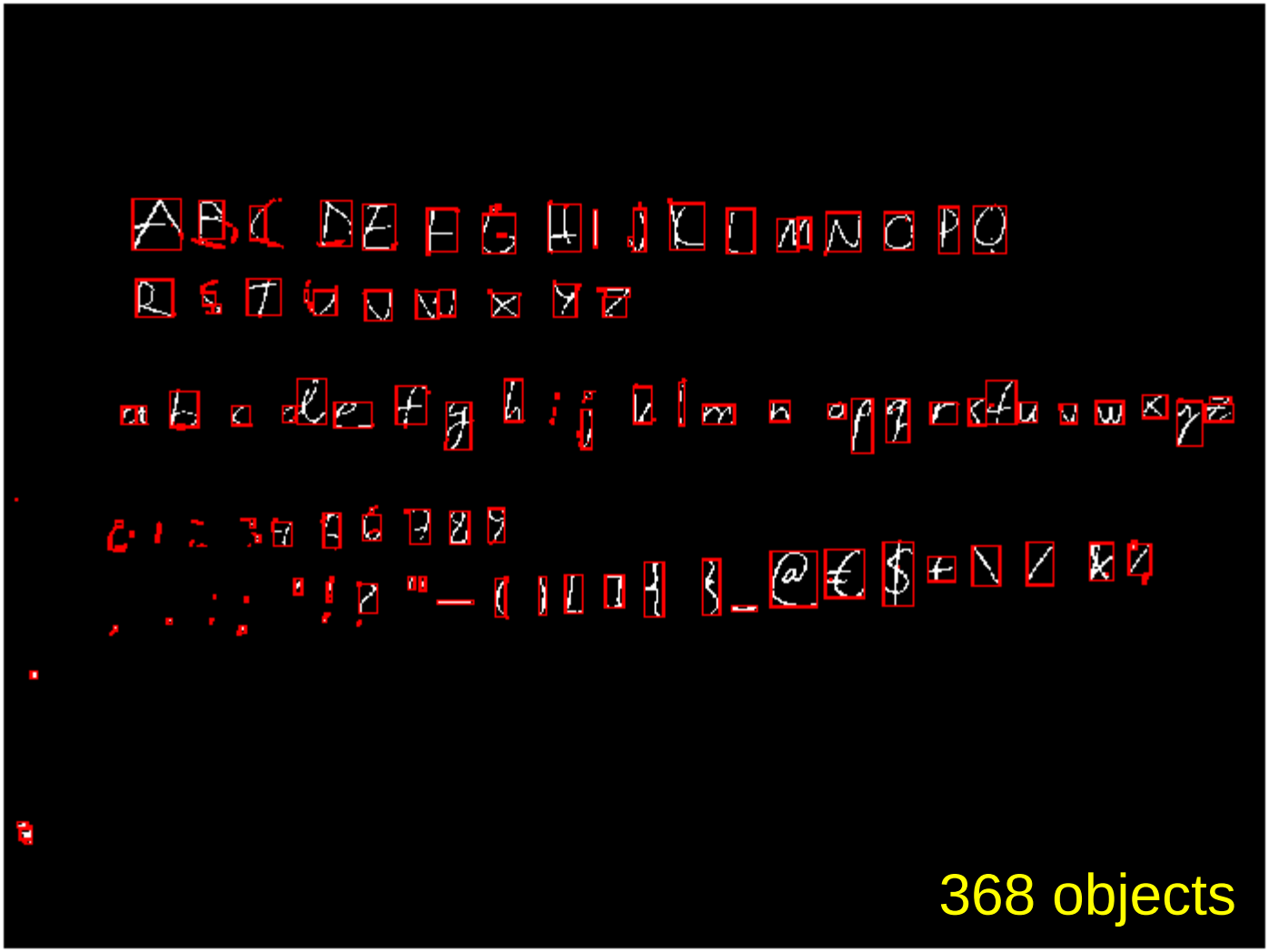
No noise removal. Sauvola with k=0.1 w=100x100



No noise removal. Sauvola with k=0.4



No noise removal. Sauvola with k=0.9



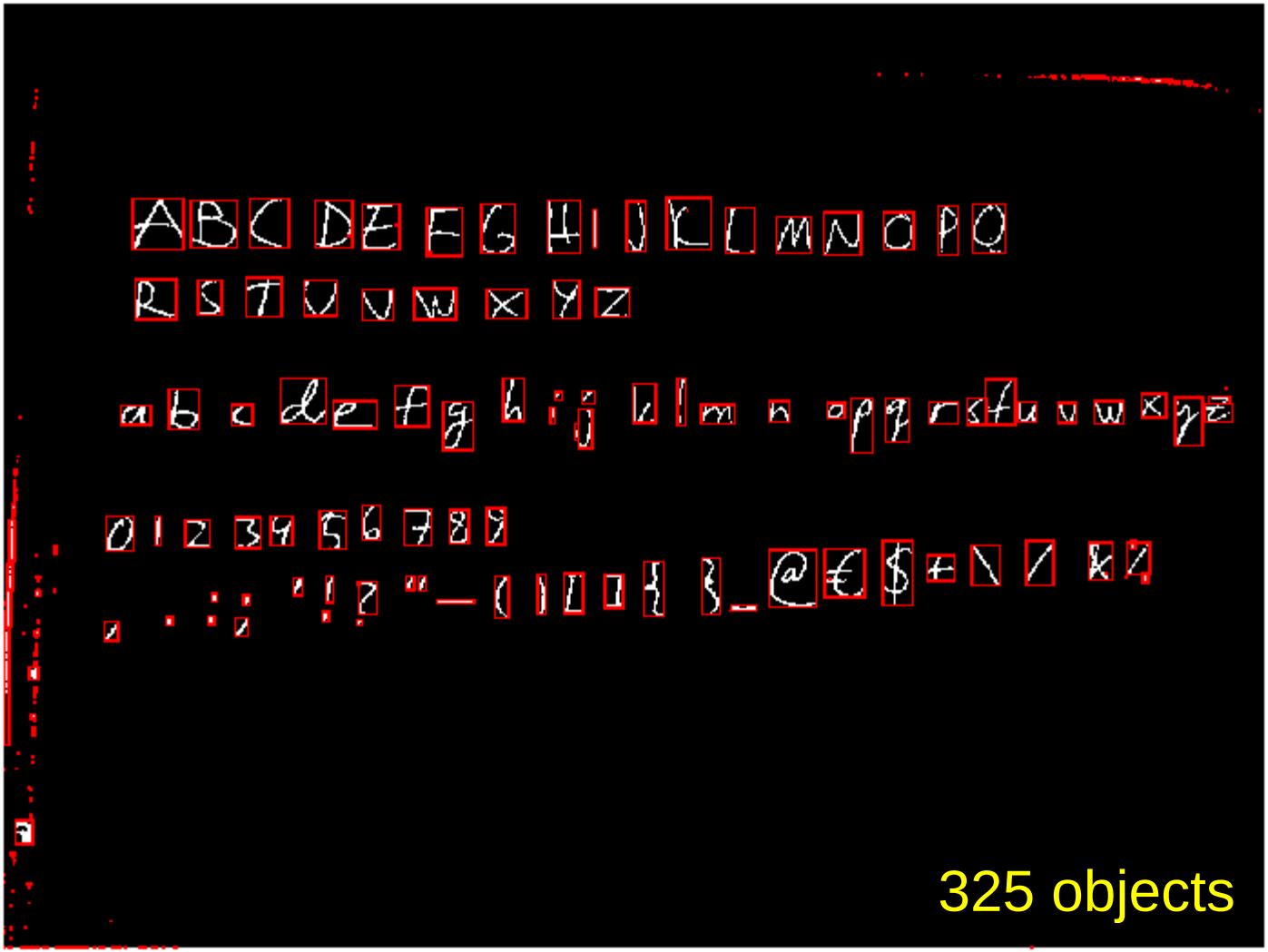
Sauvola algorithm

- For this case the threshold 0.4 works best.
  - Lower thresholds resulted in more noise. Higher thresholds result in more broken objects.
- Window size 100x100 worked best for most cases.
  - With lower window sizes more broken objects and higher window sizes resulted in more noise.

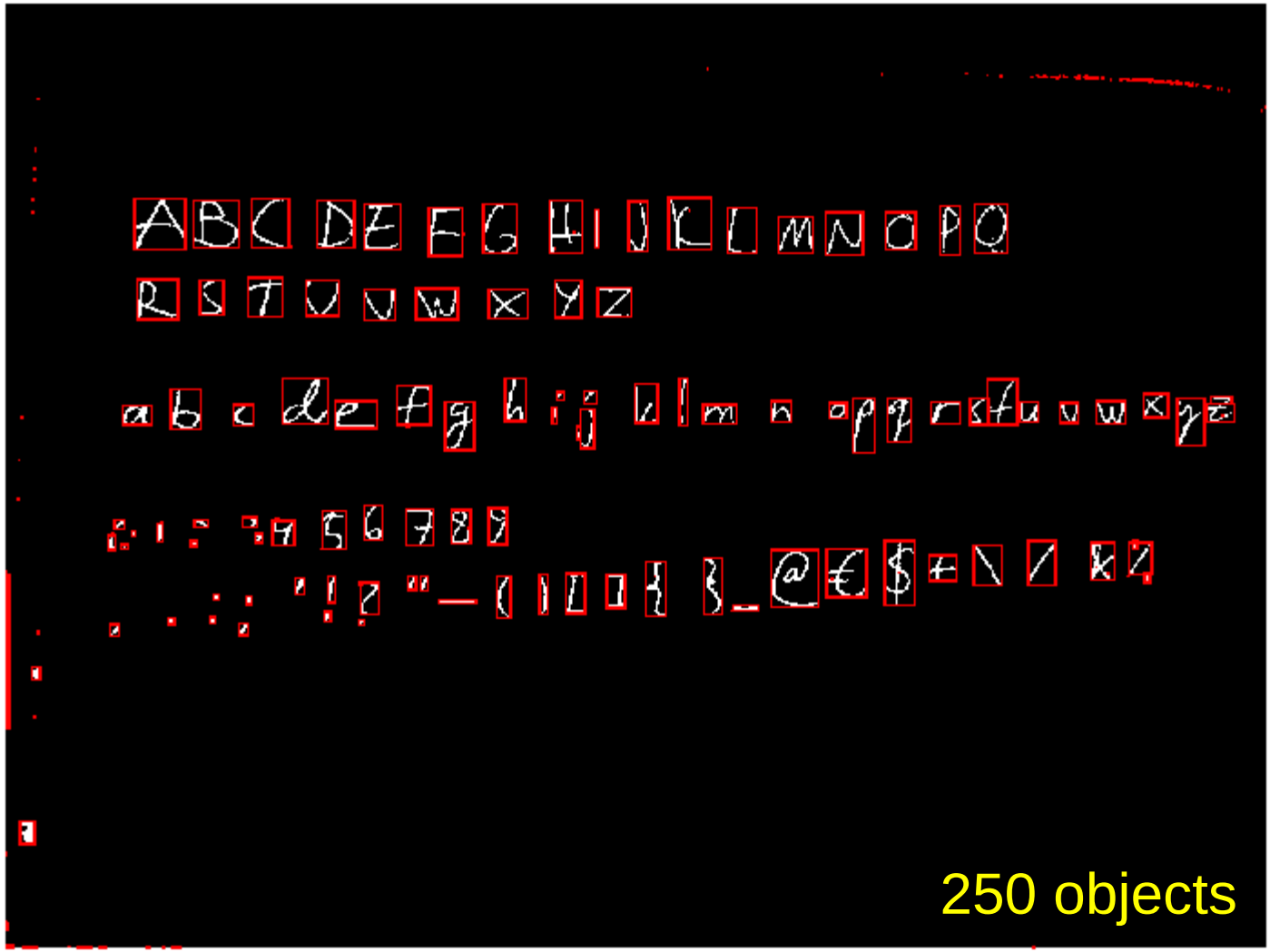
Noise removal

- Adaptive Wiener filter was used.
- Filter window size affected results
  - Chosen window size 6. Size 3 had slightly more noise and larger window sizes broke the objects.

Wiener filter size 6 (k=0.4 w=100)



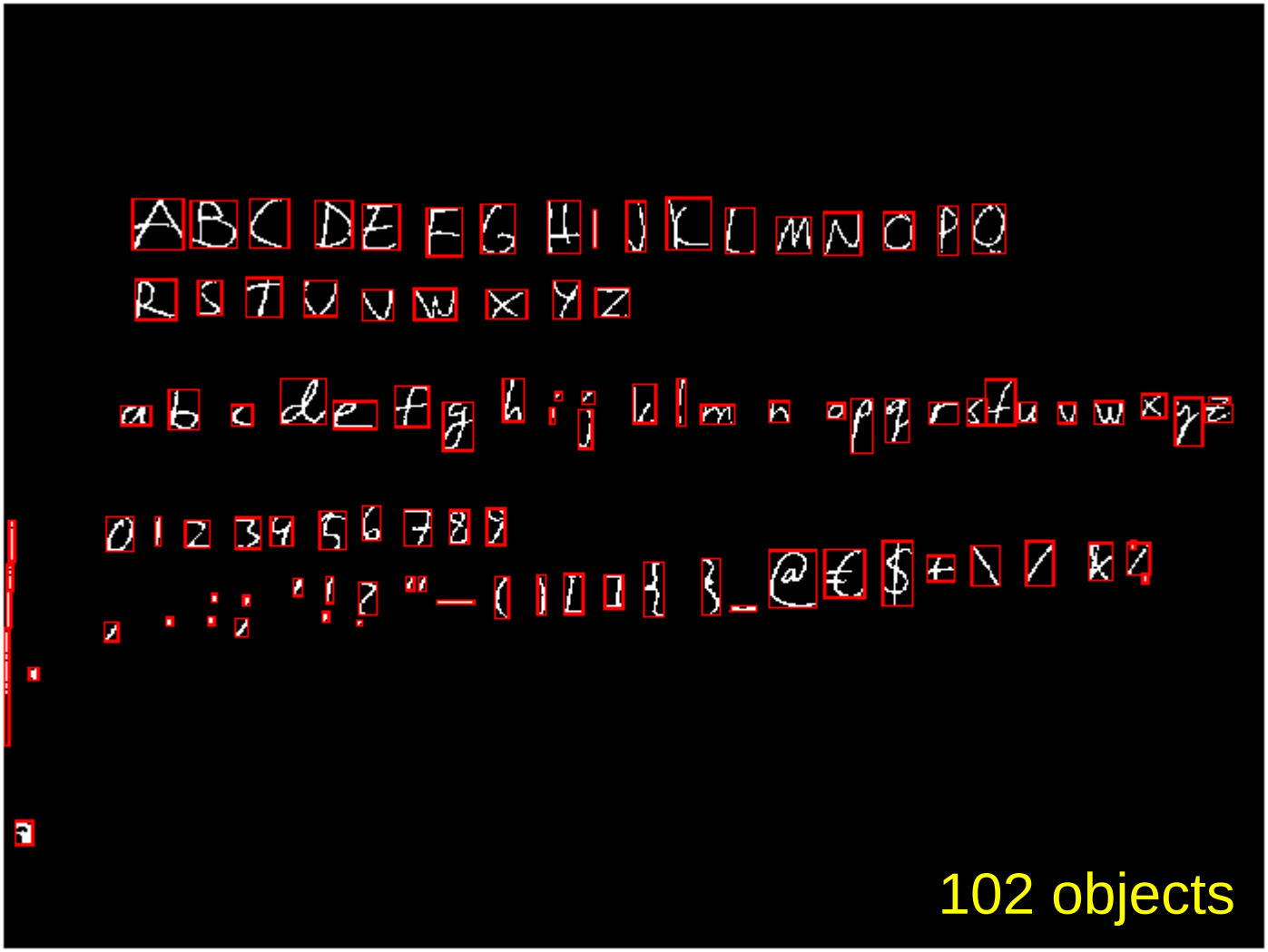
Wiener filter size 20



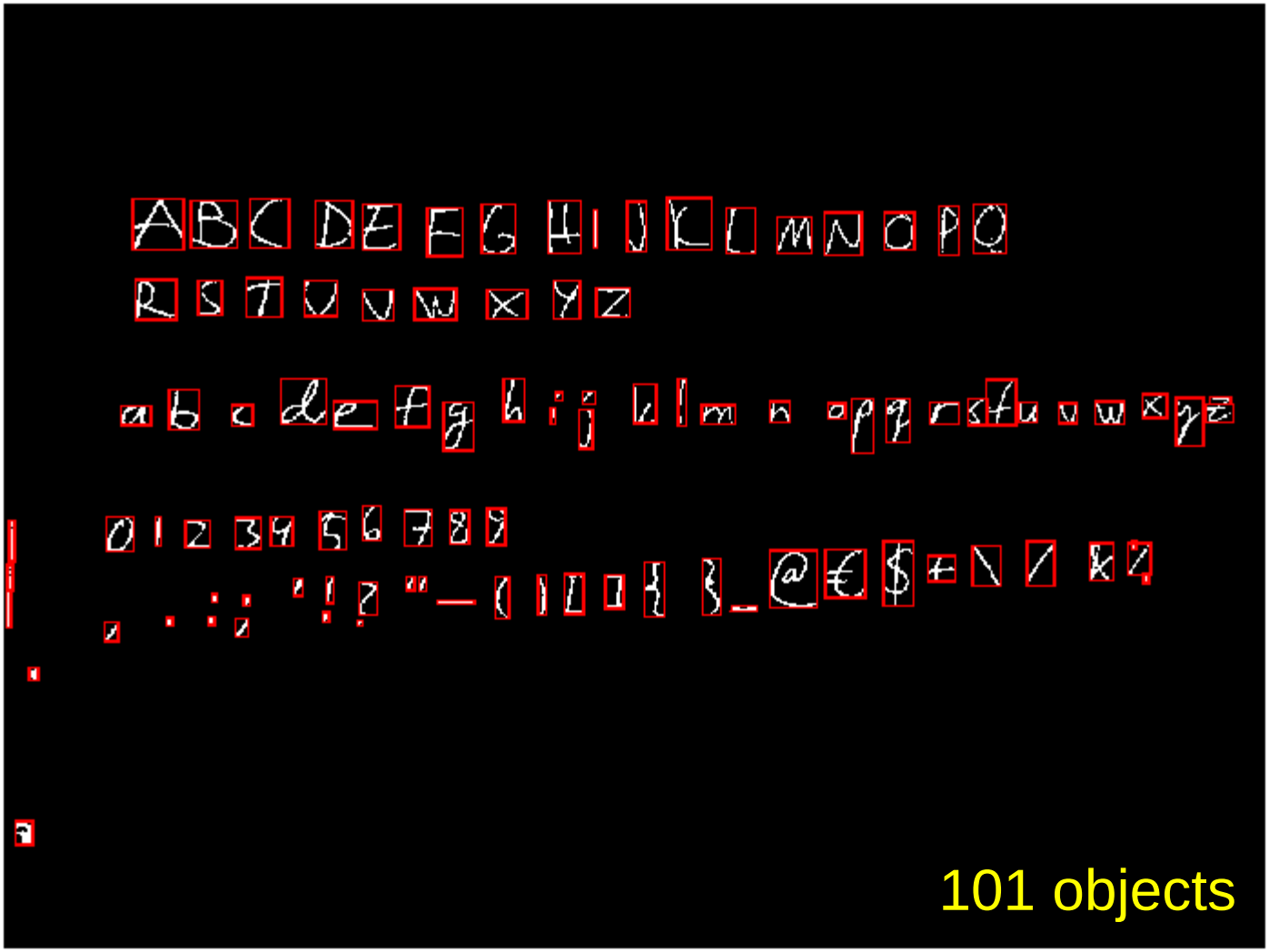
Object property analysis

- Usually some irrelevant objects still remain after binarization.
- Some object properties are analyzed and the objects are removed which have feature values outside the preferred values.
- Example object properties include area, major axis length and Euler number (Number of objects in the region minus the number of holes in those objects).
- When only relevant objects remain they can be easily extracted to smaller sub-images for further processing.

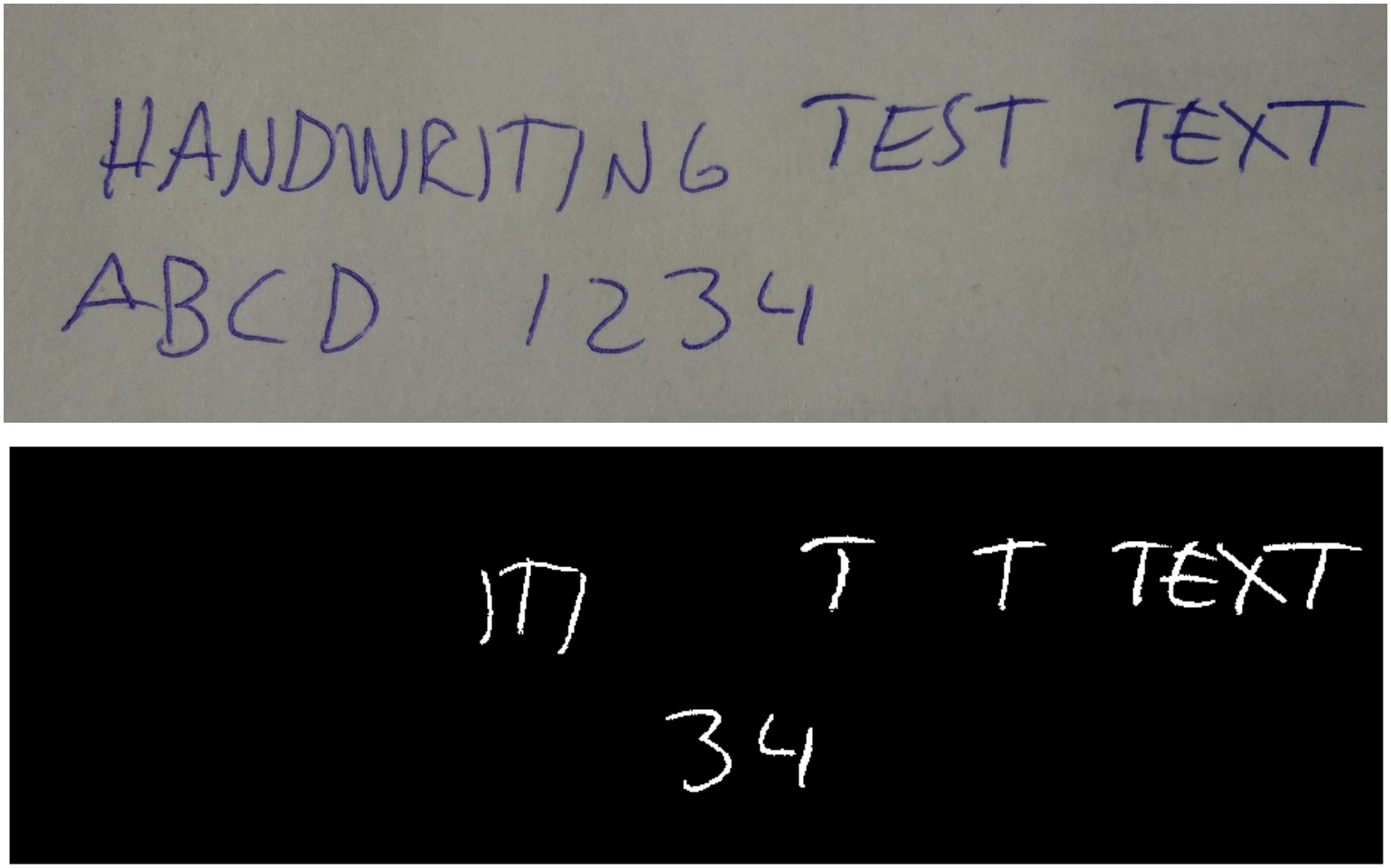
Objects smaller than 209 pixels removed.



Objects with Euler number lower than -4 removed.



Errors with different image and same parameters



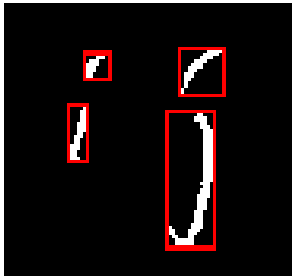
(Lighter pen caused more holes in objects during binarization which then caused Euler number to be lower than expected.)

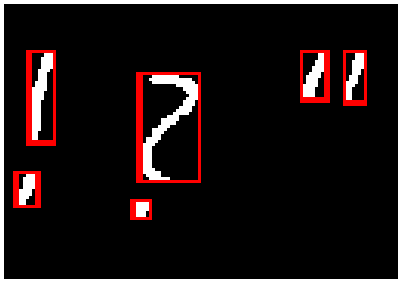
Conclusions

- Chosen methods are useful in preprocessing.
- Chosen arguments will work only for the specific case.
- Problems:
  - Human handwriting can vary a lot even with same person.
  - Different size and thickness with characters.
  - Different color pens or pencils.
  - Differing image resolutions.
- Solutions:
  - Keep resolution as constant.
  - Try to find optimal stroke width without prior information.
  - Apply morphological closing to remove small holes.

Future work 1

- Enhance the preprocessing making it less dependant on arguments.
- Extract and analyze more object properties to remove the irrelevant objects.
- Combine letters and characters containing separate elements. (i j ! ? ” = ; : %)
  - Useful feature considering Hiragana and Katakana characters. (ひらがな)

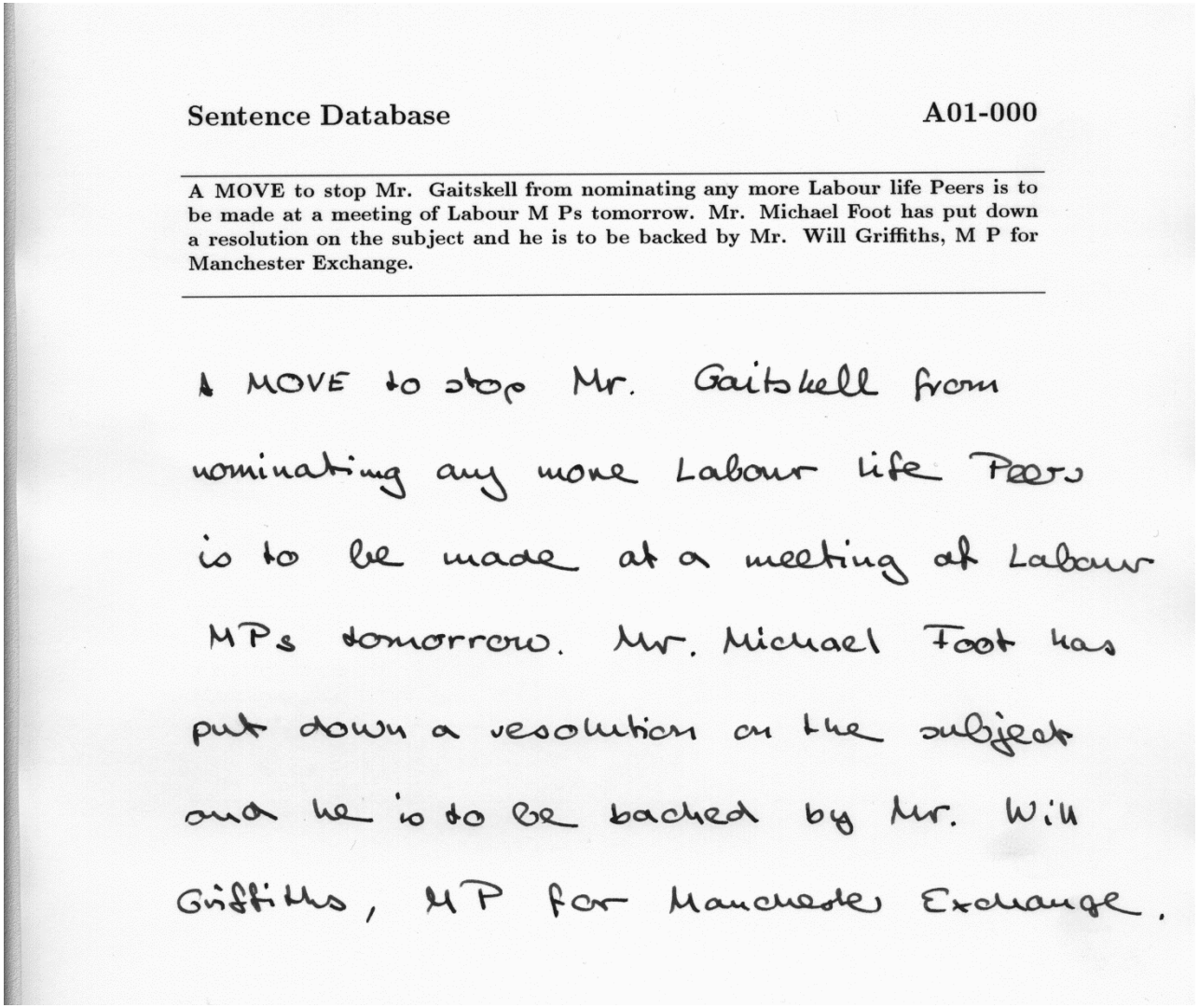




Future work 2

- Layout analysis. Detect words, rows and columns of text.
- Feature extraction:
  - Vertical and horizontal histograms.
  - Histogram of ordered gradients.
  - Topological features such as endpoints, loops and junctions.
  - Etc.
- Classification with k-nearest neighbors algorithm.
- Use IAM database for large scale tests.

Example entry in IAM handwriting database.



Sentence Database A01-000

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

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.

Questions?