Page Segmentation Algorithm

Sergey Mikhno

April 2022

1 Problem Description

Sometimes we have books which are scanned or photographed. To be able to read such books on a mobile device we have to make pages readable. We have to find all the printed symbols and separate them from images. After that the page is redrawn in a larger scale and all the symbols are reflowed.

2 Algorithm Description

A scanned page is translated into an array of pixel intesities. For the reflowed page we create a new array. Below we have the steps necessary to reflow the page image.

1. Open an image file in grayscale mode.
2. Threshold the image with OTSU and BINARY_INV.
3. Find all components containing connected non-zero pixels. See Fig. 1
4. For every component find bounding rectangles. See Fig. 2
5. Eliminate all rectangles contained inside others. See Fig. 3
6. Join all intersecting rectangles. See Fig 4
7. Make a histogram of rectangle heights.
8. The height with the highest frequency is the text height.
9. Mark or remove all components with the hight or width <than 5× most frequent text hight.
10. For every rectangle find a neighboring one to the right, it is a nearest rectangle intersecting or being inside of the interval of heights [y, y + height], where y is the ordinate coordinate of the component's left upper corner.

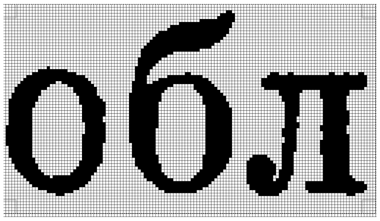


Figure 1: Connected components of pixels

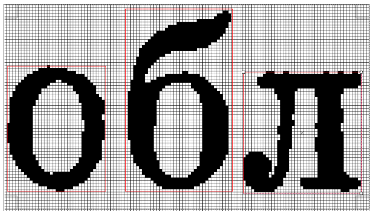


Figure 2: Components with bounding rectangles

11. Create a graph, add an edge between all components and their immediate right neighbors. See Fig 5

12. Find connected components in that graph.

13. Join intersecting ones.

14. Those components are text lines.

15. Sort the text lines and symbols inside them.

16. Calculate the average area of the connected pixel components for every text line, let us call it H_a. All the intersymbol gaps bigger than $\frac{1}{2} \times H_a$ will be the interword gaps. Use the interword gaps to split the text line into words.

17. Calculate the baseline height for every text line using the histogram of lower y coordinates. The most often occurring height is the baseline. Calculate the baseline shift for every symbol.

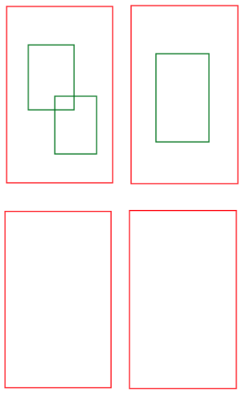


Figure 3: Inner rectangles should be removed, green rectangles are removed as seen in two lower rectangles

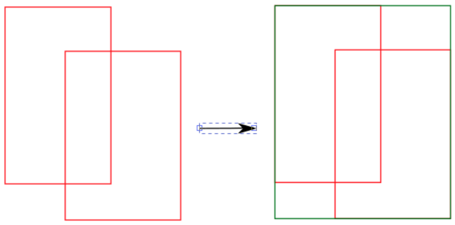


Figure 4: Two red rectangles on the left become one green rectangle on the right

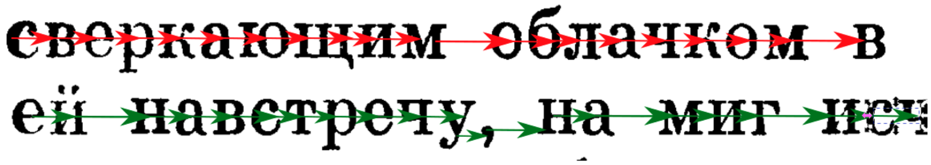
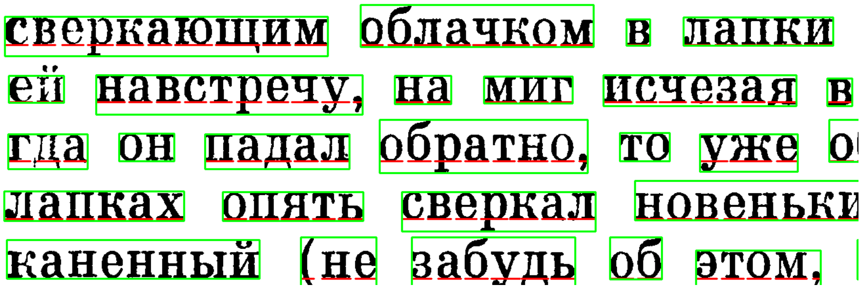


Figure 5: Connect every letter with it's right immediate neighbor



сверкающим облачком в лапки
ей навстречу, на миг исчезая в
гда он падал обратно, то уже о
лапках опять сверкал новеньки
каненный (не забудь об этом,

Figure 6: Baseline detection