

Extracting the Major Form Body Segment from Unconstrained Document Images

J. Bunch, D. Curtis, C. Jones, J. Tse, E.A. Yfantis

Department Computer Science
University of Nevada Las Vegas
Las Vegas, Nevada, United States of America

Abstract *In the analysis of any unconstrained document image it is necessary to first decide what the main area of interest is. Previous work has been done on simply removing the margins. In this paper we detail a more intensive approach to finding the main area of interest, what we call the Major Form Body Segment. We give a novel approach for determining which connected components are or are not part of the Major Form Body Segment.*

Keywords: feature extraction, image processing, pattern recognition, document image analysis

1 Introduction

When a paper document is converted to a digital image, there are affine transformation errors (translation, rotation, scaling), and introduction of noise in the form of speckles that can occur that cause the digital representation of the image to be different than the original paper document. This difference resulting from a scanning operation can be characterized by a slight or excessive rotation off the axis of the original document (skew), document margin cutoff or noise within the margin segments. A scanning error can be the cause of any or all three of these errors. Over a set of document images of the same type, documents can take on different characteristics based on these errors. Therefore, in order to perform consistent feature extraction over a set of document images, whereby, the feature extraction results can be comparable from two document images, an algorithm must be developed that can perform a structural normalization of the document image.

The major form body segmentation algorithm performs this required normalization by extracting from a document image, the major form body segment (MFBS). The MFBS has no margins and no

skew. This eliminates the contribution of noise that commonly occurs within the margins of document images. By simply eliminating the margins, the resulting document image can be comparable to other documents. This process is characterized by three general steps. The first step is applying a series of simple filters to eliminate components of no significance to the image (such as small speckles). The second step is a hierarchal construction algorithm that sequentially combines component rectangles based on several conditional properties, into body components. The final step is using these body components to construct the MFBS [1], [3].

2 Preprocessing: Noise Removal

The MFBS detection algorithm presented paper corrects for extraneous noise and other artifacts that come from the scanning process. However, if at all possible getting rid of these artifacts before the MFPS detection algorithm begins is desirable. For this reason we run a noise removal algorithm prior to the MFBS algorithm.

The noise removal algorithm we implement is based on an erosion based filter. The erosion algorithm we use looks at every black pixel's neighbors, if the pixel is completely isolated it is flipped to a white pixel. Using this simple erosion technique we can remove all completely disconnected speckled noise [2]. (See Figure 1, 2, 3)

3 Major Form Body Segment Detection

The MFBS detection algorithm works in two main phases, connected component filtering and hierarchal merging. The process starts with a list of con-

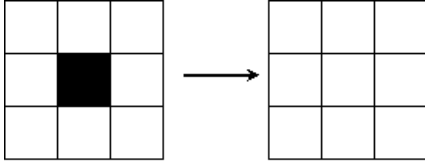


Figure 1: Pixel elimination based on the erosion algorithm.

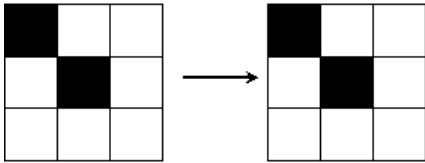


Figure 2: The erosion algorithm does not eliminate the pixel.

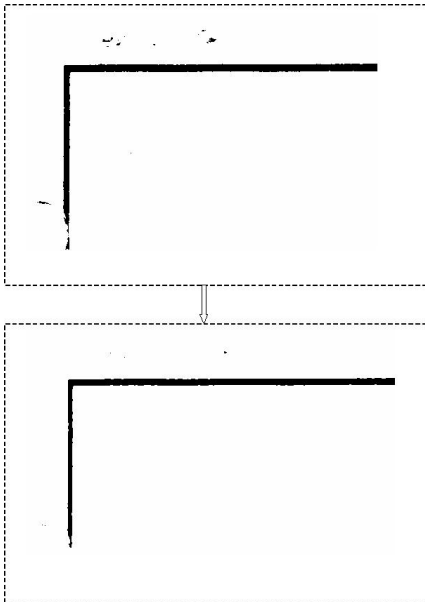


Figure 3: Example Erosion Techniques on Actual Image.

nected components and with the use of connected component filtering we get rid of connected components that obviously are not part of the MFBS. Following that we take the list of connected components and merge them in a way that will always eventually result in one segment, which becomes the MFBS.

3.1 Basic component filtering application

The first step to eliminating the margins and detecting the MFBS is to perform a component scan whereby the set C contains all the connected components in the image [4].

Then, several general filtering operations are applied. The first operation is applied to each component of C to form the set F such that

$$F := \{\forall c_i \in C \mid c_i \text{ width} > \Theta_1 \wedge c_i \text{ height} > \Theta_2\}$$

where c is an individual component, $1 \leq i \leq |C|$, $c_i \text{ width}$ and $c_i \text{ height}$ are the width and height of that component, and Θ_1 and Θ_2 are thresholds for the width and height respectively. This operation makes the assumption that components with width less than Θ_1 and height less than Θ_2 must be noise components and should not be included as a factor in the MFBS. (See Figure 4)

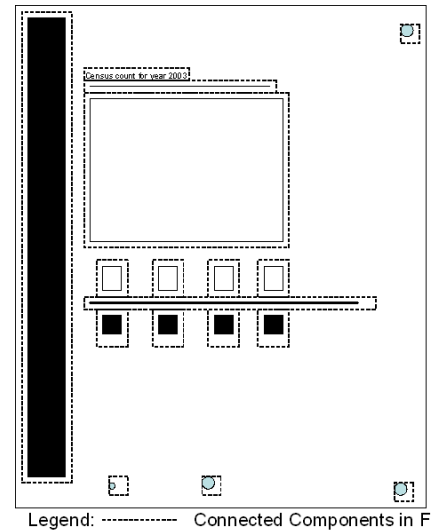


Figure 4: Connected Components in F.

The next operation forms the set P as,

$$P := \left\{ \forall c_i \in F \mid c_i \text{ width} > (3 \times c_i \text{ height}) \right. \\ \left. \vee \frac{c_i \text{ \#ofpixels}}{c_i \text{ area}} < \Phi \right\}$$

where $c_{\#ofpixels}$ is the number of pixels that compose that component, $1 \leq i \leq |F|$, c_{area} is the area of the bounding rectangular box of the component (the maximum number of pixels that could be in that component), and Φ is an upper of the pixel ratio of a component. If this ratio is above Φ the area is determined to be a large black area, which can sometimes be formed with a bad scan, and is therefore removed. The only large black areas that should be included in the MFBS are lines. This is why we check first do the check to see if the width is at least three times the height, if this is the case the component should not be removed from the MFBS. (See Figure 5)

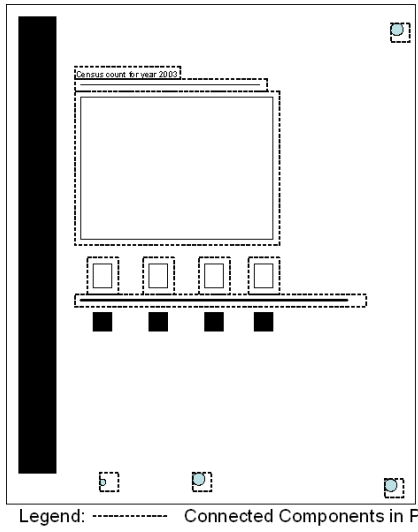


Figure 5: Connected Components in P.

Lastly we form the set W as,

$$W := \left\{ \forall c_i \in P \mid c_i \#ofpixels > \Psi \right. \\ \left. \wedge \min_{1 \leq j \leq m, i \neq j} d(c_i, p_j) < \Upsilon \right\}$$

where $d(c, P_j)$ is a measure of the distance between two components, $1 \leq i \leq |P|$, Ψ is a lower bound for the number of pixels in a component, and Υ is an upper bound for the distance that one component can be from another. This operation removes components whose pixel count is too low to be an actual character and whose minimum distance from other components is above a threshold Υ meaning it is too distant to be part of the MFBS. (See Figure 6)

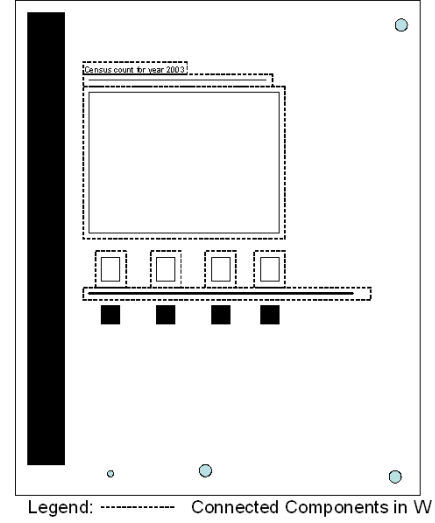


Figure 6: Connected Components in W.

3.2 Hierarchal Merging Procedures

Sort W such that

$$W := \{ \forall c_i \in W \mid c_{i-1} x_{min} \leq c_i x_{min} \leq c_{i+1} x_{min} \}$$

where x_{min} is the leftmost x coordinate of the component c_i .

Each process in this step uses a nested *FOR* loop iteration scheme, as seen in Figure 7, to compare each component with each other component. For

```

COMPONENT ITERATION TEMPLATE( )
1  for j ← 1 to |W| - 1
2  do if W_j exists
3      then for k ← j+1 to |W|
4          do if W_k exists
5              then if < conditions >
6                  then W_j = W_j ∪ W_k
7                      removed(W_k)
8

```

Figure 7: The general process of iterating through W in order to compare each component with each other component

each component compared, one or more conditions are applied to determine whether or not to merge the components by the operation

$$W_j = W_j \cup W_k \\ remove(W_k)$$

The first step is to merge any component rectangles that seem to follow one another vertically. This process is performed in order to join lines of text together into a union of these lines based on the maximum distance possible between words and lines. The algorithm merging is a merging procedure based on proximities of component rectangles. The conditions for the merging are

$$\begin{aligned} cond_1 = & [W_j y_{min} \leq W_k y_{midpoint} \wedge \\ & W_k y_{midpoint} \leq (W_j y_{min} + W_j height)] \\ & \vee [W_k y_{min} \leq W_j y_{midpoint} \wedge \\ & W_j y_{midpoint} \leq (W_k y_{min} + W_k height)] \end{aligned}$$

$$cond_2 = \frac{W_k x_{min} - (W_j x_{min} + W_j width)}{\gamma_{words}} \leq$$

The next step is a merging procedure that combines lines based on their vertical relationship, given by γ_{lines} . The conditions for this merge are

$$\begin{aligned} cond_1 = & |W_k y_{min} - (W_j y_{min} + W_j height)| < \\ & \gamma_{lines} \\ & |W_j y_{min} - (W_k y_{min} + W_k height)| < \\ & \gamma_{lines} \end{aligned}$$

$$cond_2 = \frac{|W_k x_{min} - W_j x_{min}|}{\gamma_{lineEnding}} <$$

Three constants to be defined are γ_{words} , γ_{lines} , and $\gamma_{lineEnding}$ whereby

$$\gamma_{words} = \beta_{words} \cdot \left(\frac{w}{FORM_WIDTH} \right)$$

$$\gamma_{lines} = \beta_{lines} \cdot \left(\frac{h}{FORM_HEIGHT} \right)$$

$$\gamma_{lineEnding} = \beta_{words} \cdot \left(\frac{h}{FORM_HEIGHT} \right)$$

where γ_{words} , γ_{lines} , and $\gamma_{lineEnding}$ are the respective thresholds for max distance between words and lines, w and h are the width and height respectively of the input image, $FORM_WIDTH$ and $FORM_HEIGHT$, and β_{words} and β_{lines} are normalization factors.

Once these lines have been merged, then the next merging procedure is implemented. This next procedure merges any symbols that are spaced from each other, but otherwise appear to be apart of a line. The conditions for this merge are

$$cond_1 = [W_j y_{min} \leq W_k y_{midpoint} \wedge$$

$$\begin{aligned} & W_k y_{midpoint} \leq (W_j y_{min} + W_j height)] \\ & \vee [W_k y_{min} \leq W_j y_{midpoint} \wedge \\ & W_j y_{midpoint} \leq (W_k y_{min} + W_k height)] \end{aligned}$$

$$cond_2 = \frac{\min(W_j height, W_k height)}{\max(W_j height, W_k height)} \geq \rho$$

$$cond_3 = \frac{(W_j width + W_k width)}{(W_{union_{ij}} width \times c)} >$$

where ρ is a bound for the similarity of heights between two components, $W_{union_{ij}} width$ is the width of the resulting rectangle union, and the constant c is a normalization factor for that union.

The next step merges any lines that begin and end near the same locations along the x-axis, and are relatively long lines. This binds form bodies together using and full length body lines. This combination is based on the minimum length for a complete form line computed by

$$\gamma_{completeLine} = \beta_{completeLine} \cdot \left(\frac{w}{FORM_WIDTH} \right)$$

where $\gamma_{completeLine}$ is the minimum length for a complete form line, and $\beta_{completeLine}$ is a normalization factor. The two conditions that must be met for components to be merged by this application are

$$\begin{aligned} cond_1 = & |W_j x_{min} - W_k x_{min}| < \gamma_{lineEnding} \\ cond_2 = & \max(W_j x_{min} + W_j width, W_j x_{min} + \\ & W_j width) - \min(W_j x_{val}, W_k x_{val}) > \\ & \gamma_{completeLine} \end{aligned}$$

Then, any component that has an intersection is merged.

3.3 MFBS Construction from Body Components

The final step is to combine the set of rectangular body segments, W , into one "best" segment that marks the major form body segment, m , by some merging function f where

$$m \leftarrow f(W)$$

After examining W for many images, we determined that there are two possible ways W can be divided. First, there is already one very large component, in which case this component can be made the MFBS. Second, there are many small components which have not been merged, in which case

we diagnose the image as segmented and make a bounding box that covers all of the components as the final MFBS. The first step of f is to find the components with the largest and second largest area, $W_{largestArea}$ and $W_{secondArea}$. Then the construction of m is performed as in Figure 8. After the final step, m contains the MFBS [5]. (Figure 9)

```

BODY COMPONENT MERGING( )
1   $m \leftarrow empty$ 
2  if  $W_{secondArea} < (W_{largestArea} * \phi)$ 
3    then  $m \leftarrow largest$ 
4    for  $i = 1$  to  $W_{size}$ 
5      do if  $W_i width > \omega_{width}$ 
6         $\wedge W_i height > \omega_{height}$ 
7         $\wedge W_i area > \omega_{area}$ 
8         $\wedge [\frac{W_i width}{W_i height} < \omega_{ratio}]$ 
9         $\vee W_i width > \omega_{wratio}]$ 
10     then
11        $m \leftarrow m \cup W_i$ 
12
13  else  $m \leftarrow largest$ 
14    for  $i = 1$  to  $W_{size}$ 
15      do if  $W_i area > largestArea * \omega_{arearatio}$ 
16        then  $m \leftarrow m \cup W_i$ 
17
18

```

Figure 8: Process for combining components of W into one component m given $W_{largestArea}$ and $W_{secondArea}$

4 Conclusions and Future Work

Detection of the Major Form Body Segment is an essential first step in the unconstrained document analysis process. Using the iterative methods described in this paper we created a potent algorithm for MFBS detection. The algorithm's strength lies in its ability to not include other unnecessary artifacts that emerge during the scanning process, such as large black areas and small speckled noise. Once these artifacts have been removed and the MFBS has been detected a more worry free analysis of the document will become possible.

In the future we will incorporate auto-correlation to help the noise removal process. Also we will incorporate the information about the placement and size of the margin's (from the MFBS process)

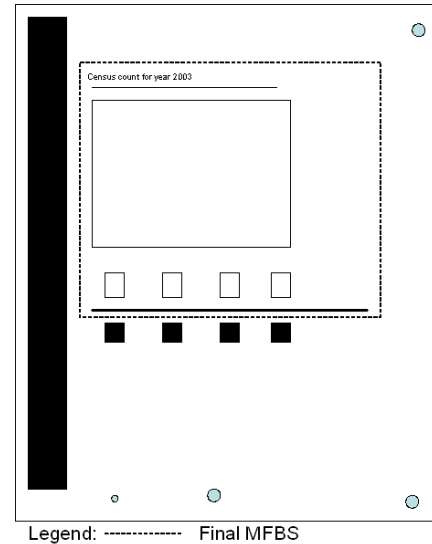


Figure 9: Final MFBS.

to help in our form classification efforts.

References

- [1] J. Sauvola and M. Pietikainen. Page Segmentation and Classification using fast Feature Extraction and Connectivity Analysis. In Proc. of the 3th International Conference on Document Analysis and Recognition pages 1127-1131, Montreal, Canada, 1995.
- [2] Thien M. Ha, H. Bunke. Image Processing Methods for Document Image Analysis. Handbook of Character Recognition and Document Image Analysis, pp.1-47. 1997, World Scientific Publishing Company.
- [3] J. Duong, M. Cote, H. Emptoz, C. Suen. Extraction of Text Areas in Printed Document Images. ACM Symposium on Document Engineering, Atlanta (USA) , 2001, pp. 157-165.
- [4] Jae Adams, E.A. Yfantis, D. Curtis and T. Pack. Feature Extraction Methods for Form Recognition Applications. *WSEAS trans. on Information Science and Applications*, Issue 3, Volume 3 March 2006 Pages 666-671.
- [5] D. Dori, D. Doermann, C. Shin, R. Haralick, I. Phillips, M. Buchman, D. Ross. The Representation of Document Structure: A Generic Object-Process Analysis. Handbook of Character Recognition and Document Image Analysis,

pp.421-456. 1997, World Scientific Publishing
Company.