Document Image Segmentation Using Dynamic Thresholds and Identification of Each Region Type

Silky Arora¹, Dharamveer Sharma², Shilpa Arora³

¹M Tech, Department of Computer Science, Punjabi University, Patiala
²Department of Computer science, Punjabi University, Patiala
³M Tech, Department of Computer Science, Punjabi University, Patiala

Abstract

Nowadays, the accumulation of paper in the life of business professional is overwhelming. Digital documents on the other hand being less expensive and more efficient are on the road to a more organized office. Also crucial for many business applications document image analysis is needed before OCR operation. Segmentation of document images into text and non-text regions is essential because OCR recognition engine produces garbage text when it gets non-text components as input. In this paper a segmentation technique is presented to decompose document image into its constituent parts, such as text blocks, pictures and tables. This technique is implemented by combining two approaches- run length smearing algorithm and recursive top down approach. Recursive top down approach works using horizontal and vertical projection profiles. Proposed technique is threshold based but threshold values are automatically calculated depending upon the geometric layout of the document. Binarization and noise removal are done as part of preprocessing. This approach works for documents in any script and in manhattan layout.

Keywords- document image segmentation, dynamic thresholds, horizontal and vertical projection profiles, OCR, recursive top down segmentation, run length smearing

Introduction

In digital format documents can easily be stored, retrieved and manipulated. Therefore nowadays many techniques are being developed for analysis and recognition of scanned documents. Documents can be in various layouts depending upon where they have been taken from (newspapers, magazines, journals, text books etc.). In this paper we have considered class of documents in manhattan layout only. When a document containing text and non-text components is directly fed to OCR, the OCR recognition engine recognizes textual components correctly but it produces garbage text on occurrence of non-text components. Thus it is important to understand structural schema of the document before OCR operation.

Page segmentation is done to analyze structure and content of the document. Various page segmentation techniques have been proposed which are broadly categorized as top down approach and bottom up approach [1]. Top down approach takes whole image as input and decompose it into hierarchy of rectangular regions while bottom up approach starts from pixel level information of the image and start combining them using connected component information. In this paper a top down segmentation technique is proposed that first isolate headings of the document using horizontal run lengths of run length smearing algorithm and after that recursively decomposes the rest of the document into non-overlapping smaller rectangular region until they share similar physical characteristics. It is assumed that skew correction is already done on the image. Binarization and noise removal are done as part of preprocessing. Threshold values are calculated dynamically based on physical layout of the document.

This paper is organized as follows: section II covers related work in document image segmentation, section III gives objectives and proposed technique in detail followed by results in section IV and section V concludes the paper.

Related Work

One of the earliest approaches of document analysis system was given by Wong et al. [2]. For segmentation and classification of digital documents it uses run length smearing approach. By using regular features of text lines, a linear adaptive classification scheme discriminates text regions from others. X-Y cut page segmentation technique [3] is based on top down approach. It uses tree based structure in which the whole document is treated as root and respective decomposed rectangular regions as leaf nodes. Decomposition is done on the basis of horizontal and vertical projection profiles of foreground pixels. White space analysis method [4] unlike many traditional segmentation approaches is independent of any threshold values. In this method white spaces runs greater than one fifth of the page are identified in both horizontal and vertical directions. Thinning algorithm is used for thinning of lines. In this way a mesh is formed by combining lines in both horizontal and vertical directions. Also many bottom-up Approaches are used for page segmentation and block identification [5, 6]. Yuan, Tan [7] designed method that makes use of edge information to extract textual blocks from gray scale document images. It aims at detecting only textual regions on heavy noise infected newspaper images and separate them from graphical regions. The White Tiles Approach [8] described new approaches to page segmentation and classification. In this method, once the white tiles of each region have been gathered together and their total area is estimated, and regions are classified as text or images. Docstrum algorithm [9] is a bottom up approach that starts from pixel level information and finds k-nearest neighborhood pixels and start converging them. This approach is also threshold based,

thresholds are based on properties of distance and angle of each connected components with k-nearest neighbors.

Proposed Technique

The objective of the proposed technique is to recursively decompose document image into hierarchy of homogeneous regions while calculating threshold values dynamically depending upon geometric layout of the document. Decomposition is followed by identification of each region type.

The proposed work can be divided into broadly four steps which are explained below:

Step 1: Preprocessing

In this step binarization and noise removal is done on the document image. After binarization noise removal is done by removing both salt and pepper noise. Every single pixel value is compared with its neighbor pixels; if neighbor pixels differ from the original pixel then original pixel value is changed to the neighbor pixel values.

Step 2: *Isolation of heading rows*

In this step input image is binarized and noise free. Horizontal run lengths of the image are drawn using run length smearing algorithm. Initial rectangles of comparable width are identified as heading of the document. Any number of heading rows can be identified using this approach.

Step 3: Decomposition into rectangular region

Excluding heading rows the rest of the image area is given as input in this step. A recursive top down technique is implemented that uses information from horizontal and vertical projection profiles for decomposing it into its constituent regions. Threshold values are calculated dynamically depending upon the physical structure and geometric location of various constituent regions. To calculate horizontal threshold value horizontal histograms are calculated for the inputted region of document. Distances are calculated in between horizontal histograms where their value is equal to or slightly greater than zero. Total distance is divided by number of gaps to get average horizontal distance which in turn is horizontal threshold value. Similarly vertical threshold value is calculated.

Say rectangle R is the input image area. Algorithm for recursive decomposition of image:

- 1. Create Horizontal Projection Profiles (HPPs) and find horizontal gaps more than or equal to horizontal threshold. Divide the rectangle R into smaller rectangles by creating dissection points using these horizontal gaps.
- 2. If number of rectangles formed using HPPs is more than 1 then create Vertical Projection Profiles (VPPs) for each of rectangle and find vertical gaps more than or equal to vertical threshold. Using these vertical gaps, final rectangles are formed.

3. If number of rectangles using HPPs is 1 then create Vertical Projection Profiles (VPPs) for that rectangle and find vertical gaps more than or equal to vertical threshold. Using these vertical gaps, final rectangles are formed. If number of rectangle formed using VPPs is more than one then create HPPs for each rectangle and horizontal gaps more than or equal to horizontal threshold are identified. Final rectangles are created using HPPs.

This algorithm works recursively until there are no horizontal and vertical gaps greater than horizontal and vertical threshold values. At the end of this step we have document image in which heading rows are separated and rest of the image is divided into its constituent blocks having rectangular boundaries.

Step 4: Identification of type of each region

Input image in this step is document image in which each constituent region is identified and has rectangular boundary around it. In this step we will identify type of each rectangular region and classify them in text blocks, picture and tables. Algorithm for region type identification is as follows:

- 1. Draw horizontal histograms for each rectangular region.
- 2. If for a rectangular region at regular gaps horizontal histogram values are not equal to or slightly greater than 0 then the rectangular region is picture.
- 3. If for a rectangular region at regular gaps horizontal histogram values are equal to or slightly greater than 0 then the rectangular region is text block or table.
- 4. To further differentiate between text and tables we will draw vertical histograms for both. If at regular gaps vertical histogram values are equal to or slightly greater than 0 then the region is table otherwise it is text.

Results

Following are the result images at the end of each step during the whole process. Fig. 1 is the original input image. Fig. 2 is the image after binarization and noise removal. Fig. 3 is the output image after step 2, that is after isolating heading of the document and fig. 4 is the output image after step 3. Fig. 5 and fig. 6 shows image after horizontal and vertical histograms are drawn for block type identification.

	A world record after 21 years Pole venit world record progression			ing crowd after his mar was erased off the books. "That's fantastic. It's his		
	lecord	Athlete	Venue	tory and I'm very happ, that Renaud does it here		
d record by clearing 6.16 6.	1.16m	Renaud Lavillenie	Donetsk, 15-02-2014	in Donetsk, my town,		
es at a meeting in etsk on Saturday. 6.	.15m	Sergey Bubka	Donetsk, 21-02-1993	said Bubka, who came o		
ld Record ! 6.16 at my 6.	1.14m	Sergey Bubka	Lievin, 13-02-1993	the track to congratulat the Frenchman.		
pt ! That's incred- 11 in the air," Lav- 6.	6.13m	Sergey Bubka	Berlin, 21-02-1992	"I'm happy because my jo		
n his official 6.	6.11m	Sergey Bubka	Donetsk, 19-03-1991	is to help athletes perform Athletics are my life. I'r		
r the French ration had 6.	6.10m	Sergey Bubka	San Sebastian, 15-03-1991	pleased for him and for ath		
	6.08m	Sergey Bubka	Volgograd, 9-02-1991	letics," added Bubka, who i the vice-president of th		
mark of	6.05m	Sergey Bubka	Donetsk, 17-03-1990	International Association		
at the same event taine in 1993. "It's taly unbelievable," But beat it so early, that's tely unbelievable," But beat it so early, that's but beat it so early, that's something else." his hands when he "I did not know what was d he had done it, happening to me. The sound was crazy. These are huge enotions. I am in a l need time to get wno earth. Itwas al record. I knew I champion and 1988 Olympic		gold medalist Bubka still holds the outdoor world record of 6.14. Bubka, who on Saturday morning had hoped Laville- nie would beat his record in his hometown of Donetsk if i had to happen one day, stood up and clapped his hands along with the cheer-	Athletics Federations. Lavillenie, who won thi Olympic title at the Lon On Games and his first indoor world title the sam year, has an outdoor best o 6.02 metres. He has so far failed to wi an outdoor world title, coi lecting one silver and tw bronze medals. — Reuters			

Figure 1 Original Image

				ing crowd after his mark
PONETSK, FEBRUARY 16 Olympic champion Renaud Lavillenic of France estab-		record after 21 whit room programic		was erased off the books. "That's fantastic. It's his-
lished a pole vault indoor	Record	Athlete	Venue	tory and I'm very happy
world record by clearing 6.16	6.16m	Renaud Lavillenie	Donetsk, 15-02-2014	that Renaud does it here, in Donetsk, my town."
metres at a meeting in Donetsk on Saturday.	6.15m	Sergey Bubka	Donetsk, 21-02-1993	said Bubka, who came on
"World Record ! 6.16 at my	6.14m	Sergey Bubka	Lievin, 13-02-1993	the track to congratulate the Frenchman.
first attempt ! That's incred- ible, I'm still in the air," Lav-	6.13m	Sergey Bubka	Berlin, 21-02-1992	"I'm happy because my job
illenie wrote on his official	6.11m	Sergey Bubka	Donetsk, 19-03-1991	is to help athletes perform. Athletics are my life. I'm
Twitter feed after the French Athletics Federation had	6.10m	Sergey Bubka	San Sebastian, 15-03-1991	pleased for him and for ath-
flagged up the new record.	6.08m	Sergey Bubka	Volgograd, 9-02-1991	letics," added Bubka, who is the vice-president of the
The previous mark of	6.05m	Sergey Bubka	Donetsk, 17-03-1990	International Association of
6.15 was set by Sergei Bubka at the same event in Ukraine in 1993. This completely unbelievable? Lavillenie, who took his head in his hands when he realised he had done it, later told French news channel BFMTV. "I will need time to get back down on earth. Itwasa mythical record. I knew I	vent "It's had the potential to do try ble." But beat it so early, that his something else." n he "I did not know what we e it, happening to me. TI news sound was crazy. These a huge emotions. I am in get imension." vas a Six-time outdoor, wor		gold medalist Bubka still bolds the outdoor world record of 6.14. Bubka, who on Saturday morning had hoped Laville- nie would beat his record in his hometown of Donetsk if it had to happen one day, stood up and clapped his hands along with the cheer-	Athletis Relevations. Lavillenie, who won the Olympic title at the Lon- don Games and his first indoor world title the same year, has an outdoor best of 6.02 metres. He has so far failed to win an outdoor world title, col- lecting one silver and two bronze medals. — Reuters

Figure 3 Image after Isolation of Heading

Lavillenie vaults over Bubka's record

			was erased off the books. "That's fantastic. It's his-	
Record	Athlete	Venue		
6.16m	Renaud Lavillenie	Donetsk, 15-02-2014		•
6.15m	Sergey Bubka	Donetsk, 21-02-1993	said Bubka, who came on	
6.14m	Sergey Bubka	Lievin, 13-02-1993		1
6.13m	Sergey Bubka	Berlin, 21-02-1992	"I'm happy because my job	-
6.11m	Sergey Bubka	Donetsk, 19-03-1991	is to help athletes perform.	1
6.10m	Sergey Bubka	San Sebastian, 15-03-1991		1
6.08m	Sergey Bubka	Volgograd, 9-02-1991	letics," added Bubka, who is	1
6.05m	Sergey Bubka	Donetsk, 17-03-1990	International Association of Athletics Federations.	1
But beat somethir "I did r happeni sound w huge en new dim Six-tim	i it so early, that's ag else." not know what was ng to me. The as crazy. These are notions. I am in a eension." we outdoor world	gold mediais: Bubka still holds the outdoor world record of 8.14. Bubka, who on Saturday morning had hoped Laville- nie would best his record in his hometown of Donetski fi thad to happen one day, stood up and clapped his hands along with the cheer-	Lavillene, who won the Olympic title at the Lon- don Games and his first indoor world title the same year, has an outdoor best of 6.02 metres. He has so far failed to win an outdoor world title, col- lecting one silver and two bronze medals.— Reuters	
	Pelo mark Record 6.16m 6.15m 6.15m 6.15m 6.16m 6.13m 6.10m 6.08m 6.08m 6.05m had the pel sound ku happeni sound ku happeni sound ku happeni sound ku	Pub wait word near projection Record Athlete 6.16m Renaud Lavillenie 6.16m Sergey Bubka 6.14m Sergey Bubka 6.14m Sergey Bubka 6.14m Sergey Bubka 6.10m Sergey Bubka 6.00m Sergey Bubka 6.05m Sergey Bubka mathematic serger Sergey Bubka 6.05m Sergey Bubka 6.05m Sergey Bubka 6.05m Sergey Bubka mathematic serger Serger mathematic serger Serger mathematic serger Serger something else " mathematic serger Serger something serger Team and mediation" Startime outdoor wordd	Bubble Donetsk, 15-02-2014 6.16m Renaud Lavillenie Donetsk, 15-02-2014 6.15m Sergey Bubka Donetsk, 21-02-1993 6.14m Sergey Bubka Lievin, 13-02-1993 6.13m Sergey Bubka Berlin, 21-02-1992 6.11m Sergey Bubka Donetsk, 19-03-1991 6.10m Sergey Bubka Donetsk, 19-03-1991 6.08m Sergey Bubka Donetsk, 19-03-1991 6.06m Sergey Bubka Volgograd, 9-02-1991 6.05m Sergey Bubka Donetsk, 17-03-1990 had the potential to do try it. gold medalist Bubka still But beat it so early, that's something else." record of 6.14. "I did not know what was happening to me. The are dimension." Bubka, who on Saturday moring had hopei Laville-nie would beat his record in huge emotions. I am in a in hometworn dDonetsk if had to happen one day, Stx-time outdoor world	Period walk word incord programmer That's fantastic. It's history and Tru very happy that Renaud does it here, in 300000000000000000000000000000000000

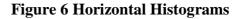
Figure 2 Image after Preprocessing

NONETSK, FEBRUARY 16 Dlympic champion Renaud avillenie of France estab-	rid record after 21 at writi roard programics	yoars	ing crowd after his mark was erased off the books. "That's fantastic. It's his-
shed a pole vault indoor Recon	d Athlete	Venue	tory and I'm very happy
ecord by clearing 6.16 6.16m	Renaud Lavillenie	Donetsk, 15-02-2014	that Renaud does it here, in Donetsk, my town."
a meeting in Saturday. 6.15m	Sergey Bubka	Donetsk, 21-02-1993	said Bubka, who came on
d!6.16 at my 6.14m	Sergey Bubka	Lievin, 13-02-1993	the track to congratulate the Frenchman.
! That's incred- in the air," Lav- 6.13m	Sergey Bubka	Berlin, 21-02-1992	"I'm happy because my job
on his official 6.11m	Sergey Bubka	Donetsk, 19-03-1991	is to help athletes perform. Athletics are my life. I'm
iter the French deration had 6.10m	Sergey Bubka	San Sebastian, 15:03-1991	pleased for him and for ath-
he new record. 6.08n	Sergey Bubka	Volgograd, 9-02-1991	letics," added Bubka, who is
ous mark of 6.05n	n Sergey Bubka	Donetsk, 17-03-1990	the vice-president of the International Association of
Lavillenie, who took his head in his hands when he realised he had done it, happer later told French news channel BFM TV.	e potential to do try it. eat it so early, that's hing else." d not know what was ning to me. The was crazy. These are emotions. I am in a imension." ime outdoor world	gold medalist Bubka still holds the outdoor world record of 6.14 Bubka, who on Saturday morning had hoped Laville- nie would beat his record in his hometown of Donetsk if it had to happen one day, stood up and clapped his	Athletics Federations. Lavillenie, who won the Olympic title at the basis indoor world title the same year, has an outdoor best of 6.02 metres. He has so far failed to win an outdoor world title, col- lecting one silver and two bronze media.— Reuters

Figure 4 Image after Recursive Decomposition

DIVERSE, FEBRUARY IS Diverse champion Based Lavillenic of Prance-setter Record Athlete		ing crowd after his <u>serio</u> was erased off the books "That's fantastic. It's tory and I'm very bar	ponersk, februar 16 Olympic champion Renaud Lavillenie of France estab- lished a pole vault indoor	A world record after 21 Pub wat world record programs Record Athlete		ing crowd after his mark was erased off the books. "That's fantastic. It's his- tory and I'm very happy
Index of pole training 6.16m Renaud Lavillenia metres at a meeting 6.15m Sergey Bubka Donetsk on Saturday	Control Office Office Office Densel 400 Carlos Densel 400 Densel 4000 Densel 400 D	that Renaud does it and in Donetsk, my torm said Bubka, who care the Frenchman. "I'm happy because pairs is to help athletes pairs pleased for him and factor pleased for him and factor help the factor of help the factor of help the factor help the help the factor help the fac	wohl record by cleaning, 6.1 metres at a meeting in Donetsk on Saturday. "World Record 16.16 at my first attempt 1 That's incred- ible, I'm still in the air," Lav- illenie wrote on his official Twitter feed after the French Athletics Federation had flagged up he new record . The -previous mark of 6.15 was set by Sergei Bubka at the same event in Ukraine in 1989. "Its completely unbelievable," Lavillenie, who took his head in his hands when he realised to had done ti, after took for the same event is at the same bad done ti, after took for the same event at the same when he realised to had done ti, after took for the same event is at the same same same same same same same at the same same same same same same same same same same same same same same	6.16m Renaud Lavillenie 6.15m Sergey Bubka 6.14m Sergey Bubka 6.13m Serley Bubka 6.13m Serley Lubka	Donetsk, 1502-2014 Donetsk, 2102-1993 Lievin, 13-02-1993 Berin, 21-02-1992 Donetsk, 19-03-1991 San Schentm, 2-04-1994 San Schentm, 2-04-1994 Schentm, 2-04-1	that Renaud does it here, in Donetsk, my town." said Bubka, who came on the track to congratulate the Frenchman. "I'm happy because my job i is to help athletes perform. A Athletics are my life. I'm pleased for him and for ath- letics," added Bubka, who is the vice president of the International Association of Athletics Pederations. Lavillenie, who won the Olympic title at the Lon- don Games and his first indoor world title the same year, has an outdoor best of a standard.

Figure 5 Vertical Histograms



Conclusion

In this paper a segmentation technique is presented that recursively divides the document image into hierarchy of homogeneous regions while calculating threshold values dynamically. This way task of choosing threshold values is removed which was a crucial step in performance of a segmentation procedure. Region type identification is done on the basis of horizontal and vertical histogram values, which is very easy and efficient to implement.

References

- [1] S. Khedekar, V. Ramanaprasad, S. Setlur, *Text Image Separation in Devanagari Documents*. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003.
- [2] K.Y. Wong, R.G. Casey and F.M. Wahl, *Document analysis system*. IBM Journal of Research and Development 1982; 26(6):647-656.
- [3] B. Kruatrachue, N. Moongfangklang and K. Siriboon, Fast *Document Segmentation Using Contour and X-Y Cut Technique*. International Journal of Computer, Information science and Engineering 2007; 1(5).
- [4] R. Garg, G. Harit and S. Chaudhury, A hierarchical analysis scheme for robust segmentation of Document Images using white-spaces. Proceedings of 1st National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics 2008.

- [5] K. Lee, Y. Choy, and S. Cho, *Geometric Structure Analysis of Document Images: A Knowledge-Based Approach*. IEEE transactions on Pattern Analysis and Machine Intelligence 2000; 22(11).
- [6] P. Mitchell, H. Yan, *Newspaper document analysis featuring connected line segmentation*. Proceedings of International Conference on on Document Analysis and Recognition, ICDAR'01, 2001.
- [7] Q. Yuan, C.L. Tan, *Text Extraction from Gray Scale Document Images Using Edge Information*. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR'01, 2001: 302-306.
- [8] A. Antonacopoulos and R. T. Ritchings "Segmentation and Classification of Document Images", The Institution of Electrical Engineers 1995.
- [9] L. O'Gorman, *The Document Spectrum for Page Layout Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence 1993; 15(11): 1162–1173.