International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1869-1875
© International Research Publications House
http://www. irphouse.com

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

Silky Arora[1], Dharamveer Sharma[2], Shilpa Arora[3]

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

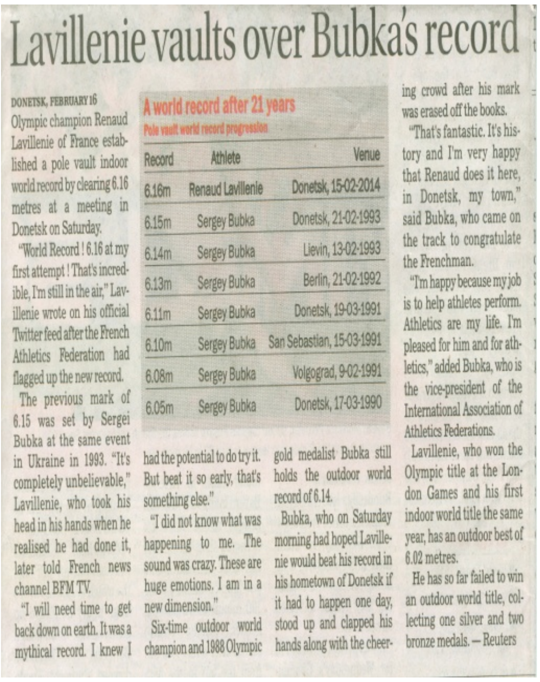
Step 4: ***Identification of type of each region***

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.



Lavillenie vaults over Bubka's record

DONETSK, FEBRUARY 16

Olympic champion Renaud Lavillenie of France established a pole vault indoor world record by clearing 6.16 metres at a meeting in Donetsk on Saturday.

"World Record ! 6.16 at my first attempt ! That's incredible, I'm still in the air," Lavillenie wrote on his official Twitter feed after the French Athletics Federation had flagged up the new record.

The previous mark of 6.15 was set by Sergei Bubka at the same event in Ukraine in 1993. "It's completely unbelievable," Lavillenie, who took his head in his hands when he realised he had done it, later told French news channel BFM TV.

"I will need time to get back down on earth. It was a mythical record. I knew I had the potential to do try it. But beat it so early, that's something else."

"I did not know what was happening to me. The sound was crazy. These are huge emotions. I am in a new dimension."

Six-time outdoor world champion and 1988 Olympic gold medalist Bubka still holds the outdoor world record of 6.14.

Bubka, who on Saturday morning had hoped Lavillenie 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 cheering crowd after his mark was erased off the books.

"That's fantastic. It's history and I'm very happy 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 is to help athletes perform. Athletics are my life. I'm pleased for him and for athletics," added Bubka, who is the vice-president of the International Association of Athletics Federations.

Lavillenie, who won the Olympic title at the London Games and his first indoor world title the same year, has an outdoor best of 6.02 metres.

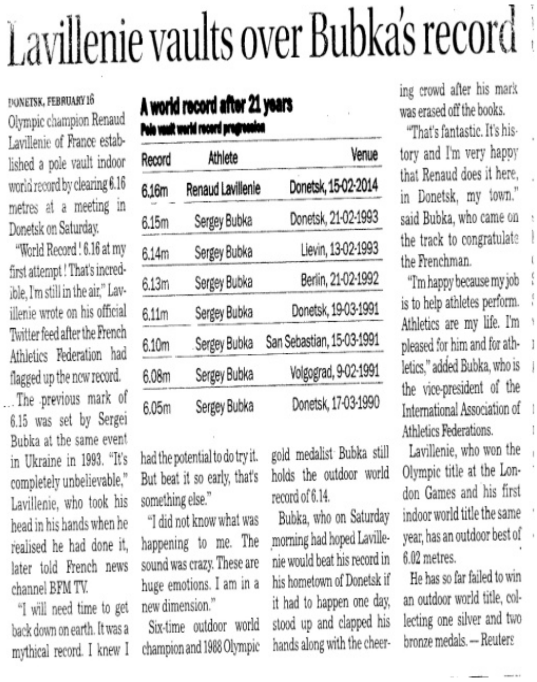
He has so far failed to win an outdoor world title, collecting one silver and two bronze medals. — Reuters

A world record after 21 years

Pole vault world record progression

Record	Athlete	Venue
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	San Sebastian, 15-03-1991
6.08m	Sergey Bubka	Volgograd, 9-02-1991
6.05m	Sergey Bubka	Donetsk, 17-03-1990

Figure 1 Original Image



Lavillenie vaults over Bubka's record

DONETSK, FEBRUARY 16

Olympic champion Renaud Lavillenie of France established a pole vault indoor world record by clearing 6.16 metres at a meeting in Donetsk on Saturday.

"World Record ! 6.16 at my first attempt ! That's incredible, I'm still in the air," Lavillenie wrote on his official Twitter feed after the French Athletics Federation had flagged up the new record.

The previous mark of 6.15 was set by Sergei Bubka at the same event in Ukraine in 1993. "It's completely unbelievable," Lavillenie, who took his head in his hands when he realised he had done it, later told French news channel BFM TV.

"I will need time to get back down on earth. It was a mythical record. I knew I had the potential to do try it. But beat it so early, that's something else."

"I did not know what was happening to me. The sound was crazy. These are huge emotions. I am in a new dimension."

Six-time outdoor world champion and 1988 Olympic gold medalist Bubka still holds the outdoor world record of 6.14.

Bubka, who on Saturday morning had hoped Lavillenie 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 cheering crowd after his mark was erased off the books.

"That's fantastic. It's history and I'm very happy 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 is to help athletes perform. Athletics are my life. I'm pleased for him and for athletics," added Bubka, who is the vice-president of the International Association of Athletics Federations.

Lavillenie, who won the Olympic title at the London Games and his first indoor world title the same year, has an outdoor best of 6.02 metres.

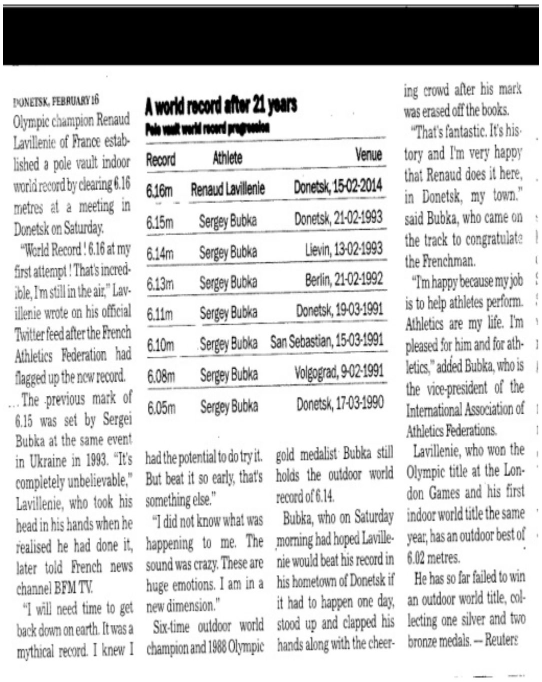
He has so far failed to win an outdoor world title, collecting one silver and two bronze medals. — Reuters

A world record after 21 years

Pole vault world record progression

Record	Athlete	Venue
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	San Sebastian, 15-03-1991
6.08m	Sergey Bubka	Volgograd, 9-02-1991
6.05m	Sergey Bubka	Donetsk, 17-03-1990

Figure 2 Image after Preprocessing



DONETSK, FEBRUARY 16

Olympic champion Renaud Lavillenie of France established a pole vault indoor world record by clearing 6.16 metres at a meeting in Donetsk on Saturday.

"World Record ! 6.16 at my first attempt ! That's incredible, I'm still in the air," Lavillenie wrote on his official Twitter feed after the French Athletics Federation had flagged up the new record.

The previous mark of 6.15 was set by Sergei Bubka at the same event in Ukraine in 1993. "It's completely unbelievable," Lavillenie, who took his head in his hands when he realised he had done it, later told French news channel BFM TV.

"I will need time to get back down on earth. It was a mythical record. I knew I had the potential to do try it. But beat it so early, that's something else."

"I did not know what was happening to me. The sound was crazy. These are huge emotions. I am in a new dimension."

Six-time outdoor world champion and 1988 Olympic gold medalist Bubka still holds the outdoor world record of 6.14.

Bubka, who on Saturday morning had hoped Lavillenie 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 cheering crowd after his mark was erased off the books.

"That's fantastic. It's history and I'm very happy 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 is to help athletes perform. Athletics are my life. I'm pleased for him and for athletics," added Bubka, who is the vice-president of the International Association of Athletics Federations.

Lavillenie, who won the Olympic title at the London Games and his first indoor world title the same year, has an outdoor best of 6.02 metres.

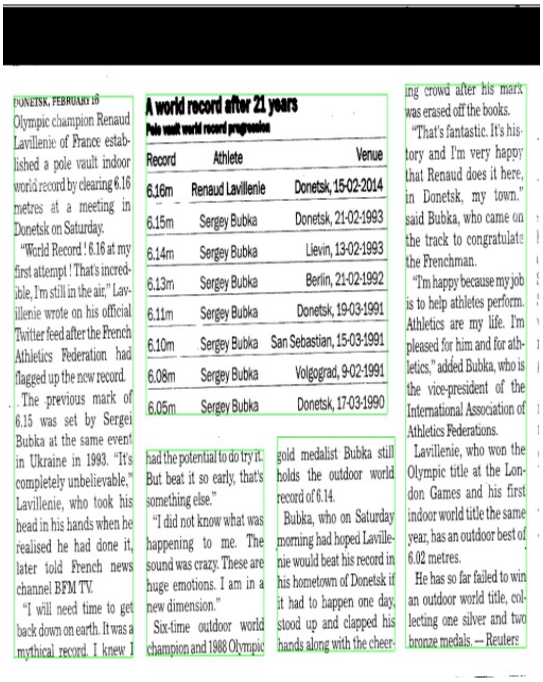
He has so far failed to win an outdoor world title, collecting one silver and two bronze medals. — Reuters

A world record after 21 years

Pole vault world record progression

Record	Athlete	Venue
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	San Sebastian, 15-03-1991
6.08m	Sergey Bubka	Volgograd, 9-02-1991
6.05m	Sergey Bubka	Donetsk, 17-03-1990

Figure 3 Image after Isolation of Heading



DONETSK, FEBRUARY 16

Olympic champion Renaud Lavillenie of France established a pole vault indoor world record by clearing 6.16 metres at a meeting in Donetsk on Saturday.

"World Record ! 6.16 at my first attempt ! That's incredible, I'm still in the air," Lavillenie wrote on his official Twitter feed after the French Athletics Federation had flagged up the new record.

The previous mark of 6.15 was set by Sergei Bubka at the same event in Ukraine in 1993. "It's completely unbelievable," Lavillenie, who took his head in his hands when he realised he had done it, later told French news channel BFM TV.

"I will need time to get back down on earth. It was a mythical record. I knew I had the potential to do try it. But beat it so early, that's something else."

"I did not know what was happening to me. The sound was crazy. These are huge emotions. I am in a new dimension."

Six-time outdoor world champion and 1988 Olympic gold medalist Bubka still holds the outdoor world record of 6.14.

Bubka, who on Saturday morning had hoped Lavillenie 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 cheering crowd after his mark was erased off the books.

"That's fantastic. It's history and I'm very happy 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 is to help athletes perform. Athletics are my life. I'm pleased for him and for athletics," added Bubka, who is the vice-president of the International Association of Athletics Federations.

Lavillenie, who won the Olympic title at the London 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, collecting one silver and two bronze medals. — Reuters

A world record after 21 years

Pole vault world record progression

Record	Athlete	Venue
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	San Sebastian, 15-03-1991
6.08m	Sergey Bubka	Volgograd, 9-02-1991
6.05m	Sergey Bubka	Donetsk, 17-03-1990

Figure 4 Image after Recursive Decomposition

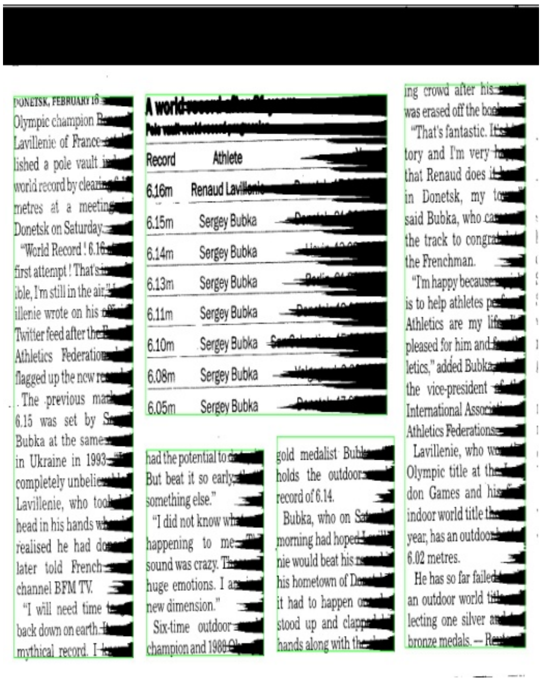


Figure 5 Vertical Histograms

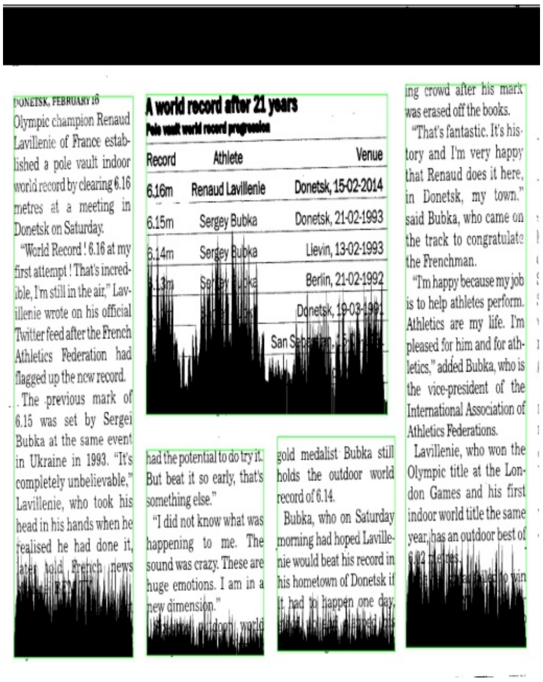


Figure 6 Horizontal Histograms

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.