

UNCONSTRAINED TIGHT STRUCTURE EXTRACTION USING VORONOI TESSELATION ON DOCUMENT IMAGES

P.Nagabhushan¹

pnagabhushan@hotmail.com

Sahana D Gowda²

sahanagowda@rediffmail.com

R.K.Bharathi³

rkbharathi@hotmail.com

1. Department of studies in Computer Science, Manasagangothri, University of Mysore, Mysore-570006, INDIA
2. Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore-560004, INDIA
3. Department of MCA, PES College of Engineering, Mandya -571401, INDIA

ABSTRACT

Document structure is the intermediary result obtained through page segmentation, which is used in the analysis of the document image. The structure serves the purpose of extracting the shape of the document from paragraph up to character level in a hierarchical exploratory methodology for understanding the layout structure of the document image. The extracted layout forms a dominant feature which plays a vital role in categorization, equivalence, ranking and retrieval of documents without reading. The main theme of this paper is to obtain the structure of the document, where entities possess unconstrained layout due to different contents. The Voronoi tessellation on document image with the points of interest being every high pixel in the image helps in recognizing the structure of the entities in the document image. New techniques known as Spring Force is being implemented to grab the point of interest present in the unbounded region and the exterior points which frame the exterior boundary of the entities in the document image to obtain a tight structure. The result of spring force technique is not only obtaining a tight structure for entities, but also to obtain the structure of the blank space in the document image. This method works successfully on all types of document images with Non-Manhattan layouts with arbitrary skew.

KEYWORDS: Unconstrained Tight Structure, Exterior Boundary Points, Tight Segmentation, Voronoi Tessellation, Spring Force Algorithm, Non-Manhattan Layout.

1. INTRODUCTION

Document Analysis or more precisely Document Image Analysis is the subfield of Digital Image Processing that performs the overall interpretation of Document Image. Document Image Analysis is the theory and practice of recovering the symbolic structure of Document Image scanned from paper or processed by computer. Basic elements or entities that are present in the Document Image include Text, Tables, Mathematical expressions, Binarized halftone, color picture and graphics [7, 9].

Analyzing the structure in Document Image is known as Document Structure Analysis. Document structure analysis consists of analyzing the Geometrical characteristics and logical consistency of entities in the Document Image [9]. The visual appearance of the Document Image may vary according to the type of the document. Hence the analysis of document structure consists of two phases: Physical layout analysis and logical layout analysis.

Physical Layout Analysis is the Analysis of the Geometrical characteristics of the entities in the Document Image which includes Position, Shape, Size, Area, Density, Sparsity, etc of the Objects [9].

The Logical Layout Analysis specifies the relationship and the Neighborhood connectivity between the entities that exists in the Document Image [9].

There are two basic types of layouts [7] such as:

- i. Manhattan Layout: where the regions are constrained polygons whose boundaries are horizontal and vertical lines. For example, Articles in magazine and news letters.

ii. Non-Manhattan Layout: where boundaries form unconstrained polygons or overlapping regions. For example, cover page of magazines, advertisements.

The result of segmentation can be used to extract the structure of the document image. Document segmentation is broadly classified into three types based on the approach such as: top-down, bottom-up and hybrid [7, 9]. Obtaining the structure of the document through any of the approach is domain dependent. For instance, articles in journals, magazines and news paper normally consist of Manhattan layout i.e. they are specified by specific general rules; they have the same layout in different editions. The main variation in the same article of same magazine in different editions is the structural variation due the existence of different contents. Hence the main aim of this paper is to extract the structure of the entity in the document image, which is obtained in a hierarchical methodology from paragraph to character level without reading or knowing the content of the article. Each character has a different structure; accurate structure extraction method is required to obtain the structure. This method helps in classification and categorization of documents based on structure of the document image.

Voronoi diagram technique: a bottom-up approach is used for document segmentation [1]. Applying the Voronoi diagram technique on document images, Voronoi tessellation is obtained: a structure which consists of bounded and unbounded cells, with each cell consisting of one point of interest [1]. In this work, as the point of interest is every high pixel, the internal structure of any entity can be precisely extracted from the document image. Voronoi tessellation consists of unbounded cells where extracting the Voronoi features is not appropriate as we consider area for a polygon as one of the Voronoi feature to decide upon where the edge which is between the two point of interest is a boundary or a superfluous edge. So, in this situation where deciding upon the edge as a boundary or superfluous edge between two regions is difficult, a new methodology known as Spring Force technique is implemented. Spring force technique is used to grab the points of interest in the unbounded cell and the cells which are in a fuzzy state to decide upon to which entity structure they belong to. This methodology results in obtaining the tight structure of the entities in the document image and also a precise structure of the blank space between the entities in the document image. This method works successfully on any document image which does not have noise. The care is taken to remove the noise using k-fill filter [6]. This paper is organized in sections, section 1 the introduction phase to understand the concept of voronoi and the limits of the existing methods. Section2 briefly describes the new methodology known as spring force technique with an example of hypothetical data set. Section 3 shows the result of spring force technique over the different document images. Section 4 discusses the limitation and novelty of spring force technique. Section 5 gives the references of the papers.

2. SPRING FORCE TECHNIQUE

Voronoi tessellation obtained from the function Voronoi [13] gives a geometrical structure of the document image. Here the input to the function Voronoi is every high pixel in the document image. In the previous works, obtaining the Voronoi tessellation the point of interests are obtained by considering the contour of the character [2] and the connected components of the characters [3]. These methods obtain point of interests from each character by contour method or connected component in a document image. Contour or connected component methodology depends on the arrangement of characters in the document image. If the characters over lap then the contour or connected component methods fail to recognize the boundary points of the individual characters.

The main purpose of [1,2] this paper is not to obtain the tight structure, but to obtain the area Voronoi for segmenting different entities in the document image without giving any importance to the blank space. The segmentation method is applied in a hierarchical methodology from paragraph to character level by removing the superfluous edges. The superfluous edges are removed based on setting the threshold manually.

The main purpose of [3,4,5] is to obtain the structure of the or shape of the entity This results in failure of obtaining the tight structure or the precise structure of the entity. The precise structure is a tight structure, in which the structure retrieves the shape of the entity. This method first extracts the bounding box around each character in the entity and voronoi tessellation is obtained for the boxes around the characters, which give a outline of the entity but not the structure of the entity. Bounding box [11,12] method itself is a bottom up approach which requires lot of computation and it also leads to miss grouping of characters which cannot be separately extracted. Hence to overcome these limitations, every high pixel in the document image is considered as the points of interest.

In the below figure the Voronoi tessellation obtained by considering point of interest on contour, connected component and every high pixel methods can be seen.

2.1 Point Voronoi Diagrams

Figure
host

Figure1: Input Image

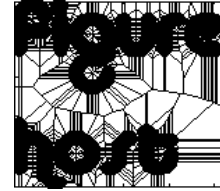


Figure2: Every High Pixel

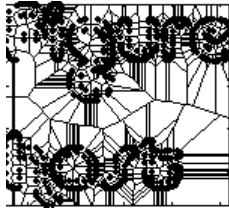


Figure3: Contour

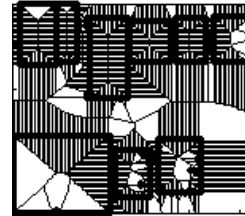


Figure4: Connected Component

Every high pixel in the document image conveys information about the structure of the entities in the document image. The structure of the document image depends on the arrangement of entities in the document image. An entity in a document image as a unique structure and this structure can be obtained only by an efficient extraction technique. To retrieve the exact structure of the entity every high pixel in the document image plays a vital role. If segmentation is only the main theme then considering ever high pixel as the points of interest becomes computational high and can relax by leaving the point of interest which lie in the unbounded region of Voronoi tessellation. The points which are in fuzzy state need not be considered, as segmentation does not concentrate on the blank space structure. Hence obtaining a tight structure is also domain dependent. The tight structure can also be obtained by ignoring the exterior point of interest which lies in the unbounded region, but the structural area will be one pixel less than the area of the exact structure of the entity.

The layout of the document image which consists of unconstrained tight structure of entities and the unconstrained tight structure of blank space can be obtained by this methodology which is more appropriate than any other methods and works successfully on any document images with Non-Manhattan layouts with arbitrary skew.

The spring force technique a new methodology which is implemented to obtain a tight structure of the entities in the document image is based the Voronoi features. The Voronoi features being extracted are shown on the hypothetical data below. Each feature obtained is based on the geometric analysis of the Voronoi tessellation obtained from applying the function Voronoi on the input image.

2.2 Hypothetical Data

For instance,

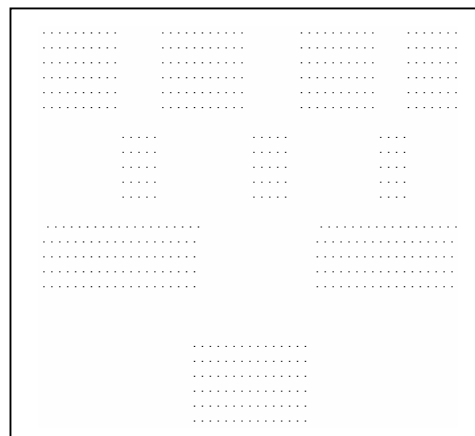


Figure5: Input Image

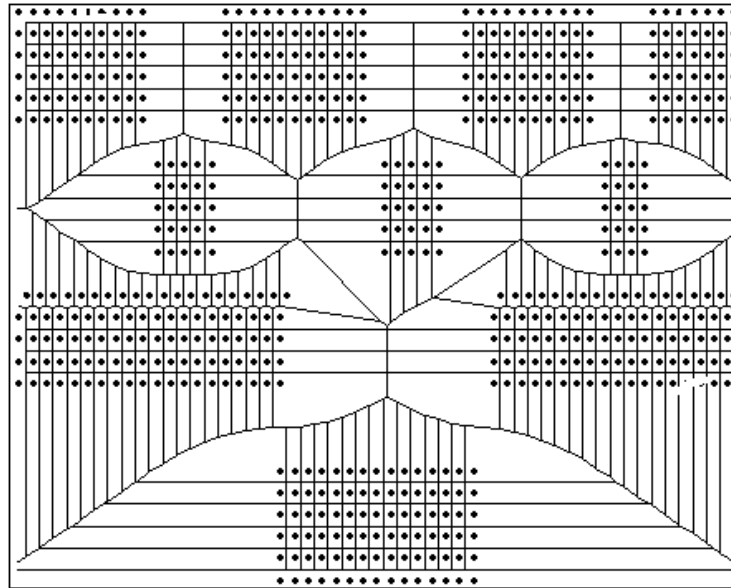


Figure 6: Voronoi Tessellation for hypothetical data

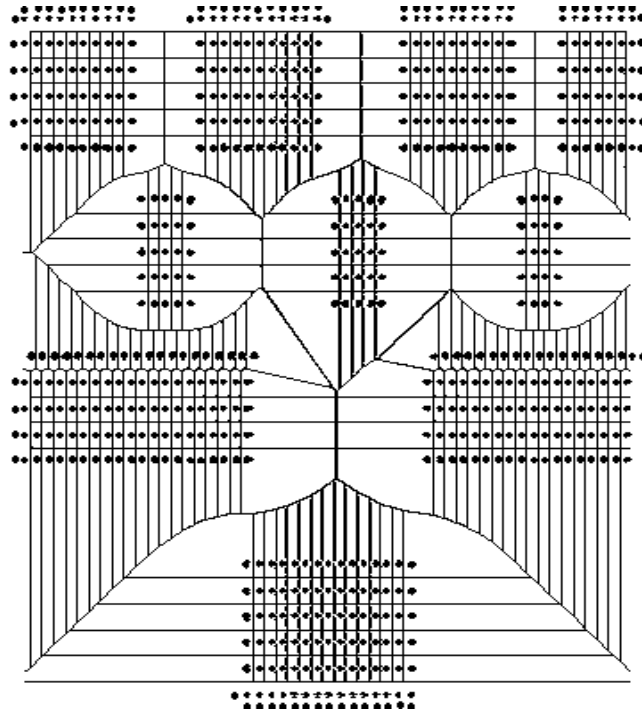


Figure 7: Output of spring force technique over Voronoi tessellation

The spring force technique grabs the exterior points and also the points of interest in the unbounded region. The dots in the next pixel width beside every point of interest can be observed in the out put of figure 7. This helps in obtaining the tight structure around entity in the document image.

2.3 Spring force technique Algorithm.

1. Calculate the Area of the bounded and unbounded cells using the function Polyarea [13]
2. Neighboring adjacency between the point of interest in both bounded and unbounded cells.
3. Neighboring adjacency is obtained by scanning the point of interest from left to right and bottom to top. Hence there are only 3 points for the boundary extraction.
4. The distance between the points of interest based on the neighboring adjacency.

- Dist = absolute (position of x coordinate – position of y coordinate);
- Neighboring adjacency with the conditions based on the ratio of the areas of polygons of two points of interests.


```

na1=A(f1);
na2=A(f2);
na3=A(f3);
rat=na1/na2;

```
 - To decide upon the boundary between the 3 points the threshold value set manually. Threshold is based on the knowledge of spacing between the entities.


```

// To consider the boundary edge on top or bottom of the point of interest
ratio=na1/na2;
if((dist1>threshold) | ratio>0.9 | ratio<1.125)
  if(ratio<0.9)
    // to draw the boundary above the pt of interest//
    plot(row_in(f2), (col_in(f2)+1),'.r');
  end
  if(ratio>1.125)
    // to draw the boundary below the pt of interest//
    plot(row_in(f1), (col_in(f1)-1),'.r');
  end
end.
//To consider the boundary edge to the left or right of the point of interest
if((dist11>threshold) | ratio1>0.9 | ratio1<1.125)
  if (ratio1<0.9)
    // to draw the boundary edge to the right of the point of interest//
    plot((row_in(f3)+1), col_in(f3),'.r');
  end
  if(ratio1>1.125)
    // to draw the boundary edge to the left of the point of interest//
    plot((row_in(f1)-1), col_in(f1),'.r');
  end
end

```

3. RESULTS

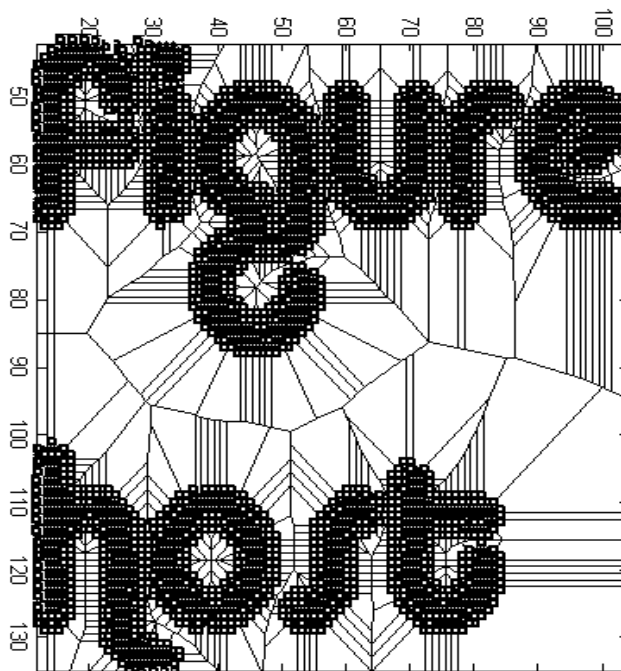


Figure 8: output of Spring Force technique over Voronoi Tessellation of a document image at a threshold of 0.9.

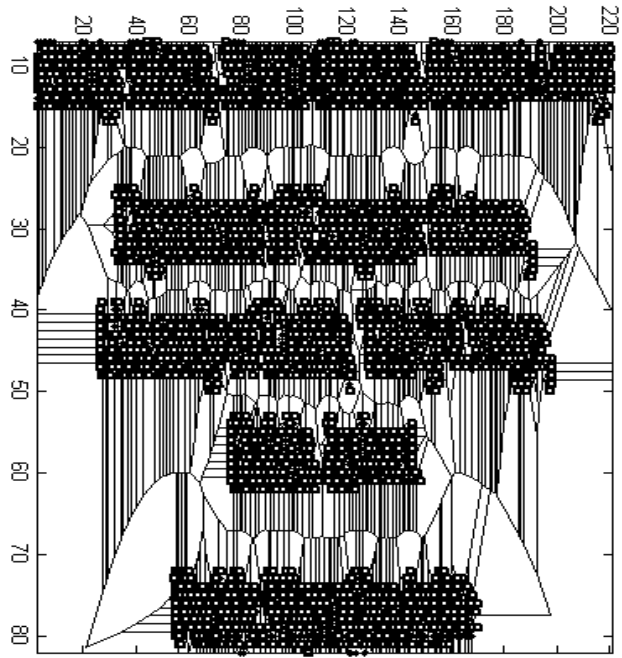


Figure 9: output of Spring Force technique over Voronoi Tesselation of a document image at a threshold of 0.6.

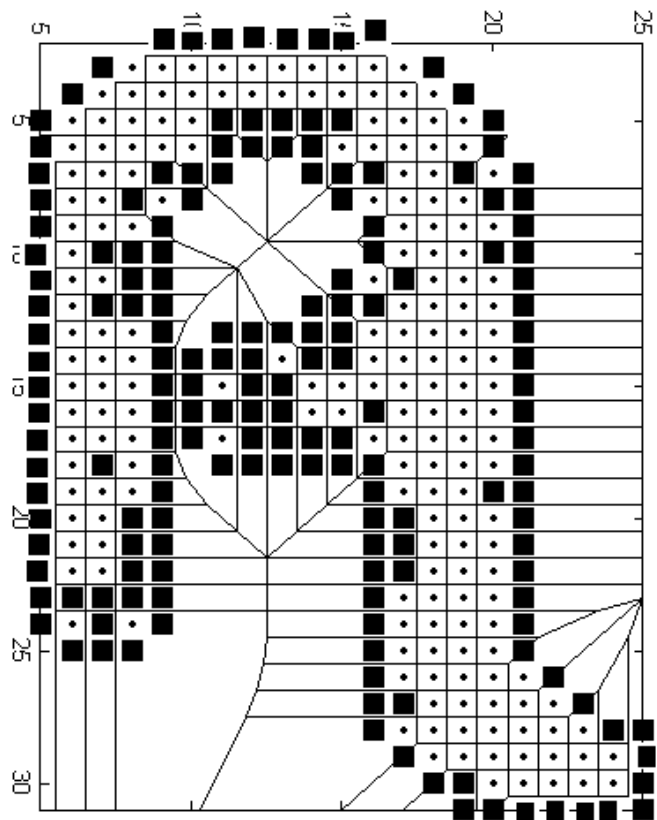


Figure 10: output of Spring Force technique over Voronoi Tesselation of a document image at a threshold of 1.2.

4. CONCLUSION

This paper presents the new technique of obtaining a tight unconstrained structure of the document entities using Voronoi tessellation. The method is computationally high, but efficient in extracting the structure of the entity. The efficiency rate is, it helps in obtaining the exact shape of the entity. Voronoi tessellation helps in obtaining the Voronoi features from the geometrical strategy. Hence features can be obtained easily by applying the geometrical characteristics such as area of the cell, distance between the two cells, bounded and unbounded ness of the cell etc. When the exterior points of interest are ignored, the structure of the entity is shrunk by one pixel width but the same structure can be obtained. But this not the case in other methods which take up only few points as points of interest, where they cannot have the exterior points of interest and at the mean time the internal structure of the document image to measure the sparsity and density of an entity would be lost. Sparsity refers to how many blank pixels are present within the entity and density refers to how many high pixels are present within the entity. Though every high pixel is considered as points of interest the Voronoi tessellation does not become complex as the pixel gap cannot be seen in densely populated regions of high pixels. Hence obtaining the internal structure becomes easier than any other methodologies. Future enhancement is to improve the analysis of the entity and blank space structure from the document image for classification and categorization of document images based on knowledge base structure analysis and to find the equivalence of document images without reading.

5. BIBILOGRAPHY

1. K.Kise, M.Iwata and K.Matsumoto , “On the Applications of Voronoi Diagrams to Page Segmentation” , Computer Vision and Image Understanding – 1998.
2. K.Kise, A.Sato and K.Matsumoto, “ Document Image Segmentation as selection of Voronoi Edges”, Proc. IEEE workshop on Document Image Analysis, San Jaun – 1997.
3. Yue Lu, C.L.Tan, “ Constructing Area Voronoi Diagram in Document Images”, Proc, 8th ICDAR’05.
4. Yue Lu, Zhe Wang, and C.L.Tan, “ Word Grouping in Document Images based on Voronoi Tessellation”, Document Image Analysis systems VI, Lecture notes in Computer Science, Vol.3161, pp.147-257, 2004.
5. Zhe Wang, Yue Lu, C.L.Tan, “ Word Extraction using Area Voronoi Diagram”, [www.comp.nus.edu.sg/~tancl/ Papers/CVPR03/DiarWangVoronoi.pdf](http://www.comp.nus.edu.sg/~tancl/Papers/CVPR03/DiarWangVoronoi.pdf)
6. Sahana D Gowda and P.Nagabhushan, “Page Segmentation and rule based Identification for technical Journal pages”, ICCAR, PP 633-639, Dec 2005.
7. L.O’ Gorman and r. Kasturi, document image analysis IEEE CS press, 1995.
8. A.K.Jain and B.Yu, “Document Representation and its application to page decomposition”, IEEE trans. PAMI, vol 20, no 3,pp 294-308, March 1998.
9. R.M. Haralick, “Document Image Understanding: Geometric and logical layout,” proc. IEEE, conf.CVPR, pp 385-390, 1994.
10. T.Saitoh, T.Yamaai and M.Tachikawa, “ Document Image segmentation and layout analysis”, IEICE Trans. Information and Systems, vol E77-D, no.7, pp 778-784, July 1994.
11. Jaekyu Ha & Robert M.Haralick, “Document Page decomposition by the bounding Box Projection Technique”, IEEE, ICDAR 1995.
12. Jaekyu Ha & Robert M.Haralick, “ Recursive X-Y Cut using bounding box of connected components”, IEEE, third ICDAR 1995.
13. MATLAB 7.