Word Extraction Using Area Voronoi Diagram

Zhe Wang, Yue Lu, Chew Lim Tan
Department of Computer Science, School of Computing
National University of Singapore, Kent Ridge, Singapore 117543
{wangzhe,luy,tancl}@comp.nus.edu.sg

Abstract

A method of word extraction based on the area Voronoi diagram is presented in this paper. Firstly, connected components are generated from the input image. Secondly, noise removal is performed including a special symbol detection technique to find some types of special symbols lying between words. Thirdly, base on the area Voronoi diagram, we select appropriate Voronoi edges which separate two neighboring connected components. Finally, words are extracted by merging the connected components based on the Voronoi edge between them. The result generated by this method is satisfactory with the ability to correctly group words of different size, font and arrangement. Experiments show that the proposed method achieves a high accuracy.

1. Introduction

In the field of image processing, the problem of text extraction from an image document remains an important issue. Many applications such as map interpretation, news articles search from microfilms, and referencing system for digitized manuscripts require text extraction.

Many works on text extraction from document images have been reported previously. Fletcher[1] described a method that uses information from each connected component in a mixed text-graphic document. In [6], Park introduced the 3D neighborhood graph model which can group words in inclined lines, intersecting lines, and even curved lines. Sobottka[7] proposed an approach to automatically extract text from colored books and journal covers. Tan[8] gave a method using irregular pyramid structure. The uniqueness of this algorithm is its inclusion of strategic background information in the analysis.

In recent years, there are many document layout analysis algorithms in the literature about applying Voronoi diagram to layout analysis. For instance, Xiao and Yan[10] described a method of text region extraction using Delaunay tessellation. Kise[4] employed area Voronio diagram to perform page segmentation. Wang[9] gave out a method to segment characters connected to graphics.

These methods focused on page segmentation by using the features of entire document page. They are not suitable for word segmentation, because the character size and intercharacter gap are different from word to word, even within the same document page. Furthermore, there exist different orientations of words in a document page. Unlike page segmentation, for the issue of word segmentation, the local information including the descriptive features of the image elements and the relative positions should be taken into account.

In this paper, we present an approach to extracting word objects from document images using the area Voronoi diagram. The area Voronoi diagram enables us to obtain neighbor relations between connected components and voronoi edges efficiently. Based on the neighbor relations, the task of word extraction is easily done by selecting appropriate Voronoi edges separating connected components in the same word and then merging those components. For this purpose, we define two characteristic features: the minimum distance of Voronoi edge and the minimum distance of connected component. We would rely on them to examine each Voronoi edge locally, and then judge whether we should select it to perform character merging or not. Experimental result on real document images shows that more than 98% of words are successfully extracted.

2. Image Preprocessing

An eight-neighbor connected component analysis algorithm is first applied to the input binary image to produce a set of connected components. Each of the connected components is bounded by a rectangular box with the coordinates of the upper left and the lower right corners of the box.

A connected component could be portion of a character or characters that are touching with each other. It is

Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'03)
1063-6919/03 $ 17.00 © 2003 IEEE

noteworthy to mention that if one connected component is encompassed by another one, they can be merged straight away because they undoubtedly belong to the same character. In addition, those with a very small area are considered as noise and thereby removed, while those with a large area are probably graphics or tables and thereby eliminated as well.

Next, we merge the overlapping connected components. If two connected components are overlapped with each other, they are merged to be one connected component. Here, we refer to these connected components as elements. As a result, all of the elements in the document image are non-overlapping, as shown in Figure 1.

(a) (b)

Figure 1. Merge overlapped components

Then, we detect two types of special symbols using a ruled-based algorithm without involving the recognition of any symbols. The first one is for the special symbols which are elongated horizontally such as dash ('-') or tilde ('∼'). The second one is for the vertically elongated symbols such as various kinds of parentheses '{', '}', '[', ']', '(', or ')'. Some special symbols are illustrated in Figure 2. Detection of the two types of special symbols is performed by a rule-based algorithm as follows:

Rule 1 For a connected component with width W_i and height H_i respectively, if it satisfies both the following two conditions, it is regarded as a special symbol of the first type:

(1) $W_i > 2 \times H_i$

(2) $H_i < T_1$, where T_1 is 30% of the median height among all connected components.

Rule 2 On the other hand, if one connected component satisfies the following four conditions, it is regarded as a special symbol of the second type:

(1) $H_i > 2 \times W_i$

(2) $D_i < 0.75 \times W_i \times H_i$, where D_i is the number of black pixels of the connected component

(3) The upper and lower parts of the connected component are symmetric and the left and right parts are not symmetric

3. Voronoi Diagrams

For completeness, we we give a brief review of the definitions of Voronoi diagrams. For further detail, readers can refer to [5]. Let $G = \{g_1, g_2, \cdots, g_n\}$ be a set of non-overlapping elements in the two-dimensional

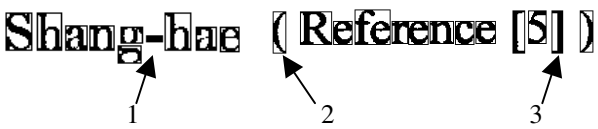


Figure 2. Special symbols

Euclidean plane, and let $d(p, g_i)$ be the Euclidean distance between a point p and an element g_i defined as

$$d(p, g_i) = \min_{q \in g_i} d(p, q),$$

where q is a point in g_i . Then the Voronoi region $V(g_i)$ and the area Voronoi diagram $V(G)$ are defined as

$$V(g_i) = \{p | d(p, gi) \leq d(p, g_j), \forall j \neq i\}$$

$$V(G) = V(g_1), \cdots, V(g_n)$$

The boundaries of Voronoi regions are called Voronoi edges. A Voronoi point indicates a point where Voronoi edges come in contact.

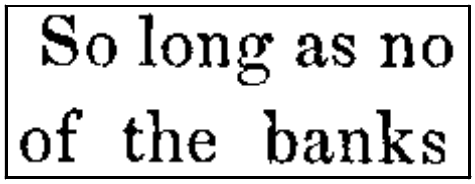
Figure 3 demonstrates the steps of constructing the area Voronoi diagram. By applying labeling and border following to the document image in Figure 3(a), the connected components are obtained as in Figure 3(b). Using these components as the generators, the area Voronoi diagram are constructed as shown in Figure 3(c).

4. Word Extraction

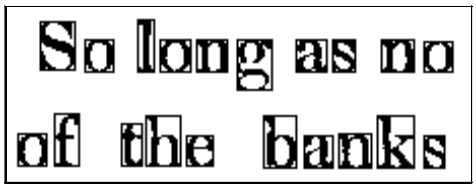
As shown in Figure 3(c), Voronoi edges lie between any adjacent connected components. In other words, every word component is represented as a set of Voronoi regions which are adjacent with one another. The process of word segmentation is, therefore, considered to be the selection of the Voronoi edges which separate two connected components potentially in the same word and then merge these two into one word component. To this end, we need criteria for selecting appropriate Voronoi edges from the area Voronoi diagram.

4.1. Features for Selection

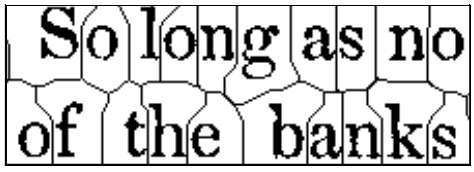
We attempt to select those Voronoi edges which are on the space between characters in order to merge the characters into words. For this purpose, we employ two characteristic features: the minimum distance of a Voronoi edge and the minimum distance of a connected component.



(a) Original image



(b) Connected components



(c) Area Voronoi diagram

Figure 3. Construction of the area Voronoi diagram

4.1.1 Minimum distance of Voronoi edge

In general, gaps between characters are narrower than those between words or text lines. From this viewpoint, we employ the minimum distance defined as follows. Let $E = \{l_1, \cdots, l_m\}$ be a Voronoi edge between two connected components g_1 and g_2, where l_i is a point in E. For each point l_i , we define $d(l_i, g_1)$ and $d(l_i, g_2)$ to be the minimum Manhattan distance from l_i to the borders of g_1 and g_2, respectively. Then, the minimum distance $d(E)$ is defined by

$$md_e(E) = \min_{1 \le i \le m} (\min(d(l_i, g_1), d(l_i, g_2)))$$

4.1.2 Minimum distance of connected component

On the area Voronoi diagram, every connected component C is enclosed by a few Voronoi edges, $\overline{E} = \{E_1, E_2, E_3, \cdots\}$. According to the definition of 4.1.1, let $M = \{m_1, m_2, m_3, \cdots\}$ to be the corresponding minimum distance value, in which m_i is the minimum distance of element E_i. We define the least value in M to be the minimum distance of this connected component:

$$md_c(C) = \min(m_i)$$

4.2. Selection of Voronoi Edges

To extract word objects, our task now becomes selecting the voronoi edges that lie between characters within words, followed by the merger of the characters that separate the corresponding characters.

By observing the area Voronoi diagram, we can find that most of the edges that should be selected have vertical trend, i.e. it is more or less parallel to y-axis. If an edge has horizontal trend (parallel to x-axis), it is most likely an inter-line boundary.

This criterion is effective, but exception does exist. For case of character i on the area Voronoi diagram, there is an edge between the dot and the lower part, and this edge should be selected although its trend is horizontal. This problem will be further discussed later.

4.2.1 Minimum distance measure

Most of the Voronoi edges in between text lines are eliminated by vertical trend judgment. Now the task we are facing is how to tell if one edge is a word boundary or not. In Fig.5, the edge with a pointing arrow has vertical trend but it should not be selected since it is a boundary of words "are" and "so". If it is wrongly selected, "are" and "so" would be merged into one single word, which is obviously an error.

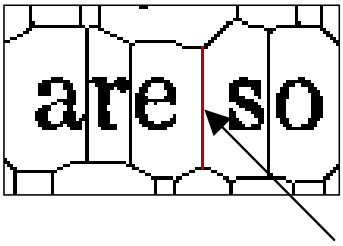


Figure 4. Voronoi edge between words

To solve this problem, we need to utilize the two important feature described in section 4.1, namely minimum distance of a Voronoi edge and minimum distance of a connected component. We set the following rule:

For a Voronoi edge E with minimum distance $md_e(E)$. We name the two elements separating E to be C_1 and C_2. Let $md_c(C_1)$ and $md_c(C_2)$ to be the minimum distance of C_1 and C_2 respectively. If $md_e(E) \le 2 \times \min(md_c(C_1), md_c(C_2))$, then we take E as an edge that we should select.

4.2.2 Punctuation detection

Punctuations should be excluded from word grouping. Comma and full stop are the two most commonly used

punctuations. For each comma or full stop, it is considered as a connected component on the area Voronoi diagram. By observation we can always find a Voronoi edge which is in between of punctuation and another character. In Figure 5, comma and character "e" are separated by an arrow pointing edge. The same thing happens to full stop case.

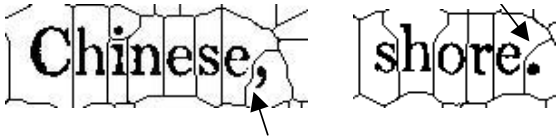


Figure 5. Comma and Full stop in area Voronoi diagram

Suppose a connected component, which is suspected to be a comma or full stop, has width W_p, height H_p, top boundary value T_p, bottom boundary value B_p and left boundary value L_p. And its neighboring character has width W_c, height H_c, top boundary value T_c, bottom boundary value B_c and right boundary value R_c.

If condition (1), (3), (4) and (5) are satisfied, it is regarded as a comma, and if condition (2), (4) and (5) are satisfied, it is regarded as a full stop:

1. 2.5*area of comma $<$ area of neighbor character
2. 5*area of comma $<$ area of neighbor character
3. $B_p > B_c + 0.25 * H_p$
4. $T_p > T_c + 1/4 * H_c$ and $L_p > R_c$
5. The ratio of vertical and horizontal distance between the centroids of punctuation and its neighboring character is inside [0.2, 1.2]

4.2.3 Special treatment of "i"

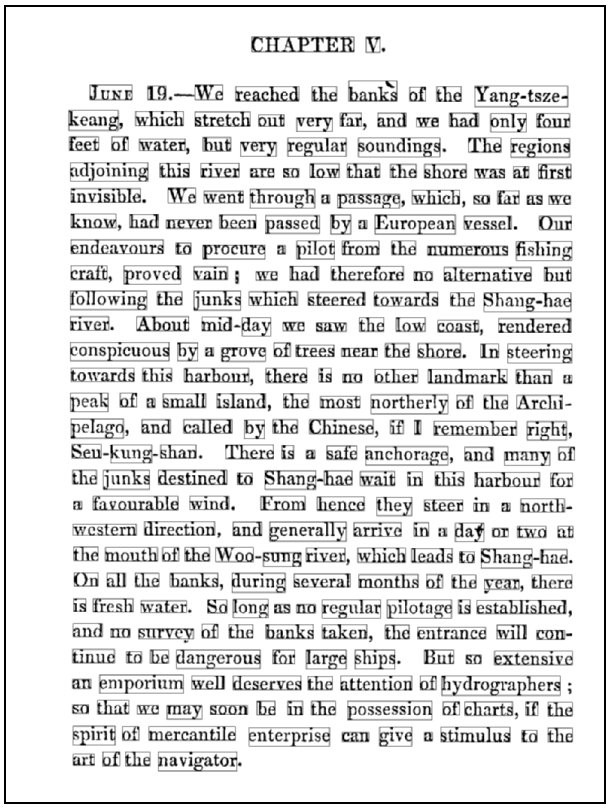
To prevent erroneous merger from different text lines, we do not allow the merging of connected components if one of them is above the other. However, character i is an exceptional case, since it has a dot on the top part. Therefore, we must pay additional attention to detect it, and recover a complete character i. The following five properties are summarized for this purpose:

1. The dot component is similar to a right rectangle
2. 4*area of dot $<$ area of lower part
3. Dot component is above the lower part
4. The range of dot component along x-axis is inside that of the lower part.
5. The ratio of width and height of lower part cannot exceed 0.7

5. Experiments

The algorithm described in this paper has been tested on more than 100 document images that were obtained from books, journals and student theses. Each document has more than 250 characters and is scanned as black and white image. The accuracy of word segmentation is tabulated in Table 1. An encouraging accuracy of over 98% has been achieved.

Figure 6 shows an example of word segmentation in which the segmented word entities are bounded using rectangle boxes. Almost all the words are correctly extracted, and most of commas and full stops are successfully detected. Figure 7 gives another example, in which the words in title and text body are correctly extracted from the image of UW document image database.



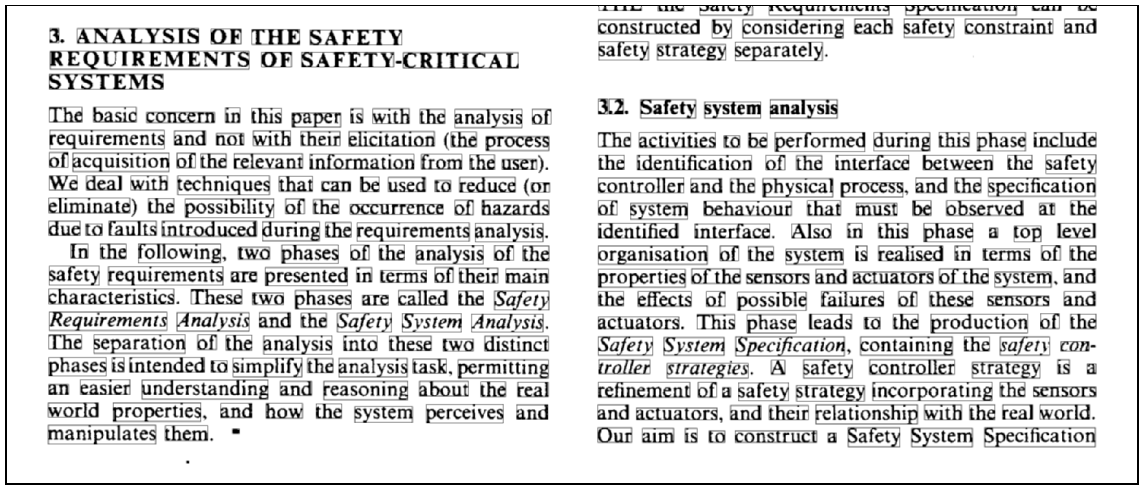
CHAPTER V.

JUNE 19.—We reached the banks of the Yang-tsze-keang, which stretch out very far, and we had only four feet of water, but very regular soundings. The regions adjoining this river are so low that the shore was at first invisible. We went through a passage, which, so far as we know, had never been passed by a European vessel. Our endeavours to procure a pilot from the numerous fishing craft, proved vain; we had therefore no alternative but following the junks which steered towards the Shang-hae river. About mid-day we saw the low coast, rendered conspicuous by a grove of trees near the shore. In steering towards this harbour, there is no other landmark than a peak of a small island, the most northerly of the Archipelago, and called by the Chinese, if I remember right, Seu-kung-shan. There is a safe anchorage, and many of the junks destined to Shang-hae wait in this harbour for a favourable wind. From hence they steer in a north-western direction, and generally arrive in a day or two at the mouth of the Woo-sung river, which leads to Shang-hae. On all the banks, during several months of the year, there is fresh water. So long as no regular pilotage is established, and no survey of the banks taken, the entrance will continue to be dangerous for large ships. But so extensive an emporium well deserves the attention of hydrographers; so that we may soon be in the possession of charts, if the spirit of mercantile enterprise can give a stimulus to the art of the navigator.

Figure 6. Example 1: Scanned book

6. Conclusions

This paper has reported a method based on area Voronoi diagram to extract words. On the diagram, every connected component is surrounded by Voronoi edges, in another words, every edge separates two neighboring components.



3. ANALYSIS OF THE SAFETY REQUIREMENTS OF SAFETY-CRITICAL SYSTEMS

The basic concern in this paper is with the analysis of requirements and not with their elicitation (the process of acquisition of the relevant information from the user). We deal with techniques that can be used to reduce (or eliminate) the possibility of the occurrence of hazards due to faults introduced during the requirements analysis.

In the following, two phases of the analysis of the safety requirements are presented in terms of their main characteristics. These two phases are called the *Safety Requirements Analysis* and the *Safety System Analysis*. The separation of the analysis into these two distinct phases is intended to simplify the analysis task, permitting an easier understanding and reasoning about the real world properties, and how the system perceives and manipulates them.

THE the Safety Requirements Specification can be constructed by considering each safety constraint and safety strategy separately.

3.2. Safety system analysis

The activities to be performed during this phase include the identification of the interface between the safety controller and the physical process, and the specification of system behaviour that must be observed at the identified interface. Also in this phase a top level organisation of the system is realised in terms of the properties of the sensors and actuators of the system, and the effects of possible failures of these sensors and actuators. This phase leads to the production of the *Safety System Specification*, containing the *safety controller strategies*. A safety controller strategy is a refinement of a safety strategy incorporating the sensors and actuators, and their relationship with the real world. Our aim is to construct a Safety System Specification

Figure 7. Example 2: UW document image

Table 1. Accuracy of word extraction

	Book	Journal	Thesis	Total
No. of words originally	26,732	11,293	14,005	52,030
No. of words extracted	26,366	11,106	13,769	51,241
Accuracy(%)	98.63	98.34	98.31	98.48

Therefore, the task of word extraction is simply to select appropriate edges and merge *closer* components that are separated by these edges. We proposed quite a few efficient rules for edge selection. The result has proven our proposed technique with the ability to correctly segment word objects from document images.

Acknowledgements

This research is jointly supported by the Agency for Science, Technology and Research, and Ministry of Education of Singapore under research grant R-252-000-071-112/303.

References

[1] L.A. Fletcher and R. Kasturi, A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, pp. 910-918, 1988.

[2] H. Goto and H. Aso, Extracting Curved Text Lines Using Local Linearity of Text Line, Int. J. Doc. Anal. Recognition 2: 111-119, 1999.

[3] S.H. Kim, C.B. Jeong, H.K. Kwag and C.Y. Suen, Word Segmentation of Printed Text Lines Based on Gap Clustering and Special Symbol Detection, International Conference on Pattern Recognition, Quebec City, Canada, Aug. 11-15 2002.

[4] K. Kise, A. Sato, and M. Iwata, Segmentation of Page Images Using the Area Voronoi Diagram, Computer Vision and Image Understanding, Vol. 70, No. 3, pp. 370-382, June 1998.

[5] A. Okabe, B. Boots, K. Sugihara, Spatial tessellations: Concepts and applications of Voronoi disgrams, 1992.

[6] H.C. Park, S.Y. Ok, Y.J. Yu and H.G. Cho, A word extraction algorithm for machine-printed documents using a 3D neighborhood graph model, Int. J. Doc. Anal. Recognition 4: 115-130, 2001.

[7] K. Sobottka, H. Kronenberg, T. Perroud and H. Bunke, Text extraction from colored book and journal covers, Int. J. Doc. Anal. Recognition 2: 163 - 176, 2000.

[8] C.L. Tan, P.O. Ng: Text extraction using pyramid. Proc. Pattern Recognition 31(1):63-72, 1997.

[9] Yalin Wang, Ihsin T. Phillips and Robert Haralick, Using Area Voronoi Tessellation to Segment Characters Connected to Graphics, Fourth IAPR International Workshop on Graphics Recognition (GREC2001), Kingston, Ontario, Canada, Sep. 2001.

[10] Y. Xiao and H. Yan, Text Region Extraction in a Document Image Based on the Delaunay Tessellation, Pattern Recognition, Vol. 36 (2003), No. 3, pp. 799-809, 2003.