Newspaper Headlines Extraction from Microfilm Images

Qing Hong Liu Chew Lim Tan
National University of Singapore
Kent Ridge Singapore 117543

ABSTRACT

Automatic indexing is important for a digital library to provide digitized manuscripts of old document images and their electronic text. As an essential step in creating such a system, this paper discusses the issue of extracting headlines from old newspaper microfilms. Most research on document layout analysis has largely assumed relatively clean images. However microfilm images of old newspapers present a challenge. Such images are usually insufficiently illuminated and considerably dirty. To overcome the problem we propose a new effective method for separating characters from noisy background since conventional threshold selection techniques are inadequate to deal with these kinds of images. A Run Length Smearing Algorithm (RLSA) is applied in the headline extraction. Experiment shows that our approach has improved the recall, precision and combined rates.

1. Introduction

There are two main functions of a digital library [1]: one is to allow a quick lookup of the desired information and the other is to enable the printing of the necessary information. Newspaper is one kind of information provided by libraries. As opposed to conventional libraries where old newspapers are persevered on microfilms, digital libraries can digitize microfilm images to allow access by the public either on the web or through other electronic means. As there is a huge amount of newspapers images, locating a news article in the image database is extremely difficult. A digitized index based on main headlines and subcategories of the old newspaper microfilm images is an instant aid to locate the needed information. Users can narrow down their search by choosing from such available headlines and subcategories. As such, an automatic indexing system based on headline extraction from digitized microfilm images has been proposed for our National Library. This task can be divided into two main parts: image analysis and pattern recognition. The first part is to extract headline areas from the microfilm images and the second part is to apply Optical Character Recognition (OCR) on the extracted headline areas and turn them into the corresponding texts. This paper focuses on the first part. Headline extraction is done through a layout analysis of the microfilm images. Most research on layout analysis has largely assumed relatively clean images. Old newspapers' microfilm images, however present a challenge. Many of the microfilm images archived in Singapore National Library are old newspapers dated over a hundred years ago. Figure 1 shows one of the microfilm images.

2. Outline of Processing

There are various preprocessing methods to deal with documents that are assumed noisy. Hideyyuki et al [2] and James L. Fisher [3] proposed hybrid methods that work on carefully captured images. O'Gorman[4] uses connectivity preserving method to binarize the document images. However as the microfilm images in our collection are very dim and considerably dirty, their methods do not work well on microfilm images. Separating text and graphics from their background is usually done by thresholding. If the text sections have enough contrast with the background, they can be thresholded directly using methods proposed so far [3][5]. However there are considerable overlaps of gray level ranges between the text, graphics and the background, in our image data. Thus applying conventional threshold techniques directly to the image will result in poor segmentation. In view of this, we propose a three stage preprocessing method, namely, histogram transformation, adaptive binarization and noise filtration. Histogram transformation is used to improve the contrast ratio of the microfilm images without changing the histogram distribution for the later preprocessing. An adaptive binarization method is then applied for converting the original image to binary image with reasonable noise removal. The last step in the preprocessing is applying a kFill filter [6] to remove the pepper and salt noise to get considerably noise-free images. To extract the headlines of the newspaper microfilm images, a Run Length Smearing Algorithm (RLSA) is applied.

3. Three Stage Preprocessing

Among a number of techniques, threshold problems can mainly be solved by three categories of methods: statistics, entropy and moment preserving method. Our research is

1051-4651/02 $17.00 © 2002 IEEE

to find the best synergy of exiting methods to solve the problem of microfilm images. We adopt the three stages preprocessing on the newspaper microfilm images.

3.1 Histogram Transformation

The gray levels of the microfilm images of the microfilms occupy only a small range. To increase the visual contrast of the image, we adopt a linear transformation that entails stretching the nonzero input intensity range, $x \in [x_{min}, x_{max}]$ to an output intensity range $y \in [0, y_{max}]$ to take advantage of the full dynamic range. As a result, the interval is stretched to cover the full range of the gray level without altering the image appearance. Fig 2 shows the result of thresholding without histogram transfer. In contrast, Figs 3 to 5 show significant improvements with histogram transformation.

3.2 Adaptive Binarization

Although the principle of binarization is simple, a number of conditions can make binarization difficult. Poor image contrast makes it difficult to resolve foreground from the background; the corresponding histogram peaks will tend to overlap. Further more, spatial non-uniformity is even worse in the background intensity in that the images appears light at some areas while dark at some other areas in one single image.

The variety of conditions under which binarization is performed requires a different approach. Thus a local adaptive binarization technique is applied to counter the effects of non-uniform background intensity values. The method divides the original image into subimages. Depending on the non-uniform degree of the original image, the image size of $N \times M$ is divided into $N / n \times M / m$ subimages of size $n \times m$. In each sub-image, we do a discriminant analysis [7] to determine the optimal threshold within each sub-image. Sub-images for which the measure of class separation is small are said to contain only one class; no threshold is calculated for these sub-images and the threshold is taken as the average of thresholds in the neighboring sub-images. Finally the sub-image thresholds are interpolated among sub-images for all pixels and each pixel value is binarized with the respect to the threshold at pixel.

Compared with several other local adaptive threshold methods [8] this method is parameter independent and also computationally inexpensive.

3.3 Noise Reduction

Binarized images often contain a large amount of salt and pepper noise and James L. Fisher's [3] study shows that noise adversely affects image compression efficiency and degrades OCR performance. A more general filter, called kFill [6] is designed to reduce the isolated noise and noise on contours up to a selected limit in size.

4. Headline Extraction

In spite of the fact that there are a number of block segmentation techniques [3][5][9] available, one of the major deficient is the large computational requirement. As there is a huge amount of images in the image store and each image size is relatively large, these methods are not quite efficient for our microfilm images.

While utilizing knowledge of the layout and structure of document results in a simple, elegant and efficient page decomposition system, such knowledge is not readily available in our present project. This is because the entire microfilm collection at the National library spans over 100 years of newspapers where layouts have changed over all these years. There are thus a great variety of different layouts and structures. Therefore we adopt our rule-based method to find the headline adaptively without any layout analysis. The following approach is then proposed that is not dependent on any particular layout.

4.1 Run Length Smearing

The approach aims at segmenting the document into regions for further processing. We begin with the run length smoothing algorithm (RLSA)[10] to produce the RLSA image.

Different values of the horizontal and vertical threshold H and V yield different types of RLSA images. A very small H value simply smoothes individual characters. Slightly larger values of H can put individual characters together to form a word (word level) and little larger values of H can smear a sentence (processing in a sentence level). An even larger value of H can merge the sentence together. Similar comments hold for the magnitude of V. Appropriate choice of the values of the thresholding parameters H and V is a thus important

4.2 Labeling

Connected components are detected in the RLSA image via a row and run tracking method. [5] Based on the RLSA image, we then establish boundaries around and calculate statistics of the regions using connected components. A rule based block classification is used for classifying each block into text, horizontal /vertical line, graphics and picture type.

Let the upper-left corner of an image block be the origin of coordinates. The following measures are applied on each block

- Minimum and maximum x and y coordinates of a block ($x_{min}, y_{min}, x_{max}, y_{max}$);

- Number of white pixels corresponding to the block of the RLSA image (N)
- Height of each block, H= y_{max} -y_{min};
- Width of each block, W= x_{max} -x_{min};
- Density of white pixels in a block, D=N/HW;

A newspaper headline often contains characters of a size and font, which are different from the text. Let H_m and W_m denote the height and width of the most likely height of connected components to be determined by thresholding. Let D_a represent the minimum density of the connected components, and d_1, d_2, d_3, d_4, e_1, e_2, e_3, and e_4 be appropriate tolerance coefficients.

- Rule1: if, the block H> e_1 H_m then it belongs to text paragraph or graphics.
- Rule2: if e_1 H_m<H< e_2 $H_{m\ and}$ e_3W_m<W< e_4 W_m then it belongs to the title or text block.
- Rule3: under rule2: if d_1 D_a<D< d_2 D_a then it belongs to the title
- Rule4: under rule2: if d_3 D_a<D< d_4 D_a then it belongs to the text block.

Rule1 is to distinguish the graphics and connected text block from the image. Rule2 is used to remove horizontal and vertical lines. Rule3 and rule4 are to differentiate the headline from the text block.

5. Experimental Observation and Discussion

Our approach has been tested on 40 images of old newspaper microfilms with the width ranging from 1800 to 2400 pixels and the height ranging from 2500 to 3500 pixels. We use three different approaches to pre-process the images before applying the headline extraction discussed in section 4. The three approaches are (1) Conventional binarization based on a normal threshold [11]; (2) Histogram transformation discussed in section 3.1 followed by Otsu method [12]. (Niblack method [13] was also experiemented but the results were not as good as that of Otsu method); and (3) The three-stage image preprocessing method described in section 3. Figs 2 to 5 show the binarization results, including the result from Niblack method.

To measure the performance of the headline extraction, we use precision and recall rates defined below:

$$\text{Precision}=\frac{\text{no. of headline characters correctly extractec}}{\text{no. of characters (total) extracted}}$$

$$\text{Recall}=\frac{\text{no. of headline characters correctly extracted}}{\text{actual no. of headline characters in the document page}}$$

Table 1 shows the experimental results of the 40 tested images. The following are some important observations:

- The method of histogram transformation has significantly improved the final output despite the extremely poor and non-uniform illumination of the microfilm images and present good results.
- Adaptive binarization approach is effective for extracting text from noisy background, even though the histogram of the image is unimodal and the gray levels of the text overlap with the background..
- The three-stage pre-processing has achieved a significant improvement in headline extraction. Fig 6 shows the final result. The average recall and precision rates are 84.4% and 89.7% as compared to those of 76.5% and 84.9% for Otsu method and 68.5% and 79.0% for the conventional approach.
- Our headline extraction method works well even with skewed images of up to 5°. Figs 7 and 8 show the examples
- The recall rate of the headline is not always 100% as shown in table 1, because some of the Headlines are too close to vertical or horizontal lines and were thus regarded as graphical or text blocks.

References

[1] M. Imai et al: Design of a digital University Libaray, Proc. Of the International Symposium on digital Libraries 1995,pp119-124, Tsububa, Japan, Aug. 1995.

[2] Hideyuki Negishi etc."Character Extraction from Noisy Background for an automatic Reference System" ICDAR 1999, pp143-146,

[3] James L.Fisher, Stuart C.Hinds .etc "A Rule-Based System for Document Image Segmentation" IEEE Trans. Pattern Analysis Machine Intelligence, 567-572,1990

[4] L.O'Gorman "Binarization and multithresholding of Document images using Connectivity" CVGIP:Graphical Model and Image Processing Vol.56, No.6 Nov, pp494-506, 1994

[5] L.A. Flecher and R.Kasturi," A robust algorithm for text string separation from mixed text/graphics images" IEEE Trans. Pattern Anal. Machine Intel. Vol. 10 no. 6,pp910-918, Nov 1988

[6] L.O'Gorman "Image and document processing techniques for the RightPages Electronic library system" in Pro.11th Int. Conf. Pattern recognition(ICPR) Aug 1992, pp260-263.

[7] Y. Liu, R. Fenrich, S.N. Srihari, "An object attribute thresholding algorithm for document image binarization, ICDAR '93, Japan, 1993, pp. 278-281.

[8] O.D. Trier, T.Taxt "Evaluation of Binarization Methods for Document Images" IEEE Trans. Pattern Analysis and Machine Intelligence Vol.17, pp312-315, March 1995.

[9] F.M. Wahl, K.Y. Wong, and R.G.Casey "Block segmentation and text extraction in mixed text / image documents", Computer vision, Graphics, Image Processing, vol 20, pp375-390, 1982.

[10] K.Y.Wong, R.G.Casey, and F.M.Wahl, "Document analysis system", IBM J.Res.Develop, vol.26,no.6, pp647-656, Nov.1983.

[11] T. Pavlidis: Algorithms for graphics and image processing, Computer Science Press, 1982.

[12] Otsu, N., "A threshold selection Method from Gray-Level Histogram" IEEE Trans. System, Man and Cybernetics, Vol. SMC-9, No. 1,pp.62-66, Jan 197

[13]W. Niblack, "An Introduction to Image Processing ", Prentice-Hall, Englewood Cliff, NJ, pp. 115-116, 1986.

Image No.	Recall Rate (%)			Precision Rate (%)		
	Conv.	Otsu	Our	Conv.	Otsu	Our
1	98.2	100	100	95.2	100	100
2	50.2	70.5	80.8	96.5	100	100
3	78.2	80.3	90.5	93.1	95.3	97.1
4	80.2	84.4	86.1	89.7	90.8	91.2
5	75.3	80.2	87.5	90.1	92.5	94.6
6	60.5	79.7	89.1	92.1	93.1	95.2
7	53.3	60.9	79.8	78.2	80.1	80.5
8	73.4	78.5	81.9	82.5	83.7	85.2
9	80.7	83.4	91.2	87.4	89.6	93.4
10	56.5	62.6	70.5	76.6	79.8	82.3
11	72.4	77.1	84.5	80.4	84.2	88.6
12	79.6	80.4	92.4	81.8	89.9	95.8
13	51.2	70.5	77.6	67.7	72.4	86.9
14	60.3	70.9	78.5	70.1	80.3	85.4
15	78.2	80.0	85.1	85.9	86.6	89.7
16	69.5	74.3	82.3	77.1	85.4	89.8
17	58.4	68.9	73.2	64.3	77.8	80.5
18	74.7	80.6	83.5	80.3	83.7	87.9
19	81.6	84.7	90.2	90.4	90.4	95.4
20	75.1	80.5	84.8	83.5	88.1	90.1
21	68.9	73.3	80	75.6	79.1	88.2
22	60.8	65.4	75.6	79.2	80.3	89.3
23	76	81.2	86.3	81.8	84.9	92.7
24	78.5	86.4	90.1	87.4	90.5	92.8
25	62.3	70.6	79.3	71.7	80.3	84.5
26	55.8	62.7	71.9	71.2	79.2	82.3
27	72.4	80.7	89.5	83.3	89.9	92.6
28	69.4	78.6	87.3	74.5	85.6	89.5
29	66.1	76.4	88.5	70.4	80.8	88.9
30	53.2	69.7	79.7	61.6	79.5	85.8
31	66.7	77.9	80.4	71.9	80.7	90.3
32	70.3	78.1	90.2	81.5	89.6	92.1
33	62.4	79	85.6	77.2	80.0	86.2
34	70.2	83.5	89.9	77.1	87.3	93.7
35	58.3	61.2	77.9	64.8	72.7	80.3
36	67.8	72.3	85.2	73.2	78.4	89.6
37	78.3	84.5	87.0	83.4	87.8	92.4
38	72.6	78.9	84.8	84.1	85.4	91.3
39	60.4	73.5	85.2	73.5	84.9	87.3
40	78.2	84.1	88.4	86.4	89.6	94.3
Ave.	68.5	76.5	84.4	79	84.9	89.7

Table1 Experiment results of three methods

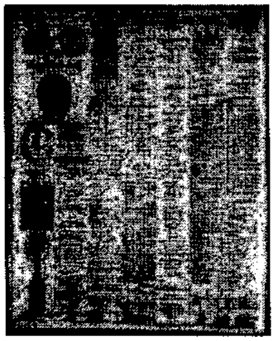

Fig 1 Image of a microfilm

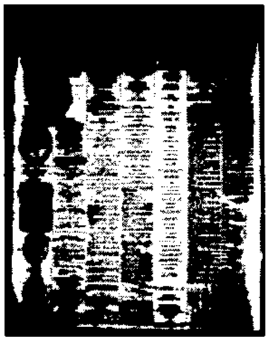

Fig 2 Conventional binarization using theshold=115

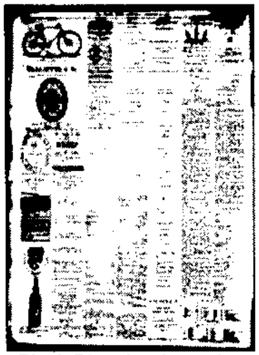

Fig 3 Binarized result by Otsu method after histogram transformation

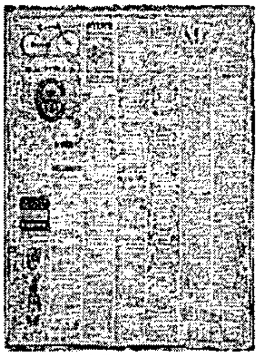

Fig 4 Binarized result by Niblack method after historgram transformation

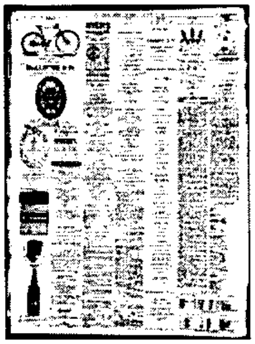

Fig 5 Binarized result by our method

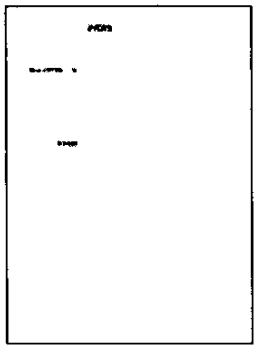

Fig 6 Extracted headlines from fig 1

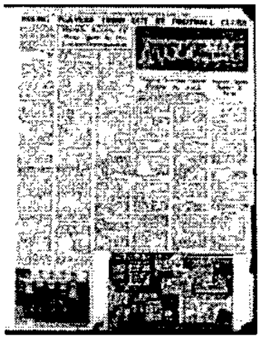

Fig 7 Skewed microfilm image

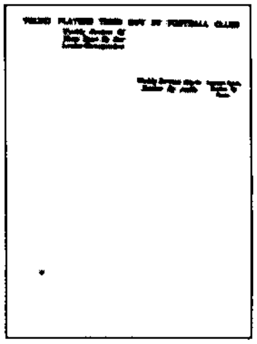

Fig 8 Extracted headlines from fig 7

Acknowledgments:

This research is supported by Agency for Science, Technology and Research and Ministry of Education, Singapore, under Grant R252-000-071-112/303. The permission from National Library Board, Singapore for the use of microfilm images is gratefully acknowledged.