Automatic Indexing of Newspaper Microfilm Images

Qing Hong Liu and Chew Lim Tan

School of Computing
National University of Singapore
Kent Ridge Singapore 117543

Abstract. This paper describes a proposed document analysis system that aims at automatic indexing of digitized images of old newspaper microfilms. This is done by extracting news headlines from microfilm images. The headlines are then converted to machine readable text by OCR to serve as indices to the respective news articles. A major challenge to us is the poor image quality of the microfilm as most images are usually inadequately 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 then applied to the headline extraction. Experimental results confirm the validity of the approach.

1 Motivation

Many libraries archive old issues of newspapers in microfilm format. Locating a news article among a huge collection of microfilms proves to be too laborious and sometimes impossible if there is no clue to the date or period of the publication of the news article in question. Today many digital libraries digitize microfilm images to facilitate access. However, the contents of the digitized images are not indexed and thus searching a news article in the large document image database will still be a daunting task. A project was thus proposed in conjunction with our Singapore National Library to do an automatic indexing of the news articles by extracting headlines from digitized microfilm images to serve as news indices. 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 for indexing. 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 dated as old as over a hundred years ago. Figure 1 shows one of the microfilm images. Adequate pre-processing of the images is thus necessary before headline extraction can be carried out. To extract the headline of the newspaper microfilm images, a Run Length Smearing Algorithm (RLSA) is applied.

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 365–375, 2002.

The remainder of the paper is organized as follows: Section 2 will describe the preprocessing for image binarization and noise removal. Section 3 will discuss our method for headline extraction. Section 4 will present our experimental results. Finally we outline some observations and conclude the paper.

2 Precrocessing

Various preprocessing methods to deal with noisy document images have been reported in the literature. Hybrid methods as proposed by Hideyyuki et al [1] and James L. Fisher [2] require adequate capturing of images. O'Gorman [3] uses connectivity-preserving method to binarize the document images. These methods were tried but found inadequate for our microfilm images because of their poor quality with low illumination and excessive noise. 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 [2][4]. However in view of the considerable overlaps of gray level ranges between the text, graphics and the background, in our image data, poor segmentation results after trying theses methods. Thus, we experimented three stages of preprocessing, 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 of the images 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 [5] to remove the pepper and salt noise to get considerably noise-free images.

2.1 Histogram Transformation

Because of the narrow range of the gray scale values of the microfilm image content, a linear transformation is adopted to increase the visual contrast. This entails the stretching of the nonzero input intensity range, $x \in [x_{min}, x_{max}]$ to an output intensity range $y \in [0, y_{max}]$ by a linear function to take advantage of the full dynamic range.

As a result, the interval is stretched to cover the full range of the gray level and the transformation is applied without altering the image appearance. Figure 2 shows the result of thresholding without histogram transfer. In contrast, figures 3 and 4 show the significant improvements with histogram transformation.

2.2 Adaptive Binarization

While the idea of binarization is simple, poor image quality can make binarization difficult. Because of the low contrast of microfilm images, it is difficult to resolve foreground from the background. Furthermore, 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.

A local adaptive binarization technique is thus applied to counter the effects of non-uniform background intensity values. Here we divide the original image into subimages. Depending on the degree of the non-uniformity of the original image, the image size of N × M is divided into N / n × M / m subimages of size n × m. In each sub-image, we do a discriminant analysis [6] to determine the optimal threshold within each sub-image. Sub-images with small measures of class separation 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.

Let $P(i)$ be the histogram probabilities of the observed gray values i, where i ranges from 1 to I:

$$P(i) = \frac{\#\{(r,c) \mid Gray - value(r,c) = i\}}{\# R \times C}. \tag{1}$$

where $R \times C$ is the spatial domain of the image. Let σ_W^2 be the weighted sum of group variances, that is, the within-group variance. Let $\sigma_1^2(t)$ be the variance of the group with values less than or equal to t and $\sigma_2^2(t)$ be the variance of the group with values greater than t. Let $q_1(t)$ be the probability for the group with values less than or equal to t and $q_2(t)$ be the probability for the group with values greater than t. Let $\mu_1(t)$ be the mean for the first group and $\mu_2(t)$ be the mean for the second group. Then the within-group variance σ_W^2 is defined by

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t). \tag{2}$$

where

$$q_1(t) = \sum_{i=1}^{t} P(i). \tag{3}$$

$$q_2(t) = \sum_{i=t+1}^{I} P(i). \tag{4}$$

$$\mu_1(t) = \sum_{i=1}^{t} iP(i) / q_1(t) \cdot \tag{5}$$

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 P(i) / q_1(t) \cdot \tag{6}$$

$$\mu_2(t) = \sum_{i=t+1}^{I} iP(i) / q_2(t) \cdot \tag{7}$$

$$\sigma_2^2(t) = \sum_{i=t+1}^{I} [i - \mu_1(t)]^2 P(i) / q_2(t) \cdot \tag{8}$$

The best threshold t can be determined by a sequential search through all possible values of t to locate the threshold t that minimizes $\sigma_w^2(t)$ .Compared with several other local adaptive threshold methods [7] this method is parameter independent and also computationally inexpensive.

2.3 Noise Reduction

Binarized images often contain a large amount of salt and pepper noise and James L. Fisher's [2] study shows that noise adversely affects image compression efficiency and degrades OCR performance. A more general filter, called kFill [5] is designed to reduce the isolated noise and noise on contours up to a selected limit in size. The filter is implemented as follows:

In a window of size $k \times k$, the filling operations are applied in the raster-scan order. The interior window, the core, consists of $(k – 2) \times (k – 2)$ pixels and $4(k –1)$ pixels on the boundary that is referred to as the neighborhood. The filling operation sets all values of the core to ON or OFF, depending on the pixel values in the neighborhood. The criteria to fill with ON (OFF) requires that all core pixels to be OFF (ON) and is dependent on three variables *m*, *g* and *c* of the neighborhood. For a fill value equal to ON (OFF), *m* equals to the number of ON (OFF) pixels in the neighborhood, *g* denotes the number of connected groups of ON pixels in the neighborhood, *c* represents the number of corner pixels that are ON (OFF). The window size *k* determines the values of *m* and *c*.

The noise reduction is performed iteratively. Each iteration consists of two sub-iterations, one performing ON fills and the other OFF fills. When no filling occurs in the consecutive sub-iterations, the process stops automatically.

Filling occurs when the following conditions are satisfied:

$$(g = 1) \text{ AND } [(m > 3k - 4) \text{ OR } \{(m = 3k - 4) \text{ AND } (c = 2)\}] \tag{9}$$

where $(m > 3k\text{-}4)$ controls the degree of smoothing: A reduction of the threshold for *m* leads to enhanced smoothing; $\{(m = 3k\text{-}4)$ AND $(c = 2)\}$is to ensure that the corners less than 90° are not rounded. If this condition is left out, greater noise can be reduced but corners may be rounded. $(g = 1)$ ensures that filling does not change connectivity. If this condition is absent, a greater smoothing will occur but the number of distinct regions will not remain constant. The filter is designed specifically for binary text to remove noise while retaining text integrity, especially to maintain corners of characters.

3 Headline Extraction

Headline extraction requires proper block segmentation and classification. Looking for possible existing methods for our current application, we found the work by Wahl *et al* [8] who placed the graphics as the same category with pictures. On the other hand, Fisher *et al* [2] made use of computation of statistical properties of connected components. In yet another approach, Fletcher and Kasturi [4] applied a Hough transform to link connected components into a logical character string in order to discriminate them from graphics, which is relatively independent of changes in font, size and the string orientation of text. Because of the huge amount of large images in the image store, the above methods prove to be too computationally expensive for our microfilm images.

At an early stage in the document understanding process, it is essential to identify text, image and graphics regions, as a physical segmentation of the page, so that each region can be processed appropriately. Most of these techniques for page segmentation rely on prior knowledge or assumptions about the generic document layout structure and textual and graphical attributes, e.g. rectangularity of major blocks, regularity of horizontal and vertical spaces, and text line orientation, etc. 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 in the image database. To address the above problems, we try to do away with the costly layout analysis. To do so, we adopt a rule-based approach to identify headlines automatically. The following approach is proposed that is not dependent on any particular layout.

3.1 Run Length Smearing

Run length smoothing algorithm (RLSA) [9] is used here to segment the document into regions. It entails the following steps: a horizontal smoothing (smear), a vertical smoothing, a logical AND operation, and an additional horizontal smoothing. In the first horizontal smoothing operation, if the distance between two adjacent black pixels (on the same horizontal scan line) is less than a threshold H, then the two pixels are joined by changing all the intervening white pixels to black ones, and the resulting image is stored. The same original image is then smoothed in the vertical direction, joining together vertically adjacent black pixels whose distance is less than a threshold V. This vertically smoothed image is then logically ANDed with the horizontally smoothed image, and the resulting image is smoothed horizontally one more time, again using the threshold H, to produce the RLSA image.

Different RLSA images are obtained with different values of H and V. A very small H value simply smoothes individual characters. Increasing the value of H can put individual characters together to form a word (word level) and further increase 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 thus important. They are found empirically through experimentation.

3.2 Labeling

Using a row and run tracking method [4], the following algorithm detects connected components in the RLSA image:

Scan through the image pixel by pixel across each row in order:

- If the pixel has no connected neighbors with the same value that have already been labeled, create a new unique label and assign it to that pixel.
- If the pixel has exactly one label among its connected neighbors with the same value that has already been labeled, give it that label.
- If the pixel has two or more connected neighbors with the same value but different labels, choose one of the labels and remember that these labels are equivalent.

Resolve the equivalence by making another pass through the image and labeling each pixel with a unique label for its equivalence class. Based on the RLSA image, we can 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 one of these types, namely, text, horizontal /vertical lines, graphics and picture.

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)

The following features are adopted for block classification:

- 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, $D=N/(H \times W)$;

Newspaper headlines often contain characters of a certain font and a larger size, 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, which can 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 $e_3 W_m < W < e_4 W_m$ 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 aims to distinguish the graphics and connected text block from the image while Rule2 is used to remove horizontal and vertical lines. Rule3 and rule4 are to differentiate the headline from the text block.

4 Experimental Result

40 images of old newspaper microfilms with the width ranging from 1800 to 2400 pixels and the height ranging from 2500 to 3500 pixels were tested in our experiments. We used three different approaches to pre-process the images before applying the headline extraction discussed in section 3. The three approaches are (1) Conventional binarization based on the normal threshold [10]; (2) Histogram transformation discussed in section 2.1 above followed by Otsu method [11](Various adaptive binarization methods including Niblack method [12] were also attempted and our final choice is Otsu Method.); and (3) The three-stage image preprocessing method described in section 3. Table 1 shows the experimental results in terms of precision and recall rates defined below:

$$\text{Precision} = \frac{\text{No. of headline characters correctly extracted by the system}}{\text{No. of characters (headline or non - headline) extracted by the system}} \qquad (10)$$

$$\text{Recall} = \frac{\text{No. of headline characters correctly extracted by the system}}{\text{Actual no. of headline characters in the document page}}\,. \qquad (11)$$

Table 1. Experiment results of three methods

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

5 Conclusion and Discussion

We propose a document analysis system that extracts news headlines from microfilm images to do automatic indexing of news articles. The poor image quality of the old newspapers presented us several challenges. First, there is a need to properly binarize the image and to remove the excessive noise present. Second, a fast and effective way of identifying and extracting headlines is required without the costly layout analysis in view of the huge collection of images to be processed.

From the experiments that we have conducted, we have the following 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 area from noisy background, even though the histogram of the image is unimodal and the gray levels of the text parts overlap with the background.
- Our headline extraction method works well even with skewed images of up to 5°. Fig 6 and 7 show the examples.
- The pre-processing step has achieved a significant improvement in headline extraction. 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% for conventional approach.
- The recall rate of the headline is not always 100% in the result shown in table 1. Because some of the headlines are too close to the vertical or horizontal line and were thus regarded as the graphical or text block.

Acknowledgements. This project is supported in part by the Agency for Science, Technology and Research (A*STAR) and Ministry of Education, Singapore under grant R-252-000-071-112/303. We thank National Library Board, Singapore, for permission to use their microfilm images.

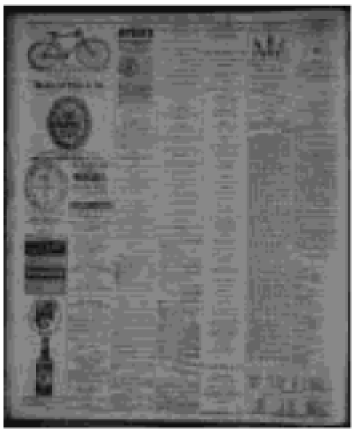

Fig. 1. One image of microfilm (T=115)

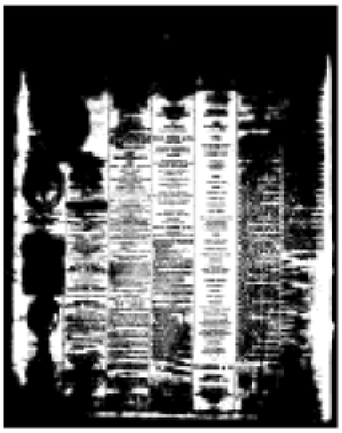

Fig. 2. Result of thresholding image of Fig. 1

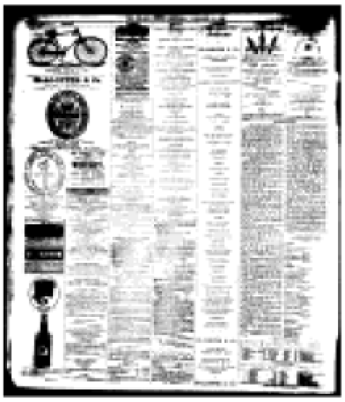

Fig. 3. Result of thresholding image of Fig 1 using Otsu method after

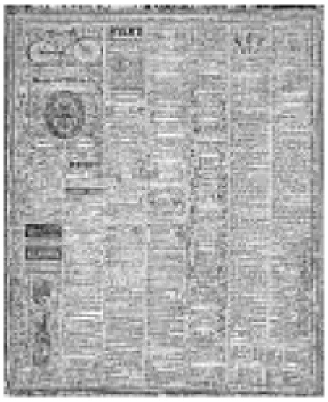

Fig. 4. Result of thresholding image of Fig 1 using Niblack method after histogram transformation

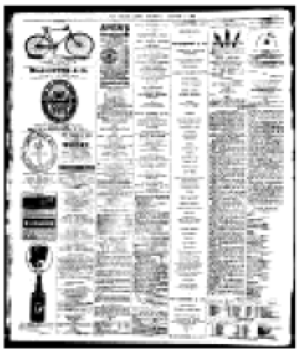

Fig. 5. Result of thresholding image of Fig 1 using our method after histogram transformation

Fig. 6. Extracted headline of the microfilm

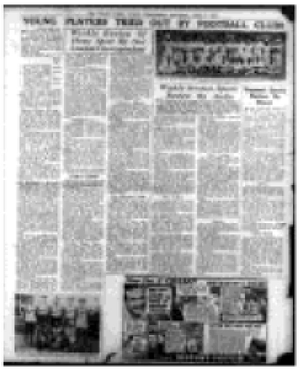

Fig. 7. Skewed newspaper microfilm image image

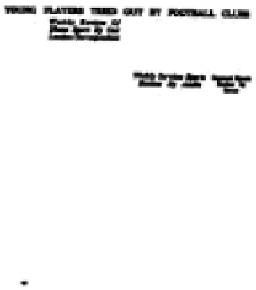

Fig. 8. Detected headline of skewed image Fig. 7

References

1. Hideyuki Negishi etc. “Character Extraction from Noisy Background for an automatic Reference System” ICDAR pp. 143–146, 1999
2. James L.Fisher, Stuart C.Hinds .etc “A Rule-Based System for Document Image Segmentation” IEEE Trans. Pattern Matching, 567–572,1990
3. L.O’Gorman “Binarization and multithresholding of Document images using Connectivity” CVGIP: Graphical Model and Image Processing Vol.56, No. 6 November, pp. 494–506, 1994
4. 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, pp. 910–918, Nov 1988
5. L.O’Gorman “Image and document processing techniques for the Right Pages Electronic library system” in Pro.11th Int. Conf. Pattern Recognition(ICPR) Aug 1992, pp. 260–263.
6. Y. Liu, R. Fenrich, S.N. Srihari, An object attribute thresholding algorithm for document image binarization, International Conference on Document Analysis and Recognition, ICDAR ’93, Japan, 1993, pp. 278–281.
7. M.A. Forrester, etc “Evaluation of potential approach to improve digitized image quality at the patent and trademark office” MITRE Corp., McLean, VA, Working Paper WP-87W00277, July 1987.
8. 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, pp. 375–390, 1982.
9. K.Y.Wong, R.G.Casey, and F.M.Wahl, “Document analysis system”, IBM J.Res.Develop, vol.26, no. 6, pp. 647–656, Nov.1983.
10. T. Pavlidis: Algorithms for graphics and image processing, Computer Science Press, 1982.
11. 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 1979
12. W.Niblack,”An Introduction to Image Processing”, Prentice-Hall, Englewood Cliff, NJ, pp. 115–116,1986.