IJDAR (2004) 6: 201–210
Digital Object Identifier (DOI) 10.1007/s10032-003-0111-2

International Journal on Document Analysis and Recognition IJDAR

Extraction of newspaper headlines from microfilm for automatic indexing

Chew Lim Tan[1], **Qing Hong Liu**[2]

[1] School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543

[2] Data Storage Institute, DSI Building, Engineering Drive 1, Singapore 117608

Received: 15 November 2002 / Accepted: 19 May 2003
Published online: 30 January 2004 – © Springer-Verlag 2004

Abstract. This paper proposes a document image analysis system that extracts newspaper headlines from microfilm images with a view to providing automatic indexing for news articles in microfilm. A major challenge in achieving this is the poor image quality of 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 this kind of image. A run length smoothing algorithm is then applied to the headline extraction. Experimental results confirm the validity of the approach.

Keywords: Headline extraction – Noise reduction – Histogram transformation – Run length smoothing

1 Motivation

Many libraries archive old issues of newspapers in the 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 publication date or period 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 the National Library of Singapore to provide 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 often done through a layout analysis of document images [7,8]. 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 the National Library are dated as old as over 100 years ago. Figure 1 shows one of the microfilm images. Adequate preprocessing of the images is thus necessary before headline extraction can be carried out. Another challenge presented to us is the variety of newspaper layouts that have changed over the years in the last 100 years of newspaper production. It is thus not possible to find a generic layout that works with microfilm images from different time periods. In fact, as our intention is mainly to extract prominent headlines to serve as news article indices, we propose a method that will extract headlines without the need for detailed layout analysis. To do so, a run length smoothing algorithm (RLSA) is applied.

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 Preprocessing

Various preprocessing methods to deal with noisy document images have been reported in the literature. Hybrid methods as proposed by Negishi et al. [5] and Fisher [1] require an adequate capture of images. O'Gorman [9] uses a connectivity-preserving method to binarize document images. We tested these methods but found them to be inadequate for our microfilm images because of the poor image quality with low illumination and excessive noise. Separating text and graphics from their background is usually done by thresholding. If the text sections have

Correspondence to: C.L. Tan (e-mail: tancl@comp.nus.edu.sg)

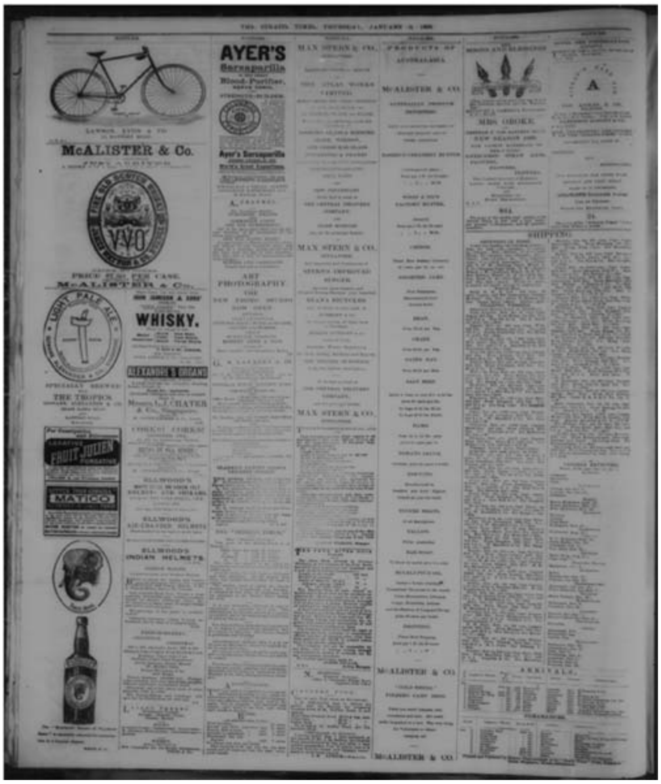

Fig. 1. A sample of newspaper microfilm image

enough contrast with the background, they can be thresholded directly using methods proposed previously [1,2]. However, because of the considerable overlap of gray-level ranges between text, graphics, and background in our image data, poor segmentation results when we try these methods. Thus we propose 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 later preprocessing. An adaptive binarization method is then applied to convert the original image to a binary image with reasonable noise removal. The last step in the preprocessing is to apply a kFill filter [9] to remove the pepper and salt noise to get considerably noise-free images.

2.1 Histogram transformation

Because of the narrow range of grayscale values of 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, Figs. 3 and 4 show the significant improvements with the histogram transformation.

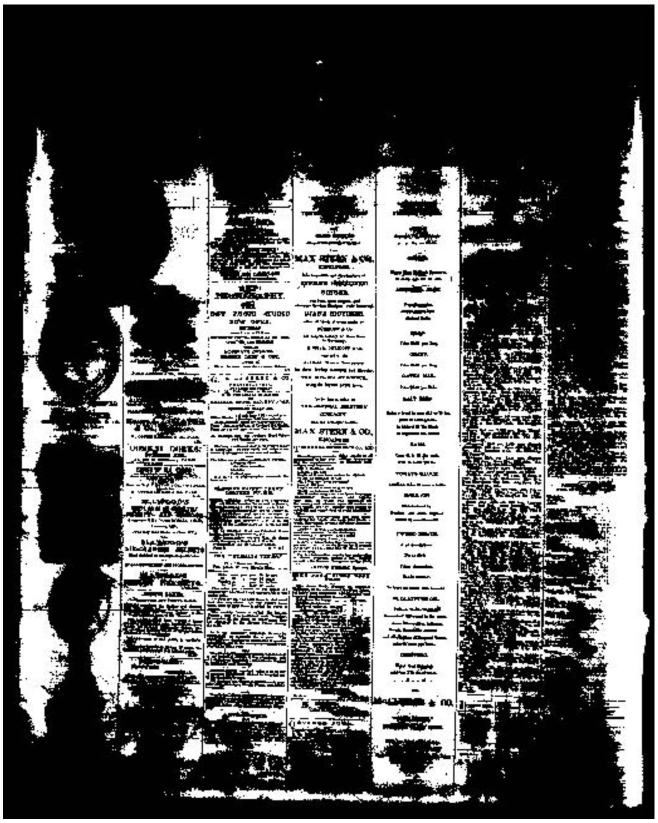

Fig. 2. Result of binarizing the Fig. 1 image with predetermined threshold ($T = 115$ based on 256 gray levels)

2.2 Adaptive binarization

While the idea of binarization is simple, poor image quality can make binarization difficult. Because of the low contrast of our microfilm images, it is difficult to resolve the foreground from the background. To deal with this problem, Otsu's method [11], a global adaptive binarization technique, is first explored. Otsu's method works by finding an optimal threshold that divides the pixels into two groups by maximizing the intergroup variance or minimizing the intragroup variance. While the method greatly improves the binarization result, the spatial nonuniformity in the intensity over the entire image presents another problem. In many cases, the image appears light in some areas but dark in other areas in a single image. Thus a global adaptive threshold found by Otsu's method may not give a perfect binarization for the entire image.

The above problem points to the need for a local adaptive binarization approach. To address this issue, Niblack's method [6], a local adaptive method that has been determined by [14] to be the best, is next explored as a possible candidate for our choice. Niblack's method works by varying the threshold over an image

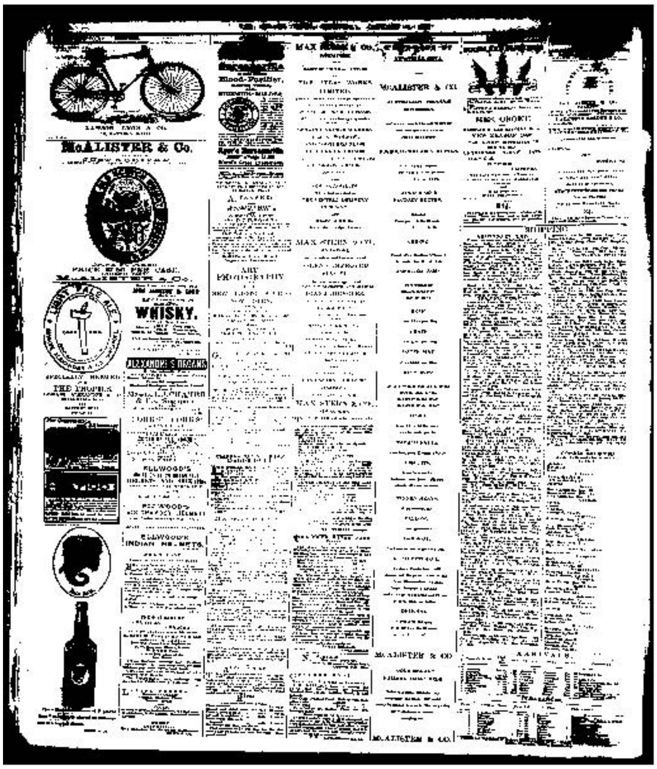

Fig. 3. Result of binarizing Fig. 1 image using Otsu's method after histogram transformation

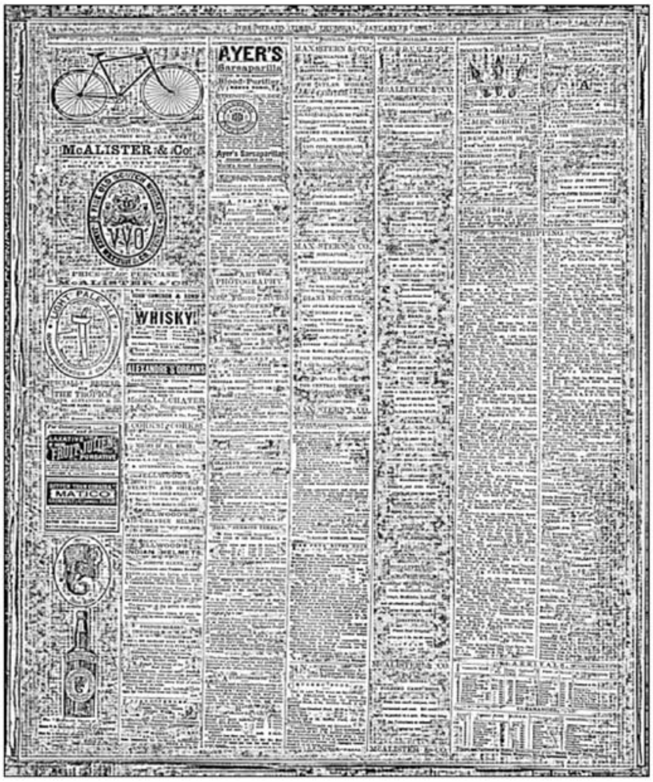

Fig. 4. Result of binarizing Fig. 1 image using Niblack's method after histogram transformation

based on the local mean, μ, and the local standard deviation, σ, computed in a small neighborhood (normally a window size of 15×15 is used) of each pixel. A threshold for each pixel at (x, y) is computed from $T(x, y) = \mu(x, y) + k.\sigma(x, y)$, where $\mu(x, y)$ and $\sigma(x, y)$ are the local mean and local standard deviation calculated in a window centered at (x, y), and k is a user-defined parameter and is negative in value. A major problem with Niblack's method is its sensitivity to the value of k for our images. It is difficult, if not impossible, to find a single k that works for all our test images. The other problem is the resultant large amount of pepper noise in the nontext areas, even if a proper k value is chosen.

In view of the above, the following local adaptive approach based on Otsu's method [11] is adopted: we first divide the original image into subimages. Depending on the degree of the nonuniformity 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 subimage, we do a discriminant analysis to determine the optimal threshold within each subimage. Subimages with small measures of class separation are said to contain only one class; no threshold is calculated for these subimages and the threshold is taken as the average of thresholds in the neighboring subimages. Finally, the subimage thresholds are interpolated among subimages for all pixels and each pixel value is binarized with respect to the threshold at the pixel.

Let $P(i)$ be the histogram probabilities of the observed gray values i, where i ranges from 1 to I, where I is the maximum gray value for the number of bits per pixel used:

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

where $G(r, c)$ is the gray value of the pixel at (r, c), R is the number of rows, and C is the number of columns. Let σ_W^2 be the intragroup variance, $\sigma_1^2(t)$ the variance of the group with gray values less than or equal to t, and $\sigma_2^2(t)$ the variance of the group with gray values greater than t. Further, let $q_1(t)$ be the probability for the group with gray values less than or equal to t, and $q_2(t)$ the probability for the group with gray values greater than t. Let $\mu_1(t)$ be the mean for the first group and $\mu_2(t)$ the mean for the second group. Then the intragroup variance σ_W^2 is defined as the following weighted sum:

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

where

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

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

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

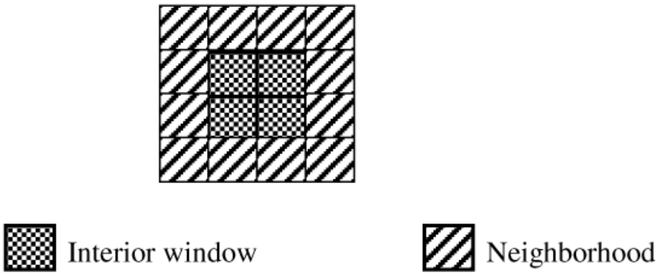

Fig. 5. Interior window and neighborhood in kFill filter

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

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

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

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 [3], 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. Fisher's [1] study shows that noise adversely affects image compression efficiency and degrades OCR performance. A more general filter, called kFill [9], is designed to reduce the isolated noise and noise on contours up to a selected size limit. The filter is implemented as follows:

In a window of size $k \times k$, the filling operations are applied in a raster-scan order. The interior window, the core, consists of $(k - 2) \times (k - 2)$ pixels and $4(k - 1)$ pixels on the boundary referred to as the neighborhood as shown in Fig. 5 for $k = 4$. The filling operation sets all values of the core to ON or OFF, depending on the pixel values in the neighborhood. The criterion to fill with ON (OFF) requires that all core pixels 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 the number of ON (OFF) pixels in the neighborhood, g denotes the number of connected groups of ON pixels in the neighborhood, and 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 subiterations, one performing ON fills and the other OFF fills. When no filling occurs in the consecutive subiterations, 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 $(m > 3k-4)$ controls the degree of smoothing: a reduction of the threshold for m leads to enhanced smoothing; $\{(m = 3k - 4)$ 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 to remove from binary text noise while retaining text integrity, especially to maintain corners of characters.

3 Headline extraction

Headline extraction requires proper block segmentation and classification. In searching for existing methods that may be applied to our current application, we found the work by Fisher et al. [1], who made use of the computation of statistical properties of connected components. On the other hand, Fletcher and Kasturi [2] applied a Hough transform to link connected components into a logical character string in order to discriminate them from graphics. The approach is relatively independent of changes in font, size, and string orientation of text. The above methods, however, have proved to be rather computationally expensive for our microfilm images.

Works directly involving newspaper headline extraction have also been studied. Niyogi and Srihari [7,8] made use of document layout analysis to find headlines in newspapers. As will be discussed later, the variety of newspaper layouts in our microfilm collection presents a problem. Takebe et al. [13] reported a method that extracts newspaper headlines mixed with some background design, a common feature found in many Japanese newspapers. This problem, however, is not present in our newspaper images.

At an early stage in the document understanding process, it is essential to identify text, image, and graphics regions as physical segmentations 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, 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 the 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; it is not dependent on any particular layout.

3.1 Run length smoothing

A run length smoothing algorithm (RLSA) [15] 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_d, then the two pixels are joined by changing all the intervening white pixels into 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_d. 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_d, to produce the RLSA image.

Different RLSA images are obtained with different values of H_d and V_d. A very small H_d value simply smooths individual characters. Increasing the value of H_d can put individual characters together to form a word (word level) and further increase of H_d can smear a sentence (processing at the sentence level). An even larger value of H_d can merge the sentence together. Similar comments hold for the magnitude of V_d. The appropriate choice of the values of the thresholding parameters H_d and V_d is thus important. They are found empirically through experimentation.

3.2 Labeling

Using a row-and-run tracking method [2], the following algorithm detects connected components in the RLSA image. Scan through the image pixel by pixel across each row in sequence:

– 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 the regions of connected components and calculate the statistics of the regions using the 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:

– The minimum and maximum x and y coordinates of a block ($x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$)
– The number of white pixels corresponding to the block of the RLSA image (N_w)

The following features are adopted for block classification:

– Height of each block, $H_b = y_{\max} - y_{\min}$
– Width of each block, $W_b = x_{\max} - x_{\min}$
– Density of white pixels in a block, $D = N_w/(H_b \times W_b)$

Newspaper headlines often contain characters of a certain font and of a larger size, which are different from the text. Let H_m and W_m denote the most likely height and width as the representative height and width of the connected components that can be determined by thresholding. Let D_a represent the minimum density of the connected components, and let d_1, d_2, d_3, d_4, e_1, e_2, e_3, and e_4 be appropriate tolerance coefficients.

– Rule 1: if the block's height $H > e_2H_m$, then the block belongs to a block of consecutive text paragraphs or a graphics block.
– Rule 2: if the block's height H is such that $e_1H_m < H < e_2H_m$ and $e_3W_m < W < e_4W_m$, then the block belongs to a title or a text block.
– Rule 3: under rule 2, if the block's density D is such that $d_1D_a < D < d_2D_a$, then the block belongs to a title block.
– Rule 4: under rule 2, if the block's density D is such that $d_3D_a < D < d_4D_a$, then the block belongs to a text block.

Rule 1 aims to identify a graphics block or a block of consecutive text paragraphs in the image, while Rule 2 serves to identify a smaller text block that could be a title or a single paragraph. Rule 2 will also remove horizontal and vertical lines. Rule 3 and 4 are to differentiate a headline from other text blocks.

For our experiment, microfilm images with different layouts and character sizes were used. Because the documents usually contain characters of a particular size and font that are in popular use for newspapers, the mean value of all the block heights approximates the most popular block height (H_p), and this can be computed automatically from the statistical features of the connected components. For each document, the mean value of height and the standard deviation S_d are derived from blocks of the most popular height $H_p.S_d$ can be computed by the following equation:

$$S_d = \sqrt{\frac{\sum_{i=1}^{N}(H_i - H_p)^2}{N_b - 1}} \qquad (10)$$

where N_b is the total number of blocks in a microfilm image. H_i is the height of each individual block.

Empirically, the most likely (H_m) text height is selected as one sixth of the most popular H_p. The ratio S_d/H_p is distributed between the range of 0.023 and 0.054 with an average of 0.038. For reliability, the tolerance of the text height is selected to be six times that of the average ratio, i.e., 0.23. Therefore, $e_1 = 1 - 0.23 = 0.77$ and $e_2 = 1 + 0.23 = 1.23$. The width tolerance parameters e_3 and e_4 are also derived in a similar way. These parameters are found to work over a wide range of microfilm images.

4 Experimental result

The parameters described in Sect. 3 were first manually set by visual inspection of the various spatial relationships. Over 60 microfilm newspaper pages from our National Library's collection were first experimented with to fine-tune the parameters. These newspaper pages were selected from a span of over 100 years with page width ranging from 1800 to 2400 pixels and page height ranging from 2500 to 3500 pixels. These selected images represent various layouts, various amounts of noise, various blurring of text lines, and a variety of symbols and text. With the parameters set as described in Sect. 3, another 40 images were chosen covering a similar spectrum of layouts and image quality for testing. To represent varying image quality of the 40 test images, the level of noise and the extent of image blurring were indicated as high, moderate, and low. Figure 1 is one of the 40 test images.

We used the following three different approaches to preprocess the images before applying the headline extraction discussed in Sect. 3.

(1) Conventional approach: this is a simple straightforward binarization using a predetermined threshold [12]. The result of binarizing the Fig. 1 image using this approach is shown in Fig. 2.
(2) Histogram transformation discussed in Sect. 2.1 above followed by Otsu's method [11], which is a global adaptive threshold discussed in Sect. 2.2. The result of binarizing the Fig. 1 image using the second approach is shown in Fig. 3. Preliminary experiments were earlier carried out to test Niblack's method [6], but this was found to produce excessive pepper noise in the nontext area. Niblack's method was thus later excluded from our experiment. Nevertheless, a sample output following Niblack's method is shown in Fig. 4.
(3) The method proposed in this paper, namely, the three-stage image preprocessing method described in Sect. 2 involving the histogram transformation, the local adaptive thresholding, and the kFill noise reduction. The result of binarizing the Fig. 1 image using the present method is shown in Fig. 6, with its final output shown in Fig. 7.

To measure the effectiveness of our method in extracting headlines, we visually inspected the final output and

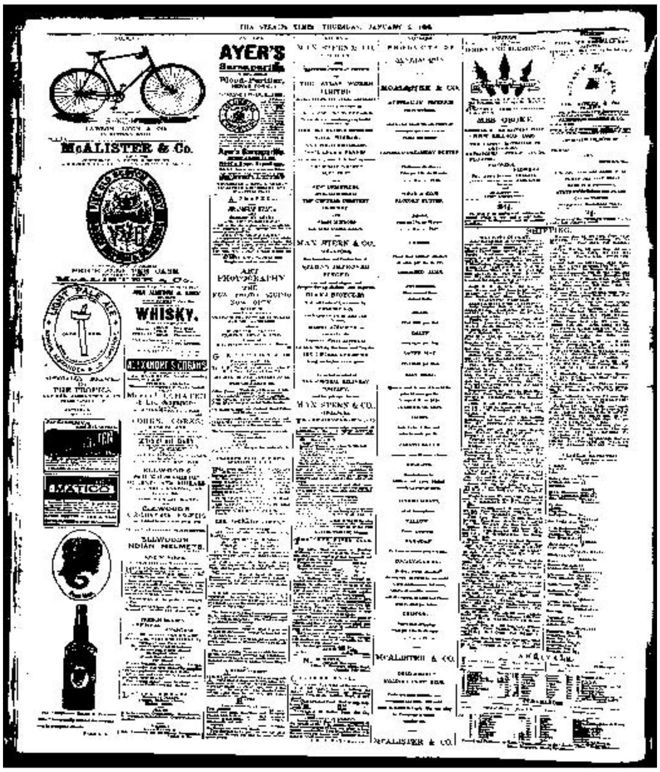
Fig. 6. Result of binarizing Fig. 1 image using the proposed three-stage preprocessing

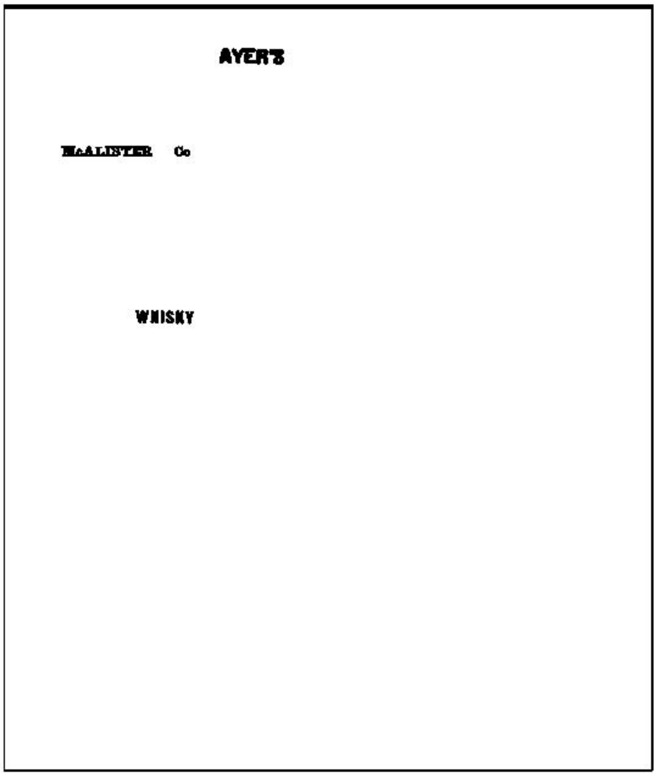

Fig. 7. Headlines extracted from Fig. 6 image

Table 1. Experimental results of three methods

Image no.	Image Degradation		Recall Rate			Precision Rate		
	Noise	Blurring	Conventional	Otsu	Present	Conventional	Otsu	Present
1	Low	Low	98.2	100	100	95.2	100	100
2	High	Moderate	50.2	70.5	80.8	96.5	100	100
3	Moderate	Low	78.2	80.3	90.5	93.1	95.3	97.1
4	Moderate	Moderate	80.2	84.4	86.1	89.7	90.8	91.2
5	Moderate	Moderate	75.3	80.2	87.5	90.1	92.5	94.6
6	High	Moderate	60.5	79.7	89.1	92.1	93.1	95.2
7	High	High	53.3	60.9	79.8	78.2	80.1	80.5
8	Moderate	Moderate	73.4	78.5	81.9	82.5	83.7	85.2
9	Moderate	Low	80.7	83.4	91.2	87.4	89.6	93.4
10	High	High	56.5	62.6	70.5	76.6	79.8	82.3
11	Moderate	Moderate	72.4	77.1	84.5	80.4	84.2	88.6
12	Moderate	Low	79.6	80.4	92.4	81.8	89.9	95.8
13	High	High	51.2	70.5	77.6	67.7	72.4	86.9
14	High	High	60.3	70.9	78.5	70.1	80.3	85.4
15	Moderate	Moderate	78.2	80.0	85.1	85.9	86.6	89.7
16	High	Moderate	69.5	74.3	82.3	77.1	85.4	89.8
17	High	High	58.4	68.9	73.2	64.3	77.8	80.5
18	Moderate	Moderate	74.7	80.6	83.5	80.3	83.7	87.9
19	Moderate	Low	81.6	84.7	90.2	90.4	90.4	95.4
20	Moderate	Moderate	75.1	80.5	84.8	83.5	88.1	90.1
21	High	Moderate	68.9	73.3	80.0	75.6	79.1	88.2
22	High	High	60.8	65.4	75.6	79.2	80.3	89.3
23	Moderate	Moderate	76.0	81.2	86.3	81.8	84.9	92.7
24	Moderate	Low	78.5	86.4	90.1	87.4	90.5	92.8
25	High	High	62.3	70.6	79.3	71.7	80.3	84.5
26	High	High	55.8	62.7	71.9	71.2	79.2	82.3
27	Moderate	Low	72.4	80.7	89.5	83.3	89.9	92.6
28	High	Moderate	69.4	78.6	87.3	74.5	85.6	89.5
29	High	Moderate	66.1	76.4	88.5	70.4	80.8	88.9
30	High	Moderate	53.2	69.7	79.7	61.6	79.5	85.8
31	High	Moderate	66.7	77.9	80.4	71.9	80.7	90.3
32	Moderate	Low	70.3	78.1	90.2	81.5	89.6	92.1
33	High	Moderate	62.4	79.0	85.6	77.2	80.0	86.2
34	Moderate	Low	70.2	83.5	89.9	77.1	87.3	93.7
35	High	High	58.3	61.2	77.9	64.8	72.7	80.3
36	High	Moderate	67.8	72.3	85.2	73.2	78.4	89.6
37	Moderate	Moderate	78.3	84.5	87.0	83.4	87.8	92.4
38	Moderate	Moderate	72.6	78.9	84.8	84.1	85.4	91.3
39	High	Moderate	60.4	73.5	85.2	73.5	84.9	87.3
40	Moderate	Moderate	78.2	84.1	88.4	86.4	89.6	94.3
	Average		68.5	76.5	84.4	79.0	84.9	89.7

counted the number of characters that had been correctly extracted by the system. Some of the outputs were found to have missed some characters in the original headlines, while others had erroneously extracted nonheadline characters. Two metrics, i.e., precision and recall [16], are used here as a measure of headline extraction by our system. The two metrics are defined as follows:

$$\text{Precision} = \frac{\begin{array}{c}\text{No. of headline characters}\\ \text{correctly extracted by the system}\end{array}}{\begin{array}{c}\text{No. of characters (headline or nonheadline)}\\ \text{extracted by the system}\end{array}}$$

$$\text{Recall} = \frac{\begin{array}{c}\text{No. of headline characters}\\ \text{correctly extracted by the system}\end{array}}{\begin{array}{c}\text{Actual no. of headline characters}\\ \text{in the microfilm page}\end{array}}$$

Note that as described in the introductory section, the present research concentrates on headline extraction; the characters extracted in the above experiments were not sent to any OCR process for conversion to text. The metrics defined above aim to measure how many characters can be correctly identified (without recognition) as headlines. A high recall rate shows the ability to extract as many headline characters as possible, i.e., a 100% recall

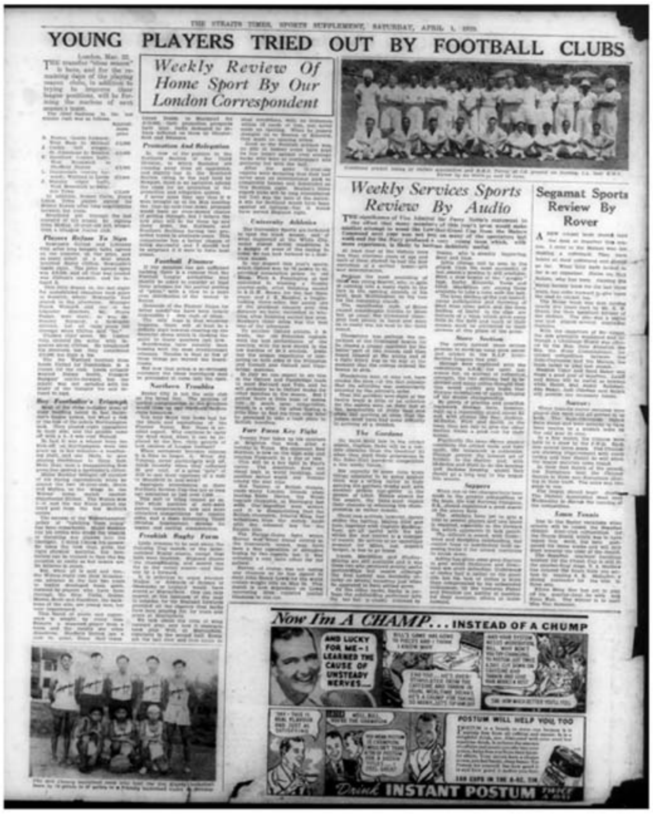
THE STRAITS TIMES, SPORTS SUPPLEMENT, SATURDAY, APRIL 1, 1939.
YOUNG PLAYERS TRIED OUT BY FOOTBALL CLUBS
Weekly Review Of Home Sport By Our London Correspondent
Weekly Services Sports Review By Audio
Segamat Sports Review By Rover
Now I'm A CHAMP... INSTEAD OF A CHUMP
Drink INSTANT POSTUM

Fig. 8. A skewed newspaper microfilm image

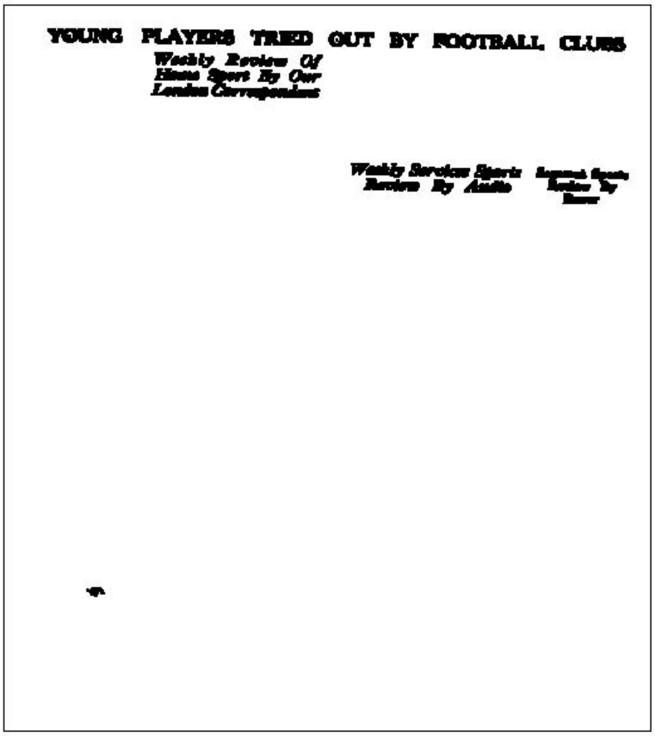
YOUNG PLAYERS TRIED OUT BY FOOTBALL CLUBS
Weekly Review Of Home Sport By Our London Correspondent
Weekly Services Sports Review By Audio
Segamat Sports Review By Rover

Fig. 9. Headlines extracted from Fig. 8 image

represents a complete extraction of all headline characters present in the microfilm page, but some of the non-headline characters may have been erroneously extracted at the same time. On the other hand, a high precision rate indicates the ability to exclude false positives as much as possible, i.e., a 100% precision means none of the non-headline characters were falsely identified as headlines, but some of the genuine headline characters may have been missed. Table 1 shows the experimental results in terms of precision and recall rates for the 40 test images. The variety of image quality in terms of extent of noise and blurring discussed earlier is also indicated in Table 1.

5 Conclusion and discussion

We propose a document analysis system that extracts news headlines from microfilm images in order to perform automatic indexing of news articles. The poor image quality of the old newspapers presented us with several challenges. First, the images must be properly binarized and excessive noise removed. 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 presented has significantly improved the final output despite the extremely poor and nonuniform illumination of the microfilm images and present good results.
- An 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 image segments overlap with the gray level of the background.
- Our headline extraction method works well even with skewed images of up to 5°. The microfilm images in the National Library were filmed using a special fixture. As such, the images are all upright with very little skew. The most serious skew is found to be within 5°, and our system has been found to work well with this skew angle. Thus no deskewing of images was done in our experiment. Figures 8 and 9 show, respectively, a skewed newspaper microfilm image and its final output using the present method.
- The preprocessing steps used in the present method have achieved a significant improvement in headline extraction. The average recall and precision rates are 84.4% and 89.7% as compared to 76.5% and 84.9% for Otsu's method and 68.5% and 79% for the conventional approach, respectively. Figures 10 and 11 show a consistent increase of recall and precision, respectively, across all the 40 test images.
- On a Pentium III 800-MHz PC, the average processing time (in seconds) of the above methods are 1.0, 2.5, 1.3, and 18.5 for histogram transformation, local adaptive binarization, noise reduction, and headline extraction, respectively.

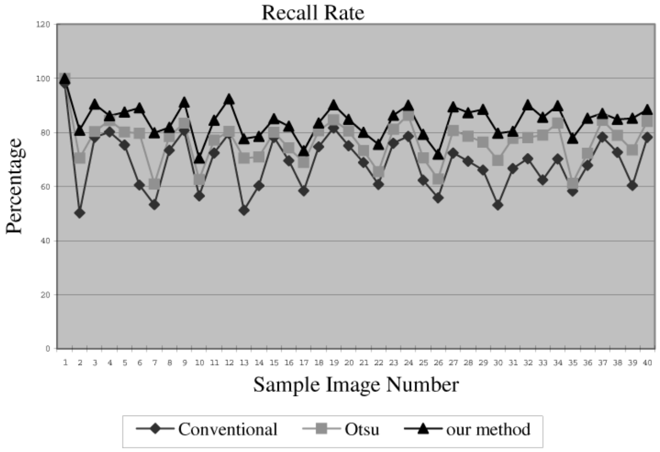

Fig. 10. Comparing recall rates of the three approaches: conventional approach, Otsu's method, and our method

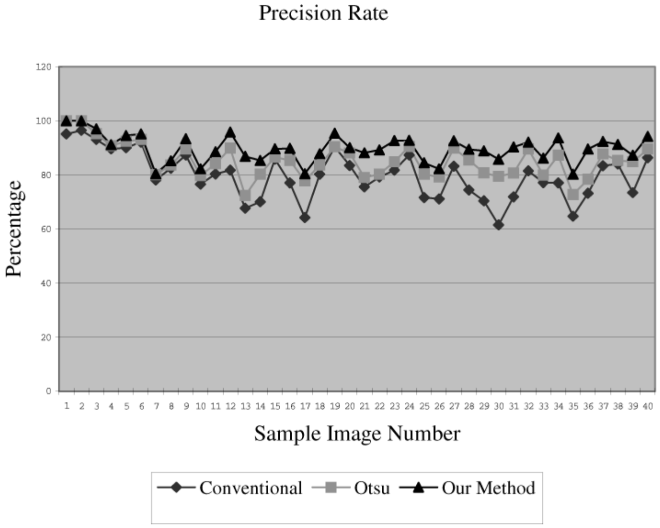

Fig. 11. Comparing precision rates of the three approaches: conventional approach, Otsu's method, and our method

– Finally, the recall rate of the headline extraction is not always 100% in the results shown in Table 1. Headlines that are too close to vertical or horizontal lines may be erroneously regarded as graphical or text blocks as shown Figs. 6 and 7. One point to note is that headlines with smaller font sizes that are outside the detection range will not be identified as headlines. They are not counted in the computation of recall and precision rates anyway. The objective in the present work is to capture only prominent headlines for automatic indexing.

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

References

1. Fisher JL, Hinds SC, D'Amato DP (1990) A rule-based system for document image segmentation. In: Proceedings of the international conference on pattern recognition (ICPR), Atlantic City, NJ, June 1990, pp 567–572
2. Fletcher LA, Kasturi R (1988) A robust algorithm for text string separation from mixed text/graphics images. IEEE Trans Patt Analysis Mach Intell 10(6):910–918
3. Forrester MA et al (1987) Evaluation of potential approach to improve digitized image quality at the patent and trademark office, MITRE Corp, Working Paper WP-87W00277, McLean, VA
4. Junker M, Hoch R, Dengle A (1999) On the evaluation of document analysis components by recall, precision and accuracy. In: Proceedings of the international conference on document analysis and recognition (ICDAR), Bangalore, India, September 1999, pp 713–716
5. Negishi H, Kato J, Hase H, Watanabe T (1999) Character extraction from noisy background for an automatic reference system. In: Proceedings of the international conference on document analysis and recognition (ICDAR), Bangalore, India, September 1999, pp 143–146
6. Niblack W (1986) An introduction to image processing. Prentice-Hall, Englewood Cliffs, NJ, pp 115–116
7. Niyogi D, Sihari SN (1997) The use of document structure analysis to retrieve information from documents in digital libraries. In: Proceedings of SPIE Document Recognition and Retrieval IV, San Jose, February 1997
8. Niyogi D, Sihari SN (1996) Using domain knowledge to derive the logical structure of documents. In: Proceedings of SPIE Document Recognition and Retrieval III, San Jose, January 1996
9. O'Gorman L (1992) Image and document processing techniques for the right pages electronic library system. In: Proceedings of the international conference on pattern recognition (ICPR), Amsterdam, August 1992, pp 260–263
10. O'Gorman L (1994) Binarization and multithresholding of document images using connectivity. CVGIP Graphical Model Image Process 56(6):494–506
11. Otsu N (1979) A threshold selection method from gray-level histogram. IEEE Trans Sys Man Cybern SMC-9(1):62–66
12. Pavlidis T (1982) Algorithms for graphics and image processing. Computer Science Press, Rockville, MD
13. Takebe H, Katsuyama Y, Naoi S (1999) Character string extraction from newspaper headlines with a background design by recognizing a combination of connected component. In: Proceedings of SPIE Document Recognition and Retrieval VI, San Jose, January 1999, pp 22–29
14. Trier OD, Taxt T (1995) Evaluation of binarization methods for document images. IEEE Trans Patt Analysis Mach Intell 17:312–315
15. Wong KY, Casey RG, Wahl FM (1983) Document analysis system. IBM J Res Develop 26(6):647–656

Chew Lim Tan is an associate professor in the Department of Computer Science, School of Computing, National University of Singapore. He received his B.Sc. (Hons) in physics in 1971 from the University of Singapore, his M.Sc. in radiation studies in 1973 from the University of Surrey, UK, and his Ph.D. in computer science in 1986 from the University of Virginia, USA. His research interests include document image and text processing, neural networks, and genetic programming. He has published more than 170 research publications in these areas. He is an associate editor of Pattern Recognition and has served on the program committees of the International Conference on Pattern Recognition (ICPR) 2002, Graphics Recognition Workshop (GREC) 2001 and 2003, Web Document Analysis Workshop (WDA) 2001 and 2003, and Document Image Analysis and Retrieval Workshop (DIAR) 2003.

Qing Hong Liu is a research engineer at the Data Storage Institute, Singapore. She received her M.Sc. in computer science from the National University of Singapore in 2002. Her research interests include image processing, information retrieval, engineering design, and developing technology.