October 20, 1998
RT0281
Computer Science 10 pages

Research Report

Restoration of decorative headline images for document retrieval

Tomio AMANO

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Restoration of decorative headline images for document retrieval

Tomio AMANO
IBM Research, Tokyo Research Laboratory,
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242, Japan,
amano@jp.ibm.com, Tel 81-462-73-4605, Fax 81-462-73-7413

October 12, 1998

Abstract

This paper describes a method for restoring decorative character images in headlines of newspapers and magazines. Although headlines contain useful keywords for document retrieval, conventional OCRs cannot always recognize them because the characters are often printed in reverse and with various background textures. We made filters that generate multiple candidate images by changing a small number of simple parameters (namely, by setting a threshold for stroke-width filtering and reversing black and white), so that one of the candidates contains a "normal" image whose characters are printed in black on a white background. If all the candidate images are recognized and an index is created, the keywords in headlines are expected to be retrieved without manual keyword entry and verification processes. In an experiment that we conducted, about 90% of characters in headline images segmented from newspapers were restored in the sense that one of the restored candidate images contained correct character images.

1 Introduction

Entry of existing paper documents is still an obstacle to the widespread adoption of electronic document management systems. Various efforts have been made to improve the accuracy of document layout analysis and character recognition. Recently, the focus has shifted to a different approach based on the functions required by applications. For example, Chen et al. [1] proposed a system for summarizing documents on the basis of images instead of coded text, which can eliminate the workload of verifying and correcting for OCR results. Senda et al.[2] investigated a retrieval scheme, related to document query by keyword,

that uses multiple candidates of character segmentation and recognition results. These studies showed that the complete results of document analysis are not always necessary to meet an application's requirements.

In this paper, we investigate a method for restoring decorative headline images that follows the multiple candidate approach. Headlines in newspapers and magazines contain useful keywords for retrieval. However, conventional OCRs cannot recognize the characters, because headlines are often printed in reverse and with background textures. To deal with such decorative images, a preprocess that restores "normal images[3, 4] and recognition algorithm that is not affected by the decorations [5] have been proposed. We think the former is superior in that it can be easily incorporated into existing systems. Our purpose is to make the preprocess approach more robust and flexible by using the multiple candidates. To generate the candidate images, we have developed a stroke-width filter whose the results are independent of the accuracy of the preceding layout analysis process and any heuristics. The behavior of the filter is defined by a small number of parameters. We generated candidate images from 50 actual newspaper headlines, varying the parameters, and recognized them by using OCR software, to confirm that they include proper images.

The rest of this paper is organized as follows: Section 2 describes an document retrieval system, and the required properties of an image-restoring method. Section 3 explains the mechanism of our stroke-width filter. Experimental results for 50 headline images are given in section 4. In section 5, we discuss the validity of the approach and possible future enhancements.

2 Image restoration for document retrieval

Figure 1 shows the assumed process flow of keyword registration using multiple candidates. First, layout analysis is applied to an input image and headline areas are segmented. For each headline image, multiple candidates (including the original image) for restored images are generated. Every candidate image is processed by a conventional OCR subsystem, and all the recognition results are used to make an index. A query word will be matched with the index. The registration tasks are performed without any manual processes for verifying and correcting the OCR results, and consequently the cost of document registration is reduced.

In view of the automated flow, the restoration method should work independently of fluctuations in the results of the preceding layout analysis. Some prior restoring algorithm normalize the image with respect to the font width or height, or extract features from the image to examine the properties of the background. The results will vary according to variations in the input, whether or not neighboring headline areas are segmented separately. If two headlines with different font sizes and different background textures are segmented as a single area (strict separation is sometimes difficult), the restoration process fails.

To generate candidates for a restored image, we adopted a simple and stable algorithm rather than an intelligent and autonomous one.

3 Generating candidate images for restoration

Figure 2 shows example of decorative character images. Character strokes are represented by black lines on white, white lines on black, or outlines. The background texture is not always uniform: Sometimes only part of the background is textured, the texture changes gradually. To deal with such variations, two filtering processes based on morphological operations are used in combination with black-and-white reversal.

1. Stroke-width filtering
2. Blurring by means of a closing operation

Since the widths of character strokes are greater than those of lines in background texture, a stroke-width filter can extract foreground character images if an appropriate range of stroke widths is given. Blurring is effective for recovering cases in which character strokes are painted with texture e.g., Figure 2(d)).

3.1 Stroke-width filters

To describe the structure of the filters, let us define some primitive operators and procedures. Opening operators O_n^h and O_n^v erase horizontal and vertical black runs when the lengths are less than a threshold value n. Closing operators C_n^h and C_n^v replace horizontal and vertical white runs with black runs when their lengths are less than a threshold value n. Let I be an original image; $O_n^h(I)$ means an opened image. Pixel-wise operators $\vee, \wedge, and \oplus$ are used to represent OR, AND, and exclusive OR, respectively. Using this notation, selection operators that extract black runs whose lengths are within a specified minimum value m and maximum value n are defined.
$S_{m-n}^h(I) \equiv O_m^h(I) \oplus O_{n+1}^h(I)$
$S_{m-n}^v(I) \equiv O_m^v(I) \oplus O_{n+1}^v(I)$

Two relaxation procedures $RP_1(S, D)$ and $RP_2(S, D)$ are used to investigate the connectivity of black runs between a source image S and a destination image D. If a run in S is judged to belong to D, the run is moved to D from S. The pseudo-code of the procedures is shown in Figure 3.

A stroke-width filter is constructed on the basis of the operators and the procedures according to following three steps (Figure 4 shows intermediate results).

Step 1: Three intermediate images, – a candidate for the characters, C, a candidate for the background, B, and unclassified image, U – are generated from

the original input image I. First, C is generated by extracting horizontal and black runs by minimum and maximum values of the width, m and n, which are given as parameters.
$C \leftarrow S^h_{m-n}(I) \vee S^v_{m-n}(I)$
The difference between I and C is set to U.
$U \leftarrow I \oplus C$
A background image is generated as areas which both the widths and heights are greater than the maximum stroke width n.
$B \leftarrow O^h_{n+1}(U) \wedge O^v_{n+1}(U)$
Step 2: Relaxation procedures $RP_1(U, C)$ and, $RP_1(B, C)$ are used to recover runs that step 1 failed to extract as parts of foreground characters. Each black run in U and B is moved to C if it is connected to more than a threshold number of pixels in C.
Step 3: In contrast to step 2, the following procedures are used to screen out surplus runs in C:
$RP_2(C, B)$
$RP_2(U, B)$
$B \leftarrow O^h_m(O^v_m(B))$
$RP2(C, B)$

Figure 5 shows examples of the restoration of stroke-width filtering for several conditions of black-and-white reversal and ranges of stroke width.

3.2 Blurring

Blurring consists of consecutive closing and opening operations.
$C \leftarrow O^v_4(O^h_4(C^v_4(C^h_4(I))))$

4 Experiments

An experiment was carried out to confirm the feasibility of the proposed approach. Plural newspapers were scanned in 200 dpi resolution, and 50 headline images were manually segmented. All the images contained various decorations, and consequently could not be recognized by conventional OCRs. Using stroke-width filtering and blurring, we generated 12 candidate images (2 (normal or black-and-white reversed)$\times$ (5 (stroke width variations) + 1 (blurring))) from each headline image. We tried five stroke width parameters: 2-16 pixels; 4-32 pixels; 8-64 pixels for both of horizontal and vertical strokes, 2-16 pixels for horizontal strokes and 4-32 pixels for vertical strokes, and 4-32 pixels for horizontal strokes and 8-64 pixels for vertical strokes. The last two combinations were added to take account of the fact that some fonts have thinner horizontal than vertical strokes.

Although some noise and lost pixels were observed, a properly restored (black

characters on white background) image was included among the candidates for 46 of the images. For more quantitative evaluation, the restored images were recognized by an OCR software product obtainable at stores. We calculated the recognition accuracy, selecting the candidate images that showed the best results. The restoration accuracy for 526 characters was 85%. Of 79 errors, 28 were caused by the OCR software itself. We expect that 90% (475/526) of decorative characters can be restored and used to make indexes for document retrieval.

Table 1 shows the distribution of parameters that gave the best results for each headline image. It shows that various decorative patterns are covered by multiple parameter sets.

Table 1: Parameters giving the best candidate images

Parameters	Stroke-width filtering										Blurring	
(Vertical and horizontal	2-16		4-32		4-32		8-64		8-64		4	
stroke widths)	2-16		2-26		4-32		4-32		8-64		4	
(Reverse/Normal)	R	N	R	N	R	N	R	N	R	N	R	N
Number	7	2	10	12	9	2	1	1	0	1	0	2

The 51 restoration errors are divided into two groups: 17 cases caused by partial noises or lost strokes, and 34 cases in which restoration failed for the overall image. Figure 6 shows examples of the second group. Figure 6(a) shows an unexpected decoration style in which character strokes are represented by outline and texture. In Figures 6(b) and 6(c), relaxation procedures did not work well for changes of background texture. A proper candidate could not be generated for the case shown in Figure 6(d), owing to the use of unsuitable blurring parameters (only one combination of blurring parameters was used in the experiment).

5 Conclusion

A method of restoring decorative headline images for document retrieval has been studied. To take account of fluctuation in the results of the preceding layout analysis, multiple candidate images are generated without parameter estimation and size normalization. In experiment using 50 actual headline images, stroke-width filtering and blurring generated candidates that could be recognized by conventional OCRs for 90% of decorative characters.

Assuming an automated method of document registration, it is more practical to use multiple candidates than a single restored image, because the decorations of headlines are arbitrarily designed by publishers. Our experiment shows that a simple stroke-width filter with a small number of parameter sets can cover various types of texture and font sizes. Even if unexpected decorations

appear, the only extra work required is to create additional filters that can deal with the decorations.

In the proposed approach, the OCR and retrieval engine have to process more data because multiple images are passed to an OCR. However, most headlines contain only about 20 characters at most. Even if 100 times more data are produced, the increase in the number of characters to be recognized will be less than 2000, which is about the number of characters on a single page of an average document. We think this is a tolerable price to pay for the advantage of being to able to retrieve headline text.

In this paper, we proposed a multiple-candidate approach to headline text retrieval, and described a mechanism for generating candidate images. We are planning to verify the approach within an automated flow (from layout analysis to character recognition process). Since the relaxation processes are performance bottlenecks of our approach, and they sometimes cause restoration failures, we also plan to enhance relaxation algorithms used in stroke-width filtering.

References

[1] Chen, R. and Bloomberg, S.: Extraction of Indicative Summary Sentences from Imaged Documents, ICDAR'97, pp. 227–232 (1997).

[2] Senda, S. Minoh, M. and Ikeda, K.: Document Image Retrieval System Using Character Candidates Generated by Character Recognition Process, ICDAR '93, pp. 541–546, (1993).

[3] Liang, S. Ahmadi, M. and Shridhar, M.: A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background images, *CVGIP:Graphical Models and Image Processing*, Vol. 56, No. 5, pp. 402–413 (1994).

[4] Lin, C. Takai, M. and Narita, S: Decorative Character Restoration by Image Processing, *Technical report of IEICE* (in Japanese) PRU94-12 (1994).

[5] Sawaki, M and Hagita, N.: Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning, *Trans. IEICE*, Vol. E79-D, No. 5, pp. 491–497 (1996).

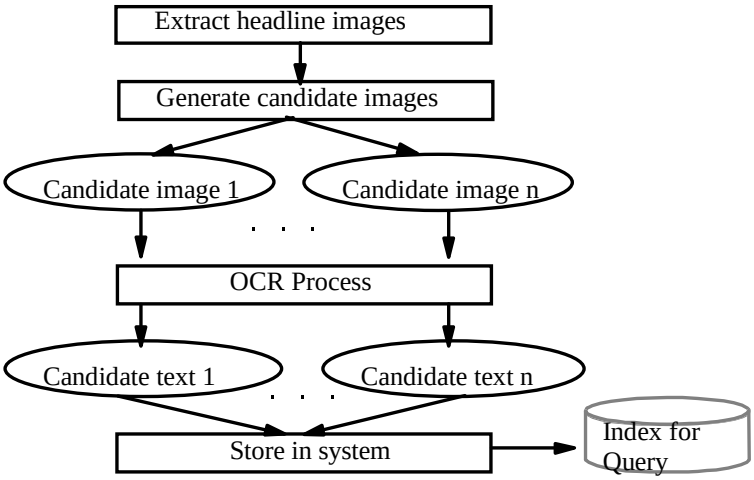

Figure 1. Registration of headlines for information retrieval.

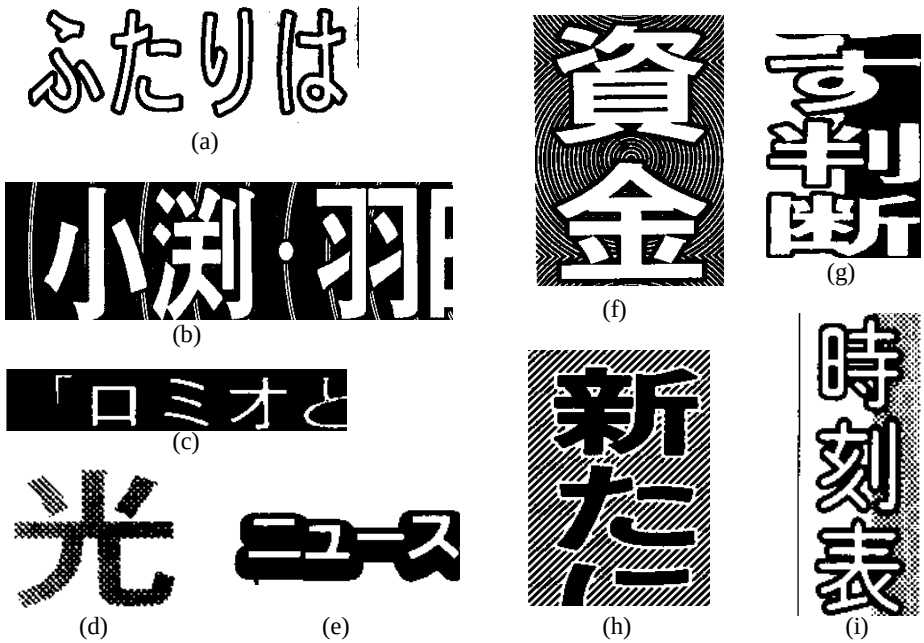

Figure 2. Examples of decorative headline images.

```
Relaxation process RP1(S, D)
{
 Perform the following processes until no run is moved {
  Raster-scan S for each observed run R_s {
   Investigate the preceding and next scan lines of R_s in D.
    Count the number of pixels (N) connected to R_s.
    If the corresponding area of the next line is filled
    with pixels belonging to S, the following scan line
    is investigated instead.
   if ((R_s is horizontally connected to D
   and N ≥length of R_s + 1)/2 ) or N ≥ length of R_s ) {
    Move R_s from S to D.
   }
  }
 }
}

Relaxation process RP2(S, D)
{
 Perform the following processes until no run is moved {
  Raster-scan S for each observed run R_s {
   Investigate the preceding and next scan lines of R_s in D.
    Count the number of pixels (N) connected to R_s.
   if ((R_s is horizontally connected to D
   and N ≥(length of R_s + 1)/2 ) or N ≥ length of R_s ) {
    Move R_s from S to D.
   }
  }
 }
 Perform same the same processes the changing main scanning direction {
  :
 }
}
```

Figure 3: Pseudo-code for the relaxation procedures

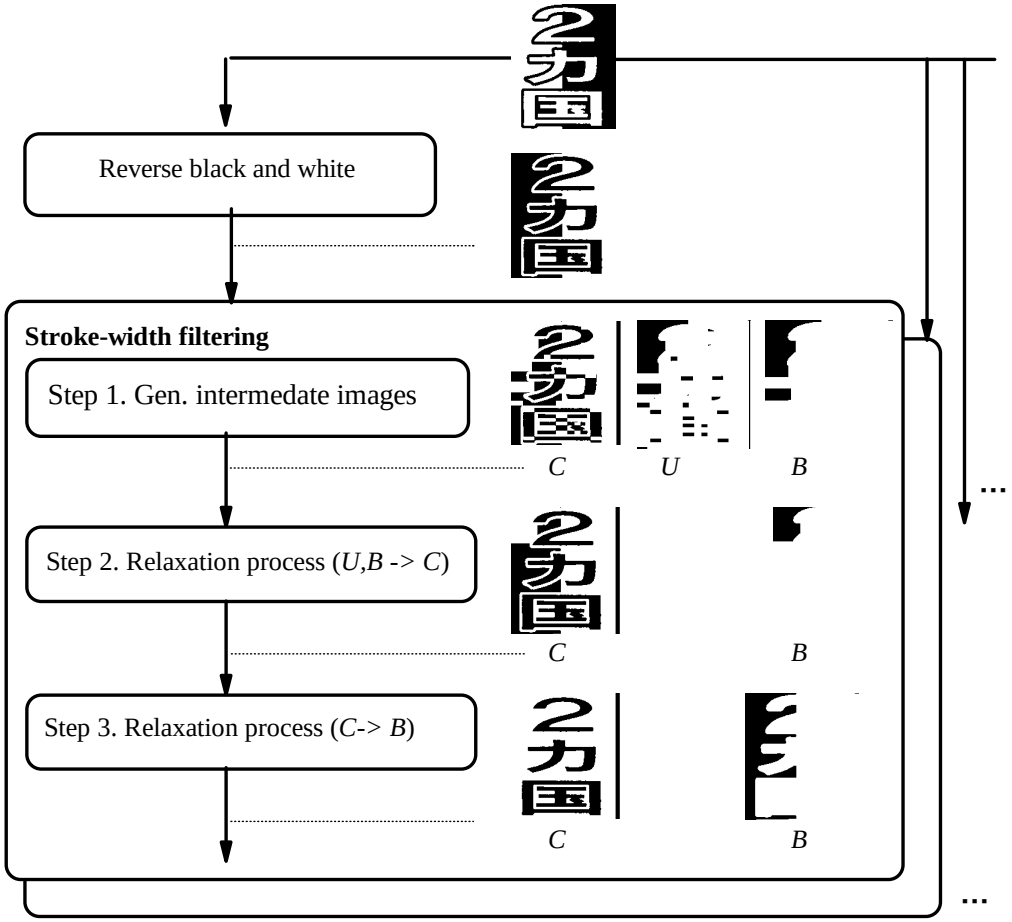

Figure 4. Restoration by stroke-width filtering

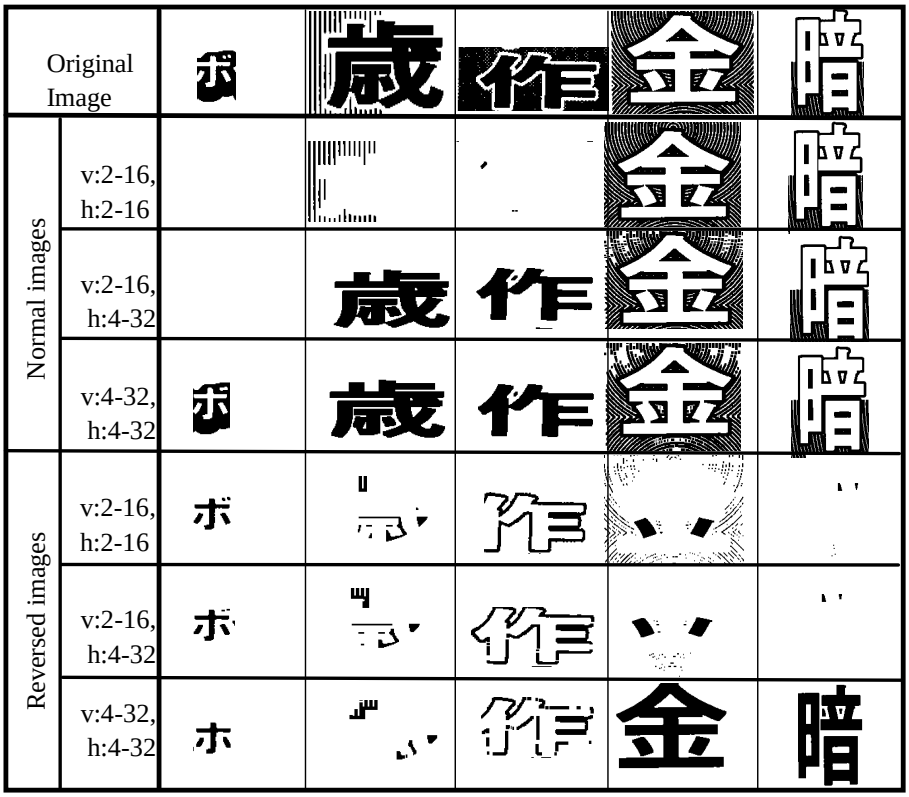

Figure 5. Examples of restoration of decorative character images

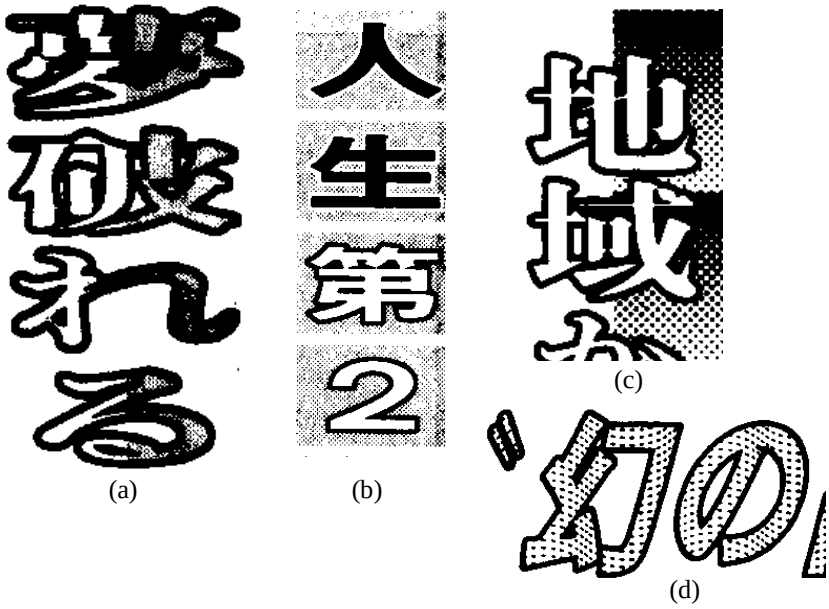

Figure 6. Examples of restoration failures