2004 IEEE International Conference on Multimedia and Expo (ICME)

The Extraction of Characters on Dated Color Postcards

[1]Shwu-Huey Yen, Mei-Fen Chen, [2]Hwei-Jen Lin, Chia-Jen Wang, Chiu-Hsiang Liu
Department of Computer Science and Information Engineering, Tamkang University
Tamsui, Taiwan, R.O.C. e-mail: {[1]shyen, [2]hjlin}@cs.tku.edu.tw

Abstract

A novel scheme is proposed to extract characters from dated postcards. The illustrations of the postcards appear in various languages and colors embedded in different backgrounds. Due to reproduction and uneven illumination, these characters suffer a severely degradation and hence extracting characters using conventional methods becomes difficult. A morphological operation is proposed to remove irrelevant backgrounds that are connecting to border edges of the postcards. As a result, characters become the most obvious objects. Followed by horizontal and vertical projections, the exact locations of the characters can be located. The proposed scheme has been executed on a set of color images of postcards and proved its efficacy.

Key words: *Color Postcards, Characters Extraction, Morphological Reconstruction, Connected Component Analysis, RLSA.*

1. Introduction

In this study, we propose a scheme to extract characters from color postcards printed around 1930 during the time Japan occupied Taiwan as one of its colonies. These postcards, collected and reproduced in the book [7], possibly the earliest postcards about Taiwan, contained photographs of mountains, rivers, countryside, famous building, etc. Figure 1 shows two color postcards. The illustrations are printed either on the upper or lower positions of the postcards with various backgrounds. Characters may appear in black, white, or red colors with underneath backgrounds of blue sky and/or white clouds, dark mountains, brownish rocks, trees with leaves and/or branches, rocky or muddy country roads, ponds, etc. These reproduced images suffer a degradation problem due to old time, uneven illuminations, or some unknown reasons. As they are precious and fragile historical documents, it is much meaningful if they can be kept in a digital library via key words searching, with helps of characters extraction, to make these historical documents available to researchers as well as the general public.

There are researches about character extraction. Wang *et al.* [4] proposed a character segmentation method of color images in which a multiscale edge detection is applied based on the color difference. But the method cannot handle characters located on complex background. In Tsai *et al.* [1], they proposed a method uses luminance and saturation features to extract possible characters on color background, but it fails when the document has inseparable luminance and saturation, or gradually decaying background. Another method in [2] is based on a combination of an adaptive color reduction technique and a page layout analysis approach. The results are not satisfied when characters are small and not in solid colors. Chen *et al.* [6] proposed a method to classify the coherent blocks as text or non-text for color technical journal's cover. It cannot extract non-uniformly colored characters from complex background.

In our study, we found out that the colors of the characters are very sensitive to the backgrounds. In particular, these characters appear in solid colors on postcards to us, but in fact they show quite a different variations on colors among their neighboring characters and sometimes are inseparable from background. Thus we use the intensity information of the postcards for extracting characters and color information for post processing only. Since illustrations are either printed in the upper or

(a) House. (b) The local vendor.

Figure 1. Two color postcards.

0-7803-8603-5/04/$20.00 ©2004 IEEE

lower 1/5 position, or no illustration at all, thus only upper and lower 1/5 of the postcards will be investigated here. The paper is organized as follows. Section 2 describes our algorithm. We discuss the effects of preprocessings in digitalilized color postcards in 2.1. Section 2.2 explains how to remove irrelevant background so that characters become principal objects in the remaining image. In 2.3, via histogram analysis, image is binarized and characters are obtained. In 2.4, the method for exact character-region detecting is described. Color contrast-enhancement and sharpening are done as the post processing is described in 2.5. Finally, experiments and discussions are in Section 3.

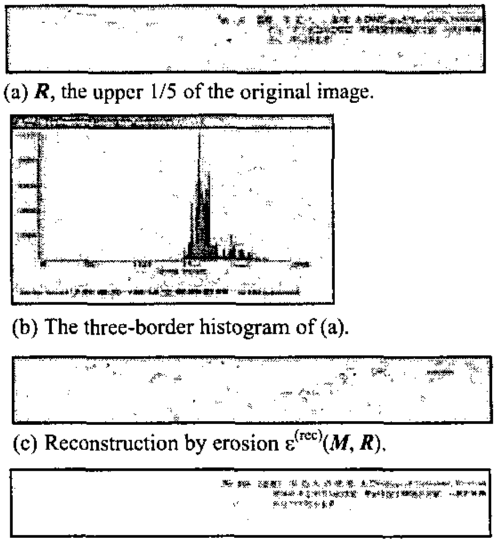
2. The algorithm

2.1. Preprocessing

Smoothing is a common preprocessing in various image processing. For example, opening-closing operation of morphology, low-pass filter of linear transformation, and diffusion filter [3] of non-linear transformation are often used as preprocessing to remove noises. In our case, since the characters printed on the postcards are relatively small, the role they play is more like noise than the principal object. Thus, the characters become more blurred than before if smoothing is applied, i.e. more information about characters is lost. Therefore preprocessing of smoothing is not suitable on this application.

2.2. Irrelevant background removal

Because characters are embedded in complex background and not connecting to the border edges of the image with rare exception, we reconstruct the background image that connecting to border edges by morphological reconstruction [5]. Followed by a subtraction from the original image then most of irrelevant background can be removed. The reference image $\boldsymbol{R}$ is the original image (fig.2 (a)) and the white (or black) marker image $\boldsymbol{M}$ is the image of the same size as $\boldsymbol{R}$ with gray values all equal to 255 (or 0) except pixels on four border edges have the same gray values as in $\boldsymbol{R}$. As in fig.2 (a), characters are dark, in order to have bright background, reconstruction by erosion with the white marker image, $\varepsilon^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$, should be used. Similarly, if $\boldsymbol{R}$ has bright characters with dark background then reconstruction by dilation with the black marker image, $\delta^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$, should be used. Since if characters are embedded in bright background, then more likely characters are printed dark and vice versa. Thus, the distribution of intensities of border pixels on three sides is used to guess the darkness/brightness of the corresponding background. The reason of only three borders used is that if the image is upper (or lower) 1/5 of the postcard then the lower (or upper) border is not a real border edge of the postcard. In our experiment, we count the number of pixels with gray value ≤ 100, if it is more than 50% then the background is more likely to be dark otherwise bright. As in fig. 2(b), the percentage of number of pixels with gray value ≤ 100 is zero. Thus $\varepsilon^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$ with white marker image $\boldsymbol{M}$ is used and the result is shown in fig.2(c). To remove irrelevant background, we do $\varepsilon^{(rec)}(\boldsymbol{M}, \boldsymbol{R})-\boldsymbol{R}$, i.e. difference the images (c) and (a), most of backgrounds are removed, only characters and some noises are remained as in (d). For the case of $\delta^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$ with black marker $\boldsymbol{M}$, $\boldsymbol{R}-\delta^{(rec)}(\boldsymbol{M},\boldsymbol{R})$ will remove background connecting to borders similarly.



(a) $\boldsymbol{R}$, the upper 1/5 of the original image.

(b) The three-border histogram of (a).

(c) Reconstruction by erosion $\varepsilon^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$.

(d) $\varepsilon^{(rec)}(\boldsymbol{M}, \boldsymbol{R})$- $\boldsymbol{R}$, the difference image of (c), (a).

Figure 2. To remove irrelevant background via morphological reconstruction by erosion.

Figure 3. The binarization of image 2(d).

2.3. Binarization of subtracted images

To extract characters, the subtracted image obtained from Section 2.2 is binarized through a threshold, *th*. An ideal subtracted image has most white background and a few dark characters or noises as in fig. 2(d). In our experiment, *th* is determined as the minimum of 225 and T (:lower 5% gray value histogram of the subtracted image), i.e. *th* = min(T, 225) where

$$\frac{\sum_{x=0}^{T} H[x]}{\sum_{x=0}^{255} H[x]} \leq 5\% \quad \text{and} \quad \frac{\sum_{x=0}^{T+1} H[x]}{\sum_{x=0}^{255} H[x]} > 5\% \quad .$$

Take fig. 2(d) as an example. Since T is evaluated as 227 and *th* = min(T, 225) = 225, the binarized image is shown in fig. 3

2.4. Detecting characters regions

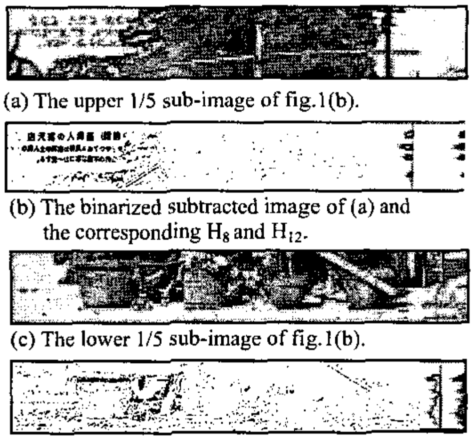
2.4.1. Horizontal projection for possible text region:

For each row of the binarized image obtained from Section 2.3 the number of crossings from black pixel to white pixel is counted. A horizontal histograms H*i* counting these crossings is constructed such that for any row *k* if the number of crossings is *m* and $m \geq i$ then $H_i(k) = m - i$ otherwise $H_i(k) = 0$. The length of a **nonzero interval** of H_i is defined to be *l* if there is a consecutive *l* nonzero rows, i.e. $H_i(j) \neq 0$, for $t \leq j \leq t+l-1$ and $H_i(t-1) = H_i(t+l) = 0$ for some t. Between any two nonzero intervals, a **zero interval** is considered, the length of a zero interval of H_i is defined similarly. H_i is to determine whether the sub-image contains illustration. As in fig.4, (a), (c) are upper/lower sub-images, (b) and (d) are the binarized subtracted images with corresponding horizontal projections H_8 and H_{12} on the right. In (b), there are four nonzero intervals and three zero intervals both in H_8 and H_{12}, but in (d) there is only one nonzero interval.

If the sub-image contains illustration then horizontal projection should exhibit regularity on heights of nonzero and zero intervals. Since there is at most one sub-image containing the illustration, this regularity property can be used for determining which sub-image does contain the illustration. If H_8 shows a regularity property, or due to possible noises affecting H_8, H_8 does not but H_{12} does, then the corresponding sub-image contains characters as in fig.4(b). On the other hand, in fig.4(d) both H_8 and H_{12} show no such regularity property, we conclude it does not contain characters.

2.4.2. Noise removal by connected component analysis:

After the sub-image containing illustration is found, connected component analysis (CCA) is applied. The resulted blocks may be characters or noises. We classify noises into three types **A**, **B**, and **C**:


(a) The upper 1/5 sub-image of fig.1(b).

(b) The binarized subtracted image of (a) and the corresponding H_8 and H_{12}.

(c) The lower 1/5 sub-image of fig.1(b).

(d) The binarized subtracted image of (c) and the corresponding H_8 and H_{12}.

Figure 4. Horizontal projections for possible character regions detection.

A. If a block with height greater than 20, density greater than 0.3, and the horizontal projection of its center-position falls on nonzero interval of H_8, then it is a type A noise. Since if the center position falls on zero interval of H_8, it is very possible that two (illustration) lines are connected into one piece, we do not want to treat them as noises.

B. If a block is not a type A noise but its center-position falls on zero interval of H_8, and the density is greater than 0.8, then it is classified as a residue noise. This kind of noises usually is the residue from the border-connecting background.

C. If a block does not belong to A, B, then it is a type C noise if it satisfies one of the following:

strict criteria:

Center position of the block is on nonzero interval of H_{12}, and there is no other block in the area by extending 3 times of the dimension of such block.

loose criteria:

Center position of the block is on nonzero interval of H_8, and there is no other block in the area by extending 2 times of the dimension of such block.

between and outside nonzero intervals :

If the horizontal projection of this block totally falls in zero interval, or above (below) the first (last) nonzero intervals of H_8.

2.4.3. Horizontal and vertical projections for exact region:

After most of noises are removed as described above, horizontal and vertical projections are formed again on the resulted image. By observing the lengths of nonzero and zero intervals, the heights of character lines and inter-line distances can be determined. Since H_8 may be affected by noises, H_{12} is used to figure out the average of height of lines and inter-line distances. Using this information false character lines can be eliminated. As in fig. 4(b), the horizontal projection shows four lines, but inter-line distance can tell us that the last line is not a real one for the illustration. The exact height of illustration is determined by extending outwards extra 5 pixels to the first pixel in first line and the last pixel in last line of H_8 to assure all characters can be included. Next, RLSA method is applied on vertical projection to determine the width of the illustration. If v(i), v(j) are nonzero and $| i\text{-}j | < c$ (: c=30) then make interval between them to be nonzero, where v(i) is the number of crossings of the column i. Finally, width of the illustration is determined by the length of longest nonzero interval.

2.5. Post processing

After the exact region determined, the quality of the characters will be improved by contrast enhancement and sharpening. The original color information of extracted characters is obtained from the color postcards. Contrast enhancement is done by suppress the first 50 and lower 50 intensity values then extend the rest of intensity to be 0 ~ 255, similarly for the saturation. To sharpen the image, we apply the high-boost filter on R, G, and B respectively for each character pixel. Figure 5 shows before and after the post processing, the quality of extracted characters is indeed improved.

3. Experimental results and Discussion

The proposed method is implemented from sample images in [7]. In "Suspension bridge" (fig. 6(b)), the original image (d) has an oblique line crossing the characters, but the characters are extracted successfully as in (c). But if the background image is really similar to characters, in addition to uneven illumination, the extracted result may be not satisfied. In "Buffalo" (fig.6(a)), illustration is on the lower right corner of the postcard with stream underneath. Although the character region (e) is found correctly but the result is not good enough due to the stream is under shadow and with ripples. How to solve this problem will be our future work.

References

[1] Chun-Ming Tsai and His-Jian Lee, "Binarization of Color Document Images via Luminance and Saturation Color Features", *IEEE Transactions on Image Processing*, Vol. 11, No. 4, April 2002, pp. 434-451.

[2] C.Strouthopoulos, N.Papamarkos, A.E.Atsalakis, "Text Extraction in Complex Color Documents", *Pattern Recognition*, Vol.35, Issue 8, 2002, pp. 1743-1758.

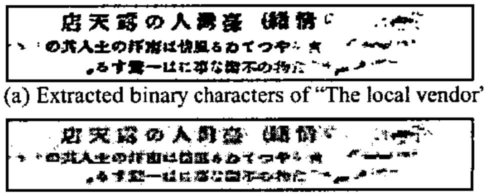
[3] Jian Ling, Alan C. Bovik, "Smoothing Low-SNR Molecular Images via Anisotropic Median-Diffusion", *IEEE transaction on medical imaging*, Vol. 21, No. 4, April 2002, pp.377-384

[4] Kongqiao Wang, Kangas, J.A., Wenwen Li, "Character segmentation of color images from digital camera", *IEEE Document Analysis and Recognition*, 2001. Proceedings. Sixth International Conference on, 10-13 Sept. 2001, pp. 210-214.

[5] Pierre Soille, "Morphological Image Analysis Principles and Applications", Springer-Verlag Berlin Heidelberg 1998.

[6] Wei-Yuan Chen, Shu-Yuan Chen, "Adaptive Page Segmentation for Color Technical Journals' Cover Images", Elsevier Science, *Image and Vision Computing* 16, 1998, pp. 855-877.

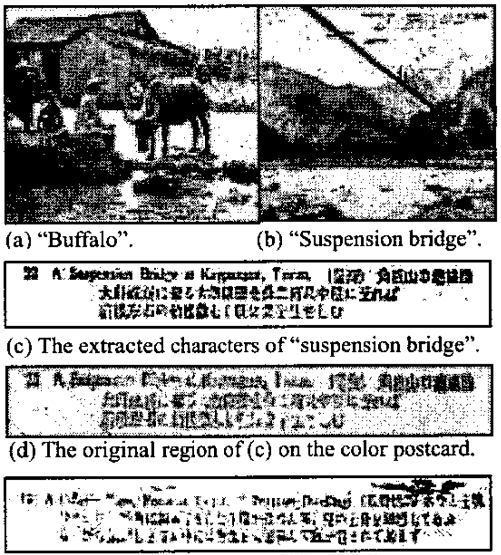
[7] Ikegami Shuuho, "Taiwan Jixing" (Journey to Taiwan), trans. by Zhang Liangze, Cianwei Pub. Inc. Taipei, 2001..



(a) Extracted binary characters of "The local vendor".

(b) The result after post processing.

Figure 5. Final result of "The local vendor".



(a) "Buffalo". (b) "Suspension bridge".

(c) The extracted characters of "suspension bridge".

(d) The original region of (c) on the color postcard.

(e) The extracted characters of "Buffalo".

Figure 6. The experimental results.