Character Region Identification from Cover Images Using DTT

Lixu Gu

Computer Science, Shanghai Jiaotong University, Shanghai, China
gu-lx@cs.sjtu.edu.cn

Abstract. A robust character region identification approach is proposed here to deal with cover images using a differential top-hat transformation (DTT). The DTT is derived from morphological top-hat transformation (TT), and efficient for feature identification. This research is considered as a fundamental study for auto-classification of printed documents for organizing a Digital Library (DL) system. The entire procedure can be divided into two steps: region classification and character region identification. In the first step, a source gray image is segmented by a series of structuring elements (SE) into sub-images using the DTT. Since the widths of regions are relative to the scales of the characters, the different scales of characters are classified into the series of sub-images. The character region identification processing is composed of feature emphasis, extraction of candidate character regions and region reconstruction processing. Feature emphasis processing reduces noises and emphasizes characters in the sub-images, and then the candidate character regions are extracted from the gray scale sub-images by a histogram analysis. Lastly, a morphological image reconstruction algorithm based on conditional dilation is introduced to make the extracted character regions distinct from noises. To demonstrate the robustness of the proposed approach, 30 gray scale cover images were tested in the experiments, which revealed that an average extraction rate of 94% has been achieved.

Keywords: Optical Character Reader, Cover Image, DTT, Region Identification.

1 Introduction

An enormous amount of digital and printed documents are produced daily. The existing printed documents are gradually replaced by digital documents (i.e., digital magazines). To store and process these documents, efficient techniques of conversion from paper to digital documents are required. An Optical Character Reader (OCR) facilitates the conversion for organizing a digital library (DL) system. But there still have many difficulties in auto-processing these source printed documents. Among them, the most errors are considered come from the step of character region identification due to the increment of variety of printing styles, especially in the cover pages.

Many character region identification systems [1-7] are proposed in broad areas of different source documents. The main categories include:

Z. Chen et al. (Eds.): ICADL 2004, LNCS 3334, pp. 388–397, 2004.

- Engineering Drawing (ED): to extract characters from design papers for the purpose of CAD/CAM.
- Texts: to segment text regions and figure regions with column structures from Newspaper, book and magazine for the purpose of the digital library.
- Maps: to extract and recognize characters from maps for the purpose of digital maps and digital navigation.
- Cover Images: to extract characters from covers of books, magazines or posters for the purpose of the digital library.
- Scene Images: to extract and recognize characters from scenes for the purpose of robotic navigation.

Comparing these researches, many similar features can be found between studies on EDs and texts [1-4], where character regions are usually separated from figures and characters usually have regular sizes, directions and positions. In these cases, feature extraction techniques (e.g. thresholding) can be efficient to distinguish character regions. Even if characters in maps [5] are in regular layout either, they usually mixed with figures, and more difficult to be extracted. The last two categories (cover images and scene images)[6-8] are even more complicated in the aspect of that character regions are composed in the figures with irregular sizes, directions and positions. More image processing techniques are required before the features can be extracted.

In this paper, we propose a novel character region identification approach using mathematical morphology. Since morphological operations are geometry basis, they can easily extract margins of characters with different shape. Several morphological approaches [9,10] are proposed for character identification where [9] are about character extraction from binary newspaper headlines, which have a patterned background. Some strict conditions are set up in these papers. For example, “background patterns are slenderer than the characters”; “background patterns are periodically arranged”. Accordingly, we employ gray-scale and color images instead of binary images to extract characters from cover images of magazines without rigid preceding conditions.

The rest of the paper is organized as follows: in section 2, a brief review of DTT method and morphological reconstruction techniques are presented. In section 3, the novel segmentation algorithm is proposed. A demonstration of the proposed algorithm alone with a validation experiment is described in section 4. The robustness and finally accuracy of our approach are discussed in section 5.

2 Morphological Top-Hat and DTT

2.1 Top-Hat Transformation

The morphological “top-hat” transformation originally proposed by Meyer [11] provides an efficient tool for extracting bright (respectively, dark) objects from an uneven background. It is denoted by $T_i^{(i)}$,and defined as:

$$T_i^{(i)} = \begin{cases} T_i & if\ \ WTT;\ where,\ T_i = F - F \circ_g r_i k \\ T^{(i)} & if\ \ BTT;\ where, T^{(i)} = F \bullet_g r_i k - F \end{cases} \quad (1)$$

Where, F stands for source image. $r_i k = k \oplus k \oplus \cdots \oplus k$ (r_i times), when r_i denotes the scale of a structure element (SE, e.g. the radius of a disk).

The gray-scale original image F opened by a SE k can removes the bright areas which cannot hold the SE, and subtracting the opened image from the original one yields an image where the bright objects clearly stand out. This transformation is called "white top-hat" transformation (WTT). A closed original image in gray-scale subtracting original one allows us to extract dark objects from bright background, which is called "black top-ha" transformation (BTT), respectively.

2.2 Differential Top-Hat Transformation

For some complicated images, especially those in which the target objects are combined in the uneven background, the TT is difficult to segment interested particles satisfactorily since parts of noise regions are also holding the top gray regions which should be extracted with objective regions by the TT. Clues for detecting features were discovered when we concentrated on the TT with different sizes of disk SE. The difference between $T_i^{(i)}$, and $T_{(i-1)}^{(I-1)}$, includes our interested objects, and that image can be easily thresholded to make features stand out.

The new morphological segmentation algorithm named "Differential Top-hats" (DTT) corresponding to WTT and BTT is defined as follows:

$$F_i = | T_i - T_{i-1} |_B - F'_{i-1}; \quad F'_i = \bigcup_{1 \le j \le i} F_j; \quad F'_1 = \phi \quad (2)$$

or

$$F_i = | T^i - T^{i-1} |_B - F'_{i-1}; \quad F'_i = \bigcup_{1 \le j \le i} F_j; \quad F'_1 = \phi \quad (3)$$

Where F_i' denotes segmented images which hold different sizes i of objects. $|\ \ |_B$ stands for a threshold operation by a gray level B, which is determined experimentally. The differences of the neighbor TT results up to i are united together in Fi' with certain size of features.

As shown in Fig.2, the DTT is significantly improved from TT in the next aspects:

1. It can automatically select appropriate sizes of structuring elements satisfying different objects. Since the DTT operation does not only employ a single structuring element like the TT operation but also utilize a series of structuring elements. Appropriate sizes of structuring elements can easily be found to fit different objective regions in the studying image.
2. We can easily identify a satisfactory threshold value for DTT results. Since the DTT algorithm emphasized the differences between the regions with steep slopes and that with gradual slopes, a common threshold value can be easily found to all the steps of the processing.

3. The DTT can reduce more noises and emphasis the features of objects. Gray levels of most noises are greatly reduced due to their trivial gradients when gray levels of objects remain high. They are easily distinguished by a threshold operation as depicted in Fig. 2.

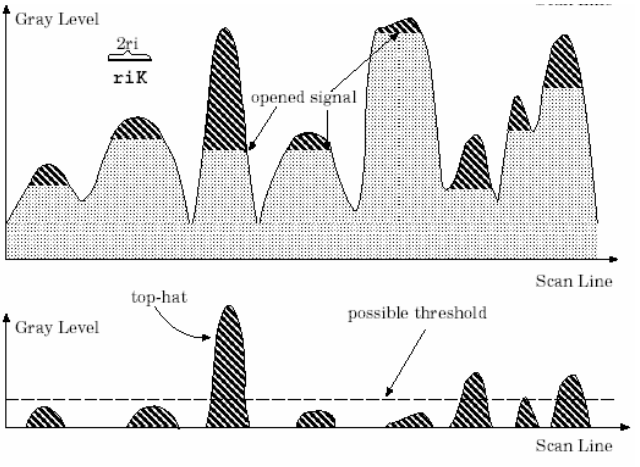


Fig. 1. The top-hat transformation. (upper) the result of opened signal and (lower) TT result T_i

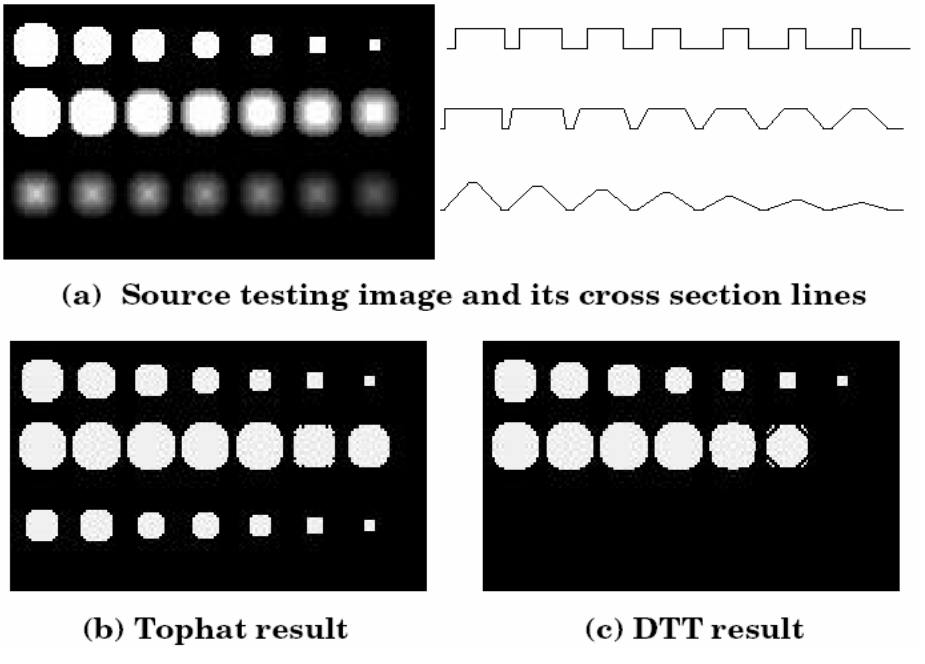


Fig. 2. A test result comparing the TT and DTT algorithms

2.3 Morphological Reconstruction

Mathematical morphology is a powerful methodology for the quantitative analysis of geometrical structures where the Morphological Reconstruction is employed to reconstruct character regions. It is defined as:

$$B_i = (B_{i-1} \oplus_g k) \bigcap | f |_G \quad (B_i \in R^3, i = 1,2,...) \tag{4}$$

In the above, i is a scale factor and K is the basic structuring element (e.g. 1 pixel radius disk). $\oplus_{gray}$ denotes a dilation operation in grayscale, and $| f |_G$, represents the *mask* of the operation, achieved via a threshold operation using a gray level G. The iteration in (4) is repeated until there is no further change between B_{i-1} and B_i. It is depicted in Fig.3.

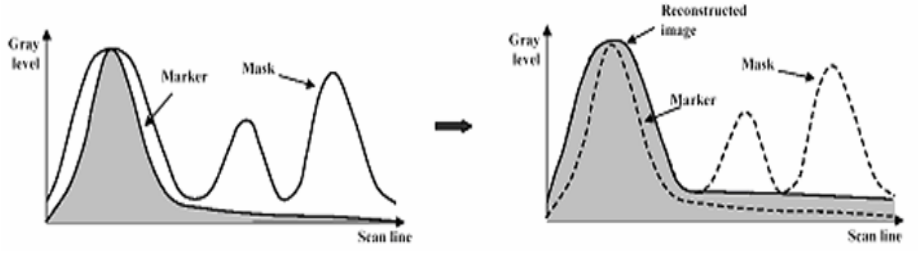


Fig. 3. Morphological Reconstruction in grayscale where regions in marker image are used to select regions of the mask image to be reconstructed

3 Character Region Identification Strategy

3.1 Prior Knowledge

Character regions have many features, which can be employed as restrictions to distinguish characters from a cover image. In this research, characters in cover image are considered satisfying the next conditions:

1. Character regions are composed of at least 3 characters;
2. Character regions are monochrome with a good contrast;
3. A character should be in a single gray level.

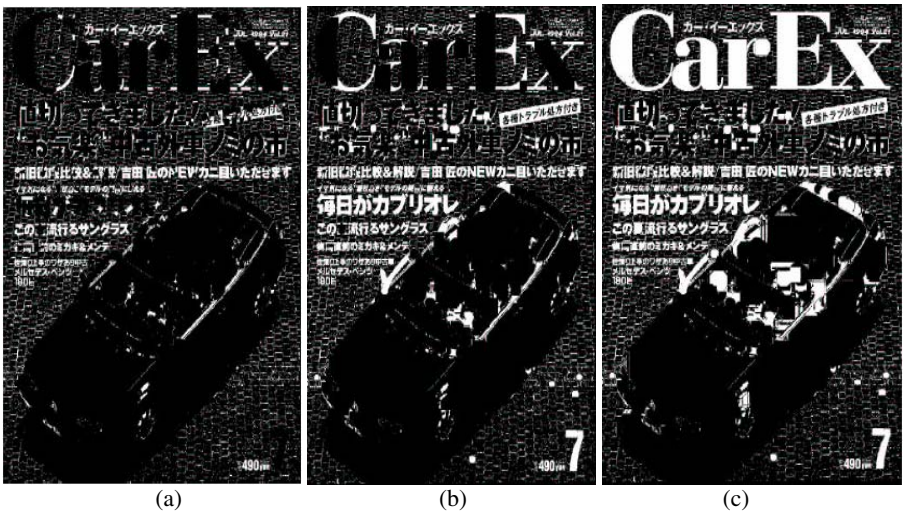
3.2 Identification Processing

The entire procedure can be divided into two steps: region segmentation processing and character extraction processing.

Region Segmentation Processing

The DTT algorithm is employed in this step. Where source image F is in gray-scale and the segmented sub-images are thresholded into binary images. Equation (2) and

(3) are performed respectively correspondence to different input images. In equation (2), all the regions which hold the same width as the disk shaped SE and brighter than the surrounding areas are detected into an image F_i. Consequently, the regions, which are smaller than specified SE and brighter than the surrounding areas (holding high gray levels) are collected together into a sub-image. F'_i In the same way, equation (2) detects all the regions which are darker than the surrounding areas (holding low gray levels). The two equations are recurrently performed i_{max} times, which is determined in the width of the largest character lines ($2\ i_{max}+1$) in the source image. The mentioned threshold value is determined experimentally. The examples of resulting sub-images are shown in Fig. 4.



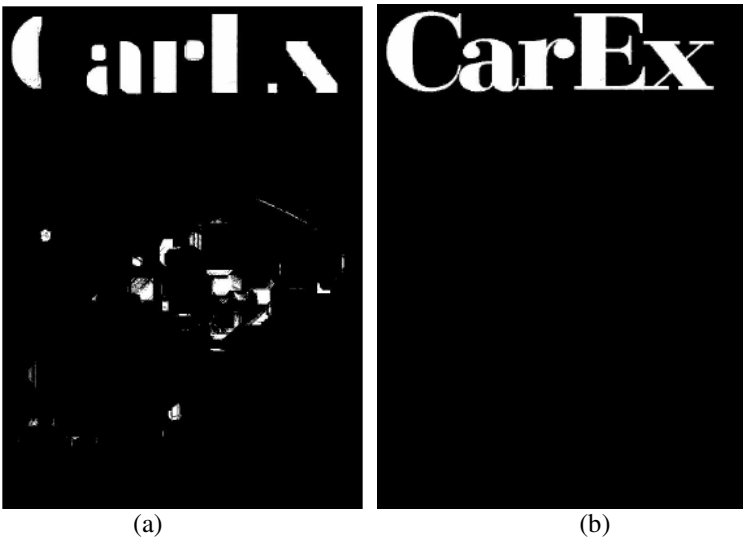
(a) (b) (c)

Fig. 4. The examples of resulting sub-images of DTT. (a) i = 4; (b) i = 10; (c) i = 30

Character Region Identification Processing

Since the character regions coexist with noises in the extracted sub-images, we firstly employ a morphological noise reduction algorithm based on dilation and opening operations to emphasize the character regions and suppress the noises. The resulted sub-images are transformed into gray-level images by an arithmetic multiplication between two same size images (source image and a sub-image).

Then, the character candidate regions are extracted by a histogram analysis technique. Since character regions hold peak gray values in the emphasized sub-images, the gray levels of character regions can be identified by detecting the peak values, which are bigger than the average value of all the peak values in their histogram curve. The detected peak value may be more than one because the character regions with different gray levels might coexist in one image. We employ all the peak values, which are bigger than the average one as the threshold levels. The sub-images are thresholded into binary images by the values.



(a) (b)

Fig. 5. The examples of identified character regions. (a) a sub-image with candidates; (b) reconstructed character region

Since there still remain some noises and part of character regions are lost, we finally reconstruct these sub-images using the morphological reconstruction algorithm in the last step. The examples are shown in Fig. 5.

4 Experimental Results

Gray-scale cover images of several kinds of magazines (a category of cover images) were employed to test our proposed algorithm. Although characters in some of them lay on a flat background, some of others are intricately composed of pictures and characters. They were regarded as a typical representative of covers with much variation.

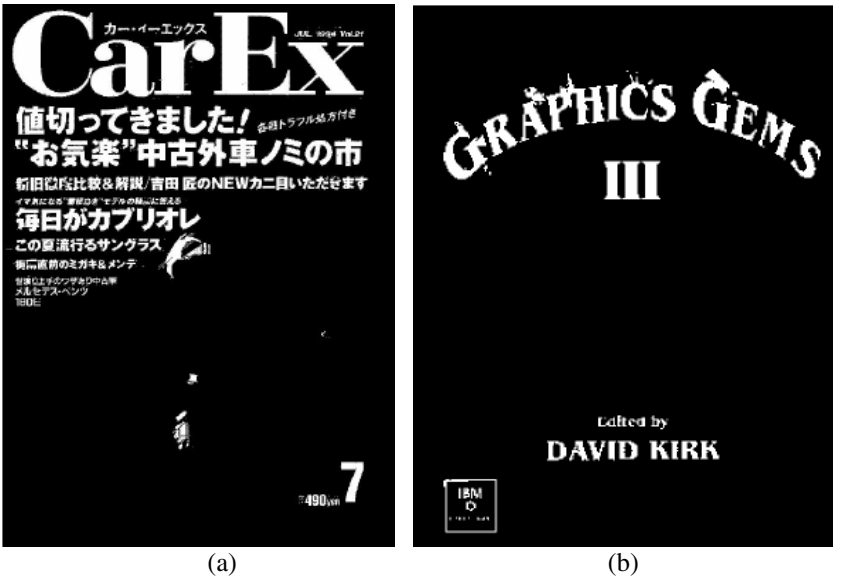
A cover image database with 30 gray scale images were constructed and they were tested in the experiment to demonstrate the robustness of the proposed approach. They were scanned in 100dpi and 1170 848 pixels. Two examples are shown in Fig.6.

The procedure of the experiment was described fully in the last section. In the segmentation processing, we defined $i_{max} = 40$ because the largest character line in our source images was smaller than 81 pixels (diameter of a disk with radius of 40) according to a statistical inspection.

As shown in Fig.7, character regions were efficiently extracted using the proposed algorithm. Some inputs, which are quite simple similar to a text page, achieved significant high extraction rate (99%) except the extremely small character regions. Only increasing the scanning resolution can solve this common problem. There are

(a) (b)

Fig. 6. Example of the source cover images. (a) CAREX; (b) Gems



(a) (b)

Fig. 7. Example of the results. (a) CAREX; (b) Gems

many other inputs in our database, which are much complicated because the characters are mixed with their background such as the "Gems" shown in Fig.6 (b). A 92% extraction rate acquired for this case. The extraction rated is defined as the

percentage between the extracted character regions and the total character regions in the inputs. Finally, among the database, an average of 94% character regions were correctly identified.

5 Conclusion

In this paper, a novel approach to identify character regions using DTT from cover images were proposed. The character regions in the cover images can be imagined as that they are ``floating on the background or sinking under it'' with a flat texture. We devoted our attention to these features and proposed the characters extraction system based on the idea of detecting “thin and long” regions. As examples, cover images of magazines were employed in the experiment. The results show that even for some complicated images, good extraction results were achieved with an average extraction rate of 94%.

As the future work, the algorithm needs improvement to deal with some character regions in the complicated images, which could not be automatically identified so far. The approach need improve to resist noises around characters in the resultant images. We are scheduling to test more cover images and finally apply the algorithm into an OCR system to facilitate the conversion of printed documents to electronic document in organizing a digital library system.

References

1. S.Liang and M.Ahmadi, “A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background Images”, Computer Vision, Graphics, Image Processing, Vol.56, No.5, pp.402-413, Sep., 1994.
2. H.Goto, H.Aso. “Character Pattern Extraction Based on Local Multilevel Thresholding and Region Growing”. Proc. of International Conference on Pattern Recognition (ICPR'00)-Volume 4: pp.4430–4433, 2000.
3. K.Marukawa, T.Hu, H.Fujisawa and Y.Shima, “Document Retrieval Tolerating Character Recognition Errors - Evaluation and Application”, Pattern Recognition, Vol.30, No.8, pp.1361-1371, Aug. 1997.
4. A.K.Jain and Y.Zhong, “Page Segmentation Using Texture Analysis”, Pattern Recognition, Vol.29, No.5, pp.743-770, May 1996.
5. P.Tofani and R.Kasturi, “Segmentation of Text from Color Map Images”, Proc. of 14th International Conference on Pattern Recognition(ICPR'98), pp.945-947, August 1998.
6. X.Ye, M.Cheriet, and C.Y.Suen,, “Stroke-Model-Based Character Extraction from Gray-Level Document Images”, IEEE Trans. on Image Processing, Vol. 10, No. 8, pp.1152-1161, August, 2001.
7. L.Gu, T.Kaneko, and N.Tanaka, “Morphological Segmentation Applied to Character Extraction from Color Cover Images”, Mathematical Morphology and Its Applications to Image and Signal Processing, Kluwer Academic Publishers, pp.367--374, 1998.
8. J.Ohya, A.Shio and S.Akamatsu, “Recognizing character in scene images”, IEEE Trans. on Pattern Anal. Machine Intell., Vol.16, No.2, pp.214-220, 1994.

9. S.Liang and M.Ahmadi, “A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background Images”, Computer Vision, Graphics, Image Processing, Vol.56, No.5, pp.402-413, Sep., 1994.
10. M.-Y.Chen, A.Kundu and S.N.Srihari, ”Variable Duration Hidden Markov Model and Morphological Segmentation for Handwritten Word Recognition”, IEEE Trans. on Image Processing, Vol. 4, No. 12, pp.1675-1688, Dec., 1995.
11. F.Meyer, “Contrast Feature Extraction”, Quantitative Analysis of Microstructures in Material Sciences, Biology and Medicine, J.-L. Chermant, ed., Special issue of Practical Metallography}, Riederer Verlag, Stuttgart, Germany, 1978.