A System to Retrieve Text/Symbols from Color Maps using Connected Component and Skeleton Analysis

Partha Pratim Roy[1], Eduard Vazquez[1], Josep Lladós[1], Ramon Baldrich[1], and Umapada Pal[2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona,08193, Bellaterra (Barcelona), Spain
[2] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata - 108, India

Abstract. Automatic separation of Text and Graphic Symbols in document images is one of the fundamental aims in Graphics Recognition. In Maps, separation of Text and Graphic Symbols involves many challenges because the text and symbol frequently touches/overlaps to the long lines of street, river, border of the regions etc. of the maps with similar color. Sometimes the colors in a simple character are gradually distributed which adds extra difficulty in the problem. In this paper we proposed a system to retrieve text/symbols from maps. Here, at first, we separate the maps into different layers according to color features and then connected component features and skeleton information are used to identify text characters, symbols from graphics on the basis of their geometrical features. From the experiment we obtained encouraging results.

1 Introduction

Various studies related to information retrieval in maps have been undertaken[1, 2]. The aim is to segment the document into two layers: a layer assumed to contain text and symbols and the other one containing the rest of graphical objects. Many methods have been proposed in binarized maps to extract text from cluttered background [5]. The problems arise due to presence of overlapping of text/symbols with lines. There exist color segmentation methods for separating different colors used in a document. But, the degradation effect of color in text layer sometimes makes it more difficult to separate them. It causes over/under -segmentation and such over/under -segmentation creates problem in separating foreground layer efficiently. To handle such situations, the objective of this paper is to combine different features like color information, connected component analysis, skeleton information etc. to get better results.

2 Text/Graphics layer Selection

The problem of foreground detection in color maps could be defined as that of detecting layers containing text/long lines. If the text layer is assumed to be of

dark color in a light background, then the problem can be solved by converting the RGB color-space to YIQ color-space, and applying a threshold in Intensity (Y channel) image. In color degraded image, this method is not efficient to separate foreground layer. For our map handling, due to image degradation, we performed the color segmentation by a combination of color feature and spatial information. This is followed by selection of foreground layer considering the features of text/line information. It is done by applying a heuristic measures on color volume and edge information of each color layers. The detail of color segmentation is as follows.

First, we apply the method of Vazquez et al.[3] to find dominant colors in an image. The creaseness operator, `MLSEC-ST`[4] is introduced in order to spurn non-representative data as well as to enhance meaningful information. Due to acquisition and image compression conditions, dominant colors may be affected by noise and to take care of this situation different maxima in creaseness distribution are joined by ridge extraction algorithm[10]. All different maxima of a representative area of a dominant color are joined. Next, using topological information obtained in the creaseness process, we find influence zone of every ridge e.g., set of data which can be unequivocally represented by a ridge. Then, the image is segmented in different representative colors for every ridge and its influence zone. In order to improve segmentation results, we apply an psychophysical method using induction process in initial stage. Induction is based on the idea of chromatic contrast and it enhances chromatic differences by means of a wavelet decomposition [7] to build the perceived image.

3 Text/Symbol separation from long lines

3.1 Connected Component Analysis with Skeleton Information

The text/symbols and graphics lines are extracted in foreground layer. The methods proposed in [1] have been developed on connected component analysis. A few criteria based on geometrical features of the connected component are good enough to group a component into one between Text or Graphics layer. But, there are some constraints. For example, the characters in some words cannot be split due to noise. We will call them as "joined characters". If joined characters touch with long lines, they would not be separated by simple rules, because their features will be different from isolated character. The stroke information is necessary to separate them from long line. We integrate skeleton information along with geometrical features to detect the long segments and to analyze them accordingly. We separate the components among 5 groups namely, Isolated Character, Joined Characters, Dash components, Long line components and Mixed components. If there exists no long segments from skeleton, they are grouped into Isolated Characters or Symbols, Connected Characters and Dash Characters. Otherwise, they are mixed or long characters. The description of each of them is given below. The fig. 1 shows different components of a foreground layer.

In *Isolated Character* normal Text alphabets and small symbols are included. *Joined Characters* consists of the components where more than one isolated character touches each other. *Dash Characters* are mainly small elongated components. These include the dash segments from the dash line along with some Isolated Characters, such as "1", "l" etc. These characters are combined into dash character group, because, at the pixel level analysis, they hold the same property as dash segments. The *Long Line* is the Graphics layer of our algorithm. The segments of these components are larger compared to the Text characters. Straight and curve both types of line can be possible and their detection procedure is discussed later. *Mixed Character* consists of the components where both long line and isolated/joined characters are present. This happens due to overlapping with each other. And at the component level analysis they cannot be separated. We will detect and mark this type of component for further analysis with other features.

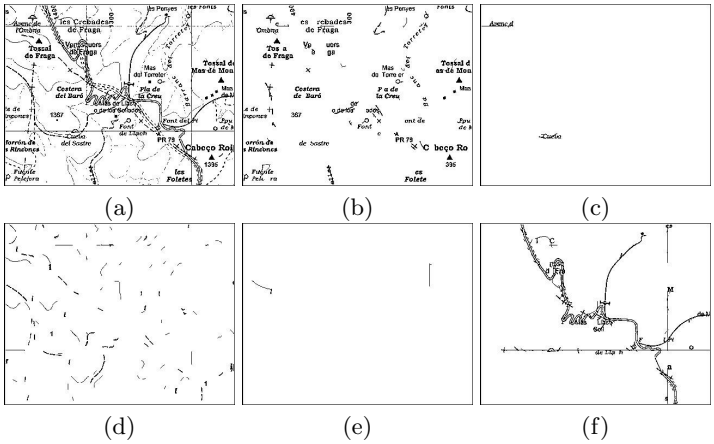


Fig. 1. (a)Foreground Layer (b)Text components (c)Joined Character (d)Small elongated components (e)Long line (f) Mixed components

3.2 Long Straight Line Removal

We perform Hough Transform to detect the straight lines present in the Binarized image. In Hough space, all the collinear pixels of a straight line will be found intersecting at the same point (ρ, θ), where ρ and θ identify the line equation. Depending on accumulation of pixels, the straight lines are sorted out. The continuous lines of a given threshold are first selected for separation. Width of each pixel in the line can be computed from the boundary pixels of the concerned line in the image. The part of the line will be removed when the corresponding width will be less than a threshold measured from the average of the total line width.

3.3 Curve Line Separation

According to text and graphics feature, it is assumed that the length of segments of the characters are smaller compared to that of graphics. The method proposed by Cao and Tan [2] which is based on the continuation of the strokes in the skeleton works well for documents, where the text and lines are of more or less thin lines. But, there are some limitations. When a line touches a symbol or text of blob like shape (dense pixels), the thinned image is always not perfect for the arrangement of segments. It needs post-processing, which is a difficult job. We started our curve line removal method in that point of view. All the segments are decomposed first at the intersection point of the thinned image. Based on the bounding box information of segments, the major axis is calculated. The segments which have large major axis compared to character height are chosen for elimination.

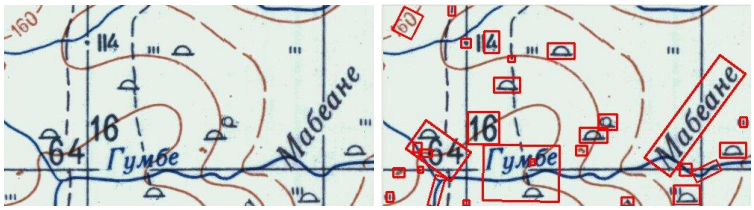
3.4 Character String Extraction

After passing through different separation methods, the mixed components will get separated. The long lines will be in the graphics layer and text/symbols will be in isolated character and joined character layer. These isolated and joined characters are combined to get all the text components. The characters of a single string can be grouped by different methods. We assume that, the gap between words T_w is larger than the gaps within words T_c and the grouping is formed by the characters of similar colors. This process consists of two-step analysis: color and proximity.

4 Experimental Results and Discussion

We have taken maps from different scripts to test our methods. 26 maps are selected from "Spanish", "English", "Russian", and "Bengali" and the average size of the test maps are 350x450 pixels. Our proposed methodology combines color information, connected component analysis and skeleton information for segmentation. The thinning was done by the algorithm proposed by Ahmed and Ward [8] which works in rotation invariant nature. The length of segments' diameter are used to separate long lines from small segments. The string construction and joining missing characters are done using morphological operation. The scale invariance is also incorporated by computing histogram analysis of the components' size, considering aspect ratio of a component. Extracted text/symbols are marked by rectangular box in Fig. 2(b).

Almost all the previous approaches in text graphics separation either used Binarized Image or converted the color image to Binarized image for this purpose. The work of Dhar and Chanda [9] used color information, which was specific to a particular map. The separation of text/graphics layer using color analysis is not an easy task. There exists no methodology to evaluate the correctness of



(a) (b)

Fig. 2. (a)Original Image (b)Segmented Text/Symbol layer. Here, segmented text / symbol parts are marked by rectangular box

color segmentation result. The validity of the results vary according to human perception and thus focus for measuring in terms of qualitative rather than quantitative. Color separation analysis using creaseness operator reduces the amount of variability of color information effectively. In our test maps, where the color degradation is less, it outperforms. But in maps of "Spanish" and "Bengali" the noise is very prominent and we got over segmentation. Here, we selected the foreground layers manually and used these layers for our text/graphics separation purpose. It is to be mentioned that in skeleton analysis, a curve line may not be fully removed, if it visits many junctions in the travel path which makes it to loose full curve information. Again, if a long line is not fully straight and it contains a sufficient straight part, this part will be detected by Hough Transform and will be removed leaving the other parts. The other segments will be classified by CC analysis and will cause false alarms in text layer construction. To get the idea of segmentation results of different scripts, see Fig. 3, where text/symbols are extracted from the color images.

References

1. K. Tombre, S. Tabbone , L. Peissier, B. Lamiroy, and P. Dosch. "Text /graphics separation revisited", *Proceedings of the 5th International Workshop on Document Analysis Systems ,Lecture Notes in Computer Science*, vol. 2423, Springer-Verlag, London, UK, pp. 200-211, 2002.
2. R. Cao and C.L.Tan. "Text/graphics separation in maps" In *Proceedings of 4th IAPR International Workshop on Graphics Recognition*, Kingston, Ontario(Canada), pp. 44-48, September 2001 .
3. Eduard Vazquez, Ramon Baldrich, Javier Vazquez and Maria Vanrell "Topological histogram reduction towards colour segmentation", *Lecture Notes in Computer Science - Pattern Recognition and Image Analysis*, pp. 55-62, 2007.
4. Antonio M. López, David Lloret, Joan Serrat and Juan J. Villanueva "Multilocal Creaseness Based on the Level-Set Extrinsic Curvature", *Computer Vision and Image Understanding: CVIU*, vol. 77, no. 2, pp. 111-144, 2000.
5. L. A. Fletcher and R. Kasturi. "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images". *IEEE Transactions on PAMI*, vol. 10, no. 6, pp. 910 - 918, 1988.

6. Partha Pratim Roy, " An Approach to Text / Graphics Separation from Color Maps", M.S. thesis. CVC, UAB, Barcelona. February, 2007
7. Otazu X, Vanrell M, "Several lightness induction effects with a computational multiresolution wavelet framework" *PERCEPTION*, vol. 35, pp. 56-56, 2006.
8. M. Ahmed and R. Ward, "A Rotation Invariant Rule-Based Thinning Algorithm for Character Recognition". *IEEE Transactions on PAMI*, vol. 24, no. 12, pp. 1672-1678, Dec. 2002.
9. D.B. Dhar and B. Chanda, "Extraction and recognition of geographical features from paper maps". *International Journal of Document Analysis*, vol. 8, no. 4, pp. 232-245, 2006,
10. Antonio M. Lopez, Juan J. Villanueva, Felipe Lumbreras and Joan Serrat, "Evaluation of methods for ridge and valley detection".*IEEE Trans. Pattern Anal. Mach.*, vol. 21, no. 4, pp. 327334, 1999.

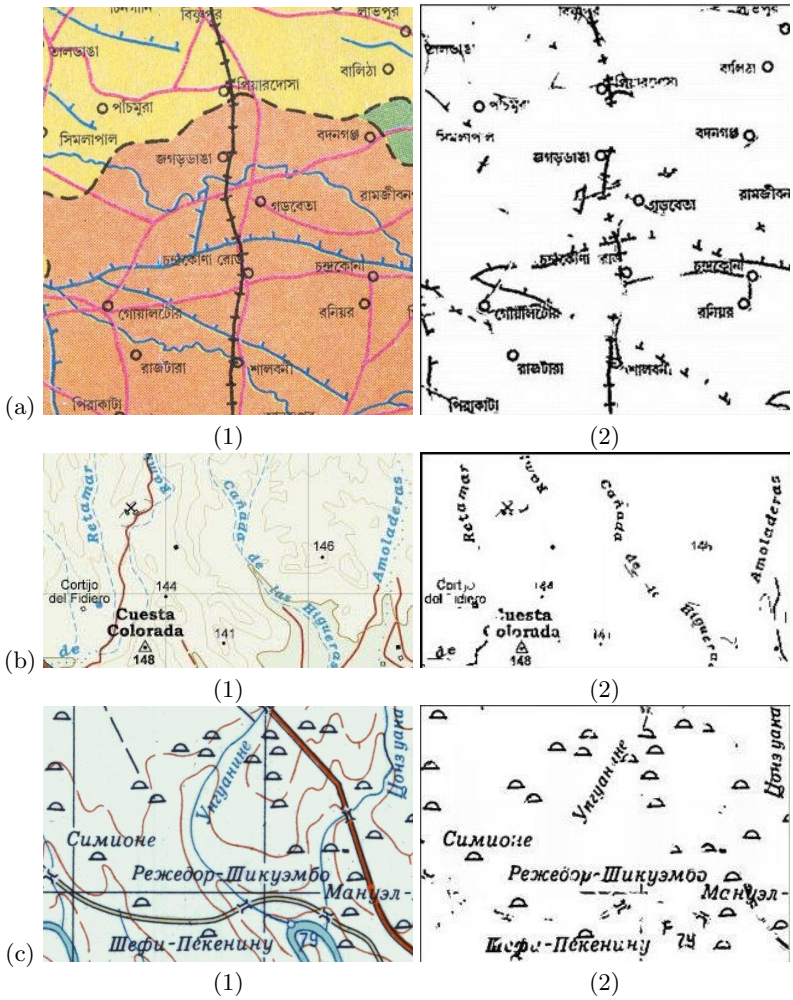


Fig. 3. Images in different scripts (a) Bengali (b) Spanish (c) Russian. In each script,(2) shows the extracted Text/Symbols of correspoding color image (1).