Computer and Information Science
www.ccsenet.org/journal.html
Vol. 1, No. 4
November 2008

# A Text Image Segmentation Method Based on Spectral Clustering

Rui Wu (Corresponding author)
School of Computer Science and Technology
Harbin Institute of Technology
PO box 352, Harbin, 150001, China
Tel: 86-451-8641-2979 E-mail: simple@hit.edu.cn
Jianhua Huang
School of Computer Science and Technology
Harbin Institute of Technology
PO box 352, Harbin, 150001, China
Tel: 86-451-8641-3631 E-mail: jhhuang@hit.edu.cn
Xianglong Tang
School of Computer Science and Technology
Harbin Institute of Technology
PO box 352, Harbin, 150001, China
Tel: 86-451-8641-3631 E-mail: txl60@public.hr.hl.cn
Jiafeng Liu
School of Computer Science and Technology
Harbin Institute of Technology
PO box 352, Harbin, 150001, China
Tel: 86-451-8641-3631 E-mail: Jeffery@hit.edu.cn

**Abstract**

We present a novel approach for solving the text segmentation problem in natural scene images. The proposed algorithm uses the normalized graph cut(Ncut) as the measure for spectral clustering, and the weighted matrices used in evaluating the graph cuts are based on the gray levels of an image, rather than the commonly used image pixels. Thus, the proposed algorithm requires much smaller spatial costs and much lower computation complexity. Experiments show the superior performance of the proposed method compared to the typical thresholding algorithms.

**Keywords:** Text segmentation, Graph cut, Spectral clustering

## 1. Introduction

Images generally contain rich messages from textual information, such as street name, construction identification, public transport stops and a variety of signal boards. The textual information assists the understanding the essential content of the images. If computers can automatically recognize the textual information from an image, it will be highly valuable to improve the existing technology in image and video retrieval from high-level semantics (Lienhart, 2002, pp.256-268). For instance, road signs and construction identification in a natural environment can be captured into images by cameras and the textual information will be detected, segmented, and recognized automatically by machines. These messages then can be synchronized as human voice to be used as instructions for visually impaired person. In addition to the example, textual information extraction plays a major role in images retrieval based on contents, cars auto-drive, vehicle plate recognition and automatics.

In general, automatic textual extraction consists of text detection, localization, binarization and recognition etc. In a natural scene texts could have different backgrounds and characters in the text message can also have variety of forms.

And, existing OCR (Optical Character Recognition)engine can only deal with printed characters against clean backgrounds and can not handle characters embedded in shaded, textured or complex backgrounds. So that characters are separated from the text in the detected region accurately is very necessary. Currently, many researchers have done a lot of work in the text detection and a lot of methods of text detection and location have been proposed. (Mariano, 2000; D. Chen,2004; Zhong,2000; X.L Chen,2004;X. Chen,2004) Compared to the text detection in natural scenes, specialized study of the characters extraction from natural environment is not more. The purpose of this paper is to extract accurate binary characters from the localize text regions so that the traditional OCR can work directly.

Most of the existing approaches are to use thresholding for binarization either global thresholds(T. Tsai,2007; Pan,2007)or local thresholds(Lienhart,2002;Wu,1999). T.Tsai(2007,pp.113-116) adopt a thresholding method suitable for segmenting the potential videotext character and modified seed-fill algorithm to extract the videotext. In Pan's(2007,pp.412-416), a simple global threshold achieved by using Otsu's(Otsu,1979,pp.62-66) thresholding technique. Lienhart(2002,pp.256-268) performed the binarization using the intensity value halfway between the intensity of the text colors and the background color as a threshold. Wu(1999,pp.1224-1229) proposed a simple histogram-based algorithm to automatically find the threshold value for each text region, making the text segmentation process more efficient. Due to its simplicity and efficiency, thresholding is a widely used method for solving this problem. But, it could not handle the cases when backgrounds have the similar color or intensity to that of the text strokes. Meanwhile, besides the changing backgrounds, texts are also changing slightly due to edge blur, image quality degrading due to video compression.

Spectral clustering has gradually gained attention from research on text classification, images segmentation and information retrieval (Shi,2000; Tao,2007). In Tao's (2007,110-118) the proposed algorithm uses the normalized graph cut(Shi,2000,pp.888-905) measure as the thresholding principle to distinguish an object from the background and a large number of examples are presented to show the superior performance of the method. But it is still a thresholding algorithm that has limitations to deal with the scene text. In this paper, we propose a new text segmentation method based on spectral clustering. In our approach, the histogram of intensity is used for the object of grouping, we partition the image into two parts using the gray levels of an image rather than the image pixels. For most images, the number of gray levels is much smaller than the number of pixels. Therefore, the proposed algorithm occupies much smaller storage space and requires much lower computational costs and implementation complexity than other similar algorithms.

The rest of the paper is organized as follows. Section 2 introduces the theory of spectral graph partition briefly. Section 3 presents the individual steps of our approach. The experimental results are given in section 4. Finally, section 5 concludes the paper.

**2. Theory of Spectral Graph**

The basic method used by image segmentation is to view an image as a weighted undirected graph $G = (V, E)$, where the nodes of the graph are the points in the feature space, and an edge is formed between every pair of nodes. The weight on each edge, $w(i, j)$, is a function of the similarity between nodes *i* and *j*. A graph $G = (V, E)$ can be partitioned into two disjoint subsets A and B, subject to $A \cap B = \phi, A \cup B = V$, by simply removing edges connecting the two parts. The degree of dissimilarity between these two pieces can be computed as total weight of edges that have been removed. In graph theoretic language, it is called the cut(Shi,2000,pp.888-905) :

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \tag{1}$$

The optimal bipartitioning of a graph is the one that minimizes this cut value. There are many criterions to measure the quality of the final partition results. Then the *Normalized Cut* value of a bipartition result can be defined as follows (Shi,2000,pp.888-905) :

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{2}$$

where $assoc(A, V) = \sum_{u \in A, v \in V} w(u, v)$ , $assoc(B, V) = \sum_{u \in B, v \in V} w(u, v)$ respectively, is the total connection from nodes in A or B to all nodes in the graph. And now, the minimal *Ncut* value is just corresponding to the optimal bipartition of the graph. In order to minimize (2), we can transform the optimization problem into solving the eigenvalue system,

$$D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} = \lambda z \tag{3}$$

where $D_{ii} = \sum_{j} w(i, j)$ ,W is a symmetric matrix with size of $N \times N$, $\lambda$ is the eigenvalue and $z$ is the corresponding eigenvector. Shi and Malik(2000,pp.888-905) have proved that the second smallest eigenvector of the eigensystem (3) is the real value solution to the normalized cut problem of (2).

When the size of an image is too big, it is difficult to solve the above eigensystem, especially if the affinity matrix W is constructed by taking each pixel as a node, the size of eigensystem would be $N \times N$ (N is the total number of pixels in an image).

**3. Our method**

Suppose

$V = \{(i, j) : i = 0,1,\cdots, n_h - 1; j = 0,1,\cdots, n_w - 1\}$, $H = \{H_0, H_1, \cdots, H_L\}$, $LL = \{0,1,\cdots, L\}$, where $n_h$ and $n_w$ is the height and the width of the image, respectively. $H$ represents the histogram of gray, $f(x,y)$ is the gray value of position $(x,y)$ .Then, $V$,$H$ and $f(x,y)$ satisfy the following formulas.

$$(x,y) \in H_l, l \in \{0,1,\cdots L\}, \qquad \forall (x,y) \in V \tag{4}$$

$$H_l = \{(x,y) : f(x,y) = l, (x,y) \in V\}, l \in LL \tag{5}$$

$$\bigcup_{l=0}^{L} H_l = V, H_i \cap H_j = \varnothing, i \neq j, \ \ i, j \in LL \tag{6}$$

Using just the intensity value of the pixels and their spatial location, we can define the graph edge weight connecting the two nodes $i$ and $j$ as:

$$w_{ij} = e^{\frac{-\|F(i)-F(j)\|_2^2}{\sigma_I}} * \begin{cases} e^{\frac{-\|X(i)-X(j)\|_2^2}{\sigma_X}}, & if \ \|X(i)-X(j)\|_2 < r \\ 0, & otherwise. \end{cases} \tag{7}$$

where $F(i)$ is a feature vector based on intensity of node $i$ ,and $X(i)$ is the spatial location at that node, $\sigma_I$ and $\sigma_X$ are scale factors used to adjust the variation of gray or spatial location between nodes, $r$ is used to decide the number of nodes from node $i$ to $j$ .

And then, we can get a bipartition $V = \{A, B\}$ corresponding to the graph $G = (V, E)$ , where

$A = \bigcup_{k \in L_A} H_k$ , $B = \bigcup_{k \in L_B} H_k$ , and $L_A \cap L_B = \varnothing, L_A \cup L_B = LL$ . Let $cut(H_i, H_j) = \sum_{u \in H_i, v \in H_j} w(u,v)$ be the total connection weights from nodes in $H_i$ with gray level $i$ to all nodes in $H_j$ with gray level $j$ , we can rewrite the above formulas as:

$$cut(A,B) = \sum_{i \in L_A} \sum_{j \in L_B} cut(H_i, H_j) \tag{8}$$

$$asso(A,A) = \sum_{i \in L_A} \sum_{j \in L_A} cut(H_i, H_j) \tag{9}$$

$$asso(B,B) = \sum_{i \in L_B} \sum_{j \in L_B} cut(H_i, H_j) \tag{10}$$

Since $asso(A,V) = asso(A,A) + cut(A,B)$ , $asso(B,V) = asso(B,B) + cut(A,B)$ , we can rewrite (2) as:

$$Ncut(A,B) = \frac{cut(A,B)}{asso(A,A) + cut(A,B)} + \frac{cut(A,B)}{asso(B,B) + cut(A,B)} \tag{11}$$

Given an image, we can construct a histogram-based matrix $M$ by computing the all weights of nodes in the corresponding graph. $M = [m_{i,j}]$ is an $L \times L$ symmetrical matrix with $m_{i,j} = cut(H_i, H_j)$ and $m_{i,j} = m_{j,i}$ ,where $L$ is the number of gray level of histogram. Now, let $M$ be the affinity matrix, we can get a complete approach of image segmentation using spectral clustering (Shi,2000,pp.888-905) . Figure 1 shows the workflow.

Note that the size of the affinity matrix $M$ depends on the number of gray-level $L$ , rather than the number of all pixels $N$ in an image. Meanwhile, the size of eigensystem to solve is $L \times L$ , rather than $N \times N$ , and usually, $L$ with a fixed size is much smaller than $N$ . Hence, the complexity of computation and spatial

cost reduce greatly.

**4. Experimental Results**

We perform a series of experiments to test the performance of this method. The samples used are gray images with the text of characters, and they are the real images from the natural environment. For illumination and other reasons, there is a clear gray difference of pixels from the same region in many images. In the following experiments the parameter settings in formula (7) are $\sigma_I = 50$, $\sigma_X = 5$, $r = 5$, $L = 100$. Our method is compared with two other methods: the *Otsu* thresholding method (Otsu,1979, pp.62-66) and the *Ncut-based* thresholding method (Tao,2007,pp.110-118). We choose them because the *Otsu* method is a simple but classic solution employed by many text segmentation schemes, while the latter is an *Ncut-based* but thresholding solution proposed recently. The aim of the three algorithms is to separate the "foreground (texts)" from the "background (non-texts)".

We present the detailed results in Figure 2~Figure 4. By the way, the actual images are much greater than that in the paper. The images are easy to separate relatively in Figure 2, we can see the latter two methods both based on *Ncut* criterion are superior or close to the *Otsu* method from the experimental results. In Figure 3, the proposed method can be the right segmentation at a reflective white spots within the part of black spots on the letter '*B*', which is difficult to achieve for the conventional thresholding methods.

It is hard to segment the images in Figure 4 because the foreground of them is not clear enough. The segmentation results of our approach look better than the others'. From the above experiments, we can see the proposed method is far superior to *Otsu* method, also to the thresholding method based on *Ncut*.

Next, we choose 150 images with clear differences and 50 images with blur text respectively to test the three methods. The precision $p$ and recall $r$ are used to evaluate the methods, which are defined as follows.

$$p = \frac{com\_\mathrm{Im}age_O \cap base\_\mathrm{Im}age_O}{com\_\mathrm{Im}age_o} \tag{12}$$

$$r = \frac{com\_\mathrm{Im}age_O \cap base\_\mathrm{Im}age_O}{base\_\mathrm{Im}age_O} \tag{13}$$

where $com\_\mathrm{Im}age_O$ is a points set of objects including texts; $base\_\mathrm{Im}age_O$ is the ground-truth text region. And then, the comprehensive assessment indicator $f$ is defined as follows (Lucas,2003,pp.682-687):

$$f = \frac{1}{a/p + (1-a)/r} \tag{14}$$

where $a$ is a relative rate between precision and recall, letting $a = 0.5$.

The result of 150 images shows in table 1, and the result of 50 images shows in table 2. From the results, we can see the proposed method has a good performance for both the two kinds of images. And our method is superior to other two methods from the point of comprehensive indicator $f$.

**5. Conclusion**

Accurate retrieval of textual information from images that exist in the real environment is critical to understand images. The key part of the research is to retrieve the characters from the text image area. And because of the complexity of backgrounds that text built in, the conventional thresholding methods often cannot separate the characters from natural backgrounds effectively. Spectral clustering can resolve the issue by using spectral graph theory. And this method controls the complexity of algorithm effectively by changing the clustering objects from pixels to gray levels. The experiment results have proved its superiority to the traditional thresholding method.

**References**

D. Chen, J.M. Odobez and H. Bourlard.(2004) "Text detection and recognition in images and video frames". *Pattern Recognition*, 37(3): 595-608.

Jianbo Shi and Jitendra Malik.(2000) "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888-905.

Lienhart R and Wernicke A.(2002) "Localizing and Segmenting Text in Images and Videos." *IEEE Transaction on Circuits and Systems for Video Technology,* 12(4): 256-268.

Lucas S M, Panaretos A, Sosa L, et al.(2003) "ICDAR 2003 Robust Reading Competition." *Proc. of 7th Int'l Conference on Document Analysis and Recognition*, pp.682-687.

Otsu N.(1979) "A threshold selection method from grey – level histograms." *IEEE Trans System. Man Cyberne*, SMC-9: 62 - 66.

Tao Wen-Bing and Jin Hai.( 2007) "A New Image Thresholding Method Based on Graph Spectral Theory." *Chinese Journal of Computers(in Chinese)*,Vol.30, No.1,pp.110-118.

Tsung-Han Tsai and Yung-Chien Chen.(2007) "A Comprehensive Motion Videotext detection Localization and Extraction Method." *Proc. of IEEE Int'l Conference on Data Engineering Workshop*, pp. 113-116.

V.Y. Mariano and R. Kasturi.(2000) "Locating Uniform-colored Text in Video Frames". *Proc. of Int'l Conference on Pattern Recognition*, 4:539-542.

W.M.Pan, T.D.Bui and C.Y.Suen.(2007) "Text Segmentation from Complex Background Using Sparse Representations." *Proc.of Int'l Conference on Document Analysis Recognition*, pp. 412-416.

Wu V, Manmatha R, and Riseman E M.(1999) "Text Finder : an automatic system to detect and recognize text in images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1224 -1229.

Xiangrong Chen and Alan L.Yuille.(2004) "Detecting and Reading Text in Natural Scenes". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.366-373.

Xilin Chen, Jie Yang, Jing Zhang, and Alex Waibel.(2004) "Automatic Detection and Recognition of Signs From Natural Scenes". *IEEE Transactions on Image Processing*, Vol.13, No.1, pp. 87-99.

Yu Zhong, Hongjiang Zhang and A. K. Jain.(2000) "Automatic Caption Localization in Compressed video". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4): 385-392.

Table 1. The segmentation result of three methods for normal images

| Method | *p* | *r* | *f* |
|---|---|---|---|
| Ostu method | 0.8089 | 0.8361 | 0.8223 |
| Ncut-based thresholding method | 0.8364 | 0.8011 | 0.8184 |
| Our method | 0.8058 | 0.8802 | 0.8413 |

Table 2. The segmentation result of three methods for abnormal images

| Method | *p* | *r* | *f* |
|---|---|---|---|
| Ostu method | 0.5307 | 0.7734 | 0.6295 |
| Ncut-based thresholding method | 0.5815 | 0.8316 | 0.6844 |
| Our method | 0.6842 | 0.8015 | 0.7382 |

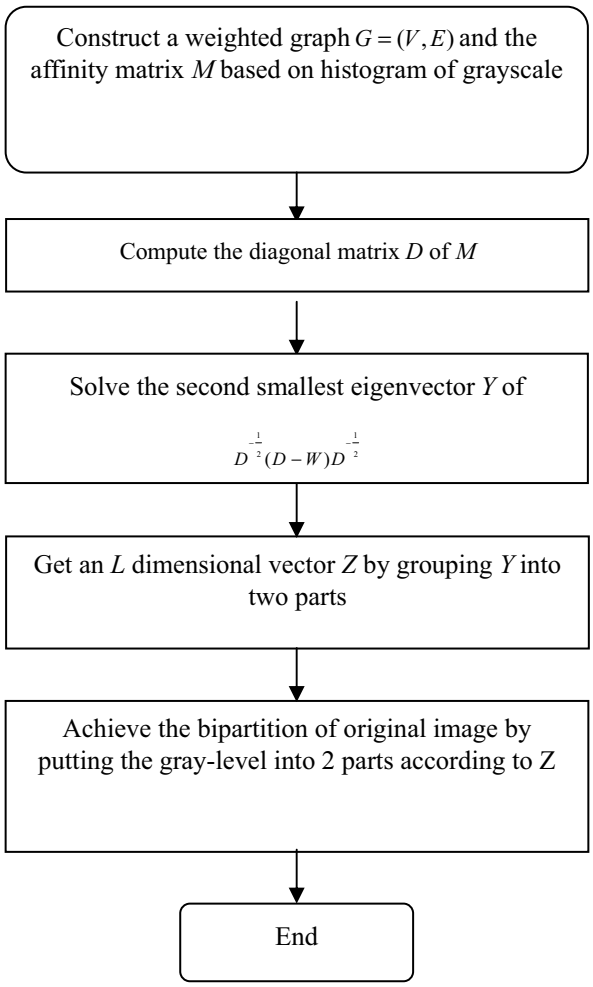
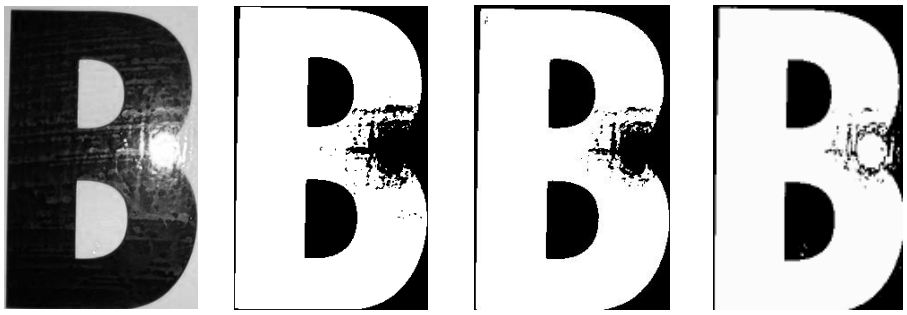



Figure 1. Workflow of Image Segmentation Based on Spectral Clustering


Figure 2. Comparison of Three Text Segmentation Methods for Normal Images



(a) (b) (c) (d)

Figure 3. Comparison of Three Text Segmentation Methods for an Illuminated Image ((a)Original image (b) Ostu result (c) Ncut-based result (d) Our result)


Figure 4. Comparison of Three text Segmentation Methods for Abnormal Images