COLOR TEXT IMAGE BINARIZATION BASED ON BINARY TEXTURE ANALYSIS

Bin Wang *Xiang-Feng Li* *Feng Liu* *Fu-Qiao Hu*

Instit. Image Processing & Pattern Recognition
Shanghai Jiaotong University
Shanghai 200030, People's Republic of China

ABSTRACT

In this paper, a novel binarization algorithm for color text images is presented. This algorithm effectively integrates color clustering and binary texture analysis, and is capable of handling situations with complex backgrounds. In this algorithm, dimensionality reduction and graph theoretical clustering are first employed. As the result, binary images related to clusters can be obtained. Binary texture analysis is then performed on each candidate binary image. Two kinds of effective texture features, run-length histogram and spatial-size distribution related respectively, are extracted and explored. Cooperating with an LDA classifier, the optimal candidate of the best binarization effect is obtained. Experiments with images collected from Internet has been carried out and compared with existing techniques, both show the effectiveness of the algorithm.

1. INTRODUCTION

Text is a kind of important feature for information retrieval applications. For these applications, as the first step, text in the image should be recognized for its semantic meanings. But current optical character recognition (OCR) technologies are mostly restricted to recognize text against clean backgrounds. Thus binarization techniques, which aim to separate text from image backgrounds and obtain a clean representation, are usually adopted as an indispensable preprocessing.

Most existing binarization techniques are thresholding related. Basically, these techniqeus can be categorized into two categories: global and local or adaptive. Global methods tempt to binarize the image with a single threshold. Among some most powerful global techniques, Otsu's algorithm can achieve high performance in terms of uniformity of thresholded regions and correctness of segmentation boundaries [1]. In [2], Liu and Srihari used Otsu's algorithm to obtain candidate thresholds. Then, texture features were measured from each thresholded image, based on which the best threshold were picked. In contrast to global ones, adaptive or local methods change the threshold dynamically over the image according to local information. In [3], an adaptive algorithm was developed by Wellner for the DigitalDesk. The method calculated threshold values at each point of estimated background illumination based on a moving average of local pixel intensities. For images with low contrast, variable background intensity and noise, local algorithms work well. However, it appeared that techniques of both above categaries perform poorly under complex backgrounds.

In this paper, a novel binarization algorithm is proposed. It is designed to overcome the limitations of existing techniques for color text images. The proposed algorithm efficiently integrates color clustering and binary texture feature analysis. Two kinds of features capable of effectively characterizing text-like binary textures, including run-length histogram and spatial-size distribution related features, are explored. In addition, a combination strategy is implemented among the binary images obtained by color clustering. The integration of these techniques enable the algorithm survives those complex background conditions, as existing techniques tend to fail.

2. BINARY TEXTURE ANALYSIS

Texture is one of the main features intensively utilized in image processing and pattern recognition. Other than the numerous existing texture description methods based on second order statistics, two kinds of new features are taken into account in this paper, one is based on run-length description, another spatial-size distribution related. Both these features are capable of effectively characterizing text-like textures in binary images.

2.1. Run-length based features

A run is a maximum contiguous set of constant-gray-level pixels located in a scan line, which can then be described by gray-level a, length r, and direction θ, denoted as $B(a,r,\theta)$. Let $B(a,r)$ be the number of runs in all directions of length r and gray-level a. Texture description features can be obtained from computation of continuous probability of the length and gray-level of runs in the image.

In this paper, study objects are binary images, and only foreground horizontal runs are of interest; thus $a \equiv 1$ and $r \in [1 \ldots L]$, where L is the maximum acceptable run-length. Then it is convenient to shorten *B(a,r)* to *B(r)*. Denote *B(R)*, *R*=[1 . . . *L*], the one-dimensional array contains the frequencies of each run, called run-length histogram. Examples of run-length histograms for different binay images are shown in Figure 1.

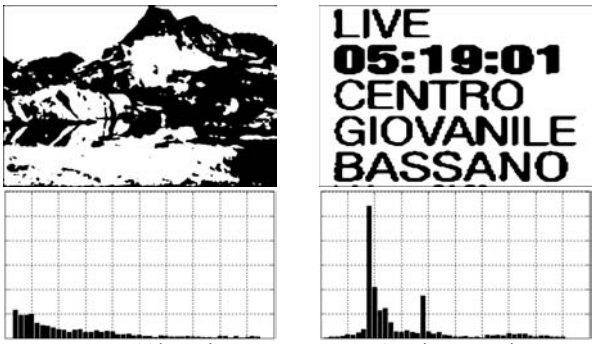
Features of interet are the maximum probability and the stroke width. The maximum probability is defined as the highest frequency in the run length histogram, excluding the unit run-length, which is always related to noise. That is,

$$MAXPROB = max_{r \in R} B(r), r \neq 1. \quad (1)$$

And the run-length of the highest frequency is:

$$SW = arg\ max_{r \in R} B(r), r \neq 1. \quad (2)$$

It actually reflects the average stroke width of the dominant text in the image, thus in [2], it was directly called stroke width feature.



(a) Normal binary image (b) Binary text image

Fig. 1. Different binary images and their corresponding run-length histogram

2.2. Spatial-size distribution related features

Binary images are the main concern of this paper. Thus, there is no variation of intensity as in general gray-level texture analysis. The most important here are the objects or components in the binary image and their spatial arrangement patterns.

The run-length histogram is a 1-D global representation of the image. Though it contains the essential information for stroke-like features, it is still incompetent to convey the true 2-D distribution pattern in the image. Mathematical morphology has been proverbially recognized as a powerful tool for binary image analysis. The granulometric size distribution, or pattern spectrum [4], is one of the main mathematical morphology related approaches to texture description. But limitation of pattern spectrum is that it conveys only the object size distribution in the image, and very different textures may share the same pattern spectrum. This has been overcomed by a new spatial size distribution (SSD) descriptor [5], through incorporating spatial distribution information into its conventional counterparter.

The *(p,q)*-SSD of the image G with respect to a convex and compact set U containing the origin is defined as [5]:

$$F_{G,U}(\lambda_1, \cdots, \lambda_p, \mu_1, \cdots, \mu_q) = \frac{1}{v(G)^{q+1}} \cdot \int_{\mu_q U} \cdots \int_{\mu_1 U} v(G \cap G + h_1 \cap \ldots \cap G + h_q) - v(\tilde{\Psi}(G) \cap \tilde{\Psi}(G) + h_1 \cap \ldots \cap \tilde{\Psi}(G) + h_q) dh_1 \cdots dh_q \quad (3)$$

where *v(•)* stands for the area, and

$$\tilde{\Psi}(G) = \Psi^{(1)}_{\lambda 1}(\cdots(\Psi^{(1)}_{\lambda_1}(G))\cdots), \quad (4)$$

is the composition of the different granulometries. The joint density function associated with $F_{G,U}$ is:

$$f_{G,U}(\lambda_1, \ldots, \lambda_p, \mu_1, \ldots, \mu_q) = \frac{\partial^{p+q} F_{G,U}(\lambda_1, \cdots, \lambda_p, \mu_!, \cdots, \mu_q)}{\partial \lambda_1 \ldots \partial \lambda_p \partial \mu_1 \ldots \partial \mu_q}. \quad (5)$$

For discrete sets or discrete images, it can be written as:

$$f_{G,U}(\lambda_1, \ldots, \lambda_p, \mu_1, \ldots, \mu_q) = \sum_{(u,v) \in N} (-1)^{sgn(u,v)} F_{(G,U)}(u, v) \quad (6)$$

where

$$N = \{(u, v) : u_i = \lambda_i \text{ or } u_i = \lambda_i - 1; \; v_j = \mu \text{ or } v_j = \mu_j - 1\}$$

and

$$sgn(u, v) = \sharp\{u_i : u_i = \lambda_i - 1\} + \sharp\{v_j : v_j = \mu_j - 1\}.$$

In this paper, SSD features of order (1,1) and (2,1) are of interest. The value of element size (λ) is set the same as that of the stroke-width feature of the image; the support-set size (μ) varies from 0 to 9 in increments of 3. Thus total 8 SSD features can be obtained for each binary image.

3. COLOR TEXT IMAGE BINARIZATION

The proposed binarization method for color text images consists of four main steps: color space dimensionality reduction, color clustering, texture feature extraction, and selection of the optimal binary image. The flowchart is shown in Figure 2.

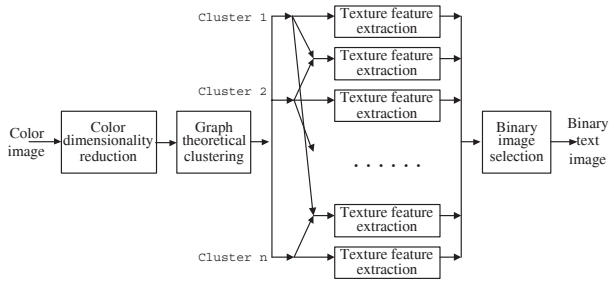


Fig. 2. Flowchart of the proposed algorithm

3.1. Dimensionality reduction and clustering

Considering properties of human vision, there is large amount of redundancy in the 24-bit RGB representation of color images. In [6], Jones et al reported that 77% of the possible 24-bit RGB colors were never encountered. In this paper, their research result is employed to serve as effective preprocessing. In our application, we found that representing each of the RGB channels with only 4 bits introduced little, or even no perceptible visual degradation, as shown in Figure 3(b). Another attractive feature of this operation is its conveniency, simply through performing a 4-bit right-shifting on each RGB channel.

Though the dimensionality of the color space has been dramatically reduced, it is still of 16x16x16; an unsupervised graph theoretical clustering [7] is employed for further information congregation. Unlike other techniques, the graph theoretical clustering does not act on image pixels directly, but on the condensed 3-D color histogram. Thereby, it suffers little from variation of image sizes. Figure 3 gives an example of the dimensionality reduction and graph theoretical clustering: image (a) is the original color image of 49,789 colors; image (b), dimensionality reduction reduces the number of colors to 816; after further clustering, in image (c), only 5 colors (clusters) remain.

(a) Original image (b) Dimension reduction (c) Color clustering

Fig. 3. Color dimensionality reduction and grap-theoretical clustering

3.2. Binary image generation and feture extraction

For each of the clusters obtained by the graph-theoretical clustering, a binary image is constructed. Additional binary images are built through combination between/among different original binary images. Before combination take effect, connected component analysis is performed on the original binary images, which can wipe off some background components from them.

According to different combination strategies and different input images, we can obtain teens to tens of binary images. The problem becomes how to identify the one giving the best binarization among these candidates. To evaluate the goodness of the candidates, the aforementioned two kinds of texture features are extracted and analyzed.

3.3. Optimal binary image selection

In Liu's method [2], a simple decision tree was employed to identify the best threshold. In the proposed method, limitations of threshold based methods, including Otsu's and Liu's, are overcomed through combination among clusters, instead of selecting a single cluster related binary image as the final output.

Multiple steps are implemented to obtain the best binary image. First, initial binary images with small stroke-width frequencies, or the maximum-probability feature, are discarded. Thus only clusters of prominent stroke-like features remain for further processing. But it is often that non-text binary images can possess prominent stroke-like features in terms of run-length description, and then pass the initial evaluation. This is the reason why Liu's method [2] still suffers from complex backgrounds. In our method, SSD features of the remained images are fed into a linear discriminant analysis (LDA) classifier for further verification. Images pass the classier are combined again to generate the final result.

Comparison between results with and without combination is shown in Figure 4. It is obvious that integrating the combination strategy into the algorithm can bring more satisfactory binarization, especially for images of several text blocks with different colors.

(a) Original image (b) No combination (c) With combination

Fig. 4. The effect of the combination strategy

4. EXPERIMENTAL RESULTS

The proposed algorithm has been evaluated with a data set of 520 color images gathered from Internet. All these images contain prominent text blocks, but with varying complexity of image backgrounds. Results were compared with two existing algorithms, Otsu's method and the run-length histogram based thresholding method by Liu et al [2].

An example of the binarization result is shown in Figure 5. As expected, Otsu's algorithm, Figure 5(b), did not perform well in separating text from complicated backgrounds. Method by Liu, Figure 5(c), performed better; but, same as Otsu's algorithm and being single-threshold technique, it still suffered much from complex backgrounds. Our method, as shown in Figure 5(d), however, can survive such situations and yield satisfactory results. Furthermore, in Liu's method, only text of uniform stroke width was considered; while in our algorihtm, text with various stroke width in a same image can be efficiently handled.

(a) Input image (b) Otsu (c) Liu's method (d) Proposed

Fig. 5. Comparison among different algorithms

The statistics for the experimental results are shown in Table 1. They are devided into two categories, correponding to two evaluation strategies. One strategy is holistic image visual judgement, the other accurate word level evaluation. It shows that, the Otsu's algorithm only gains an accuracy of 67.9% in the visual judgement, a better accuracy of 72.3% for Liu's method; superior to both the two techniqus, the proposed algorithm reaches an accuracy of 85.8%. In the word level evaluation, accuracy for Otsu's algorithm is 68.1%, and 69.4% for Liu's method; both are under that of the proposed algorithm, whose is 81%, with a wide gap of more than 10%. The experimental results showed that, compared to existing techniques, the proposed algorithm can efficiently extracte text from images, and is capable of obtaining cleaner binarization results.

5. CONCLUSION

Binarization is an important processing in computer vision, especially for document or text image related applications. Most existing techniques for document image binarization utilize thresholding, either globally or locally. For these techniques no satisfactory binarization results are guaranteed as applied to images with complex backgrounds. In this paper, we have presented a new scheme for color text image binarization which efficiently integrats color clustering and binary texture analysis. Compared with existing ones, the proposed method is robust enough to survive those complicated situations. First results on our data set of above 500 color text images collected from Internet are very encouraging.

Scheme	Visual	Word-level
Total	520 images	3519 words
Otsu	353 (67.9%)	2397 (68.1%)
Liu's [2]	376 (72.3%)	2441 (69.4%)
Proposed	446 (85.8%)	2850 (81%)
The data set consists of 520 color text images.		

Table 1. Statistics of the experimental results

6. REFERENCES

[1] P.K. Sahoo, S. Soltani, and A.K.C. Wong, "A survey of thresholding techniques," *Computer Vision, Graphics and Image Processing*, vol. 41, no. 2, pp. 233–260, February 1988.

[2] Y. Liu and S. N. Srihari, "Document image binarization based on texture features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 540–544, May 1997.

[3] P. D. Wellner, "Adaptive thresholding on the digitaldesk," Tech. Rep. EPC-93-110, Rank Xerox Research Center, Cambridge Laboratory, 1993.

[4] P. Maragos, "Pattern spectrum and multiscale shape representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 701–716, July 1989.

[5] G. Ayala and J. Domingo, "Spatial size distributions: Applications to shape and texture analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1430–1442, December 2001.

[6] M.J. Jones and J.M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, January 2002.

[7] J. Matas and J. Kittler, "Spatial and feature space clustering: Applications in image analysis," in *Proceedings of the 6th International Conference on Computer Analysis of Images and Patterns*, Prague, Czech Republic, September 1995, pp. 162–173.