분산맵을 이용한 웹 이미지 텍스트 영역 추출

Text Region Segmentation from Web Images using Variance Maps

정인숙, 오일석

전북대학교 전자정보공학부 컴퓨터공학・영상정보신기술연구소

InSook Jung(msisjung@chonbuk.ac.kr), Il-Seok Oh(isoh@chonbuk.ac.kr)

요약

분산맵은 텍스트 영역이 주변과의 색상 혹은 밝기 변화가 심하다는 특징을 이용하는 방법으로 특히 잦은 포맷 변환에 의하여 해상도가 낮거나 일정하지 않은 웹 이미지의 텍스트 영역을 추출하는 데 적용할 수 있다. 그러나 이전의 분산맵을 적용한 방법들은 입력 영상 전역에 고정된 마스크를 한 번만 적용하는 광역 분산맵을 사용하므로 텍스트 크기가 매우 작거나 큰 경우, 획의 색상에 gradation효과가 있는 경우, 각도, 위치, 색상 등이 복잡한 경우 텍스트 추출 성능이 안정적이지 못하다. 본 논문은 2단계 분산맵을 사용하여 Web 이미지에서 텍스트 영역을 안정적으로 추출하는 방법을 제안한다. 제안된 방법은 광역 및 지역 분산맵이 각 단계에서 적용되며 서로 계층적 관계를 가진다. 1단계는 텍스트 영역 추출 재현율을 높일 수 있도록, 충분히 큰 글자 혹은 작은 글자도 추출할 수 있는 일정한 마스크 크기를 가진 광역의 수직 및 수평 색 분산맵을 적용하여 유사 텍스트 영역을 추출한다. 2단계에서는 1단계의 각 연결요소영역에 새로운 마스크 크기를 가진 명암 분산맵을 적용하여 최종적인 텍스트 영역을 추출한다. 2단계 분산맵 적용에 의하여 1단계에서 구한 유사 텍스트 영역에 남아 있는 배경 부분이 많이 사라지게 되어 추출 정확률이 높아진다. 제안한 방법을 400개의 Web 이미지에 적용한 결과 배경이 복잡해도 비교적 안정적으로 텍스트 영역을 추출하는 것을 확인할 수 있었다.

■ 중심어 : | 텍스트 영역 | 텍스트 추출 | 웹 이미지 | 2단계 분산맵 |

Abstract

A variance map can be used to detect and distinguish texts from background in images. However, previous variance maps work at one level and they suffer a limitation in dealing with varieties in text size, slant, orientation, translation, and color. We present a method for robustly segmenting text regions in complex color Web images using two-level variance maps. The two-level variance maps work hierarchically. The first level finds the approximate locations of text regions using global horizontal and vertical color variances with the specific mask sizes. The second level then segments each text region using intensity variance with a local mask size, which is determined adaptively. By the second process, backgrounds tend to disappear in each region and segmentation can be accurate. Highly promising experimental results have established the effectiveness of our approach.

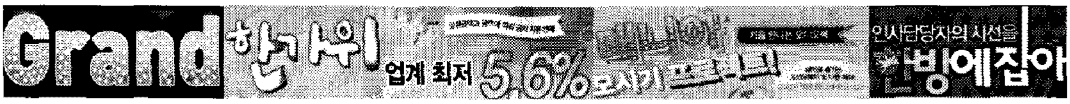
■ keyword : | Text Location | Text Segmentation | Web Images | Two-level Variance Maps |

* This research was financially supported by the Ministry of Education, Science Technology (MEST) and Korea Institute for Advancement of Technology(KIAT) through the Human Resource Training Project for Regional Innovation.

접수번호 : #090722-004　　심사완료일 : 2009년 08월 11일

접수일자 : 2009년 07월 22일　　교신저자 : 오일석, e-mail : isoh@chonbuk.ac.kr

I. 서 론



(a) Images with multi color text over multi color background

(b) Images containing 3D-effects, shadows and outlines text

(c) Images containing text not always on a straight line

Figure 1. Text in web images

While existing search engines index a page on the text that is readily extracted from its HTML code, an increasing amount of the information on the Web is embedded in images. In extreme cases,

all of the text on a page might be present solely in image format. Existing Web commercial search engines are limited to indexing the raw ASCII text they find in the HTML–they cannot recover image text. This situation presents new and exciting challenges for the fields of Web document analysis and text information retrieval, because Web image texts are typically rendered in color and at very low spatial resolutions.

Considering traditional optical character recognition (OCR), one may initially think that Web images containing texts present some advantages over scanned documents, such as the lack of digitization-induced noise and skew. However, the task is considerably difficult for traditional OCR for a number of reasons. First, these (often complex) Web images tend to be of low resolution (just good enough for display and usually 72 dpi), and the font size used for text is very small (about 5~7pt) or very large (about 50~70pt). Such conditions clearly pose a challenge to traditional OCR, which works with 300dpi images and typical character sizes of 10pt. Moreover, Web images tend to have various artifacts (anti-aliasing, colour quantization, and lossy compression), colour schemes (multicolour text over multicolour background [Figure 1](a) and character effects (3D effects, shadows, outlines [Figure 1](b), characters not always on a straight on a line [Figure 1](c), etc.).

The goal of locating the texts in image form can be split into two objectives: Text segmentation. and Text extraction.

In view of the difficulties posed by image and text characteristics, the segmentation stage is by far more challenging. The performance at that stage affects quite crucially the degree of success in subsequent recognition. This paper presents a new approach to the segmentation, especially in complex Web images (e. g. those in [Figure 1]).

II. 관련연구

There are two primary methods for segmenting text in images: colour representation-based (or region-based) methods ([1-4]) and texture-based methods ([5][6]).

It is argued that the colour representation-based methods (commonly used by previous approaches) are ill-suited to this particular task for Web images. Their methods for text segmentation and extraction are based on clustering in the colour space. However, they are inappropriate for low resolution images and those and with various character effects. They depend on the effectiveness of the segmentation method, which should guarantee that a character is segmented into a few connected components (CC) separated from other objects and the background. These methods produce good results for relatively simple images, but fail when more complex images are encountered for the following reasons. These approaches mostly deal with a very small number of colors (they do not work on full-color, e. g. JPEG, images). They also assume a practically constant and uniform color for text ([7][8]) and fail when this is not the case. In practice, there are many situations where a gradient or multicolour texts are present (see [Figure 1]).

The situation in which dithered colors are present (especially in GIF images) has received some attention ([9][10]), but such colours are only found in a relatively small number of Web images. Furthermore, the background may also be complex (in terms of colour), so the assumption that it is the largest area of (almost) uniform colour in the image ([11]) does not necessarily hold.

Unlike the above colour representation, the texture-based methods employ distinct properties of texts not shared with their backgrounds. In these methods, the textual properties of an image are often detected using the techniques of Gabor filters, wavelets, spatial variance, etc.

If a document or an image containing texts is viewed at a certain distance far from a person, he sees a blurred image of the document but is still able to detect the different blocks of the document. Detection is possible since each block has a specific texture pattern. These patterns correspond to regions of text, regions of graphics, and regions of pictures. Thus document images can be segmented into regions of text, and regions of graphics and/or pictures using the texture of low-resolution images.

The assumption of the variance map is that each part of a document image has a different texture. Text regions have different textures from that of graphics and pictures. Graphic regions have different textures from that of texts and pictures, and the same may be applied to regions with pictures. Under this consideration document segmentation may be considered as a texture segmentation problem.

One of the characteristics of this method compared with others is that it maps the variance around the pixel to just one value. This value represents the variance of the texture around the pixel of interest. Also, in the computation of the variances, a mean value is used instead of the actual value of the pixel at the center of the mask. The texture at a pixel is defined as the average value of the variance of the neighbors of the pixel in 4 or 8 different directions, vertical, horizontal, left diagonal, and right diagonal. Thus the document image is mapped in to the texture image. Despite its simplicity, variance maps have been shown to be effective and robust. In [12], Murguia proved that a document image can be segmented into regions of texts, and regions of graphics and/or pictures using gray-level spatial variance of low resolution images. It was designed to work with free format documents, text in background other than white, and skew greater than 10 degrees. Moreover, it requires less computation than the segmentation methods using the other textures described in other papers.

Karatzas and Antonacopoulos abandons analysis by the RGB colour clustering and adopts a segmentation method based on analysing differences in colour and lightness that is closer to how humans perceive distinct objects [13]. However, the current methods using differences in colours fail for solving the problem of segmenting characters in colour web images containing texts (in headers, titles, banners etc.) when the image size is not small or the character size is too large. Song et al. proposed an extraction method to detect text regions from the images using the intensity variance and color variance, but the method has difficulties in locating large text or text with severe illumination changes [14].

In such a global approach, the mask of variance maps was applied to a single value for an entire image. A global mask has a good performance when the text has a uniform size and the character width is less than that of the mask. Very often, Web images do not satisfy these criteria.

Although very effective in text localization, the variance maps have some shortcomings:

(1) difficulties in designing a locality of the mask information (type, size) to satisfy the wide variations of text size.

(2) inability to ensure accurate location of texts. For example, the resulting variance map suffers from a great amount of backgrounds in cases in which there is non-uniform size of text and/or the text is not in straight lines.

III. 제안하는 방법

In this paper, variance maps are used to detect and distinguish texts from the background in Web images. We propose a method to deal with local information for masks in variance maps using two levels. Local information may guide the adaptive mask size for the local text region only.

Our method is to improve the variance map to overcome shortcomings in previous variance maps and then to manage accurately complicated cases such as multicolor text and/or texts of various sizes or not arranged in straight lines.

Our approach is to segment the image into text/non-text regions as effectively as possible and then let the OCR system do the detailed refinement. In other words, we would like to find all text areas and as few spurious non-text areas as possible without actually classifying the characters.

Section 4 provides a detailed description of our method and section 5 shows the experimental results. The conclusion is given in Section 6.

IV. 2단계 분산맵 분석 알고리즘

The two-level variance maps work hierarchically. The first level variance finds the approximate locations of text regions using horizontal and vertical color variances with specific mask sizes to ensure extraction of both large and small text. At the first level, we can increase the recall rate using the color variance map with the specific mask size to approximately segment-text like regions and apply CC analysis (CCA). It segments text components in these regions using local thresholds; a new mask size is determined adaptively in each. As a second level, the automatic and non-heuristic gray variance map using the new mask size is applied to each region. By the second process, backgrounds tend to disappear in each region and segmentation can be accurate. In the second level, we can also increase the precision rate using an intensity variance map with an adaptive mask size to find accurate text-like regions. This technique has been widely used in image analysis because it exhibits better performance in segmenting the objects from an image that contains spatially uneven texture features.

[Figure 2] shows an overview of our method.

First, the global variance of the input image is computed to segment text regions in the input image. This is because text usually occurs with sufficient differences with the background. Applying CCA, every CC of the variance image is binarized using a local threshold. Each CC is then processed separately. This image is passed through a Laplacian filter, which is useful for extracting variances which occur in texts of CC region. Next, for every CC, we determine the new mask size based on the Laplacian map and then apply a local variance map with the new mask size to segment the text region accurately. The method is finished when the local variance is computed and applied to the entire image area. (After this step, the resulting image will have two level variance maps.)

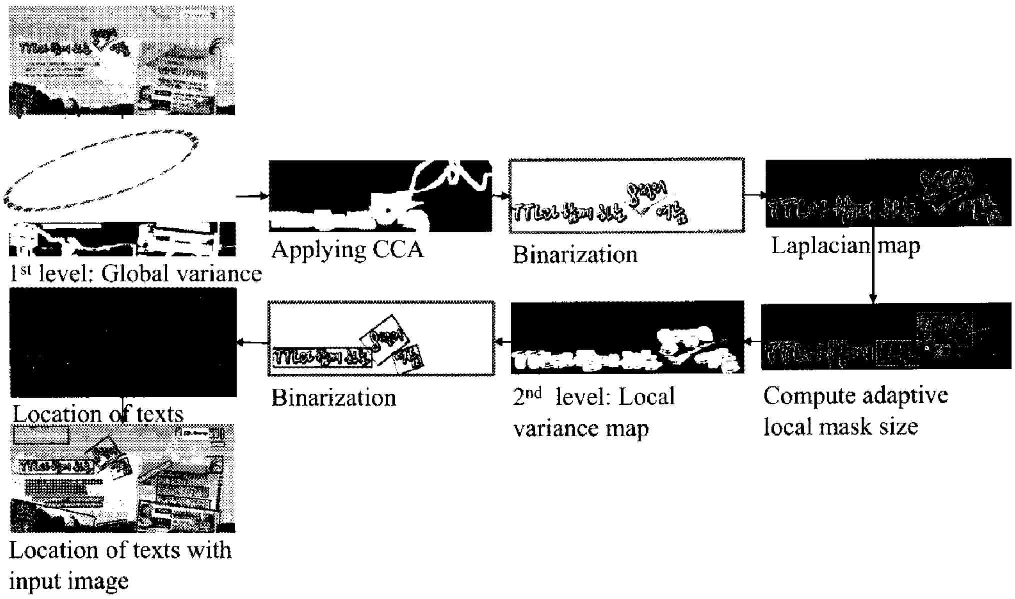


Figure 2. Overview of our method

1. Parameterization of variance map

We obtain the variance map by applying masks to each pixel (such as [15]). We then transform the variance map into a binary image. The pixels that were assigned to 0 are regarded as background and excluded in subsequent processing. In this processing, we have various maps that depend on mask shape, mask size, and the combination of masks if multi masks are used. We define them parametrically as follows.

Θ=(m, s, op)

m=subset of {rectangle, vertical form, horizontal form, cross, diagonals from the upper-right and upper-left corner, diagonal cross,.....}

s= mask size, ($3 \leq s$)

op ={and, or,..}

The possible shapes, m, of the mask are rectangle, horizontal, vertical, right-upper diagonal, left-upper diagonal, diagonal cross, etc. They can have different sizes according to shapes. Sizes must be greater than at least 3 to obtain its variances. Size s is defined as the number of pixels in the mask.

$|m|$ is the number of masks used. It is necessary to combine $|m|$ maps into one when $|m|$ is larger than 1. We need combination operations are required. Usually bitwise-and or bitwise-or is used. In this paper, we use the following parameters:

m= {rectangle, vertical, horizontal}

s={3x21 rectangle, 17x3 rectangle, 1x21 horizontal, 17x1 vertical}
op ={ and }

In the RGB colour space, the variance at each pixel (x,y) in applying the horizontal mask is defined as follows. The variances by the other masks are defined in the similar way. Horizontal mask operation:

$\sigma_C^h(x,y)$=variance of the (x,y) pixel in applying the horizontal mask to the pixel in the C plane (with 21 neighbour pixels), where C is one of R, G, and B.

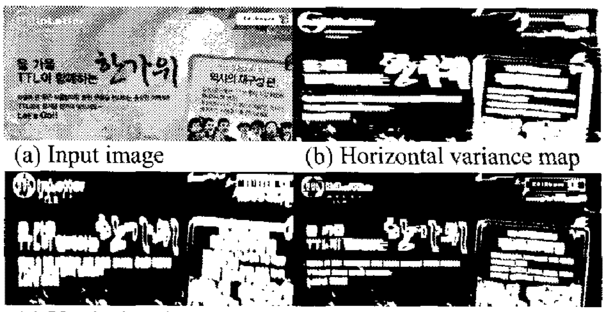
$$\sigma^h(x,y)=1/3*(\sigma_R^h(x,y)+\sigma_G^h(x,y)+\sigma_B^h(x,y)) \quad (1)$$

$\sigma^h(x,y)$ and vertical variance map $\sigma^v(x,y)$ obtained in this way.

Next we transform the two variance maps to a variance map $T(x,y)$ using a threshold value t and a combination operation 'and'.

$$T(x,y)=\begin{cases}1, & \sigma^h(x,y)>t \text{ and } \sigma^v(x,y)>t \\ 0, & \text{otherwise}\end{cases}$$

[Figure 3] shows an input image and its variance maps. [Figure 3](b) and [Figure 3](c) are the horizontal variance map σh(x,y) and vertical variance map $\sigma^v(x,y)$ correspondently. [Figure 3](d) displays the variance map obtained from the two variance maps.



(a) Input image (b) Horizontal variance map
(c) Vertical variance map (d) Variance map

Figure 3. Input image and its variance maps

2. First level: Global analysis

In our method, we first compute the global horizontal and vertical variances applying equation (1) with 3x21 horizontal and 19x3 vertical masks. The mask sizes were determined empirically in previous studies. At this step, we obtain a rough estimation of text regions. Our intention is to proceed to an initial segmentation of text (foreground) and non-text (background) regions that will provide a superset of the correct set of foreground pixels. This is refined at a later step using the so-called local variance map. [Figure 4] shows the image of the variance map.

(a) Input image (b) Variance map

Figure 4. Global variance map

We see that text areas are high in variance (white pixels in [Figure 4](b)). To remove further the noisy regions, we also apply morphological operations. The operations have structuring elements with size 2x5 for dilation and 3x3 for erosion to emphasize more horizontal texts. Morphological closing and opening are applied.

[Figure 5](a) shows the results of the morphological operations, and [Figure 5](b) shows the connected components (CC) (which are randomly colored).

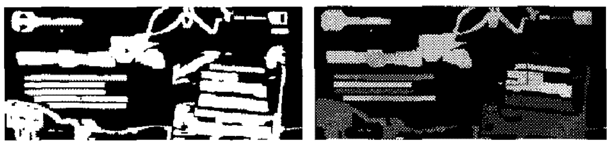


Figure 5. (a) Morphological operation applied (left) and (b) CCs (right)

As shown in [Figure 5](b), every CC usually contains image text, but only, however, in a sub-optimal fashion: Some CCs span more than one line and/or column of text, others contain no text, while in many the background comprises a large portion of the pixels. Fortunately, these shortcomings can be overcome by the next step, applying the local variance map.

3. Second level: Local analysis

3.1 Binarization

We generate a bounding box around each CC region on the variance map of the first level. Once the bounding box of each CC is obtained, each region in the bounding box is binarized using Otsu thresholding. This step produces binary text regions to be used as inputs to the local variance map (It is possible to run local color images directly, but we obtain substantially worse performance if we do so).

Every CC region provided by the variance map is expected to contain only text, implying that the background and foreground should be easily separable through thresholding.

To ensure the correct labelling of both dark-on-light and light-on-dark text, the proportion of pixels which fall above and below the thresholds is considered. Since in a block of text there is always a larger area of background than text elements, the group of pixels with the lower proportion is labelled as text, and the other group as background.

3.2 Laplacian map and CC

In our approach, the local binary image is passed through a Laplacian filter which is useful in applying the local variance map. The Laplacian edge detector produces closed edge contours because edge strength is not considered, so even the slightest, most gradual intensity transition produces differences in pixel values. Next, CCA is generated based on the Laplacian image.

CCA detects character candidates and enables us to estimate their size and the spacing between them. These estimates will be used to apply the local variance map with adaptive mask for text directly to the regions.

We show results for the Laplacian map and CCs in [figure 6] for the top-left text region shown in [Figure 5].

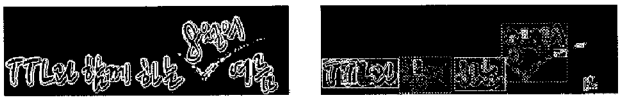


Figure 6. Laplacian map and CCs

3.3 Estimation of adaptive mask

In applying the local variance map, the key problem is the determination of the mask type and size. Determination of a adaptive mask size is very important and perhaps the most sensitive part of any image segment scheme of variance maps because a wrong size of mask may result in discarding of some text information (an object can be considered part of background and vice versa).

The well-known local adaptive method uses mean and standard deviation to compute mask size over a CC text region (e. g. bounding box of every CC in [Figure 6]). We use the mean and the standard deviation along with widths of all bounding boxes in every CC to compute another mask size for the CC region. In other words, we compute the adaptive mask size to apply the local variance map, which uses adaptive contribution of mean and standard deviation in determining local mask size.

3.4 Local variance map

The local variance map is an adaptive one in which a mask size is determined over a CC region. The local method performs better in case of nonlinearly aligned texts and over a wide variation of text sizes. The segmentation quality depends on the above adaptive mask size.

We apply intensity variance map to the CC region presented by the Laplacian filter with adaptive mask size. Our method computes the local horizontal variance using only equation (1) with the mean and the standard deviation as a mask size for every CC region.

We show results for the local variance map and the red bounding boxes of CCs in [Figure 7].

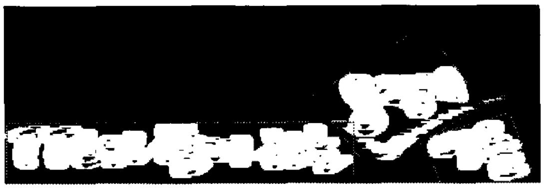


Figure 7. Local variance map and CCs

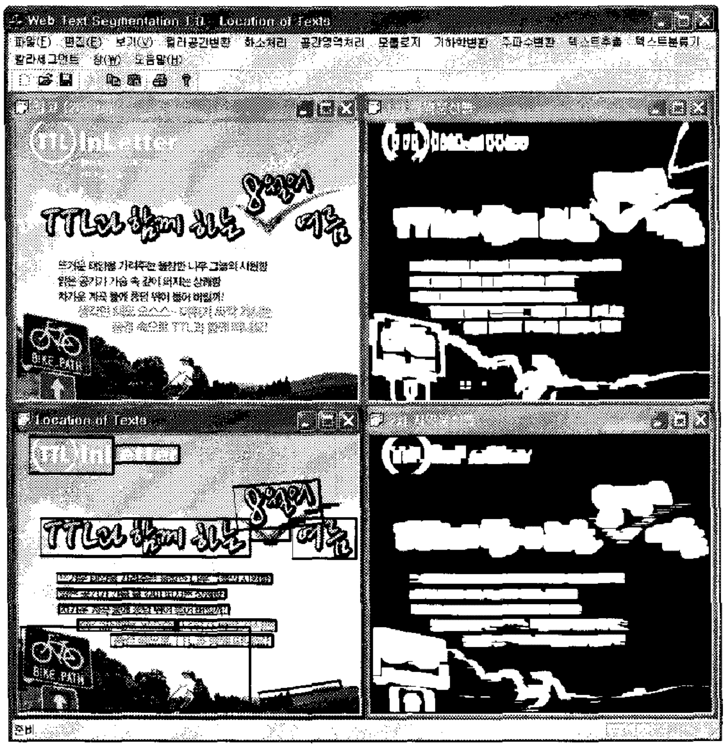


Figure 8. User Interface of the proposed text extraction method.

V. 실험 결과

We conducted experiments with a PC having a CPU (Intel Core 2 Duo E7200 2.53GHz) with Window XP using VC++6.0. The proposed method was tested using 400 web images selected randomly from the Web whose sizes are varied greatly. Some of them contain extremely large or small text, non homogeneously colored text, and/or non-horizontal layouts. It is notable that all of our experiments were performed without apriori knowledge about the input image. We did not have any information about the font size or format of the text in the image. All images had low resolution, and the sizes of characters varied from 4pt to 72pt.

The user interface of the proposed text segmentation method is illustrated in [Figure 8]. The windows in this figure show the results of the text segmentation process. The bottom-right window in this figure shows the results of the text segmentation process. In the results, segmented texts are represented as black rectangular bounding boxes.

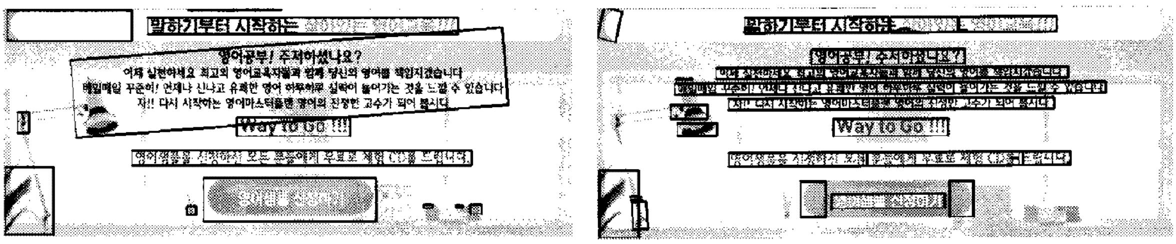
In order to compare our method with other global variance-based methods, we evaluated the performance of our method and the method of [14] with the same test images. The method of [14] uses the only global variance map with the different mask sizes (3x21 horizontal and 19x3 vertical masks). [Figure 9] illustrates the results of segmenting texts in Web images using the method of [14] (left) and our method (right), and the improvement can be seen at right sides of the resulting images. In all of these experiments, we use the same parameters. Based on visual criteria, we have seen that our method outperforms that of [14].

[Figure 9] shows the results that our method has removed the background as much as possible without disturbing the text area. This is an improvement over the method of [14]. By looking at results of our method, we have observed that, by the local variance, we can get the unnecessary pixels eliminated from the image background in a better way while preserving characters and vice versa. In this way, we have found adaptive local mask sizes to be appropriate for this kind of Web image.

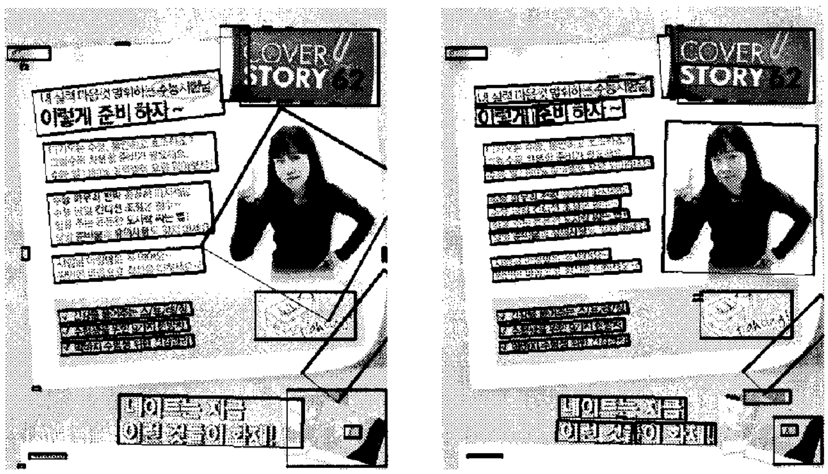
Additionally, our method serves the text line segmentation effect. [Figure 9] (a), (b), (c), and (d) present the line segmentation results of our method. Due to the size of variation of characters, connected words, curvilinear text lines, non uniform gaps between neighboring text lines in Web images, and text line segmentation for Web images still presents a significant challenge ([16-18]). Therefore, our method can be the intermediary between text extraction and line segmentation.

In the experiments, we found that fragmentation often appears at those text lines that isolated both IN horizontal and vertical orientations. Because headlines vary in size greatly, some headline components are erroneously segmented into the body of text components. The fragmentation rate of headline regions is higher than that of body text regions.

Our method may fail to extract texts when texts in the input image have severe colour gradation and too small differences of intensity. False alarms are currently ignored.



(a)



(b)

(c) (d) (e) (f)

Figure 9. Images of different segmentation results, contained in the rectangles: (Left) for Song et al.([14]) method; (right) for our proposed method.

VI. 결 론

In this paper, we propose a local adaptive approach using variance maps to segment texts in Web images. The proposed method is less sensitive to user parameters and can deal with segmentations where shadows, non-uniform character sizes, low resolution, and skewing occur. After the local approach, our method demonstrates superior performance on Web images using visual criteria. Our method has the additional advantage that it can be applied directly to the line segment. Further research will focus on developing the text or non-text classifier and character segmentations.

참 고 문 헌

[1] H. K. Kim, "Efficient Automatic Text Location Method and Content-Based Indexing and

Structuring of Video Database," Journal of Visual Communication and Image Representation, Vol.7, No.4, pp.336-344, 1996.

[2] R. Lienhart and F. Stuber, "Automatic Text Recognition In Digital Videos," Proceedings of SPIE, vol.2666, pp.180-188, 1996

[3] Y. Zhong, K. Karu, and A. K. Jain, "Locating Text Complex Color Images," Pattern Recognition, Vol.28, No.10, pp.1523-1535, 1995.

[4] R. Lienhart and W. Effelsberg, "Automatic Text Segmentation and Text Recognition for Video Indexing," Technical Report TR-98-009, Praktische Informatik IV, University of Mannheim, 2000.

[5] K. C. Jung and J. H. Han, "Hybrid Approach to Efficient Text Extraction in Complex Color Images," Pattern Recognition Letters, Vol.25, pp.679-699, 2004.

[6] K. C. Jung, K. I. Kim, and A. K. Jain, "Text Information Extraction in Images and Video: A Survey," Pattern Recognition, Vol.37, No.5, pp.977-997, 2004.

[7] J. Zhou and A. D. Lopresti, "Extracting Text from WWW Images," Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR'97), Vol.1, pp.248-252, 1997.

[8] A. Antonacopoulos and F. Delporte, "Automated Interpretation of Visual Representations: Extracting Textual Information from WWW Images," *Visual Representations and Interpretations,* R.Paton and I.Neilson(eds.), Springer, London, 1999.

[9] J. Zhou, A. D. Lopresti, and T. Tasdizen, "Finding Text in Color Images," Proceedings of the IS&T/SPIE Symposium on Electronic Imaging, SanJose, California, Vol.3305, pp.130-140, 1998.

[10] A. D. Lopresti and J. Zhou, "Locating and Recognizing Text in WWW Images," Information Retrieval, Vol.2, pp.177-206, 2000.

[11] A. K. Jain and Y. Bin, "Automatic Text Location in Images and Video Frames," Pattern Recognition, Vol.2, pp.1497-1499, 1998.

[12] M. I. C. Murguia, "Document Segmentation Using Texture Variance and Low Resolution Images," Image Analysis and Interpretation, 1998 IEEE Southwest Symposiumon, pp. 164-167, 1998.

[13] D. Karatzas and A. Antonacopoulos, Colour "Text Segmentation in Web Images Based on Human Perception," Image and Vision Computing, Elsevier, Vol.25, pp.564-577, 2007.

[14] Y. J. Song, K. C. Kim, Y. W. Choi, H. R. Byun, S. H. Kim, S. Y. Chi, D. K. Jang, and Y. K. Chung, "Text Region Extraction and Text Segmentation on Camera Captured Document Style Images," Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR'05), Vol.1, pp.172-176, 2005.

[15] I. S. Jung, D. S. Ham, and I. S. Oh, "Empirical Evaluation of Color Variance Method for Text Retrieval from Web Images," Proceedings of the 19th Workshop on Image Processing and Image Understanding (IPIU'08), pp.36-41, 2008.

[16] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.30, No.8, pp.1313-1329, 2008.

[17] M. Makridis, N. Nikolaou, and B. Gatos, "An Efficient Word Segmentation Technique for Historical and Degraded Machine-Printed Documents," Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR'07), Vol.1, pp.178-182, 2007.

[18] I. Nwogu and G. H. Kim, "Word Separation of Unconstrained Handwritten Text Lines in PCR Forms," Proceedings of the Eight International Conference on Document Analysis and

Recognition (ICDAR'05), Vol.2, pp.715-719, 2005.

저 자 소 개

정 인 숙(InSook Jung) 정회원

- 1989년 2월 : 이화여자대학교 수학과(이학사)
- 1991년 2월 : 포항공과대학교 수학과(이학석사)
- 2005년 3월 ~ 현재 : 전북대학교 전자정보공학부 컴퓨터공학전공 박사과정

<관심분야> : 패턴인식, 영상처리, 컴퓨터비젼

오 일 석(Il-Seok Oh) 정회원

- 1984년 2월 : 서울대학교 컴퓨터공학과(공학사)
- 1992년 2월 : KAIST 전산학과(공학박사)
- 1992년 ~ 현재 : 전북대학교 전자정보공학부 컴퓨터공학전공 교수

<관심분야> : 문서영상처리, 패턴인식, 컴퓨터비젼