공간주파수를 이용한 장면영상에서 텍스트 검출
(Text Detection in Scene Images using spatial frequency)

신 봉 기 † 김 선 규 ††
(Bong-Kee Sin) (Seon-Kyu Kim)

요 약 장면 영상 속의 문자 영역에는 다른 부분과는 구분되는 특징적인 공간주파수가 있다. 이 특징은 직관적이며 또한 유용한 정보로서의 가치가 있다. 본 논문에서는 장면 영상에서 수평 텍스트를 찾는 방법을 제안한다. 수직 및 수평 방향으로 걸친 edge 픽셀의 빈도수와 푸리에 변환에 의한 기본 주파수의 두 가지 특징을 이용한 방법이다. 두 가지 특징을 독립적으로 활용하여 그 결과를 결합하거나 연속하여 적용하여 원하는 결과를 얻을 수 있다. 이와 같은 특징은 대체로 언어 또는 문자에 무관함을 확인하였다. 이에 추가하여 Hough 변환을 이용한 장면 속의 사각형을 탐색하였다. 여러 사람들에게 유용한 정보는 보통 강한 색상대비로 눈에 잘 띄는 색깔의 사각형 안에 쓰여있는 경우가 보통이므로 사각형의 탐색함으로써 보다 효과적으로 문자를 탐색할 수 있다.

키워드 : 패턴인식, 텍스트 검출, 공간 주파수, 언어 식별

Abstract It is often assumed that text regions in images are characterized by some distinctive or characteristic spatial frequencies. This feature is highly intuitive, and thus appealing as much. We propose a method of detecting horizontal texts in natural scene images. It is based on the use of two features that can be employed separately or in succession: the frequency of edge pixels across vertical and horizontal scan lines, and the fundamental frequency in the Fourier domain. We confirmed that the frequency features are language independent. Also addressed is the detection of quadrilaterals or approximate rectangles using Hough transform. Since texts that is meaningful to many viewers usually appear within rectangles with colors in high contrast to the background. Hence it is natural to assume the detection rectangles may be helpful for locating desired texts correctly in natural outdoor scene images.

Key words : Pattern recognition, Text location, Spatial frequency, Language identification

1. Introduction

One of the most significant sources of information in natural outdoor scene images is text. Text in images appears in diverse ways in various conditions. In fact it is often the most important source to an intelligent automatic navigation system, but it is not so easy to locate all and only relevant texts in the outdoor environment. Fortunately, however, it is not impossibly hard because most important texts are painted in colours of high contrast to and readily stand out from the background. Furthermore they are almost always contained in rectangular boxes, as in sign boards, in order to grab attention of as many people as possible.

Most of the previous researches have been concerned with extracting texts from document images[1] or captions from video streams utilizing interframe redundancy and the continuity of motion [2, 3]. In other researches the targets were Web images[4], book cover or CD title cover images[5]. Their methods are based on color or texture. Color or gray scale intensity based methods in general

• 본 연구는 한국과학재단 목적기초연구(과제번호 2001-1-30300-007-1) 지원으로 수행되었음.
† 종신회원 : 부경대학교 전자컴퓨터정보통신공학부 교수
bkshin@mail.pknu.ac.kr
†† 비 회 원 : (주)한국머틴테크
sk7944@mutaltech.com
논문접수 : 2002년 7월 27일
심사완료 : 2002년 10월 31일

defines connected components of pixels and labels individual regions according to a given criterion[2]. Texture-based method, on the other hand, takes a root in image processing field. Typically used are the tools of Gaussian filter[6], Gabor filter[1], and spatial variance[5]. There is a recent surge of research and reports around the world[7,8,9,10]; most of them are, with little exception, based on the pixel clustering and connected components. Our research is distinguished from them in the features used for evaluating the degree of text-likeness.

This paper describes a method of finding horizonal text regions in scene images using frequency-based features estimated in both space domain and frequency domain. In the space domain we count the number of edge pixels or *edgels* in edge images as a good indicator of the presence of texts along the scan line, either horizontal or vertical(Section II). Although the extent of its utility is not certain, the proposed features are language-independent and we expect that it will work well in text images of other languages. The second feature of our research is the use of Fourier spectrum(Section III) to estimate the fundamental frequency of the text images, which is deemed a better measure subject to a formal analysis. Then we devoted Section IV to describing a practical method of detecting rectangles that can help locating milestone texts in natural scene images. Finally, Section V presents experimental results. In this Section we will also consider the method of identifying the language of the located texts. In Chapter VI we will conclude the paper.

2. Frequency of Scan Line Edgels

This section discusses the regularity of character strokes in a given dimension, which has led us to assume that the edge pixel occurrences along a scan line in edge enhanced images show some characteristic frequencies.

2.1 Horizontal edge crossings

A character pattern consists of strokes usually in high contrast, in color or intensity, to the background. Therefore the edges of the strokes will have higher intensity gradients than some reasonable threshold. We will count the number of high-gradient pixels across the horizontal and vertical scan lines.

First we count the number of significant edges in an edge-enhanced image. For edge enhancement any of the well-known edge detection techniques can be employed. We have taken the simplest one, Roberts operator. In this step every scan line is evaluated by

$$h(y)=\sum_{x=1}^{N}\delta(\Delta(x,y),\theta) \tag{1}$$

where

$$\Delta(x,y)=I(x+1,y)-I(x,y)$$

and

$$\delta(\Delta,\theta)=\begin{cases}1 & \Delta\geq\theta\\ 0 & \Delta<\theta\end{cases}$$

Then a profile of significant edgels along a horizontal scan line can be obtained as follows:

$$H=\{y\mid h(y)>\Theta_H\} \tag{2}$$

for an experimental threshold Θ_H based on image dimension and statistics. The profile usually consists of a set of consecutive integers which correspond to the y-coordinates of the horizontal scan lines containing a number of edgels. They constitute a set of horizontal belts as in the sample result shown in Figure 1.

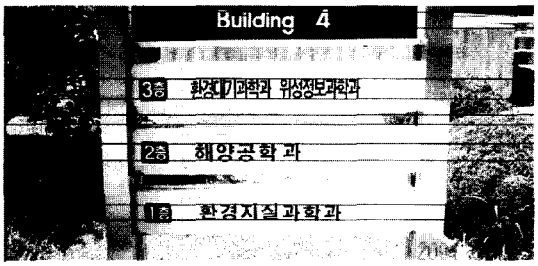


Fig 1 Horizontal text belts

2.2 Vertical text detection

The horizontal belt sections of the previous step are believed to contain texts somewhere therein. Vertical detection step aims to locate the exact regions. The underlying method is similar to that of the preceding step. But since most text usually

runs horizontally, the vertical heights of the belts are not big enough for reliable identification of text.

In order to support the decision, we estimate blocks of eight pixels wide at a time instead of single vertical scan lines. This has the effect of scanning eight times the height of texts. The successive blocks are set to lie four pixels apart and overlap by a half. The detection procedure starts from the calculation of vertical and horizontal gradients.

$$\Delta_H(x,y) = I(x+1,y) - I(x,y)$$

$$\Delta_V(x,y) = I(x,y+1) - I(x,y)$$

Then,

$$h(y \mid B(l,k)) = \sum_{x \in B(l,k)} \delta(\Delta_H(x,y), \Theta_H)$$

$$v(x \mid B(l,k)) = \sum_{y \in B(l,k)} \delta(\Delta_V(x,y), \Theta_V) \quad (3)$$

Here $B(l, k)$ denotes the k-th block in the l th horizontal belt. $h()$ and $v()$ are the numbers of significant edgels, horizontal and vertical direction, inside the block $B(l,k)$. The identification of text block is based on the sum of the two values. The result is the set of text blocks identified as such:

$$T = \{ B(l,k) \mid h + v > \Theta \} \quad (4)$$

In the preceding equations, Θ's corresponds to the thresholds applied for identifying text blocks.

3. Frequency Spectrum Periodicity

Counting edgels, although simple, does not provide a sufficient characterization of frequency features of text patterns. This section will explore this point for a more rigorous characterization

A character pattern consists of strokes with characteristic complexity (depending on language) including font, width('i', and 'm') and height('b', 'e' and 'p'), and stroke thickness. These are the features that cannot be observed from other non text objects like trees and building windows.

3.1 Fourier Spectrum

Let $B(l, k)$ be an S(T block of pixels. By linking all the horizontal scan lines in order, we create a single line of $S(T=N$ pixels. This being a one dimensional sequence, we can apply 1-D Fast Fourier Transform(FFT) to obtain a frequency spectrum of the block.

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(-\frac{j2\pi kn}{N}\right) \quad k = 0, \cdots, N-1 \quad (5)$$

where x_n is a pixel and X_k is the resulting Fourier descriptors. The spectrum of a typical text block is given in Figure 2(a). Compare it with the non-text spectrum of Figure 2(b). Unlike natural scene objects, most text image has similar characteristics. According to our study of 2D characterization, 2D FFT showed a strong presence of vertical and horizontal frequencies.

3.2 Harmonic frequency of text image

Auto correlation is a useful measure of periodicity or similarity between patterns which are separated by some distance k temporally or spatially.

$$R(k) = \sum_{n=-\infty}^{\infty} x_n x_{n+k} \quad (6)$$

This will have maximum when k = 0, and there are additional smaller peaks as k approaches the integer multiple of fundamental frequency, if they exist.

The autocorrelation of the Fourier spectrum of Figure 2 is shown in Figure 3. The peaks arealigned along the slopes at an equal distance. Unlike those of text blocks, non-text blocks do not possess such a sequence of peaks. We can decide

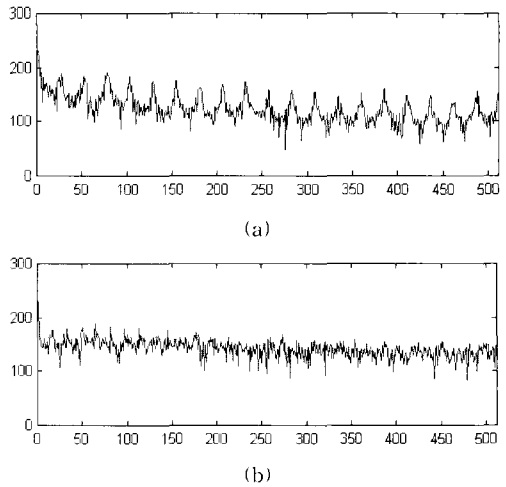


Fig 2 Fourier spectrums of (a) a text and (b) a non-text blocks.

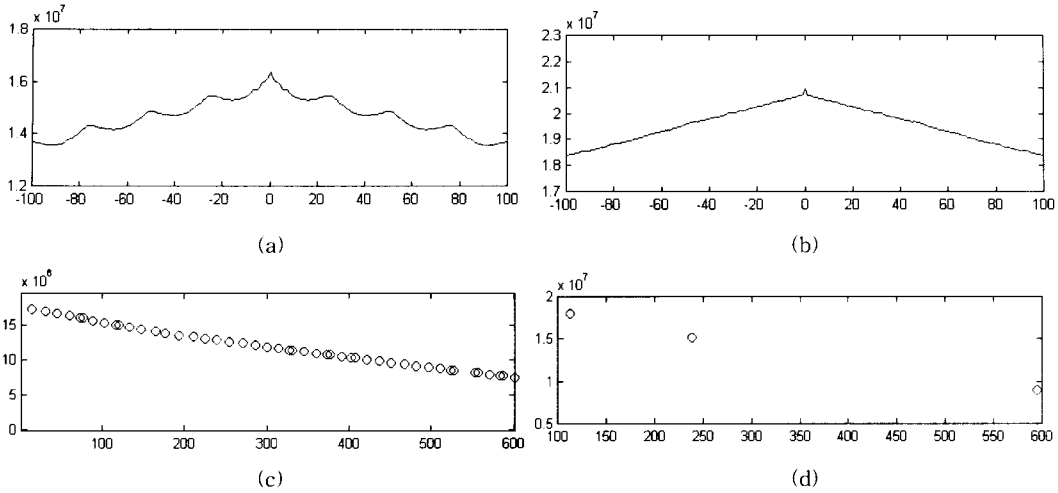


Fig 3 Autocorrelation test of the Fourier spectrum in Figure 2. (a) Text spectrum AC, (b) non-text spectrum AC, (d) text AC peaks, and (d) non-text AC peaks.

that a block is textual or not according to the number of peaks. Compare the number of peaks extracted from Figure 3(a) and (b) in which peaks are marked by small circles. The result of this autocorrelation test is used to decide or confirm that the block under consideration has text or not.

3.3 Fundamental frequency of text image

The preceding section discusses the new method of confirming the periodicity of the text images in the frequency spectrum. As will be shown by experiments in the next chapter, the feature is quite useful as a measure for locating and/or confirming text in scene images. But it is an indirect measure which does not readily lend itself to our understanding.

In this section a generalized signal superposition technique of homomorphic filtering is applied to capture the fundamental frequency of text images. Specifically, we apply inverse discrete Fourier transform(IDFT) to the logarithm of the magnitude of the Fourier transform

$$| IDFT(\log | DFT(x_n) |) | \quad (7)$$

The result is called the cepstrum of the input $\{x_n\}$. The cepstrum shows a strong peak at a space/distance equal to the period of the input sequence. This corresponds to the fundamental frequency. Widely used in the analysis of speech signals, the cepstrum serves as a very good tool for the discrimination of vowels from consonants. In the text image domain, it is conjectured that the cepstrum can be used to characterize the differences embedded in the geometrical shape of script. In the experiments we found that this feature of fundamental frequency is a useful tool for not only discriminating text from non-text but also making detailed decisions such as language identification.

4. Rectangle Search

Although not directly related to preceding presentation, a method of finding rectangles will be described in this section. It is primarily due to the fact that the presence of a rectangle in general is a good indicator to the presence of a significant text inside.

The overall process consists of edge detection using Sobel operators, line detection by Hough transform, and extraction of line segments by referring to the edge image. Then we search for rectangles by appropriate combination of line segments. Rectangle search is based on the search for corner hypotheses generated by the end points

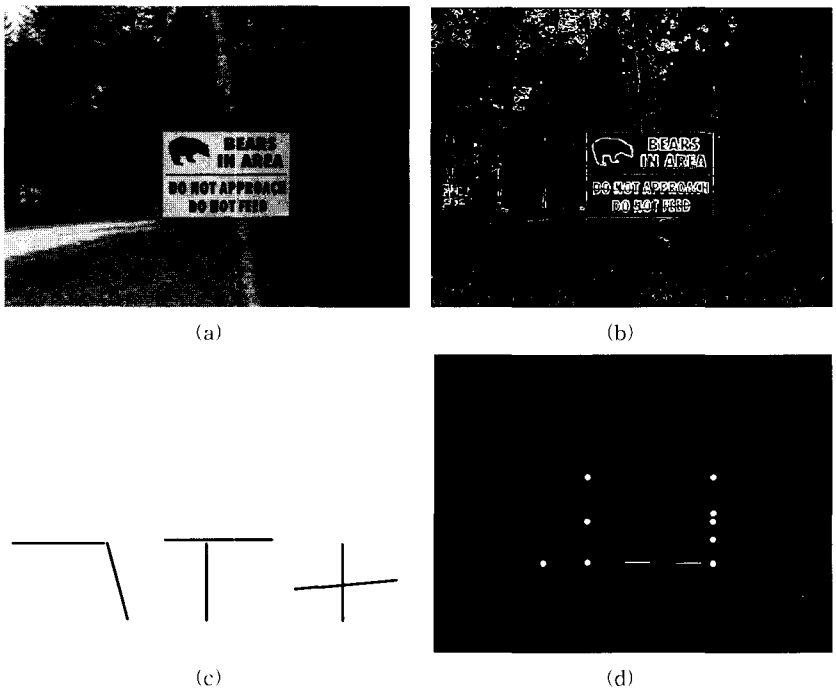


Fig 4 Rectangle detection. (a) an input image, (b) lines from Hough transform, (c) conditions for rectangle corners, and (d) the resulting rectangle corner hypotheses marked in white dots.

of line segments. Figure 4(c) shows the conditions for generating a corner hypothesis. Sample corner hypotheses are shown in Figure 4(d) highlighted in white dots. In this we simply choose the biggest rectangle hypothesis.

Hough transform is a well-known generic method of detecting a parametric curve in an image. Since its inception it has been widely used for various computer vision tasks. One is the study reported by Lee et al.[11]. Similar research has been carried out as a part of our research. But it is distinguished from that in a few aspects. A rectangle in our study need not be a regular box; rather an approximate rectangular tetragon is the target here.

Due to the importance of and interest involved in the task, many researchers have reported their own methods, most of which are based on Hough transform. The same is true of this research. But one novel feature of the proposed method is in searching procedure; the search is based on the hypotheses of end points of line segments. This is advantageous in that it has a greater probability of matching even when there is no line segments that cover whole side of a hypothetical rectangle.

5. Experiments

The proposed method of text locating has been tested using a prototype system. It has been developed using Visual C++ in Pentium III.

The test data include about one hundred images, a half collected from the World-Wide Web and the other half taken around the campus using a digital camera. Their size is about 420x300 on the average. The processing time for each image is well below a fraction of a second, with a slight increase due to Hough transform. Thus the system can be used in real time recognition.

In the experiment the first problem we have encountered is the choice of performance measures.

Unlike other pattern recognition or vision tasks, it is not easy to make reliable objective measurements for text location. The primary source of uncertainty is the presence of very small and undecipherable characters and very big and/or partially occluded characters. There are no universally agreed criteria as yet. In this study we have adapted a measurement used in others domains by including partial or incomplete text regions. The measure is

$$\frac{H}{H + FP + FN + H'}$$

the number of correct hits relative to the total number of hits and misses which may be partial or not; H is the number of hits correctly labelling a text block as text, and FP and NP are respectively the number of wrong positive labelling of non-text blocks and the number of omitted text blocks, and finally H' is the number of blocks extracted partially and unrecognizably.

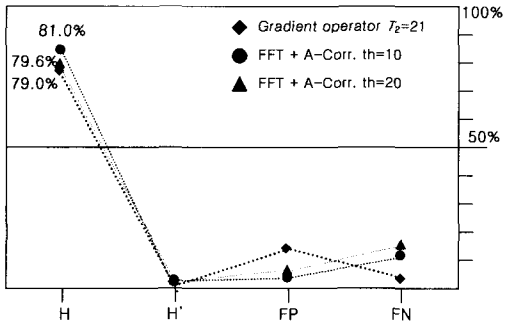


Fig 5 Text detection performance.

The graph shows high rates of successful hits, near 80 percent. By incorporating FFT and AC, the error rate has reduced by up to ten percent, while the false positives and false negatives changed detection rates each other.

The proposed feature is not only a good criterion for text vs non-text discrimination, but also a language-independent measure by nature. Hence we have tried locating English and Chinese texts in WWW images. The test sets include about 40 Korean images used in the previous test, 30 English text images, and sixteen Chinese text images. The result is summarized in the following table:

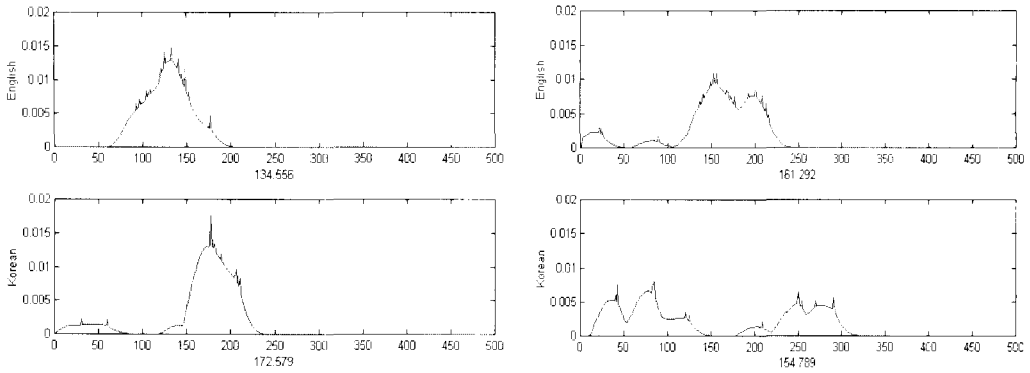
Table 1 Text block location performance by languages

	Gradient only	with FFT and AC
Korean	78.3 %	79.8 %
English	74.4 %	75.6 %
Chinese	75.7 %	78.6 %

The difference in the performance can be explained in the following terms: English texts contain low case letters with ascenders and descenders which confuses the frequency statistic along scan lines; Chinese characters are often highly complex and too compact to show characteristic spatial frequency. Nevertheless, the table shows that the proposed feature is quite reliable and works well when compared to other methods thus far published.

The preceding feature of harmonics, although simple and useful for locating text blocks, is not easily understood. Furthermore it is our conjecture that the problem of its language independence is due to the lack of the power in resolving languages. Every periodic or quasi-periodic signal posses a certain fundamental frequency that characterizes the input signal. This observation has led us to exploiting the very measure for detecting the generic category of the signals.

According to the study of signal processing, the phase information of speech is not important to our ears; likewise it is also conjectured that the phase of the text image is also not absolutely necessary. In order to test this point we have developed the idea of using the inverse Fourier transform of the logarithm of the phase-free magnitude of the Fourier transform of the signal. For this we have used text images of font size seven points and ten points scanned at 600 dpi. The interpretation of the result can be extended, with an acceptable range of variations, to the text images found in natural



(a) (b)

Fig 6 The fundamental frequency distribution of English and Hangul, respectively for font sizes (a) seven and (b) ten. Horizontal axis represents the spatial frequency with a number at the center below each box being the mean of the distribution

scenes.

The current test is limited to only two languages, Korean Hangul images and English words of Roman alphabet. Prior to the test of language identification, the model parameters for the fundamental frequencies of each language have been measured. We trained the model parameters using about twenty representative samples, followed by simple linear(or rates proportional to distance) smoothing to avoid insufficient training. Refer to Figure 4 and compare the resulting probability distribution for each font size.

The distributions of Figure 6 show a good separation of domain distributions. Using this statistic parameters we have test on another collection of text images containing both Hangul and English. The result is summarized in Table 2.

Table 2 Hangul-English classification using the fundamental frequency

	Labelling success rate
7 points	76.9%
10 points	78.5%

Although not satisfactory with this performance, the proposed method of language identification can aid in accelerating the recognition of characters in natural scene images of multi lingual environment. Once the language has been identified, the corresponding language's character recognizer is invoked, thus eliminating the need for invoking both/all recognizers and saving the computation time proportionately.

The final set of experiments involves the identification of the language of the text images located in scene images. It is the final set of experiments, which is not extensive and has not yet passed any parameter tuning step. The goal of the task is identifying the language of given image texts. We have prepared three sets of images: each with twenty images containing Korean, English or Chinese characters of arbitrary context and condition. The statistical model parameters have been trained using additional sets of images in similar numbers.

The resulting density distributions are illustrated in Figure 7. According to the figures, we can infer that the fundamental frequency of the English and Chinese is highly separated while those of Korean and English are similar to each other. Table 3 shows a simple summary of the experiment. The performance in Korean English discrimination recorded a little less than seventy percent. This is stimulating. Considering that the fundamental

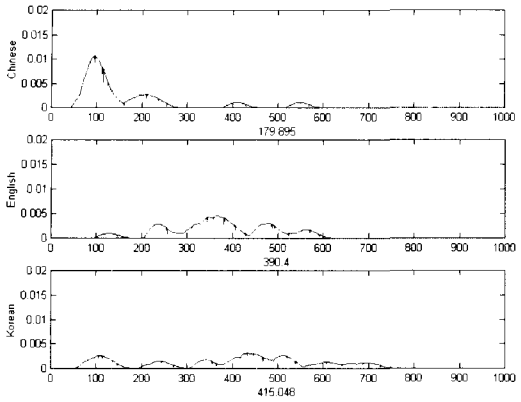


Fig 7 The frequency distributions for three language texts. Sample images were normalized by their height

Table 3 Language discrimination in natural scene images

Korean and English	67.4%
Korean, English, and Chinese	60.3%

frequency alone is too much a simplification, we consider that the spatial frequency feature alone is quite a powerful tool for deciding whether a block region of image is text or not, and, if it is, what language it is. We believe that the proposed method can be used reliably with some elaboration of measurements. One of the most notable facts is that the proposed feature of spatial frequency [measurements] is highly generic and universal. However, this is not to say that it is complete per se; we need future study for a accurate understanding of the frequency features

6. Conclusion

This paper proposed a method for locating text blocks and identifying the language in natural outdoor scene images. It uses highly intuitive and consistent characteristic of spatial frequency, a measure of capturing the repetitive structure of characters along a scan line. The experimental results showed sufficiently high results, and supported our assertion that the feature is useful across many languages.

Another notable point of the paper is rectangle location in the context of text location. Currently we are locating approximately rectangular blocks containing some text. The idea is deemed very useful for many practical applications. One future extension will be arbitrary quadrilaterals for estimating perspective distortions and recognition of poor quality characters in the block images.

References

[1] A. K.Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," Machine Vision and Applications, Vol.5, pp.169-184, 1992.

[2] A.K.Jain and B.Yu, "Automatic text location in images and video frames," Pattern Recognition, Vol.32, No.12, pp.2055-2076, 1998.

[3] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," IEEETrans.PAMI, Vol.22, No.4, pp.385-392, 2000.

[4] J. Zhou and D. Lopresti, "Extracting text from WWW images," Proc.ICDAR '97V, pp.248-252, August 1997.

[5] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color image," Pattern Recognition, Vol.28, No.10, pp.1528-1535, 1995.

[6] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: an automatic system to detect and recognize text in images," IEEE Trans. PAMI, Vol.21, No.11, pp.1224-1229, Nov. 1999.

[7] X. Wang, X. Ding, C. Liu, "Character extraction and recognition in natural scene images," in Proc. ICDAR02, Seattle, USA, pp.1084-1088, Sept. 2001.

[8] M.-C. Roh, Y.-W. Choi, and S.-W. Lee, "Scene text extraction of natural scenes in video frames," in Proc. KISS Autumn Workshop on CVPR 2001, Seoul, Korea, pp.161-162, Nov. 2001. (in Korean)

[9] K. Jung, K.-I. Kim, and J. H. Han, "Efficient scene text extraction on planar planes," in Proc. KISS Autumn W. CVPR 2001, Seoul,

Korea, pp.165-166, Nov. 2001.

[10] I.-Y. Jang, B.-C. Ko, K.-C. Kim, and H.-R. Byun, "Automatic text extraction in video images using Morphology," in Proc. KISS Autumn W. CVPR 2001, Seoul, Korea, pp.169-170, Nov. 2001. (in Korean)

[11] H.-S.Lee and J.-H. Lee, "Tetragon detection using Hough transform," in Proc. KISS Autumn W. CVPR 2001, Seoul, Korea, pp.189-190, Nov. 2001. (in Korean)

신 봉 기

1985년 서울대학교 자원공학 학사. 1987년 한국과학기술원 전산학 석사. 1995년 한국과학기술원 전산학 박사. 1987년~1999년 한국통신 멀티미디어연구소. 1999년~현재 부경대학교 전자컴퓨터정보통신공학부 전임강사. 관심분야는 인공지능, 패턴인식, 지능형 에이전트, 정보검색

김 선 규

1999년 부경대학교 정보통신공학 학사. 2002년 부경대학교 정보통신공학 석사. 2002년 ~ 현재 ㈜한국머털테크 근무. 관심분야는 정보통신, 패턴인식, PDA