# PROPOSAL OF THE HYBRID SPECTRAL GRADIENT METHOD TO EXTRACT CHARACTER / TEXT REGIONS FROM GENERAL SCENE IMAGES

*Yoichiro BABA and Akira HIROSE*

Dept. Electron. Eng., The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

## ABSTRACT

We propose the spectral gradient method that is a novel method to extract the character / text regions from general scene images. We obtain the distribution of the degree of likelihood of character / text region by calculating spatial variation of texture. We evaluate the texture variation by the gradient of local spatial spectra. A characteristic Fourier transform process, named Hybrid spectral gradient method, is also developed to achieve a high extraction performance. This method is based on the human foveation and can be applied for a wide range of languages and letters.

## 1. INTRODUCTION

Character / text region extraction is one of the most important and difficult steps in recognizing texts in general scene images. In this field, there have been many researches reported before [1]-[8]. The most typical approach to cut out character / text regions is a rule based method [1]-[4]. The technique assumes many empirical rules. For example, characters in a string are located closely one another, they have an identical color, outward form, size, and so on. Though the computational time can be small, the performance is often affected by some illness of the relation between the characters and the background.

Another approach is called the texture-based method [5]-[6] which is more robust than the rule based one. The general process of the texture-based method is to segment an image by paying attention to texture. It divides and classifies the image into several regions based on texture difference. Then it evaluates whether each region includes characters or not.

However, the texture method has some problems. First, it requires a large computational cost. In general, the texture features are extracted in the frequency domain. Therefore, the texture-based method requires a two-dimensional Fourier transform whose computational time and memory size are quite large. Secondly, in not a few actual cases, the statistical feature of texture in a character region is very similar to that in non-character regions. That is to say, the texture *itself* is not a good index to discriminate character and non-character regions.

To solve these problems, we propose in this paper a novel method that detects continuous textural change in space. We name it the spectral gradient method. We find that the spatial change of texture, which characters also possess inherently, attracts attention. We introduce a unique definition of texture variation that realizes high-resolution texture change detection. It is shown that our method extracts character / text region successfully and very robustly almost independent of character kinds and background nature.

## 2. ALGORITHM

Let us consider an image $I$ whose pixels are located at nodes of a rectangular grid $(x, y)$ and the number is totally $X \times Y$. In our method, there are two steps: (1)First we calculate the degree of likelihood that a pixel is included in a character region and yield a likelihood distribution image. (2)Then we determine and extract the character / text region.

Firstly, we evaluate the spatial texture change by calculating on the spectral gradient. Then we generate an evaluation image that shows the likelihood distribution that each pixel is included in a character region. We call this process *the spectral gradient method*. In the image $I$, $s^{\mathrm{col}}(x, y)$ denotes the pixel value where the superscript $\mathrm{col}$ has three color elements $\{r, g, b\}$ (red, green and blue) for example. In the case of a gray scale image, the image data is degenerate in colors and, therefore, we can neglect the superscript. In this paper, we consider color images. The color element number is $Col = 3$. The following algorithm is applicable both of color and gray-scale images without modification. That is, the system does not need to decide if the image is color or gray-scale. Unless, if we choose $Col = 1$, then the calculation cost is reduced to $1/3$. The following two subsections describe two possible algorithms that the first step performs to detect spatial spectral change.

### 2.1. Two-dimensional spectral gradient method

We calculate the degree of likelihood that a pixel $s^{\mathrm{col}}(x, y)$ in an image $I$ is included in a character region as follows. We prepare a window $W$ of $L \times M$ size whose center position is $(x, y)$. Pixels in the window is denoted as $(x', y')$ $(\in W)$.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

We choose the Hanning window $w(x'-x, y'-y)$ as the weight profile for example. We obtain the local spatial frequency spectrum $S^{\mathrm{col}}_{uv}(x,y)$ by applying the 2-dimensional Fourier transform to the local image in the window $W$ where $u$ and $v$ are spatial discrete frequencies.

Next we calculate the spatial gradient of the power spectrum $\nabla|S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2$. The gradient is a vector of $Col \times U \times V$ in the color and frequency domain and defined spatially discretely as

$$\nabla|S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2 \equiv \left(|S^{\mathrm{col}}_{uv}(x+1,y,u_p,v_q)|^2 - |S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2\ ,\right. \\ \left. |S^{\mathrm{col}}_{uv}(x,y+1,u_p,v_q)|^2 - |S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2\right) \quad (1)$$

where $u_p$ and $v_q$ denote discrete spatial frequency in $x$ and $y$ directions, respectively. The number of them are $U$ and $V$, respectively. Our introspection suggests that, if the absolute value (vector length) of the spectral gradient $\nabla|S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2$ is large, the pixel $s(x,y)$ is considered to have a high possibility to be included in a character region. That is, the degree of likelihood $A_{2\mathrm{D}}$ is determined as

$$A_{2\mathrm{D}}(x,y,u_p,v_q) \equiv \sum_{\mathrm{col}}\sum_{u_p}\sum_{v_q} l\left(\nabla|S^{\mathrm{col}}_{uv}(x,y,u_p,v_q)|^2\right) \quad (2)$$

where $l(\cdot)$ denotes a norm of a vector. Each summation is executed for frequencies and colors. Suffix 2D means the 2-dimensional Fourier transform. Accordingly the estimation $A_{2\mathrm{D}}$ expresses a total texture change including colors.

This method deals with the spatial change of the 2-dimensional power spectra $|S^{\mathrm{col}}_{uv}|^2$. We name this method the 2-dimensional spectral gradient method. Though this method works quite well to determine human attention regions, it is also found that the improved method mentioned below is more effective in particular for character region extraction.

### 2.2. Hybrid Spectral Gradient Method

The 2-dimensional spectral gradient method estimates that the pixel has a higher attraction ability to human being when the window at the pixel has a different texture from those of the neighboring pixels. In our experimental examinations of the method, we find that, when the window size is smaller than the character area, the system does not detect the texture change. That is, the system is insensitive to central part of a character region when the region is large.

To realize a fine texture change detection, the system has to have a high spatial-resolution Fourier transform. However, the Fourier resolution has a trade-off relation with the spatial window width. The frequency domain feature reflects a statistical property of certain number of pixels in the window area. Therefore, the 2-dimensional spectral gradient method has a difficulty to detect the fine structure.

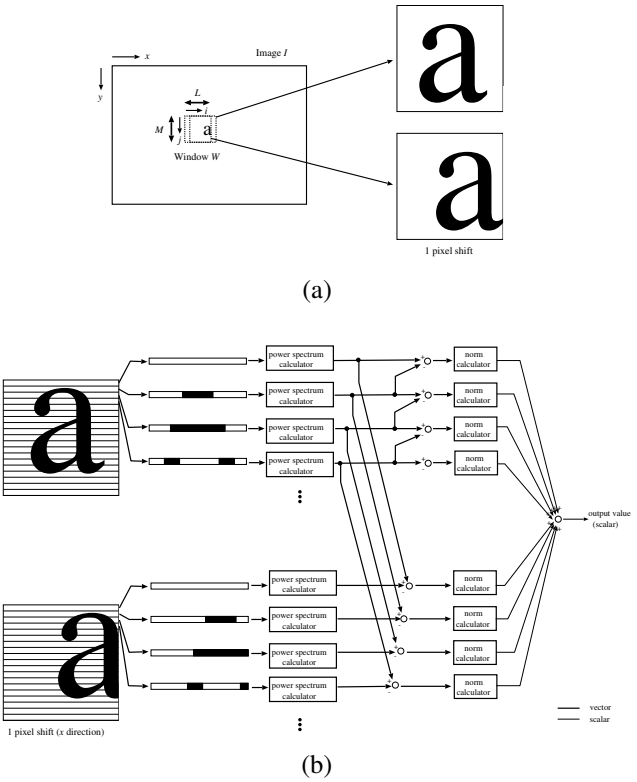


**Fig. 1**. (a)Window and its shift in $x$ direction in image $I$ and (b)schematic illustration of the hybrid spectral gradient method consisting of power spectrum, gradient, norm and summation calculations.

In this paper, we propose a characteristic frequency analysis process that is based on the 1-dimensional Fourier transform. We employ the 1-dimensional Fourier transform in $x$ direction, for example, and then we calculate the gradient in $y$ direction. We also perform the same analysis in another directions, i.e., $y$-direction Fourier transform and $x$-direction gradient calculation, to reduce the anisotropy. We name this method the hybrid spectral gradient (HSG) method where the frequency-domain feature in one direction is combined with the spatial gradient in another direction. In this method, it is found in the experiment reported below that a high spatial resolution can be compatible with a high frequency resolution in an effective manner for the character region determination.

Figure 1 illustrates the HSG method. First we apply the 1-dimensional Fourier transform to pixels in a $x$-directional line in the $(x,y)$-centered window $W$ to obtain the frequency spectrum $S^{\mathrm{col}}_{u}(x,y)$ ( $\in$ complex number) where $u$ denotes $x$-directional discrete frequency. Then we calculate the gradient in $x$ and $y$-directions to get the spatial variation of the power spectrum $d|S^{\mathrm{col}}_{u}(x,y)|^2$ where $d$ denotes the differential in space. In the same manner, we calculate the counterpart in another direction and obtain $d|S^{\mathrm{col}}_{v}(x,y)|^2$.

Next we calculate the summation of these orthogonal hybrid spectral gradient values also by taking into consideration the color domain. Finally we regard the summation value as the degree of the likelihood that the pixel is included in a character / text region, i.e., the attracting region. The process is expressed by the following equations.

The pixel values $s^{col}(x', y')$ on a line in $x$ direction is 1-dimensionally Fourier transformed into $S_u^{col}(x, y, y', u)$, while those on a line in $y$ direction is into $S_v^{col}(x, y, x', v)$. We first consider the gradient of $S_u^{\mathrm{col}}(x, y, y'_j, u_p)$ in $x$ and $y$ directions. Note that the spectra $S_u^{\mathrm{col}}$ is a function in the spatial and frequency domains, whereas the gradient is determined in the space in $x$ and $y$ directions.

$$\frac{\partial |S_u^{\mathrm{col}}(x,y)|^2}{\partial x} = l_{y'_j}\left(l_{u_p}\left(|S_u^{\mathrm{col}}(x+1,y,y'_j,u_p)|^2 - |S_u^{\mathrm{col}}(x,y,y'_j,u_p)|^2\right)\right) \quad (3)$$

$$\frac{\partial |S_u^{\mathrm{col}}(x,y)|^2}{\partial y} = l_{y'_j}\left(l_{u_p}\left(|S_u^{\mathrm{col}}(x,y+1,y'_j,u_p)|^2 - |S_u^{\mathrm{col}}(x,y,y'_j,u_p)|^2\right)\right) \quad (4)$$

where $l_{y'_j}(\cdot)$ and $l_{u_p}(\cdot)$ means the vector length (i.e., norm) in the space formed by $y'_j$ ($j$=1, 2, 3, ...) and $u_p$ ($p$=1, 2, 3, ...), respectively. Then we sum them up to yield the total differential $d|S_u^{\mathrm{col}}(x,y)|^2$ as

$$d|S_u^{\mathrm{col}}(x,y)|^2 = \frac{\partial |S_u^{\mathrm{col}}(x,y)|^2}{\partial x}dx + \frac{\partial |S_u^{\mathrm{col}}(x,y)|^2}{\partial y}dy \quad (5)$$

We execute the same process also in the orthogonal direction. The result is written as $d|S_v^{\mathrm{col}}(x,y)|^2$.

Lastly, we determine the degree of likelihood of character / text region $A_{\mathrm{H}}$ as the summation of $d|S_u^{\mathrm{col}}|^2$ and $d|S_v^{\mathrm{col}}|^2$ and evaluate the norm in the color space col. Though there are various possible definitions of the summation of $u$ and $v$ elements, we use a norm-like operator $l_{u+v}(\cdot)$ for this summation. The norm in the color domain is also expressed as $l_{\mathrm{col}}(\cdot)$ following the formulation above.

$$A_{\mathrm{H}}(x,y) \equiv l_{\mathrm{col}}\left(l_{u+v}\left(d|S_u^{\mathrm{col}}(x,y)|^2, d|S_v^{\mathrm{col}}(x,y)|^2\right)\right) \quad (6)$$

In the following experiments, we adopt the absolute summation and the Manhattan distance to reduce the calculation cost in software or hardware implementation. In this case, (6) are rewritten as

$$A_{\mathrm{H}}(x,y) \equiv \sum_{\mathrm{col}}\sum_{u_p}\sum_{v_q}\left[\sum_{j=1}^{M} d\left(\frac{\partial |S_u^{\mathrm{col}}|}{\partial x}, \frac{\partial |S_u^{\mathrm{col}}|}{\partial y}\right) + \sum_{i=1}^{L} d\left(\frac{\partial |S_v^{\mathrm{col}}|}{\partial x}, \frac{\partial |S_v^{\mathrm{col}}|}{\partial y}\right)\right] \quad (7)$$

where $d(\cdot)$ denotes the Manhattan distance.

Incidentally, it is found that the calculation cost of (7) can be reduced by changing the calculation order. That is, we can modify (7) as

$$A_{\mathrm{H}}(x,y) \equiv \sum_{\mathrm{col}}\left[\sum_{j=1}^{M}\sum_{u_p} d\left(\frac{\partial |S_u^{\mathrm{col}}|}{\partial x}, \frac{\partial |S_u^{\mathrm{col}}|}{\partial y}\right) + \sum_{i=1}^{L}\sum_{v_q} d\left(\frac{\partial |S_v^{\mathrm{col}}|}{\partial x}, \frac{\partial |S_v^{\mathrm{col}}|}{\partial y}\right)\right] \quad (8)$$

In this way, the tangled $j$ summation and the $u$ and $v$ summations are separated and the redundancy of the Fourier transform is avoided. The calculation cost is reduced by a factor of $M$. In the same manner, the second term of the right-hand side of (7) is reduced by a factor of $L$ in (8).

### 2.3. Adaptive character / text region determination

The second step determines adaptively the areas that are estimated to be character / text regions in the output image of the first step $A_{\mathrm{H}}$. Details will be presented.

## 3. EXPERIMENT

### 3.1. Experimental condition and calculation time

We carried out the following experiments for color still and movie images of 512×384 pixels. We scan an image for all the pixels $s(x,y)$ to get the likelihood image $A_{\mathrm{H}}(x,y)$, though we cannot touch those in the image edge regions within the half length of the local window size $L = M = 32$ from the edges.

The total processing time for a single image is about 3 seconds on a 2.4GHz Pentium 4 PC. The present system is installed as software. However, if we realize hardware, we can deal with movie stream in real time (e.g., 30 frames / second for 512×384 pixels). In this paper, the movie experiment below is conducted off-line.


**Fig. 3**. Result example for an image including English and Japanese simultaneously.


**Fig. 2**. Result example: (a)likelihood image $A_{\mathrm{H}}$, (b)B/W image obtained by a threshold process, (c)extracted character / text regions (surrounded by pink curves) superimposed on the original image.

**Fig. 4**. Result example for a movie image stream. Pink curves indicate extracted character / text regions.

### 3.2. Results and discussion

**Still scene image:** Figure 2 shows the results obtained for a still scene image. Figure 2(a) is the normalized likelihood image $A_{\mathrm{H}}(x, y)$ obtained by (8) where the brightness corresponds to the degree of the likelihood that the pixel is included in the character / text region. Figure 2(b) is the binarized image of (a) with a threshold of 0.5. Figure 2(c) is the superimposition of (b) on the raw image in which the pink curves surround the high likelihood region. That is to say, the inside of the pink curves are the estimated character / text regions. It is found in the result Fig.2(c) that the proposed method extracts the characters successfully.

**English and Japanese mixture image:** Figure 3 on the previous page gives an example where we have to process both the roman letters and the Chinese / Japanese characters simultaneously. The image contains English and Japanese texts. It is found that both of them are extracted successfully. For this image, if the system uses the 2-D spectral gradient method, it often segments faultily the high contrast parts, such as the display edges, as well as the texts. However, with the HSG method (8), it is not attracted by such a high contrast. Accordingly, the performance is not affected by the variation of languages, characters and letters. This fact is a significant merit of our proposal.

**Movie frame stream:** Figure 4 presents the result for a movie frame stream. The present system deals with the sequential images independently from each other. That is, each image is a still scene with different expansion / reduction, translational shift and view angle. In Fig.4, the signboard approaches to the video camera. Therefore the ambiguous and collapsed characters become gradually larger and clearer. Even in this case, we find the extraction is steadily successful.

## 4. CONCLUSION

We have proposed the Hybrid Spectral Gradient (HSG) Method to extract character / text regions from scene images. The experiments have demonstrated that the HSG system extracts robustly the character regions.

## 5. REFERENCES

[1] Anil K.Jain and Bin Yu, Pattern Recognition, 31, 12 (1998) 2055-2076

[2] Yu Zhong, Kalle Karu, and Anil K. Jain, Pattern Recognition, 28, 10 (1995) 1523-1535

[3] Karin Sobottka, Horst Bunke, and Heino Kronenberg, in Proc. ICDAR'99 (1999) 57-62

[4] J.Ohya, A.Shio, and S.Akamatsu, IEEE Trans. Pattern Analysis and Machine Intelligence, 16 (1994) pp. 214-220

[5] Victor Wu, Raghavan Manmatha, and Edward M. Riseman, IEEE Trans. Pattern Analysis and Machine Intelligence, 21 (1999) 1224-1229

[6] Huiping Li, David Doermann, and Omid Kia, IEEE Trans. Image Processing, 9, 1 (2000) pp. 147-156

[7] Xiaoou Tang, Xinbo Gao, Jianzhuang Liu, and Hongjiang Zhang, IEEE Trans. Neural Netw., 13, 4 (2002) 961-971

[8] C.Garcia and X.Apostolidis, in Proc. IEEE International Conference, Speech, Signal Processing, 4 (2000) 2326-2329