Text Detection in Natural Scene Images by Stroke Gabor Words

Chucai Yi
Dept. of Computer Science
The Graduate Center, City Univ. of New York
New York, U.S.A.
e-mail: CYi@gc.cuny.edu

Yingli Tian
Dept. of Electrical Engineering
The City College and Graduate Center, City Univ. of New York, New York, U.S.A.
e-mail: ytian@ccny.cuny.edu

***Abstract*—In this paper, we propose a novel algorithm, based on stroke components and descriptive Gabor filters, to detect text regions in natural scene images. Text characters and strings are constructed by stroke components as basic units. Gabor filters are used to describe and analyze the stroke components in text characters or strings. We define a suitability measurement to analyze the confidence of Gabor filters in describing stroke component and the suitability of Gabor filters on an image window. From the training set, we compute a set of Gabor filters that can describe principle stroke components of text by their parameters. Then a *K*-means algorithm is applied to cluster the descriptive Gabor filters. The clustering centers are defined as Stroke Gabor Words (SGWs) to provide a universal description of stroke components. By suitability evaluation on positive and negative training samples respectively, each SGW generates a pair of characteristic distributions of suitability measurements. On a testing natural scene image, heuristic layout analysis is applied first to extract candidate image windows. Then we compute the principle SGWs for each image window to describe its principle stroke components. Characteristic distributions generated by principle SGWs are used to classify text or non-text windows. Experimental results on benchmark datasets demonstrate that our algorithm can handle complex backgrounds and variant text patterns (font, color, scale, etc.).**

Keywords- Gabor Filter; Stroke Component; Suitability Measurement; Stroke Gabor Words; SGW Characteristic Distributions

I. INTRODUCTION

Camera captured text information in natural scene images can serve as indicative marks in many image-based applications such as assistive navigation, auxiliary reading, image retrieval, scene understanding, etc. Different from the scanned document images [1, 11], extracting text from natural scene images is a challenging problem because of complex backgrounds and large variations of text patterns such as font, color, scale, and orientation.

Many optical-character-recognition (OCR) systems, either open source or commercial software, have been developed to recognize text by taking character corners or junctions as feature points of learning and matching on scanned documents. But these OCR systems cannot automatically filter out variant background outliers in natural scene images. When a raw natural scene image is input into an OCR system, the text recognition rate is often very low. Therefore, to extract text from natural scene images, text detection is an essential step to compute the image sub-regions containing text characters or strings.

Many rule-based algorithms have been proposed for text detection [6, 12, 18]. They extracted text characters and strings by using gradient-based and color-based local features, including minimum size, aspect ratio, edge point density, gradient distribution, color uniformity and stroke width consistency. But these features are sensitive to variant text patterns and background outliers that resemble text characters. Many researchers applied a machine learning model to solve the problems of text detection. Chen *et al.* [3] developed an Adaboost learning framework by using selected Haar features, joint histogram of intensity and intensity gradient, and distribution of edge points as features to train classifiers. Pan *et al.* [16] extracted segments of character boundaries as features and employed a K-SVD based learning model to detect text. Hu *et al.* [7] presented an adaptive Frechet Kernel based support vector machine (SVM) for text detection. Kumar *et al.* [9] established a set of globally matched wavelet filters as feature descriptors and used SVM and Fisher classifier to classify image windows as text or non-text. Generally, the patterns of tangible objects, such as face, human body and car, are stable for learning-based object detection. The dissimilarity between different training samples and testing samples is small enough or can be lowered by alignment and normalization. However these benefits are not applicable to text in natural scene images.

As basic element of text character and text string, stroke provides robust features for text detection in natural scene images. Text can be modeled as a combination of stroke components with a variety of orientations, and features of text can be extracted from combinations and distributions of the stroke components. In this paper, a novel algorithm is proposed to detect text regions by using Gabor filter responses to model the stroke components of text. It is able to handle complex backgrounds and variant text patterns. The contributions of this paper are: (1) A suitability measurement of Gabor filter to measure its confidence in stroke component description and its suitability for an image window. (2) Stroke Gabor Words for universal descriptions of stroke components. (3) A classification algorithm based on the characteristic distribution of suitability measurements generated by principle SGWs. The flowchart of our algorithm is presented in Fig. 1. In this paper, we use image window, a rectangle image sub-region with fixed aspect ratio, as a basic processing cell.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.44

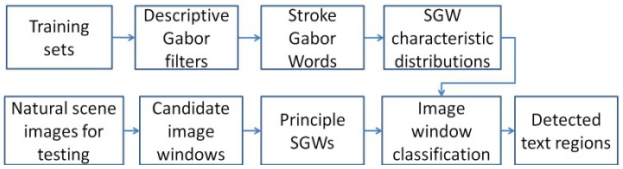

Figure 1. The flowchart of our algorithm.

II. DESCRITPIVE GABOR FILTERS

A. Gabor filter Descriptions of stroke components

We use Gabor filters to describe the stroke components in text characters or strings. Gabor filter was widely used for texture analysis and image representation [10, 20]. Gabor filter was also employed for segmentation in the document image [8, 17, 19]. In [4], Gabor filter is applied to obtain local features for text recognition after text detection and affine rectification. In [17], Gabor filter is used for script identification. A 2-D Gabor filter is a Gaussian kernel modulated by a sinusoidal carrier wave, as in (1) and (2). It gives responses to structure of line segment in scene images.

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (1)$$

where

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta \quad (2)$$

A Gabor filter bank is built in accordance with the 5 parameters $<\lambda,\theta,\psi,\sigma,\gamma>$. Given an image window $w(x,y)$, Gabor filter response r is obtained by convolution with a Gabor filter g from the filter bank. A Gabor filter g can be used to describe a stroke component as long as it generates the maximum energy of response among all the Gabor filters in the bank.

B. Suitability Measurement

The stroke component can be approximately reconstructed by its descriptive Gabor filter, but there exists errors between Gabor filter description and actual stroke component in image. To model the confidence of Gabor filters in describing stroke component, we define a suitability measurement based on binary Gabor filter response map $B_{GR}(w)$ and binary stroke map $B_{SM}(w)$. The Gabor filter response map $B_{GR}(w)$ is obtained by inserting an additional parameter t ($0 \le t \le 1$) into the Gabor filter to binarize the Gabor filter response map $\boldsymbol{r}$, by (3).

$$B_{GR}(\boldsymbol{w};\boldsymbol{g}) = B_{GR}(\boldsymbol{r}) = \begin{cases} 1 & if\ \boldsymbol{r} \ge t\cdot\boldsymbol{r}_{max} \\ 0 & if\ \boldsymbol{r} < t\cdot\boldsymbol{r}_{max} \end{cases} \quad (3)$$

where $\boldsymbol{r}_{max}$ represents the maximum value on the Gabor filter response map. Thus a Gabor filter $\boldsymbol{g}$ is denoted by a vector of the 6 parameters $\langle\lambda,\theta,\psi,\sigma,\gamma,t\rangle$ (see Fig. 2(c-d)). The stroke map $B_{SM}(\boldsymbol{w})$is obtained by labeling the pixels located in the torso of strokes. According to the definition in [6], stroke is a set of pixels in connecting paths of two edge pixels with approximately equal gradient magnitude and opposite gradient directions. On the basis of edge map and gradient map, we construct probe rays at each edge pixel along the gradient direction to find out the satisfied connecting paths. Then stroke map is obtained by assigning foreground value to pixels in these paths, as shown in Figure 2(b).

The suitability measurement D is defined as the pattern correlation between $B_{GR}(\boldsymbol{w};\boldsymbol{g})$ and $B_{SW}(\boldsymbol{w})$ by (4).

$$D(\boldsymbol{w};\boldsymbol{g}) = \frac{2\times|B_{SW}(\boldsymbol{w})\ \&\ B_{GR}(\boldsymbol{w};\boldsymbol{g})|}{|B_{GR}(\boldsymbol{w};\boldsymbol{g})| + |B_{SW}(\boldsymbol{w})|} \quad (4)$$

where the $|\cdot|$ represents the number of pixels with value 1 in the binary map. $D(\boldsymbol{w};\boldsymbol{g})$ is in the range $[0,1]$.

On a stroke component, D models the confidence of Gabor filters in stroke description. However, a text window usually includes a number of different stroke components. $D(\boldsymbol{w};\boldsymbol{g})$ is used to model suitability of Gabor filter $\boldsymbol{g}$ for image window $\boldsymbol{w}$. A large $D(\boldsymbol{w};\boldsymbol{g})$ will be obtained as long as Gabor filter $\boldsymbol{g}$ has high confidence in describing a stroke component with high frequency of occurrence in image window $\boldsymbol{w}$. This Gabor filter is said to be suitable for $\boldsymbol{w}$, and this stroke is defined as principle stroke component of $\boldsymbol{w}$. Text detection can be transformed into detection of principle stroke components.

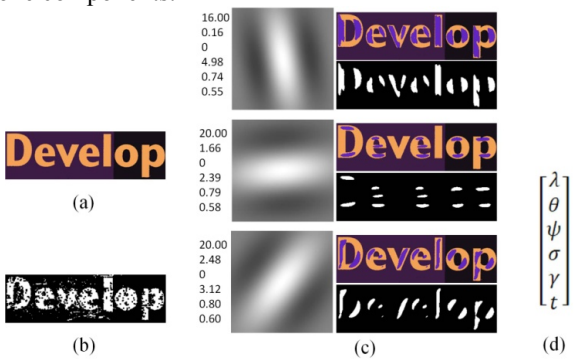

Figure 2. (a) Original image; (b) binary stroke map; (c) three Gabor filters marked with the parameter vectors and the corresponding binary Gabor filter response maps to display stroke components; (d) ordering of Gabor filter parameters listed in (c).

III. STROKE GABOR WORDS

A. Training Set

From natural scene images with text regions manually labeled, we generate positive training samples by slicing each ground truth region vertically into overlapped windows with width-to-height ratio 2:1, as shown in Fig. 3(a). Each sample contains text in regular print patterns and they are normalized into fixed size, height 48 and width 96 pixels. From the same natural scene images, we generate negative training samples by the bounding boxes of non-text object boundaries. In addition, we take 1402 images covering multiple background scenes where text information might exist. But they do not include any text information.

Negative samples are generated by randomly cutting out image windows with height 48 and width 96 pixels. The training set contains 2711 positive samples and 9208 negative samples in total. To ensure that maximum Gabor responses are generated at stroke components, all positive samples of this training set have higher stroke intensity than background intensity.

B. Stroke Gabor Words

We build a set of Gabor filters for universal description of stroke components from the training set. At first, the descriptive Gabor filters of principle stroke components are computed for each training sample $\boldsymbol{w_m}$. Principle stroke components serve as main bodies of text characters in an image window, so they are described by Gabor filters whose responses are compatible with stroke map of this window (see Fig. 2(c)). From the Gabor filter bank, we find out a set of Gabor filters generating maximum values of suitability measurement. In (5), the most suitable Gabor filter $\boldsymbol{g}^*$ describes the most frequent principle stroke component in window $\boldsymbol{w_m}$. Then we build an ordered list of Gabor filters from the text window according to the suitability measurements. A threshold T_g is set to calculate the descriptive Gabor filters of principle stroke components that construct text in $\boldsymbol{w_m}$ by (6).

$$\boldsymbol{g}^* = \text{argmax}_{\boldsymbol{g}}\, D(\boldsymbol{w_m}; \boldsymbol{g}) \tag{5}$$

$$G_m = \left\{ \boldsymbol{g_i} \middle| D(\boldsymbol{w_m}; \boldsymbol{g_i}) / D(\boldsymbol{w_m}; \boldsymbol{g}^*) \geq T_g \right\} \tag{6}$$

We combine the descriptive Gabor filters from all positive samples, and weighted K-means is applied to cluster the parameter vectors of descriptive Gabor filters. The clustering centers are defined as Stroke Gabor Words (SGWs) denoted by S. Fig. 3(b) presents some example SGWs. The SGWs are Gabor filters serving as stroke basis for universal description of principle stroke components in the positive training samples.

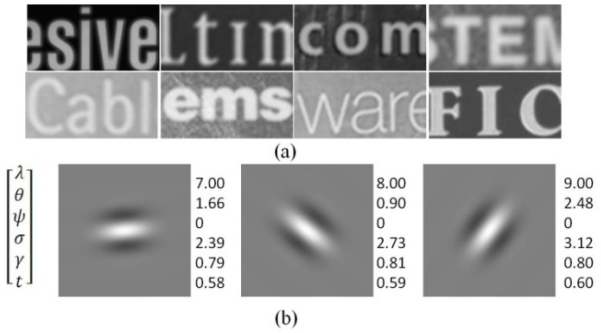

Figure 3. (a) Examples of text windows from positive training samples; (b) examples of SGWs with parameter lists on the right.

Based on the suitability measurements, an image window can be mapped to an ordered list of SGWs. We can find out a subset of SGWs that describes the stroke components in the image window. They are defined as principle SGWs of image window $\boldsymbol{w_t}$, which have larger suitable measurements than the other SGWs, and we calculate them by (7) and (8).

$$\boldsymbol{s}^* = \mathbf{argmax}_{\boldsymbol{s}} D(\boldsymbol{w_t}; \boldsymbol{s}) \tag{7}$$

$$S_t = \{ \boldsymbol{s_i} | D_G(\boldsymbol{w_t}; \boldsymbol{s_i}) / D_G(\boldsymbol{w_t}; \boldsymbol{s}^*) \geq T_{dc} \} \tag{8}$$

where S_t is the set of principle SGWs for $\boldsymbol{w_t}$ and $\boldsymbol{s}^*$ is the most suitable SGW for $\boldsymbol{w_t}$. The set of principle SGWs compose a subset of the K SGWs. In our experiments, we set $T_g = 0.95$, $K = 25$, and $T_{dc} = 0.975$ to achieve the best performance. The principle SGWs and corresponding suitability measurements are features of an image window, which are used to classify it as either a text or non-text window.

C. Classification Algorithm Based on Principle SGWs

Text region detection requires localization to obtain preliminary image windows and classification to determine text windows and non-text windows. Heuristic layout analysis is performed to partition the scene image into a set of candidate image windows. It is based on the magnitude gradient difference in Laplacian map [18] and the adjacent character grouping to find out all possible fragments of text strings, which are three or more edge boundaries with approximately equal heights, distances and horizontal alignment [21].

Then we propose a novel classification algorithm based on distribution of suitability measurements of each SGW in training set to classify the candidate image windows as either text or non-text windows. Each SGW is able to give a vote of image window classification according to statistics of its suitability measurements on training samples. On the positive samples, most values of SGW suitability measurements are distributed in the range (0.55, 0.7). Inspired by the Rayleigh nature of Gabor filter outputs in the texture analysis [2], we employ a mirror reversed Rayleigh distribution to model the statistical results of suitability measurements from positive samples.

$$P_{\boldsymbol{s}}(D) = \frac{1-D}{{\sigma_P}^2} \exp\left(-\frac{(1-D)^2}{2{\sigma_P}^2} \right) \tag{9}$$

where N_P is the size of positive training samples and σ_P is a parameter whose maximum likelihood is $\hat{\sigma}_P = \sqrt{(1/2N_P)\sum_{p=1}^{N_P}(1-D)^2}$. On the negative samples, suitability measurements are irregularly distributed in the range [0, 1] because the negative training samples contains multiple patterns without any constraints. Thus we use the Gaussian distribution to model the statistical results of suitability measurements from negative samples.

$$N_{\boldsymbol{s}}(D) = \frac{1}{\sqrt{2\pi{\sigma_N}^2}} \exp\left(-\frac{(D-\mu_N)^2}{2{\sigma_N}^2} \right) \tag{10}$$

where μ_N and σ_N are mean and variance which can be estimated by maximum likelihood. As shown in Fig. 4, $P_{\boldsymbol{s}}(D)$ and $N_{\boldsymbol{s}}(D)$ are characteristic distributions of the SGW.

They indicate the probability that an image window will be classified as text window or non-text window by SGW $\boldsymbol{s}$.

If an image window $\boldsymbol{w_t}$ contains text information, its principle SGWs give positive votes. The corresponding suitability measurements should be mapped to high probability in the mirror reversed Rayleigh distributions and to low probability in the Gaussian distributions. If $P_s\big(D(w_t; s_i)\big) \geq P_s(\mu_P - 0.3\sigma_P)$ where $\mu_P = 1 - 1.253\sigma_P$ or $N_s\big(D(w_t; s_i)\big) \leq N_s(\mu_N \pm \sigma_N)$, the principle SGW s_i gives a positive vote value 1, otherwise it gives a negative vote value -1. A weighted sum of the votes of all principle SGWs is calculated to make image window classification by (11).

$$H(\boldsymbol{w_t}) = sgn\left(\sum_i \beta_i V_i\right) \tag{11}$$

where V_i denotes the vote values and $\beta_i = s_i / \sum_{s \in S_t} s$ denotes the weights.

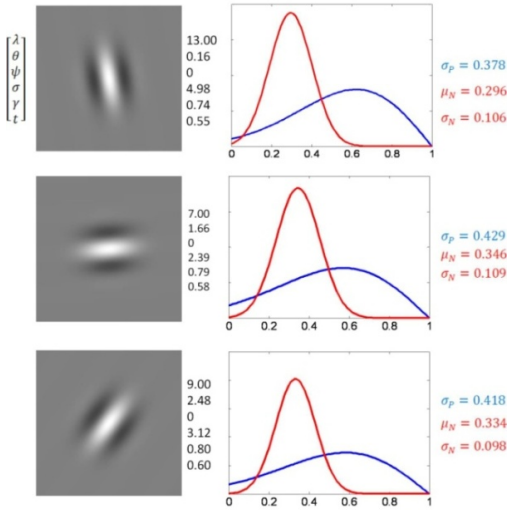

Figure 4. Three SGWs and corresponding characteristic distributions. The mirror reversed Rayleigh distribution on positive training samples is marked in blue and the Gaussian distribution on negative training samples is marked in red.

Background outliers such as bricks, window grids or some stripe texture are probably classified as text regions because they also generate similar SGW suitability measurements as text characters. To filter out the false positive detection, we employ the number of dominant colors and foreground connected components in an image window to make further classification. Color reduction based on [15] is performed to group the pixels with similar colors together, and the color corresponding to the most pixels is set as background color. In general, a true positive text window contains exactly two colors including background and foreground characters, and each character forms a connected component. In our algorithm, we define that text window after color reduction contains no less than 2 and no more than 4 colors, and the number of foreground connected components should be greater than 1 and smaller than 7. When a group of neighboring image windows had been classified as text windows, they would be merged into text regions as results of text detection.

IV. EXPERIMENTS

A. Datasets

We evaluate the proposed algorithm on two datasets. The first one is ICDAR 2003 Robust Reading Dataset. It contains 509 images in total, in which the first 258 images are used for training and the rest 251 images are used for testing. The image sizes range from 640×480 to 1600×1200. There are 2258 ground truth text regions in total. The second dataset is provided by [6]. There are 307 images containing 1981 ground truth text regions in total. The image sizes range from 1024×764 to 1024×1360. The natural scene images in this dataset present a more challenging background. The height of 901 ground truth text regions is less than 20 pixels.

B. Results and Discussions

We evaluate the performance of our algorithm by comparing the detected text regions with the ground truth text regions. We define "*precision*" as the ratio of area of the true positive extracted text regions to area of the detected regions, and "*recall*" as the ratio of area of the true positive extracted text regions to area of the ground truth regions. Here area means the number of pixels in the image region. "*f-measure*" is defined as the combination of precision and the recall by the harmonic mean.

On the Robust Reading Dataset, the testing images are used to evaluate the performance of our algorithm. Since adjacent character grouping in the process of layout analysis cannot cover text strings with less than three character members, we eliminate images whose ground truth text regions contain less than three text characters. Thus 236 testing images are chosen for performance evaluation. To keep consistent with higher stroke intensity than background in training samples, two rounds of text detections are performed for each testing image based on gray image and inverse gray image respectively. The better result is used to evaluate algorithm performance. The evaluation results are calculated from average measures of all testing images, *precision* 0.64, *recall* 0.76, and *f-measure* 0.68. By comparison with the state-of-the-art algorithms, our method achieves the best performance of recall and f-measure in this experiment. Some examples of detected text regions are shown in Fig. 5.

Figure 5. Some example results of text string detection on the Robust Reading Dataset. The detected regions are marked in cyan.

On the other dataset, text windows obtained from the ground truth regions of the first 150 images are added into the training set as positive samples. The text detection is then performed on the whole 307 images of this dataset. By using the same measures, we obtain precision 0.49, recall 0.60, and f-measure 0.42. Our results are close to precision 0.54, recall 0.42, and f-measure 0.47 presented in [6]. To improve the precision, a more robust model of SGW evaluations on negative samples should be developed to handle the complex background outliers that resemble text structure in the future, rather than a naive normal distribution. Fig. 6 presents some examples of detected text regions.

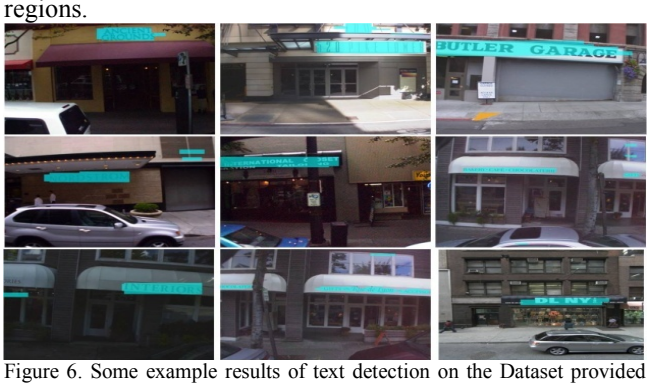

Figure 6. Some example results of text detection on the Dataset provided by [6]. The detected regions are marked in cyan.

C. Conclusion and future work

We have presented a novel algorithm to detect text regions in natural scene images. First, we use Gabor filter to describe stroke component and define a suitability measurement to model the confidence of Gabor filter description of strokes. Second, we carried out the statistical analysis on the stroke components of text from training sets to obtain SGWs, which are used as the universal description of principle stroke components. Third, characteristic distributions are established for each SGW by using the Rayleigh model to describe suitability statistics on positive training samples and the Gaussian model to describe suitability statistics on negative training samples. Image window classification is performed based on characteristic distributions of the principle SGWs. The experimental results demonstrated that our algorithm performed well on backgrounds and variant text patterns, and outperforms the state of the art algorithms for text extraction from natural scene images. Our SGW model demonstrates the statistical stationarity of the stroke components of text. In the future, we will develop more effective suitability measurements and more robust models to describe the SGW suitability statistics on the negative training samples. Furthermore, we will extend our algorithm to detect text with non-horizontal orientations or on deformed surfaces.

ACKNOWLEDGMENTS

This work was supported in part by NIH Grant 1R21EY-020990, NSF Grant IIS-0957016, and ARO Grant W911NF-09-1-0565.

REFERENCES

[1] J. Banerjee, A. M. Namboodiri and C. V. Jawahar, "Contextual restoration of severely degraded document images," *Proceedings of IEEE CVPR*, pp.517-524, 2009.

[2] S. Bhagavathy, J. Tesic, B. S. Manjunath, "On the Rayleigh nature of Gabor filter outputs," in *ICIP* 2003.

[3] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," In *CVPR*, vol. 2, pp. II-366 – II-373, 2004.

[4] X. Chen, J. Yang, J. Zhang and A. Waibel, "Automatic detection and recognition of signs from natural scenes," In *IEEE Transactions on image processing*, Vol. 13, No. 1, pp. 87-99, 2004.

[5] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Res. 20 (10): 847–56, 1980*.

[6] B. Epshtein, E. Ofek and Y. Wexler, "Detecting text in natural scenes with stroke width transform," In *CVPR*, pp. 2963-2970, 2010.

[7] S. Hu and M. Chen, "Adaptive Frechet kernel based support vector machine for text detection," In *ICASSP'05*, Vol. 5, pp. 365-368, 2005.

[8] A. K. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," In *Machine Vision and Applications*, Vol. 5, Issue 3, pp. 169-184, 1992.

[9] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model," *IEEE Trans on Image Processing*, Vol. 16, No. 8, pp. 2117-2128, 2007.

[10] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Transactions on PAMI*, Vol. 18, No. 10, 1996.

[11] J. Liang, D. DeMenthon and D. Doermann, "Geometric Rectification of Camera-captured Document Images," *IEEE Transactions on PAMI*, Vol. 30, No. 4, pp. 591-605, 2008.

[12] Q. Liu, C. Jung, and Y. Moon, "Text Segmentation based on Stroke Filter," *Proceedings of International Conference on Multimedia*, pp. 129-132, 2006.

[13] S. M. Lucas, A. Panaretos, L. Sosa, "A. Tang, S.Wong and R. Young. ICDAR 2003 Robust Reading Competitions," In *ICDAR*, 2003.

[14] S. M. Lucas, "ICDAR 2005 text locating competition results," *Proceedings of the ICDAR*, Vol. 1, pp 80–84, 2005.

[15] N. Nikolaou and N. Papamarkos, "Color Reduction for Complex Document Images. *International Journal of Imaging Systems and Technology*," Vol.19, pp.14-26, 2009.

[16] W. Pan, T. D. Bui and C. Y. Suen, "Text detection from scene images using sparse representation," in *ICPR*, 2008.

[17] W. Pan, C. Suen and T. Bui, "Script Identification Using Steerable Gabor Filters," in *Proc. of ICDAR*, 2005.

[18] T. Phan, P. Shivakumara and C. L. Tan, "A Laplacian Method for Video Text Detection," *In Proc. of ICDAR*, pp.66-70, 2009.

[19] Y. L. Qiao, M. Li, Z. M. Lu and S. H. Sun, "Gabor filter based text extraction from digital document images," In *IIH-MSP'06*, pp. 297-300, 2006.

[20] R. Sandler and M. Lindenbaum, "Optimizing Gabor Filter Design for Texture Edge Detection and Classification," In *Int. Journal Computer Vision 84: 308–324*, 2009.

[21] C. Yi and Y. Tian, "Text string detection from natural scene s by structure-based partition and grouping", In *IEEE Transactions on Image Processing*, *DOI: 10.1109/TIP.2011.2126586*, 2011.