Text Extraction in Digital News Video Using Morphology

Hyeran Byun[1], Inyoung Jang[1], and Yeongwoo Choi[2]

[1]Visual Information Processing Lab., Dept. of Computer Science, Yonsei University, 134, Shinchon-Dong, Seodaemun-Gu, Seoul, Korea, 120-749
{hrbyun, stefano}@cs.yonsei.ac.kr
[2]Image Processing Lab., Dept. of Computer Science, Sookmyung Women's University, Chungpa-Dong 2, Yongsan-Gu, Seoul, Korea, 140-742
ywchoi@sookmyung.ac.kr

Abstract. In this paper, a new method is presented to extract both superimposed and embedded scene texts in digital news videos. The algorithm is summarized in the following three steps : preprocessing, extracting candidate regions, and filtering candidate regions. For the first preprocessing step, a color image is converted into a gray-level image and a modified local adaptive thresholding is applied to the contrast-stretched image. In the second step, various morphological operations and Geo-correction method are applied to remove non-text components while retaining the text components. In the third filtering step, non-text components are removed based on the characteristics of each candidate component such as the number of pixels and the bounding box of each connected component Acceptable results have been obtained using the proposed method on 300 domestic news images with a recognition rate of 93.6%. Also, the proposed method gives good performance on the various kinds of images such as foreign news and film videos.

1 Introduction

In recent years the amount of digital video used has risen dramatically to keep pace with the increasing use of the internet and consequently an automated method is needed for indexing digital video databases. Up to now most of digital video indexing is done by human operators. This is an inefficient process due to the need of time and manpower for dealing with massive digital videos. Also, there is a room for making errors on account of the subjective decisions of the human operator. To avoid these inefficiencies and errors, automatic shot segmentation and text extraction have been studied [7,11,12,14]. However, since the automatic methods of shot segmentation alone have limitations for the complete digital video indexing, the research on the text extraction is also needed. Textual information, both superimposed and embedded scene texts, appearing in a digital video can be a crucial clue for helping the video indexing [1–5]. Also, there have been various approaches [1–5,7,11,12,15] for the correct extraction of textual information. In general, typical obstacles are variations in

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 341–352, 2002.

size and font, the orientation and positioning of the characters, different textures, unconstrained illumination, and irregular background and color gradients on the character stroke [5,6]. To overcome some of these difficulties, this paper proposes morphological operations and Geo-correction method to extract text regions.

2 Related Works

There are various approaches for extracting and recognizing texts on the image. If the text is in binary images, such as book pages, it can be segmented by identifying the foreground pixels in each horizontal line [8]. But, to extract text regions in complex video frames or scene images we need more advanced approaches. Various advanced methods have been proposed to extract texts from the complex images.

Zhong *et al*. [9] presented two methods for extracting texts from the complex color images, and combined the methods together. The first method used color image with color quantization and the connected components analysis on each quantized color plane. Heuristic filters for removing non-text components are developed and used. The second method used spatial variance on a gray-level image by assuming that the spatial variance in the background region is lower than that in the text regions. In their approach, the ascending and descending characters are not well detected.

Ohya *et al*. [10] presented a method to extract text in scene images. In their approach, several assumptions on the characters are used: they should be upright without slant or skew, distinctive between texts and background regions, and uniform in their gray values. Their method first segments image using a local thresholding method for detecting patterns of candidate characters. Then, the differences in the gray values between text and background regions are evaluated. Finally, the similarity to a set of character categories is measured, and component merging is performed by using a relational operation. They tested several kinds of images such as road signs, automobile license plates, and signboards of shops with various sizes, and gray-level under unconstrained illumination conditions. This method is not independent on text slant or tilt.

H. K. Kim [11] presented automatic text detection and location method for color video frames. In his paper, characters are assumed to lying on a horizontal way with uniform color and size. The algorithm first performs color segmentation by quantizing the image using the color histogram. The most dominant color is segmented by clustering colors. In color clustering, selecting color space and distance metrics are critical factors for the results. Then, heuristic filtering is applied to the candidate text regions. This method used too many ad-hoc thresholds.

M. A. Smith [12] proposed a method for extracting textual information from consecutive video frames. In this approach, the characters are assumed to lying on a horizontal line. The algorithm first extracts the high contrast regions using thresholding and vertical edge detection. Then, the non-text regions are removed and the broken regions are merged using a smoothing filter, and the candidate text regions are detected. Finally, the filtering is performed by considering the following criteria: the

pixel density of the candidate text region, the ratio of the minor to the major axis of the bounding box, the local intensity histogram. In this approach, texts with different contrast are not well extracted.

P. K. Kim [14] presented a text location method in complex color images. A local color quantization is applied for each color separately. The algorithm consists of four phases: converting the input color image into a 256 color image, contour following using a local color quantization, extracting a connected component, and filtering. An intermingled text regions with backgrounds, which may happen in the global color quantization, is excluded by using local color quantization. This method improved the detection rate of texts, but it requires plenty of processing time.

3 The Proposed Method

The proposed method is composed of three steps as shown in Fig 1: preprocessing (step 1), extracting candidate text regions (step 2), and filtering candidate text regions (step 3). In the preprocessing step, a color image is converted into a gray-level image and applies histogram stretching method to enhance the contrast of the image. Then, a modified local adaptive thresholding is applied to the contrast enhanced image. In the text extraction step, various image processing methods based on morphological operations are applied to remove non-text components while retaining the text components. In the final filtering step, the characteristics of each connected component are used.

In this paper, morphology operation is applied to the modified local adaptive thresholded image to emphasize the false positive component, which can be easily decided as a text while not a text region. Then, morphology operation and Geo-correction filtering which is proposed in this paper are applied to the modified local adaptive thresholded image to emphasize both text and text-like component. Text components, which we want to extract, are mainly remained and false positive components are mostly removed by means of obtaining difference image. Opening with 3x3 structuring element and (OpenClose + CloseOpen)/2 are used for extracting the false positive component, and (OpenClose + CloseOpen)/2 and Geo-correction are used for extracting both text and text-like components. In this paper, text region is assumed as comprising at least three consecutive characters and lies on a horizontal direction.

3.1 Characteristics of Text Appearing in News Video

Text may appear as a superimposed text or a scene text in digital video. The superimposed text is added to a video after finishing the video shooting in a post processing stage by artificially, while the scene text is recorded with the scene without

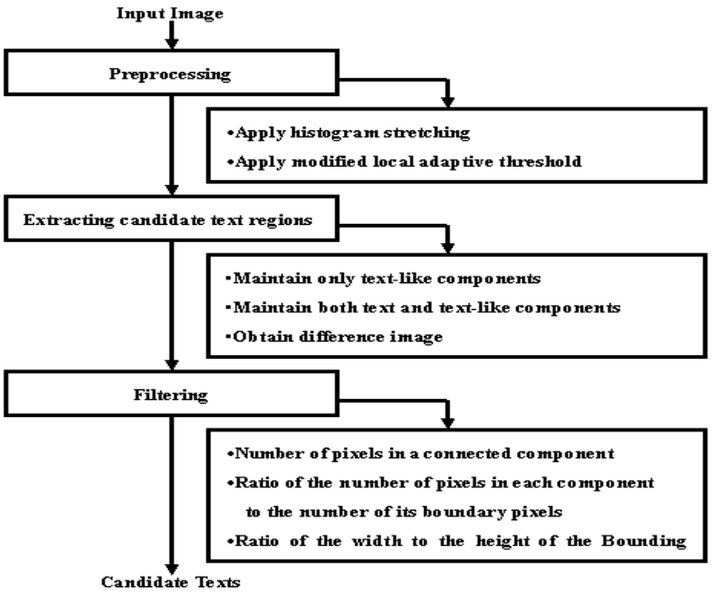

Fig. 1. Steps of the proposed method

any intention. Superimposed text extraction is more important than scene text extraction. The reason is that the superimposed text in video implies a mostly condensed and important content, whereas scene text usually appears without any intend. Thus, superimposed text is more appropriate for indexing and retrieval than scene text. However, text extraction in scenery image, needless to say scene text is the most important and very difficult to extract as much owing to infinite diversity of its appearance in direction, slant, occlusion and unconstrained illumination. Though appearance of superimposed text also has as much of the same difficulty as scene text, extraction of the superimposed text is less complicated than scene text since, the superimposed text has made for the purpose of reading and catching a viewer's eye easily. In this paper, extraction of the superimposed text has been focused and its features are described as follows [15]:

- Contrast between text as foreground and non-text as background is high because the superimposed text has made with intend for easy reading.
- Text is upright and monochrome in most cases.
- Each character piles up into a text line at a uniform interval with a horizontal direction.
- Text has the following restrictions in size: smaller than entire image size and bigger than a certain size as it can be seen.

3.2 Preprocessing

In text extraction, input images for preprocessing are color [10,12,13,15] or gray level images [11,14]. Zhong et al. [9] employed both the color and the gray-level images for their hybrid approach. The shortcoming on using the color image is that text regions and background can be merged [14]. Therefore, the proposed method uses the gray-level image for text extraction and will use the contrast between text and background and the shape of text. The RGB components of the input color image are converted into gray-level image.

A histogram stretching method is applied to the gray-level image to enhance the contrast of the input image. This is done to emphasize the brightness difference in the text region between text strokes and their background. A modified local adaptive thresholding is applied to the contrast enhanced image. The block size used is (Width/30)*(Height/30). The local threshold, T, is computed by considering mean, m, and standard deviation, σ, of the pixel values in each block as given in (1). User input variable, k, can be controlled by the types of video sources such as news, sports, cinemas, commercials, etc. The modification of our method to the original local adaptive thresholding is in setting the thresholded values: when the pixel value is less than T, it is set to zero, but when the value is greater than or equal to T, the gray value of the pixel is remained.

$$T = m + \sigma * k \tag{1}$$

3.3 Extracting Candidate Text Regions

3.3.1 Maintaining Only Text-Like Components

In this paper, text components are exactly the text, which is appearing on an image, and text-like components are something, which can be easily presumed as a text even not a text. In this subsection, morphological opening and (OpenClose + CloseOpen)/2 operations are consecutively applied to the semi-thresholded image. First, the semi-thresholded image is binarized by converting non-zero values to 255 to apply binary morphological operations. 3x3 structuring element is used for opening. This 3x3 structuring element is selected based on analyzing the character width on a video image. In the opening process, the erosion is applied to remove noises and text-like components, and then the dilation is performed to recover the remaining objects that can be damaged during the erosion operation, as shown in Fig 4 (c) and (d). Then the gray value of semi-thresholded image is copied into the pixels with the gray value of 255 in binary image which is the result of the opening operation. (OpenClose + CloseOpen)/2 operation is applied to the above result using 1x5 structuring element as shown in Fig 2. (OpenClose + CloseOpen)/2 operation is a gray level morphology operation. Opening and closing operations are consecutively applied to the input image and closing and opening operations are consecutively applied to the input image apart from the former, after that an average is calculated between these two images.

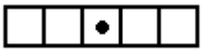

Fig. 2. Structuring element for OpenClose or CloseOpen

Erosion and dilation used in (OpenClose + CloseOpen)/2 operation to the gray level image in this paper are shown in Fig 3. In the erosion, subtract pixel values in structuring element from pixel values in input image. After substraction, pixel values smaller than the central pixel value are remained as it is and bigger than the central pixel value are substituted as the central pixel value as shown in Fig 3-(a).

In the dilation, the pixel value bigger than the central pixel value are remained as it is and smaller than the central pixel value are substituted as the central pixel value as shown in Fig 3-(b) after the summation of pixel values in input image and pixel values in structuring element. Therefore,the image will be more brightened after applying dilation and more darkened after applying erosion in the gray level image.

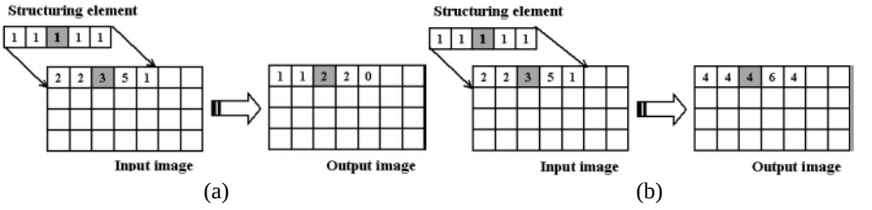

Fig. 3. Gray level morphology operation used in this paper, (a) Erosion in gray-level morphology, (b) Dilation in gray-level morphology

The (OpenClose + CloseOpen)/2 operation based on the above morphological operations will reduce the noise in the background by OpenClose and fill out the holes in the remaining objects by CloseOpen operation [13]. This operation is performed only in a horizontal direction using the 1x5 structuring element to detect the horizontal text lines. The results of each morphological operations are shown in Fig 4.

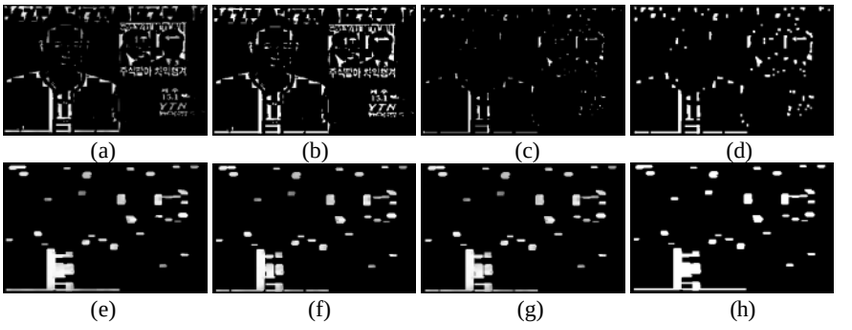

Fig. 4. Results for each morphological operation, (a) Semi-thresholded image, (b) Binarized image, (c) Erosion applied, (d) Dilation applied (e) OpenClose applied, (f) CloseOpen applied, (g) (OpenClose+CloseOpen)/2, (h) Binarization of (g)

3.3.2 Maintaining Both Text and Text-Like Components

In this step, (OpenClose + CloseOpen)/2 and Geo-correction method are applied. First, (OpenClose + CloseOpen)/2 operations are applied to remove noise in the background and to fill out the holes in the object. Then, the result image is binarized by setting a non-zero gray value to 255 and then the Geo-correction filtering is performed (refer to Fig 6-(e)). The Geo-correction shown in Fig 5 is needed for further recovering the text candidate components by connecting the separated components that can be resulted by the erosion operation in the previous morphological operations. It is performed along the horizontal and vertical directions. In the proposed method, the threshold for the vertical direction is set smaller than the threshold for the horizontal direction, since most of texts are lying on horizontal direction in news image. The threshold value 20 for the horizontal and the value 5 for the vertical direction chosen by experiments are used.

The Geo-correction filtering fills and connects the intermediate pixels by 255 as follows. Each pixel is scanned from left to right and top to bottom. If the pixel value is 255, begin scanning the consecutive pixels continuously until zero appears, and then count the number of zeros until reaching out pixel value of the next 255. If the number of pixel values having zero is below the given threshold, these pixels are converted into 255 in order to connect two lines. And if the number is above the threshold, maintain the pixel value as it is.

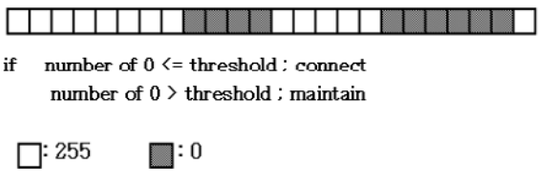

Fig. 5. Applying Geo-correction filtering

3.3.3 Obtaining Difference Image

The binary image obtained by extracting false positive text components using morphology has mostly non-text components, and the binary image obtained by maintaining both text and text-like components using morphology has both the text candidates and the non-text components. Text components are mainly remained and false positive components are mostly removed as shown in Fig 6-(f) by means of obtaining difference between these two images. We set negative pixel values to 0.

3.4 Filtering Text Candidates

The non-text or the noise components are needed to be removed. A connected component labeling is performed first, and then the filtering is applied. Three features are used for filtering and the proper threshold values in each filtering are selected by the experiment. Since these values are measured as ratios, they are not affected by the

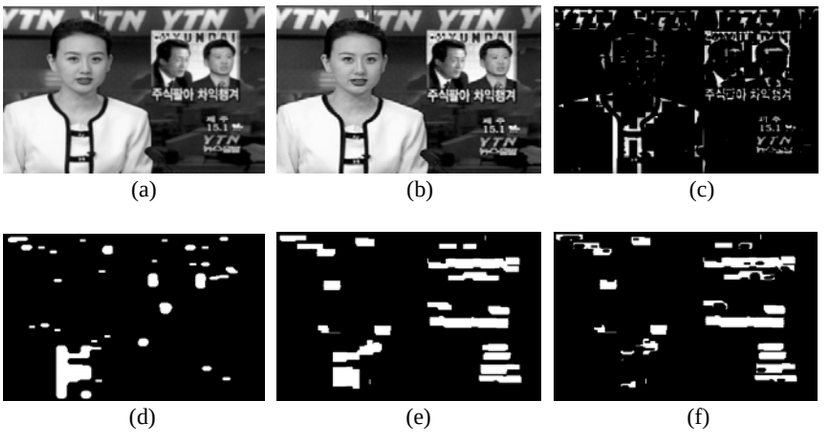

Fig. 6. Snapshots of the preprocessing and text region extractions, (a) Input image, (b) Contrast stretched result, (c) Semi-thresholded image, (d) Sub step 1, (e) Sub step 2, (f) Subtracting (d) from (e)

entire image size. First, the number of pixels*(NPi)* in each connected component*(Ci)* is considered. When the number of pixels in each connected component is too small as shown in equation (2), which is considered as non-text region, this component is removed. Since text region is assumed as comprising at least three consecutive characters in this paper, the proposed method uses the experimental results about the average character width appearing in news video. The average pixel ratio of three consecutive characters are lower than 4% in the entire image pixel.

$$\text{If } NPi \leq 0.04 \text{ then } Ci \text{ is removed} \qquad (2)$$

$$\text{else } Ci \text{ is retained}$$

Second, the ratio between the number of pixels in each component and the number of its boundary pixels*(NBi*) is used. This is based on the experimental result that the ratio between the number of pixels in each component and the number of its boundary pixels below some ration is assumed as text component. The ratio 0.23 is obtained by the experiment and the component with the ratio less than 0.23 is removed.

$$\text{If } NBi \leq 0.23 \text{ then } Ci \text{ is removed} \qquad (3)$$

$$\text{else } Ci \text{ is retained}$$

Third, the ratio between the width and the height of the bounding box of each component is used. If this ratio is less then the given threshold, the component will be removed. This is based on the assumption that the text line is assumed as comprising at least three consecutive characters and most of characters appearing in news video lies on a horizontal direction. Thus, the ratio 0.66 is obtained after the experiment about width and height of character appearing in news video.

$$\text{if } Height/Width \leq 0.66 \text{ then } Ci \text{ is removed} \quad (4)$$

$$\text{else } Ci \text{ is retained}$$

4 Experimental Results

The proposed algorithm has been implemented using Microsoft Visual C++ 6.0 on PC with 866MHz Pentium processor. The proposed algorithm is evaluated using 300 color images of Korean news clips (MBC, SBS, KBS, YTN) and 100 color images of English news clips (CNN, BBC, Bloomberg). Also, 100 film video clips, which include subtitles, and 50 commercial TV clips are evaluated to make a comparison among the different kinds of video. The image size is 320*210. Each image data is randomly captured on the consecutive frames. Three performance criteria are used to evaluate the results: correct extraction rate, practical extraction rate, and error rate. The practical extraction rate is a rate of finding text regions within a permitted tolerance of region judged by human. The error rate is defined as a rate of finding non-text components as texts and missing texts together. The extracted text line is counted as one text. These three criteria are stated as follows.

$$\text{Correct extraction rate} = Nct / Tn \quad (5)$$

$$\text{Practical extraction rate} = (Nct + Npt) / Tnt$$

$$\text{Error rate} = (Nnr + Nmt) / Tnt$$

Where, *Tnt*: the total number of texts in test images

Nct: the number of texts correctly extracted

Npt: the number of texts extracted in a permitted tolerance

Nnr: the number of extracted non-text regions

Nmt: the number of missing texts

(a) (b) (c) (d)

Fig. 7. Examples of three performance criteria, (a) Correct extraction, (b) Practical extraction, (c) Finding non-texts as text, (d) Missing texts

Fig 7-(a) shows the example of correct extraction, which is the case of finding all text line appearing in an image. Fig 7-(b) shows the example of practical extraction,

which is the case of finding text line and keyword for video indexing even though there are some partial errors. The partial errors are due to not recovering sufficiently while applying morphology operations. Figure 7-(c) shows the example of finding non-text component as text. This error is caused by high contrast between foreground and background, and these high contrast components have not removed during the modified local adaptive threshold phase. Figure 7-(d) shows the example of missing texts. This error is caused by low contrast between foreground and background, and these low contrast components have removed during the modified local adaptive threshold phase. In the proposed method, the main reasons of the error are due to the tiny text in the image, the low contrast between the text and the background, and also due to the large horizontal distance between the same text components.

Table 1. Recognition rate of test data

	Nct		Npt		Nnr + Nmt	
	Number	%	Number	%	Number	%
Korean news	**533**	**77.5%**	**533+110**	**93.6%**	**46+42**	**12.8%**
English news	**291**	**76.5%**	**291+43**	**88.1%**	**6+45**	**13.4%**
Film video	**132**	**73.7%**	**132+43**	**97.7%**	**2+4**	**3.3%**
Commercial video	**66**	**49.3%**	**66+29**	**70.9%**	**8+39**	**35.1%**

The reason why the proposed method classifies the test data as Korean and English news in the evaluation phase is that the height of each character in English news video can be different while the height in Korean news video is almost the same. This characteristic may cause the error during the modified local adaptive threshold and also has a close relation with the extraction rate. As we can see the above test result, the extraction rate of the Korean news was a little bit higher than the English news. In the film video images, the practical extraction rate was higher and the correct extraction rate was lower than the news video images. This is because of the complicated backgrounds in the film video images even though those images have a regular character size than the news video images. In the commercial video images, the extraction rate was not better than the news video images, since the commercial video images easily contain the texts with variations in size and font, skewed, different textures, unconstrained illumination, *etc*. The commercial video images contain many of the scene texts and thus the modified local adaptive threshold has not segmented the image well. Consequently, the proposed method gives higher extraction rate on the superimposed texts in the news videos, but it has limitations on finding the scene texts with complex backgrounds.

Fig. 8. Text extraction results

5 Conclusions

A new method has proposed for extracting text specially well for the superimposed texts from digital video news images. In the proposed method, several morphological operations are used: eliminating text-like components by applying erosion, dilation, and (OpenClose + CloseOpen)/2 operations, maintaining text components using (OpenClose + CloseOpen)/2 and Geo-correction operations, and subtracting two result images. The OpenClose, CloseOpen and their combined operations can reduce noises and remove holes in each connected component. The proposed Geo-correction method is also efficient for conserving and compensating the candidate text components that can be damaged by the morphological operations. The experimental results show that the proposed method shows good performance in extracting texts in news video with the correct extraction rate of 77.5%, the practical extraction rate of 93.6% and the error rate of 12.8%. The proposed method also has tested with movies and commercial videos by adjusting the structuring elements of the morphological operations to the width of the character, and the initial results are very promising. We need to develop this research to modify and refine for the extraction of the scene texts with various kinds of images and video frames.

Acknowlegdements. This research was supported as a Brain Neuroinformatics Research Program sponsored by Korean Ministry of Science and Technology (M1-0107-00-0009).

References

1. Jae-Chang Shim, Chitra Dorai, and Ruud Bolle, Automatic Text Extraction from Video for Content-Based Annotation and Retrieval, *Proceedings of Fourteenth International Conference on Pattern Recognition*, Vol. 1, pp. 618–620 , 1998.
2. Anil K. Jain and Bin Yu, Automatic text location in images and video frames, *Pattern Recognition*, Vol. 31, No. 12, pp. 2055–2076, 1998.
3. H.Kuwano, Y.Taniguchi, H.Arai, M.Mori, S.Kuraka-ke, and H.Kojima, Telop-on-demand: video structuring and retrieval based on text recognition, *IEEE International Conference on Multimedia and Expo*, Vol. 2, pp. 759-762, 2000.
4. U. Gargi, S. Antani, and R. Kasturi, Indexing text events in digital video databases, *Proceedings of Fourteenth International Conference on Pattern Recognition*, Vol. 1, pp. 916–918, 1998.
5. Sameer Antani, Ullas Gargi, David Crandall, Tarak Gandhi, and Rangachar Kasturi, Extraction of Text in Video, *Dept. of Computer. Science and Eng., Pennsylvania State Univ., Technical Report*, CSE-99-016, 1999.
6. S. Messelodi and C.M. Modena, Automatic identification and skew estimation of text lines in real scene images, *Pattern Recognition*, Vol. 32, pp. 791–810, 1999.
7. S. Antani, D. Crandall, and R. Kasturi , Robust extraction of text in video, *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 1, pp. 831–834, 2000.
8. Y. Lu, Machine printed character segmentation-An overview, *Pattern Recognition*, Vol. 28, pp. 67–80, 1995.
9. Y. Zhong, K. Karu, and A. K. Jain, Locating text in complex color images, *Pattern Recognition*, Vol. 28, pp. 1523–1535, 1995.
10. J. Ohya, A. Shio, and S. Akamatsu, Recognizing characters in scene images, *IEEE Trans on Pattern Analysis and Machine Intelligence*. PAMI-16, pp.214–220, 1994.
11. H. K. Kim, Efficient automatic text location method and content-based indexing and structuring of video database, *J. Visual Commun. Image Representation*, Vol. 7, pp. 336–344, 1996.
12. M. A. Smith and T. Kanade, Video skimming for quick browsing base on audio and image characterization, *Technical Report CMU-CS-95-186, Carnegie Mellon University*, July 1995.
13. J. Serra, Image Analysis and Mathematical Morphology. New York: Academic, 1982.
14. Pyeoung-Kee Kim, Automatic Text Location in Complex Color Images using Local Color Quantization, *TENCON 99. Proceedings of the IEEE Region 10 Conference*, Vol. 1, pp. 629–632, 1999.
15. R. Lienhart and F. Stuber, Automatic text recognition in digital videos, *SPIE Image and Video Processing IV*, pp 2666–2669, 1996