DOI: 10.3745/KIPSTD.2010.17D.1.59

# 카메라 영상 위에서의 문자 영역 추출 및 OCR

신 현 경[†]

## 요 약

기존의 OCR 엔진은 보정된 환경에서 읽혀진 서류 영상에 맞게 설계되어있다. 스마트 폰을 비롯한 검정 화면 거리가 보정되지 않은 기기에서 읽혀진 영상에서는 삼차원 원근 투시에 의한 찌그러짐 또는 곡면상에서의 찌그러짐 등이 핵심적인 문제점들로 여겨진다. 휴대용 단말기에서 읽혀진 영상들에서의 OCR 기능에 대한 요구가 증가일로에 있는 시점에서, 본 논문에서는 문제점들을 세 가지로 구분하고 - 회전에 무관한 문자 영역 추출, 폰트 등의 크기에 무관한 문자 선 영역 추출, 3차원 매핑 이론 - 이를 해결하기위한 방법을 제시하였다. 이러한 방법론을 통합하여 카메라 영상 위에서의 OCR을 개발하였다.

**키워드 : 문자 인식, 문자 영역 탐색, 문자 열 추출, DCT, Thresholding, 비 선형 매핑, 카메라 영상 문자 인식, 직각화 함수**

# Text Region Extraction and OCR on Camera Based Images

Shin Hyun Kyung[†]

## ABSTRACT

Traditional OCR engines are designed to the scanned documents in calibrated environment. Three dimensional perspective distortion and smooth distortion in images are critical problems caused by un-calibrated devices, e.g. image from smart phones. To meetthe growing demand of character recognition of texts embedded in the photos acquired from the non-calibrated hand-held devices, we address the problem in three categorical aspects: rotational invariant method of text region extraction, scale invariant method of text line segmentation, and three dimensional perspective mapping. With the integration of the methods, we developed an OCR for camera-captured images.

**Keywords : OCR, Text Region Detection, Text Line Segmentation, DCT, Thresholding, Nonlinear Mapping, Camera-based 3D OCR, Rectification Mapping**

## 1. Introduction

Range of applications of text recognition and TTS (text-to-speech) with the hand-held devices is vast, e.g., a cell phone can read the labels on the medicine bottles for the blind [1], a GPS device can read the road signs for the car drivers, etc. However, traditional OCR engines are designed to the calibrated devices and are not accurate enough for the images acquired by the un-calibrated digital devices such as smart phones [2]. Poor performance of standard OCR on camera based image is caused by blurring from mal-adjusted focal length and by the circumstance of text occurrence on perspective plane or curved surfaces.

In the industry, while scanner based OCR applications are shifting to new platforms such as XEROX PC-cam OCR suite [3] and DigitalDesk [4], the current trend is focused on pre-processing the input images and re-using the existing OCR engines rather than developing new recognition technologies. The pre-processing is generally consisted of the three steps: text region segmentation, rectification or de-warping [5], and normalization including deblurring and upscaling.

Text segmentation is a task to identify topologically connected polygonal regions containing text characters in the images. Text segmentation techniques engage four stages: text detection; text localization; text line extraction; and text region grouping. Rectification or dewarping of text regions extracted from camera-captured image is required. For the better performance of OCR, the text contents from the rectified regions are investigated to be normalized through resampling in terms of its font sizes.

※ 본 연구는 2009년 경원대학교 교내연구비 지원을 받아 시행된 연구임.

† 종신회원 : 경원대학교 수학정보학과 조교수

논문접수 : 2009년 11월 6일
수 정 일 : 1차 2009년 12월 29일
심사완료 : 2009년 12월 29일

The structure of this paper is organized as follows. In section 2, the previous related research works are summarized. In section 3, the underlying algorithms employed in this paper are described. In section 4, the experimental results are presented. Section 5 concludes the paper with discussions.

## 2. Related Works:

Progress of camera-based OCR is summarized by Doerman et al. [2], where they address issues of low resolution, blur, and perspective distortion from camera-based image. In the paper, discussions are concentrated on document image.

Text detection is a process build edge map from continuous color scale input image. Due to combined effect of low resolution and blur, traditional text detection is not usually efficient enough. For fast and robust edge map transformation, DCT has been employed by many researchers. Chaddha et al. [6] studied lists of DCT coefficient used for best coefficient search and proposed a list of eight combinations. Zhong et al. [7] defined DCT energy functions using horizontal and vertical spatial intensity. However, they assumed the texts are aligned front-parallel occurrence. In order for rapid derivation of binary edge map from the video, Lee et al. [8] access to compressed domain of I-frame to obtain DCT of MPEG directly. An edge extraction algorithm based on the correlation between AC coefficients finds text block region.

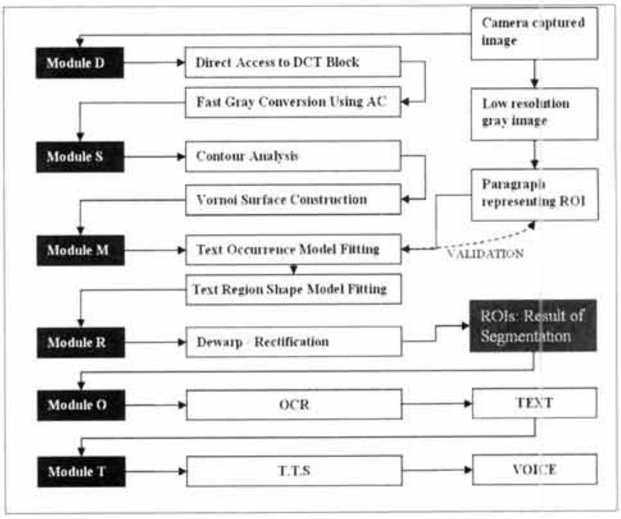
Text localization, a process of grouping the edges acquired from text detection into texts, is divided into texture based and connected component based methods. Connected component based method is straightforward. In texture based approaches are Gabor filter [9], stroke filter [10], pixel value variance [11], and multi-resolution feature [12].

Text line segmentation processes the localized text into line and word. Traditionally, challenges of line segmentation are characterized by three issues: line proximity, line fluctuation, and fragmentation [13]. Line segmentation methods can be divided by five categories listed as follows: Projection based methods [14-16] commonly used for printed document segmentation; Run length smoothing algorithm [17] for separation of touching text lines; Grouping by joining nearest neighbors [18] by edge map of variance or level set[19]; Hough transform based; stochastic methods [20, 21].

To utilize exiting OCR engine geometric rectification is necessary. Distortion of text regions in camera-captured image is primarily caused by perspective projection and non-planar shape. In case of planar surface, rectification process only involves 3D perspective rotation which requires vanishing point analysis; however rectification can be achieved 2D data [22-24]. In case of text occurred on curved surface, estimating 3D information is critical to find text lines. 3D method is used for opened book [25] while 2D warping method is used [26, 27] based on local linearity. 2D method usually has problem in character level.

## 3. System Design and Underlying Algorithms

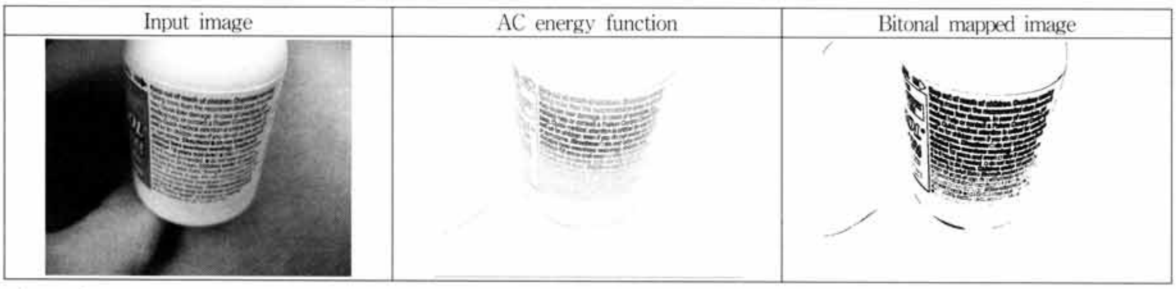
Our system is consisted of six pipe-lined modules as seen in (Fig. 1). Module D represents binary conversion module by accessing block DCT embedded in JPEG, Module S represents text paragraph region detection module by contour analysis, Module M represents shape model fitting module, Module R represents rectification module, Module O represents OCR module, Module T represents TTS module. More details are described throughout the following sub-sections 3.1 - 3.6.


Camera captured image
Module D
Direct Access to DCT Block
Fast Gray Conversion Using AC
Low resolution gray image
Module S
Contour Analysis
Vornoi Surface Construction
Paragraph representing ROI
Module M
Text Occurrence Model Fitting
VALIDATION
Text Region Shape Model Fitting
Module R
Dewarp / Rectification
ROIs: Result of Segmentation
Module O
OCR
TEXT
Module T
T.T.S
VOICE

(Fig. 1) OCR system (for camera captured image) consisted of the pipelined modules

### 3.1 Module D: binary conversion module

For the bitonal conversion, our method is in two stages: gray conversion using DCT and thresholding. We use DCT information of YCC channels. Suppose a[n,m] indicate element of DCT, we define AC energy as $|a[0,1] + a[1,0] + a[1,1]|$. After accessing three values of AC energies from each of YCC channels, we take the mean of three energies to represent pixel value of the corresponding DCT block. For gray scale representation,



(Fig. 2) processes in Module B (binarization). Input color image is transformed to gray scale of AC energy (AC energy function). The gray image is thresholded to bitonal mapped image

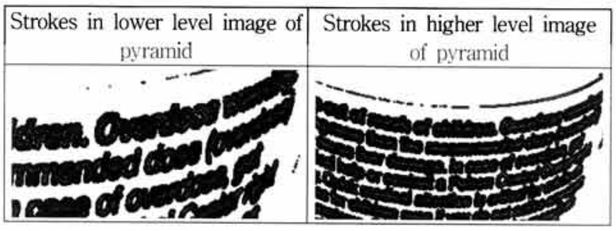
the pixel values are scaled into [0, 255]. (Fig. 2) illustrates the process: input color image is transferred to grayscale (the middle column). The gray scale image obtained from the AC is converted to bitonal image using simple thresholding with a threshold value of 215 (the right most column).

### 3.2 Module S: text region detection module

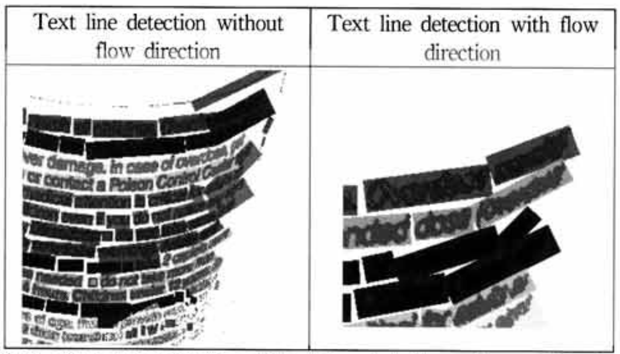
Text regions are detected through contour retrieval method on the image pyramids. For the contour retireval methods we use OpenCV API from Intel.

The bitonal conversion described above is applied to the image pyramid of three levels. Considering smoothing effect of low resolution of higher levels in pyramid, we use different threshold values of 205 for the third level. For each gaussian level of the pyramid, we apply the distance map transformation on foreground (black) pixels with Euclidean distance. Instead of finding full medial axis, we find the local maximal points in the resulting label image transformed by distance map. The list of local maximal points is a subset of medial axis. We connect two local maximal points if they are located within a 5x5 window. We call the connected line fragment as 'stroke'. The strokes shaped in different level of pyramid are illustrated in (Fig. 3). At the left panel red pixels represent strokes at level zero while at the right panel strokes at level three. At the higher level the strokes show trace of text lines more clearly than at the lower level.

Major axis of a stroke can be estimated by finding the minimum rectangle with Sklansky's algorithm. The minimum rectangle is represented by the offset, the size and rotated angle. We take the median of the rotated angles as text flow direction. In order for separation of merged text lines as can be seen in the top-right of (Fig. 3) and of (Fig. 4), we apply directional morphology along with the direction estimated by the average (i.e. median) rotated angle of major axis of strokes. The effect of directional morphology is illustrated in (Fig. 4). At the left panel in (Fig. 4) text lines are found in which the same color indicates same line. At the right top corner two words at the first line and the second are stuck together, while at the right panel two words are separated after the morphology operation.


Strokes in lower level image of pyramid
Strokes in higher level image of pyramid

(Fig. 3) Demonstration of 'stroke' retrieval. Strokes in higher level image (right) show indication of text flow direction

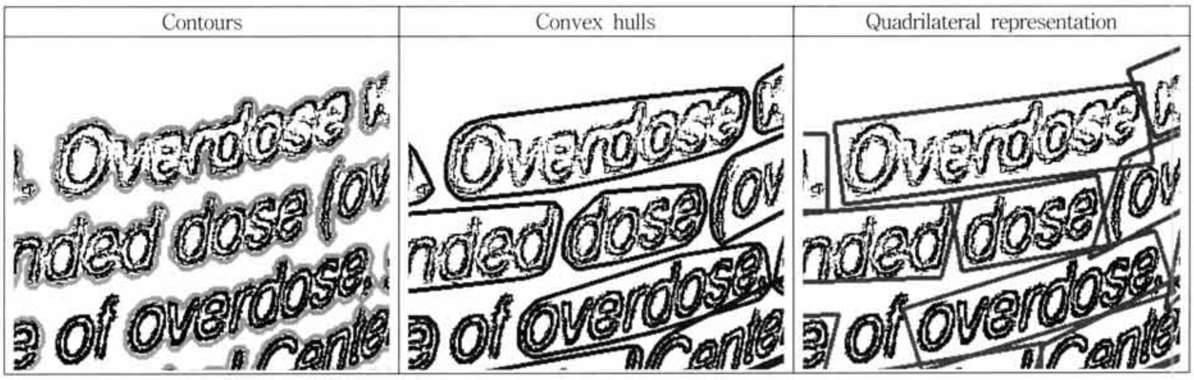

Text line detection without flow direction
Text line detection with flow direction

(Fig. 4) A demonstration of text line detection with or without information of flow direction

### 3.3 Module M: Text region shape model fitting module

As a generative modeling approach, in this paper we define text region model as combination of text occurrence model and text shape model. Text occurrence model has parameters of rotated angle for x-, y- and z-axis, respectively. Text shape model has parameter of shape list: rectangle, curved, free-form.

Text region (paragraph) is conceptually defined as linked-list of text lines. As can be seen in (Fig. 4), text line is linked list of word-level object so that a line can be easily determined either linear or locally piecewise linear (i.e., curved). In order for estimating direction of



(Fig. 5) A demonstration of contour, convex hull, and quadrilateral representation of text words.

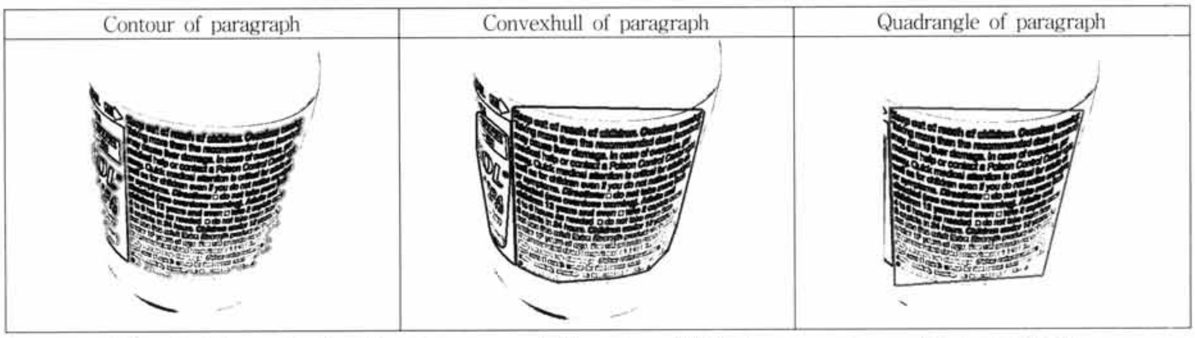
words consisting of line, we retrieve a contour first and construct its convex hull using Sklansky's algorithm. Based on the convex hull, we construct the minimum area rectangle and a quadrilateral. By investigating the angles of minimum rectangle, we can fit the shape model. Quadrilateral representation is for estimating the parameter of text occurrence model including the vanishing point. (Fig. 5) demonstrates contour, convex hull and quadrilateral representation for text words, respectively.

### 3.4 Module R: Rectification module

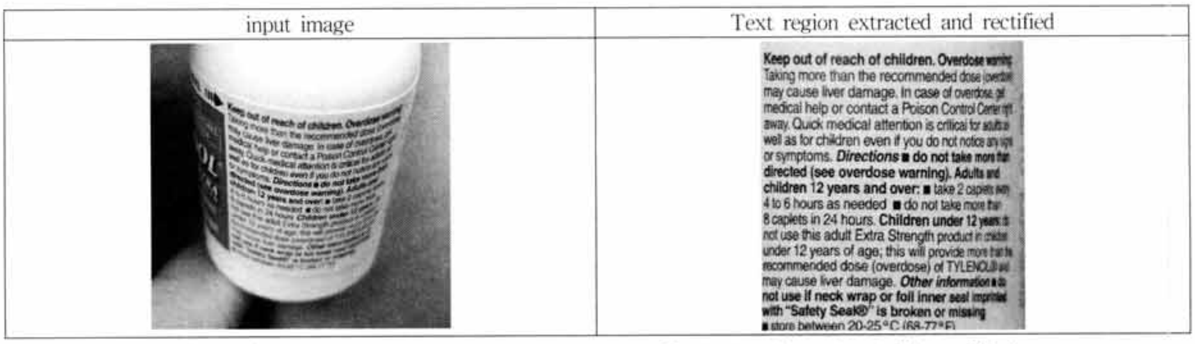
3D perspective matrix is created using the oriented quadrilateral, i.e., a map from the oriented quadrilateral to the destination rectangle. We use OpenCV API for 3D rectification mapping.

Once text lines are found as can be seen in (Fig. 4), text direction is also achieved by it. Based on the text direction, a text paragraph is constructed by gathering parallel text lines. In (Fig. 6) collection of lines is presented. At leftmost panel contour of text paragraph, at middle convex hull, and at rightmost panel quadrilateral representation. In order to apply 3D perspective map requiring four corner points, the convex hulls are further approximated to form quadrilaterals. Additionally, among the four sides of the resulting quadrilateral, one of the sides should be selected as the top-side. Text flow direction is re-used to find the top-side.

Quadrilateral representation of text paragraph is used to three-dimensional perspective transformation or to coon's patch transformation if the region is classified as linear or curved, respectively. (Fig. 7) demonstrates a result of rectification process.



(Fig. 6) A demonstration of contour, convex hull, and quadrilateral representation of text paragraphs



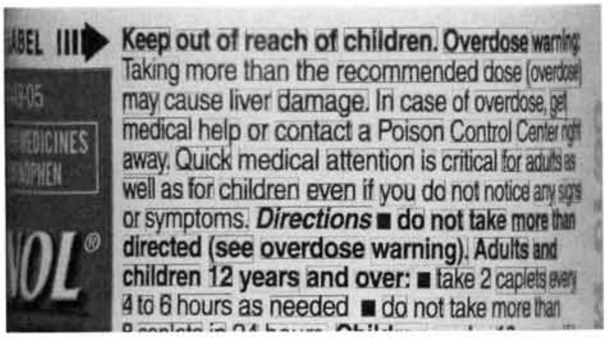
Keep out of reach of children. Overdose warning
Taking more than the recommended dose (overdose)
may cause liver damage. In case of overdose, get
medical help or contact a Poison Control Center right
away. Quick medical attention is critical for adults as
well as for children even if you do not notice any signs
or symptoms. *Directions* ■ do not take more than
directed (see overdose warning). Adults and
children 12 years and over: ■ take 2 caplets every
4 to 6 hours as needed ■ do not take more than
8 caplets in 24 hours. Children under 12 years: do
not use this adult Extra Strength product in children
under 12 years of age; this will provide more than the
recommended dose (overdose) of TYLENOL® and
may cause liver damage. *Other information* ■ do
not use if neck wrap or foil inner seal imprinted
with "Safety Seal®" is broken or missing
■ store between 20-25°C (68-77°F)

(Fig. 7) a demonstration of rectification process. The text region extracted is rectified

### 3.5 Module O: OCR module

For OCR module we use API's of SDK from Nuance and open Tesseract from Google.

For the current project, we invoke the latest version of Nuance engine, named OmniPage 16.2. (Fig. 8) shows an output from running the OCR where the input is the rectified regions as seen at the right panel in (Fig. 7). For the purpose of visualization, the output shows the red bounding boxes indicating the text words identified by the engine.



Keep out of reach of children. Overdose warning: Taking more than the recommended dose (overdose) may cause liver damage. In case of overdose, get medical help or contact a Poison Control Center right away. Quick medical attention is critical for adults as well as for children even if you do not notice any signs or symptoms. **Directions** ■ do not take more than directed (see overdose warning). Adults and children 12 years and over: ■ take 2 caplets every 4 to 6 hours as needed ■ do not take more than

(Fig. 8) an out put of OCR. Red bounding boxes indicate text words identified by OCR engine

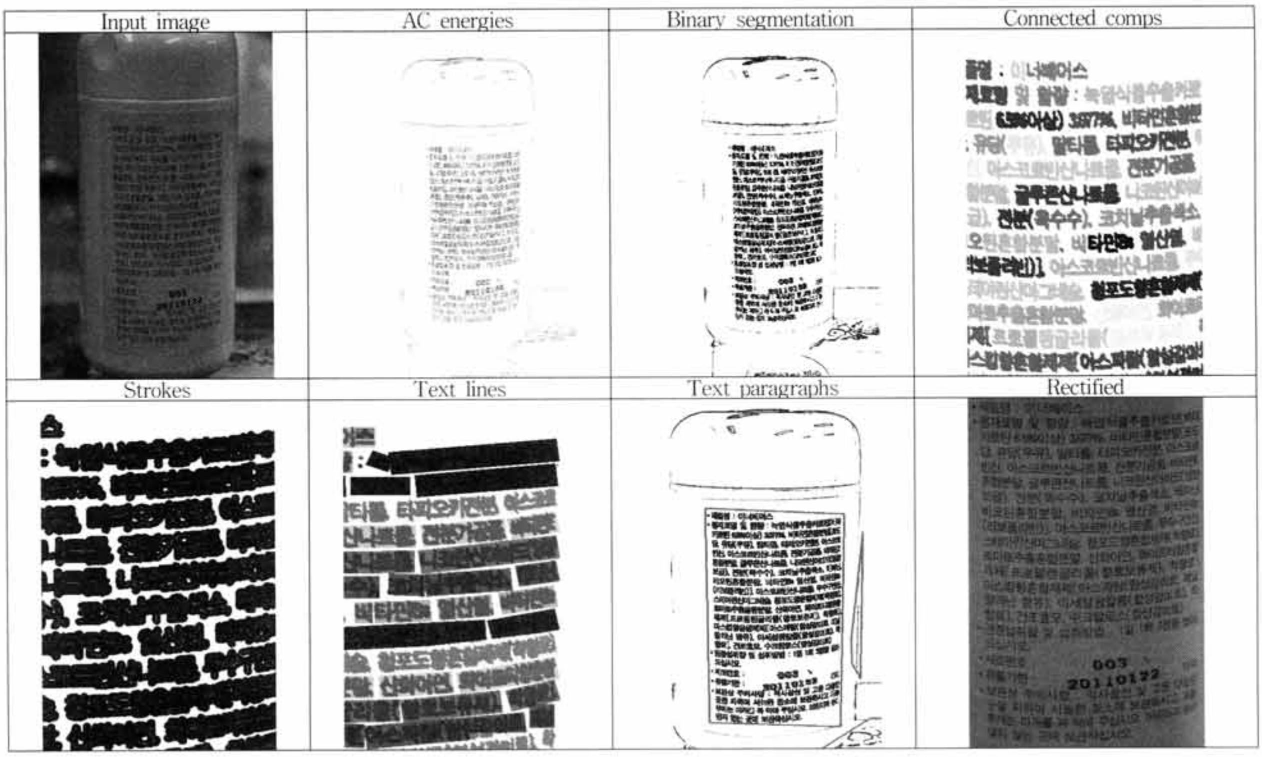
### 3.6 T.T.S module

We employ the Microsoft Text-To-Speech API which is available in any desktop PC with Windows XP for speech conversion process.

## 4. Experimental Results

The critical problems in performing OCR for camera-based image are mainly categorized into the two groups. The one is the distortion by the three dimensional perspective map and the other is the curved surface transformation. In this section we select two sample images corresponding to the group, which described as below. We performed OCR with many of the images matched with the categories. The result of OCR is briefed as follows: OCR performance is very limited for the curved texts and OCR engines could not catch any texts for the perspective distortion especially when the free parameters of the distortion is two.

In (Fig. 9), the sample image has the texts occurred in curved surface. Without rectification, current OCR engines fail to recognize the texts in the original image. In the figure, we show the intermediate results from the pipelined stages consisting of the whole rectification process. From the input image (the leftmost at the first row), AC energies are estimated (the second column at the first row), the energy values are thresholded (the third column at the first row) into binary image, a result of connected components retrieval process is shown (the rightmost column at the first row), the stroke features are extracted (the leftmost at the second row) from the connected components, the stroke information helps to find text line



Input image

AC energies

Binary segmentation

Connected comps

Strokes

Text lines

Text paragraphs

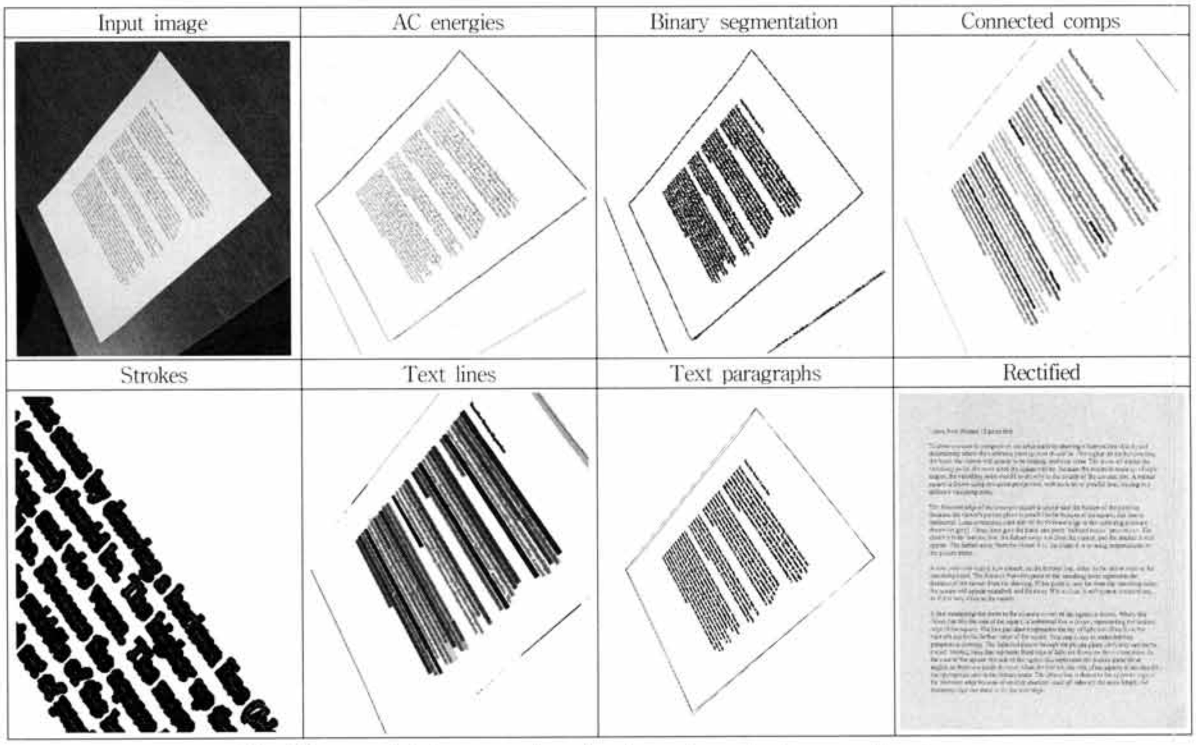
Rectified

(Fig. 9) case of texts occurred on the curved surface structure (the second column at the second row), text paragraph formation (the third column at the second row) is achieved by linking of text lines, rectification result by Coon's patch method (the rightmost at the second row).

In (Fig. 10), the sample image has the texts occurred in three dimensional perspective distortion. Without rectification, this type of input images also fail current OCR engines. As the same as the curved case, we show the intermediate results. From the input image (the leftmost at the first row), AC energies are estimated (the second column at the first row), the energy values are thresholded (the third column at the first row) into binary image, a result of connected components retrieval process is shown (the rightmost column at the first row), the stroke features are extracted (the leftmost at the second row) from the connected components, the stroke information helps to find text line structure (the second column at the second row), text paragraph formation (red box in the third column at the second row) is achieved by linking of text lines, rectification result by perspective map (the rightmost at the second row).

There are no known public database for three dimensional OCR. We created database ourselves consisting of 100 street images, 100 document images captured by cell-phone with various orientation and perspective angles (see Fig. 10), 100 multiple plane images (such as serial boxes), 100 smooth surface images (See Fig. 9).

In order for accessing OCR performance, we build a word dictionary containing about 190,000 words. The texts obtained from images by OCR are matched with the dictionary to count the matched words. The results are presented in Table 1. In the table, 'Count of words from OCR' represents the count of words recognized from OCR and 'matched words' are the words listed in the dictionary. 'Seg & Rect' represents segmentation and rectification processed images while 'Original' indicates non-preprocessed images.



(Fig. 10) case of texts occurred on the three dimensional perspective plane

<Table 1> comparison of OCR performance before and after applying sgementation and rectification

| | Count of matched words | | Count of words from OCR | |
|---|---|---|---|---|
| | Seg & Rect | Original | Seg & Rect | Original |
| Street images | 745 | 9 | 791 | 11 |
| 3D document images | 18399 | 166 | 19401 | 191 |
| Multi-plane | 854 | 3 | 924 | 5 |
| Smooth surface | 4206 | 174 | 5024 | 265 |

As seen in the table, without segmentation and rectification, OCR is no use on the three dimensional images.

## 5. Discussion

For text recognition system on camera based images, especially mobile applications, the main problem is the distortion from perspective projection and non-planar surfaces. The less significant factors are more or less resolved by state-of-the-art OCR, e.g., blur, low resolution, uneven lightening. We solve the distortion problem by employing text region detection with modeling of 3D shape for text region. OCR with automatic rectification enhances the performance greatly: for example, traditional OCR engines cannot recognize any text appeared on perspective plane.

There are limitations in the method employed in this paper which will be addressed in the future: verification module of text region is not implemented, no other languages except English for TTS module are supported, and the current system is only designed for single image rather than stereo images which will provide better solution.

## Reference

[1] A. Zandifar, R. Duraiswami, A. Chahine, L. Davis, "A Video Based Interface to Textual Information for the Visually Impaired," IEEE 4th icmi, pp.325-330, 2002.

[2] D. Doermann, J. Liang, H. Li, "Progress in Camera-Based Document Image Analysis," ICDAR. 2003.

[3] W. Newman, C. Dance, A. Taylor, S. Taylor, M. Taylor, T. Aldhous, "CamWorks: A Video-based Tool for Efficient Capture from Paper Source Document," Proc. In the ICMCS, pp.647-653, 1999.

[4] P. Wellner, "Interacting with Paper on the DigitalDesk," Comm. ACM, Vol.36, No.7, pp.87-96, 1993.

[5] J. Liang, D. DeMethon, D. Doermann "Geometric Rectification of Camera-Captured Document Images," IEEE Trans. PAMI. 2006.

[6] N. Chaddha, R. Sharma, A. Agrawai, A. Gupta, "Text Segmentation in Mixed Mode Images," in Proc. Asilomar Conf. Signals, Syst., Comput., Vol.2, pp.1356-1361, 1994.

[7] Y. Zhong, H. Zhang, A.K. Jain, "Automatic Caption Localization in Compressed Video," IEE Trans. PAMI., Vol.22, No.4, pp. 385-392, 2000.

[8] S. Lee, Y. Kim, S. Choi, "Fast Scene Change Detection Using Direct Feature Extraction from MPEG Compressed Videos," IEEE Trans. on Vol.2, Issue4, Dec., 2000 pp.240-254.

[9] A. Jian, S. Bhattacharjee, "Text Segmentation Using Gabor Filters for Automatic Document Processing," Machine Vis. Applicat., Vol.5, pp.169-184, 1992.

[10] C. Jung, Q. Liu, J. Kim, "A New Approach for Text Segmentation Using a Stroke Filter," Signal Processing, 88, pp.1907-1916, 2008.

[11] V. Wu, R. Manmatha, E. Riseman, "Textfinder: An Automatic System to Detect and Recognize Text in Images," IEEE. Trans. Pattern Anal. Mach. Intell., Vol.21, No.11, pp. 1224-1229, 1999.

[12] M. Guarnera, G. Messina, E. Ardizzone, L. Agro, "Text localization from photos," Digest of Technical Papers International Conference on Consumer Electronics, pp.1-2, 2009.

[13] L.L. Sulem, A. Zahour, B. Taconet, "Text Line Segmentation of Historical Documents: a Survey," IJDAR 2007.

[14] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic hand-written text-line extraction," ICDAR 2001.

[15] R. Ryue, J. Song, M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction," IEEE Trans. CSVT, 2005.

[16] R. Manmatha, N. Srimal, "Scale space technique for word segmentation in handwritten manuscripts," PAMI, 2005.

[17] Shi, Z., Venu Govindaraju, "Line separation for complex document images using fuzzy runlength," Proceedings. First International Workshop, 2004.M. Lyu, J. Song, M.

[18] M. Feldback, K.D. Tonnies, "Line Detection and Segmentation in Historical Church Registers," ICDAR, 2001.

[19] Y. Li, Y. Zheng, D. Doermann, "Script-independent Text Line Segmentation in Freestyle Handwritten Documents," IEEE Trans. PAMI., 2008.

[20] E. Oztop et al, "Repulsive attractive network for baseline extraction on document Images," IEEE Signal proceesing. 1997.

[21] Tseng, Lee, "Recognition based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm," PR Letter, 1999.

[22] S. Pollard, M. Pilu, "Building cameras for capturing documents," IJDAR, Vol.7, pp.123-137, 2005.

[23] P. Clark, M. Mirmehdi, "Estimating the orientation and recovery of text planes in a single image," in Proc. BMVC, pp.421-430, 2001.

[24] G. Myers, R. Bolles, Q. Luong, J. Herson, H. Aradhye, "Rectification and recognition of text in 3-D scenes," IJDAR, Vol.7, pp.147-158, 2005.

[25] A. Ulges, C. Lampert, T. Breul, "Document image dewarping using robust estimation of curled text lines," Proc. ICDAR, pp.1001-1005, 2005.

[26] C. Wu, G. Agam, "Document image de-warping for text/graphics recognition," in SPR2002, Int. Workshop on

Stat. and Struc. Pattern Recognition, Lecture Notes in Computer Science, Vol.2396, pp.348-357, 2002.

[27] Z. Zhang, C. Tan, "Correcting document image warping based on regression of curved text lines," ICDAR, Vol.1, pp. 589-593, 2003.

**신 현 경**

e-mail : hyunkyung@kyungwon.ac.kr

2003년 State University of New York at Stony Brook 응용수학과(공학박사)

2008년~현 재 경원대학교 수학정보학과 조교수

관심분야 : Neural network, Machine Learning, Image processing.