Improved Hybrid Binarization based on Kmeans for Heterogeneous document processing

Mahmoud Soua, Rostom Kachouri, Mohamed Akil

Improved Hybrid Binarization based on Kmeans for Heterogeneous document processing

Mahmoud Soua
soua.mahmoud@esiee.fr

Rostom Kachouri
rostom.kachouri@esiee.fr

Mohamed Akil
mohamed.akil@esiee.fr

Université Paris-Est, Laboratoire d'Informatique Gaspard-Monge, Equipe A3SI, ESIEE Paris, France

***Abstract*—Nowadays, more and more scanned documents are converted into editable electronic representation. This proceeding relies on the Optical Character Recognition (OCR) tool-chain. Generally, an OCR system is based on the important binarization step that separates character strokes from the background document. In this context, one of more robust binarization methods is the recently proposed Hybrid Binarization based on Kmeans (HBK). It handles effectively scanned documents which includes text on simple background. Nevertheless, in Heterogeneous documents, HBK ends up with some issues when extracting foreground text from complex background images. Moreover, HBK assumes to have a dark foreground against a clear background. Otherwise, it fails to render correct binarization colors. In this paper, we propose to improve the HBK method for handling efficiently Heterogeneous documents. Indeed, our proposal employs a layout analysis process that classify document regions into text and image. Image regions are enhanced with Gamma Correction (GC) before HBK binarization. Text regions are treated directly with HBK, keeping its effectiveness on text with homogeneous background. To ensure a robust and independent color rendering in the binarized documents, we control the labeling polarity of text and background through a pixel density-based technique. According to our experiments on LRDE and ICDAR datasets, we demonstrate that I-HBK outperforms HBK when dealing with Heterogeneous documents in both F-measure and OCR accuracy.**

***Keywords*—OCR, Binarization, Heterogeneous Documents, HBK, Gamma Correction**

I. Introduction

Optical Character Recognition (OCR) is a great interest in pattern analysis field. Its goal is to recognize the characters in a document to form a digital text file which can be edited and processed. Binarization, is a common and important first step in OCR systems. It converts the pixel values of document images into two-level representations for text and non-text regions.

Several binarization methods were proposed in the literature [5]. According to our knowledge, the Hybrid Binarization based on Kmeans (HBK) [7][1] outperforms state of the art methods such as SauvolaMS$_{kx}$ [6] and Niblack [4]. HBK scores a character recognition rate of 91% when performing on magazine documents [7]. Actually, its hybrid approach ensures a robust binarization on scanned documents including text on homogeneous background. Indeed, it produces a local binarization for characters and decreases artifacts thanks to its global correctness. However, regarding heterogeneous documents including text, drawing and natural scene images, HBK fails to extract correctly text from complex background. Moreover, similar to conventional binarization methods, HBK makes the assumption to have dark foreground against clear background and leads sometimes to generate inappropriate binary document colors. To overcome these limits, we propose in this paper an Improved HBK binarization that we call I-HBK.

To deal with text extraction from complex background, documents can be enhanced to increase the image visibility and details which helps to distinguish characters. In this context, some algorithms perform image enhancement before the binarization step [11]. For example, we can find the Histogram Equalization [9], Low Pass Filter [12], High Pass Filter [12], Hommorphic Filter [13] and image Gamma Correction (GC) algorithms [10]. Compared to the above methods, Gamma correction overcomes effectively the problem of uneven illumination [16]. In the other hand, the more efficient algorithm employed in the literature to handle the text and the background polarity problem, is the pixel density-based technique [17].

In the following, we study the binarization process of heterogeneous documents in section 2. Then, we describe our proposed I-HBK method in section 3. Next, in section 4, obtained results are shown and discussed. Finally conclusion is drawn in section 5.

II. Heterogeneous Document Binarization

Scanned text documents are categorized into three main kinds: documents with Simple text images, documents with Caption text images and documents with Scene text images [15]. Furthermore, documents with heterogeneous content images can be considered as a fourth kind of scanned documents. It is generally named Heterogeneous Document and includes at least two of the above mentioned document types. Figure 1 illustrates an example of Heterogeneous Document that contains both text regions (Figure 1-a) and image regions (Figure 1-b: Caption text image, Scene text image).

Generally extracting text from Scene text images is more challenging than in Caption text ones. For this, we consider in the following only Simple text images and Scene text images to study the performance of robust binarization methods and discuss their limits on Heterogeneous documents. In the literature, our recently proposed Hybrid Binarization based on Kmeans (HBK) [7][1] outperforms well known binarization methods such as Niblack [4] and Sauvola [5]. It ensures high local binarization robustness and reduces the possible appearing artifacts using global correctness. In this binarization method, the

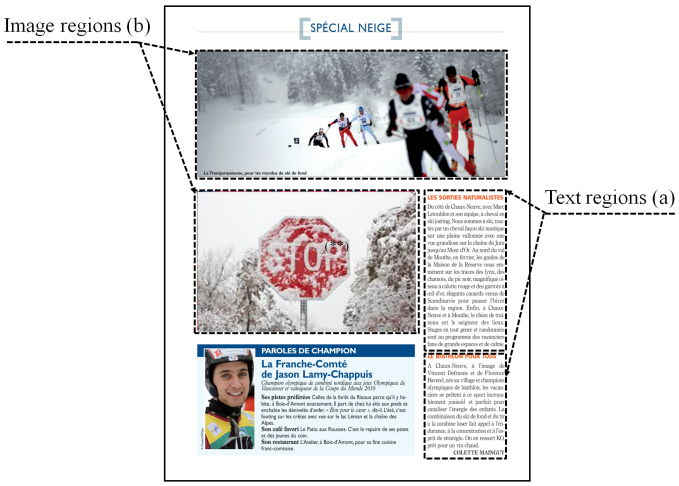


Figure 1: Heterogeneous Document: a. Text regions, b. Image regions.

input image is divided into equal sized blocks that can be set to small, medium or large according to the used character sizes in the document. Then, the Kmeans clustering algorithm [3] is applied on each block independently. The global phase gathers the obtained pixel values and number belonging to the related clusters in each block. The algorithm performs a correctness loop until convergence. According to Figure 2-(e,b-*), we can see that HBK provides high binarization quality on Simple text images, where characters are well separated. However, it does not do well on Scene text images with complex background, see Figure 2-(g,b).

To overcome the problem of extracting text from complex background images, some methods enhance the image before the binarization process. In this context, we consider approaches that look for a gamma value to help correct separation between foreground and background. The Gamma Correction Method (GCM) [8] is one of the most robust methods in this field. It aims to suppress non-text background details from scene text images by applying appropriate gamma value. For this, the GCM [8] transforms the input image to gamma-transformed images using a range of gamma values from 0.1 to 10.0 with increments of 0.1, resulting in 100 different images. A graylevel co-occurrence matrix for each image is computed to extract contrast and energy. Extracted textural feature measures of co-occurrence matrices with four orientations are averaged and threshold value is computed for each image using the Otsu algorithm [2]. Finally, Gamma, Contrast, Energy and Threshold values are examined to determine the appropriate value of gamma. Actually, GCM outperforms HBK on Scene text images with complex background and produces an efficient binarization quality Figure 2(g,c). However, according to Figure 2(e,c-*), GCM provides less accurate binarization quality than HBK on Simple text images. Thus, binarized characters are slightly touched due to the gamma transformation process. Table I illustrates the OCR Accuracy of HBK and GCM methods for Simple and Scene text images. The obtained results support the previous discussed visual observations.

We can see that, HBK outperforms GCM in processing

Table I: OCR Accuracy (%) of GCM and HBK methods on Heterogeneous documents: Simple and Scene text images

Methods	Simple text image (LRDE [6])	Scene text Images (ICDAR [18])
HBK [7]	**94.11**	50
GCM [8]	93.88	**98.38**

Simple text images by scoring 94.11% of OCR Accuracy. However, GCM is better than HBK in extracting text from scene text images when reaching 98,38% of OCR rate. Hence, we conclude that GCM and HBK methods can be complementary in Heterogeneous document binarization process.

In an another hand, we note that GCM and HBK methods yield to inverted foreground and background colors on Simple text image (Figure 2-(e,c-**), Figure 2-(e,b-**)) and on Scene text images (Figure 2-(f,c), Figure 2-(f,b)). We illustrate in Table II the effectiveness of the two studied methods. We note by D_f and C_f respectively dark and clear foreground. Likewise, C_f and D_b mean respectively dark and clear background.

Table II: Evaluation summary of HBK and GCM effectiveness on Simple and Scene text images

Methods	Simple text images				Scene text Image	
	D_f & C_b		C_f & D_b		-	-
	Single	Multiple	Single	Multiple	Single	Multiple
HBK [7]	**Yes**	**Yes**	No	No	No	No
GCM [8]	No	No	No	No	**Yes**	No

We observe from the Table II that on single regions HBK and GCM have a cross effectiveness when processing simple or Scene text regions. However, GCM always fails to provide high quality binarization when processing multiple regions. Otherwise, we remind that both HBK and GCM fail in some cases to extract the appropriate foreground and background colors. Considering the previous study, we describe our proposed I-HBK method in the following section.

III. Improved Hybrid Binarization based on Kmeans (I-HBK)

Our proposed I-HBK method aims to improve HBK by providing a better binarization quality and OCR Accuracy on Heterogeneous documents. In this type of documents, the textual information can be included in text or image regions. For this reason, our idea is to benefit from the cross advantages of HBK and GCM on both region types. Indeed processing HBK on text regions and GCM on image regions improves the final binarization quality. As shown in Figure 3, our proposed method includes four major stages. Firstly, a Layout analysis is performed to identify text and image regions. Secondly, a binarization algorithm based on Gamma Correction (GC) is tailored to handle image regions. We name it Gamma Correction-based HBK (GCHBK). Similarly, HBK is performed on text regions. The third stage consists on a Pixel Density-based Technique (PDT), where a foreground and a background color study is performed. Finally, binarized regions are gathered in one single document.

Figure 2: Visual binarization: a. Single and Multiple regions of scanned documents, b. HBK results, c. GCM results

A. *Layout Analysis*

It is important to recognize accurately whether a document region is text or image to perform the adequate treatment. For this, we perform a layout analysis on the input documents. We use the Page Layout Analysis (PLA) algorithm [19], part of the well known Tesseract engine 3.02 [19]. Actually, the PLA uses bottom-up methods, including binary morphology and Connected Component (CC) analysis, to estimate the type of the CCs. In addition, the method uses some measures to detect the tab-stops of column boundaries, indents, table columns. The column layout of the pages is determined from the detected tab-stops. The column layout constrains the construction of partitions of the page that are then gathered into text regions. A similar process is applied to detect image regions.

B. *Binarization stage (HBK and GCHBK)*

After extracting regions from the heterogeneous document, we employ the appropriate binarization algorithm to process respectively text and image regions. Indeed, we improve the HBK method processing on image documents by including a pre-processing image enhancement. For this, we use a Gamma Correction approach (GC) that helps to separate foreground

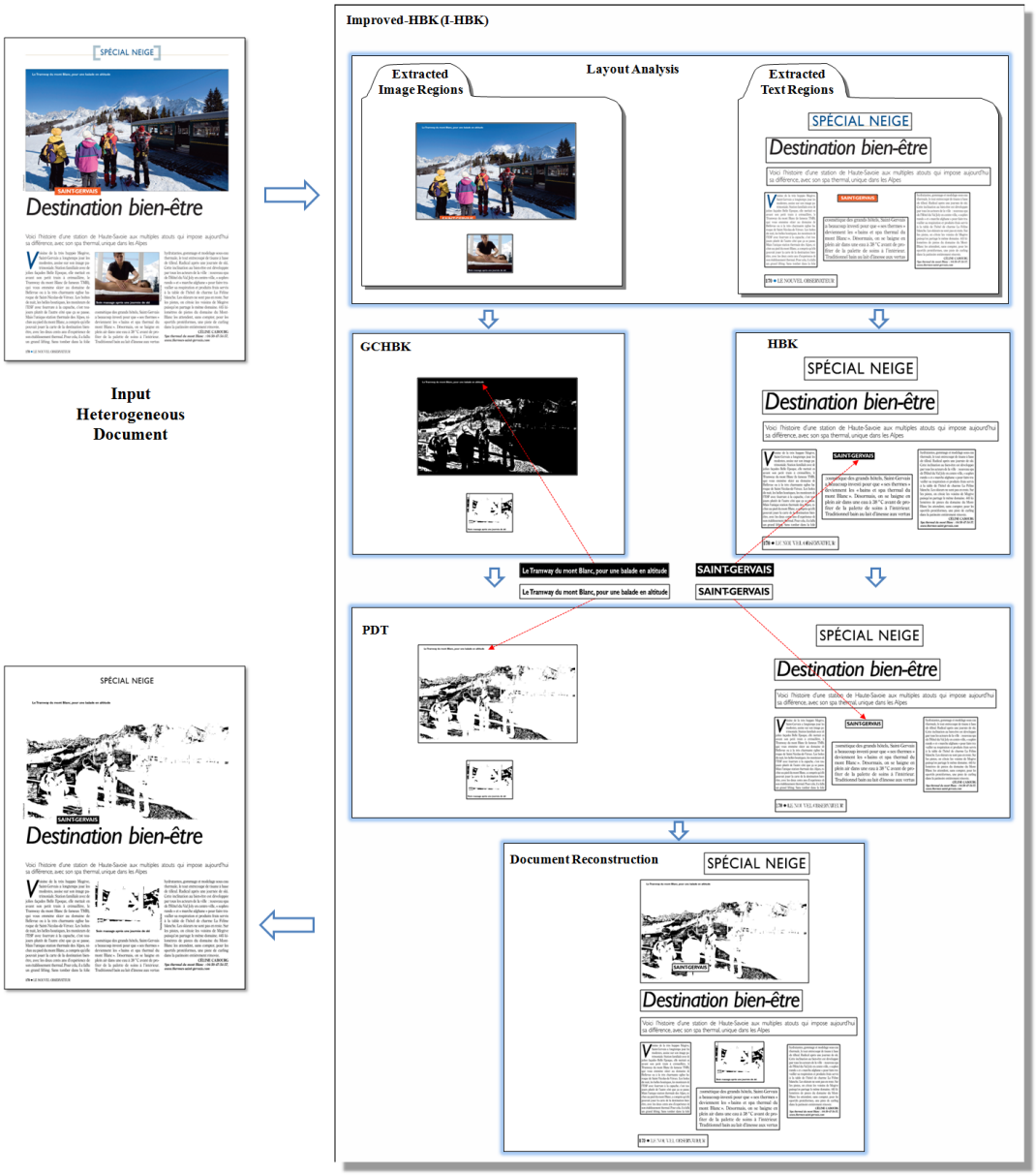


Figure 3: Flow diagram of the proposed I-HBK method

from complex background in text scene images. Actually, we employ GCM [8] Gamma Correction stage and we replace the Otsu binarization method with the HBK one. According to the literature [7], binarizing with HBK instead of Otsu is more effective scoring 91% of OCR Accuracy against 84%. We name this approach Gamma Correction-based HBK (GCHBK). Our evaluation of the proposed GCHBK is shown in the experimental results section.

C. Pixel Density-based Technique (PDT)

According to our Knowledge several binarization methods [4][5][7] assume to have a dark foreground against a clear background. Otherwise, they fail to render correct binarization colors. To overcome this problem some methods control the labelling polarity of text and background. For example, Clark [17] uses a simple decision logic based on the assumption that the background pixel number is greater than the text pixel one. This method fails in case characters are thick and occupy a significant area of the used window technique. So instead of using a window approach, we apply the same concept of pixel density-based technique on the extracted regions. This way, we control the foreground and background color polarity according to Equ 1.

$$P_i = \begin{cases} 0 & if N_{Bp} > N_{Wp} \\ P_i & else \end{cases} \quad (1)$$

with P_i designs a pixel, i in $[0, N_{Bp} + N_{Wp}]$, and N_{Bp}, N_{Wp} are respectively the number of black and white pixels in the studied region. The final reconstitution stage gathered the binarization of the two methods into one single document according to the recorded region coordinates.

IV. Experimental results

In the following, we evaluate our proposed I-HBK method. Firstly, we introduce the employed materials. Then, we compare the OCR Accuracy of the GCM and our proposed GCHBK technique to show its effectiveness when binarizing scene text images. Following, we evaluate our proposed I-HBK when comparing it with HBK according to Fmeasure and Visual quality. Finally, we study I-HBK using the OCR Accuracy.

A. Materials

The proposed I-HBK method was evaluated on LRDE-DBD[1] [6] and ICDAR[2] [18] datasets. In the following experiments, we fix the HBK block size to 32x32 pixels. According to the literature, this block size gives an acceptable binarization quality [7]. We use the Fmeasure metric to evaluate the binarization accuracy. In addition, Tesseract 3.02 was employed to perform the character recognition study. Following, we evaluate our GCHBK binarization approach on natural scene images.

B. GCM and GCHBK Evaluation

We evaluate our proposed enhancement of HBK binarization on image documents using the Gamma Correction approach (GC). Indeed, we perform the image enhancement part of GCM then we use HBK as a final binarization process in place of Otsu (see section III-B). Table III compares GCM and our GCHBK technique on ICDAR scene text images. We note that GCHBK provides better character recognition

Table III: Comparison of GCM and GCHBK methods based on OCR Accuracy for ICDAR scene text images

Methods	OCR Accuracy (%)
GCM: GC + Otsu	84.4
GCHBK: GC + HBK	**85.2**

accuracy by scoring 85.2%. We explain this result with the fact that HBK method outperforms Otsu in the binarization step. Indeed, the employed hybrid processing strategy takes into account both local and global pixel information to provide efficient binarization result. We provide a visual result in Figure 4. As we can see, characters are better separated using GCHBK.

C. I-HBK evaluation based on Visual quality

Following, we show in Figure 5 the binarization results of HBK, GCHBK and our I-HBK methods when processing a sample Heterogeneous document from the LRDE-DBD dataset. We note that I-HBK gives clearly the best binarization result. HBK produces high binarization quality on text regions. However, it fails in rendering the appropriate colors of text and background because it assumes that text is in dark color

(a) (b) (c)

Figure 4: Binarization quality of GCM and GCHBK on text scene image from ICDAR dataset. a. Original image, b. GCM binarization, c. GCHBK binarization

and background is in clear one. GCHBK succeed in binarizing image regions. However it shows information loss in text regions with homogeneous background.

Following we evaluate HBK and I-HBK methods on both Simple text images and Heterogeneous documents.

D. Comparison of HBK and I-HBK methods based on analytic analysis

The HBK method is robust binarization method on Simple text images. Following, we demonstrate in Table IV the reliability of I-HBK on LRDE-DBD images that include only text information.

Table IV: Comparison of HBK and I-HBK methods based on Binarization-Accuracy on 125 Simple Text LRDE-DBD images

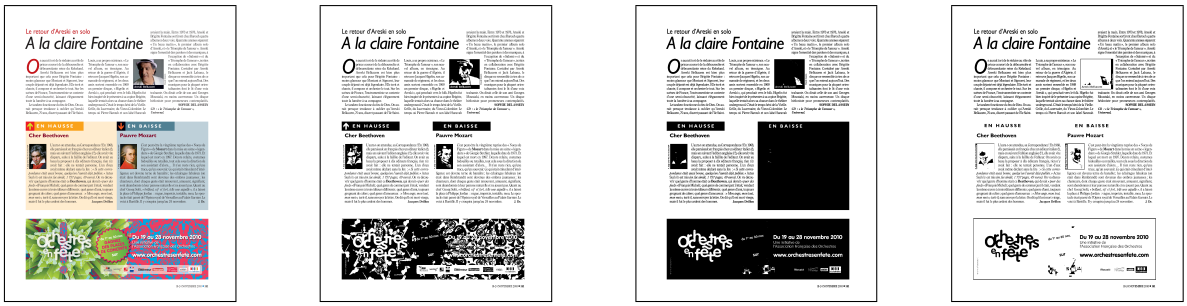
Methods	FMeasure (%)
HBK [7]	98,21
I-HBK	**98,84**

As expected, I-HBK and HBK have a close Fmeasure result on simple text images. Indeed, the processed documents contain only text regions that include, in most cases, dark text on light foreground. We observe a small improvement as I-HBK handles properly regions with light text on dark background. Thus, text and background colors are rendered correctly thanks to the used pixel density-based technique. In addition, some text regions include non-homogeneous background, on which I-HBK gives better text binarization quality. Next, we demonstrate the effectiveness of our proposal using the OCR Accuracy on Heterogeneous documents.

To process Heterogeneous documents, we choose a set of LRDE-DBD documents including Scene text images. For this purpose, we construct their OCR ground-truth as they are not available in the LRDE-DBD Dataset. Figure 6 shows the OCR accuracy of our proposed I-HBK and HBK methods on these documents. We observe that our I-HBK proposal improves considerably the OCR accuracy of HBK. This is due to the good workload distribution between HBK and GCHBK methods. Furthermore, GCHBK allows to extract well the text in Scene text images contrary to HBK that mixes foreground and background colors.

[1]Copyright(c) 2012. EPITA and Development Laboratory (LRDE) with permission from Le Nouvel Observateur. LRDE-DBD is available on-line on the web site: $http : //www.lrde.epita.fr/cgi - bin/twiki/view/Olena/DatasetDBD$

[2]ICDAR 2013 dataset is available on the web site $http : //dag.cvc.uab.es/icdar2013competition$



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

Figure 5: Visual evaluation of the I-HBK method: a. Original document, b. HBK binarization and c. GCHBK binarization, d. IHBK binarization

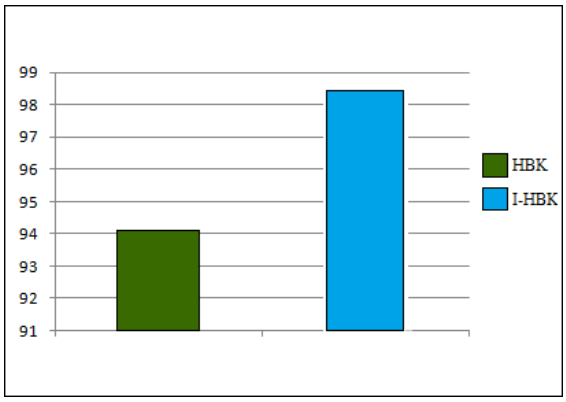


Figure 6: Comparison of HBK and I-HBK OCR accuracy on Heterogeneous LRDE-DBD documents

V. CONCLUSION

Binarization is an important step in any task of image processing and analysis. Recently we proposed the Hybrid Binarization based on Kmeans method (HBK). To process heterogeneous documents, we proposed in this paper a new binarization method named Improved-HBK (I-HBK). In addition of HBK text binarization, our proposal performs layout analysis scheme, then employs a modified Gamma Correction approach GCHBK to extract text from complex background. In addition, it studies the image regions to control the final color rendering of text and background. According to experiments, we demonstrated that I-HBK outperforms HBK in both FMeasure, and OCR accuracy when processing LRDE and ICDAR documents.

REFERENCES

[1] M.Soua, R.Kachouri, M.Akil GPU parallel implementation of the new hybrid binarization based on Kmeans method (HBK) Journal of Real-Time Image Processing, Springer: 1-15, 2014.

[2] N. Otsu. A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man and Cybernetics, 9(1): 62-66, 1979.

[3] S. P. LLOYD, Least square quantization in PCM, IEEE Transactions on Information Theory 28(2): 129-137, 1982.

[4] W. Niblack, An Introduction to Digital Image Processing, Standberg Publishing Company, 1985.

[5] J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition, 33: 225-236, 2000.

[6] G. Lazzara, T. Graud, Efficient Multiscale Sauvolas Binarization, International Journal on document analysis and recognition, 2013.

[7] Soua, M., Kachouri, R., Akil, M.: A new hybrid binarization method based on Kmeans. In: IEEE International Symposium on Communications, Control, and Signal Processing: 118-123, 2014.

[8] C. P. Sumathi and G. Gayathri Devi, Automatic Text Extraction From Complex Colored Images Using Gamma Correction Method, Journal of Computer Science, 10(4): 705-715, 2014

[9] Pizer, S. M., E.P. Ambum and J.D. Austin, Adaptive histogram equalization and its variation, Computer Vision, Graphics, and Image Processing 39(3): 355-368, 1987.

[10] Asadi Amiri, S., Hassanpour ,H. and Pouyan, A.,Texture Based Image Enhancement Using Gamma Correction, Middle-East Journal of Scientific Research. 6: 569-574, 2010.

[11] Gilboa, G., Sochen, N. and Zeevi, Y. Y., Image enhancement and denoising by complex diffusion processes, IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(8): 1020-1036, 2004.

[12] Guillon, S., Baylou, P., Najim, M. and Keskes, N., Adaptive nonlinear filters for 2D a nd 3D image enhancement, Signal Processing, 67(3): 237-254, 1998.

[13] Fries, R. and Modestino, J., Image enhancement by stochastic homomorphic filtering, IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(6), 625-637, 1979.

[14] Farid, H., "Blind inverse gamma correction", IEEE Transactions on Image Processing, 10: 1428-1433, 2001.

[15] C.p Sumathi. A survey on various approaches of text extraction in images. International Journal of computer sciences engineering 3(4): 15-19, 2012.

[16] K.Kaur, N.Kanwal, J.S.Bhullar. A Technique for Enhancement of Gray Image using Local Gamma Correction. International Journal of Computer Applications. 105(5):36-39 ,2014

[17] P. Clark and M. Mirmhedi. Rectifying perspective views of text in 3D scenes using vanishing points. Pattern Recognition, 36(11): 2573-2686, 2003.

[18] Karatzas, D. and F.Shafait, S.Uchida, M.Iwamura. ICDAR Robust Reading Competition. Document Analysis and Recognition (ICDAR): 1484-1493, 2013.

[19] R. Smith, An Overview of the Tesseract OCR Engine. International Conference on Document Analysis and Recognition, 2007.