Text extraction in document images: highlight on using corner points

Vikas Yadav
ECE department
Visvesvaraya National Institute of Technology
Nagpur, India,
vikas.yadav11021995@gmail.com

Nicolas Ragot
Université François Rabelais Tours
Laboratoire Informatique (LI EA6300)
Tours, France
nicolas.ragot@univ-tours.fr

***Abstract*—During past years, text extraction in document images has been widely studied in the general context of Document Image Analysis (DIA) and especially in the framework of layout analysis. Many existing techniques rely on complex processes based on preprocessing, image transforms or component/edges extraction and their analysis. At the same time, text extraction inside videos has received an increased interest and the use of corner or key points has been proven to be very effective. Because it is noteworthy to notice that very few studies were performed on the use of corner points for text extraction in document images, we propose in this paper to evaluate the possibilities associated with this kind of approach for DIA. To do that, we designed a very simple technique based on FAST key points. A first stage divide the image into blocks and the density of points inside each one is computed. The more dense ones are kept as text blocks. Then, connectivity of blocks is checked to group them and to obtain complete text blocks. This technique has been evaluated on different kind of images: different languages (Telugu, Arabic, French), handwritten as well as typewritten, skewed documents, images at different resolution and with different kind and amount of noises (deformations, ink dot, bleed through, acquisition (blur, resolution)), etc. Even with fixed parameters for all such kind of documents images, the precision and recall are close or higher to 90% which makes this basic method already effective. Consequently, even if the proposed approach does not propose a breakthrough from theoretical aspects, it highlights that accurate text extraction could be achieved without complex approach. Moreover, this approach could also be easily improved to be more precise, robust and useful for more complex layout analysis.**

Keywords—text extraction; corner points; FAST (Features from Accelerated Segment Test); multilingual documents; historical documents; handwritten documents.

I. INTRODUCTION

To be able to access and make the navigation inside the increased mass of digitized documents easier, text extraction is essential. Text segmentation is indeed part of the more complex process of layout analysis and it is an important preliminary step for processes such as skew detection and correction, line and word extraction, OCR, word spotting, etc. Indeed, many skew detection methods, line and word segmentation approaches and OCR systems are giving erroneous results in the presence of non-textual elements (graphical, etc.) along with text elements. This problem has been widely studied in the field of Document Image Analysis (DIA) and many approaches have been designed for contemporary documents (newspaper, books, administrative documents and letter, etc.) as well as for historical documents (handwritten documents, old books, etc.). As it will be shown in section 2, many approaches rely on more or less complex methods based on several image processing techniques like preprocessing (binarization, filtering, etc.), image transform/filters to get components, edges or pixel descriptions, analysis of these elements to decide whether it is text or not.

Recently, with the increased use of new media and technological devices, many systems are elaborated for recognizing text from news channels headlines, traffic signals instructions, camera or mobile videos, etc. In such systems, text is often merged inside images. To extract it, several approaches rely on corner points. The reason is that corner points were designed to be less sensitive to noise, variations of illumination, rotation, resolution, and so they are particularly adapted to images coming from videos.

Surprisingly, contrary to text extraction inside videos or natural scene images, corner points were not much studied for text extraction in document images. Reasons for that are probably because, from theoretical point of view, classical corner points in text images are not really meaningful. Indeed, since ink is based on black to white transitions, many points are often extracted and their position is not very robust: repeatability and position cannot be guaranteed especially on noisy and black and white documents (which is often the case in DIA since many processes starts with binarization).

Nevertheless, by this paper, we would like to show/remind that corner points could be of high interest for text extraction in document images. Note here, that this study is focused on extraction of text blocks as it was studied in previous ICDAR competitions [6],[8],[15] We are neither expecting segmentation of textual elements at pixel level, nor trying to recover reading order. In such scenario, relative positions of corner points and their repeatability is not needed itself and our study is showing that interesting results could be achieved even with a very simple and generic method that is contrasting with other complex approaches of the literature. We also think that improving the method could allow to perform more complex layout analysis task easily and efficiently.

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.67

The proposed approach relies on the same assumption as the one used for text extraction in videos or natural images: it is based on the observation that corner points are more densely spread over text regions, compared to other regions in the image. Here we simply extract key points detected by FAST method [10][11] and analyze their density inside blocks of fixed sized over the image to keep the more dense ones as text regions. For this, we have used only one threshold parameter that we kept fixed for all images whatever their content (handwritten, typewritten, script and language, font size, kind of document, etc.). Even with such simple approach, the results in terms of recall and precision on various kind of documents is very interesting and show its robustness to language, tilt and size of text, as well as illumination variations, degradations and resolution, thanks to corners properties. Finally, the computational complexity of this approach could be of interest in comparison with other methods based on connected components analysis for example.

In the remaining, section 2 provides an overview of existing text extraction methods from both videos/natural images and document images application domain. The complete description of the used technique is given in the section 3. Section 4 provides experimental results on various kind of images, some of them coming from classical benchmarks like Saint Gall and Tobacco-800 datasets.

II. Related Work

Text extraction from document images was studied for a long time. It is generally part of the more complex process of layout analysis. Sometimes authors are focusing on text/graphic separation[1] and sometimes the focus is much more put on text extraction in text documents with no graphics: from line/word extraction to text segmentation at pixel level. Consequently, techniques used generally depend on the full flow and final objective, which could explain relative complexity of most of approaches. Several surveys were published on the topic such as [13][14][15]. Existing techniques for text extraction can be broadly considered into connected components analysis, edge-based, filter-based and texture-based approaches.

Most frequent approaches are the one based on components analysis [21][22][23][6][7]. In [24], authors have used an hybrid approach with connected components analysis as well as edges. In [25], connected components analysis was used as well as several transforms (curvelets and undecimated wavelets).

For filter-based approaches, we can refer to [2] in which authors have proposed techniques using Gabor filters and log Polar wavelets transform for text extraction. [5][19] have also used Gabor filter for text extraction. [16][18] based their method on digital filter using Haar wavelet. Zaravi *et al.* [17], in their case, used discrete wavelets transform and dynamic thresholding to extract regions of interest. In [20], authors have used Kalman filter for handwritten text line extraction in low resolution images.

Edge-based methods were mainly used for color images or camera captured images [26][27][28][29] but not only [30][31]. Some were also used as well with connected components analysis [32].

Finally, other methods are based on pixel classification: with clustering [37]; after a characterization at pixel level by textures [33][34]; using Markov Random Field [35]; or statistical methods [36]. In [8], authors have used inpainting.

By this literature review, we can see that very few methods are based on corner points, even very recently in ICDAR competitions [6][12]. Exceptions are [1] in which authors have used corner points obtained with Harris method to extract text regions but it was tested only on Chinese script. This method also relies on many threshold values to classify text, image and background or noisy regions. [2] is also another example in which corner points are characterized by Gabor filters.

On the other side, many approaches have been studied to extract text from scene images or document images captured by mobile or camera. Among them preliminary ones were based on same techniques as for classical DIA [38] . But recently, the use of corner points has received more attention [1][3][4][9].

III. A Simple Approach Based On Corner Detectors

As explained previously, the goal of this study was not to propose a breakthrough from theoretical aspects. In fact we have just designed here a simple approach based on key points to extract and localize text blocks. The output of proposed approach could be further refined and processed for a complete layout analysis at text level, but existing approaches could do that efficiently on the resulting text blocks. We also think that corners could be used further for these tasks (paragraph extraction, line segmentation especially), instead of using another technology, but this is outside the scope of this study.

In fact, the proposed approach uses the density and distribution of corner points inside windows, to accurately decide whether it is text or not (we have combined zoning technique with corners extraction to improve accuracy).

Flowchart in Fig. 1 explains the different steps of the system. These steps are detailed below:

Step 1] Smoothing of input image by Gaussian filter:

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{\frac{-(x^2+y^2)}{2\sigma^2}}, \qquad (1)$$

where σ is the standard deviation of Gaussian filter, and x, y are coordinates of pixels inside the image.

[1] Note here that this text/graphic separation task in graphical documents where text components are touching graphical elements is a specific case which is out of the scope of this study.

Step 2] Determine corner points by FAST corner detection technique [10][11]. In FAST algorithm, a pixel 'C' is choosed to be a keypoint depending on its intensity I_c and the one of its 16 neighbor's: if the intensities of a minimum of 12 pixels out of the 16 surrounding ones are either above or below a specified threshold, then it is a key point. The threshold decided by E Rosten and Drummond was 20% of I_c. We have taken the same threshold.

Step 3] Divide the image in 32*32 blocks (non-overlapping) and calculate the number of corner points inside each block.

Step 4] From the block which has the maximum number of corner points (*Nmax*), define a threshold T1 (the only threshold used) as 0.2**Nmax*. This threshold is also a relative value and hence it works even if resolution or size of image changes. We have taken 20% of the maximum density as the threshold from experimental evaluation (performed on different images than the one used for evaluation below).

Step 5] Blocks having more number of corner points than this threshold may belong to text regions, and blocks having less, belong to other regions (image, background, noise).

Step 6] After detecting text blocks in previous step, we check for connectivity of these blocks (8-connectivity) to build text regions.

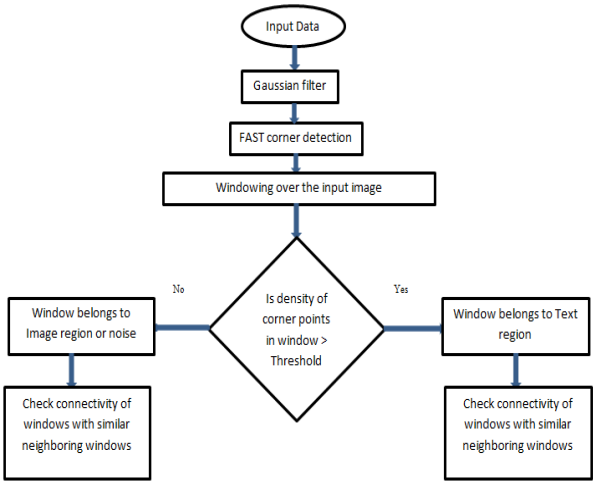


Figure 1. Flowchart of the corner-based text extractor.

IV. Experimental Evaluation

We have tested our system on many kind of different images, with different scripts, resolutions, orientations, layout, printed and handwritten, etc. Examples of results are given in following figures in which red blocks correspond to detected text blocks and blue blocks (when shown) are blocks not considered by the approach as text. Following figures illustrates some samples results.

Fig. 2 shows the ability to work on handwritten text whatever the orientation.

Fig. 3 is showing results on images at different resolutions. We can see that the precision (boundaries) of text area is better when resolution is higher. Nevertheless, the approach is still working well at low resolutions.

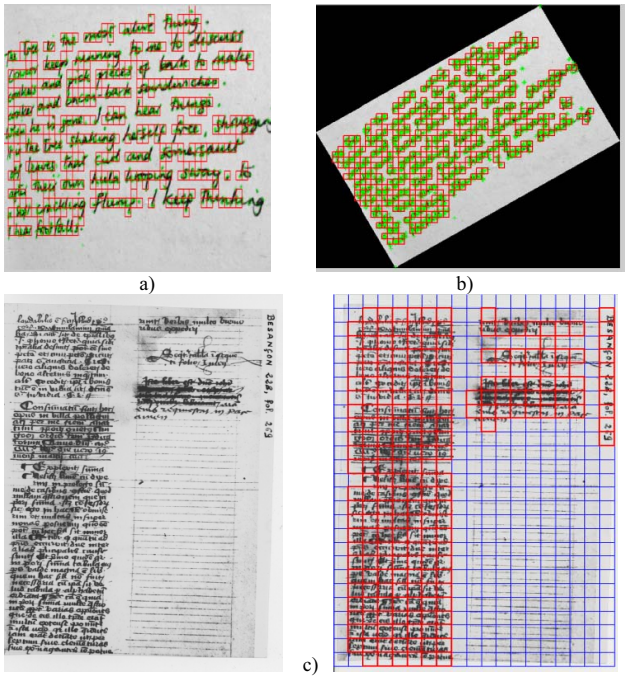
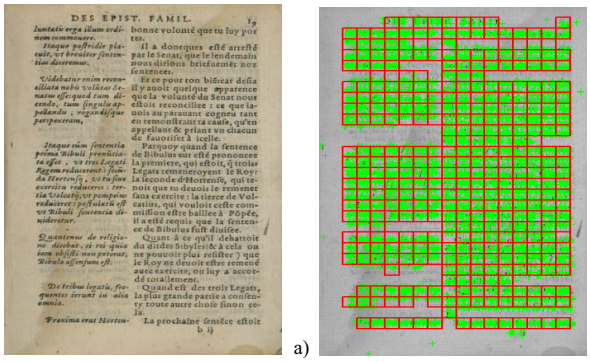
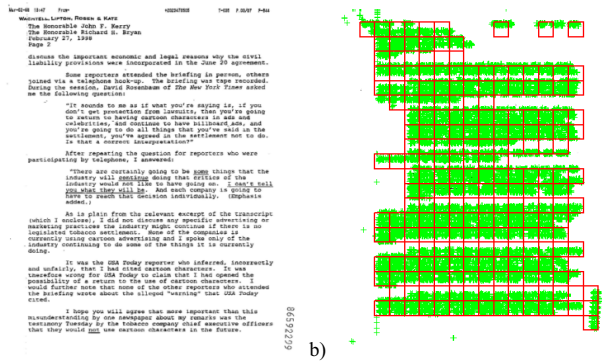


Figure 2. Examples of text blocks extracted inside handwritten documents: a) document at 96 dpi; b) 30° skewed image at 96 dpi; c) historical document from BVMM[2] at 300 dpi (left original; right detected text blocks).



[2] BVMM (Bibliothèques Virtuelles des Manuscrits Médiévaux) / IRHT (Institut de recherche et d'histoire des textes): http://bvmm.irht.cnrs.fr



b)

Figure 3. Impact of resolution; a) Image from BVH[3] book (300 dpi); b) Image from Tobacco-800 dataset (reduced at 72 dpi).

We have also tested our system on texts in various languages. Fig. 4 shows as example text extraction from an image having both Telugu and Arabic text.

Figure 4. Text extracted on mixed scripts (96 dpi).

Finally, Fig. 5, also on handwriting, is showing the ability to handle noise and illumination problems (see text detected inside black ink area at bottom right).

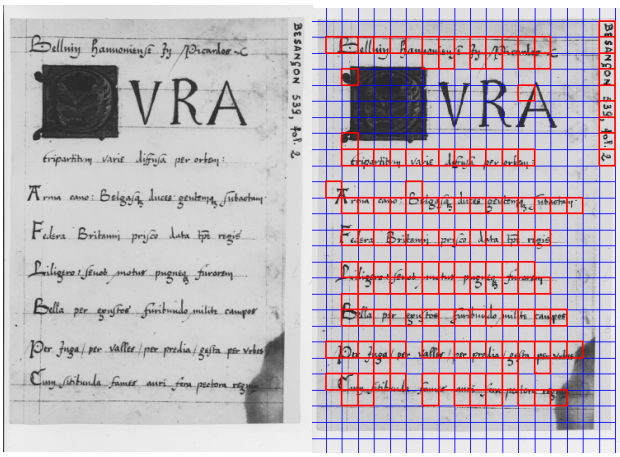


Figure 5. Results on noisy document image from BVMM[2] sparsely handwritten (300 dpi).

We have also evaluated our engine on handwritten George Washington letters and HDLAC2011 dataset. Handwritten letters with different font sizes and in various languages required image to be divided in 16*16 blocks for achieving best accuracy. Figure 6 shows output with different sizes of blocks.

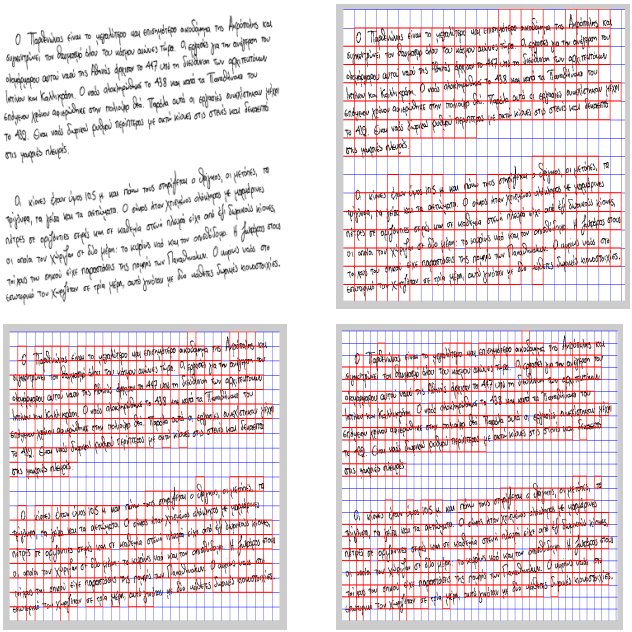


Figure 6: Text extraction results by dividing image in 16, 20 and 24 blocks

Figure 7: Text extraction results on HDLAC2011 dataset by dividing image in 16 blocks.

Our engine also has drawbacks for detecting text regions on newspaper with dot images. As dot images has large number of corner points on image region also, extraction region with this technique gives low accuracy.

To obtain a quantitative evaluation of this approach, we have calculated the accuracy in terms of precision and recall of textual elements on several images coming from classical benchmarks (Tobacco-800, Saint Gall) as well as other documents. The precision and recall are computed by (2) and the results are given in Table 1 for different categories.

[3] BVH (Bibliothèques Virtuelles Humanistes) / CESR (Centre d'Etude Supérieur de la Renaissance): http://www.bvh.univ-tours.fr

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

TP= Number of red blocks correctly spotting text region

FN= Number of blue blocks containing text region

FP= Number of red blocks containing non-text region

In this table, we can see that even with a fixed window size and threshold T the results are not much decreased when resolution is going from 300 to 72 dpi. We can see also that most of the time, precision and recall are quite high (around 95%) except for some cases where it decreases at 87%. This is showing the robustness of corner points for such kind of problems. This is also clear that by adapting a little bit the method we can easily increase the accuracy by modifying parameters and/or by doing some neighboring inspection. We can also see, on Fig. 2b), 3, and 5 that with little new rules text layout can be extracted by detecting lines, end of paragraphs and columns. Nevertheless, some problems are still remaining, especially for big fonts titles (Figure 5) for which few corners could be extracted and also for some documents where illustration are specific such as with arabesques and lot of details.

V. CONCLUSION

In this paper, we have shown that using corner points on document images (whatever their resolution, language, level of degradation), it could be very simple to obtain an accurate text extractor at low cost and without bothering a lot about parameters (standards ones are working fine in nearly all cases). The method simply extract FAST keypoints in image and then, with a zoning approach, it keeps as text blocks the zones in which the density of corners is over 20% the more dense block in the image. Experimental results are showing that with such simple methods, precision and recall are over 90% (most often around 95% in average). Table 1 delineates high precision and recall values for multilingual documents and noisy text documents. The technique fails for big size fonts as well as for some specific illustrations for which corners are responding too much. But at the same time, it is quite fast (and could be easily parallelized) and could be improved further, which could make it competitive with other state of art approaches while being less complex. Finally, this method seems also to be usable to extract more complex layouts such as paragraphs, and lines.

Datasets	Tobacco -800, 72 dpi (50 images)	Tobacco-800, 300 dpi (50 images)	Saint Gall, 300 dpi (60 images)	BVH[3] French book, 300 dpi (8 images)	Telugu and Arabic text images, 96 dpi (12 images)	Handwritten text images in other languages, 96dpi and 300 dpi (10 images)	Handwritten Washington letters (15 images)	HDLAC2011 dataset (15 images)
Recall	93.21%	94.74%	89.65%	91.50%	86.91%	92.30 %	90.40%	87.45%
Precision	97.36%	97.80%	94.12%	96.65%	90.72%	95.82%	92.34%	89.70%

TABLE I. PRECISION AND RECALL OF PROPOSED METHOD ON SEVERAL DATASETS.

REFERENCES

[1] Fanfeng Zeng, Guofeng Zhang and Jin Jiang, "Text Image with Complex Background Filtering Method Based on Harris Corner-point Detection", *Journal of Software*, Vol. 8, No 8, pp. 1827-1834, 2013.

[2] Farshad Nourbakhsh, Peeta Basa Pati, A. G. Ramakrishnan, "Document Page Layout Analysis Using Harris Corner Points", *Intelligent Sensing and Information Processing*, 2006.

[3] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu and Thomas S. Huang, "Text From Corners: A Novel Approach to Detect Text and Caption in Videos", IEEE *Transactions On Image Processing*, Vol. 20, n° 3, pp. 790-799, march 2011.

[4] Narasimha Murthy K N, Dr. Y S Kumaraswam, "Robust Model for Text Extraction from Complex Video Inputs Based on SUSAN Contour Detection and Fuzzy C Means Clustering", IJCSI *International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 3, September 2011.

[5] Rainer Herzog, Arved Solth and Bernd Neumann, "Text Block Recognition in Multi-Oriented Handwritten Documents", Report, http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/207/, 2014.

[6] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, **"**ICDAR2013 Competition on Historical Book Recognition – HBR2013", 12th *International Conference on Document Analysis and Recognition*, 2013.

[7] Zhixin Shi, Srirangaraj Setlur and Venu Govindaraju, **"**Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map", Proceedings of the 8th *International Conference on Document Analysis and Recognition*, pp. 794-798, Aug. 29- Sept. 1, 2005.

[8] Yen-Lin-Chen, "Automatic text extraction, removal and inpainting of complex document images", *International Journal of Innovative Computing, Information and Control*, Vol. 8, n° 1(A), pp. 303-327, 2012.

[9] Xin Liu, Jin Dai, Yuanyuan Jia and Rubin Liu, "Caption Region Detection in Video Images by Improved Corner Detector", *International Journal of Signal Processing, Image Processing and Pattern Recognition,* Vol.7, No.4, pp.409-420, 2014.

[10] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection", Computer Vision – ECCV 2006, *Lecture Notes in Computer Science,* Vol. 3951, pp. 430-443, 2006.

[11] Edward Rosten, Reid Porter, and Tom Drummond, "Faster and better: a machine learning approach to corner detection", IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32 , Issue 1, pp. 105-119, 2008.

[12] A. Antonacopoulos, S. Pletschacher, D. Bridson and C. Papadopoulos, "ICDAR2009 Page Segmentation Competition", 10th *International Conference on Document Analysis and Recognition,* 2009.

[13] Keechul Jung, Kwang In Kimb, Anil K. Jain, "Text information extraction in images and video: a survey", *Pattern Recognition*, Vol. 37, Issue 5, pp. 977-997, 2004.

[14] Deepika Ghai, Neelu Jain, "Text Extraction from Document Images- A Review", *International Journal of Computer Applications*, Vol. 84, n° 3, 2013.

[15] David Doermann, Karl Tombre (Eds.), *Handbook of Document Image Processing and Recognition*, Springer, 2014.

[16] S.J. Ha, B. Jin, N.I. Cho, "Fast Text Line Extraction in Document Images", 19th IEEE *International Conference on Image Processin*g, pp. 797-800, Orlando, Sept. 30-Oct 3 (2012).

[17] D. Zaravi, H. Rostami, A. Malahzaheh, S.S Mortazavi, "Journals Subheadlines Text Extraction Using Wavelet Thresholding and New Projection Profile", World Academy of Science, *Engineering and Technology*, 49, pp. 686-689, 2011.

[18] S. Audithan, RM. Chandrasekaran, "Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform", *European Journal of Scientific Research*, 36, pp. 502-512, 2009.

[19] Y.L. Qiao, M. Li, Z. M. Lu, S.H. Sun, "Gabor Filter Based Text Extraction from Digital Document Images", Proceedings of the 2006 *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 297-300, USA, 2006.

[20] A. Lemaitre, J. Camillerapp, "Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image", Proceedings of the 2nd *International Conference on Document Image Analysis for Libraries*, pp. 45-52, Lyon, April 27-28, 2006.

[21] V.K. Koppula, N. Atul, U. Garain, "Robust Text Line, Word And Character Extraction From Telugu Document Image", 2nd *International Conference on Emerging Trends in Engineering and Technology*, pp. 269- 272, Dec. 16-18, 2009.

[22] Z. Shi, S. Setlur, V. Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines", 10th *International Conference on Document Analysis and Recognition*, pp. 176-180, Barcelona, July 26-29, 2009.

[23] A.R. Chaudhuri, A.K. Mandal, B.B. Chaudhuri, "Page Layout Analyser for Multilingual Indian Documents", Proceedings of the *Language Engineering Conference*, 2002.

[24] P. Nagabhushan, S. Nirmala, "Text Extraction in Complex Color Document Images for Enhanced Readability", *Intelligent Information Management*, 2 , pp. 120-133, 2010.

[25] T.V. Hoang, S. Tabbone, "Text Extraction From Graphical Document Images Using Sparse Representation", *International Workshop on Document Analysis Systems*, pp. 143-150, June 9-11, 2010.

[26] S. Grover, K. Arora, S. K. Mitra, "Text Extraction from Document Images using Edge Information", IEEE *India Conference* (INDICON), pp. 1-4, Gujarat, Dec. 18-20, 2009.

[27] S.S. Bukhari, T.M. Breuel, F. Shafait, "Textline Information Extraction from Grayscale Camera- Captured Document Images", ICIP Proceedings of the 16th IEEE *International Conference on Image Processing*, pp. 2013 – 2016, Cairo, Nov. 7-10, 2009.

[28] Y.J. Song, K.C. Kim, Y.W. Choi, H.R. Byun, S.H. Kim, S.Y. Chi, D.K. Jang, Y.K. Chung, "Text Region Extraction and Text Segmentation on Camera-captured Document Style Images", Proceedings of the 2005 Eight *International Conference on Document Analysis and Recognition,* pp. 172-176, Aug. 29-Sept. 1, 2005.

[29] H.M. Suen, J.F. Wang, "Text string extraction from images of colour-printed documents", IEEE Proceedings of *Vision, Image and Signal Processing*, 143, pp. 210- 216, 1996.

[30] A. Negi, N. Kasinadhuni, "Localization and Extraction of Text in Telugu Document Images", Proceedings of the 7th *International Conference on Document Analysis and Recognition*, pp. 749-752, Oct. 15-17, 2003.

[31] Q. Yuan, C.L. Tan, "Text Extraction from Gray Scale Document Images Using Edge Information", Proceedings. Sixth *International Conference on Document Analysis and Recognition,* pp. 302-306, Washington, Sept. 10-13, 2001.

[32] S.V. Seeri, S. Giraddi, Prashant B.M, "A Novel Approach for Kannada Text Extraction", Proceedings of the *International Conference on Pattern Recognition, Informatics and Medical Engineering*, pp. 444-448, Tamil Naidu, Mar. 21-23, 2012.

[33] Nicholas Journet, Rémy Mullot, Véronique Eglin, Jean-Yves Ramel, "Document Image Characterization Using a Multiresolution Analysis of the Texture: Application to Old Documents", *International Journal on Document Analysis and Recognition*, pp.9-18, Springer Verlag, 2008.

[34] M. Mehri, P. Gomez-Krämer, P. Héroux, A. Boucher, and R. Mullot, "A Texture-based Pixel Labeling Approach for Historical Books". Pattern Analysis and Applications, Springer-Verlag, pages 1-40, 2015.

[35] H. Kawano, H. Orii, H. Maeda, N. Ikoma, "Text Extraction from Degraded Document Image Independent of Character Color Based on MAP-MRF Approach", IEEE *International Conference on Fuzzy Systems*, pp. 165-168, Jeju Island, Aug. 20-24, 2009.

[36] W. Boussellaa, A. Bougacha, A. Zahour, H.E. Abed, A. Alimi, "Enhanced Text Extraction from Arabic Degraded Document Images using EM Algorithm", 10th *International Conference on Document Analysis and Recognition*, pp. 743-747, Barcelona, July 26-29, 2009.

[37] K. Sobottka, H. Bunke, H. Kronenberg, "Identification of Text on Colored Book and Journal Covers", *International Conference on Document Analysis and Recognition*, pp. 57-62, Bangalore, Sept. 20-22, 1999.

[38] Keechul Junga, Kwang In Kim, Anil K. Jain, "Text information extraction in images and video: a survey", Pattern Recognition, 37, pp. 977-997, 2004.