Text Extraction from Document Images- A Review

Deepika Ghai
E & EC Deptt.
PEC University of Technology
Chandigarh, India

Neelu Jain
E & EC Deptt.
PEC University of Technology
Chandigarh, India

ABSTRACT
Text extraction in an image is a challenging task in the computer vision. Text extraction plays an important role in providing useful and valuable information. This paper discusses various approaches such as Adaptive Local Connectivity Map (ALCM), Expectation Maximization (EM), Maximization Likelihood (ML), Markov Random Field (MRF), Spiral Run Length Smearing Algorithm (SRLSA), Curvelet transform etc. for extracting text from scanned book covers, journals, multi-color document, handwritten document, ancient document and newspaper document images. Text line segmentation is a major component for document image analysis. Text in documents depend upon various factors such as language, styles, font, sizes, color, background, orientation, fluctuating text lines, crossing or touching text lines. This paper provides performance comparison of several existing methods suggested by researchers in document text extraction on the basis of recall rate, precision rate, processing time, accuracy etc.

Keywords
Optical Character Recognition (OCR), Morphological Component Analysis (MCA), Undecimated Wavelet Transform (UWT), Discrete Wavelet Transform (DWT), Connected Component Analysis (CCA), Adaptive Local Connectivity Map (ALCM), Expectation Maximization (EM), Maximum Likelihood (ML), Spiral Run Length Smearing Algorithm (SRLSA), Resolution Enhancement (RE), Markov Random Field (MRF), Maximum A-posteriori Probability (MAP), Block Energy Analysis (BEA), Support Vector Machine (SVM), Thin Line Coding (TLC), Constrained Run Length Algorithm (CRLA).

.

1. INTRODUCTION
Text in images contains meaningful and useful information which can be used to fully understand the contents of the images. Text extracted from images play an important role in document analysis, vehicle plate detection, video content analysis, document retrieval, blind and visually impaired users etc. A document image contains various information such as texts, pictures and graphics i.e., line-drawings and sketches. They are developed by scanning journals, historical document images, degraded document images, handwritten text images, printed document images, multi-color book covers and newspaper etc. There are many challenges which are faced in scanned documents that are low contrast, low resolution, color bleeding, complex background and unknown text color, size, position, orientation, layout etc. as shown in Fig. 1. In case of multi-color document image, extraction of text is difficult and unreadable due to color mixing of foreground text with the background as shown in Fig. 2. Even very good quality OCR system performs poorly during text extraction in case of problems such as low resolution, complex background, blurring, multiple layouts and stray marks or voids kind of images. Generally OCR works on clean background images.

Handwritten historical document collected from library and museums around the world also face many challenging problems in its analysis such as degraded images, low quality images, fluctuating and crossing text lines as shown in Fig. 3. Text line identification for handwritten document like text on cards, papers etc. is more challenging task as compared to printed text document. Problems such as skew, fluctuating lines, touching or crossing lines, presence of small symbol between text lines, crowded writing etc. are faced in handwritten documents as shown in Fig. 4. Newspaper is a document with multiple layouts as shown in Fig. 5.

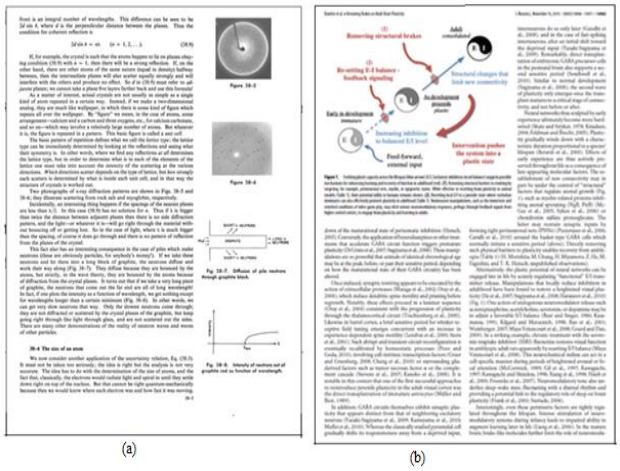


Fig. 1: Grayscale document images: (a) Single-column text from a book, (b) Two-column text from journal

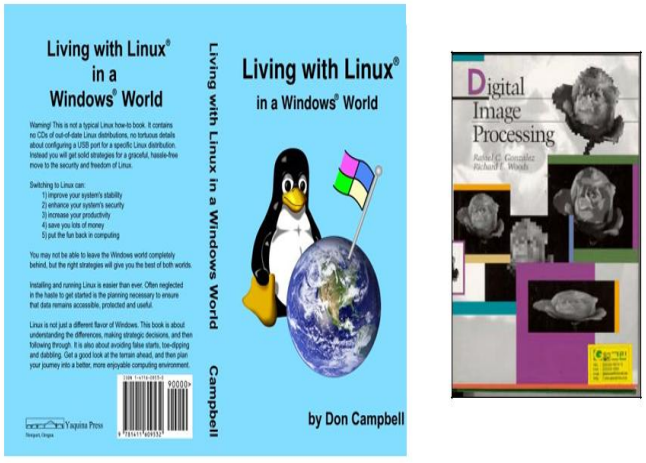


Fig. 2: Multi-color book cover images

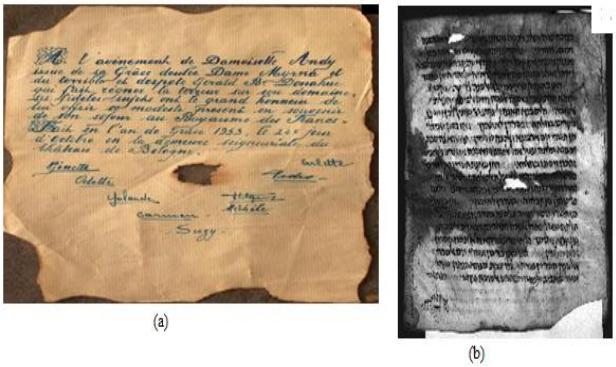


Fig. 3: Historical document images: (a) Ancient document image, (b) Degraded image

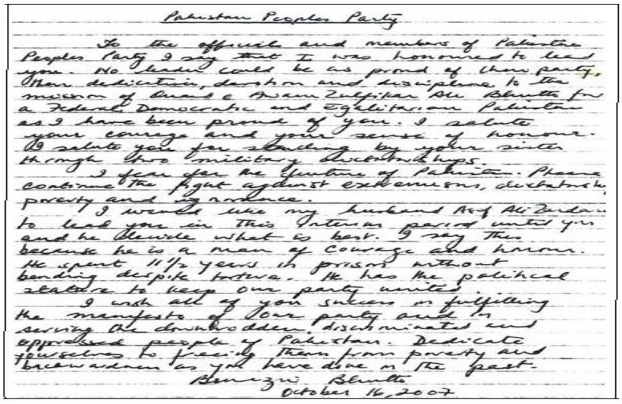


Fig. 4: Handwritten text document image

Fig. 5: Newspaper document image

2. REVIEW OF TEXT EXTRACTION TECHNIQUES

A large number of techniques have been used for text extraction of document images. These are discussed as follows

Anupama et al.[1] suggested an algorithm based on multiple histogram projection using morphological operators to extract text from the Telugu document image. Morphological operator is used to remove noise. Binary image is produced from gray-scale image by applying threshold value. Then, histogram in y-direction and x-direction are obtained in order to perform line segmentation and word segmentation respectively as shown in Fig. 6. In 2012, Malakar et al.[2] proposed a spiral run length smearing algorithm (SRLSA) for text extraction from handwritten document pages. Firstly, document image is divided into number of equal width vertical portions. Then, SRLSA is applied to identify all text line segments present in vertical portions. Finally, text line segment are analysed and merged in order to form text lines present in handwritten document image. Ha et al.[3] proposed an integral image approach for fast text line extraction in document image instead of binary image. Firstly, document image is converted into integral image. Digital filter (like Haar wavelet) are used to detect the text region in the integral image. Next, non- maximum suppression technique is applied to the location to find the center points in text regions. Finally, the centre points are grouped together to extract text lines as shown in Fig. 7.

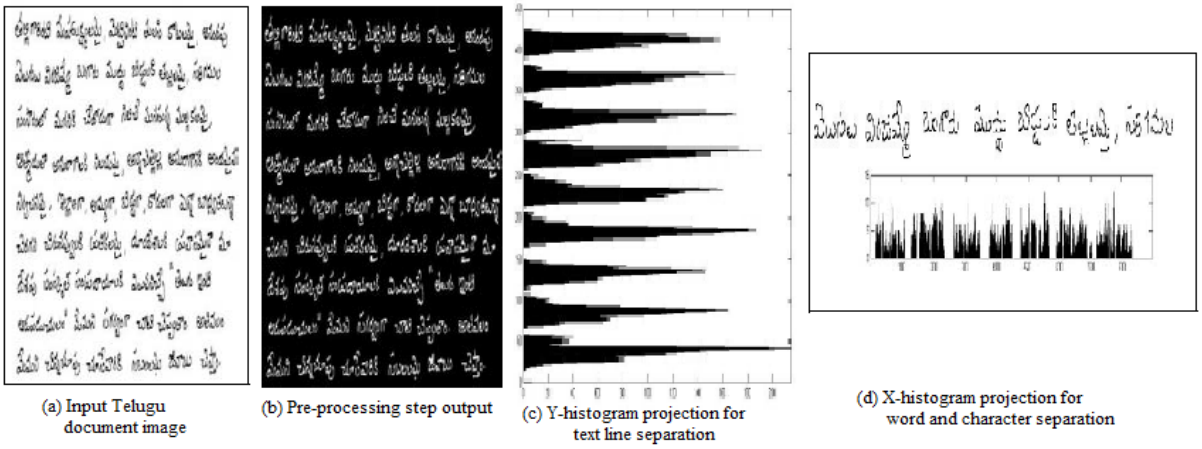
Seeri et al.[4] proposed an edge detection approach for Kannada text extraction. In pre-processing stage, input color image is converted to gray-scale image. This image is then filtered by using median filter to remove noise. Next, sobel edge detector is applied to extract strong edges and separate the background from the object. Edges are mapped to connect edges of same object. Order static filter is used to merge the neighbouring edges to form a single text region. Some of the non-text region is removed based on the structural property of the text i.e. specific height and width of the text. Finally, binary image is processed, labelled and text is extracted.

In 2011, Li et al.[5] proposed an effective interpolation-based resolution enhancement (RE) algorithm for low resolution document images. Firstly, low resolution (LR) image is obtained by multiple observations of same document with small camera motion between these documents. Image registration algorithm aligns many low resolution document images in a sequence. It generates geometric transformation parameters such as translation shifts and rotation angle. Then, Interpolation algorithm maps each LR image onto high resolution (HR) grid according to its geometrical transformation. Further, an iterative weighted average process is used to fill in holes of the interpolated HR image with their neighbourhood values. Finally, the reconstructed HR document is passed as input to OCR for character recognition. Zaravi et al.[6] proposed a method based on wavelet transform to extract text from colored book cover and journals. Firstly, wavelet transform decomposes the image into three detail sub-bands i.e. horizontal, vertical and diagonal. Next, edges were detected by using dynamic thresholding from these detail sub-bands. Region of interest (ROI) was applied to achieve binary image. Finally, text boxes were extracted with projection profile.

In 2010, Hoang et al.[7] proposed a text extraction method from graphical document images by applying morphological component analysis (MCA). MCA allows the separation of features contained in an image by promoting sparse representation of these features in two chosen dictionaries. Curvelet transform and Undecimated Wavelet Transform (UWT) dictionary is used for graphics and text respectively. The problem of touching between text and graphics is overcomed by this approach and is applicable to all images having different font sizes, styles, orientation etc. Nagabhushan et al.[8] suggested a hybrid approach which was combination of Connected Component(CC) and texture feature analysis for text extraction in complex background color document images. In this method, canny edge detector detects the edges. After dilation operation on the edge image,

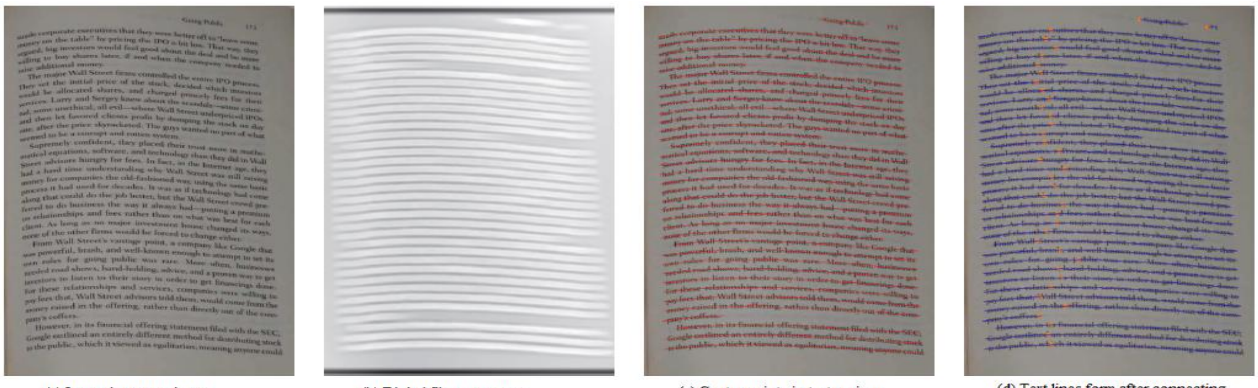
holes were created in the connected component. Connected component without holes (non-text region) are removed. Other non-text components are removed by computing and analysing standard deviation of each connected component. The text from the background is separated by using local thresholding. The noisy text region are identified and reprocessed to further improve the readability of the retrieved foreground text as shown in Fig. 8.

In 2009, Koppula et al.[9] suggested a method for extraction of text from Indian document images of Telugu script. In this method, vertical spatial relation and nearest neighbour algorithm is used to create relationship between connected component and connected component are clustered to form text line. In word segmentation, the space between nearby character is computed and clustered into word. Santos et al. [10] proposed a reliable algorithm to segment handwritten text on the basis of morphology and histogram projection. Firstly, Morphological operator is used to extract the feature of the input image and produces a binary image. Then, text line positions are detected by Y-histogram projection. Thresholding is applied for text line separation to remove noises and false text lines. The line region recovery step recovers some losses of the text line area. Further, X-histogram projection is applied in horizontal direction to detect word separation. Another threshold is applied in X-direction to remove false words. Finally, text line regions are obtained. Kawano et al.[11] formulated a background and foreground estimation algorithm in Markov Random Field (MRF) and Maximum A-Posteriori Probability (MAP) approach for binarization of degraded document images. Firstly, MAP estimator gives a solution which maximizes a-posteriori probability for a given gray-scale image. MRF shows that a conditional probability at a pixel depends only on the neighbouring pixels. Median filter functions as background estimator and it removes noise. After subtraction of median filtering image from gray-scale image, the degree of irregularity in the original image is calculated with text estimator.



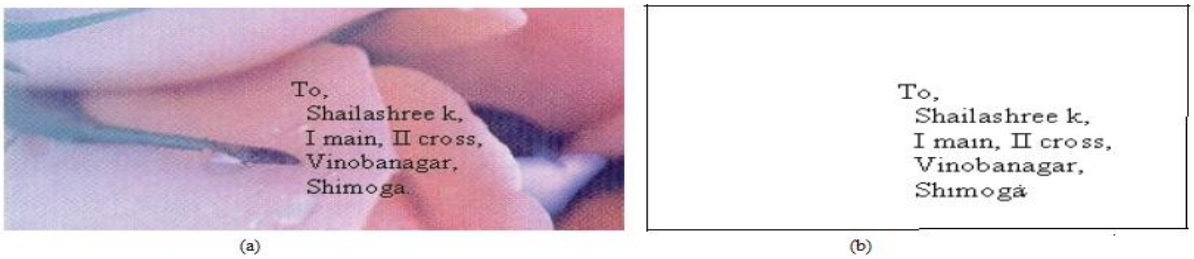
(a) Input Telugu document image (b) Pre-processing step output (c) Y-histogram projection for text line separation (d) X-histogram projection for word and character separation

Fig. 6: Multiple Histogram projection method



(a) Input document image (b) Digital filter response (c) Center points in text regions (d) Text lines form after connecting center points

Fig. 7: Integral image method for fast text line extraction



(a) (b)

Fig. 8: Hybrid method (a) Input color image (b) Text separation from complex background

Grover et al.[12] proposed edge based feature for detection of text embedded in complex colored document images. The complex colored image was converted into gray-scale image. Then edge detection was performed on the gray-scale image by convolving the image with sobel mask, separately for horizontal and vertical edges. Further, weak edges are removed with thresholding. After this, the edge image is divided into small non-overlapping blocks of pixels, where depends on image resolution. Then, block classification is done using pre-defined threshold which will distinguish text from the image. Audithan et al.[13] described an effective method for text extraction from document images using Haar discrete wavelet transform (DWT). 2D-Haar DWT detect all the horizontal, vertical and diagonal edges. The non-text edges are further removed by using thresholding technique. Then, morphological dilation operators are used to connect the text edges in each detail component. Further, logical AND operator are used to remove the remaining non-text regions.

Bukhari et al.[14] proposed a method for curled textline detection from gray-scale camera captured document images. The gray-scale textline was promoted by using multi-oriented multi-scale anisotropic Gaussian smoothing. Ridges were used for detection of curled textline region. This method is based on differential geometry, which used local direction of gradient and derivative to measure curvature. Hessian matrix is used for finding direction of gradient and derivative. This information helped in detection of ridges. For estimating X-line and baseline pairs from detected textline, modified coupled snake model was used. Boussellaa et al.[15] proposed an expectation maximization (EM) algorithm for text extraction from gray-scale and colored degraded document images on the basis of probabilistic models. Color document image was changed to YIQ color image and operated on Y luminance channel. The initial parameters like mean and standard deviation of EM algorithm are calculated by using k-means clustering algorithm. After the parameter estimation, the document image is partitioned into text and background classes by means of maximum likelihood (ML) approach.

Shi et al.[16] suggested a method based on adaptive local connectivity map (ALCM) using steerable directional filter for extracting handwritten Arabic text. Firstly, steerable filter is applied to the document image and it is converted into ALCM. Next, adaptive thresholding is applied on ALCM to get binarized image which gives a rough estimate of text line location. Connected component analysis (CCA) is used for grouping the connected component into location mask for each text line. The text lines are extracted by superimposing the text line pattern in ALCM on the original document image as shown in Fig. 9.

Sarkar et al.[17] described bottom-up approach of line segmentation from handwritten Bangla text. Firstly, the input document image is divided into number of squares $(n \times n)$ pixels each. If there are more than 50% black pixels in a square, the square is totally filled with black pixels to achieve smoothing image. When a white square is surrounded by three or more black squares, it is also blackened to have better smoothing. Then, the height of each component of the smooth image is calculated with the help of Gaussian distribution. Further, a rectangular block is created with width and height and its total number of black pixels is counted. If this counted number is greater than number of previous position of the block, then the previous information is discarded. Otherwise, the previous position of block is taken as centreline or mid-line. This process is continued till the scanning of whole image is completed. Finally, text line is extracted by joining and linking of mid-lines.

In 2006, Qiao et al.[18] suggested a Gabor filter based method to extract text from document images. Firstly, the document image is passed through Gabor filter, the Gabor filtered image is processed and fused at different orientations and scales to extract text regions directly in the form of rectangle boundary. Text regions are quite rich in the middle and have high frequency component. Next, binarization operation is applied to eliminate some non-text region. Mending operation is performed to connect gap between small regions such as characters and words. Two parameters i.e. standard rate (SR) and high frequency content (HFC) are performed to make text region separable from non-text region. Lemaitre et al.[19] suggested a method for text line extraction in handwritten document with Kalman filter applied on low resolution images. Text line extractor based on the theory of Kalman filtering is used to detect skewed, curved or overlapped text lines. This method was also used to extract text from ancient damaged documents of the 18^{th} and 19^{th} century.

In 2005, Shi et al.[20] proposed an algorithm using adaptive local connectivity map (ALCM) for text extraction from complex handwritten historical document. The gray-scale document image is transformed into ALCM. Thresholding is applied on ALCM to give the text line pattern in terms of connected component. By using grouping algorithm, the connected components are grouped into location masks for each text line. Text line was extracted by mapping location mask back onto binary image to collect the text line components. Further, splitting algorithm is applied to overcome touching multiple lines problem. Song et al.[21] suggested a text segmentation approach based on color clustering method for camera captured document images. The two level (low and high) features are combined hierarchically. This method detects text region from the images using two low level features i.e. intensity and color variation. Text candidate regions obtained in each method are combined .Text region is verified through high level text stroke feature. When large region is overlapped, it is considered as a text region and no verification is necessary. When no or small overlapping exists, verification is done with Support Vector Machine (SVM). Binarization is done with k-means clustering method.

In 2004, Raju et al.[22] formulated a texture based and connected component analysis approach for text extraction from document image. Firstly, anisotropic diffusion filtering is performed to clear background noise. Then, the noise free document image is passed through Gabor filter, the Gabor filtered image is processed and fused at different orientations and scales to extract text regions directly. Text regions are quite rich in the middle and have high frequency component. Next, binarization is done on document image by selecting local threshold value and binarized image is fed to connected component analysis (CCA). CCA distinguishes the text and graphics portion based on the size of the component. Block energy analysis (BEA) separates text and non-text area.

In 2003, Negi et al.[23] proposed top-down and bottom-up approach for localisation and extraction of text in Telugu document images. Firstly, sobel operator obtains the gradient magnitude of the gray-scale image. By applying threshold and median filter, image is binarized to clarify the text and noise is removed respectively. Next, Hough transform is applied to locate circular features and fill them. Recursive XY Cuts

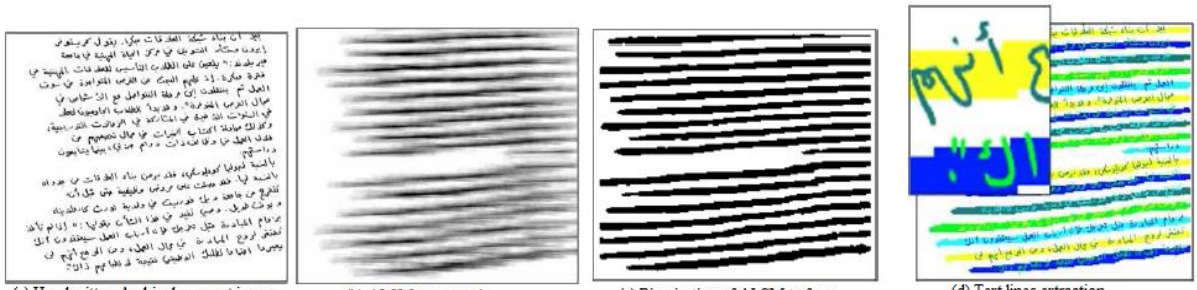
(RXYC) divides the region into paragraphs, lines and words region. Bounding box are generated around each word and are binarized using histogram based binarization method to highlight the text. This method finds an optimum threshold value by finding a valley between two peaks which represent object and background. Further, connected component merging is done to gather individual words. The final output is passed as input to OCR for character recognition as shown in Fig.10.

In 2002, Chaudhuri et al.[24] proposed a page layout analyser for multi-column Indian document languages such as Bangla and Devanagari (i.e. Hindi). Firstly, gray-scale image is converted into binary image by thresholding. Next, the component labelling i.e. (i) headline positioning, (ii) headline zoning, (iii) upper, middle and lower part zoning is performed by recursive neighbourhood search and tagging. On the basis of headline positioning and zoning, bounding boxes are formed. Further, block absorption and block merging is applied and bigger block is generated.

In 2001, Yuan et al.[25] proposed a method using edge information to extract text from gray-scale document images i.e. heavy noise infected newspaper. Firstly, canny edge detector is used to detect edges from gray-scale image. Next, edge merging is performed by connecting the small horizontal edge component to form longer lines. There is larger spatial variance in text as compared to background. Then, Thin Line Coding (TLC) is applied to eliminate noise and false segment from image. TLC operates on chains and lines instead of pixels. Finally, parallel straight line segments are grouped to form bounding box in which text is enclosed.

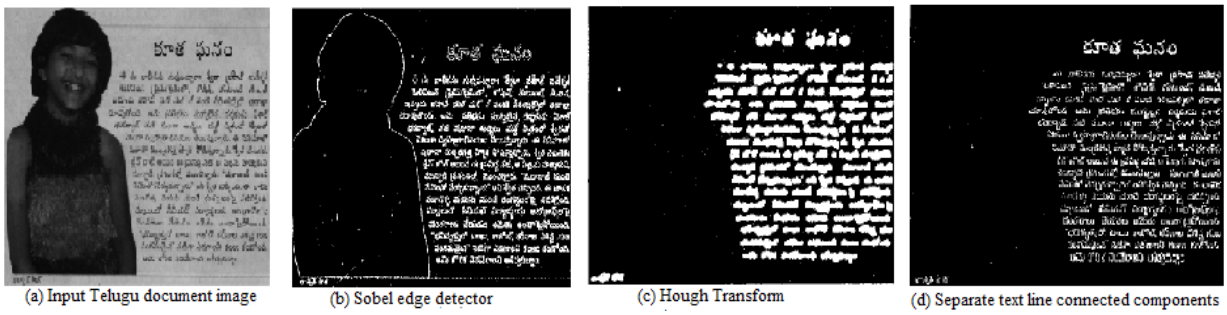
In 1999, Sobottka et al.[26] proposed a method based on top-down and bottom-up analysis for text location and identification on colored book and journals. Firstly, clustering algorithm was applied to reduce the color variations. In top-down analysis, splitting of image takes place in horizontal and vertical direction alternatively and the output is obtained in the form of rectangular blocks. The bottom-up analysis detects homogeneous region using a region growing method. Beginning with a start region, pixels are merged if they belong to the same cluster. Finally text and non-text regions are separated as shown in Fig. 11.

In 1996, Suen et al.[27] proposed a method for extracting text strings from color printed document in a 24-bit true color image. The processing is time consuming due to very large amount of data in 24-bit color image. Firstly, sobel edge detector transforms the color document image into binary image. Then, constrained run length algorithm (CRLA) divides document image into blocks, each of which contains only one type of objects i.e. text or graphics. This technique is applied row by row as well as column by column on binary image. Finally, all the text lines are segmented successfully. The final image is a white background/ black text binary image.



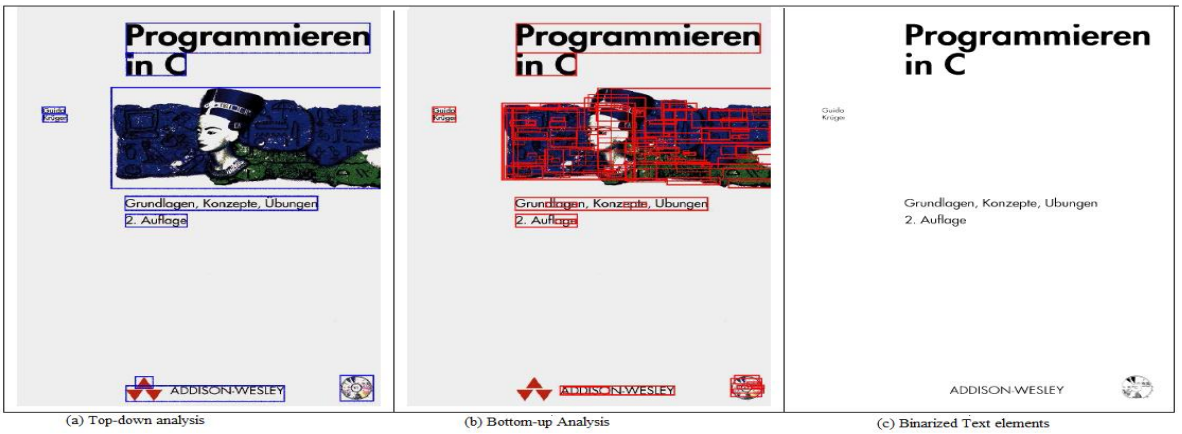
(a) Handwritten Arabic document image with varying skew (b) ALCM generated (c) Binarization of ALCM to form text lines patterns (d) Text lines extraction

Fig. 9: Adaptive Local Connectivity Map (ALCM) method



(a) Input Telugu document image (b) Sobel edge detector (c) Hough Transform (d) Separate text line connected components

Fig. 10: Top- down and Bottom-up method for text localization and extraction in Telugu document image



(a) Top-down analysis (b) Bottom-up Analysis (c) Binarized Text elements

Fig. 11: Clustering method

3. COMPARATIVE ANALYSIS

The detailed analysis of text extraction techniques used for document images is shown in Table 1.

Table 1: comparison of document images

Author	Year	Approach	Features	Images	Parameters
Anupama et al.[1]	2013	Morphology operators, Histogram Projection (X and Y histogram)	(1) Insensitive to skew. (2) Fail in case of touching characters and overlapping lines.	Handwritten Telugu document images.	Detection rate (98.54%), Recognition Accuracy (98.29%)
Malakar et al. [2]	2012	Spiral Run Length Smearing Algorithm (SRLSA), partitioning based approach	(1) Insensitive to skew and languages. (2) Fails due to presence of touching or overlapping of text lines. (3) Undersegmentation and oversegmentation of text lines may occur.	Handwritten document pages written in Bangla script with English words.	Accuracy (89.35%)
Ha et al.[3]	2012	Digital filter using Haar wavelet, Non-maximum suppression technique.	Insensitive to shadow (or reflection), font size and layouts.	Document images	Mean rate (95.29%), Mean time (1.17 sec), variance (0.11)
Seeri et al.[4]	2012	Median filter, sobel edge detector, connected component labelling, order static filter.	(1)Robust to font size, small orientation and alignment of text. (2) Fails to extract very small characters. (3) Fails to identify the characters due to disjointness.	Kannada text images	Precision rate (84.21%), Recall rate (83.16%), Accuracy (75.77%)
Li et al.[5]	2011	Interpolation-based resolution enhancement (RE) algorithm, Iterative weighted average method.	(1)Insensitive to font size. (2) Fail when shadow or light present in the images.	Low resolution document images.	Recognition error rate (RER) (4%), Run time (1.04 sec)
Zaravi et al. [6]	2011	Discrete Wavelet Transform, Dynamic thresholding, Region of Interest (ROI)	Robust to noise	Colored book and journal cover sheets	Bit rate (91.20%)
Hoang et al. [7]	2010	Sparse Representation, Morphological Component Analysis (MCA), Curvelet Transform, Undecimated Wavelet Transform (UWT), Adaptive thresholding	(1) Overcome the touching problem between text and graphics. (2) Insensitive to different font size, styles, orientation.	Graphical document images like engineering drawings sheets.	Recall Rate (96.2%)
Nagabhushan et al. [8]	2010	Hybrid approach (connected component (CC) and texture analysis), canny edge detector, dilation operation, local	(1) Tilted line detection. (2) Foreground text separation from complex background. (3)Edge detection in gray-	Color document images, printed postal address images.	Character recognition rate (CRR) (98.53%) in text rich document and CRR (83%) in printed postal

		thresholding	scale document images resulted in loss of text edge pixels to certain extent. (4) Fails to separate foreground text when contrast between foreground and background is poor. (5) Fails to detect single letter word which neither contain a hole nor created hole by dilation.		document, Accuracy (97.12%).
Koppula et al. [9]	2009	Connected component Analysis, vertical spatial relation and nearest neighbour algorithm.	(1) Works well on good quality (good print), high text density documents. (2) Errors for degraded (broken and touching characters) documents.	Telugu document images	Extraction rate (99.86%) for good quality, (98.1%) for high text density and (89.9%) for degraded document images.
Santos et al. [10]	2009	Morphology-based, Histogram Projection, Dynamic thresholding.	Insensitive to font, size, style, color, orientation.	Handwritten document.	Detection rate (98%), Missed detection rate (2 %)
Kawano et al. [11]	2009	Markov Random Field (MRF), Probabilistic model (Gaussian distribution probability), Median filter.	Used for (1) black text and white background, (2) white text and black background.	Degraded gray-scale document images.	
Grover et al. [12]	2009	Sobel edge detector, thresholding, block classification	(1) Insensitive to color, different fonts and languages. (2) Fails when gradient of intensity of text and image are similar.	Document Images	Sensitivity (99%), false alarm rate (4%)
Audithan et al.[13]	2009	Haar DWT, Morphological Dilation operator, logical AND operator, Dynamic thresholding	Independent of contrast	Document images	Detection rate (94.8 %)
Bukhari et al. [14]	2009	Multi-oriented multi-scale Anisotropic Gaussian smoothing, Ridges based on differential geometry, Hessian matrix, X-line Baseline Pairs Estimation: Snakes	(1) Accurately track x-line and baseline of curled textlines even in the presence of large number of capital letters, numbers and quotation marks. (2) Robust to high degree of curl.	Gray-scale camera-captured document images.	Accuracy (91%)
Boussellaa et al. [15]	2009	Expectation Maximization (EM) algorithm, k-means clustering method, Maximum likelihood (ML) segmentation method.	Work well on old degraded document image.	Color and gray-scale Arabic degraded document images.	Precision (96%), Recall (93%), mean error (0.9%)
Shi et al. [16]	2009	Steerable filter, Adaptive Local Connectivity Map (ALCM), Adaptive thresholding, Connected Component Analysis (CCA).	Overcome the problem of fluctuating, touching or crossing text lines.	Handwritten Arabic text lines	Accuracy (99.5%)
Sarkar et al. [17]	2009	Bottom-up approach, Graphical Smoothing Algorithm (GSA) along with block creation.	Insensitive to skew, touching or overlapping of handwritten text.	Bangla handwritten text	Accuracy (92%)
Qiao et al. [18]	2006	Gabor filter, thresholding, Mending operation, Standard Rate (SR), High Frequency Content (HFC).	Insensitive to different styles, languages, fonts and skew.	Chinese document images.	
Lemaitre et al.[19]	2006	Kalman filter	Insensitive to noise, skew, touching or overlapping lines.	Low resolution handwritten document images	
Shi et al.[20]	2005	Adaptive Local Connectivity Map (ALCM),	(1) Insensitive to fluctuating, crossing or touching text	Gray-scale historical document	Accuracy (95%)

		Grouping algorithm, Splitting algorithm.	lines. (2) Fails when images with severe damage, noise and low visual readability. (3) Fails to short text lines such as abbreviation and page numbers.	images	
Song et al. [21]	2005	Intensity variation and color variance method, canny edge detection, Support Vector Machine (SVM), Multi-resolution wavelet transform, k-mean color clustering method.	Insensitive to fonts, styles, sizes, colors and skews.	Camera-captured document style images	Precision rate (95.86%), Recall rate (92.98%), processing time (1.2968 sec)
Raju et al. [22]	2004	Anisotropic filtering, Gabor filter, connected component analysis (CCA), thresholding, Block Energy Analysis (BEA).	Insensitive to script, font size, layout, noise, style and skew.	Complex, scanned and camera-captured document images, newspapers, books, handwritten text images.	Processing time (30 sec)
Negi et al. [23]	2003	Sobel operator, Median filter, Thresholding, Hough Transform, Recursive XY Cuts (RXYC) approach, Histogram based.	Detect the circular nature of the script.	Telugu document images	
Chaudhuri et al.[24]	2002	Thresholding, connected component labelling	Insensitive to font sizes.	Devanagari (Hindi) and Bangla multicolumn document images.	Processing time (1.35 sec)
Yuan et al. [25]	2001	Canny edge detection, adaptive thresholding, Thin Line Coding (TLC) approach.	(1) Insensitive to skew. (2) Fails when documents with multiple layouts. (3) Detection rate decreases when textual paragraphs interwined heavily with irregular graphical blocks.	Gray-scale document image, heavy noise infected newspaper images.	
Sobottka et al.[26]	1999	Clustering algorithm, Top-down and Bottom-up Analysis	(1) Applicable to homogeneous, multi-colored and textured background. (2) Insensitive to color, font, size, style and noise. (3) Undersegmentation and oversegmentation may occur for characters	Colored books, Journal covers web and video images.	Processing time of top-down analysis is 1.05 sec amd bottom-up analysis is 20.26 sec.
Suen et al. [27]	1996	Sobel edge detector, Constrained Run Length Algorithm (CRLA), logical AND operator, connected component labelling.	Insensitive to sizes, fonts and colors.	Chinese and English color printed document images	Processing time (72 sec)

4. CONCLUSION

In this paper, various text extraction techniques such as Adaptive Local Connectivity Map (ALCM), Expectation Maximization (EM), Maximum Likelihood (ML), Spiral Run Length Smearing Algorithm (SRLSA) etc. have been discussed. The performance comparison of these methods for document text extraction on the basis of accuracy, precision rate, recall rate, processing time has been done. It is observed that accuracy (98.53%) is best for hybrid (connected component (CC) and texture Analysis) approach and edge based text extraction techniques. Precision (96%) and recall rate (93%) is best in case of EM algorithm and ML segmentation method. Processing time (1.17 sec) is best in case of digital filter using Haar wavelet. For handwritten text document images, accuracy (99.5%) is best for ALCM and histogram projection based method.

5. REFERENCES

[1] N. Anupama, C. Rupa, E.S. Reddy, *Character Segmentation for Telugu Image Document using Multiple Histogram Projections*, Global Journal of Computer Science and Technology Graphics and Vision, 13 (2013) 11-16.

[2] S. Malakar, S. Halder, R. Sarker, N. Das, S. Basu, M. Nasipuri, *Text line Extraction from Handwritten Document pages using spiral run length smearing algorithm*, International Conference on communications, Devices and Intelligent Systems, Kolkata, Dec. 28-29 (2012) 616-619.

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

[4] 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, Tamil Naidu, Mar. 21-23 (2012) 444-448.

[5] Z. Li, J. Luo, *Resolution Enhancement from Document Images for Text Extraction*, 5th International Conference on Multimedia and Ubiquitous Engineering, Loutraki, June 28-30 (2011) 251-256.

[6] 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 (2011) 686-689.

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

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

[9] 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, Dec. 16-18 (2009) 269-272.

[10] R.P.D. Santos, G.S. Clemente, T.I. Ren, G.D.C. Calvalcanti, Text Line Segmentation Based on Morphology and Histogram Projection, 10th International Conference on Document Analysis and Recognition, Spain, July 26-29 (2009) 651-655.

[11] H. Kawano, H. Orii, H. Maeda, N. Ikoma, *Text Extraction from Degraded Document Image Independent of Character Color Based on MAP-MRF Approach*, IEEE, Jeju Island, Aug. 20-24 (2009) 165-168.

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

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

[14] 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, Cairo, Nov. 7-10 (2009) 2013 – 2016.

[15] 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, Barcelona, July 26-29 (2009) 743-747.

[16] 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, Barcelona, July 26-29 (2009) 176-180.

[17] D. Sarkar, R. Ghosh, *A Bottom-Up Approach of Line Segmentation from Handwritten Text*, (2009).

[18] 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, USA, (2006) 297-300.

[19] 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, Lyon, April 27-28 (2006) 45-52.

[20] Z. Shi, S. Setlur, V. 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, Aug, 29- Sept. 1 (2005) 794-798.

[21] 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, Aug. 29-Sept. 1 (2005) 172-176.

[22] S. Raju S, P.B. Pati, A.G. Ramakrishnan, *Gabor Filter Based Block Energy Analysis for Text Extraction from Digital Document Images*, Proceedings of the Ist International Workshop on Document Image Analysis for Libraries, (2004) 233-243.

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

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

[25] Q. Yuan, C.L. Tan, *Text Extraction from Gray Scale Document Images Using Edge Information*, Washington, Sept. 10-13 (2001) 302-306.

[26] K. Sobottka, H. Bunke, H. Kronenberg, *Identification of Text on Colored Book and Journal Covers*, Document Analysis and Recognition, Bangalore, Sept. 20-22 (1999) 57-62.

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