Text Graphic Separation In Indian Newspapers

Ritu Garg
Deptt. of EE
IIT Delhi, India
ritu2721a@gmail.com

Anukriti Bansal
Deptt. of EE
IIT Delhi, India
anukriti1107@gmail.com

Santanu Chaudhury
Deptt. of EE
IIT Delhi, India
schaudhury@gmail.com

Sumantra Dutta Roy
Deptt. of EE
IIT Delhi, India
sumantra.dutta.roy@gmail.com

ABSTRACT

Digitization of newspaper article is important for registering historical events. Layout analysis of Indian newspaper is a challenging task due to the presence of different font size, font styles and random placement of text and non-text regions. In this paper we propose a novel framework for learning optimal parameters for text graphic separation in the presence of complex layouts. The learning problem has been formulated as an optimization problem using EM algorithm to learn optimal parameters depending on the nature of the document content.

Categories and Subject Descriptors

I.7.0 [**Computing Methodologies**]: Document and Text Processing - General; 1.5.4 [**Computing Methodologies**]: Pattern Recognition—*Application*

General Terms

Text Document Image Classification System

Keywords

Text Graphic Separation, Indian Newspaper, Complex Layout, Parameter Estimation

1. INTRODUCTION

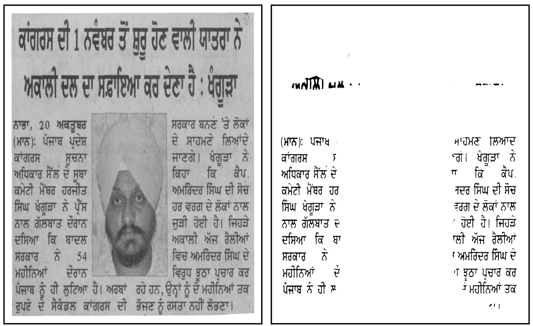
The process of converting physical document page into digital format is important for its preservation and archival. Digitization of books for search and indexing applications has been extensively studied. Similar work on newspaper articles can be done along with other upcoming applications like aligning TV newscast and newspaper reports for effective retrieval of related articles across different media.

A major step in digitization of document images is identification of homogeneous regions of text and graphics. Since the text carries most of the information it becomes necessary to locate text within the document image and recognize it to extract hidden information. Text graphic separation has been the most challenging problem in document image analysis. In literature, methods for text graphic separation have mostly focused on document images with simple manhattan layouts. These methods tend to fail on complex layout document images such as newspapers, where the text and graphic components are randomly positioned and wide variation in font size and style exists. Utilizing texture properties alone is not enough for obtaining satisfactory results. Exploiting content specific information i.e. features based on script characteristics [24] can be readily used in combination to distinguish between text and graphics. Hence, combining contextual and content related information over a local neighbourhood can provide an effective solution to text graphic separation.

Most of the existing work on text graphic segmentation show satisfactory results on document images available in Latin scripts. However, these algorithms may not be effective for document images in Indic scripts due to script complexity. Indian scripts can be visually discriminated by observing the curliness in the script or the presence of horizontal line at the top of the words along with other dominant vertical strokes. Scripts like Devanagari, Bangla, Gurumukhi etc use a four ruled standard due to the presence of shiro-rekha. While scripts like Telugu, Tamil, Malayalam etc. lack such standards as the components overlap with the boundaries of the neighbouring lines due to which the block/line level segmentation becomes complex and inefficient. Thus, scripts based features can provide valuable information for locating text regions within a document image. Moreover, most of the text graphic segmentation algorithms rely on one or more parameters for their successful execution. The values assigned to these parameters are usually application specific or based on some heuristics. Hence, estimating optimal values instead will help in generalizing the algorithm as well as provide accurate results.

This paper presents a modification of the earlier work on text graphic separation [1] that exploits the nature of the document image content for learning optimal parameters for binarization and effective text graphic separation. The approach utilizes local spatial texture properties along with the script characteristics to enhance the performance of the text graphic separation. In our earlier work, for segmentation optimal values for P/N ratio were estimated to distinguish between text and graphics using a fixed window

MOCR '13, August 24 2013, Washington, DC, USA
Copyright 2013 ACM 978-1-4503-2114-3/13/08 ...$10.00.



ਕਾਂਗਰਸ ਦੀ 1 ਨਵੰਬਰ ਤੋਂ ਸ਼ੁਰੂ ਹੋਣ ਵਾਲੀ ਯਾਤਰਾ ਨੇ
ਅਕਾਲੀ ਦਲ ਦਾ ਸਫ਼ਾਇਆ ਕਰ ਦੇਣਾ ਹੈ : ਖੰਗੂੜਾ

ਨਾਭਾ, 20 ਅਕਤੂਬਰ (ਮਾਨ): ਪੰਜਾਬ ਪ੍ਰਦੇਸ਼ ਕਾਂਗਰਸ ਸੂਚਨਾ ਅਧਿਕਾਰ ਸੈੱਲ ਦੇ ਸੂਬਾ ਕਮੇਟੀ ਮੈਂਬਰ ਹਰਜੀਤ ਸਿੰਘ ਖੰਗੂੜਾ ਨੇ ਪ੍ਰੈੱਸ ਨਾਲ ਗੱਲਬਾਤ ਦੌਰਾਨ ਦਸਿਆ ਕਿ ਬਾਦਲ ਸਰਕਾਰ ਨੇ 54 ਮਹੀਨਿਆਂ ਦੌਰਾਨ ਪੰਜਾਬ ਨੂੰ ਹੀ ਲੁਟਿਆ ਹੈ। ਅਰਬਾਂ ਰੁਪਏ ਦੇ ਸੈਕੰਡਲ ਕਾਂਗਰਸ ਦੀ ਸਰਕਾਰ ਬਨਣ 'ਤੇ ਲੋਕਾਂ ਦੇ ਸਾਹਮਣੇ ਲਿਆਂਦੇ ਜਾਣਗੇ। ਖੰਗੂੜਾ ਨੇ ਕਿਹਾ ਕਿ ਕੈਪ. ਅਮਰਿੰਦਰ ਸਿੰਘ ਦੀ ਸੋਚ ਹਰ ਵਰਗ ਦੇ ਲੋਕਾਂ ਨਾਲ ਜੁੜੀ ਹੋਈ ਹੈ। ਜਿਹੜੇ ਅਕਾਲੀ ਅੱਜ ਰੈਲੀਆਂ ਵਿਚ ਅਮਰਿੰਦਰ ਸਿੰਘ ਦੇ ਵਿਰੁਧ ਝੂਠਾ ਪ੍ਰਚਾਰ ਕਰ ਰਹੇ ਹਨ, ਉਨ੍ਹਾਂ ਨੂੰ ਦੋ ਮਹੀਨਿਆਂ ਤਕ ਭੱਜਣ ਨੂੰ ਰਸਤਾ ਨਹੀਂ ਲੱਭਣਾ।

(a) (b)

Figure 1: Failure case of using fixed neightbood of size 350×350 on newspaper data

size. However, adapting the same approach did not work for newspaper images because of the following issues 1) random placement of text and graphics 2) close proximity of graphics and text regions 3) wide variation in text font type and style. Figure 1 shows a sample newspaper image where using a fixed neighborhood of size 350×350 results in removal of text surrounding the image. While figure 2(o) shows an improvement in the output when an optimal neighborhood size estimated by the proposed algorithm is applied. In this paper, we modify the earlier approach by adaptively estimating the local neighbourhood size along with the P/N ratio for better identification of text graphics regions in newspaper images. The earlier work was demonstrated on document images from books where the layout is simple as compared to newspapers and the font size and style remains constant throughout. Here we evaluate the performance of the text graphic sepration approach on newspaper articles.

The organization of the paper is as follows. Section 2 gives a brief summary of the related work. Section 3 discusses our approach in details. The experiments performed are shown in Section 4. Finally the conclusion of the paper is discussed in Section 5.

2. RELATED WORK

Over the years several document layout analysis algorithms have been proposed in [4, 19, 21, 26]. Many methods have been proposed to address the problem of text-graphics separation in document images [3, 23, 27, 25, 22]. The most commonly used approach for text/graphic separation in document images [5, 6] is based on connected component analysis. These approachs tend to fail on low resolution and noisy images with complex layout as they lack clear separations. Wavelet based approach [15] works well for identifying halftones, but due to similar spatial characteristics of text and graphics the graphic patterns get classified as text. In [12], authors show use of Gabor filter for text graphic separation. In [2, 28], authors segment the document image into blocks using run-length smearing and classify each block as either text or graphic using local statistical features. Such approaches fail on document images with complex layout where the text and graphic regions are placed in a random fashion. In [13], author use orientation distribution using autocorrelation over a fixed window for locating text graphic in ancient document images. The algorithm fails if the window happens to contain single text line.

The random placement of different components (text, graphics and half-tones) and variation in the gap between components prohibits the application of document image segmentation algorithms on newspapers. The problem of layout analysis on newspaper data has been addressed by few researchers [8, 7, 9, 10, 17, 18, 20, 16]. Gatos *et al* [8] proposed a two stage technique for layout analysis of newspaper page. In first stage, various regions are extracted using smearing and connected component labelling. A rule based approach is applied in second stage to identify various regions. Liu *et al* [16] presented a component based bottom-up algorithm for analysing complex newspaper layout. This algorithm is based on layout rules which are designed heuristically. In [29], presents classification of newspaper image block using textural features. The technique proposed assumes homogeneous rectangular blocks are extracted using RLSA and Recursive XY-cut. The blocks are then classified based on statistical textural features and feature space decision techniques. Most of the work reported have experimented with newspapers available in Latin, Arabic and Chinese scripts. In contrast, very few authors have addressed the text non-text separation for Indian script. In [14], authors use projection-profile based algorithm to separate text blocks from non-text blocks in a Devanagari document. Due to the presence of shiro-rekha the horizontal profile possesses regularity in frequency, orientation and spatial cohesion for text blocks. This helps in separating text blocks from non-text blocks.

In contrast to earlier approaches, we present a technique which can be applied to documents with different layouts. The proposed framework is a modification of our earlier work [1]. The approach utilizes local spatial texture features to adaptively learns parameters for text graphic separation in documents with complex layouts. An EM based estimation framework has been formulated to learn parameters based on the nature of the document image content. The characteristics of the content are represented using Edge Direction Histogram (EDH) features. The evaluation results establish the efficiency of the proposed approach.

3. TEXT GRAPHIC SEPARATION IN NEWSPAPER ARTICLES

We propose an approach that adaptively learns parameters based on the content of the document for text graphic separation in newspaper articles. The approach utilizes the spatial texture properties over a local neighbourhood, the dimensions of which are also one of the parameters that are learned during optimization. Each script is represented as a distribution over edge direction histogram (EDH) features [24]. These distributions are used to model the script characteristics that are learned from the training data using Expectation Maximization (EM) algorithm. We now describe the parameter optimization and text graphic separation methodologies in detail in the following sections.

3.1 Parameter Optimization

EM algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models. Our EM based parameter estimation framework relies on the nature of the document image content. The documents in the training set are ground-truth images in which the regions of the text are marked. Each document from the training set is represented as $(E_i, r_i, W_i(m,n))$, where i denotes the documents in the training set. E_i is the global EDH of the i^{th} ground-truth image responsible for modelling the document content. The edge direction features are computed by convolving the document image with Sobel mask in horizontal and vertical directions. We then compute the direction of edges to generate a histogram of edges with b-bins to obtain a b-dimensional feature vector. EDH bin size b = 32 is chosen empirically. r_i and $W_i(m,n))$ are the parameters to be optimized, where r_i is the P/N ratio computed over a local neighbourhood $W_i(m,n))$ respectively. The EM framework is summarized in table 1.

Table 1: EM based Optimization Framework

```
Input   $X_i = (X_1, X_2......, X_n)$ be n training images
Output  Optimized paramter : $r_i$ and $W_i(m,n))$
Procedure
1) Each training image is represented as $X_i = (E_i, r_i, W_i(m,n))$, $E_i$ is the EDH and $r_i$ and $W_i(m,n)$
   are the parameters to be estimated ($r_i$: P/N ratio and $W_i(m,n)$: local neighbourhood of size (m×n))
2) $OP(X_i)$ gives the optimum $r_i$ and $W_i(m,n)$ for any $X_i$
3) $d(X_j, X_i)$ be the Kullback Leibler divergence between two distribution $(X_j, X_i)$. Where $X_j$ be the latent
   variable for $j^{th}$ document from test set.
4) Iterate E and M steps untill convergence
   4a) E-Step: To begin with, assume some seed value for $r^t$, $W^t(m,n)$ and get the edge descriptor histogram
       $E^t$ to obtain $Xj = (E^t, r^t, W^t(m,n))$.
   4b) M-Step: Obtain $r^{t+1}$ by
       $r^{t+1} = OP\{arg\ min_{i\ \in\ [1,n]} d(X_j, X_i)\}$.
```

3.2 Adaptive Segmentation

For a given gray-scale document image, it is binarized using the method described in [1]. We first perform connected component labelling that provide connected regions of varying sizes. Generally it is observed that the larger connected regions are graphics which should be removed to restrict the further processing of components that are probable text strings. An appropriate threshold is selected based on the relative frequency of occurrence of components as a function of their size. This connected component based processing does not work for all cases, because the graphics may or may not be completely connected.

The adaptive segmentation scheme presented here does not require block segmentation, but works at pixel level. We now apply the parameters learned using EM for text graphic separation. The text/graphics segmentation problem is modelled as an optimization problem where we estimate optimal values for P/N ratio and local neighbourhood size $W_i(m,n)$.

In this paper we detect the pseudo-periodic pattern that efficiently discriminates text from graphics, due to difference in the spatial distribution of black pixels. The horizontal projection for text regions depict pseudo-periodicity in contrast with graphics. The pseudo-periodic pattern is the autocorrelation of the horizontal projection profile over a local neighbourhood. Based on this we define a parameter called P/N ratio. Where P is the cumulative height with positive slope and N is the cumulative height with negative slope in the autocorrelation plot. Let y[n] be the autocorrelation, then P and N are given by equation 1

$$\begin{aligned} P &= \sum |y[n+2] - y[n]|, \text{if}(y[n+2] - y[n]) > 0 \\ N &= \sum |y[n+2] - y[n]|, \text{if}(y[n+2] - y[n]) < 0 \end{aligned} \tag{1}$$

Generally it is observed that the P/N ratios shows a clear distinction between text and graphics. However, for cases where window overlaps text and graphics pixels in unequal proportions the P/N ratio estimated is not optimal. Hence, adaptive tuning of P/N ratio over different local neighbourhood size is necessary. In contrast to our earlier work [1], where a fixed neighbourhood of size 350 × 350 was used, we learn optimal neighbourhood size to improve the segmentation in newspaper images.

For each image in the training dataset we have its corresponding ground-truth where we have the text and non-text regions marked at pixel level. For segmentation, we slide a window of different sizes over the image with the pixel to be classified placed at the centre of this window. For every window size we calculate the P/N ratio (r_i) as described above. If r_i is less than P/N ratio threshold, the pixel is marked as graphics and vice-versa. For training dataset P/N ratio is determined iteratively over the range $[0.01, 0.9]$ in steps of 0.05. Different segmentation outputs are obtained for different window size and P/N ratio. The P/N ratio and window size which gives the best match between the segmentation output and the ground-truth is selected. To find the best match we compute the KL divergence between the EDH of the segmentation output and ground-truth. Thus, for all images in the training dataset, we store EDH $(E_{(r^i, W^i(m,n))})$ and the corresponding P/N ratio threshold (r^i). For segmentation of test images, we start with 0.01 as seed value for P/N threshold and window of size 100 × 100. The next value in the iteration is determined using the trained EM framework as summarized in table 2.

Table 2: Segmentation: Optimization Framework

```
Inputs:  Binary test image I, training dataset,
         $\{E_j, r_j, W_i(m,n)\}$ where $j = 1, ..., n$.
Outputs: Segmented binary image containing text
Steps:
1)  Set $i \leftarrow 1$ and let $r^{(1)} \leftarrow 0.01$ and $W_i(m,n) \leftarrow (100 \times 100)$ be the initial values.
2)  Get the segmented image $I^{(i)}$ with $P/N$ ratio threshold $r^i$ and $W^i(m,n)$ for $i^{th}$ iteration.
3)  Compute EDH $Z$ of the resultant image $I^{(i)}$.
4)  Apply $(r^\star, W^\star(m,n)) = OP\{arg\ min_{j\epsilon[1,n]} d(Z, E_j)\}$,
    where d is KL divergence measure and $OP\{j\} = (r_j, W^\star(m,n))$ is the optimal $P/N$ ratio.
5)  Check convergence criterion, If $(r^\star, W^\star(m,n)) = (r^i, W^i(m,n))$, then exit.
6)  Set $(r^{i+1}, W^{i+1}(m,n)) \leftarrow (r^\star, W^\star(m,n))$ and go to step 2.
```

4. RESULTS AND DISCUSSION

In this section we present the segmentation results obtained by the proposed framework. The performance of our segmentation algorithm is tested on a set of newspaper articles collected from six different Indian language newspapers, namely, Gurumukhi, Hindi, Kannada, Malayalam, Tamil and Telugu. For experiments, we have prepared a training set with 600 newspaper images, 100 images taken from each script. Ground-truth images were generated by manual labelling at pixel level. Our test-set focused on the newspaper images with both graphics and text. We collected 45 such images along with few images with only text regions.

The effectiveness of our algorithm is demonstrated by the results obtained on the examples shown in first column of figure 2. The second column shows the results of adaptive binarization. Text-graphics segmentation results are shown in the third column. It is worth noting that the complex layout of newspaper image is segmented properly. The algorithm learns the P/N ratio on the running text in which headlines comes very rare. Therefore, headlines and boldface fonts are segmented as graphics. However, when the headlines are of similar size as text, they are segmented correctly as text.

The method used for evaluating the performance of our algorithm is based on counting the number of matches between the pixels segmented by the algorithm and the pixels in the ground truth [11]. We calculate the intersection of the pixel sets of the obtained segment blocks (OB) and the ground truth blocks (GB), and compute the following accuracy measure:

$$\frac{\text{no. of pixels common to both OB and GB}}{\text{maximum area of the two OB or GB}} \qquad (2)$$

The accuracy of the segmentation algorithm proved to be 87-100% when evaluated against the ground-truth for our test set. The accuracy was 100% for pages with only text regions. For the pages with both text and graphics, we have been able to correctly segment 39 images out of 45. The errors were mostly due noisy and degraded newspaper images. We anticipate to improve the segmentation results by experimenting with differnt pre-processing routines to handle degraded images. In addition, the performance and accuracy of our algorithm can be improved by adding more newspaper images, with different layout and scripts with different fonts size and style in the training set.

5. CONCLUSION

This paper contributes a unique technique for segmenting text graphics in Indian newspapers. The proposed technique can be easily adapted for a variety of complex layouts and scripts. The technique has been tested on a variety of document images from different newspapers and books with different page layouts. The results show that it works satisfactorily. Grouping of different segmentation components into areas that have similar function is part of our future work.

6. ACKNOWLEDGEMENT

This work is funded by MCIT, Government of India under the project entitled *Development of Robust Document Image Understanding System for Document in Indian Scripts.*

7. REFERENCES

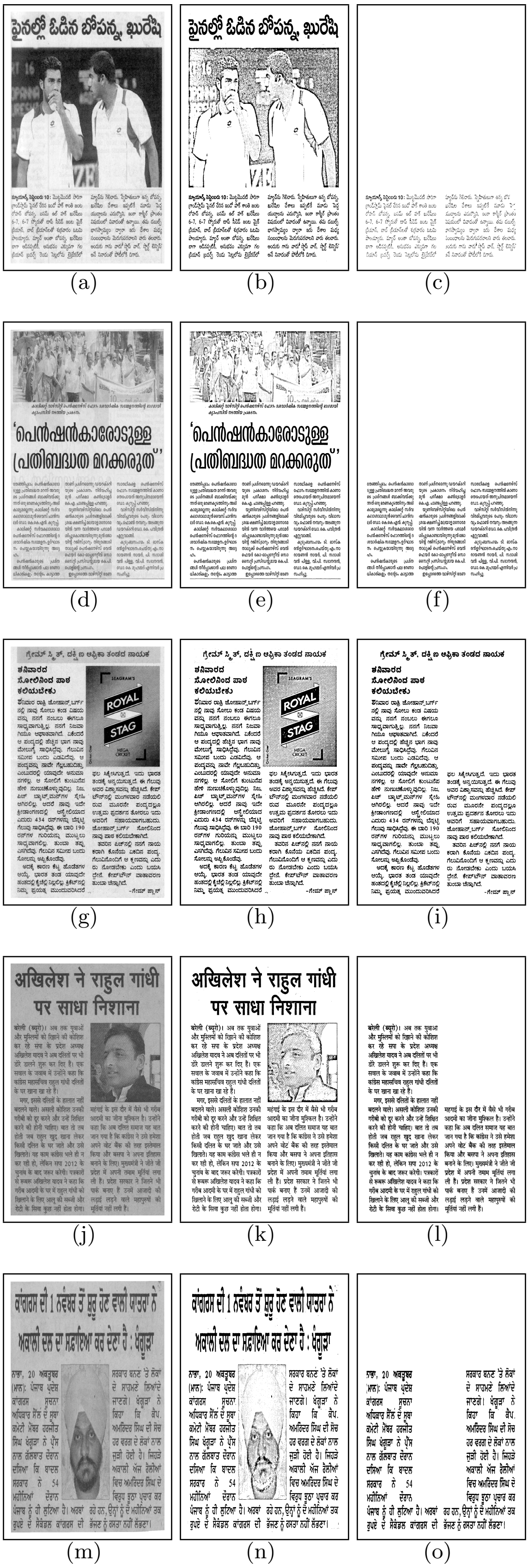

ఫైనల్లో ఓడిన బోపన్న, ఖురేషి
(a)
(b)
(c)
'പെൻഷൻകാരോടുള്ള പ്രതിബദ്ധത മറക്കരുത്'
(d)
(e)
(f)
ROYAL STAG
(g)
(h)
(i)
अखिलेश ने राहुल गांधी पर साधा निशाना
(j)
(k)
(l)
(m)
(n)
(o)

Figure 2: (a, d, g, j, m) show original newspaper article images, (b, e, h, k, n) shown the corresponding binarized images and (c, f, i, l, o) show the results obtained by our approach

[1] S. Aggarwal, S. Kumar, R. Garg, and S. Chaudhury. Content directed enhancement of degraded document images. In *Proceeding of the workshop on Document Analysis and Recognition*, pages 55–61, 2012.

[2] K. C. Fan, C. H. Liu, and Y. K. Wang. Segmentation and classification of mixed text/graphics/image documents. *Pattern Recognition Letters*, 15(12):1201–1209, 1994.

[3] R. Cao and C. L. Tan. Text/graphics separation in maps. In *Fourth International Workshop on Graphics Recognition Algorithms and Applications*, pages 167–177, London, UK, UK, 2002. Springer-Verlag.

[4] R. Cattoni, S. M. T. Coianiz, and C. M. Modena. Geometric layout analysis techniques for document image understanding: a review. *Technical report, IRST*, 1998.

[5] S. Chowdhury, S. Mandal, A. Das, and B. Chanda. Segmentation of text and graphics from document images. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, pages 619–623, Washington, DC, USA, 2007. IEEE Computer Society.

[6] L. A. Fletcher and R. Kasturi. A robust algorithm for text string separation from mixed text/graphics images. *IEEE Transaction Pattern Analysis Machine Intelligence*, 10(6):910–918, 1988.

[7] B. Gatos, S. L. Mantzaris, and A. Antonacopoulos. First international newspaper segmentation contest. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 1190–1194, 2001.

[8] B. Gatos, S. L. Mantzaris, K. V. Chandrinos, A. Tsigris, and S. J. Perantonis. Integrated algorithms for newspaper page decomposition and article tracking. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, 1999.

[9] K. Hadjar, O. Hitz, and R. Ingold. Newspaper page decomposition using a split and merge approach. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 1186–1189, 2001.

[10] K. Hadjar and R. Ingold. Arabic newspaper page segmentation. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ICDAR '03, 2003.

[11] G. Harit, R. Garg, and S. Chaudhury. Syntactic and semantic labeling of hierarchically organized document image components of indian scripts. In *Advances in Pattern Recognition, 2009. ICAPR '09. Seventh International Conference on*, pages 314–317, 2009.

[12] A. K. Jain and S. Bhattacharjee. Texture segmentation using gabor filters for automatic document processing. *Machine Vision and Application*, 5:169–184, 1992.

[13] N. Journet, V. Eglin, J. Ramel, and R. Mullot. Text/graphic labelling of ancient printed documents. In *Proceedings of International Conference on Document Analysis and Recognition*, volume 2, pages 1010–1014, August 2005.

[14] S. Khedekar, V. Ramanaprasad, S. Setlur, and V. Govindaraju. Text - image separation in devanagari documents. In *Proceedings of the Seventh ICDAR*, pages 1265–1269, 2003.

[15] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. *IEEE Transactions of Image Processing*, 16:2117–2128, August 2007.

[16] F. Liu. A new component based algorithm for newspaper layout analysis. In *Proceedings of the Sixth ICDAR*, ICDAR '01, 2001.

[17] J. Liu, Y. Y. Tang, and C. Y. Suen. Chinese document layout analysis based on adaptive split-and-merge and qualitative spatial reasoning. *Pattern Recognition*, 30(7):1265–1278, 1997.

[18] Z. M.-H. H. X.-Z. Liu Dong-Rong, Wang Ke-Jian. Chinese newspaper layout analysis with antecedent compartmental lines. In *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pages 2771–2774, 2003.

[19] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. *Proc. SPIE Electronic Imaging*, page 197âĂŞ207, 2003.

[20] P. E. Mitchell and H. Yan. Newspaper layout analysis incorporating connected component separation. *Image Vision Comput.*, 22(4):307–317, 2004.

[21] G. Nagy. Twenty years of document image analysis in pami. *IEEE Trans. PAMI*, 22(1):38–62, 2000.

[22] P. P. Rege and C. A. Chandrakar. Text-image separation in document images using boundary/perimeter detection. *ACEEE International Journal on Signal and Image Processing*, 03(1):10–14, 2012.

[23] P. P. Roy, J. Llados, and U. Pal. Text/graphics separation in color maps. In *Proceedings of the International Conference on Computing: Theory and Applications*, pages 545–551, Washington, DC, USA, 2007. IEEE Computer Society.

[24] G. Sharma, R. Garg, and S. Chaudhury. Curvature feature distribution based classification of indian scripts from document images. In *Proceedings of the International Workshop on Multilingual OCR*, pages 3:1–3:6, 2009.

[25] C. L. Tan and P. O. Ng. Text extraction using pyramid. *Pattern Recognition*, 31:63–72, 1998.

[26] Y. Y. Tang, S.-W. Lee, and C. Y. Suen. Automatic document processing: A survey. *Pattern Recognition*, 29(12):1931 – 1952, 1996.

[27] K. Tombre, S. Tabbone, L. Pélissier, B. Lamiroy, and P. Dosch. Text/graphics separation revisited. In *Proceedings of the 5th International Workshop on Document Analysis Systems V*, pages 200–211, London, UK, UK, 2002. Springer-Verlag.

[28] F. M. Wahl, K. Y. Wong, and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. In *Computer Graphics and Image Processing*, volume 20, pages 375–390, 1982.

[29] D. Wang and S. N. Srihari. Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics, and Image Processing*, 47(3):327 – 352, 1989.