A CRF Based Scheme for Overlapping Multi-Colored Text Graphics Separation

Ritu Garg, Ehtesham Hassan, Santanu Chaudhury, M Gopal
Department of Electrical Engineering
Indian Institute of Technology Delhi, New Delhi
{*ritu2721a, hassan.ehtesham*}*@gmail.com,* {*santanuc, mgopal*}*@ee.iitd.ac.in*

***Abstract*—In this paper, we propose a novel framework for segmentation of documents with complex layouts. The document segmentation is performed by combination of clustering and conditional random fields (CRF) based modeling. The bottom-up approach for segmentation assigns each pixel to a cluster plane based on color intensity. A CRF based discriminative model is learned to extract the local neighborhood information in different cluster/color planes. The final category assignment is done by a top-level CRF based on the semantic correlation learned across clusters. The proposed framework has been extensively tested on multi-colored document images with text overlapping graphics/image.**

***Keywords*-Document image analysis; Conditional random fields; Complex layout analysis;**

I. Introduction

Document image segmentation is a crucial pre-processing step for context extraction. Document images typically consists of text, graphics/half-tones and background. Several documents such as magazines and brochures contain very complex layout. Segmentation and layout understanding of such documents is presents a challenging task because of the following reasons:

- i Random placement of figures and text.
- ii Complex (textured/colored) backgrounds.
- iii Text overlaid on images/graphical patterns.
- iv Variation in text formatting in terms of font properties (font type, size and color) and orientations.
- v Irregular text regions (non-rectangular)

Figure 1 shows few example document pages with complex layout which have been considered for the current work.

(a) (b) (c)

Figure 1. Sample document images with complex layout

In this paper, we describe a new approach to decompose the document images with complex layouts, and label the segmented regions as text, graphics or background. Our approach defines a hierarchical architecture for segmentation. Color information represents primary cue to separate the graphics and text regions from the background. Hence, we identify the dominant color planes in the image and process pixels corresponding to these planes using individual classifiers. Each classifier is trained by exploiting the local neighborhood properties of the respective color plane. Finally the contextual relationship across different color planes is learned by CRF to smoothen the plane-wise label assignment for final segmentation.

The paper organization is as follows: Section II presents an overview of the related work reported in literature. The details of the proposed document segmentation framework is presented in section III. Experimental results and conclusion are presented in section IV and section V.

II. Related Work

Over the years several document layout analysis algorithms have been proposed in [1], [2]. These approached can be broadly divided into two categories namely top-down and bottom-up methods. Top-down approach use global properties such as white space separations in the document page thereby partitioning the document image into component blocks. Such techniques fail while handling documents with complex layout as they lack clear separations. Conversely, Bottom-up approaches start at pixel level and propagate the local information at global level to perform segmentation. Approach proposed in [3] uses a combination of top-down and bottom-up methodology for handling documents with complex layouts. Lin et. al. [4] presents a hybrid approach to segment and classify document content as text, picture and background. Mukherjee et. al. [5] discusses a multi-scale clustering based technique for document segmentation. Also these approaches are based on classical document segmentation methodologies that are only applicable for gray-scale or binary images. Many model-guided segmentation and layout analysis schemes [6], [7] are also reported in literature. Mighlani et. al. [8] performs color histogram analysis for automated layout segmentation of documents. Layout analysis using color information have been proposed in [9]–[11] to handle color document images with complex layouts such as forms, text overlaid on image, posters etc. These approaches work on either RGB distribution or use

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.245

some color reduction algorithms to use an optimal set of color for text extraction only. [12] describes a methodology for extracting text rendered in uniform color. The algorithm uses connected component analysis based on color similarity in the RGB color space. Various researchers [13], [14] have reported usage of wavelet based techniques for document image understanding (extracting text, picture and background). Such schemes require a lot of prior knowledge of similar documents and are computationally intensive. In contrast to earlier approaches, we propose a technique which works through self-organizing decomposition process in the color space. The hierarchical framework enables to separate out different logical components despite being overlapped in image space. Further, we have used CRF based learning to model the contextual dependencies in image space by using a new formulation of neighborhood across clusters in color space and spatial proximity.

III. Framework for Document Segmentation

The complete framework (Figure 2) proposed for text, graphics and background separation in documents with complex layouts consists of the following steps:

- Pre-processing
  - Color analysis
  - Clustering
  - Feature extraction
- Initial segmentation
- CRF based smoothening

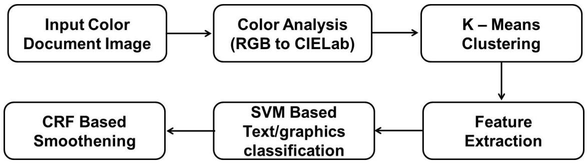
Input Color Document Image
Color Analysis (RGB to CIELab)
K – Means Clustering
Feature Extraction
SVM Based Text/graphics classification
CRF Based Smoothening

Figure 2. Architecture of the Document Segmentation Framework

A. Color Analysis

In many document images, different image components (text, graphics/image and background) appear in different colors. Additionally, in many cases the text overlaps pictures/graphics. Hence it becomes essential to extract local color information in uniform color space. Here we first convert the image from RGB color-space to a uniform color space such as *CIELab space* [15]. *Lab color space* has separate lightness and chroma channels that are approximately perceptually uniform and serves as a device independent color model. CIELab formulae is derived from CIEXYZ, therefore conversion from RGB to CIELab will involve:

i Conversion from RGB to XYZ

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \star \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

ii Converting XYZ to LAB Space

$$L^\star = 116 f(Y/Y_n) - 16$$

$$a^\star = 500\left[f(X/X_n) - f(Y/Y_n)\right]$$

$$b^\star = 200\left[f(Y/Y_n) - f(Z/Z_n)\right]$$

Here X_n, Y_n and Z_n are the CIE XYZ tristimulus values of the reference white point (the subscript n suggests “normalized“), and f is defined as

$$f(t) = \begin{cases} t^{1/3} & \text{if } t > (\frac{6}{29})^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 t + \frac{4}{29} & \text{otherwise} \end{cases}$$

B. Clustering

The initial segmentation is carried out by identifying the color modes using K-means clustering. Each cluster identifies a color plane representing pixels with similar color properties. Ideally, number of clusters should be equal to number of categories. However, the objective of the present work is to analyze document images having multi-colored and overlapping text and non-text (graphics/picture) regions. Based on initial observation, we selected K = 8 as the optimum value, as very large K gives disconnected noisy regions decreasing the over-all segmentation performance. Figure 4 shows the different cluster planes obtained after K-mean clustering. Features extraction for each cluster is discussed in the following section.

C. Feature Extraction

Using the K-clusters we decompose the image into K-color planes. We extract the following local features in each of these color planes:

a. ***Gabor Features***: Texture features are based on the local power spectrum that are computed using 2D gabor filter. Such filters are local and linear, characterized by the localization properties in both spatial domain and spatial frequency domain [16]. The 2D gabor filter is defined as follows:

$$g_{\lambda,\theta,\varphi,\delta,\gamma}(x,y) = Kexp(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2})\cos(2\pi\frac{x'}{\lambda}+\varphi) \quad (1)$$

Here $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$ and λ is the wavelength of the cosine factor of the gabor filter kernel, θ is the orientation, φ is the phase offset with φ = [0 $\pi/2$], δ is the standard deviation of Gaussian function and γ is the aspect ratio ($\gamma = 0.5$). A set of values for $\theta = [0, \pi/4, \pi/2, 3\pi/4]$ and $\lambda = [2, 4, 8]$ are used in our experiments that lead to generation 12 gabor filters.

b. ***Edge Features***: Graphics/images are predominantly defined by high spatial frequency components whereas background regions are defined by low frequency components. In general, the textual content in document images are defined by frequency range lying between the high and low frequencies defining background and graphics regions. Image $I(x, y)$ is convolved using horizontal and vertical masks defined by Sobel, to obtain gradient components G_x and G_y. The gradient magnitude ($Grad_{mag}$) and orientation (θ) are computed at each pixel as:

$$Grad_{mag} = |\sqrt(G_x^2 + G_y^2)|$$

$$\theta = \tan^{-1}(\frac{G_y}{G_x})$$

Two *local gradient histogram* features have been extracted:

b.i *Direction Histogram*: Edge Direction Histogram [17] feature extracts the statistical distribution of curvature found in text (in different scripts/languages) as compared to graphics/images. Most of the scripts have either horizontal and vertical straight lines or curly construction with almost no straight lines. On the other hand, for graphics regions edges distribution is random. The normalized edge direction histogram is computed locally for each pixel in 5×5 neighborhood. We detect the edges in the pixel neighborhood by convolving with horizontal and vertical Sobel masks. Next we compute the edge directions at each pixel in the neighborhood. Finally we compute edge histogram using 12 quantization levels for our experiments.

b.ii *Magnitude Histogram*: Gradient computed over an image gives the information of directional change in intensity values of the image. Local histogram of the gradient magnitudes is computed for every pixel in 5×5 neighborhood. The gradient magnitudes are quantized into T (T = 10 in our experiments) bins to form a T-dimensional features vector corresponding to each pixel. The histogram is normalized by its sum to apply it for further classification purposes.

D. Initial Segmentation

Initial document segmentation is performed over each color plane by SVM based supervised learning. The classification performs a coarser segmentation of text, graphic/picture and background regions using the combination of features discussed above. The combination is done by concatenating various features. The multi-class SVM is architectured in Decision Directed Acyclic Graph (DDAG) [18] framework. The Decision DAG SVM consists of $N(N-1)/2$ nodes, here each node is associated with a binary classifier for a pair of classes. The figure 3 demonstrates prediction process for test point x in 4-class problem in Decision DAG framework. In the case of N class process, the prediction process follows evaluation of $N - 1$ nodes, therefore reducing the required number of kernel computations. The path followed by Decision DAG for a test point is called the evaluation path. The average kernel computation for complete test dataset is obtained by averaging over the count of unique support vectors over the evaluation path for test data points.

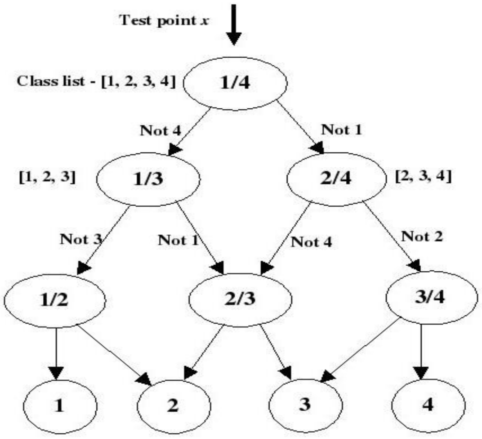

Figure 3. 4-class classification with DAG SVM

Figure 4 shows the resultant images for each cluster plane after SVM labeling. Green, red and blue pixels indicate picture/graphics, text and background regions respectively.

E. CRF based Post-processing

SVM based classification performs the coarse level segmentation of the document. The segmentation can be improved by utilizing the contextual information between pixels. The contextual relationship can be efficiently learned by a probabilistic graphical model. The CRFs are an efficient tool for such problems. The objective of CRF is to apply smoothening over the hard labels assigned by SVM which does not consider the contextual relationship between pixels for classification.

CRF is a discriminative modeling tool for segmenting and labeling the sequential data [19]. The CRFs work under the maximum entropy principle and observe the features as sequence data. Let $y_i \in 1, 2, \ldots L$ denotes the label, and $\underline{x}_i$ denote a D-dimensional feature sequence provided for the smoothening in the neighborhood of i^{th} pixel.

The conditional probability of label y_i is modeled as

$$p(y_i|\underline{x}_i, \underline{\lambda}) = \frac{1}{Z_\lambda} \prod_{c \in \mathrm{C}(G)} exp(\lambda_c, f(c)) \qquad (2)$$

Here G is undirected graph for capturing the spatial dependencies, $\mathrm{C}(G)$ is the set of the cliques in th egraph. $\underline{\lambda}$ are the parameters for cliques in the graph. $Z_\lambda = \sum_{y_i} \prod_{c \in \mathrm{C}(G)} exp(\lambda_c, f(c))$. We have defined two types of cliques:

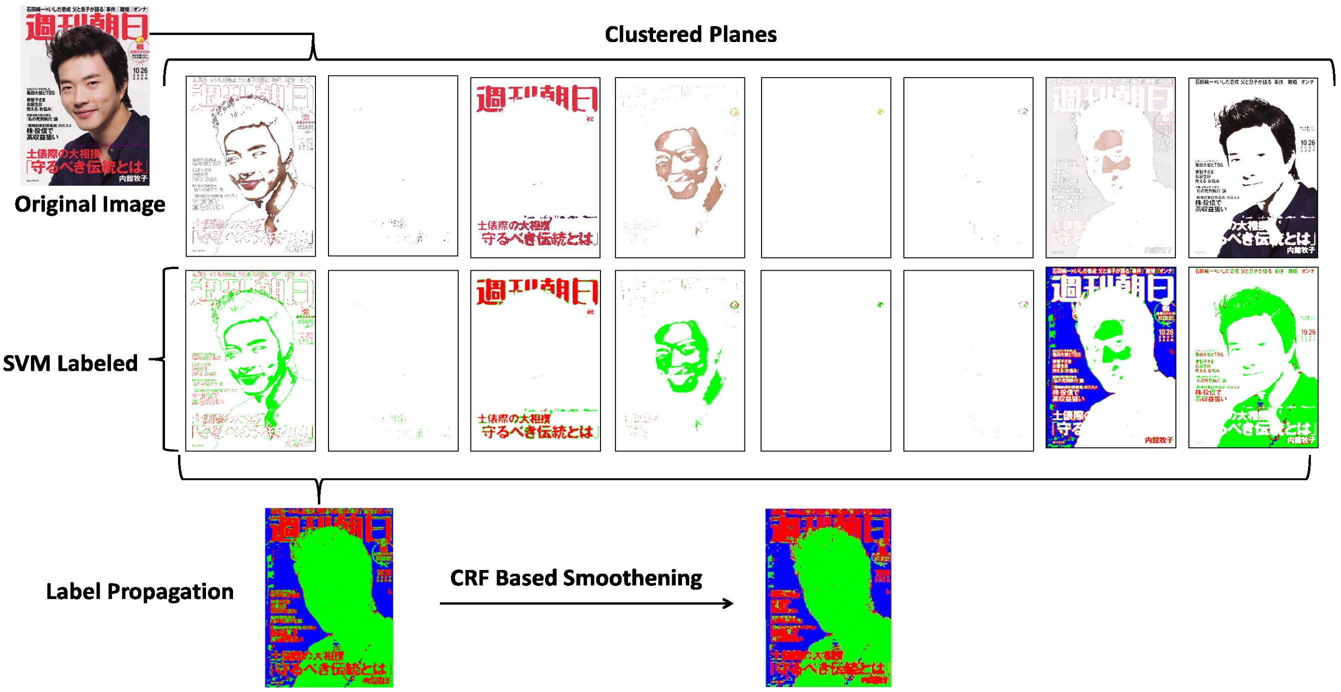

Figure 4. Text Graphic Separation framework

i Single node-clique composed at each node i

$$f_1(y_i, \underline{x}, i) = \sum_{j=1}^{D} d_1(y, x_j)$$

Here d_1 represents the L1 distance.

ii Spatial node-clique composed of spatial edges in the given neighborhood of node i. This composes spatial compatibility by encouraging the assignment of the similar color label in a neighborhood.

$$f_2(y, y^{'}, \underline{x}_y, \underline{x}_{y'}, i, j) = d_1(\underline{x}_y, \underline{x}_{y'})$$

Here $j \in nbh(i)$.

The labels from each cluster plane are propagated to a single plane as shown in figure 4. This plane is then subjected to CRF based smoothening, as the SVM based deterministic labels do not consider the contextual relationship across the pixel neighborhood. For our experiments we have considered a 5×5 neighborhood. The results after CRF smoothening can be seen in figure 4.

IV. Experimental Results

Our dataset is composed of document images consisting of the three classes: Text, Graphics/Image and Background. The collection comprises of document from magazines, articles and brochures, typically in non-manhattan layout. Typically we have picked magazine front-pages which have text rendered in different styles, color and fonts and orientation and also overlaid on graphics.

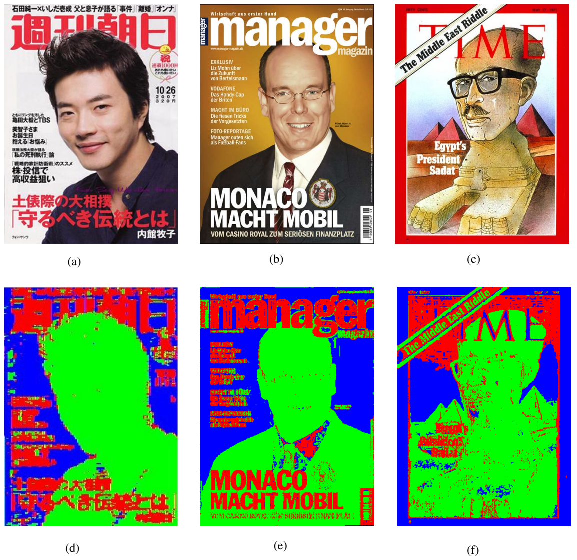

Figure 5. (a), (b), (c) are the original images, and (d), (e), (f) are the final segmented results

We have tested our approach on 20 images with size varying from 400×600 to 900×1200. For all the images, we have prepared the ground truth carefully by manual segmentation. We performed training over 4 randomly selected images and rest of the images were used for testing. The experimental results have been presented as average of 3 iterations. The SVM classifiers for each color plane

Table I
FINAL SEGMENTATION ACCURACIES WITH SVM AND CRF SMOOTHENING

Original Image	Final Segmentation Result	SVM	After CRF Smoothening
figure 5(a)	figure 5(d)	84%	86%
figure 5(b)	figure 5(e)	87%	89%
figure 5(c)	figure 5(f)	81%	83%

were trained by randomly sampling 5% of the total number of the training pixels corresponding to a color plane. We achieved 83.7% classification accuracy by SVM based classification. In general, for all the training images CRF based smoothening increased average classification accuracy by 4%-6%. After application of CRF, final classification accuracy computed over three iterations, improved from 83.7% to 89%.

V. CONCLUSION

We have demonstrated a novel approach for document segmentation by defining a hierarchical framework. The hierarchical framework uses unsupervised learning over color properties for doing the coarse level document segmentation. The top level of the hierarchy combines SVM and CRF for doing the final classification. Our framework can successfully identify text, graphics and background regions in multicolored documents. The efficacy of proposed framework is supported by the experimental results presented above. The experimental evaluation of the proposed framework for wider class of documents is part of future works.

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*.

REFERENCES

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

[2] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: a literature survey," *Proc. SPIE Electronic Imaging*, p. 197207, 2003.

[3] S. W. Lam, "A local-to-global approach to complex document layout analysis," *IAPR Workshop on Machine Vision Applications*, pp. 13–15, 1994.

[4] M.-W. Lin, J.-R. Tapamo, and B. Ndovie, "A texture-based method for document segmentation and classication," *In Joint Special Issue Advances in end-user data-mining techniques*, pp. 49 – 56, 2006.

[5] D. P. Mukherjee and S. T. Acton, "Document page segmentation using multiscale clustering," *In Proceedings of International Conference on Image Processing*, vol. 1, pp. 234 – 238, 1999.

[6] S. Chaudhury, M. Jindal, and S. D. Roy, "Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field," *In Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, pp. 375–380, 2009.

[7] T. Watanabe and T. Sobue, "Layout analysis of complex documents," *In Proceedings 15th International Conference on Pattern Recognition*, vol. 4, pp. 447 – 450, 2000.

[8] D. Mighlani, A. Hennig, N. Sherkat, and R. J. Whitrow, "A visual vocabulary for flower classification," *In Proceedings of IEEE TENCON '97. Speech and Image Technologies for Computing and Telecommunications.*, pp. 191 – 194, 1997.

[9] W. S. Wong, N. Sherkat, and T. Allen, "Use of colour in form layout analysis," *Sixth International Conference on Document Analysis and Recognition, 2001*, pp. 942 – 946, 2001.

[10] C. Strouthopoulos, N. Papamarkos, A. Atsalakis, and C. Chamzas, "Text identification in color documents," *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, pp. 702 – 705, 2003.

[11] M. Berardi, O. Altamura, M. Ceci, and D. Malerba, "A color-based layout analysis to process censorship cards of film archives," *In Proceedings of Eighth International Conference on Document Analysis and Recognition*, pp. 1110 – 1114, 2005.

[12] A. Clavelli and D. Karatzas, "Text segmentation in colour posters from the spanish civil war era," *10th International Conference on Document Analysis and Recognition*, pp. 181 – 185, 2009.

[13] 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 on Image Processing*, vol. 8, pp. 2117 – 2128, 2007.

[14] M. Acharyya and M. K. Kundu, "Document image segmentation using wavelet scalespace features," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1117 – 1127, 2002.

[15] S. K. Shevell, *The Science Of Color*. Elsevier Science & Technology, July 2003.

[16] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on gabor filters," *IEEE Transactions on Image Processing*, pp. 1160 – 1167, 2002.

[17] 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*, 2009.

[18] J. C. Platt, N. Cristianini, and J. S. Taylor, "Large margin dags for multiclass classification," *In Advances in Neural Information Processing Systems*, vol. 12, pp. 547–553, 2000.

[19] F. P. John Lafferty, Andrew McCallum, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," *Proceedings of the International Conference on Machine Learning, Morgan Kaufmann*, pp. 282 – 289, 2001.