Newspaper Article Extraction Using Hierarchical Fixed Point Model

Anukriti Bansal*, Santanu Chaudhury*, Sumantra Dutta Roy* and J.B. Srivastava†
*Department of Electrical Engineering
†Department of Mathematics
Indian Institute of Technology Delhi, New Delhi, India
Email: {anukriti1107,schaudhury,sumantra.dutta.roy,jbsrivas}@gmail.com

***Abstract*—This paper presents a novel learning based framework to extract articles from newspaper images using a Fixed-Point Model. The input to the system comprises blocks of text and graphics, obtained using standard image processing techniques. The fixed point model uses contextual information and features of each block to learn the layout of newspaper images and attains a contraction mapping to assign a unique label to every block. We use a hierarchical model which works in two stages. In the first stage, a semantic label (heading, sub-heading, text-blocks, image and caption) is assigned to each segmented block. The labels are then used as input to the next stage to group the related blocks into news articles. Experimental results show the applicability of our algorithm in newspaper labeling and article extraction.**

I. Introduction

Article extraction is very useful for storage, retrieval, transmission, preservation and archival of newspaper images in digital form. The problem of article extraction is immensely challenging due to a wide range of layouts and random placements of different components of newspaper image. Algorithms for article extraction have been proposed by authors in the past, however, most of these methods are based on a set of heuristic rules which are designed manually. If the number of distinct layout is large, the number of rules that need to be manually created also becomes large, thereby decreasing the recognition speed and accuracy. In contrast to earlier methods, we propose a learning based framework which uses a Fixed-Point Model [15] to automatically learn the layout and structure of a newspaper document. It assigns a unique label (heading, sub-heading, text-block, image and caption) to each component of a newspaper image and then group the related components into news article.

Fixed Point Model as proposed by Li *et al* [15] was used for structured labeling tasks by capturing the dependencies between observed data. The structured input is denoted as graph. The objective of structured labeling task is to jointly assign labels to all nodes as a joint output. In computer vision, images are considered as a structured input of all the pixels, and the structured output is corresponding labels of these pixels. Edges between the nodes denotes the correlations among structured outputs. The correlation may occur between neighboring nodes, or the nodes relatively distant. Fixed Point Model provides the labeling of each node by using the features of node and labeling of the neighboring nodes. The motivation to use fixed point model for the labeling and article extraction task arises from the spatial inter-dependencies of different regions of a newspaper document. Fixed point model uses the context information and attains a contraction mapping to assign a unique label to each region of newspaper image and determines the logical relationships between different regions to group them into news article.

II. Related Work

Over the years, several interesting survey papers have been published on layout analysis and article extraction [4], [7], [17]–[20], [22]. Here we review the literature specific to region labeling and article extraction. Approaches vary from using rule based techniques to the use of classifiers, textual features and graph theory.

Most of the earlier work on layout analysis is done on journal pages which does not have complex layout where text and graphic regions are placed in a random fashion. The problem of layout analysis for newspaper images is addressed by few authors. Gatos *et al* [11] proposed an integrated methodology for segmenting newspaper page and identifying newspaper article. In the first stage, various regions are extracted using smearing and connected component labeling. A rule based approach is applied in the second stage to extract various regions and newspaper articles. Liu *et al* [16] presented a component based bottom-up algorithm for analyzing newspaper layout. This algorithm is based on a distance measure and layout rules which are designed heuristically.

Few authors [3], [10] have also worked on article extraction based on similarity of text from the text blocks generated from segmenting newspaper images. Furmaniak [10] identified paragraphs in the newspaper page and then measured similarity is measured between neighboring OCR'ed paragraphs. Wang *et al* [23] classified newspaper image block using textual features. The technique proposed assumes homogeneous rectangular blocks extracted using RLSA and Recursive X-Y cuts. The blocks are then classified based on statistical textual features and space decision techniques.

In [12], [13] authors have used split and merge techniques for the decomposition of newspaper document into blocks. Hadjar *et al* [13] employed a split and merge algorithm, which splits the entire image along detected lines and then attempts to merge these small zones into larger zones.

Learning based methods have also been reported in literature for block labeling, layout analysis and other document processing tasks. Bukhari *et al* [6] proposed a layout analysis technique for Arabic historical document images.

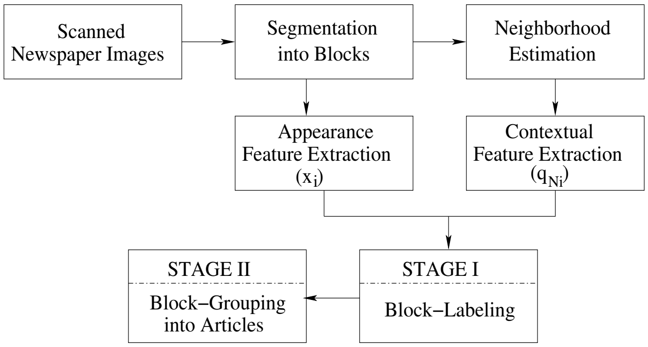

Scanned Newspaper Images
Segmentation into Blocks
Neighborhood Estimation
Appearance Feature Extraction (x_i)
Contextual Feature Extraction (q_{Ni})
STAGE II
Block–Grouping into Articles
STAGE I
Block–Labeling

Fig. 1. Block diagram of the system showing steps involved in article extraction process

They extracted and generated feature vector in a connected-component level. Multi-layer perceptron classifier was used to classify connected components to the relevant classes of text. A voting step is applied for the final classification and refinement. Markov Random Fields [24] and Condition Random Fields [8], [14], [21] have been used to capture the contextual information. However, due to heavy computational burden in their training and testing stages, MRF and CRF are often limited to capturing a few neighborhood interactions and thus, limiting their modeling capabilities. Here we have used a fixed point model which is capable of learning the context in efficient fashion within reasonable time. The model also avoids a cascaded approach with deep layers. However, it can use rich contextual information through effective and efficient inference. The use of Fixed point model for automatically learning the rules for block labeling and article extraction is the key contribution of this paper.

A. System Overview

Article extraction includes grouping together the blocks that belong to the same article. We obtain text-graphics blocks by segmenting the image using morphological image processing techniques as described in [5]. There exists a correlation between these blocks, eg., a text block below a graphic region can be classified as a caption if it has proper dimensions, font size, and alignment with the graphic block. As shown in Figure 1, in the first stage, a fixed point model determines the relationship between neighboring blocks and classify them as heading, sub-heading, caption and text-blocks. These labels are used as input to the next stage where another fixed point model learns the rules to group the blocks which belong to same news article. This work proposes a unified learning-based solution for the problem of labeling and article extraction in newspaper images having complex layout.

The organization of the paper is as follows. Section III gives a brief description of the fixed point model. Section IV discusses our approach in detail. The results of the proposed approach are shown in Section V. Finally we conclude the paper in Section VI.

III. Fixed Point Model Description

Li *et al* [15], introduced a contextual prediction function for each node. It gives the labeling of an individual node as output and takes its features and labeling of neighbouring nodes as input. The fixed point function is the vector form of the contextual prediction function of all nodes and is trained with the property of contraction mapping so that a unique label is obtained during the iterative prediction process.

Input to the fixed point model is an image of non-overlapping text and graphics blocks as shown in figure 2, which represents a 2D graph structure and is denoted as $\mathcal{I} = (\mathcal{B}, \mathcal{L})$. Each block $b_i \in \mathcal{B}$ corresponds to a non-overlapping block with its features denoted as x_i. The link parameter $\mathcal{L}$ decides the neighborhood of each block. For instance, the neighborhood $\mathcal{N}_i$ of block b_i in a 2D graph is specified by m blocks in each of the 4 directions: top, bottom, left and right. We use m to denote the number of neighbors a block can have in the neighborhood specification. Labels for all the blocks is denoted as $\mathbf{y} = (y_i : 1..|\mathcal{C}|)$ where, $y_i \in \mathcal{C}$ and $\mathcal{C}$ = {Heading, Sub-heading, Text-columns, Graphics, Caption}, is the label space. $\mathbf{q}_{\mathcal{N}i}$ is the labeling of the neighborhood of b_i and $\mathbf{q}$ denotes the labeling of all the blocks in $\mathcal{I}$. A context prediction function is used to predict the labeling $y_i \in \mathcal{C}$ for the text blocks. For each block b_i, the function f takes in both b_i's appearance features (x_i) and contextual features ($\mathbf{q}_{\mathcal{N}i}$). Appearance features try to associate each block to a label using the characteristics of that block alone. Contextual features, on the other hand, associate a block to a label using information from the neighboring blocks. The contextual prediction function is described as below (following the scheme proposed in [15])

$$q_i = f(\mathrm{x}_i, \mathbf{q}_{\mathcal{N}i}; \boldsymbol{\theta}) \tag{1}$$

where f is a regression function within range [0,1], and θ is the parameter of the function. From Equation 1, the labeling $\mathbf{q}$ of all text blocks can be written in a vector form,

$$\mathbf{q} = \mathbf{f}(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}, \mathbf{q}; \boldsymbol{\theta}), \tag{2}$$

As seen from equation 2, both output and input contains labeling $\mathbf{q}$. Given the labeling $\mathbf{q}$ and features $\{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}\}$ of the training data, we learn the parameters $\boldsymbol{\theta}$. The logistic regression function holds the property of contraction mapping, i.e., it has a stable state (an attractive fixed point) for each structured input. There exists a stable status for ground-truth labeling in the training process which leads to the fixed-point iteration in the prediction process: $\mathbf{q}^t = \mathbf{f}(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}, \mathbf{q}^{t-1}; \boldsymbol{\theta})$ and $\mathbf{q}^t \to \mathbf{q}$ as $t \to \infty$. The function $\mathbf{f}$ with such property is termed as fixed point function.

We have performed experiments using SVM (Support Vector Machines) and KLR (Kernel Logistic Regression) as contextual prediction function. For each block, we train the function on the basis of appearance feature x_i and contextual feature $\mathbf{q}_{\mathcal{N}i}$. Once learned, the contraction mapping is applied iteratively to the new structured inputs. During testing, for a new structured input, the label q_i of a block $b_i \in \mathcal{B}$ is initialized with a value. We have initialized it with 0 in our experiments.

IV. Methodology

We have used scanned images of English newspapers for our experimentation. Each document provides for a number of blocks, typically an average of 60 blocks exists in a single newspaper image. We use leptonica [1] to obtain these blocks.

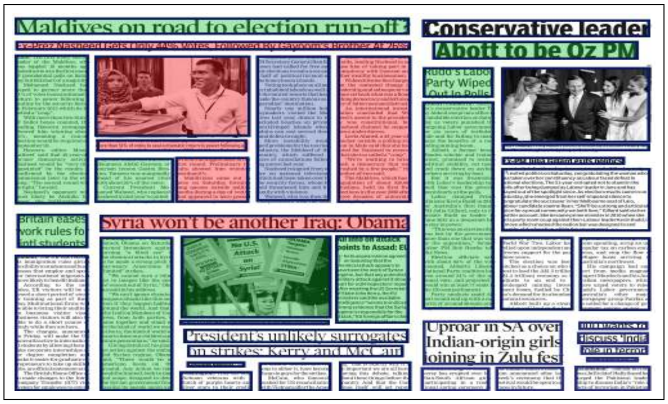

Maldives on road to election run-off?
Conservative leader
Abott to be Oz PM
Syria won't be another Iraq: Obama
Britain eases work rules for intl students
Uproar in SA over Indian-origin girls joining in Zulu fest
President's unlikely surrogates on strikes: Kerry and McCain

Fig. 2. Neighborhood Selection: Current block marked in blue, neighbors at m = 1 marked in red, neighbors at m = 2 marked in red and green

A brief overview of the algorithm is given in Section IV-A. The method to obtain the linking (neighborhood) information of each block of an image is described in Section IV-B. Section IV-C explains the process of block labeling and IV-D describes the article extraction using fixed point model. The operational steps involved in labeling and grouping of blocks are shown in Figure 1.

A. Preprocessing and Text Graphics Segmentation

The gray scale image is first binarized using the method described in our earlier work [2]. Horizontal and vertical lines are then removed on the basis of aspect ratio of connected components. Next we segment the binarized image into text and graphic regions. Efficient methods are present in literature for this purpose. In the present work we use the leptonica library designed by Bloomberg [5] for segmenting the newspaper image into homogeneous regions of text and graphics. Advantages of using leptonica for the segmentation are three folds: i) It segments the image without any *a priori* knowledge, ii) It is computationally very efficient, and iii) gives homogeneous blocks for text and graphics regions as shown in Figure 2. The result of segmentation is that each of the printed blocks on the image is isolated and categorized as text or graphics.

B. Neighborhood Estimation

Once all the blocks are obtained for a newspaper image, the neighboring blocks are identified. The neighborhood of the i^{th} block is defined by all the adjacent blocks to its right, left, top and bottom. We use a neighborhood parameter m, which specifies the span of context/neighborhood we want to capture. Figure 2 shows the result of segmentation on a small region of document, wherein the i^{th} block is shown in blue. The neighborhood of text block for m = 1 are the blocks highlighted in red. And for m = 2 neighborhood blocks are all the blocks highlighted in green and red. In our experiments, we empirically choose m = 6 for labeling the blocks and m = 9 for article extraction.

C. Block Labeling

In this stage, all the segmented blocks of newspaper image are labeled as heading, sub-heading, text-block, caption and graphics. The non-overlapping blocks corresponds to a structured input $\mathcal{B}$ and the i^{th} block corresponds to b_i. For each block, the features listed in Table I are grouped to form appearance feature vector x_i. A normalized histogram of neighboring blocks is used as contextual feature $\mathbf{q}_{\mathcal{N}i}$. The number of bins in a histogram is equal to the number of class labels. The frequency of occurrence of a particular class label within the neighborhood of a block is assigned to the corresponding bin. For each block b_i, this creates a 4 x m x $\mathcal{C}$ dimensional contextual feature vector. Here, $\mathcal{C}$ is the number of class labels, m is the span of the context and 4 specifies the neighborhood in four directions (upper, lower, left and right). During testing, at each iteration, the label for which convergence (contraction mapping) is achieved, is assigned to the corresponding block. When contraction mapping is attained, the block labels do not change in subsequent iterations. In our experiments, we set the number of maximum iterations to be 5. We use Kernel Logistic Regression (KLR) with RBF kernel as the contextual prediction function. We have also experimented with L1 regularized support vector machine (SVM-L1), provided in the Liblinear software package [9] as the classifier.

TABLE I. APPEARANCE FEATURES FOR BLOCK CLASSIFICATION

Appearance Feature	**Feature Description**
Height	Maximum Height of the block
Width	Maximum width of the block
Second maximum component height	Second maximum height of the connected components in a block
Aspect Ratio	Width/Height of the block
Ratio of black and white pixels	Used to determine inverse text

D. Block Grouping and Article Extraction

At this stage we have a list of all the blocks which have been classified and labeled. However, the goal is to extract the information unit of newspaper document, i.e., news articles. So, the final step is to group the blocks which belong to same news article. Here, all the news articles in a newspaper image correspond to the structured input $\mathcal{B}$ and the i^{th} article represents b_i. The label space $\mathcal{C}$ = {article, non-article}. The proposed method follows an assumption that an article will always be associated with a 'heading' tag. The method for article extraction is outlined below:

1) Learning Phase:
   a) A block with label 'heading' is chosen.
   b) All the blocks which belong to the same article which comprises the heading are labeled as 'article' blocks.
   c) Other blocks which do not belong to the article region corresponding to the chosen heading are labeled as 'non-articles'.
   d) The Fixed-Point model is learned for all such articles present in the training images.
   e) This phase is expected to learn the grouping rules whereby neighboring blocks get associated with a heading to form an article.
2) Testing Phase:
   a) For each given test image, we process the blocks labeled as 'heading' one by one.

TABLE II. EXPERIMENTAL RESULTS OF NEWSPAPER BLOCK LABELING

Block Label	**# blocks**	**Average Accuracy using KLR (in %)**	**Accuracy using SVM (in %)**
Heading	359	96.32	90.762
Sub-heading	145	82.4	76.279
Text-block	1416	97.93	94.81
Caption	122	83.74	78.3
Total Blocks	2042	96.0	91.76

TABLE III. EXPERIMENTAL RESULTS OF ARTICLE EXTRACTION

Context prediction function	**Total # articles**	**Average Accuracy (Percentage of articles in test set)**		
		Articles with correct labeling	Articles with false positive error	Articles with false negative error
SVM	359	76.315	18.42	5.26
KLR	359	71.4	21.05	7.89

   b) For a given heading, we label all the blocks in the test image as article/ non-article.
   c) The blocks which are closer to the given heading get labeled as 'article' and all other blocks get labeled as 'non-article'.

3) The appearance feature vector for each block in an article is formulated as: (i) percentage overlap with the heading block, (ii) the horizontal and vertical distance to the heading block, (iii) its block-label. The contextual features are the class label (article or non-article) and the block label (heading, sub-heading, caption, image or text-block) of neighboring blocks.

V. RESULTS AND DISCUSSION

The proposed method for newspaper labeling and article extraction is tested on a dataset comprising 45 images from different English newspapers, which are characterized by a variety in page layouts. Each image was of size 4000 x 6500 pixels. Figure 3(a) shows one of the newspaper images from our dataset. We manually created the ground truth at block level for the quantitative evaluation of our labeling and article extraction methods. From 45 images, we got 2268 blocks of text and graphics which form a dataset for block labeling process. 384 articles are used for the validation of article extraction methods. We do 10-fold cross validation for evaluating the results of learning. Performance of our algorithm is summarized in Table II and Table III.

We achieved 96% accuracy for assigning a semantic labeling to each block. The dataset was divided in nearly 10 equal parts for cross validation and average accuracy over all iterations is presented in Table II. Additionally, accuracy is calculated for all the different type of blocks found in labeling process. Significant accuracy is recorded in case of blocks of type 'heading' and 'text-block'. The possible reason is that the contextual and appearance features of these block-labels are very strong which make them distinct from neighboring blocks. Long captions get confused with text-blocks and in some cases the features (like font size) of sub-heading block are similar to heading blocks.

Table III shows the results of article extraction. An article is regarded to be correctly identified when all its blocks are grouped together (labeled) correctly. Additionally, two measures are computed for the incorrectly identified articles:

1) False Positive Error: (Incorrectly classified article region outside the Ground-Truth region)/Actual article region
2) False Negative Error: (Incorrectly classified article region inside the Ground-Truth region)/Actual article region

Performance of article identification step is adversely affected by the errors in the previous steps of text-graphics segmentation and block labeling. The block-level accuracy for article/non-article labeling comes around 90% during article extraction.

We have implemented the fixed point model using SVM and KLR as contextual prediction function. As evident from tables, KLR gives better results than SVM for the block-labeling task while SVM works better in case of article identification. Both learn the rules differently and attains contraction mapping to assign a unique label to each block.

Figure 3 show the results of our method on a newspaper image from test set. It can be seen from figure 3(b), that features of some of the 'heading' and 'sub-heading' blocks are similar but they are correctly labeled due to the use of rich contextual information. Figure 3(c) show the results of article extraction on different layouts of news articles.

VI. CONCLUSION

This paper presents a novel learning based framework for the problem of article extraction in newspaper images. The Fixed point model captures contextual information to predict the label of each block. The approach is very general and is not based on a set of rules. It can be used to train various other layouts. Automatically learning the rules to identify the logical units of a document image is the key contribution of this paper. The experimental results show that our method is efficient in correctly identifying the news articles and provide an alternative to most of the present heuristic rule based systems. Accuracy can be further improved by adding more newspaper images with different layout in the training set. Future work includes a comprehensive analysis of the applicability of the proposed approach to larger datasets of newspapers and other types of documents.

REFERENCES

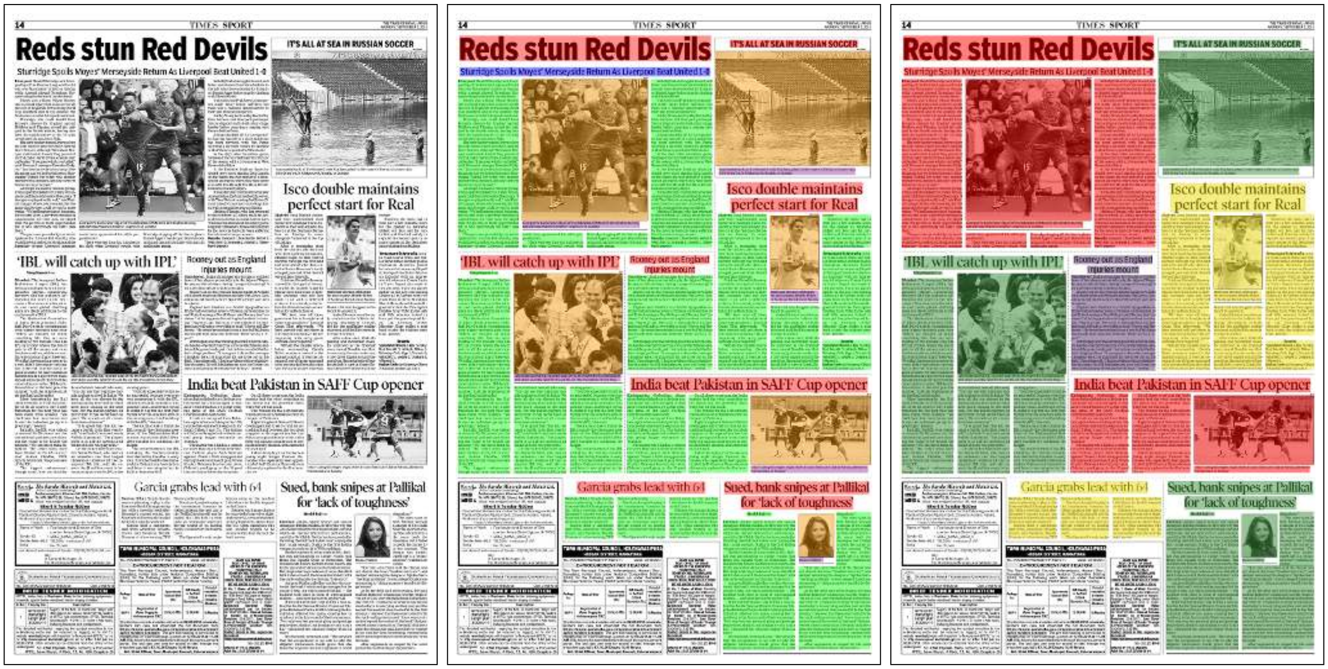
[1] http://www.leptonica.com/.

[2] 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*, DAR '12, pages 55–61, 2012.

[3] M. Aiello and A. Pegoretti. Textual article clustering in newspaper pages. *Applied Artificial Intelligence*, 20(9):767–796, 2006.

[4] R. Beretta and L. Laura. Performance evaluation of algorithms for newspaper article identification. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, ICDAR '11, pages 394–398, 2011.

[5] D. S. Bloomberg. Multiresolution morphological approach to document image analysis. In *Proceedings of the 1991 International Conference on Document Analysis and Recognition*, ICDAR '91, 1991.


14
TIMES SPORT
Reds stun Red Devils
Sturridge Spoils Moyes' Merseyside Return As Liverpool Beat United 1-0
IT'S ALL AT SEA IN RUSSIAN SOCCER
Isco double maintains perfect start for Real
'IBL will catch up with IPL'
Rooney out as England injuries mount
India beat Pakistan in SAFF Cup opener
Garcia grabs lead with 64
Sued, bank snipes at Pallikal for 'lack of toughness'

(a) (b) (c)

Fig. 3. Results of article extraction: (a)the original newspaper image; (b)Results of Labeling each block as heading (red), sub-heading (blue), text-block (green), image (orange), caption (violet); (c)Different articles identified.

[6] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana. Layout analysis for arabic historical document images using machine learning. In *ICFHR*, pages 639–644, 2012.

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

[8] S. Chaudhury, M. Jindal, and S. D. Roy. Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field. Lecture Notes in Computer Science, pages 375–380. Springer, 2009.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[10] R. Furmaniak. Unsupervised newspaper segmentation using language context. pages 1263–1267, 2007.

[11] B. Gatos, S. L. Mantzaris, K. Chandrinos, A. Tsigris, and S. J. Perantonis. Integrated algorithms for newspaper page decomposition and article tracking. pages 559–562. IEEE Computer Society, 1999.

[12] V. Govindaraju, S. W. K. Lam, D. Niyogi, D. B. Sher, R. K. Srihari, S. N. Srihari, and D. Wang. Newspaper image understanding. volume 444 of *Lecture Notes in Computer Science*. Springer, 1989.

[13] K. Hadjar, O. Hitz, and R. Ingold. Newspaper page decomposition using a split and merge approach. pages 1186–1189. IEEE Computer Society, 2001.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, 2001.

[15] Q. Li, J. Wang, Z. Tu, and D. P. Wipf. Fixed-point model for structured labeling. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 214–221, 2013.

[16] F. Liu, Y. Luo, D. Hu, and M. Yoshikawa. A new component based algorithm for newspaper layout analysis. pages 1176–1180. IEEE Computer Society, 2001.

[17] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. volume 5010 of *SPIE Proceedings*, pages 197–207. SPIE, 2003.

[18] G. Nagy. Twenty years of document image analysis in pami. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):38–62, Jan. 2000.

[19] A. Namboodiri and A. Jain. Document Structure and Layout Analysis. pages 29–48. 2007.

[20] O. O. and P. M. A survey of texture-based methods for document layout analysis., 2000. In Texture Analysis in Machine Vision (Ed. M. Pietikäinen), Series in Machine Perception and Artificial Intelligence, Vol. 40, World Scientific, 165-177.

[21] S. Shetty, H. Srinivasan, S. Srihari, S. Shetty, H. Srinivasan, M. Beal, and S. Srihari. Segmentation and labeling of documents using conditional random fields, 2007.

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

[23] D. Wang and S. N. Srihari. Classification of newspaper image blocks using texture analysis. 47:327–352, 1989.

[24] Yefeng Zheng, Huiping Li, and David Doermann. Machine Printed Text and Handwriting Identification in Noisy Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):337–353, March 2004.