Model-Guided Segmentation and Layout Labelling of Document Images Using a Hierarchical Conditional Random Field

Santanu Chaudhury, Megha Jindal, and Sumantra Dutta Roy

Dept of Electrical Engg, IIT Delhi, Haux Khas, New Delhi - 110 016, India
{schaudhury,megha1jindal}@gmail.com, sumantra@cse.iitd.ac.in

Abstract. We present a model-guided segmentation and document layout extraction scheme based on hierarchical Conditional Random Fields (CRFs, hereafter). Common methods to classify a pixel of a document image into classes - text, background and image - are often noisy, and error-prone, often requiring post-processing through heuristic methods. The input to the system is a pixel-wise classification based on the output of a Fisher classifier based on the output of a set of Globally Matched Wavelet (GMW) Filters. The system extracts features which encode contextual information and spatial configurations of a given document image, and learns relations between these layout entities using hierarchical CRFs. The hierarchical CRF enables learning at various levels - 1. local features for text, background and image areas; 2. contextual features for further classifying region blocks - title, author block, heading, paragraph, etc.; and 3. probabilistic layout model for encoding global relations between the above blocks for a particular class of documents. Although the work has been motivated for an automated layout analyser and machine translator for technical papers, it can also be used for other applications such as search, indexing and information retrieval.

1 Introduction

Automatic segmentation and layout analysis of documents can be used for interpretation and machine translation of technical documents, search and information retrieval, in general. Common approaches have often been heuristic in nature, for instance [1]. A learning-based approach is more general than using assumptions about document layouts. Further, this can make it tunable for a particular class of documents. An earlier work [2] presents a learning-based scheme for extraction of text, image and background pixels using a learning-based approach - globally matched wavelets (GMWs, hereafter), and a Markov Random Field (MRF, hereafter) model for smoothing the results. This works at a pixel level, and does not consider the problem of layout analysis. Knowledge of the layout itself can remove page segmentation errors. Shafait et al. [3] present a statistical learning-based mechanism for layout analysis. They overcome the exponential computational cost of optimal geometric parsing of methods relying on probabilistic grammars [4], [5], [6]. They model a page as a mixture of layout

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 375–380, 2009.

structures, and use a probabilistic matching algorithm for find the most probable layout. Our work here uses a much more general structure - conditional random fields (CRFs, hereafter), which avoid the limitations of generative models such as MRFs [7]. MRF-based layout modelling [8] and generative zone models [9] typically need large amounts of labelled training data [10].

Xuming He et al. [7] propose the use of multi-scale CRFs for image labelling. CRFs have been used for document segmentation and labelling e.g., Shetty et al. [11]. The authors use a CRF with simple features such as height of a patch, component width, density, etc. In contrast, our work uses a unified CRF learning-based approach to document layout segmentation at three different levels:

1. A CRF framework for filtering a Fisher classifier output on GMW filters applied on document images.
2. Contextual features for classifying text blocks, and
3. A CRF-based probabilistic method for encoding global relations between the above blocks, for different document classes.

Fig. 1(a) gives a graphical overview of the ideas in this paper. To the best of our knowledge, no other work encompasses a general learning-based procedure at all levels of a segmentation and layout understanding task. Sec. 2 gives an overview of the GMW-based pixel-wise classification of the image. In the next section, we enumerate the first level of application of CRFs - to the output of the Fisher classifier of Sec. 2, for smoothing/rectification. Secs. 3 and 4 work on the idea of CRFs for regional and global features. We then show representative results of extensive experimentation, and list some of our planned future extensions.

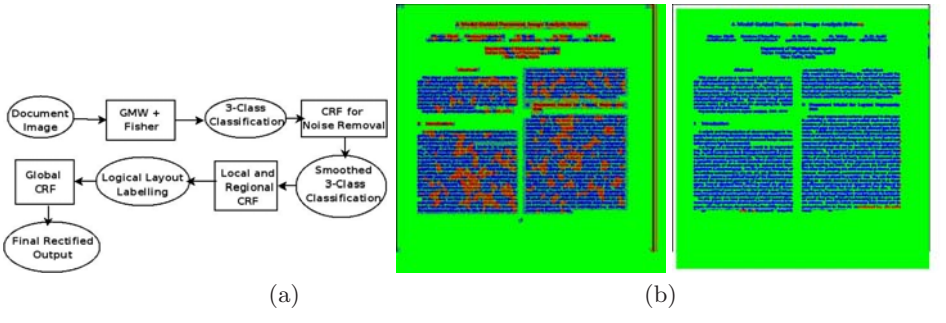


Fig. 1. (a) An overview of our multi-level CRF learning-based document segmentation and layout analysis method, and (b) GMW-based pixel-level Fisher classification of a document page, and its CRF-filtered output: Sec. 2.1

2 Pixel Classification Using Globally Matched Wavelets

We use the outputs of globally matched wavelet (GMW) filters which learn parameters to classify a pixel as being from a text, background or image region - at

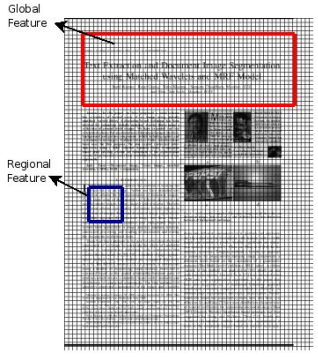
different scales (See our earlier work [2] for details.) The GMW filter responses are features, which we pass through three Fisher classifier, optimised for a two-class problem (Text-Image, Image-Background, and Text-Background classification). Each classifier gives a confidence value. We combine the outputs of all above classifiers to make the final decision. Independently at each pixel, the Fisher classifier produces a distribution over different class variable given filter outputs which is often noisy due to the overlap between neighboring classes. The next section describes a novel CRF based learning procedure for removing this noise.

2.1 CRF-Based Rectification of Fisher Classified GMW Responses

The input to this module is the output of the Fisher classifier of the previous section (Sec 2). We divide an image into a series of overlapping regions. We define features of size 3×3 which encode these error conditions and contextual information. The local content of the image at each pixel is encoded by representing binary values of gray level, the average gray level of neighbouring pixels, the gradient in horizontal and vertical direction and the output of the Fisher classifier. These extracted features are given to CRF for learning. Fig. 1(b) shows a representative example of a document page with pixel-level classification (text in blue, image in red, and background in green), and its corresponding CRF-filtered version. The CRF formulation for this stage is similar to formulation of CRF model based on local and regional image features. (The next section, Sec. 3 discusses this in detail.) *It is important to note that the present CRF-based approach scores over our original MRF-based approach [2] - in encoding contextual information better than a generative model such as an MRF, which typically encodes continuity information. Further, an MRF typically needs a large amount of training labelled data [10].*

3 Hierarchical CRFs for Regional and Global Document Features: Regional Features

We use a hierarchical CRF model to capture both regional and global features [7]. *Two image patches can be indistinguishable at a local scale, but layout relationship can provide the context for correct labelling.* We represent images as rectangular grids of patches by overlaying a grid of fixed size and then associate a hidden class label with each patch. (Fig. 2(a) illustrates the above idea, on a typical 480×640 image, with 8×8 patches. The size of a patch is often a compromise between processing complexity reduction, and labelling quality.) The CRF sequentially combines predictions or probability distributions corresponding to each image patch. (Fig. 2(b) shows templates for some regional features.) The local image content of each patch is encoded using the labels produced by local classifier f_l and position descriptors f_p. The Regional image content of each patch is encoded using regional feature f_r that encode certain patterns within a Image. We use the cell index of the location of a patch, as its position feature.



B	B	B
X	T	X
X	X	X

T	T	T
X	T	X
X	X	X

HT	HT	HT
HT	HT	HT
HT	HT	HT

I	I	I
I	I	I
I	I	I

HT	B	B	HT
HT	B	B	HT
HT	B	B	HT

HT	HT	HT	B
HT	HT	HT	B
HT	HT	HT	B

MT	MT	MT	
X	MT	X	
X	X	X	

HT	HT	X	
HT	HT	X	
X	X	X	

X	X	X
LT	LT	LT
X	X	X

HT	B	HT
HT	B	HT
HT	B	HT

HT	HT	X
HT	HT	X
HT	HT	X

X	HT	HT
X	HT	HT
X	X	X

Fig. 2. (a) An example of regional and global features for a typical document, and (b) Templates for some regional features. B: background, T: Medium Text density, I: Image, LT: Low Text density, HT: High Text density, X: Don't care.

These regional feature patterns are matched at each patch in the image. If particular pattern of feature is matched at patch in a image, then value of feature at that patch is taken as 1, otherwise 0.

Each patch is thus coded by these binary vectors along with position descriptor and labels produced by the feature classifier. We define CRF observation functions as linear functions of these binary vectors. These modalities are modelled as being independent given the patch label. Now, we provide a formulation for conditional model for patch labels that incorporates both local patch level features and regional features aggregated over neighbourhood patches. Let $x_i \in \{1, ..., L\}$ denote the label of patch i, y_i denote the $D-$ dimensional concatenated binary indicator vector of its local patch content, position descriptor and regional features. $D = (f_l, f_r, f_p)$. The conditional probability of the label x_i is then modelled as

$$p(x_i = l|y_i) \propto exp\left(-\sum_{d=1}^{D} \alpha_{dl} y_{id}\right)$$

where α_{dl} is a $D \times L$ matrix of coefficients which need to be learnt using CRF training. Thus at each patch we get a distribution over the label variables. The output of this stage classifies and segments out the logical regions of document image but it is somewhat noisy and thus needs some post-processing. The following section describes our global features based CRF learning for the same.

4 CRF Based on Global Document Features

This stage takes the classification output of the previous stage and focus on removing the errors by using global features and probabilistic layout model. For our experiments, we define a hierarchical CRF model with seven output labels,

and the parameters are learned on fully labelled images. The output labels are title block, author blocks, headings, background, paragraph, column-separator and figure. *Our models take the global image context into account by including feature functions based on probabilistic layout model of documents.This global CRF aims to remove the ambiguities that arise when patches are classified using local and contextual image features only.* For the domain of technical papers, one needs to learn the probabilistic layout of the region labels using a CRF. These global feature represent the relationship between different output labels e.g., an author block cannot come above the title block. In this way, we tried to remove the ambiguities that could arise. The Image is again considered to be divided into overlapping patches. These global feature are learnt using CRF learning. The posterior distribution over the label variables given the hidden variables, can be written as for regional features. The conditional probability of the label x_i using probabilistic layout model k is modelled as:

$$p(x_i = l|k) \propto exp\left(-\sum_{d=1}^{D} \beta_{dl} k_d\right)$$

where β_{dl} is $D \times L$ matrix of coefficients which we need to learn.

5 Experimental Results and Discussion

We have experimented with a large number of document images of technical papers, of different layouts. We show some representative results in this section. In Fig. 3, we show the three important outputs of our layout analysis procedure: the original image, the image prior to the use of the global CRF, and that after the application of the global CRF. The colour coding is as follows: Blue - Title block, Sky Blue - Author block, Green - Figure block, Red - Background, Pink - Heading, White - Paragraph, and Yellow - Column Separator. For numerous experiments with the system, the layout segmentation is almost always correct, with very few exceptions. We measure the performance of the labelling system

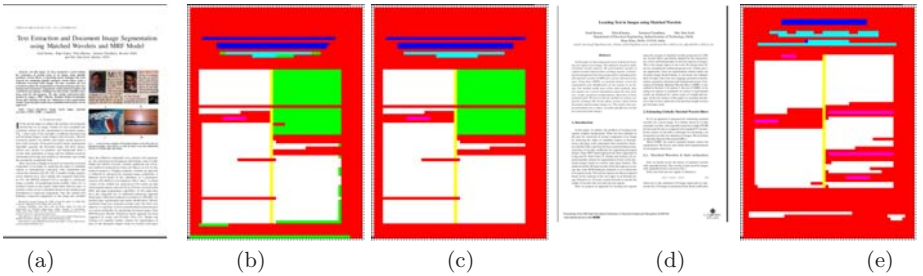


Fig. 3. (a) A document page (b) the corresponding labelled regions before, and (c) after the global probabilistic layout model application. (d) Another document page, and (e) the final segmentation output.

by considering manual ground truth (manually labelled images), and examining ROC parameters - the precision and recall rates. In most case, both these parameters are above 95%.

6 Discussion and Conclusions

This paper presents a novel integrated scheme for a general CRF-based learning framework for document image segmentation and layout analysis at many different logical levels. This starts from filtering and smoothing of the output of Fisher classified GMW outputs, and goes up to learning regional and global features, and their inter-relationships, which further aid in top-down layout analysis and document image segmentation. An interesting extension of our work will be to examine the use of Markov field aspect models [12] and CRFs [13] for a first-level pixel labelling which may be incorrect, or incomplete.

References

1. Gupta, G., Niranjan, S., Shrivastava, A., Sinha, R.M.K.: Document Layout Analysis & Classification and its Application in OCR. In: Proc. IEEE EDOCW (2006)
2. Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., Joshi, S.D.: Text Extraction and Document Image Segmentation. IEEE Transactions on Image Processing 16(8), 2117–2128 (2007)
3. Shafait, F., van Beusekom, J., Keysers, D., Bruel, T.M.: Background Variability Modeling for Statistical Layout Analysis. In: Proc. ICPR (2008)
4. Kanungo, T., Mao, S.: Stochastic language models for style-directed layout analysis of document images. IEEE Transactions on Image Processing 12(5) (2003)
5. Shilman, M., Liang, P., Viola, P.: Learning Non-generative Grammatical Models for Document Analysis. In: Proc. IEEE ICCV, pp. 962–969 (2005)
6. Tokuyasu, T., Chou, P.A.: Turbo Recognition: A Statistical Approach to Layout Analysis. In: Proc. SPIE Document Recognition and Retrieval, pp. 123–129 (2001)
7. He, X., Zemel, R.S., Carreira-Perpinan, M.A.: Multiscale Conditional Random Fields for Image Labeling. In: Proc. IEEE CVPR, pp. II:695–II:702 (2004)
8. Liang, J., Haralick, R.M., Phillips, I.T.: A Statistically based, Highly Accurate Text-Line Segmentation Method. In: Proc. ICDAR, pp. 551–555 (1999)
9. Gao, D., Wang, Y., Hindi, H., Do, M.: Decompose Document Image using Integer Linear Programming. In: Proc. ICDAR, pp. 397–401 (2007)
10. Nicolas, S., Dardenne, J., Paquet, T., Heutte, L.: Document Image Segmentation using a 2D Conditional Random Field Model. In: Proc. ICDAR, pp. I:407–I:411 (2007)
11. Shetty, S., Srinivasan, H., Beal, M., Srihari, S.: Segmentation and labeling of documents using conditional random fields. In: Proc. SPIE Document Recognition and Retrieval (2007)
12. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: Proc. IEEE CVPR (2007)
13. Verbeek, J., Triggs, B.: Scene segmentation with conditional random fields learned from partially labeled images. In: Proc. NIPS (2008)