Document Page Structure Learning for Fixed-layout E-books Using Conditional Random Fields

Xin Tao[a], Zhi Tang[ab] and Canhui Xu[ab]

[a]Institute of Computer Science and Technology, Peking University, Beijing, China;
[b]State Key Laboratory of Digital Publishing Technology, Beijing, China

ABSTRACT

In this paper, a model is proposed to learn logical structure of fixed-layout document pages by combining support vector machine (SVM) and conditional random fields (CRF). Features related to each logical label and their dependencies are extracted from various original Portable Document Format (PDF) attributes. Both local evidence and contextual dependencies are integrated in the proposed model so as to achieve better logical labeling performance. With the merits of SVM as local discriminative classifier and CRF modeling contextual correlations of adjacent fragments, it is capable of resolving the ambiguities of semantic labels. The experimental results show that CRF based models with both tree and chain graph structures outperform the SVM model with an increase of macro-averaged F_1 by about 10%.

Keywords: Page structure, Logical labeling, Conditional Random Fields, Fixed-layout document

1. INTRODUCTION

Recently, mobile reading has become an increasing requirement of document application. Due to the variety of device display sizes, electronic book (e-book) reading lacks practical usability for traditional documents like PDF, which is a widely used format for document generation and web publishing. The growing need for reflowable documents has fueled the research on converting the fixed layout documents like legacy PDF so as to enhance the reading experience. A reflowable format allows the reader to adjust the contents to various display devices with different display areas.

To ensure successful conversion of PDF to reflowable formats like EPUB or CEBX,[1] reliable document structure analysis is a crucial procedure. Document structure analysis consists of sequential stages such as physical segmentation involving blocks or lines and logical analysis including assignment of semantic labels to the segments and determination of their relationships. The task of logical structure analysis is still an open problem not only for analysis of traditional image based documents but also for born-digital documents. As an important subtask of logical structure analysis, logical labeling aims to infer the intrinsic semantic purpose of each segment. Though one may guess the logical role of an individual document segment independently without regarding others, the contextual interactions are assumed to be additionally informative.

In this paper, we take full consideration of the inherent PDF attributes including raw content streams, spatial coordinates, text patterns and typesetting information to characterize document page fragments. Besides of local evidence, inter-fragment relationship is modeled to improve performance of logical labeling. The intra-page dependencies are learnt by applying 2D CRF framework over neighborhood based graph structures. While local features exhibit less discriminative ability, the contextual information can profit the logical classification.

Further author information: (Send correspondence to Zhi Tang)
Xin Tao: E-mail: jolly.tao@pku.edu.cn
Zhi Tang: E-mail: tangzhi@pku.edu.cn, Telephone: +86 (10)82 52 97 25

Document Recognition and Retrieval XXI, edited by Bertrand Coüasnon, Eric K. Ringger,
Proc. of SPIE-IS&T Electronic Imaging, Vol. 9021, 90210I · © 2014 SPIE-IS&T
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2039492

2. RELATED WORK

With regard to layout analysis of legacy PDF document, there exist several pioneering groups during the last decade. DIVA research group proposed a reverse engineering tool XED[2] to analyze the embedded resources of PDF files and generate their physical structures in a format XCDF.[3] Based on the application of XCDF format, another interactive system Dolores[4] was presented to recover logical structure of newspaper through neural network learning mechanism. Marinai described a rule based system to identify the table of contents[5] and the notes in the text for converting certain PDF books into a reflowable XHMTL based format.[6] Chao developed a heuristic method to extract outlines, style attributes and contents which are expressed in XML for the reuse or modification of PDF document page.[7] Tang focused on the conversion between fixed-layout and fluid document with research results involving paragraph recognition,[8] mathematical formula identification,[9] graphic component recognition[10].[11] Déjean exploited different streams contained in PDF files to organize empirically the documents in blocks by XY-cut segmentation and then converted them to XML structured files.[12] Most existing geometric layout analysis are specialized for certain types of documents through heuristic top-down, bottom-up or hybrid methods.

Compared with well researched segmentation based geometric layout analysis, logical structure recovery has far less available literature due to its inherent complexity. In the field of image-based document analysis, there were various attempts in logical structure recovery with exploitation of document geometric layouts. Tsujimoto transformed the geometrical layout tree into a logical layout tree using generic rules for multi-columned documents like technical journals and newspapers.[13] Recent researches regarding logical layout analysis have considered machine learning methods as alternative remedy to avoid the inflexibility and rigidity of manually built rule systems. Rangoni used an transparent artificial neural network and resolve ambiguous results through a feedback mechanism.[14] Montreuil[15] adpoted CRF to extract logical layout of unconstrained handwritten letters. Shetty[16] used CRF to label segments of scanned documents as machine-print, handwritting and noise. It is claimed that the logical layout analysis methods have no standardized benchmarks or evaluation sets,[17] which is highly desired in this field.

Among the machine learning based extraction of structural information methods, it is noteworthy that conditional random fields (CRF) is reported to gain better performance than Hidden Markov Models,[18] or Support Vector Machines[19][20] in the fields of text processing and handwriting recognition. Other published linear chain CRF based logical structure detection also declares its effectiveness for documents in scholarly digital libraries.[21]

3. CONDITIONAL RANDOM FIELDS

3.1 Probabilistic Framework

Since the goal of document logical layout analysis is to assign a correct label for each physical fragment in a page, we can formulate this task as a classification problem. Let the fragments be indexed by i, Y_i be the multinomial random variable indicating the logical role of a fragment whose value can be taken from a label set $\mathcal{L}$, and X_i be the observations characterizing the fragment. The model $P(\boldsymbol{Y}|\boldsymbol{X})$ then describes the distribution of logical labels $\boldsymbol{Y} = \{Y_i\}$ given observations $\boldsymbol{X} = \{X_i\}$. With each vertex associated with a random variable Y_i and edges connecting correlated random variables, a graph is defined as $G =< V, E >$, where V and E denote the vertices and edges respectively. $(\boldsymbol{X}, \boldsymbol{Y})$ is a conditional random field if the variables $\boldsymbol{Y}$, when conditioned on $\boldsymbol{X}$, satisfy the Markov property with respect to G:

$$P(Y_i|\boldsymbol{X}, Y_{V\setminus i}) = P(Y_i|\boldsymbol{X}, Y_{N_i}), \tag{1}$$

where $N_i = \{j|(i,j) \in E\}$ is vertex i's neighborhood. An assignment to $\boldsymbol{X}$ is denoted by $\boldsymbol{x}$, and an assignment to a clique $c \in G$ is denoted by $\boldsymbol{x}_c$. The notations are similar for $\boldsymbol{Y}$. By the Hammersley and Clifford theorem, the conditional probability distribution $P(\boldsymbol{Y}|\boldsymbol{X})$ factorizes over G into unnormalized potential functions $\Psi_c(\boldsymbol{x}_c, \boldsymbol{y}_c)$ on maximal cliques

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{c \in G} \Psi_c(\boldsymbol{x}_c, \boldsymbol{y}_c), \tag{2}$$

where $Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}'} \prod_{c \in G} \Psi_c(\boldsymbol{x}_c, \boldsymbol{y}_c)$ is the partition function that sums over all possible assignments to $\boldsymbol{Y}$. Taking the log linear model, the potential function can be parameterized as

$$\Psi_c(\boldsymbol{x}_c, \boldsymbol{y}_c; \boldsymbol{\lambda}_c) = \exp\left\{\sum_k \lambda_{ck} f_{ck}(\boldsymbol{x}_c, \boldsymbol{y}_c)\right\} \tag{3}$$

where $\{f_{ck}(\boldsymbol{x}_c, \boldsymbol{y}_c)\}$ are feature functions of clique c associated with weights $\{\lambda_{ck}\}$. In order to reduce model complexity, the cliques can be further grouped into a set of clusters $\mathcal{C}$, where $C_p \in \mathcal{C}$ is called a clique template. Cliques belonging to a clique template C_p share the same parameters $\boldsymbol{\lambda_p} = \{\lambda_{pk}\}$.

3.2 Parameter Estimation and Inference

Given parameterization of CRF defined above, the conditional log likelihood of a dataset with known labels is formulated as

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \left(\sum_{C_p \in \mathcal{C}} \sum_{\Psi_c \in C_p} \log \Psi_c(\boldsymbol{x}_c^{(i)}, \boldsymbol{y}_c^{(i)}; \boldsymbol{\lambda}_p) - \log Z(\boldsymbol{x}^{(i)}) \right) \tag{4}$$

where n is the number of data instances.

Parameter estimation is performed by maximizing the penalized conditional log likelihood with respect to $\boldsymbol{\lambda}$. The maximization is accomplished by a quasi-Newton optimization method L-BFGS,[22] which gradually adjusts the weight vector iteratively until convergence. Inference is required to calculate $Z(\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$. We chose Tree Belief Propagation because exact likelihood can be obtained within tractable time with our model. The inference is also used to acquire the marginal probabilities of random variables $\boldsymbol{y}$ of dataset with unknown labels. Then the labels are predicted as the assignments that maximize the marginal probabilities of each y_i.

4. MODELING FOR LOGICAL ANALYSIS

Logical analysis is carried out on the scale of fragments, which are attainable in the procedure of physical segmentation by grouping the primitive contents with homogeneity and geometric proximity. In most cases, a fragment is an aggregation of adjacent basic elements, whose size is no larger than a text line. Taking fragments as the model input is an appropriate granularity for the reasons that they provide richer feature description when compared with characters and require less sophisticated physical segmentation algorithms than blocks.

It is believed that a fragment provides its inherent local observation, and expresses contextual dependencies together with its neighbors. To incorporate both the local and contextual evidence in our CRF model, two types of clique templates are introduced to determine the potential functions, namely unary potentials and pairwise potentials which are expounded in following subsections.

4.1 Unary Potentials

The unary potentials describe how likely a logical label should be assigned to a fragment isolately. Given the features of a fragment i, its unary potential can be parameterized in the form of Equation 3, where $\boldsymbol{y}_c$ in this case contains a single variable y_i. The selection of $\{f_{ck}\}$ has significant impact on discriminability of the whole model. In this work, we derive feature functions from outputs of an effective local classifier. Support Vector Machine (SVM) is known to be one of the best state-of-the-art local classifiers for its advantages of generalization properties and maximum margin nature. We convert scores of SVM classifier to posterior probabilities $p_{svm}(y_i = l|\boldsymbol{x})$ by Platt's method,[23] and define the local feature functions as

$$f_{s,l}(y_i, \boldsymbol{x}) = \mathbb{1}\{y_i = s\} \log(p_{svm}(y_i = l|\boldsymbol{x}))$$

where $s, l \in \mathcal{L}$, and $\mathbb{1}\{y_i = s\}$ denotes an indicator function which equals 1 if $y_i = s$ and 0 otherwise. We also share the parameters $\{\lambda_{s,l}\}$ over all the fragments.

To obtain $\log(p_{svm}(y|x))$, we need to train a SVM model by providing concrete local observations. We distill basic attributes from PDF files using a commercial parser (open source tools like PDFBox[24] can serve the same

purpose). Four types of observations are further derived from these PDF attributes , including spatial coordinates, text patterns, typesetting and raw content streams. Analogous to image document analysis, geometric observations are extracted from fixed-layout documents. Besides, fixed-layout documents offer precise text of textual contents. Typesetting information such as font sizes are included to enrich discriminating capacity for classifiers. We choose a set of 31 observations to characterize fragments, which are standardized with means of 0 and standard deviations of 1. These observations are described detailly in Table 1

Table 1. Observations for unary potentials

Observation type	*Observation name*	*Description*
Geometric	Height	normalized height
	Width	normalized width
	Area	normalized area
	Aspect ratio	min(width, height) / max(width, height)
	Position	relative position of each fragment within a page
Textual	Has digit	whether text contains digit
	All digit	whether all letters are digits
	Is uppercase	whether all letters are uppercase
	Math	whether text contains math symbols or greek letters
	Digital number	detect the pattern of containing or being digits
	Figure caption pattern	whether text has figure caption pattern
	List item pattern	whether text has list item pattern
	Above fragment text pattern	fragment above has certain text pattern
Typesetting	Font size	greater/smaller/equal compared with dominant font size
	Indent level	discretized indent level
Content type	Source type	raw content type of fragment, e.g. text, image or path
	Is above fragment image	fragment above belongs to raw content type image

4.2 Pairwise Potentials

The pairwise potentials reflect the semantic dependencies between connected fragments, conditioned on the observations. In graph G, pairwise potentials are defined on cliques involving two adjacent random variables. It is expected that the interactive influence will regularize the probabilities estimated by unary potentials. Given that two random variables i and j are connected in the graphical model, their pairwise feature functions are defined as

$$f_{s,t,k}(y_i, y_j, \boldsymbol{x}) = \mathbb{1}\{y_i = s, y_j = t\} g_k(\boldsymbol{x})$$

where $s, t \in \mathcal{L}$, $\{g_k(\boldsymbol{x})\}$ are the pairwise observations indexed by k. $\mathbb{1}\{y_i = s, y_j = t\}$ is an indicator function which equals 1 if $y_i = s$ and $y_j = t$. With parameters of the feature functions shared across the pairwise clique template, the potentials are also parameterized in the form of Equation 3

A set of $K = 6$ pairwise observations are extracted for each pair of adjacent fragments like geometric relationships, typesetting and raw content streams. These observations are normalized between 0 and 1. The specific pairwise observations g_k are described in Table 2. The pairwise features are required for modeling the affinity of logical labels of a fragment pair in our application. Generally, when the transition of labels is a common occurrence, the learning process would assign the corresponding weight a larger value to maximize the objective function. Therefore, the weights are trained to prefer label combinations conforming to known data.

4.3 Graph Structure

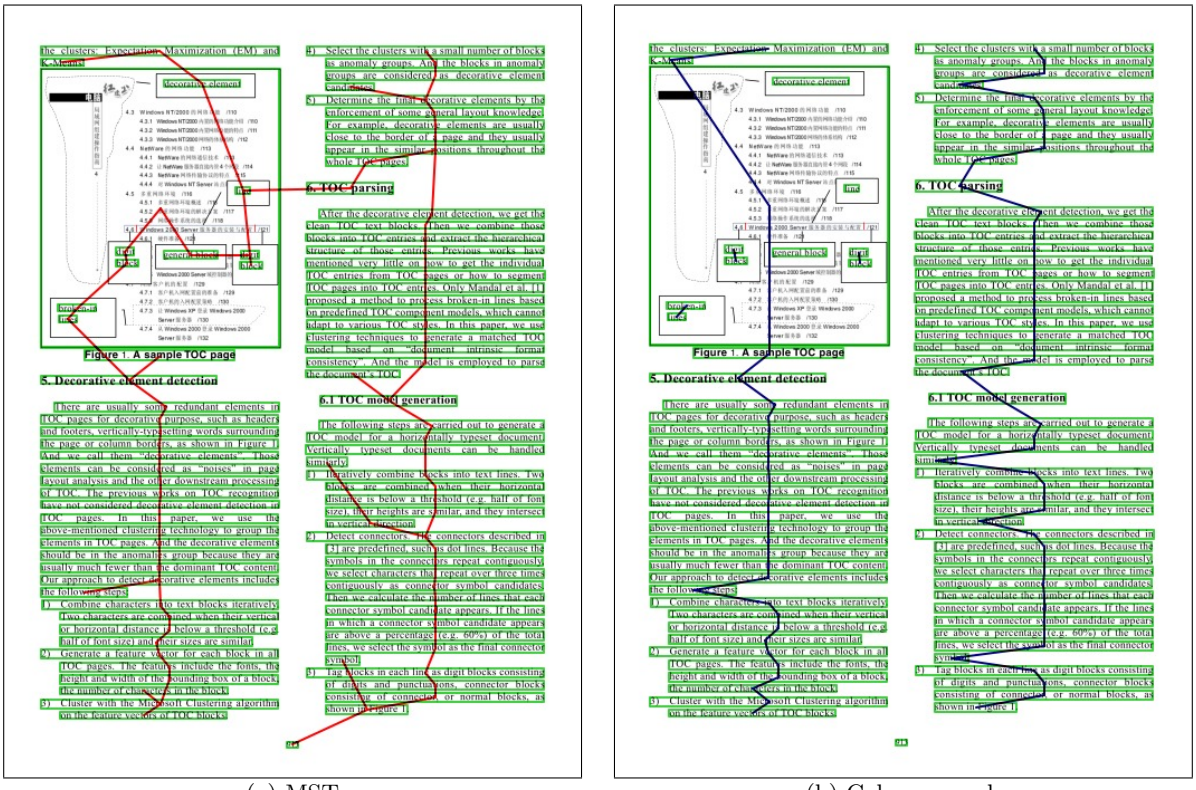
The legitimacy of graph structure is also important to build the model. Heterogeneous edges in clique templates can confuse parameter estimation and weaken the discernment of model. As for the problem of logical labeling, we focus on the locality of fragments within a document page. The intuition is that either neighboring fragments follow semantic consistency, or there exist certain regularities in transitions of their logical roles.

Table 2. Observations for pairwise potentials

Observation type	*Observation name*	*Description*
Geometric	Alignment	Alignment properties including left, right or central alignment
	Y distance	Distance of the two fragments along y axis
	Height ratio	ratio between the heights of the fragments
	Width ratio	ratio between the widths of the fragments
Typesetting	Font size	whether the two fragments have same font size
Content source	Source type	whether the two fragments have same raw content type

To treat a page as a graph, centroids of bounding boxes of the fragments are extracted as vertices, and edges are created between these vertices. With the edges measured by Euclidean distance, a minimal spanning tree (MST) is constructed to establish neighborhood of each fragment within the page. The minimal spanning tree is a global optimum that ensures the sum of the edges distances is minimal among all possible spanning trees of the same graph.

An alternative graph structure, called column graph, is chain-based rather than tree based graph like MST. It is observed that in most languages, document contents have vertical aligned layout. For each fragment, we search for a nearest neighbor that is below the current fragment and overlaps it along the horizontal orientation to establish an edge. The column graph is more locally constructed compared with the minimum spanning tree, aiming to mimic the natural reading order.



(a) MST (b) Column graph

Figure 1. An example of MST and column graph structure construction within a PDF document page. Fragments are bounded with rectangle boxes. The graph structure is depicted by solid lines connecting the centroids of fragments.

We build the CRF model from these two structures: minimum spanning tree graph and column graph. Figure 1 visualizes the construction of both MST and column graph on bounded fragments of a two column

PDF document page. The minimum spanning tree graph expresses geometric adjacency of the fragments, while column graph simulates the logical order.

5. EXPERIMENTAL RESULTS

5.1 Experimental Setup

To accomplish precise quantitative evaluation for logical structure extraction methods, a representative data set with complete physical and logical ground-truth information is indispensable, though its construction can be very time-consuming. A ground-truthing tool can facilitate the labeling process. In our work, a GUI application based on wxpython is developed to accelerate manual annotation of the dataset.

With regard to data source, 124 PDF document pages are selected from 25 e-books in English and Chinese at the proportion of 1:1. Chinese books are provided by Founder Apabi digital library, and English books are selected among books crawled from web. The layouts of the selected pages within each e-book have single-columned and double-columned styles with the distribution of 1:1. The types of these books vary from social or scientific library books to academic journals and magazines. Using the ground-truthing tool, we manually marked 4642 fragments in total with logical labels. There are three separate content streams parsed from PDF: text, image and path primitives. Every fragment contains only one kind of content stream source, e.g. body text fragments include only primitives from text content stream. The positions of bounding boxes, unique IDs, and all children IDs of fragments are recorded in an XML file.

A set of total 13 semantic labels are defined, including body text, title, figure, figure caption, figure caption continuation, list item, list item continuation, equation, page number, footer, header, footnote, and marginal note. Each fragment is assigned with a corresponding semantic label. The label "figure caption" actually indicates the first line of a figure caption, distinguishing from its continuations. The label "list item" is similarly defined. Marginal note here refers to complementary texts at the left or right margin within the page. Footnote contains a note of reference, explanation or comments beginning with numerical or customized marks such as $*$ or † near the bottom of the page. The original PDF documents, along with their physical and logical ground-truth are accessible publicly from `http://www.icst.pku.edu.cn/cpdp/data/marmot_data.htm`.

5.2 Evaluation

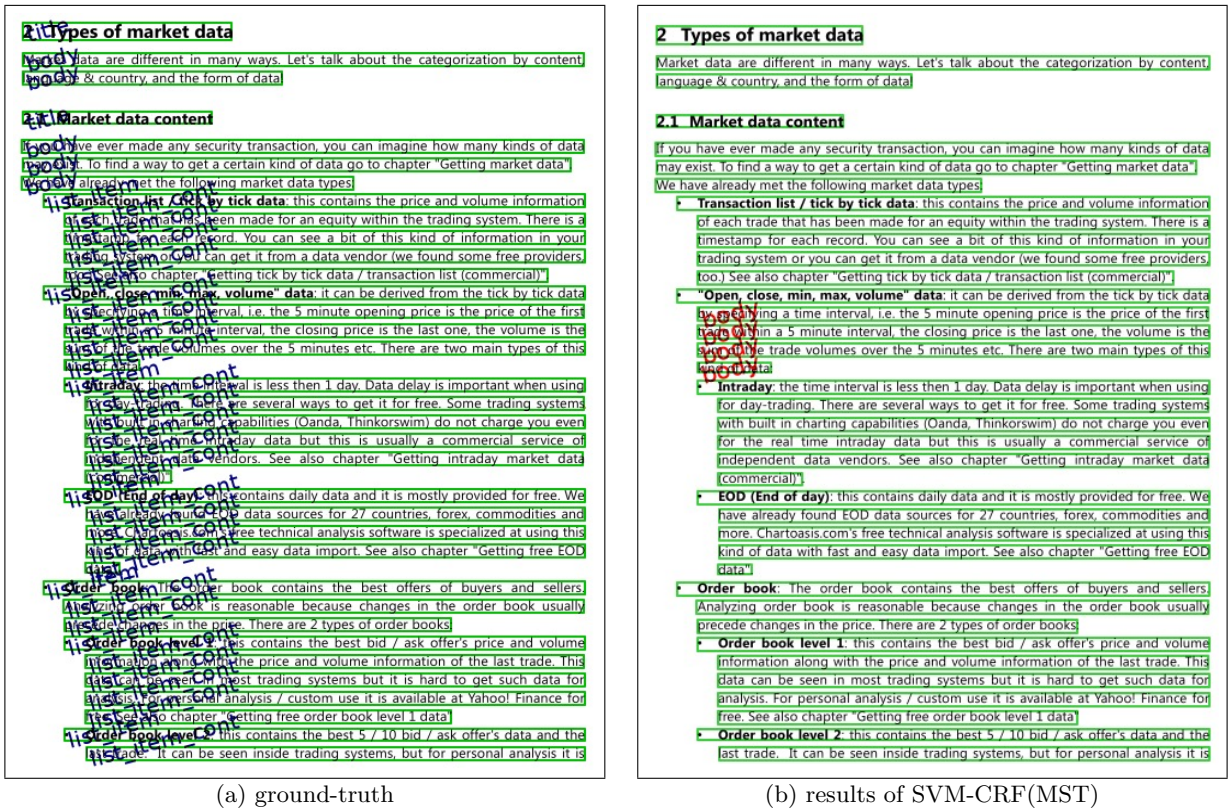
The performance is evaluated on the fragments using precision P, recall R and F_1-measure defined as $\frac{2 \cdot P \cdot R}{P+R}$. Among 13 semantic labels, as can been seen from Table 3, the distribution of fragments over each semantic label is highly imbalanced. Majority of the fragments belong to body text, which in this experimental setting possess a percentage of 65.5%. Hence, accuracy measure results can be misleading. More comprehensive metrics including macro- and micro-averaged F_1 are used respectively. Macro-averaged metrics weigh each label equally and compute their arithmetic mean, and micro-averaged metrics weigh each fragment equally and calculate the arithmetic mean.

To evaluate the proposed method, both unstructured and structured classifiers, including SVMs and CRFs are compared. The SVM models are trained in a one-against-all manner for multi-class recognition with Radial Basis Function (RBF) kernel. Probability estimates of the SVM models are calculated using five-fold cross-validation, and then fed to CRF models to generate unary potentials. CRF models are trained employing the graph structures described in 4.3. Given that the unary and pairwise feature sets are kept unchanged, the performance is able to reflect the effectiveness of the two graph structures. All the results regarding precision, recall and F_1-measure are averaged over 10 trials. Different SVM and CRF models are trained and tested across the trials. For each trial, the total 124 PDF document pages are divided randomly into training and testing sets in a ratio of 2:1.

Table 3 summarizes the performance of the models mentioned above. As can be seen from the table, body text dominates most other semantic labels. Under such condition, the marco-averages are able to provide more informative results, among which the CRF model clearly proves its improved overall performance over SVM model about 10% for precision, recall as well as F_1-measure. The micro averaged metrics increased by around 4%, which implies most improvements occurred in minority labels other than body text. Figure 2 illustrates ground-truth and classification results of SVM-CRF(MST) on a sample page.

Table 3. Comparative performance between SVM and SVM-CRF methods with different graph structure

Label	#Frag	SVM			SVM-CRF(COL)			SVM-CRF(MST)		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Body	3041	89.49	95.37	92.33	92.90	97.00	**94.91**	91.42	96.28	93.79
Equation	196	92.55	93.29	**92.92**	90.87	93.77	92.30	88.98	87.48	88.23
Figure	208	62.48	68.43	65.32	68.89	64.69	66.72	73.96	70.69	**72.29**
FigureCap	92	81.48	66.67	**73.33**	79.24	62.69	70.17	79.64	62.86	70.26
FigureCapCont	154	37.15	25.98	30.57	41.02	41.02	**41.02**	36.77	38.03	37.39
Footer	41	51.72	30.61	38.46	73.57	70.07	71.78	75.97	79.67	**77.78**
Header	125	62.13	55.21	58.46	74.94	80.39	77.57	81.28	79.05	**80.15**
ListItem	159	97.67	83.78	**90.19**	97.08	81.33	88.51	97.74	70.47	81.89
ListItemCont	188	43.38	24.10	30.99	87.79	63.93	73.99	90.04	74.45	**81.51**
Marginal	100	86.67	82.76	84.67	98.14	84.08	**90.57**	98.05	81.13	88.79
Note	83	57.44	39.86	47.06	83.12	45.55	58.85	83.83	49.30	**62.08**
PageNum	106	86.93	86.44	86.69	84.62	90.11	**87.28**	84.55	90.13	87.27
Title	149	78.31	76.97	77.63	83.11	80.81	**81.95**	83.44	78.97	81.14
Micro-Averages	-	84.59	84.59	84.59	88.78	88.78	**88.78**	87.93	87.93	87.93
Macro-Averages	-	71.34	63.81	66.82	81.18	73.51	76.58	81.97	73.74	**77.12**



(a) ground-truth (b) results of SVM-CRF(MST)

Figure 2. (a) Ground-truth of a sample page. The logical labels are displayed by rotated blue text. (b) Results of SVM-CRF(MST) model. Only classification errors are displayed in red.

Some of the logical labels have instances with explicit features. Consequently, they are accurately recognized by the SVM models. Whereas other semantic classes are easily confused with body text due to lack of local characteristics. The ambiguities are significantly alleviated when CRF model is adopted. For example, it is hard to tell list item continuations from body texts, except that they are continuous in logical order and follow an list item line initialized with an obvious bullet. SVM model classifies these fragment independently, and has a relatively lower performance as expected (with F_1 measure of 30.99% for label ListItemCont). It is noteworthy that this affinity of neighboring labels is better captured by the CRF model (with F_1 measure of 73.99% and 81.51% for label ListItemCont). Similar tendency can be observed over other labels.

Though both the minimum spanning tree and column CRF models precede the SVM model, neither of them obviously defeats the other on our dataset. We attribute this result to the limited differences between their graph structures. It is expected that combining various purposed graph structures could contribute to further performance improvement.

6. CONCLUSION AND FUTURE WORK

This paper has proposed a conditional random field method for the logical structure analysis of born-digital fixed-layout documents. In addition to local evidence of individual fragment, relationship between fragments is also incorporated in the CRF model. The feature engineering is carried out by exploiting only the inherent PDF attributes. The experimental results reveal that CRF model significantly outperforms the non-structured SVM model. Though the logical labels are highly imbalanced, CRF model still benefits from neighboring dependencies and achieve remarkable reduction of confusions between ambiguous semantic classes. By virtue of the generality and flexibility of CRF model, we believe that it is promising to achieve better performance by extending feature sets and exploring higher-level dependencies.

ACKNOWLEDGMENTS

This work was supported by National Basic Research Program of China(NO. 2012CB724108). The authors would like to thank Dr. Adrien Delaye for his insightful advice.

REFERENCES

[1] Qiu, R., Tang, Z., Gao, L., and Yu, Y., "A novel xml-based document format with printing quality for web publishing," in [*IS&T/SPIE Electronic Imaging, Imaging and Printing in a Web 2.0 World; and Multimedia Content Access: Algorithms and Systems IV*], 75400J–75400J (2010).

[2] Hadjar, K., Rigamonti, M., Lalanne, D., and Ingold, R., "Xed: a new tool for extracting hidden structures from electronic documents," in [*Proceedings of International Workshop on Document Image Analysis for Libraries*], 212–224 (2004).

[3] Bloechle, J., Rigamonti, M., Hadjar, K., Lalanne, D., and Ingold, R., "Xcdf: A canonical and structured document format," in [*Document Analysis Systems VII*], 141–152, Springer (2006).

[4] Bloechle, J., Rigamonti, M., and Ingold, R., "Ocd dolores-recovering logical structures for dummies," in [*10th IAPR International Workshop on Document Analysis Systems (DAS)*], 245–249 (2012).

[5] Marinai, S., Marino, E., and Soda, G., "Table of contents recognition for converting pdf documents in e-book formats," in [*Proceedings of the 10th ACM symposium on Document engineering*], 73–76 (2010).

[6] Marinai, S., Marino, E., and Soda, G., "Conversion of pdf books in epub format," in [*International Conference on Document Analysis and Recognition (ICDAR)*], 478–482 (2011).

[7] Chao, H. and Fan, J., "Layout and content extraction for pdf documents," in [*Document Analysis Systems VI*], 213–224 (2004).

[8] Fang, J., Tang, Z., and Gao, L., "Reflowing-driven paragraph recognition for electronic books in pdf," in [*IS&T/SPIE Electronic Imaging, Document Recognition and Retrieval XVIII*], 78740U–78740U (2011).

[9] Lin, X., Gao, L., Tang, Z., Lin, X., and Hu, X., "Mathematical formula identification in pdf documents," in [*International Conference on Document Analysis and Recognition (ICDAR)*], 1419–1423 (2011).

[10] Xu, C., Tang, Z., Tao, X., Li, Y., and Shi, C., "Graph-based layout analysis for pdf documents," in [*IS&T/SPIE Electronic Imaging, Imaging and Printing in a Web 2.0 World IV*], 866407–866407 (2013).

[11] Xu, C., Tang, Z., Tao, X., and Shi, C., "Graphic composite segmentation for pdf documents with complex layouts," in [*IS&T/SPIE Electronic Imaging, Document Recognition and Retrieval XX*], 86580E–86580E (2013).

[12] Déjean, H. and Meunier, J., "A system for converting pdf documents into structured xml format," in [*Document Analysis Systems VII*], 129–140 (2006).

[13] Tsujimoto, S. and Asada, H., "Major components of a complete text reading system," *Proceedings of the IEEE* **80**(7), 1133–1149 (1992).

[14] Y. Rangoni, Y. and Belaïd, A., "Document logical structure analysis based on perceptive cycles," in [*Document Analysis Systems VII*], 117–128 (2006).

[15] Montreuil, F., Grosicki, E., Heutte, L., and Nicolas, S., "Unconstrained handwritten document layout extraction using 2d conditional random fields," in [*International Conference on Document Analysis and Recognition (ICDAR)*], 853–857, IEEE (2009).

[16] Shetty, S., Srinivasan, H., Beal, M., and Srihari, S., "Segmentation and labeling of documents using conditional random fields," in [*IS&T/SPIE Electronic Imaging, Document Recognition and Retrieval XIV*], 65000U–65000U, International Society for Optics and Photonics (2007).

[17] Paaß, G. and Konya, I., "Machine learning for document structure recognition," in [*Modeling, Learning, and Processing of Text Technological Data Structures*], 221–247 (2012).

[18] Peng, F. and McCallum, A., "Information extraction from research papers using conditional random fields," *Information Processing & Management* **42**(4), 963–979 (2006).

[19] Han, H., Giles, C., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E., "Automatic document metadata extraction using support vector machines," in [*Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*], 37–48 (2003).

[20] Delaye, A. and Liu, C., "Context modeling for text/non-text separation in free-form online handwritten documents," in [*IS&T/SPIE Electronic Imaging, Document Recognition and Retrieval XX*], 86580C–86580C (2013).

[21] Luong, M., Nguyen, T., and Kan, M., "Logical structure recovery in scholarly articles with rich document features," *International Journal of Digital Library Systems (IJDLS)* **1**, 1–23 (2010).

[22] Liu, D. and Nocedal, J., "On the limited memory bfgs method for large scale optimization," *Mathematical programming* **45**(1-3), 503–528 (1989).

[23] Platt, J., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers* **10**(3), 61–74 (1999).

[24] "Apache pdfbox - a java pdf library." `http://pdfbox.apache.org`.