2014 11th IAPR International Workshop on Document Analysis Systems

Logical Labeling of Fixed Layout PDF Documents Using Multiple Contexts

Xin Tao, Zhi Tang, Canhui Xu and Yongtao Wang
Institute of Computer Science and Technology
Peking University
Beijing, China
Email: {jolly.tao, tangzhi, wyt}@pku.edu.cn, ccxu09@yeah.net

***Abstract*—The task of logical structure recovery is known to be of crucial importance, yet remains unsolved not only for image based document but also for born-digital document system. In this work, the modeling of contextual information based on 2D Conditional Random Fields is proposed to learn page structure for born-digital fixed-layout documents. Heuristic prior knowledge of Portable Document Format (PDF) content and layout are interpreted to construct neighborhood graphs and various pairwise clique templates for the modeling of multiple contexts. By integrating local and contextual observations obtained from PDF attributes, the ambiguities of semantic labels are better resolved. Experimental comparisons for six types of clique templates has demonstrated the benefits of contextual information in logical labeling of 16 finely defined categories.**

***Keywords*—logical labeling, Conditional Random Fields, born-digital fixed-layout document**

I. Introduction

The conversion of fixed-layout documents like Portable Document Format (PDF) files to reflowable documents like ePub has become increasingly indispensable to enable the usability of born-digital fixed-layout documents for mobile reading. The conversion commonly involves segmenting the documents into smaller pieces and reorganizing them according to their semantic roles. The semantic roles have to be recognized and forwarded to the reflowable documents. Logical labeling is a prerequisite step toward semantic recovery of fixed-layout documents. The task of logical labeling can be regarded as exploring the intrinsic semantics of document page contents. E-book pages, for instance, have a set of distinguishable logical classes such as titles, body text, figures and tables, etc..

In the past decades, majority of research has been devoted to achieving reliable segmentation, especially for image based document pages, and there also exists increasing attention on born-digital fixed-layout documents. With regard to conversion of legacy PDF documents, DIVA research group proposed a reverse engineering tool XED [1] to analyze the embedded resources of PDF files and generate their physical structures in a format XCDF [2]. Based on the application of XCDF format, another interactive system Dolores [3] 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 [4] and the notes in the text for converting certain PDF books into a reflowable XHMTL based format [5]. Recently, Tang focuses research on the conversion between fixed-layout and fluid document with research results involving paragraph recognition [6], mathematical formula identification [7], graphic component recognition [8] [9].

However, the crucial determination of the semantic roles of the contents, also known as logical labeling, remains an open problem. Compared with well researched segmentation, logical labeling has far less available literature due to its inherent complexity. Rangoni used an transparent artificial neural network and resolve ambiguous results through a feedback mechanism [10]. With the help of an OCR engine, Luong uses a linear chain based CRF model to detect logical structures of documents from scholarly digital libraries [11]. It is claimed that the logical labeling methods have no standardized benchmarks or evaluation sets[12], which is highly desired in this field.

Intuitively, it is possible to infer the semantic roles of some document contents independently. For instance, a number with no more than three digits in the corner of a page tends to be identified as a page number. It is more often the case that semantics are clarified more likely by relative relationships with their neighboring contents. On one hand, the intersections between latent semantics reflect which logical classes conventionally appear together. On the other hand, adjacent fragments exhibit relative similarities and diversities in visual perception, serving the purpose of readability. But these prior knowledge in interaction relationship of neighbors is generally not considered in local unstructured classifiers.

We hypothesize that relational dependencies of interconnected variables in logical labeling problem can be effectively characterized by structured models. As a special form of probabilistic graph model, Conditional Random Fields (CRF) is explored in this scene to model dependencies between random variables, can naturally fit the aforementioned intuitions. CRF has several appealing aspects: it can directly model the conditional probability distribution of labels given the observed data; and it considers dependencies between the labels. CRF has already achieved extensive successes in various application fields like natural language processing [13], computer vision [14] and image document analysis [15] [16]. Usually, research on application of CRF specify structures like linear chain for sequence labeling or 2D lattice for image processing.

In our case, the contextual graph structure needs to be explored, which is the aim of this paper. It is also our concern that whether incorporating contexts will have more beneficial performance than using a local classifier alone. Provided with adequate prior knowledge description, we expect this framework modeled with several contexts is more expressive

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.54

for the problem of logical labeling of born-digital fixed-layout documents. In this work, a 2D CRF framework is applied to model the hidden semantics of document page fragments with considering the joint classification of independent semantic labels classification. We introduce six types of contexts specifying the cooperation of prior knowledge into graph structure. The CRF framework and problem formation is introduced in Section 2. The prior knowledge based contexts definition is proposed in Section 3. Its application on a ground-truthed PDF document dataset and performance comparison is presented in Section 4. Finally, section 5 concludes the paper.

II. Conditional Random Field

A. CRF 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 $\boldsymbol{X} = \{X_i\}$. A graph $G =< V, E >$ can be built with each vertex associated with a random variable Y_i. $(\boldsymbol{X}, \boldsymbol{Y})$ is a conditional random field if the variables $\boldsymbol{Y}$, when conditioned on $\boldsymbol{X}$, satisfy the Markov property with respect to G:

$$P(Y_i|\boldsymbol{X}, \boldsymbol{Y}_{V\backslash i}) = P(Y_i|\boldsymbol{X}, \boldsymbol{Y}_{N_i}), \tag{1}$$

where $V\backslash i$ denotes all vertices except i, and $N_i = \{j|(i,j) \in E\}$ is i's neighborhood. 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 summing over all possible assignments of $\boldsymbol{Y}$.

B. Modeling for Logical Labeling

A document page is considered as a graph where the vertices are the content fragments and the edges stand for their relationships. We take both local and contextual evidence into consideration in order to obtain a discriminative model. The local observations and neighboring interactions are correspondingly expressed with unary and binary cliques in the CRF model. A label set $\mathcal{L}$ of total 16 semantic logical labels is defined, including body text, equation, figure, figure annotation, figure caption, figure caption continuation, list item, list item continuation, footer, header, marginal, notes, table cell, table caption, page number, title. Each fragment can be assigned with a corresponding logical label through model inference.

1) Unary Potentials: Unary potentials are estimates based on only local observations over the random variables. In the problem of logical labeling of document page, we adopt a two step strategy. In the first step, an SVM classifier is first trained as the local classifier. Local observations are derived from PDF attributes, including spatial measures (height, width, area, aspect ratio and position), text patterns (digit, uppercase, math symbols, explicit keywords), typesetting information (font size, indent level) and primitive types. A set of 73 observations are extracted from each fragment and its neighbors as input to train the SVM classifier. The output of SVM is transformed using Platt's method to provide posterior probability estimates $p_{svm}(y_i|\boldsymbol{x}_i)$. Then feature functions for CRF are derived through combinations of labels and probabilities as:

$$f_{s,l}(y_i, \boldsymbol{x}_i) = \mathbb{1}\{y_i = s\} p_{svm}(y_i = l|\boldsymbol{x}_i) \tag{3}$$

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. The potential function of unary cliques can be parameterized with log-linear model as

$$\Psi(y_i, \boldsymbol{x}_i) = \exp\{\sum_{s,l \in \mathcal{L}} \lambda_{s,l} f_{s,l}(y_i, \boldsymbol{x}_i)\} \tag{4}$$

where $\lambda_{s,l}$ is shared across all unary cliques.

2) Pairwise Potentials: In addition to local observations, pairwise relationships are also exploited to capture possible dependencies between fragments. In order to control complexity of the model, we choose feature functions related to only the logical labels for each context type:

$$f_{s,t,k}(y_i, y_j, \boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{1}\{y_i = s, y_j = t\} \tag{5}$$

where k indexes the type of the context. Feature functions in this form are expected to capture frequent co-occurrences of logical labels without regard to the observations over fragments.

Similar to the unary situation, the potential function is parameterized as

$$\Psi_k(y_i, y_j, \boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\{\sum_{s,t,k} \lambda_{s,t,k} f_{s,t,k}(y_i, y_j, \boldsymbol{x}_i, \boldsymbol{x}_j)\} \tag{6}$$

Unlike in the unary potentials, an edge in the graph structure may carry information of multiple clique types. As a result, the parameters are shared across only the *same* context type. The context types are explained detailly in the next section.

C. Inference and Parameter Estimation

In the construction of graph using multiple contexts, an edge may represent two or more relationships. Because the reuse of edges will lead to a cyclic graph, Loopy Belief Propagation (LBP) is adopted as the inference method during prediction. The classification is performed by choosing the labeling that maximizes the joint probability of the model in log-space.

Parameter estimation is carried out by maximizing the Pseudo-Likelihood (PL) which is an approximate objective function with respect to $\boldsymbol{\lambda}$. The PL is an local approximation of the exact likelihood, following the assumption that the objective function depends on conditional probabilities of single variables. Normalization constants of each random variable in PL involves only its neighborhood, thus the computation is very efficient. The maximization of PL is accomplished by a quasi-Newton optimization method L-BFGS[17], which gradually adjusts the parameter vector iteratively until convergence.

III. Prior Knowledge Based Contexts

Rather than isolated from each other, the document contents in a page are correlated implicitly, resulting from conventions of writing styles and typesetting systems. Furthermore, relationships between text and text are intuitively different from those between non-text and text. With the flexibility of the CRF framework, various contexts motivated from prior knowledge can be incorporated in the model. In PDF documents, the primitive contents are categorized as text, images or graphics. The fragments are determined as text or non-text according to the primitive types of their components. In this section, we derive multiple contexts from both PDF primitive types and common layouts.

A. Basic Context

The most straightforward relationships are those between neighboring fragments. Two types of neighborhood are adopted in this work, namely *minimum spanning tree graph (MST)* and *column graph (COL)*. With the edges measured by Euclidean distance, the minimum spanning tree graph is a *globally* optimal tree structure that ensures the sum of the edge distances is minimal among all possible spanning trees over the page graph. The column graph tries to simulate the relationship of contents in the *local* scale of logical flow. As illustrated in Figure 1(a) and Figure 1(b), these two graph structures represent general relationships without knowledge other than geometric adjacency.

B. Context Between Text and Text

The majority of document contents are text. In modern typesetting systems, it is natural to gather continuous text contents and arrange them in a neat manner so that their homogeneity is easily recognized by the readers. On the other hand, change of logical roles may also leave obvious sign in page layout to avoid semantic confusion (enhance readability).

One useful layout pattern related to this prior knowledge is alignment. It is observed that adjacent fragments belonging to the same paragraph are often aligned on the left or right, while a “zigzag” indicates possible transition of semantics. Two graph structures are derived from alignment between neighboring text as *text and text alignment (TT-A)* and *text and text non-alignment (TT-N)*. They are supposed to express the maintenance and transition of logical roles respectively. Figure 1(c) and 1(d) illustrate them respectively.

C. Context Between Non-Text and Text

Although non-text contents occupy relatively less portion in documents, interactions between non-text and text contents are informative as well. The co-occurrence of non-text and text in some degree implies the logical roles of both ends. In this work attention is mainly paid to non-text contents that are possible to serve as figures.

A common combination of non-text and text contents is figure and its caption, if available. In such combination, the non-text content has significant area and explicit pattern like “figure” are expected to appear in the text content. This prior knowledge is modeled as *non-text and text cascade (NT-C)*. In the situation of figure with annotations, non-text contents may also overlap with text contents, which is modeled as *non-text and text overlap (NT-O)*. Examples of NT-C and NT-O are shown in Figure 1(e) and Figure 1(f).

IV. Experimental Results and Discussion

A. Experimental Setup

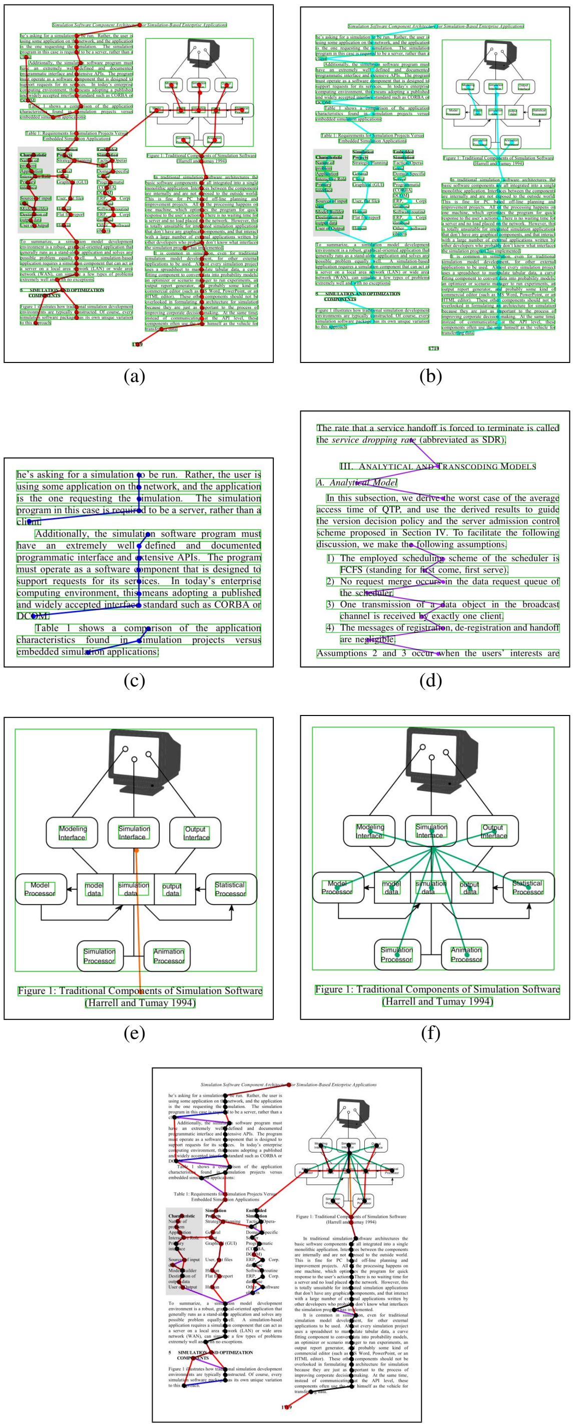
Our experimental data consists of a collection of 244 PDF document pages selected from 35 e-books in English and Chinese, which cover a wide range of layout styles for assessing the learning ability of the proposed model. Chinese books come from Founder Apabi digital library, and English books are selected among books crawled from web. It is known that there exists no standardized benchmarks or evaluation sets, which is time consuming to construct. However, we provide a ground-truthing tool to fill this gap in labeling process. A GUI application based on wxpython is developed to facilitate manual annotation of the dataset, which is accessible publicly from http://www.icst.pku.edu.cn/cpdp/data/marmot_data.htm. Total 11347 fragments are manually marked by using the ground-truthing tool, and the physical fragments are further tagged with a set of 16 semantic logic labels, including body text, title, figure, figure annotation, figure caption, figure caption continuation, list item, list item continuation, table cell, table caption, equation, page number, footer, header, footnote, and marginal note. The total 244 PDF document pages are divided randomly into training and testing sets in a ratio of 2:1. The SVM classifier is trained with Radial Basis Function (RBF) kernel, and 5-fold cross validation is used to obtain its probability estimates.

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 16 semantic labels, as can been seen from the table, 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 50.5%. Hence, accuracy measure results can be misleading. More comprehensive metrics including macro- and micro-averaged F_1 are used respectively. Macro-averaged weigh each label equally and compute their arithmetic mean, and micro-averaged weigh each fragment equally and calculate the arithmetic mean.

B. Effects of Contexts

The contexts proposed in section III are experimentally compared over their classification performance. Total 10 configurations of contexts, listed in Table II are evaluated to explore their contribution. For each configuration, performance is measured by micro-and macro-averaged F_1 over all the labels. The detailed performance over all the labels of configuration 9 which achieves the highest micro-averaged F_1 is given in table I.

The first experiment does not use any pairwise context. Experiment 2-7 explore the contribution of individual contexts. Experiment 8-11 show the performance of merged MST and COL graph structure, with or without additional contexts. As can be seen, the *MST*, *COL* and *TT-A* contexts are relatively more beneficial, because of their similarities to original logical flow in most part of the document. Effects of the *NT-C*



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Contexts motivated from prior knowledge. (a) minimum spanning tree (MST). (b) column graph (COL). (c) text and text alignment (TT-A). (d) text and text non-alignment (TT-N). (e) non-text and text cascade (NT-C). (f) non-text and text overlap (NT-O). (g) combination of all contexts.

TABLE I. PERFORMANCE OF CRF-SVM METHOD WITH MUTIPLE CONTEXTS

Label	#Frag	CRF-SVM		
		Precision	Recall	F_1
Body	5752	94.88	94.29	94.58
Equation	438	91.67	94.48	93.05
Figure	265	95.51	98.84	97.14
FigureAnnot	474	67.69	94.62	78.92
FigureCap	243	89.86	88.57	89.21
FigureCapCont	223	79.66	55.29	65.28
Footer	39	83.33	93.75	88.24
Header	262	95.18	91.86	93.49
ListItem	198	89.55	81.08	85.11
ListItemCont	320	83.72	88.52	86.06
Marginal	121	100.00	100.00	100.00
Note	69	33.33	20.69	25.53
PageNum	235	97.33	98.65	97.99
TableCap	64	100.00	46.67	63.64
TableCell	2395	94.36	96.23	95.28
Title	249	87.21	90.36	88.76
Micro-Averages	-	92.70	92.70	92.70
Macro-Averages	-	86.45	83.37	83.85

TABLE II. PERFORMANCE OF CONFIGURATIONS WITH MUTIPLE CONTEXTS

configuration	basic	texttext	non-texttext	micro	macro
1	-	-	-	90.44	79.64
2	MST	-	-	91.80	82.88
3	COL	-	-	91.91	82.13
4	-	TT-A	-	92.06	82.41
5	-	TT-N	-	91.01	80.29
6	-	-	NT-C	90.70	79.36
7	-	-	NT-O	90.67	79.60
8	MST+COL	-	-	92.61	84.52
9	MST+COL	TT-A+TT-N	-	92.70	83.85
10	MST+COL	-	NT-C+NT-O	92.55	84.17
11	MST+COL	TT-A+TT-N	NT-C+NT-O	92.67	83.90

and *NT-O* contexts are barely noticeable due to their scarce occurrences.

Experiment 7 demonstrates a further improvement by combining the basic contexts of type *MST* and *COL*. The explanation of this result may be that the *COL* context works well in most part of the page while the *MST* context remedies its defects where chain-like structure is not proper to capture the layout.

Unfortunately, the collaboration of TT-A, TT-N, NT-C and NT-O does not seem clear. Among configuration 7-11, none of them absolutely defeats the others in both micro- and macro-averaged f_1 measure. It is believed that lack of appropriate features other than the bias constant for these contexts should be the reason of the oscillating results. Also the types of contexts can be extended by other prior knowledge. For example, distance between fragments is a reasonable source to derive new contexts.

V. CONCLUSION

This paper has proposed a 2D Conditional Random Fields model with multiple contexts for logical labeling of born-digital fixed-layout documents. The baseline is developed using an SVM classifier without contextual information. The effects of contexts are examined by incorporating pairwise potentials derived from various prior knowledge. The experimental results show that combinations of contexts can benefit the performance. It is expected that describing contexts with richer features can bring further improvements.

Acknowledgment

This work is supported by National Natural Science Foundation of China, (No. 610300061).

References

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

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

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

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

[5] ——, “Conversion of pdf books in epub format,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 478–482.

[6] J. Fang, Z. Tang, and L. Gao, “Reflowing-driven paragraph recognition for electronic books in pdf,” in *IS&T/SPIE Electronic Imaging*, 2011, pp. 78 740U–78 740U.

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

[8] C. Xu, Z. Tang, X. Tao, Y. Li, and C. Shi, “Graph-based layout analysis for pdf documents,” in *IS&T/SPIE Electronic Imaging*, 2013, pp. 866 407–866 407.

[9] C. Xu, Z. Tang, X. Tao, and C. Shi, “Graphic composite segmentation for pdf documents with complex layouts,” in *IS&T/SPIE Electronic Imaging*, 2013, pp. 86 580E–86 580E.

[10] Y. Y. Rangoni and A. Belaïd, “Document logical structure analysis based on perceptive cycles,” in *Document Analysis Systems VII*, 2006, pp. 117–128.

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

[12] G. Paaß and I. Konya, “Machine learning for document structure recognition,” in *Modeling, Learning, and Processing of Text Technological Data Structures*, 2012, pp. 221–247.

[13] 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*, San Francisco, CA, USA, 2001, pp. 282–289.

[14] X. He, R. S. Zemel, and M. A. Carreira-Perpinán, “Multiscale conditional random fields for image labeling,” in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–695.

[15] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, “Document image segmentation using a 2d conditional random field model,” in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1. IEEE, 2007, pp. 407–411.

[16] S. Shetty, H. Srinivasan, M. Beal, and S. Srihari, “Segmentation and labeling of documents using conditional random fields,” in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 65 000U–65 000U.

[17] D. Liu and J. Nocedal, “On the limited memory bfgs method for large scale optimization,” *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.