인보이스 서류 영상의 테이블 헤더 문자 분류를 통한 구매 정보 추출 모델

신현경*

Purchase Information Extraction Model From Scanned Invoice Document Image By Classification Of Invoice Table Header Texts

Hyunkyung Shin*

요 약 스캔된 인보이스에 특화된 서류 관리 자동화 시스템 구축에있어서 추출된 금전적 데이터의 정확도에대한 엄격한 요구는 인보이스 테이블을 위한 발생적 모델 설계에서 자체 인증 절차를 포함하는 것을 필요로 한다. 가격 = 단가 x 구매수량과 같은 내부적 관계식을 활용한 단순한 인증 절차를 사용하는 것이 전형적 방법론이다. 본 논문에서는 영상내 테이블 헤더 부분의 탐색과 탐색된 헤더의 컬럼 구분자를 활용하는 개선된 자동 인증 절차를 갖춘 인보이스내 정보 추출 모델을 제안한다.

주제어 : 기계 학습, 문자인식, 문자 선 세그멘테이션, 문자 분류, 서류 영상 처리

Abstract Development of automated document management system specified for scanned invoice images suffers from rigorous accuracy requirements for extraction of monetary data, which necessiate automatic validation on the extracted values for a generative invoice table model. Use of certain internal constraints such as “amount = unit price times quantity” is typical implementation. In this paper, we propose a noble invoice information extraction model with improved auto-validation method by utilizing table header detection and column classification.

Key Words : machine learning, ocr, text line segmentation, text classification, document image processing.

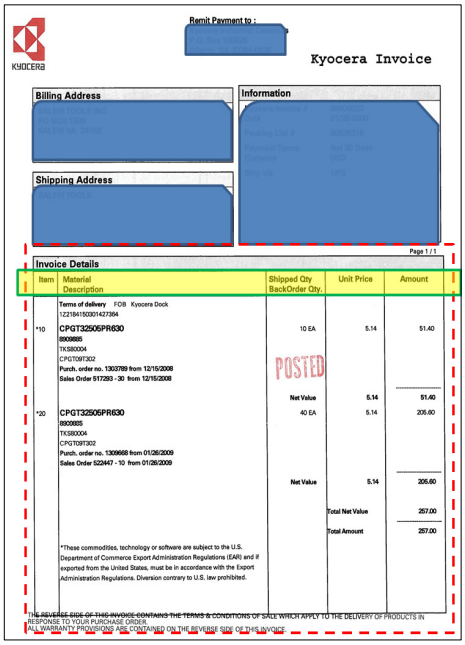
1. Introduction

An automated scanned document management system is consisted of the data storage server and the data processing client. The former is for storing the images updated on daily basis and the latter is for performing OCR and information extraction.

As the outputs of system, the extracted information is converted to XML format which provides user friendly validation/monitoring GUI procedure. In case the document management system is specified to invoices, the extracted informations are purchase orders and purchase details. The purchase order includes vendor name, billing address, and phone number which can be extracted using the regular expression and its variations. On the other hand, extraction of the purchase details requires more advanced method than the context free grammar.

The [Figure 1] illustrates a typical example of the invoice document images, where the cropped region by the red dashed line represents an invoice table. Inside the cropped region, the yellow colored annotated box indicates the header for the invoice table

※ 본 논문은 2011년 가천대학교 교내연구비 지원을 받아 시행된 연구임
*가천대학교 수학정보학과 조교수
논문접수: 2012년 11월 12일, 1차 수정을 거쳐, 심사완료: 2012년 12월 10일



[Figure 1] an example of invoice document containing the invoice table marked by the red dashed line.

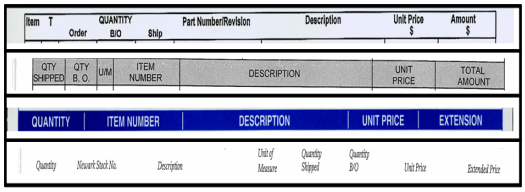
The purchase information is a list of the amounts corresponding to quantity, unit price, part number, and description.

The Table 1 below illustrates an example of XML converted output of the invoice purchased details extracted from Figure 1.

〈Table 1〉 an example of xml converted output of extracted information on the purchase details

```
<table>
  <table_row number = "1">
    <quantity><data 1.0/><position left b32 top
1232/></quantity>
  <unitcost><data 23.82/><position left 1808 top
1232/></unitcost>
  <descript><data 3200TIALN.51/2/><position left 922
top 1232/></descript>
  </table_row>
  <table_row number="2">
    <quantity><data 10/><position left 632 top
1332/></quantity>
    <unitcost><data 42.63/><position left 1808 top
1332/></unitcost>
    <descript><data 3200TIALN.62/> <position left
922 top 1332/></descript>
  </table_row>
  ...
</table>
```

Locations of the specific target values to be found (e.g., “quantity” and “unit price” in Figure 1) can be deduced from the column name placed in the table header. The problems of the approach are two folds: how to find the position of table header and how to classify the text in columns. In this paper we propose a machine learning based method for both finding header location and classifying table columns. Figure 2 demonstrates different types of invoice table header. Due to unstructured format of the header, the order of columns is not predictable, which prevents using pre-defined column ordering. The terms may vary: e.g., “quantity” and “QTY”.



[Figure 2] various formats of invoice table header

This paper is organized as follows: in section 2 previous researches are explained, in section 3 both localization of invoice table header and table column classification are described, the experimental results are summarized in section 4.

2. Previous Research

Text classification has been a basic technology for information extraction from general documents including the invoice type document, this can be seen in [1][4][8][10]. For the classification techniques, Belaïd proposed morphological tagging approach for invoice document analysis, which is bottom-up without a-priori template [2]. Nielson and Barrett presented a template based layout zoning method [13] and Kotsiantis considered induction classification algorithms [11], concluded that SVM and MLP were

superior to logic based (non-LDA) tree methods when dealing with multi-dimensional continuou (ordinal) features[3][9]. Cesarini, et. al. introduced case dependent domain knowledge and applied to invoice document as a case-study [5]. Shin developed a fast and robust text line segmentation as a pre-processing stage for invoice recognition [16].

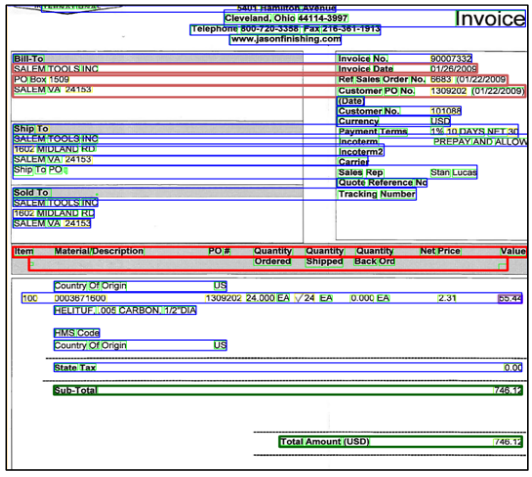
Among the researches on the cosntruction of document management system specializing on invoice recognition, Ming et al. proposed whole block moving method for slant image to deal with chinese invoice, the pre-processed invoice is processed using pre-compiled template library [12]. Hamza et al. developed a case based reasoning for document invoice analysis which is basically an auto templating method [7]. Sako et al. studied form data identification problem with the target ROI extraction using keyword matching, knowledge base character string recognition [15]. Chen and Blostein reported survey of document image classification emphasizing three components: problem statements, classifier architecture, and performance evaluation [6].

3. Table Header Localization and Header Column Classification

The raw texts obtained by OCR from invoice document images are unstructured and are not efficient for information extraction. without a semi-structured transformation. We take a text line as information unit entity which is obtained by text line segmentation.unstructured raw text data into the collection of semi-structured text lines vertically aligned. In this paper we used a DCT based projection profile method [18]. An an example of text line segmentation is illustrated in Figure 3

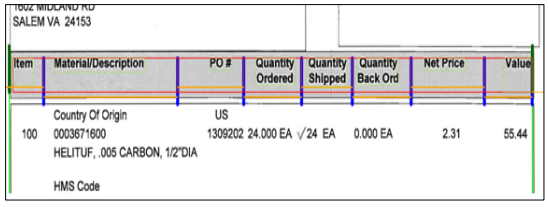
Each entity of the segmented text lines is then subjected to be classified into the several categories including table header line.

Figure 3 shows a typical output of text line segmentation in which the text lines have different colors for the purpose of representing individual target categories such as purchase order information, purchase line item, table header, and purchase summary line.



[Figure 3] illustration of text line segmentation

For this classification task, we adopted a supervised training based machine learning method by utilizing CART as explained comprehensively in [17]. Primary gain from localization of table header provides positional information of purchase items. Moreover, if text line can be splitted into columns, text words in the column can be used for column classification, which offers improved auto validation method.



[Figure 4] illustration of column separation in table header

In this paper, we develop text word segmentation for the texts contained in the table header line. Before performing segmentation, word grouping needs to be

applied as seen in Figure 4. As seen in the Figure, at the forth column from the left, “Ordered Quantity” should be a single entity of words rather than the two separate words.

For this task, we use vertical projection profile method similar to [17].

From the text line identified as a table header (e.g., the red colored line at the center area in Figure 3), each of the text words in the line is categorized as follows: we created the pre-defined categories as the followings: EXTENSION, QUANTITY, UNIT COST, DISCOUNT, UNIT OF MEASURE, and DESCRIPTION. We create the pre-defined keywords associated to the category. The keywords includes bi-grams and tri-grams. For the keyword matching, Levenshtein edit distance is used with tolerance of 5%. For each word group, for example “Item”, “Material/Description”, “PO#”, “Quantity Ordered”, “Quantity Shipped”, “Quantity backordered”, “Net Price”, and “Value”, we create an evaluation matrix. The element of the matrix is the count of keyword matched. For each of the text words maximum weighted category is assigned. The values corresponding to UNIT COST, QUANTITY, DISCOUNT, and EXTENSION are auto validated by the formula of EXTENSION = UNIT COST * QUANTITY * DISCOUNT.

4. Experiments

From a big data set, we randomly selected 11 data sets with total of 2,307 invoice images. For the purpose of validation, we manually built the ground truth data for the images[14]. For the measure of fit or error on the data extracted from an invoice image, the total number of table_row number, refer to Table 1, must equal to the ground truth and then the both values of quantity (quantity) and unit cost (unitcost) should also same. In some cases the contents in the description (descript) column also needs to be compared, but for the conciseness of this paper, we omit the error analysis on the description. The Table 2 shows the results of ground truth matching with the fit measures defined as above.

〈Table 2〉 results validation for extraction of invoice information

SET	DATA	MATCH	ERR LINE	ERR QTY	ERR COST
S-01	273	255 (93.4%)	8 (2.9%)	6 (2.2%)	9 (3.3%)
S-02	348	317 (91.1%)	10 (2.9%)	11 (3.2%)	21 (6.0%)
S-03	209	194 (92.8%)	9 (4.3%	4 (1.9%)	4 (1.9%)
S-04	164	151 (92.1%)	6 (3.7%)	6 (3.7%)	6 (3.7%)
S-05	231	204 (92.1%)	6 (2.6%)	16 (6.9%)	18 (7.8%)
S-06	171	157 (88.3%)	6 (3.5%)	4 (2.3%)	8 (4.7%)
S-07	192	169 (91.8%)	10 (5.2%)	10 (5.2%)	13 (6.8%)
S-08	157	139 (88.5%)	7 (4.5%)	10 (4.5%)	10 (10.2%)
S-09	177	146 (82.5%)	13 (7.3%)	14 (7.3%)	18 (4.0%)
S-10	50	47 (94.0%)	1 (2.0%)	1 (2.0%)	2 (4.0%)
S-11	335	303 (90.5%)	13 (3.9%)	15 (4.5%)	19 (5.7%)
TOT	2,307	2,082 (90.3%)	89 (3.9%)	97 (4.2%)	128 (5.5%)

In the table, the error is page based, in other word, in order for an invoice to be called MATCHED, all the numbers in the invoice corresponding UNIT COST, QUANTITY, DISCOUNT, and EXTENSION are correct.. ERR LINE, ERR QTY, and ERR COST indicate the percentage of invoices mistakenly identified the number of line items, quantity, and unit cost, respectively.

Reference

[1] Baird, H., & Lopresti, D., & Davison, B. & Pottenger, W., “Robust document image understanding technologies”, Proc. of ACM HDP

Workshop, USA, 2004, pp. 9-14.

[2] Belaïd, Y., & Belaïd, A., "Morphological Tagging Approach in Document Analysis of Invoices," Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)

[3] Breiman, L & Friedman, J. H., & Olshen, R. A., & Stone, C. J., "Classification and regression trees," Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[4] Büttcher, S., & Clarke, C. L. A., & Cormack, G. V., "Information Retrieval: Implementing and Evaluating Search Engines," MIT Press, 2010.

[5] Cesarini, F. & Francesconi, E., & Gori, M., & Soda, G., "Analysis and Understanding of Multi-Class Invoices," IJDAR, 2003.

[6] Chen, N., & Blostein D., "A survey of document image classification: problem statement, classifier architecture and performance evaluation," IJDAR, vol. 10, pp. 1-16, 2007.

[7] Hamza, H., & Belaid, Y., & Belaid, A., "Case-Based Reasoning for Invoice Analysis and Recognition," LECTURE NOTES IN COMPUTER SCIENCE, num. 4626, pp. 404-418, 2007.

[8] Hand, D, & Mannila, H. & Smyth, P. "Principles of Data Mining," Cambridge: MIT Press, 2001.

[9] Haykin, S., "Neural Networks-A Comprehensive Foundation," second ed. Prentice-Hall Inc., 1998.

[10] Ishitani, Y., "Model-based information extraction method tolerant of OCR errors for document images," Int. J. Comput. Proc. Oriental Lang., 15(2):165 - 186, 2002.

[11] Kotsiantis, S. B., "Supervised Machine Learning: A Review of Classification Techniques," Informatica, Vol. 31 (2007), pp. 249-268.

[12] Ming, D., & Liu, J., & Tian, J., "Research on Chinese financial invoice recognition technology," Pattern Recognition Letters, Volume 24, Issues 1-3, January 2003, Pages 489-497

[13] Nielson, H. E. & Barrett, W. A., "Consensus-Based Table Form Recognition ", ICDAR, Edinburgh (Scotland), pp. 906-910 , 2003.

[14] Richard, P & Dennis, C., "Cross-Validation of Regression Models". Journal of the American Statistical Association 79 (387): 575 - 583, 1984.

[15] Sako, H., & Seki, M., & Furukawa, N., & Ikeda, H., & Imaizumi, A., "Form Reading based on Form-type Identification and Formdata Recognition", In International Conference on Document Analysis and Recognition, Edinburgh (Scotland), pp. 926-930, 2003.

[16] Shin, H., "Fast Text Line Segmentation Using DCT For Color Image", KIPS, 2010.

[17] Shin, H., "Machine Learning Based Automatic Categorization Model for Text Lines in Invoice Documents", JKMS, 2011.

[18] Witten, I., & Moffat, A., & Bell, T. C., "Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition," Morgan Kaufnann Publishiers, New York, 1999.

신 현 경

- 2002년 8월: State University of New York at Stony Brook. 대학원 응용수학과(Ph.D.)
- 2007년 8월 현재: 가천대학교 수학정보학과 조교수
- 관심분야: Image Processing, Neural Network, Machine Learning.
- E-Mail: hyunkyung@gachon.ac.kr