Invoice Detection Processing and Analysis

Bhavesh Patil[1], Ruhita Patil[2], Sumant Patil[3], Prof Vishvayogita Savalkar[4]

[1,2,3] Dept. of Computer Engineering, M.G.M's College of Engineering and Technology, Kamothe, Navi Mumbai, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, M.G.M's College of Engineering and Technology, Kamothe, Navi Mumbai, Maharashtra, India

---

Abstract - *Invoice processing and financial information extraction have been popular topics among researchers for decades. Corporations like participation banks process invoices manually. Extraction of important information from invoices like product, price, amount etc. is a prerequisite for these banks. Image edge information is essentially one of the most significant information in an image, which can describe the target outline, its relative position within the target area, and other important information. Edge detection is one of the most important process in image processing, and the detection results directly affects the image analysis. In this, we propose a novel technique for processing invoice image tables automatically. Invoice images we process are mostly sent by customers as scanned or fax images which have low image quality. In order to process these invoices we run different methods several times with different parameters. Results from each method are fused to get candidate tables. The proposed methods are robust to the character set used in a document, the image resolution and the noise ratio of the document image, and can perform detection operations in a highly effective manner. In addition to success in low quality images, this method can be applied both on tables with and without borders. The quantitative results obtained by applying this method on real business invoices have very favourable results. By interpreting invoices with the OCR-engines, it results in the output text having few spelling errors. However, the invoice structure is lost, making it impossible to interpret the corresponding fields. If Naïve Bayes is chosen as the algorithm for machine learning, the prototype can correctly classify recurring invoice lines after a set of data has been processed. Machine learning with Naïve Bayes works on invoices if there is enough previously processed data. The findings in this thesis concludes that machine learning and OCR can be utilized to automatize manual labor.*

***Key Words*: Machine Learning, Naïve Bayes, OCR, OCRopus, Tesseract, Invoice handling**

1. INTRODUCTION

The main reason for the study was to ease the workload of the economic team and to save time and money for the company. If processes like invoice handling can be automated, humans can put their attention on other more important problems. While the solution presented in this thesis may need some manual interaction, the machine learning part of this thesis might help decrease the amount of supervision needed as more data is processed.

1.1 PROBLEM STATEMENT

Extracting important data from processed image and then the recognition of text into the specific format to generate output is an important feature of Invoice Detection. It can improve efficiency and minimalize the friction of handling systems.

It should detect the bill/invoice from the images by Edge detection, cropping, flattening, enhancement of cropped image and compression such that text is extracted neatly from processed image so that no background noise be present.

After converting text from image, data should be stored in prescribed output format. Model remains compatible to support various forms of image and processing time shouldn't exceed 5 seconds.

1.2 GOALS

The main goal was to create a prototype with the ability to evaluate data from invoices and determine whether the invoice was correct or not. To achieve the goal, the work was divided into the following steps.

The survey should:

- Evaluate existing OCR-engines on how they perform in terms of number of correctly recognized words on invoices,
- Evaluate how text matching can be applied to extract structured data from plain text and correct OCR-generated errors through a survey of suitable theories and related works, and

- Determine how the process of interpreting invoices would be automated using machine learning on the invoice specific data to conclude whether the invoice is correct or not.

The OCR-engine implementation should:

- Make few spelling errors,
- Produce readable plain text, and
- Retain the invoice structure.

The machine learning implementation should:

- Demonstrate whether the data in invoices is correct or should be supervised and corrected manually,
- Be modular to make it easier to adapt the solution to future fields, and
- Determine if an invoice is correct or not with a certain percentage.

The analysis should:

- Evaluate how existing OCR-engines performs on different invoices,
- Conclude if the machine learning prototype in fact becomes better as more data has previously been processed,
- Calculate with how much certainty the machine learning prototype will deliver the answer, and
- Evaluate if the technique is a working alternative to human supervision.

1.3 FLOW DIAGRAM

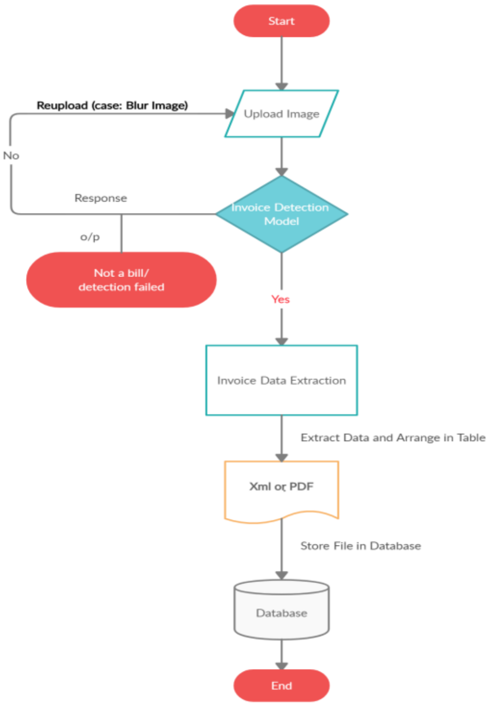


Fig-1: Flow Chart

2. ALGORITHMS AND TECHNIQUES

2.1 Current system

The current system uses is reached via an internal website within the company network, where users can upload invoices, which the system then interprets. The invoice is sent to a server where it is processed. The system can interpret the invoice if a structure extraction template has been made for the specific invoice format. If the system can interpret the invoice with a template, a result is presented to the user where the invoice head and invoice lines are presented in tables. The user has options to save the interpreted invoice to a database. What the current system does not include is the functions of telling the user if the content in the invoice is correct or not, and it does not remind the user if an invoice payment period is expiring and needs renewal.

2.2 OCR

Optical Character Recognition (OCR) is a technique used to interpret scanned documents into computer readable text. This thesis focusses on using OCR to interpret invoices and their content. The OCR-process had to result in the invoice structure to be relatively alike how it was structured before the process. It was of uttermost importance that the structure of the invoice did not differ after the OCR process.

Conclusion on OCR-scanning invoices:

To further evaluate how accurate these engines behaved when handling such documents, an evaluation was needed. Due to the fact that no previous work was found using these engines to scan invoices, an evaluation will be conducted in later chapters. This was done to gain knowledge about how well the engines performed scanning data from invoices.

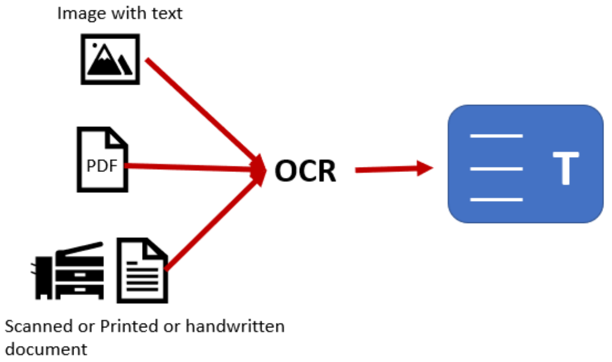


Fig-2: OCR

2.3 Text matching

Text matching or string pattern matching is described by Aoe [3], and is well used in computer science due to several needs, such as interpreting real handwritten documents and digital invoices, matching a string to data patterns in a data source. In other words, text matching is used in almost every case of text processing. There are several different algorithms developed that can be used in the purpose of matching text, two of them are presented below.

2.3.1 Longest common substring

Longest Common Substring is a text-matching algorithm, which in two or more strings finds the longest common string.

2.3.2 Levenshtein distance

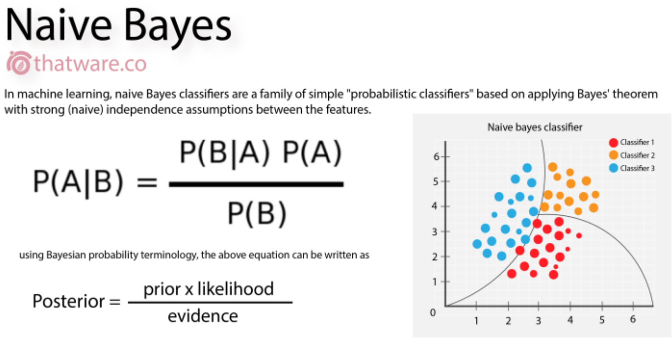
Another text matching algorithm is Levenshtein Distance. Levenshtein Distance is used to compute the distance between two strings similarity as stated by Haldar et al.

2.4 Machine learning

The theory behind machine learning is applying a previous solution, to solve a future problem as stated by Kulkarni [8]. The solution needs to be calculated based on previous solutions and other relevant information to make a qualified guess that may or may not be correct. There are several different algorithms and techniques used in machine learning. Not all of them are applicable on invoice data validation, and as such only those relevant to invoice data validation were evaluated.

2.5 Naïve Bayes

The Naïve Bayes algorithm has been widely used and derives from Bayesian decision theory. The algorithm is based upon features which the algorithm uses to classify the specified text. Rennie et al. [11] states that the algorithm does not take into consideration the relations between the features, which may be seen as a simple model. Due to features not having any relation between each other, it may result in patterns with more complex relations will not be found by the algorithm. According to Rish [12], Naïve Bayes is best utilized when the problem can be constructed as a one-to-one function. Naïve Bayes algorithm determines the probability between different cases that something will happen, based on previously encountered facts. It determines the probability of each feature independently and they are multiplied with each other for all possible outcomes, and the outcome with the highest probability is chosen. Therefore, if the probability of one or more of the independent features for a specific outcome are zero, then the probability of the outcome is zero.



Naive Bayes

thatware.co

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

Fig-3: Naïve Bayes

3. CONCLUSIONS

The OCR-engine test concluded that Tesseract was the better engine to use due to tedious training of OCRopus, and the spelling errors made by Tesser act still made the plain text readable. The time spent on training OCR opus did not present an adequate decrease of errors. Still, the problem with structure remained which made it impossible for a human to interpret the invoice correct. This concluded that the goal for OCR-engines was not reached as without a structure template, the scanning would not retain the invoice structure. It was however an important study nonetheless, because it proved that it was not possible to only use an OCR-engine without extraction templates for scanning invoice to usable plain text. The prototype using machine learning with Naïve Bayes was able to automate the process of invoice handling. It improved over time but if there was any change in an invoice line already saved in the data set, it would present a problem to change this data, resulting in having to remove it and start over again. The use of machine learning instead of conventional programming methods could prove advantageous if a lot of similar data was processed. The conclusion on machine learning with the prototype was that it could determine if data in an invoice was correct or not. It presented the user with a warning when the classification percentage was not high enough. It was also modular and easy to change the classification algorithm if there was a desire to test another algorithm. The use of machine learning to solve similar problems could also automate processes within companies which could make them more efficient. This showed that machine learning could be a valid choice of technology.

REFERENCES

[1] Springmann U, Najock D, Morgenroth H, Schmid H, Gotscharek A, Fink F. "OCR of historical printings of Latin texts: problems, prospects, progress". Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14). 2014.

[2] Hamza H, Belaïd Y, Belaïd A. "A case-based reasoning approach for invoice structure extraction". Document

Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. 2007 Oct: p. 327-331.

[3] Rish I. "An empirical study of the naive Bayes classifier". IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001: p. 41-46.

[4] McCallum A, Nigam K. "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI-98 workshop on learning for text categorization. 1998 Jul: p. 41-48.

[5] Kulkarni P. "Knowledge Augmentation: A Machine Learning Perspective". In Reinforcement and Systemic Machine Learning for Decision Making.: WileyIEEE Press; 2012. p. 209 - 236.

[6] Rennie JDM, Shih L, Teevan J, Karger DR. "Tackling the poor assumptions of naive bayes text classifiers". ICML. 2003 Aug: p. 616-623.

[7] Panigrahi KP. "A Comparative Study of Supervised Machine Learning Techniques for Spam E-mail Filtering". Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. 2012 Nov

[8] Sorio E, Bartoli A, Davanzo G, Medvet E. "A Domain Knowledge-based Approach for Automatic Correction of Printed Invoices". Information Society (iSociety), 2012 International Conference on. 2012 Jun: p. 151-155.