A Part Based Modeling Approach for Invoice Parsing

Enes Aslan[2], Tugrul Karakaya[1,2], Ethem Unver[2] and Yusuf Sinan Akgul[1]

[1]Dept. of Computer Eng., Gebze Technical University, GIT Vision Lab, Kocaeli, Turkey

[2]R&D Dept., Kuveyt Turk Participation Bank, Kocaeli, Turkey

{enes.aslan, tugrul.karakaya, ethem.unver}@kuveytturk.com.tr, akgul@gtu.edu.tr

Keywords: Invoice Processing, Part Based Modeling, Page Segmentation, Document Analysis, Information Extraction.

Abstract: Automated invoice processing and information extraction has attracted remarkable interest from business and academic circles. Invoice processing is a very critical and costly operation for participation banks because credit authorization process must be linked with the real trade activity via invoices. The classical invoice processing systems first assign the invoices to an invoice class but any error in document class decision will cause the invoice parsing to be invalid. This paper proposes a new invoice class-free-parsing method that uses a two-phase structure. The first phase uses individual invoice part detectors and the second phase employs an efficient part-based modeling approach. At the first phase, we employ different methods such as SVM, maximum entropy and HOG to produce candidates for the various types of invoice parts. At the second phase, the basic idea is to parse an invoice by parts arranged in a deformable composition similar to face or human body detection from digital images. The main advantage of the part-based modeling (PBM) approach is that this system can handle any type of invoice, a crucial functionality for business processes at participation banks. The proposed system is tested with real invoices and experimental results confirm the effectiveness of the proposed approach.

1 INTRODUCTION

Extraction of financial data from digital images of documents, which is very popular among academic (Cesarini, Francesconi, Gori, Marinai, Sheng and Soda, 1997) and business (Comay, 2014) worlds, is a very critical and costly operation for participation banks. Participation bank credit operations are mostly cost-plus profit financing transactions (Hardy, 2012), (Khan, 2010). Proof of purchase is required for these transactions and invoices are required as a purchase evidence (Kuwait Finance House, 2015). Although electronic invoices are getting more popular, many smaller businesses still issue paper based invoices. In addition, it is still a common practice to keep paper copies of these electronic documents. Kuveyt Turk participation bank manually processes around 1000 invoices per day each of which takes 6 minutes to complete. Therefore, automatic invoice processing would offer a number of advantages such as less labor, faster response time, and higher reliability. According to a survey it costs 9 Euros to process per invoice (Klein, Agne and Dengel, 2004). Similarly, it is predicted that a reliably automated invoice processing application will save Kuveyt Turk 3250 man-day for each year.

Invoices are one of the most unstructured financial document types due to their variations on the issuing company, product type, transaction type, etc. If the main structure of the invoice is already known, then parsing of these documents becomes easier. As a result, most of the studies focus on classifying invoices or extracting information depending on previously known invoices types. (Sorio, Bartoli, Davanzo and Medvet, 2010) uses an SVM based classifier to find new classes which are not known before. A new invoice is either assigned to an existing class or a new class is created. Image level features are utilized to match the given invoice with the previously known invoice types while ignoring smaller differences, such as stamps and signatures.

Another group of invoice processing systems are rule based. smartFIX (Klein, Dengel and Fordan, 2004) is such a system that classifies documents using extraction rules which are either specific to issuer or generic for all types of documents (Forcher,

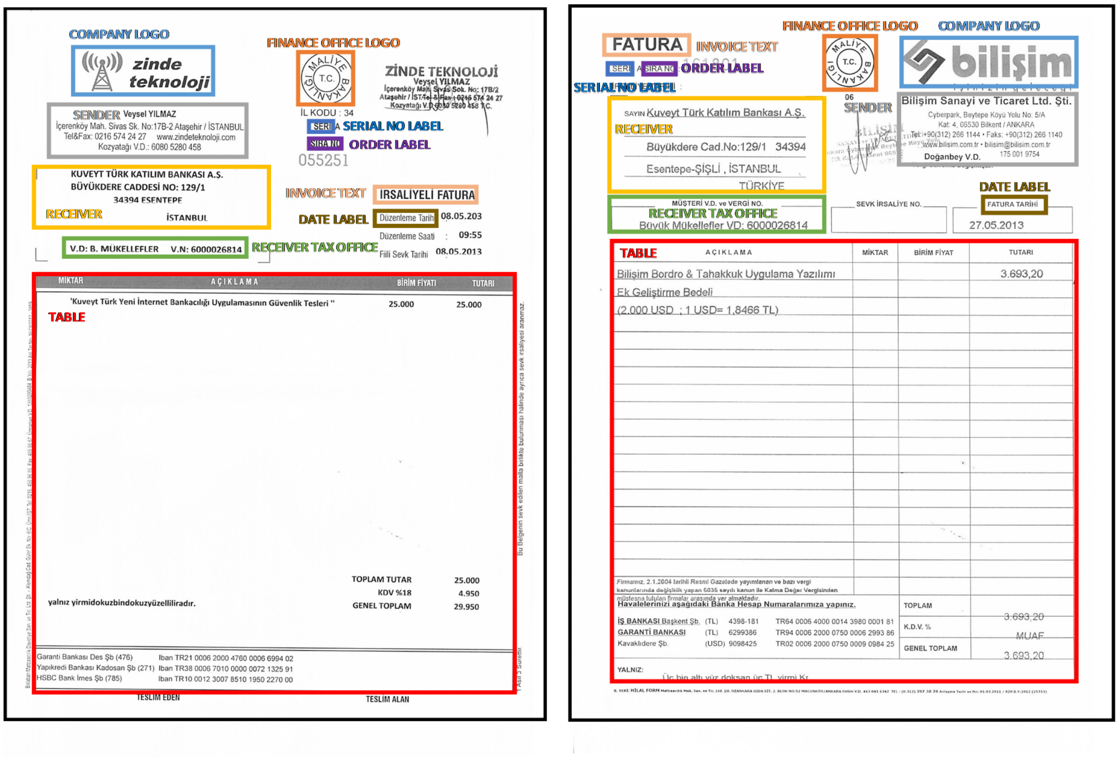

COMPANY LOGO
FINANCE OFFICE LOGO
SENDER
SERIAL NO LABEL
ORDER LABEL
RECEIVER
INVOICE TEXT
DATE LABEL
RECEIVER TAX OFFICE
TABLE
FINANCE OFFICE LOGO
COMPANY LOGO
FATURA
INVOICE TEXT
ORDER LABEL
SERIAL NO LABEL
SENDER
RECEIVER
DATE LABEL
RECEIVER TAX OFFICE
TABLE

Figure 1: Two sample invoices with marked parts (colored rectangles). Note the variations of the part positions.

Agne, Dengel, Gillmann and Roth-Berghofer, 2012). For known and new document type classification, smartFIX uses CBR (similar to (Hamza, Belaïd and Belaïd, 2007)) to extract information from unknown invoices using the most similar document class.

Generally, the invoice processing systems of the literature assume that there are classes of documents. Once the document class is decided then the parsing of the invoice becomes trivial. As a result, the decision on the document class label is the critical point of the whole process. Any errors in the document class decision will cause the invoice parsing to be invalid. For practical applications, the number of invoice classes are usually very high, which makes the class decision problem even more error prone. The tasks of keeping the classes and dynamically adding new ones to the system makes the whole process very complicated and difficult to maintain. Furthermore, it is a very demanding task to maintain rule based systems which use rules to make classification and parsing decisions.

This paper proposes a new invoice parsing method that eliminates invoice classes. We view the invoice documents as a single generic class. We model the invoice parsing task as a generic object detection process, such as face or human body detection from digital images. The variation of human body geometry, the skin color, clothing, body articulation, and the camera position makes the people detection problem very difficult. Parsing of invoices is not very different. Finding the invoice parts of senders, dates, articles, tax numbers (See Figure 1 for all the invoice parts) to process an invoice is similar to finding hands, faces, and arms in an image to detect body shape because for both cases these parts change in appearance and relative positions.

Computer vision community has been effectively employing Part Based Modeling (PBM) (Fischler and Elschlager, 1973), (Felzenszwalb and Huttenlocher, 2005) to address the above problem of object detection. PBM assumes that objects are composed of different parts arranged in a deformable configuration. Each part is individually detected and the candidates for each part are later combined under the deformable object model trained on an image set. We propose to employ the same idea with novel modifications for invoice parsing. The candidate positions of individual invoice parts are first detected and these candidates are combined under a deformable model optimization framework. This approach eliminates many problems about document classes listed above such as document class decisions, class layouts, high number of classes, and adding new

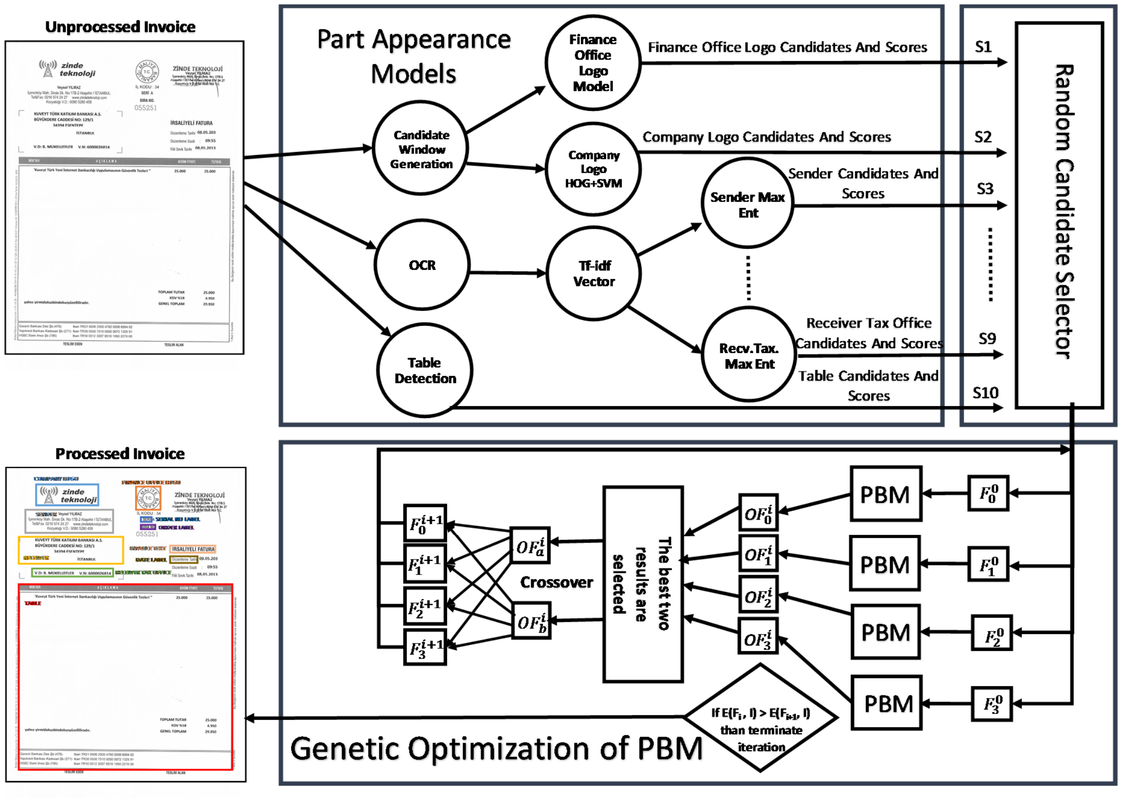

Unprocessed Invoice
Part Appearance Models
Finance Office Logo Model
Finance Office Logo Candidates And Scores
S1
Candidate Window Generation
Company Logo HOG+SVM
Company Logo Candidates And Scores
S2
Sender Max Ent
Sender Candidates And Scores
S3
OCR
Tf-idf Vector
Recv.Tax. Max Ent
Receiver Tax Office Candidates And Scores
S9
Table Detection
Table Candidates And Scores
S10
Random Candidate Selector
Processed Invoice
PBM
PBM
PBM
PBM
F_0^0
F_1^0
F_2^0
F_3^0
OF_0^i
OF_1^i
OF_2^i
OF_3^i
The best two results are selected
OF_a^i
OF_b^i
Crossover
F_0^{i+1}
F_1^{i+1}
F_2^{i+1}
F_3^{i+1}
If $E(F_i, I) > E(F_{i-1}, I)$ than terminate iteration
Genetic Optimization of PBM

Figure 2: The main architecture of the system.

classes dynamically. Furthermore, since PBM trains its model on an image data set, it does not need any complicated maintenance tasks.

We previously applied the idea of part based modeling to the invoice parsing problem (Aslan, Karakaya, Unver and Akgul, 2015). This work employs two new part based model optimization methods for the invoice parsing. Rest of this paper is organized as follows. Section 2 introduces the proposed two level invoice parsing method. Section 3 gives details about the validation work and finally we provide concluding remarks in Section 4.

2 THE INVOICE PARSING FRAMEWORK

The proposed system has a two phase architecture (See Figure 2). The first phase of the system models each invoice part as a separate part appearance detection problem. Each part detection module runs on a given invoice and returns the candidate part positions along with the scores for the positions. Note that this phase of the system does not consider any information about the absolute or relative positions of the parts. It only uses the image appearance information of the parts. The second phase of the system uses the candidate positions and scores of the first phase to decide final positions of each invoice part. This phase uses the PBM optimization framework to decide the final invoice part positions. PBM uses training information about how the part positions might change with respect to each other and with respect to the invoice itself.

2.1 Part Appearance Models

The individual appearance models of each invoice part might be different. We have four different appearance models. The first appearance model uses the TF-IDF vectors (Salton and McGill, 1983) of the OCR results from the invoice images. The OCR engines return groups of words along with their enclosing rectangles from the invoice images. We calculate the TF-IDF vectors for each group of words and run 7 different maximum entropy classifiers (Berger, Pietra and Pietra, 1996) for 7 invoice parts (sender, receiver, date, serial label, order label, invoice date, and receiver tax office). The scores from

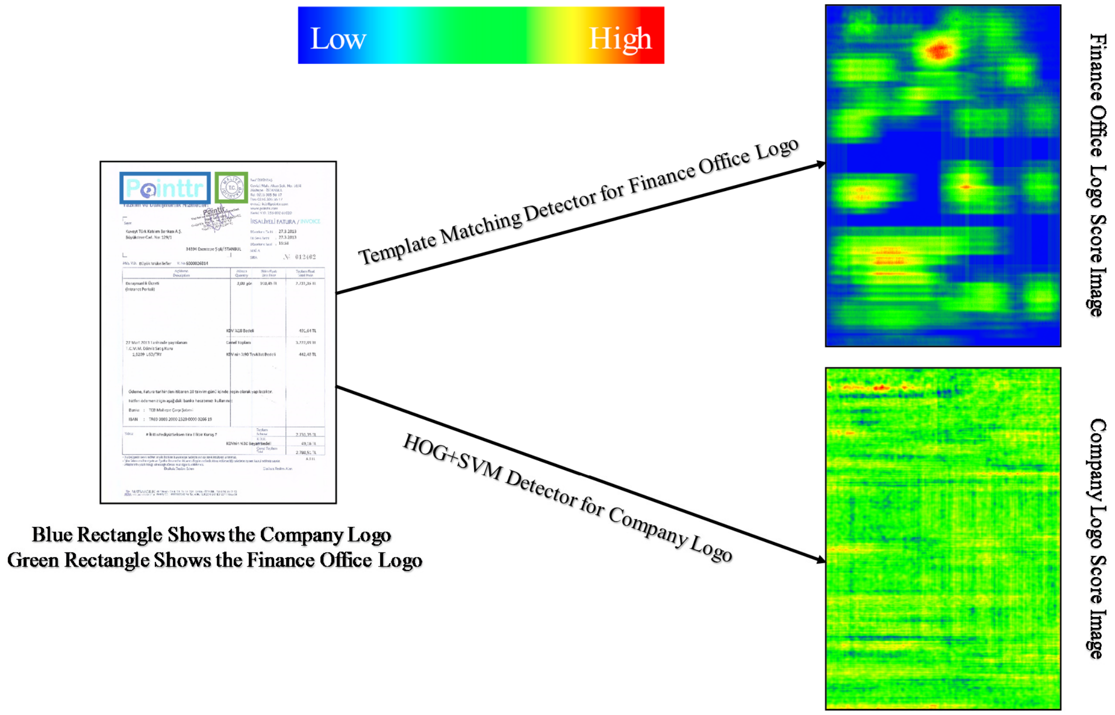

Low
High
Template Matching Detector for Finance Office Logo
HOG+SVM Detector for Company Logo
Finance Office Logo Score Image
Company Logo Score Image
Blue Rectangle Shows the Company Logo
Green Rectangle Shows the Finance Office Logo

Figure 3: Finance and company logo detectors score images.

these classifiers are fed to the PBM phase of the system.

The second appearance model is for the finance office logo, which legally has to appear exactly the same for all invoices. Although parsing the finance logo positions is not needed as the final result, their positions would greatly help in finding the position of other useful invoice parts at the PBM optimization phase. We use standard image template matching algorithm (Tanimoto, 1981) for the finance logo model. The third appearance model is for the company logo part of the invoice. Unlike the finance office logo, the company logo part can have very different appearances depending on the issuing company of the invoice. For this appearance model, we use a popular object detection algorithm (Dalal, Triggs, 2005) that employs Histograms of Oriented Gradient (HOG) vectors with SVM classifiers.

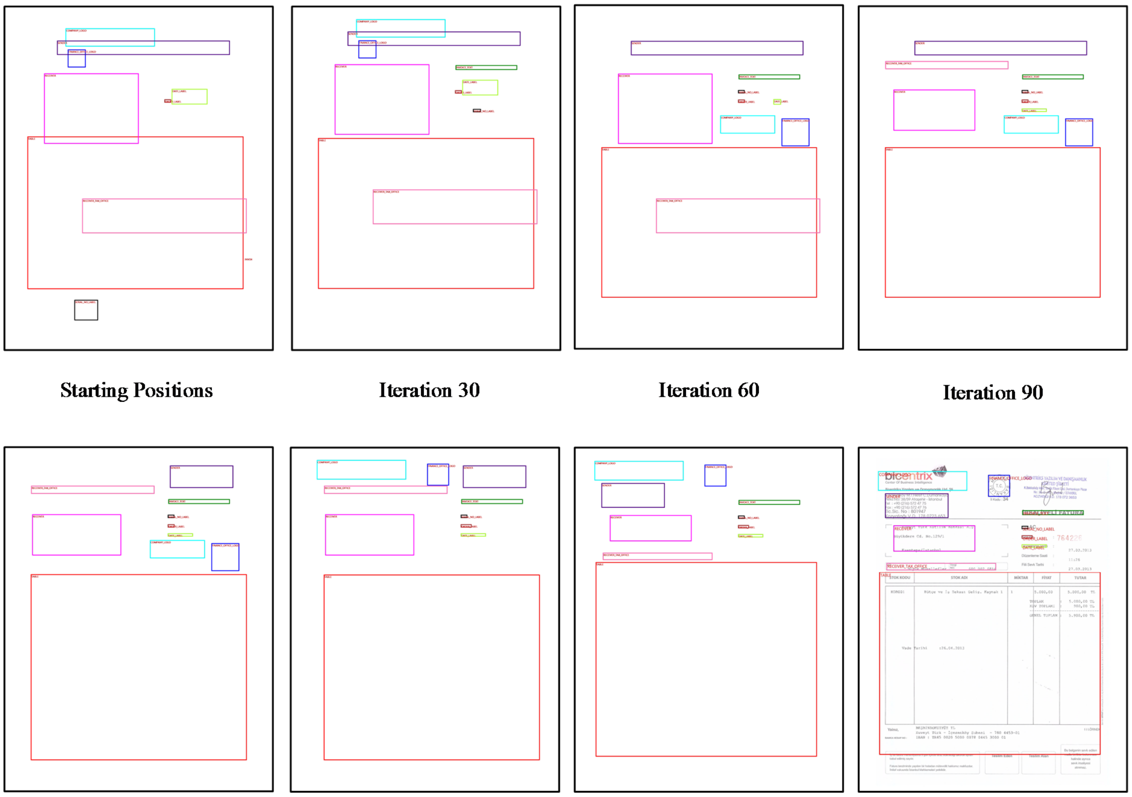
The last appearance model is for the content table, which is very large and has its own internal structure (column headers, rows, etc.). We developed a specialized feature fusion based table detector for the invoices (Unal, Unver, Karakaya and Akgul, 2015). This detector returns a number of candidate table areas along with their scores.

2.2 Part Based Model (PBM) Framework

The second phase of the system (PBM) is very generic. This module is used to choose the best configuration of the candidate invoice parts generated by the part appearance models. A PBM can be expressed by a graph *G=(V,E)* where the vertices $V = \{p_1,p_2,\ldots,p_n\}$ represent the invoice parts. Each part p_i represents a rectangular area on the invoice, $p_i=(x_i, y_i, w_i, h_i)$, where x_i, y_i represent the position of the part, and w_i, h_i represent the size of the part. There is an edge $(p_i, p_j) \in E$ between each part pairs of the invoice. A configuration of the parts on an invoice is shown by $F=\{p_1,p_2,\ldots,p_n\}$, where n is the number of parts on the invoice. PBM defines an energy function for a given configuration F, and finding a configuration that minimizes this function is called invoice parsing. The energy of a particular configuration of parts (invoice layout) is strongly related with parts' individual location and how well the relative location of the parts are positioned. More information about PBM can be found at (P. F. Felzenszwalb and D. P. Huttenlocher, 2005).

For a given configuration F on an invoice I, we define the PBM energy function as

$$E(F,I) = \sum_{i=1}^{n}\left(\alpha_i S_i(p_i) + \sum_{i\neq j}\beta_{ij}R_{ij}(p_i,p_j)\right), \quad (1)$$



Starting Positions

Iteration 30

Iteration 60

Iteration 90

Iteration 120

Iteration 150

Iteration 180

Final Result

Figure 4: Invoice layout at each optimization step.

where S_i is the normalized appearance model score function for part *i*, R_{ij} is the geometric relation function between parts p_i and p_j. The α and β values are weights. The parse of a given invoice can be estimated by

$$\underset{F}{\operatorname{argmin}}\, E(F, I), \tag{2}$$

The function R estimates the similarity of the relative positions of a given part pair to the learned geometric relations between these pairs. We use a Gaussian Mixture Model (GMM) to represent the geometric relation between part pairs.

$$R_{ij} = \prod_{l=1}^{4} \gamma_l^{ij}\, r_l^{ij}(p_i, p_j), \tag{3}$$

where each r_l^{ij} represents the individual relations between parts i and j. We use the normalized differences between the part widths (r_1^{ij}), heights (r_2^{ij}), x positions (r_3^{ij}) and y positions (r_4^{ij}) in the form GMM expressions as

$$r_l^{ij}(p_i, p_j) = \sum_{k=1}^{M} \emptyset_{ij}\, N(\mu_{ij}, \Sigma_{ij}), \tag{4}$$

where M is the number Gaussians in the GMM, $\emptyset$ are the weight parameters, μ and $\sum$ are the means and covariances, respectively.

Our employment of the PBM is different from the typical PBM framework because typically PBM uses the same type of detector for each part while we employ most suitable detector type in order to fit the needs of the application.

2.3 PBM Optimization

We used three different optimization methods: local, sequential, and genetic. The local optimization, finds the best position for each part independently. Since it does not consider the part position dependencies, it runs faster than the other optimization methods. However, the accuracy performance of this method would be lower than the other methods. Figure 3, shows the heat maps generated by the two invoice part finders on a sample invoice. After producing such heat maps, the final positions of the parts are determined by using the heat map of the corresponding detector.

For the sequential optimization, each element of F is first assigned a random candidate position from the appearance model scores. The Equation (2) is optimized by changing the configuration F parts one by one starting from the parts that have better appearance models such as table and finance logo. This process continues until there is no improvement in the results. Note that during the optimization we penalize parts that overlap. This is a hard constraint that should be satisfied by all parsed invoices.

At the beginning of genetic optimization (Figure 2) four starting part combinations (F_0^0,F_1^0,F_2^0,F_3^0) are selected from candidates by "Random Candidate Selector". Then we apply PBM to each of those part combinations to find four new PBM optimized invoice part combinations (OF_0^i, OF_1^i,OF_2^i,OF_3^i). Two of the combinations that have higher energies are eliminated. By crossing two remaining combinations four new combinations are created (F_0^{i+1},F_1^{i+1},F_2^{i+1},F_3^{i+1}). We use these combinations as starting part combinations to re-apply whole process from the beginning. This process continues while the best score of current combinations is lower than previous combination. Otherwise, the lowest PBM scored combination is selected as final combination.

Figure 4 shows a sample run of the proposed genetic algorithm on an invoice. The initial positions of the invoice parts start from random positions and at each iteration, they move to a new positon. At the end of the iterations, each part finds its final position.

2.4 System Training

Our system needs training for both Part Appearance Model phase and Part Based Model phase. The annotated invoice images are taken as inputs to both phases. The appearance scores for each part are also input to the PBM phase so we first have to train the appearance models. The learned parameters for the appearance models include the SVM parameters for the logo model and the maximum entropy parameters. The positive and negative window samples are chosen depending on the machine learning algorithm. For the logo detection, the positive training windows are the company logo regions of the invoices and the negative training windows are the most likely

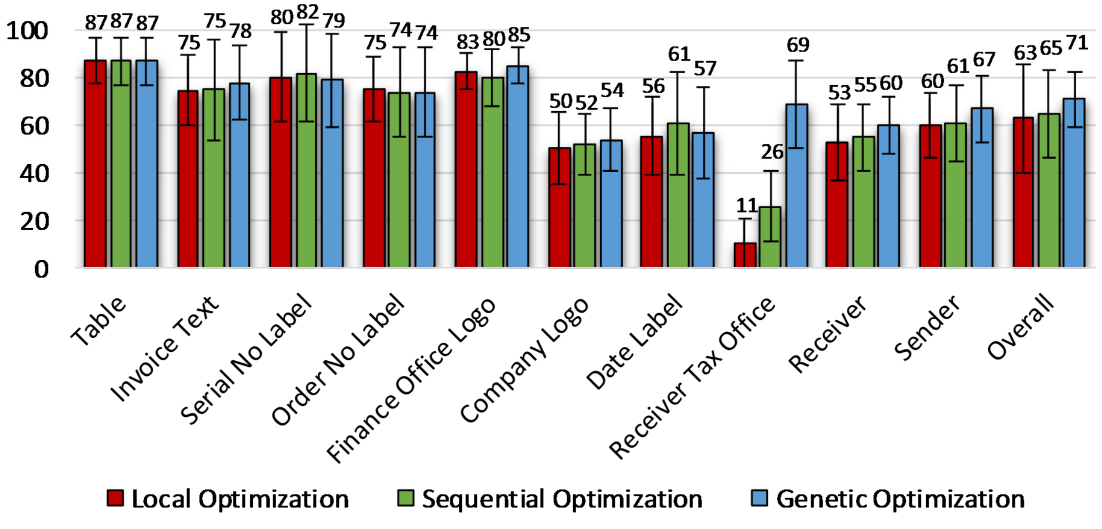


Figure 5: Means and standard deviations of intersections of sequential and genetic optimization results and real fields.

confusion zones for the logos such as the finance office logo, company signature area, etc. We determined most of the negative window locations by running the detectors on the invoice image set used for the training.

For the text areas (maximum entropy detectors), we need only positive window areas. OCR engines produce many erroneous recognition results so the TF-IDF vectors have high amounts of noise. In order to deal with this problem, we use edit distance metric to measure the distance between the words from OCR engine and regular dictionary words. In other words, we run a specialized spell checker for the invoice documents.

The learned parameters for the PBM include the GMM parameters of the R functions that define geometric relations between the invoice parts. The main parameters learned are the means and covariances of Equation (3).

3 EXPERIMENTS

For the system verification, we did not use standard English invoice sets because our application is specialized to Turkish banking and invoicing system. However, we built a representative set of invoices on which we show our results.

We performed a number of experiments to validate the proposed system. In all the experiments we used the same weigh parameters for Equation (1) ($\beta_{\text{ij}} = 1$, α_i=1). In addition, we kept the training and the testing invoice sets completely different to avoid any memorization problems of machine learning algorithms. The training set includes about 320 invoices and the test set includes 80 invoices. The invoices in the test set are all from different issuer companies so that can be considered as 80 distinct classes. For the OCR engine, we used a commercial product with the same parameter set.

For the performance metric we use Part Match Scores (PMS) that returns the amount of matching between a detected part and the annotated part.

$$PMS_{\text{g}} = \frac{2*\text{Area}(IA)}{Area(p_i) + Area(p_A)} \tag{5}$$

$$PMS_{\text{t}} = \frac{2*\text{NoW}(IA)}{NoW(p_i) + NoW(p_A)} \tag{6}$$

where p_i is the detected part, p_A is the annotated part. IA means Intersection Area between p_i and p_A. PMS$_\text{g}$ is used for graphic-based parts (company logo etc.) and PMS$_\text{t}$ is for text-based parts (receiver etc.). NoW calculates the Number of Words in the given area.

The genetic based algorithm performs mostly better than the sequential method. For some of the invoice parts, it improves the results considerably (e.g., tax office field), while the results for some fields are slightly worse (e.g., data label)

Although genetic optimization method is four times slower than the sequential optimization, it produces an overall performance around 6% better as shown in Figure 5. Note that the local optimization results are mostly worse than the other methods as expected. Note also that these results are comparable to state of the of art commercial invoice processing products that use known invoice classes, which makes our system very promising.

4 CONCLUSIONS

We presented a novel method for invoice parsing. The proposed method does not use any invoice classes and each invoice is considered as a new case. We employed ideas from Part Based Modeling approaches of general object detection to handle the high variations between the invoices. The proposed method can be extended with new part detectors conveniently without modifying the main optimization framework. The experiments performed on the real invoice data show the applicability of the method for the real life employment. For the future work, we plan to use a more sophisticated optimization methods and augment the text field detectors with image based features to handle OCR engine problems.

ACKNOWLEDGEMENTS

This work is supported by TUBITAK TEYDEB project number 3130882.

REFERENCES

L. Hardy, "The Evolution of Participation Banking in Turkey." Al Nakhlah Online Journal of Soutwest Asia and Islamic Civilization (2012).

F. Khan, "How ‘Islamic’is Islamic banking?" Journal of Economic Behavior & Organization 76.3 (2010), pp.805-820

"Payment of Supplier’s Due Amounts." Kuwait Finance House. Kuwait Finance House. Web. 26 Jan. 2015. <http://www.kfh.com/en/commercial/murabahaa/payment-of-suppliers-due-amounts.aspx>.

E. Sorio, A. Bartoli, G. Davanzo, & E. Medvet, (2010, September). Open world classification of printed invoices. In Proceedings of the 10th ACM symposium on Document engineering (pp. 187-190). ACM.

H. Hamza, Y. Belaïd, & A. Belaïd, (2007). Case-based reasoning for invoice analysis and recognition. In Case-Based Reasoning Research and Development (pp. 404-418). Springer Berlin Heidelberg.

B. Klein, A. R. Dengel, & A. Fordan, (2004). smartFIX: An adaptive system for document analysis and understanding. In Reading and Learning (pp. 166-186). Springer Berlin Heidelberg.

M. A. Fischler, & R. A. Elschlager, The representation and matching of pictorial structures. IEEE Transactions on Computers, 22(1) (1973), 67-92.

P. F. Felzenszwalb, D. P. Huttenlocher. "Pictorial structures for object recognition." International Journal of Computer Vision 61.1 (2005): 55-79.

B. Forcher, S. Agne, A. Dengel, M. Gillmann, & T. Roth-Berghofer, "Towards understandable explanations for document analysis systems." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

B. Klein, S. Agne, and A. Dengel. "Results of a study on invoice-reading systems in Germany." Document Analysis Systems VI. Springer Berlin Heidelberg, 2004. 451-462.

F. Cesarini, E. Francesconi, M. Gori, S. Marinai, J. Q. Sheng, G. Soda, "Rectangle labelling for an invoice understanding system." Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on. Vol. 1. IEEE, 1997.

G. Salton, & M. J. McGill, (1983). Introduction to modern information retrieval.

A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. "A maximum entropy approach to natural language processing." Computational linguistics 22.1 (1996): 39-71.

S. L. Tanimoto, "Template matching in pyramids." Computer Graphics and Image Processing 16, no. 4 (1981): 356-369.

N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

O. Comay, "Form data extraction without customization." U.S. Patent No. 8,660,294. 25 Feb. 2014.

U. S. Unal, E. Unver, T. Karakaya, Y. S. Akgul, “Invoice Content Table Detection and Analysis with Feature Fusion”, Signal Processing and Communications Applications 2015.

E. Aslan, T. Karakaya, E. Unver, Y. S. Akgul, “An Optimization Approach For Invoice Image Analysis”, Signal Processing and Communications Applications 2015.