Field Extraction from Administrative Documents by Incremental Structural Templates

Marçal Rusiñol*†, Tayeb Benkhelfallah* and Vincent Poulain d'Andecy*
*ITESOFT
Parc dAndron, Le Séquoia
30470 Aimargues, France.
†Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain.

***Abstract*—In this paper we present an incremental framework aimed at extracting field information from administrative document images in the context of a Digital Mail-room scenario. Given a single training sample in which the user has marked which fields have to be extracted from a particular document class, a document model representing structural relationships among words is built. This model is incrementally refined as the system processes more and more documents from the same class. A reformulation of the tf-idf statistic scheme allows to adjust the importance weights of the structural relationships among words. We report in the experimental section our results obtained with a large dataset of real invoices.**

I. Introduction

The automatic processing of administrative documents has been one of the applications that has received most attention from the dawn of document image processing. Nowadays, all these applications can be enclosed into the digital mailroom paradigm. Digital mailroom applications have to deal usually with incoming documents that need to be treated in a close to real time manner. The incoming documents usually have an heterogeneous nature and it is not rare to receive a mixed document flow that contains letters, faxes, invoices, forms, contracts, etc. More often than not we face a large scale problem as well, since the amount of correspondence to deal with might be massive. Manually processing this data is a really costly task and the industry and the market needs have led an important amount of research and development in the context of automatic processing such administrative documents.

Depending on the particular document type, different workflows are usually defined. The most common tasks that are embedded into digital mailroom services contain document classification [1], multipage document treatment [2], document understanding [3], [4], routing [5], and information extraction [6], [7], [8]. In this paper we focus in the information extraction problem which can be defined as the task of extracting relevant structured information from semi-structured document images. If we focus on the specific case of processing incoming invoices, the aim of the task is that given an incoming image from a known provider to be able to extract and store in a database all the relevant fields such as the date, the total amount, the invoice number, etc.

Many commercial products offering such services are based on the definition of fixed spatial templates that basically map the locations in which the OCR has to read the particular fields to extract. Such basic strategies might perfectly work in simple scenarios but start to cause problems as soon as we face large scale scenarios and if the document layout varies along time. Research papers usually propose structural templates that encode the local context of the information to extract such as in the works by Ishitani [7], [8] or Peanho et al. [6]. Such approaches learn a local layout structure from a sample document that is then registered to the test images in order to identify the fields to extract. However, in such approaches, the user is asked to mark several semantically meaningful layout entities in order to build a good model.

In this paper we propose an information extraction method that also relies on registering portions of layout elements. The proposed model is though extremely simple and requires really a minimum human intervention. In addition, our main contribution is that although the template model is learned using a single annotated image, the subsequent processed images incrementally adjust this model by a reformulation of the tf-idf statistic. The main advantage of such method is that the learned model can absorb variations of the invoice layout through time and learns in an unsupervised way which are the discriminative layout elements that help the most to extract a certain field. In the next Section we present an overview of the proposed system as well as the organization of the rest of the paper.

II. System Overview

As a previous stage of our system, all the incoming images are preprocessed by a set of cleaning steps and finally OCRed. Thus, the input of our system is the image to process together with the OCR output containing both word transcriptions and bounding-boxes and spatial positions of the read words. In this paper we assume that a previous classification step has been able to categorize the incoming image as one of the possible providers that the system expects. In the next Section we will however provide a short glimpse on how the documents are classified. We can see in Fig. 1 an overview of the proposed incremental field extraction system.

Whenever a new provider arrives, the user has to manually select which are the important fields to be retrieved from this type of documents. A structural model is generated for each of the fields. These structural models encode the local context of

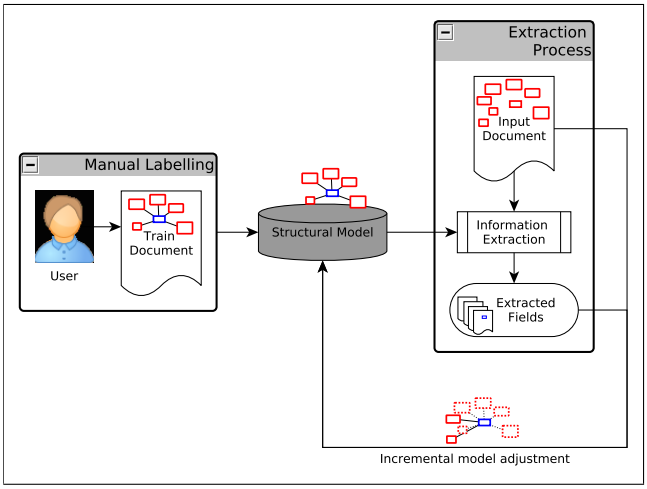

Fig. 1. System Overview

the target field to retrieve by storing the spatial relationships between pairs of words. We detail in Section IV this learning step. In the extraction process, an unseen image of a given provider arrives and all the structural models are applied. By matching word transcriptions and applying the structural relationships previously stored in the model, a voting scheme determines which candidate word is likely to be the field to extract. The extraction process is detailed in Section V.

Finally, if the confidence of the extracted field is high enough, an incremental step is applied in order to adjust the learned structural model. Some words that have been previously learned might start loosing importance if the processed documents cease to contain them, whereas some other words that are proven to be very discriminative along time must be reinforced. Section VI provides the details on this proposed iterative procedure. We finally provide the experimental results in Section VII.

III. Document Classification

The document classification step per se is beyond the scope of this paper, however, for the sake of generality we will provide a brief overview on how the invoice providers are identified. Sine there is a large amount of document classification techiques within the document analysis field [9], [1], any of such strategies could be applied in order to correctly identify the providers.

In our specific scenario, invoice images already contain information about which is the provider for such invoice. Whenever the OCR readings are correct, the classification step is done by looking at the OCR transcriptions for specific fields that would identify it, such as addresses, emails, phone numbers, tax numbers and so on. The provider's identity is cross-checked with all those fields in order to avoid false alarms by just relying on one single field read.

Whenever the OCR readings do not provide a correct identification or not evidenced enough, a visual descriptor codifying the invoice visual appearance such as [10] is applied. Afterwards any off-the-shelf statistical machine learning classifier can be used in order to provide a posteriori probabilities that the invoice under analysis corresponds to a specific provider.

IV. Learning Structural Models

For each different provider, the user is asked to manually label all the fields to extract from a single image. A different model for each of the fields is created. The structural model encodes pairwise word relationships between the target field to extract and the rest of the words that appear in the document. This structural template is modeled by a star graph S_n [11] that links all the words w_i in the document with the target field to extract t_j, as we can see in Fig. 2.

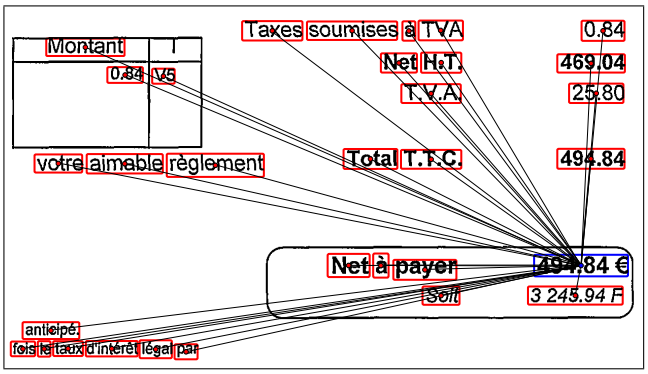

Fig. 2. Example of the structural template

In each node of S_n we store the word transcription $\mathrm{t}(w_i)$. Whereas in each edge we store the spatial relationship between the word w_i and the target field t_j in polar coordinates (θ_{ij}, r_{ij}). We can see an example of such relationship in Fig. 3.

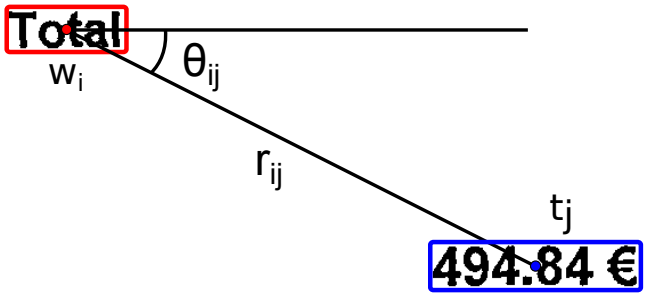

Fig. 3. Spatial relationship between a word and the target field.

V. Field Extraction

Once a single image for a given provider is annotated, we are already able to start extracting its relevant information. The input document is OCRed but we do not know which transcribed word might be the relevant one to extract. At this point the previously learned structural template is used in order to pinpoint which bounding-box is in the right position.

Given a word w_i of the incoming document having the centroid of its bounding-box located at coordinates (x, y), we retrieve all the nodes of the graph S_n having exactly the same transcription $\mathrm{t}(w_i)$. From this list of matched nodes, we obtain a set of polar coordinates that indicate a possible spatial relationship between the word w_i and the target field to extract.

The candidate target coordinates (x_t, y_t) are obtained for every tuple (θ_{ij}, r_{ij}) as follows:

$$\begin{aligned} x_t &= x + (r_{ij} \times \cos(\theta_{ij})) \\ y_t &= y + (r_{ij} \times \sin(\theta_{ij})) \end{aligned} \quad (1)$$

Finally a voting scheme is applied to locate the position of the most plausible field to extract. Each word of the incoming document might match several nodes of the structural template and thus have associated several spatial relationships. Each of these spatial relationships casts a vote in the candidate coordinates (x_t, y_t). The closest bounding-box to the maximum peak in the voting space is selected as the field to extract.

VI. Iterative Term Weighting

Up to now, the model of the invoice structure has been learned by a single annotated example from the user. However, when a document is processed, if the system yields a high confidence value that the extracted field is correct, we could use this non-annotated document in order to learn a better model for processing further images.

The first step of this incremental model adjustment is to add all the words in the new processed document that were not present in the learned one. A second step devoted to weight the different terms is then applied. This weighting wants to emphasize the following points:

- Words that appear many times in the same document are less informative since they introduce several spatial relationships for the same transcription
- Words that are always present given a provider are more discriminative than words which presence is irregular
- Words which are closer to the target are often more informative since they model the local context

A reformulation of the classic tf-idf [12] statistic scheme allows us to adjust the importance weights of different word transcriptions. We propose then to combine the *inverse term frequency* and the *document frequency* in order to highlight the transcriptions that are not frequent in a single page but that constantly appear in invoices from the same provider. Given a word transcription w and a document set d, the itf-df weight is computed as:

$$\text{itf-df}_{w,d} = \text{itf}_{w,d} \times \text{df}_w, \quad (2)$$

where $\text{itf}_{w,d}$ is the inverse term frequency and df_w is the document frequency of a given term, computed both as follows:

$$\text{itf}_{w,d} = \frac{1}{\text{f}(w,d)}, \quad (3)$$

where $\text{f}(w,d)$ is the raw frequency, i.e. the number of times that term w occurs in document d. If the word w appears just once in the document d, then $\text{itf}_{w,d} = 1$ and its value approaches 0 the more times the w appears in d.

The document frequency is defined as:

$$\text{df}_w = \log_{10}\left(1 + \frac{9 \times \text{g}(w)}{N}\right), \quad (4)$$

TABLE I. Reached accuracies for the different fields to extract with and without the use of the iterative framework.

	Field to extract		
	Amount	Date	Invoice number
Without iterative model learning	87.41%	90.09%	93.2%
With iterative model learning	92.09%	95.56%	93.94%

where $\text{g}(w)$ is the number of documents where the term w appears and N the total number of documents in the corpus. Here a logarithmic scale is used, and $\text{df}_w = 1$ if the word w appears in all the documents d and tends to 0 as soon as we see documents that do not contain any instance of w.

Both $\text{itf}_{w,d}$ and df_w are normalized so as $\text{itf-df}_{w,d}$ falls within the $[0, 1]$ range.

Finally, in order to weight the local context of the target word, we define a step function sid_w proportional to the inverse distance between the word w and the target word t, so words that are closer to the target receive more weight than words that are farther.

Summarizing, after the user labels the train image, word transcriptions can be weighted by using both $\text{itf}_{w,d}$ and sid_w, so the votes they cast in the voting space are proportional to their respective weights. After processing a test image, new words are added to the graph S_n and the weights of each word w are updated by $\text{itf-df}_{w,d} \times \text{sid}_w$. Let us see in the experimental results the effects of this iterative term weighting.

VII. Experimental Results

Our dataset consists of a total of 1020 invoice images coming from 20 different providers. We manually labeled three different fields to extract for each invoice in order to build the ground-truth data. Two of those fields, the date of the invoice and the invoice number, were mostly spatially stable within the same class. The amount field however has not a fixed position in some providers' invoices. For each provider a single image is taken to build the model for each field. The rest of images from this provider are used as the testing set. We can see some thumbnail versions of the used invoices in Fig. 4.

We detail in Table I the obtained accuracies for each of the fields to extract. We first evaluated the performance of the structural templates as our baseline. The obtained results ranged between a 87.41% and a 93.2% accuracy depending on the field to extract. As expected, the amount field is the one presenting more errors since in some cases it might move from an invoice to another. The date field is usually present in a cluttered zone of the invoice whereas the invoice number is usually in a less crowded zone and thus is easier to extract.

The use of the iterative framework presented in Section VI yields an increase of the accuracies that range between 0.74 and 5.47 percentage points with respect to the baseline system. Weighting both the notion of the local context of the target field and the discriminative power of the surrounding words largely improve the results in the most difficult cases.

However, building a model in an incremental fashion might as well present some drawbacks. We show in Fig. 5 the evolution of the amount of words considered in the model after several iterations. As expected, the more images the system

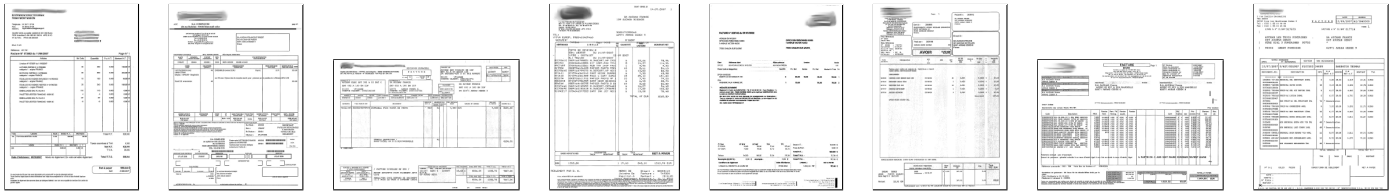

Fig. 4. Some examples of the invoices in our test scenario.

TABLE II. ACCURACIES FOR THE DIFFERENT PROVIDERS IN OUR DATASET.

Provider	Field to extract					
	Amount		Date		Invoice number	
	Regular	Iterative	Regular	Iterative	Regular	Iterative
01	100%	100%	100%	100%	100%	100%
02	100%	100%	94.33%	**100%**	98.14%	98.14%
03	100%	100%	**100%**	97.32%	**100%**	90.17%
04	**100%**	94.44%	96.29%	96.29%	100%	100%
05	100%	100%	**100%**	96.55%	100%	100%
06	58.18%	**100%**	63.63%	**100%**	46.78%	**75.22%**
07	100%	100%	100%	100%	100%	100%
08	100%	100%	100%	100%	100%	100%
09	**100%**	95.23%	75%	75%	100%	100%
10	100%	100%	100%	100%	**100%**	97.22%
11	94.44%	**100%**	24.13%	**100%**	100%	100%
12	100%	100%	100%	100%	100%	100%
13	**42.1%**	40.35%	**53.19%**	44.68%	100%	100%
14	100%	100%	**100%**	84%	100%	100%
15	100%	100%	95.23%	95.23%	**73.33%**	20%
16	**90.9%**	85.45%	100%	100%	100%	100%
17	**23.52%**	20.58%	100%	100%	**100%**	88.23%
18	100%	100%	100%	100%	100%	100%
19	71.42%	**100%**	100%	100%	100%	100%
20	100%	100%	100%	100%	80%	80%

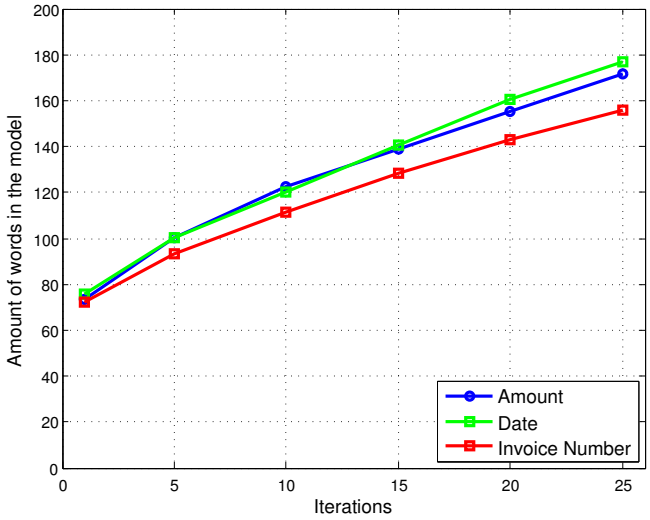

Fig. 5. Evolution of the amount of words in the model after different iterations.

processes, the more words the model stores. Although we did not noticed any decay of the system performance after several iterations, to indefinitely keep the model growing might end up corrupting it. A pruning step of the words with lower itf-df scores has to be envisaged.

It is however interesting to note that in Fig. 5, the different growth tendencies of the model depending on the position of the field to extract. The date field is usually located within a local context that tends to variate more, e.g. address blocks, whereas the invoice number is usually located in a less crowded and where the contents are more static. We can see that even if they start with models having very similar sizes, after several iterations the date model contains more than 20 more words than the invoice number.

Finally, in Table II we detail the reached accuracies for the twenty different providers in our dataset when using the baseline or the iterative procedure. First of all, although the overall behavior of the system is satisfactory, we can observe that specific layouts might cause severe problems to some particular fields, e.g. the date field for provider 11 or the amount field for the provider 17. We can also observe that provider-wise the use of the iterative process does not always benefits the system's performance. There are some cases in which incrementally adjusting the model causes errors that were not present when using the baseline. However, in overall, the gain obtained when using the incremental framework is bigger than the losses we might suffer.

VIII. CONCLUSION

In this paper we have presented an iterative framework for information extraction from administrative document images. Structural templates that model the relationships among word positions have been used to lead the field extraction process. An iterative term weighting scheme inspired by the tf-idf statistic allowed an incremental adjustment of the model that showed an interesting increase in the system performance.

The proposed field extraction method requires a minimum human intervention when building the model. The iterative

framework allows the discriminative portions of the model to be enforced as the system processes incoming images. This model adaptation is also suitable to absorb variations of the processed invoices through time.

ACKNOWLEDGMENT

This work has been supported by the European 7th framework project FP7-PEOPLE-2008-IAPP: 230653 ADAO. The work has been partially supported as well by the Spanish Ministry of Education and Science under projects RYC-2009-05031, TIN2011-24631, TIN2009-14633-C03-03, Consolider Ingenio 2010: MIPRCV (CSD200700018) and the grant 2009-SGR-1434 of the Generalitat de Catalunya.

REFERENCES

[1] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 1–16, June 2006.

[2] M. Rusiñol, D. Karatzas, A. Bagdanov, and J. Lladós, "Multipage document retrieval by textual and visual representations," in *Proceedings of the International Conference on Pattern Recognition*, 2012, pp. 521–524.

[3] H. Hamza, A. Belaïd, and B. Chaudhuri, "An end-to-end administrative document analysis system," in *Proceedings of the International Worksop on Document Analysis Systems*, 2008, pp. 175–182.

[4] S. Baumann, M. Ali, A. Dengel, T. Jager, M. Malburg, A. Weigel, and C. Wenzel, "Message extraction from printed documents – a complete solution," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1997, pp. 1055–1059.

[5] P. Viola, J. Rinker, and M. Law, "Automatic fax routing," in *Proceedings of the International Worksop on Document Analysis Systems: Document Analysis Systems VI*, ser. Lecture Notes on Computer Science, 2004, vol. 3163, pp. 484–495.

[6] C. Peanho, H. Stagni, and F. da Silva, "Semantic information extraction from images of complex documents," *Applied Intelligence*, vol. 37, no. 5, pp. 543–557, December 2012.

[7] Y. Ishitani, "Model-based information extraction method tolerant of OCR errors for document images," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2001, pp. 908–915.

[8] ——, "Model-based information extraction and its applications for document images," in *Proceedings of the Workshop on Document Layout Interpretation and its Applications*, 2001.

[9] D. Doermann, "The indexing and retrieval of document images: A survey," *Computer Vision Image Understanding*, vol. 70, no. 3, pp. 287–298, June 1998.

[10] P. Héroux, S. Diana, A. Ribert, and E. Trupin, "Classification method study for automatic form class identification," in *Proceedings of the Fourteenth International Conference on Pattern Recognition*, 1998, pp. 926–928.

[11] F. Harary, *Graph Theory*. Addison-Wesley, 1994.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.