Unsupervised Recognition of the Logical Structure of Business Documents Based on Spatial Relationships

Louisa Kessi[1,2(✉)], Frank Lebourgeois[1,2], and Christophe Garcia[1,2]

[1] Université de Lyon, CNRS, Lyon, France
{louisa.kessi,franck.lebourgeois,
christophe.garcia}@liris.cnrs.fr, l.kessi@orpalis.com
[2] INSA-Lyon, LIRIS, UMR5205, 69621 Lyon, France

Abstract. This paper presents the very first unsupervised and automatic system which can recognize the logical structure of business documents without any models or prior information about their logical structure. Our solution can process totally unknown new models of documents. We consider the problem of recognition of logical structures as a problem of detection, because we simultaneously have to localize and recognize the logical function of blocks of text. We assume that any document is composed of parts from several other models of documents. We have proposed a part-based spatial model suited for partial voting. Our proposed model presents the concept of Spatial Context (SC) as a spatial feature, which locally measure the distribution of spatial information around a point of reference. Our method is based on a Gaussian voting process providing a robust mechanism to detect elements of any logical structure. Our solution is suited for non-rigid structures and works well with a reduced number of images. This excellent property is not shared by the supervised approaches, especially methods based on neuronal networks.

Keywords: Unsupervised recognition · Spatial relation · Voting · Logical structure recognition · Business documents

1 Introduction and Context

The main objective of this work is the automatic reading of business documents. In the Document Image Analysis (DIA) domain, it is a direct application of the recognition of the logical structure of documents. In contrast to the layout analysis of documents named physical structure that study the documents' appearance, the recognition of the logical structure aims to localize and recognize the logical function of text blocks. It is one of the most challenging problems in DIA because of the causality dilemma between recognition and localization. There is only a limited number of research on the recognition of the logical structure of documents and only very few studies concerning business documents. All previous works in logical structure recognition use supervised approaches applied on homogeneous documents with structures known a priori and that have a very standardized

N. Tsapatsoulis et al. (Eds.): CAIP 2021, LNCS 13053, pp. 57–72, 2021.
https://doi.org/10.1007/978-3-030-89131-2_6

predetermined rigid model. The systems developed by means of this research can only process documents with a model. They are not applicable to business documents.

Today companies have numerous suppliers and customers that exchange a great number of commercial documents. Each year, millions of companies go bankrupt and disappear; simultaneously almost the same number of new companies are created. Each company creates its own model of documents with a specific layout and logical structure because there is no regulation to design a model of business documents. Manual processing of all ingoing and outgoing documents is too expensive. Large and medium-sized companies need automatic solutions to process their business documents on a daily basis. The automatic processing of administrative documents is an important business that generates large profits for private companies. Most existing commercial solutions are systems based on rules or templates for each model of business document. Software based on templates requires manual modelization of each type of document. For new or unknown types of documents, an operator must define the location and the label of each metadata to read. The conception of a document model is time consuming and expensive work. Some companies provide an annotation tool to costumers so that they design the new models by themselves. The new models are shared among the costumers, which is a kind of crowdsourcing solution. Systems based on rules try to generalize the modelization of several different models of documents and can process some of documents having different layouts and structures. But these rule-based systems cannot outperform the software based on templates.

Research in this field rallied much later than research in the field of automatic recognition of characters. Indeed, the characterization of the different structures of a document presents various difficulties. Some OCR software maintains the physical structure and preserves the typography and the organization of documents, but the function of blocks of text and thus the logical structure cannot be analyzed. Today, there is no commercially available structure recognition system that is completely automatic.

The business documents are so heterogeneous that it becomes impossible to modelize millions of templates, for each company or administration. Our proposed work aims to develop an automatic system which can read any business or administrative document without a model of these documents. It is one of the more challenging developments in DIA.

This paper is organized as follows: Sect. 2 presents the related state of the art in the domain of the logical structure recognition. Section 3 details our proposal and the new concepts we have introduced. Section 4 describes a novel spatial feature; we have called "Spatial Context". Section 5 introduces the Gaussian voting mechanism and the final decision stage. The last section gives the results on a database of real invoices.

2 State of the Art

Most of related works about logical structure recognition has been introduced between 1990 and 2010. Specific session about logical structure recognition has existed in the ICDAR until recently. The sessions "Segmentation and Layout Analysis" of the last ICDAR do not concern the logical structure understanding. A specific Workshop DLIA

(Document Layout Interpretation Analysis) introduced during ICDAR99 has disappeared. Only a very small number of previous works concerns business documents. We split the related works into four categories:

Data-Driven Approaches: Usually, they analyze the layout by using rules, grammars, or heuristics in order to retrieve the tabular structure of the document [1–5]. Few papers concern business documents. In [6] Klein used the headers of the tables as the first solution for locating tables. Furthermore, header extraction works only if similar headers exist in a header database related to the extraction system. [7] proposed to localize tables using a grammar (EPF) and an associated analyzer. Moreover, its major disadvantage lies in the fact that the user must himself formalize the grammar relating to the type of documents before starting the information extraction.

Model-Driven Approaches: These systems are based on a model of the document to extract information. The model can be built automatically or manually [8–16]. The work presented in [13, 14] uses keywords, and other areas of interest such as logos and horizontal and vertical lines, in the context of manual document modeling. All these elements are extracted manually. The proposed final model is a labeled and oriented graph. However, the ability to generalize this system is not really demonstrated. The tests established on 138 documents are insufficient to aspire to a generalization. In addition, if a completely new invoice case arises and the information in the knowledge bases does not cover this specific case, it becomes very difficult to find an interpretation of this invoice. This is because each keyword is analyzed independently of the others. Esser et al. [7] builds a database of absolute positions of fields for each template. The work proposed in [15, 16] aimed to develop a system of recognition of heterogeneous documents from observations already memorized. The modeling of the structure of a document is performed from the text obtained by OCR and the relative position of the textual fields between them. The model is generated semi-automatically from keyword and pattern structures described in a spatial relationship graph. The modeling and recognition system uses the Case Reasoning (RpC) mechanism. However, this system is not suitable for the process of totally unknown new models of documents not registered in the database. Very few papers report quantitative results about logical structure recognition [17–22] on various documents like patents, newspapers, books, magazines, scientific papers, table of contents.

Deep Learning-Driven Approaches: [41] concerns only web wrappers and not the logical structure of invoices for the digitized business document recognition. The analysis of Web page is easier because it is OCR errors free. They recognize only one metadata: the field "price". The authors doubt about their ability to recognize a second metadata. The works of [42–44] are however not applicable in our task as we do not have access to the representation of source markup for the documents we process.

Information extraction from business documents for problematics like named entity recognition and relation extraction take advantage from recent advances in deep learning [31, 35–37, 47], however, these techniques are not directly applicable to our task on logical structure recognition. [38–40, 46] didn't deal with the logical structure recognition (i.e. the logical function of text blocks) but the layout analysis (description of the layout

in terms of figure, table, section, caption, list paragraph) which is a different problem. Layout analysis can use the visual appearance (font style and size, color, alignments, texture...) to recognize the components of the layout. For the logical structure of invoices, we must use the spatial relationships between text blocks.

Commercial systems exists but they are limited to regular documents having a layout that rarely change. For unknown document, each company design its own documents and create a new layout (color, logo, fonts, style…) for their business documents.

Industrial Known Systems: The works of DocuWare [34] and the work by ITESOFT [29] require the creation of a database of templates in order to extract keywords and positions for each field. A template based system and a rule-based approach for unknown documents which are not recognized by the models is processed using heuristic and machine learning classifiers. The work of smartFIX [30] uses a manually programed rules for each template. ABBYY FlexiCapture [33] processes business documents and can extract data from forms. Some manual checking of data is done before the import into business databases. ReadSoft [32] Match zones from templates designed by users (for free!). For each new document, a user modelize manually a template which is shared automatically to the other users in the world. A manual verification mechanism reduces the recognition errors. CloudScan [31] is an invoice analysis system using recurrent neural networks. The authors takes a PDF file as the input and extract the words and their positions. Each line is analyzed as a vector of n-grams which limit the accuracy. A contextual features based on the closest four entities and an Long Short-Term Memory (LSTM) is used for classification.

3 Proposal

We describe the high-level stages of a more complex system which process automatically business document. The low level processing stages (separation between added text and preprinted text, color segmentation, layout extraction, character restoration) have already been published [48–51]. We introduce new concepts and make several assumptions during the development of the final stages of our recognition system:

- We assume that any document with an unknown model can be recognized by using parts of the logical structures of other known documents. We introduce a part by part recognition approach and a part-based model of the structure of documents which can recognize the logical structure of any document without a model of this document.
- We introduce the concept of micro-structure suited for the recognition part-by-part of any document. We define a micro-structure all pairs of text blocks that have a logical link. For business documents, most of the logical text blocks to detect, called "metadata", are mostly associated with a label called "caption" which are vertically or horizontally aligned. We define a micro-structure by the pairs of text blocks (*Caption* → *Metadata*) with → which describes the spatial relation between the caption and the metadata to retrieve. ("Due date" → 13/02/2018) ("Total Net" → "*135,00€*") ("VAT → *19,6%*") are some example of micro-structures. The class of possible caption is given by the matching of the OCRresults into a dictionary of all possible captions

found in business documents in Europe. Spatial horizontal or vertical alignment is also an important feature which links the metadata and its caption.

- We claim that an unsupervised system is better suited for this problem than a supervised approach. Millions of new models of business documents are created each year, and supervised approaches must be retrained each time a single new model is introduced or deleted. This explains our choice to focus our work only on an unsupervised approach which allows to add new models without a retraining.
- The recognition of the logical structure of a document is not possible without taking the spatial relation into account as the main feature. We introduce the concept of Spatial Context (SC) as a spatial feature, which describes the relative positions of metadata or caption in a Neighborhood around each word of interest.
- We introduce an original voting process in the spatial space that allows to localize and recognize a researched metadata and its logical function. The voting process is a statistical unsupervised approach which accumulates concordant information according to different parameters. Voting approaches has already been used in computer vision to find straight lines [23] shapes [24, 25], arbitrary lines [26], objects detection [27, 28]. This approach is well-known to be robust to noise and missing information due to the partial occultation of an object. For our application, a recognition system based on voting is an unsupervised approach that does not require any training and can manage part-based models and spatial mutual information into a single scheme.

4 Spatial Contexts

We define the Spatial Context (SC) of a neighborhood N centered into a point C, the pairs of text block E_i and a vector $\vec{U}_i$ which define the spatial relation between E_i and the center C of the neighborhood. Because the spatial structure of a document essentially varies horizontally and vertically, $\vec{U}_i$ is expressed in Cartesian coordinates (dx,dy) (1) (Fig. 1).

$$SC_{N(C)} = \left\{ \left(E_i, \vec{U}_i \right) \; E_i \in N(C) \; \vec{U}_i = \overrightarrow{CE_i}, \right\} \tag{1}$$

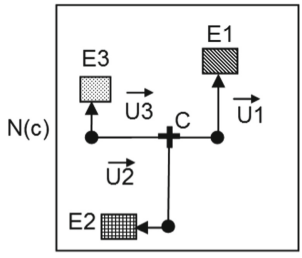
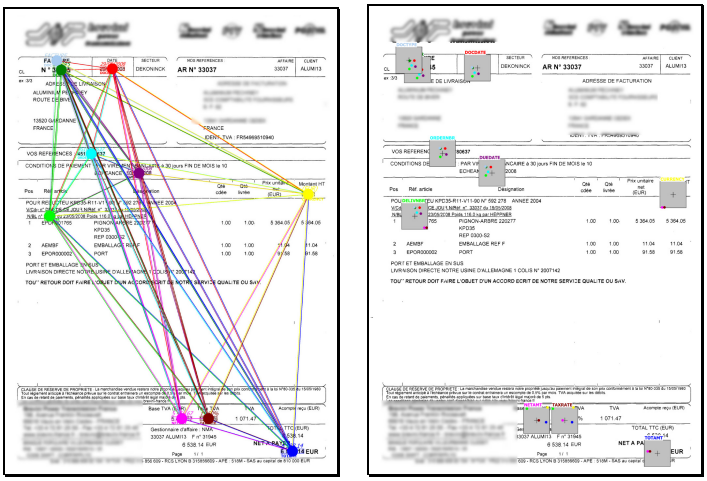


Fig. 1. A Spatial Context (SC).

We propose a novel spatial structure model and replace a classical spatial structure (Fig. 2a) by a set of Spatial Contexts which describe locally the neighboring elements



a) Classical structure model *b)Part-based Spatial Structure*

Fig. 2. Comparison between classical spatial model and our part-based spatial structure.

of the structure (Fig. 2b). This model allows to analyze a spatial structure part-by-part and reduce the structure complexity.

We introduce two types of spatial contexts:

- Metadata to Captions Spatial Context (MCSC)
- Metadata to Metadata Spatial Context (MMSC)

Because the documents may have different sizes and different resolutions, we normalize all positions of text block by dividing all coordinates by the size of the image. During the detection process, we multiply all coordinates by the size of the current image. This normalization guarantees that the spatial relations are always suited to the current document, whatever its size.

4.1 Metadata to Captions Spatial Context (MCSC)

The Metadata to Captions Spatial Context $MCSC_k$, centered on the metadata M_k from the logical class $n°k$, measures the spatial distribution of the possible captions described by the words W_i which are vertically or horizontally aligned with M_k and belonging to the lexical dictionary of captions of class $n°\ k$ (2). With a window size of 50% of the size of the image to define the neighborhood, the MCSC is a local spatial feature suited for body part processing and part-by-part recognition.

It is the main information that must be used first during the recognition step. But the MCSC may be empty if there is no word W_i found in the neighborhood of the metadata n°k, with the lexical class k. In this case, the metadata n°k cannot be recognized with

only this spatial context.

$$MCSC_k = \left\{ W_i. = \begin{pmatrix} W_i.x \\ W_i.y \end{pmatrix}, \vec{U}_i = \overrightarrow{M_k W_i} = \begin{pmatrix} W_i.x - M_k.x \\ W_i.y - M_k.y \end{pmatrix} / f(W_i) = k,\ W_i \to M_k \right\} \tag{2}$$

$$W_i \to M_k \Leftrightarrow \{W_i \textit{ is aligned to the metadata } M_k\}$$

$$f(W_i) = \{\textit{Lexical Class of the word } n^\circ i\ W_i\}$$

4.2 Metadata to Metadata Spatial Context (MMSC)

The Metadata to Metadata Spatial Context ($MMSC_k$) measures the spatial distribution with the other i^{th} metadata M_i which appear in the neighborhood $N(M_k)$ centered on M_k (3). The MMSC must be used after the prior localization of metadata by using the MCSC. This spatial context assumes that the metadata make statistically recurrent micro-structures. Voting by using the MMSC will reinforce the correct prior detection of the metadata and reduce false detections.

$$MMSC_k = \left\{ M_i = \begin{pmatrix} M_i.x \\ M_i.y \end{pmatrix}, \vec{U}_i = \overrightarrow{M_k M_i} = \begin{pmatrix} M_i.x - M_k.x \\ M_i.y - M_k.y \end{pmatrix} / M_i \in N(M_k),\ i \neq k \right\} \tag{3}$$

5 Voting and Detection Stages

For the recognition of the logical structure of documents, we use a 2-dimensional voting space defined by the parameters (*xc*,*yc*) coordinate which localize the metadata to detect. The structure of business documents is based on very few text blocks. The voting process for the recognition of the logical structure of documents is achieved part-by-part and not model by model. We have a 2D-pool for each metadata to retrieve.

After several experiments, we chose a pool with a variable size that depends on the size of the document. The reduction factor α is also important for the precision of the localization of the text blocks to detect. After several experiments on our image database, the optimal choice of the reduction factor α equals 16. This value makes sense because it approximately corresponds to the average height of characters and text lines with 400 dpi of resolution. We quantify the parameters (*xc*,*yc*) into $W \times H$ bins with W = ImageWidth/α and H = ImageHeight/α. We introduce the Gaussian voting which consists to vote spatial Gaussian functions instead of using a classical variation of parameters generally apply during voting process (Fig. 3). A classical vote of dirac functions, followed by a smoothing of the pool by a Gaussian function in order to detect local maxima in the pools, doesn't work in our case.

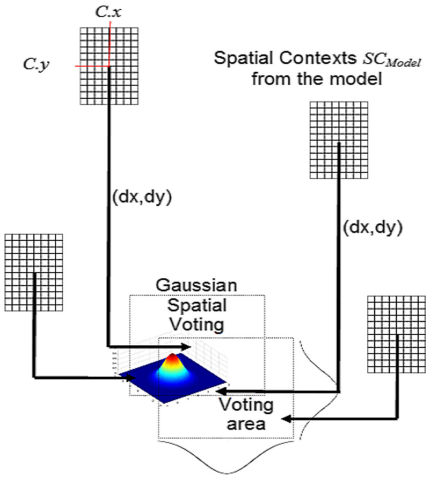


Fig. 3. Illustration of the Gaussian voting process

We choose to use four different voting processes into four different pools:

- Voting by using Metadata position
- Voting by using Metadata to Captions Spatial Context
- Voting by using the metadata format
- Voting by using Metadata to Metadata Spatial Context

These four voting processes are complementary. Metadata aligned with a caption can be detected by using their relative positions' possibilities, the spatial relations with their captions, the spatial relations with the other metadata, and the metadata format. The metadata that are not described by a caption are detected by using the metadata's possible position, the metadata format and the possible relationships between the other metadata.

5.1 Voting by Using Metadata Position

We use the relative position for a preliminary vote in order to coarsely localize each metadata M_k within the image. For each model from the training, we sum the Gaussian function with high standard deviation values $\sigma x = 0.4 \times W$ and $\sigma y = 0.2 \times H$ because the localization of the metadata by using their relative position is imprecise. We sum the 2D Gaussian function for all (a,b) within the limits of the pool $H \times W$ (4). The 2D Gaussian function has a width two times larger than its height because the positions of text blocks in documents vary more horizontally than vertically.

$$Pool[k][a][b] += \sum_{MMSC_k} \sum_{a=xc-3\sigma x}^{a=xc+3\sigma x} \sum_{b=yc-3\sigma y}^{b=yc+3\sigma y} e^{-\frac{(a-xc)^2}{2\times\sigma x^2} - \frac{(b-yc)^2}{2\times\sigma y^2}} \tag{4}$$

$$(a, b) \in [0..W-1] \times [0..H-1] \; (xc, yc) = (M_k \cdot x, M_k \cdot y)$$

5.2 Voting by Using the Metadata to Captions Spatial Context

The Metadata to Captions Spatial Context n°k ($MCSC_k$) allows the detection of the metadata M_k from the possible captions localized by the words W_i with the lexical class of the captions of the metadata n°k and aligned with M_k. We also use a Gaussian function with small standard deviations $\sigma y = TextHeight/2$ and $\sigma x = \sigma y \times 2$ that depend on the text height of the word W_i. For each word W_i which potentially is a caption having the lexical class of the metadata n°k, for each $MCSC_k$, we compute all possible positions of the metadata (xc,yc) by using the word position W_i and the vectors $-\overrightarrow{U_i}$. Then for each (a,b) coordinate in the limits of the pool, we sum the Gaussian function values (5).

$$Pool[k][a][b] += \sum_{Wi} \sum_{MCSC_k} \sum_{\vec{U}_i} \sum_{a=xc-3\sigma x}^{a=xc+3\sigma x} \sum_{b=yc-3\sigma y}^{b=yc+3\sigma y} e^{-\frac{(a-xc)^2}{2\times\sigma x^2} - \frac{(b-yc)^2}{2\times\sigma y^2}} \tag{5}$$

$$(a, b) \in [0..W-1] \times [0..H-1]\ Lexical\ Class(W_i) = k$$

$$(xc, yc) = (W_i \cdot x - U_i \cdot dx,\ W_i \cdot y - U_i \cdot dy)$$

The voting applied for document structure recognition has a very low complexity in comparison to the voting for object detection in the computer vision domain. The voting process is fast because there are only 2 parameters in a 2D spatial pool and a reduced number of words.

5.3 Voting by Using the Metadata Format

Most of the metadata are described by a format or a regular expression. Among all formats, we only selected 4 regular expressions or formats of metadata that match the 10 metadata to detect.

- ***Date:*** DOCDATE, DUEDATE
- ***Number:*** DOCNBR, ORDERNBR, DELIVNBR
- ***Amount:*** TOTAMT, NETAMT
- **Percentage:** TAXRATE

The format of a word W is detected by the regular expression regex(W) or by heuristics if regex() fails. We start a vote around all words W_i that have a regular expression or a format compatible with the metadata to detect. Like the other voting stages, we use a Gaussian function with small standard deviations $\sigma y = TextHeight/2$ and $\sigma x = \sigma y \times 2$ (6).

$$Pool[k][a][b] += \sum_{Wi} \sum_{a=xc-3\sigma x}^{a=xc+3\sigma x} \sum_{b=yc-3\sigma y}^{b=yc+3\sigma y} e^{-\frac{(a-xc)^2}{2\times\sigma x^2} - \frac{(b-yc)^2}{2\times\sigma y^2}} \tag{6}$$

$$FormatOfMetadata(W_i) = k$$

$$(xc, yc) = (W_i \cdot x,\ W_i \cdot y)$$

DOCTYPE, CURRENCY, TAXRATE have no votes because there is no text format for these metadata. These metadata will be detected in the voting stage by MCSC or MMSC.

5.4 Voting by Using the MMSC

Voting with the MMSC requires the coarse localization of each metadata with previous voting stages. For each metadata class k_1, for each local optima in position (*xc*,*yc*) with a normalized value *Pool2[k1][xc][yc]/Max{Pool2[k1]}* superior to a threshold ε, and for all metadata class k_2 different from k_1, we sum the Gaussian weights around position (*xc*,*yc*) (7). These coordinates are deduced from the spatial relation between the metadata k_1 and the metadata k_2 contained in the $MMSC_{k1}$ and the possible position of the metadata of k_1 localized by (*xc*,*yc*). To avoid interference we analyze pool2, which is a copy of the original pool. We use small standard deviations $\sigma y = TextHeight/2$ and $\sigma x = 2 \times \sigma y$. The threshold ε is fixed to a very high value with $\varepsilon = 0.9$. A vote is started for each local maximum value of the pool superior to 90% of the absolute highest value of the pool. This important step predicts the possible positions of the metadata from previous votes.

$$Pool[k2][a][b] += \sum_{k1=0}^{k1<NM} \sum_{\substack{Optima \\ Pool2[k1]}} \sum_{\substack{k2=0 \\ k1\neq k2}}^{k2<NM} \sum_{a=xc-3\sigma x}^{a=xc+3\sigma x} \sum_{b=yc-3\sigma y}^{b=yc+3\sigma y} e^{-\frac{(a-xc)^2}{2\times\sigma x^2} - \frac{(b-yc)^2}{2\times\sigma y^2}}$$

$$(xc, yc) = argmax\{Pool2[k1][x][y] / Pool2[k1][xc][yc] > \varepsilon \times Max\{Pool2[k1]\}\} \quad (7)$$

5.5 Detection Stage

We have 10 metadata to retrieve from the logical structure [DOCDATE, DOCNBR, TOTAMT, NETAMT, DOCTYPE, CURRENCY, DUEDATE, ORDERNBR, DELIVNBR, TAXRATE]. The description of the metadata is given in the Table 1. Figure 4 shows the pools contents for the 10 metadata in the same order of the list and for the 4 voting stages in the order of the description.

These pools have been computed from the image of the invoice Fig. 6. The four voting stages vote in the same pool for each of the 10 metadata to detect (Fig. 5).

The detection stage builds a map "Classmap" of possible metadata locations from the 10 final pools. For each coordinate (*x*,*y*) in the image, we compute the list of classes of metadata having normalized pool values that exceed a threshold $\gamma = 0.7$ (8). Each word in position (*x*,*y*) from the document is automatically detected with the metadata returned by non empty values of *ClassMap[x][y]*. If the word overlaps several metadata in the *ClassMap*, we select the metadata that are found more frequently inside its bounding

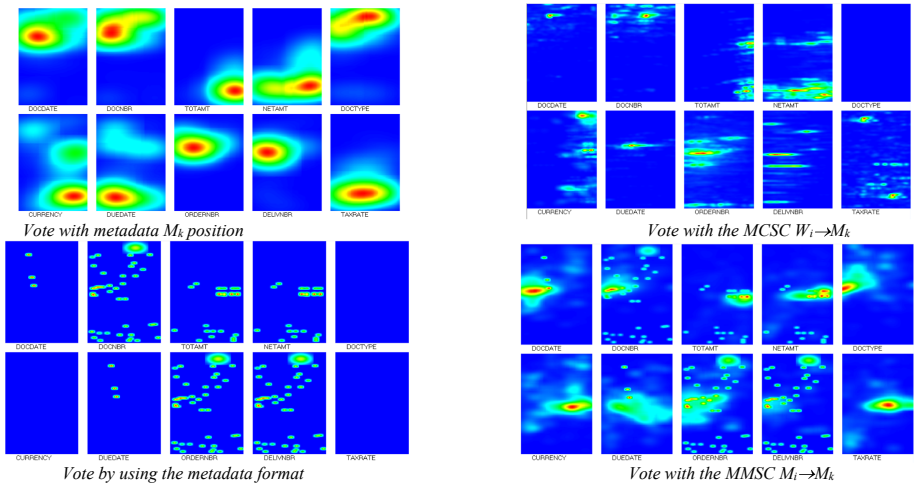


Fig. 4. Pools for the 10 metadata and the 4 voting stages

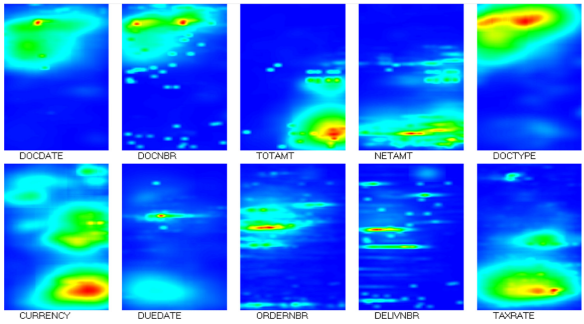


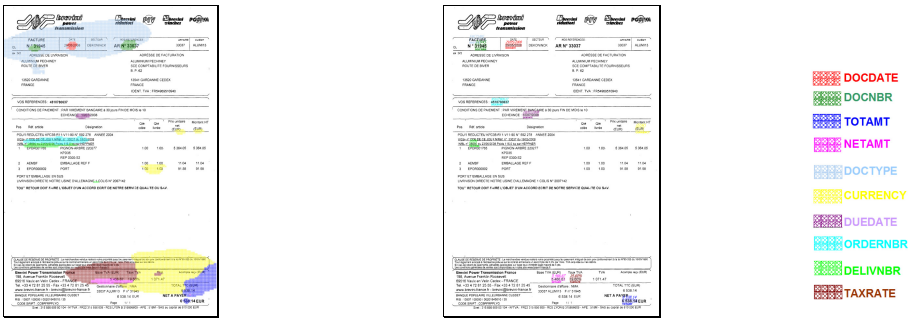
Fig. 5. Pools Results after the voting by the four stages

box. The word is also detected if its format *(IsaNumber(), IsAdate(), IsAnAmount()…)* is compatible with the detected metadata.

$$ClassMap[x][y] = \underset{k=1..NbrOfMetadata}{AllArgMax} \left\{ \frac{Pool[k][x/\alpha][y/\alpha]}{\underset{(i,j)}{Max}\{Pool[k][j][i]\}} > \gamma \right\} \tag{8}$$

$$(x, y) \in [0..ImageWidth] \times [0..ImageHeight]$$

Figure 6 shows the maxima of the combination of the pools from the 4 voting stages and the ground truth. Seven metadata are correctly detected (*DOCTYPE, TOTAMT, TAXRATE, CURRENCY, DELIVRNBR, DUEDATE, DOCDATE)* and three metadata are not correctly detected (*ORDERNBR, DOCNBR, NETAMNT).*



ClassMap[x][y] γ=70% *Ground truth*

Fig. 6. Superposistion of the maxima of the combination of the pools $> \gamma$ compare to the ground truth.

6 Results

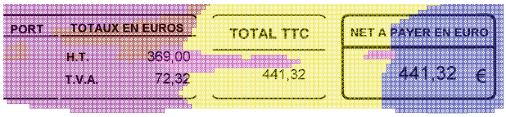
The company which grants this work provides a database of 474 annotated invoices. We cut the database into two equal parts, 237 images for the conception of the part-based structure and 237 different images for the evaluation. We found 228 different templates for 474 images. Most of the templates are represented by only one or two images. Only ten templates are represented by a dozen of images in average. For that, our base is very heterogeneous.

It is absolutely impossible to train a supervised method with only a few hundred samples. But for an unsupervised detection system a knowledge database can be generated with a small number of samples. Because several metadata are repeated several times in different places in the document, the operator arbitrarily chooses only one text block for each metadata. Unfortunately, the same operator can choose different text blocks for the same metadata and the same model of document. Therefore, this database is difficult to use for the construction of a reliable knowledge database even for an unsupervised detection system. Moreover, it impacts the evaluation of the system because the text block chosen by the operator for each metadata may be different from those detected by our system (Fig. 7).

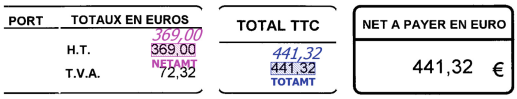
Because the ground truth contains, for each class of metadata, only one text block chosen randomly and not all occurrences of the same metadata, our results will be under evaluated. The database also shows a unknown number of annotation errors. We consider the errors negligible for the evaluation.

The results (Table 1) are encouraging if we consider that they are under evaluated by the annotation of the ground truth and the arbitrary choice of only one sample among several occurrence of repeated metadata.

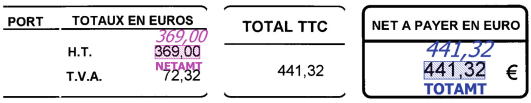
Our detection system is completely unsupervised and works well with a reduced number of images. Several hundred of images from hundreds of different models of invoices are sufficient to build a good knowledge of all spatial relationships. This excellent property is not shared by the supervised approaches, especially methods based on neuronal networks that require from thousands to millions of images for their training. Moreover, our system is highly scalable and perfectible by adding new images from



ClassMap shows that TOTAMT is detected in the right



TOTAMT is chosen in the center in the ground truth



TOTAMT is falsely detected by the evaluation system

Fig. 7. Example of metadata correctly detected but considered as wrongly detected

Table 1. Results of the logical structure recognition on invoices

Label	Description	Detection rate	Nbr of objects
DOCDATE	The date of the document	74,56%	228
DOCNBR	The document number	85,46%	227
TOTAMT	The total amount after taxes	93,69%	222
NETAMT	The net amount	88,88%	225
DUEDATE	The date of payment	82,77%	180
DOCTYPE	The type of document	76,54%	226
ORDERNBR	The order number	90,82%	229
DELIVNBR	The delivery number	92,18%	64
CURRENCY	The currency of the amount	68,49%	219
TAXRATE	The tax rate applied	74,52%	212
TOTAL		**82,79%**	**2032**

numerous other models of invoices. The more images and models of invoices that are provided to the system, more performant the detection will be. This property is explained by the robustness of the partial voting. The classes of metadata {DOCDATE, DOCTYPE, CURRENCY, TAXRATE} are the more difficult to detect, which is explained by the fact that they are not always associated with a caption.

Unfortunately we cannot compare ourselves directly to the works described as the datasets used are not publicly available and the evaluation methods are different. It is also difficult to compare our self to other industrial works for the same reason. It is also hard to create our own dataset due to privacy restrictions. We sincerely believe that such a dataset if exists in the future will contribute to advance the domain significantly.

However, all previous works use supervised approaches to recognize documents with a rigid logical structure, which never change spatially. These approaches are trained on the specific models of each document to read. In contrast to existing systems for business documents, our recognition rate is absolutely given without any heuristics, post-processing steps and contextual enhancements. The results confirm the several assumptions we made at the beginning of the work Sect. 3.

7 Conclusion

Structure recognition is an emerging field that is beginning to break into effective and sufficiently generic platforms. If we consider the same time scale that has been needed to develop OCR into an unmarked industrial product, the development of document-structure recognition software will require many years of research. However, the need for automatic recognition of structures is increasingly urgent in the face of current digitization projects. Difficult problems remain to be solved. The modeling and recognition of the logical structure remains the Achilles' heel of recognition systems.

At present, each system developed in public or private laboratories operates on specific documents, with structure that is regular, rigid, and is either predictable or already known in advance. Therefore, even today, there are no structure-recognition systems that can automatically decode the structure of any text.

In this paper, we have proposed the very first automatic and scalable system which can recognize the logical structure of business documents without any models or prior information about their logical structure. Our solution can process totally unknown new models of documents. Our detection system can also deal with non-rigid structures.

We have proposed *a part-based spatial model suited for partial voting*. Our proposed model introduce the concept of Spatial Context (SC) as a spatial feature, which describes the relative positions of metadata or caption in a Neighborhood around each word of interest. We have introduced two different types of SC: *the Metadata to Captions Spatial Context (MCSC) and the Metadata to Metadata Spatial Context (MMSC).* The MCSC memorizes the spatial relation between possible captions detected by the lexical classification of words and the metadata to detect. The MMSC measures the spatial relations between neighboring metadata.

We introduce an original and robust Gaussian voting process in the spatial space that allows to localize and recognize automatically a researched metadata and its logical function. Our voting process is robust against missing information, OCRerrors and annotation errors. Our detection system is completely unsupervised and is working well with a much-reduced number of images. This excellent property is not shared by the supervised approaches, especially methods based on neuronal networks.

In future works, we want to explore other applications in DIA of our part-based model of detection.

Acknowledgement. This work was granted by ITESOFT and LIRIS Lab from INSA-LYON for the project DOD.

References

1. Srihari, N., et al.: Name and address block reader system for tax form processing. In: ICDAR, pp. 5–10 (1995)
2. Mao, J., et al.: A system for automatically reading IATA flight coupons. In: ICDAR97, pp. 153–157 (1997)
3. Cesarini, F., et al.: Trainable table location in document images. In: ICPR (3), pp. 236–240 (2002)
4. Gatos, B., Danatsas, D., Pratikakis, I., Perantonis, S.J.: Automatic table detection in document images. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3686, pp. 609–618. Springer, Heidelberg (2005). https://doi.org/10.1007/11551188_67
5. Coüasnon, B., et al.: Dmos, a generic document recognition method: application to table structure analysis in a general and in a specific way. In: IJDAR, pp. 111–122 (2006)
6. Klein, B., et al.: Three approaches to "industrial" table spotting. In: ICDAR, pp. 513–517 (2001)
7. Coüasnon, B., et al.: DMOS, It's your turn! In: 1st International Workshop on Open Services and Tools for Document Analysis. ICDAR17
8. Mao, J., et al.: A model-based form processing sub-system. In: ICPR (1996)
9. Ting, A., et al.: Business form classification using strings. In: ICPR 96, p. 690
10. Héroux, P.: Etude de méthhodes de classification pour l'identification automatique de classes de formulaires. In: CIFED (1998)
11. Duygulu, P.: A hierarchical representation of form documents for identification and retrieval. IJDAR **5**(1), 17–27 (2002)
12. Ishitani, Y., et al.: Model based information extraction and its application to document images. In: DLIA (2001)
13. Cesarini, F., et al.: INFORMys: A Flexible Invoice-Like Form-Reader System. In: IEEE PAMI, pp. 710–745 (1998)
14. Cesarini, F., et al.: Analysis and understanding of multi-class invoices. IJDAR **6**(2), 102–114 (2003)
15. Hamza, H., et al.: Incremental classification of invoice documents. ICPR, pp. 1–4 (2008)
16. Hamza, H., et al.: Application du raisonnement à partir de cas à l'analyse de documents administratifs. Nancy2 University, France (2008)
17. Tateisi, Y., et al.: Using stochastic syntactic analysis for extracting a logical structure from a document image. In: ICPR, pp. 391–394 (1994)
18. Belaïd, Y., et al.: Form analysis by neural classification of cells. In: DAS, pp. 58–71 (1998)
19. Tsuji, Y., et al.: Document recognition system with layout structure generator. In: Proceedings of the MVA (1990)
20. Yamashita, A., et al.: A model based layout understading method for the document recogntion system. In: ICDAR, pp. 130–138 (1991)
21. LeBourgeois, F., et al.: Document understanding using probabilistic relaxation: application on tables of contents of periodicals. In: ICDAR, pp. 508–512 (2001)
22. Lebourgeois, F.: Localisation de textes dans une image 'a niveaux de gris. In: CNED 1996, pp. 207–214
23. Hough, P.V.C.: Method and means for recognizing complex patterns, U.S. Patent 3,069,654, December 18 (1962)
24. Duda, R.O. et al.: Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM **72**, 11–15
25. Ballard, et al.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Recogn. **13**(2), pp. 111–122 (1981)

26. Medioni, G., et al.: 3-D structures for generic object recognition. In: ICPR, pp. 1030–1037 (2000)
27. Opelt, A., Pinz, A., Zisserman, A.: A boundary-fragment-model for object detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 575–588. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_44
28. Leibe, B., et al.: Robust object detection with interleaved categorization and segmentation. Int. J. Comp. Vis. **77**(1–3), 259–289 (2008)
29. Rusinol, M., et al.: Field extraction from administrative documents by incremental structural templates. In: ICDAR, pp. 1100–1104 (2013)
30. Dengel, A.R., Klein, B.: smartFIX: a requirements-driven system for document analysis and understanding. In: Lopresti, D., Hu, J., Kashi, R. (eds.) DAS 2002. LNCS, vol. 2423, pp. 433–444. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45869-7_47
31. Palm, R.B., et al.: Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In: ICDAR, pp. 406–413 (2017)
32. https://www.kofax.com/-/media/Files/Datasheets/EN/ps_kofax-readsoft-invoices_en.pdf
33. https://www.abbyy.com/media/16413/fcadminguide0.pdf
34. Schuster, D., et al.: Intellix – end-user trained information extraction for document archiving. In: ICDAR, pp. 101–105 (2013)
35. Liyuan, L., et al.: On the variance of the adaptive learning rate and beyond. In: ICLR (2020)
36. Katti, A.R., et al.: Chargrid: towards understanding 2d documents. In: EMNLP, pp. 4459–4469 (2018)
37. Zhao, X., et al.: CUTIE: learning to understand documents with convolutional universal text information extractor (2019)
38. Denk, T.I., et al.: Bertgrid: Contextualized embedding for 2d document representation and understanding. CoRR,abs/1909.04948 (2019)
39. Xiaojing, L., et al.: Graph convolution for multimodal information extraction from visually rich documents. In: NAACL, pp. 32–39 (2019)
40. Majumder, B.P., et al.: Representation learning for information extraction from form-like documents. In: ACL, pp. 6495–6504 (2020)
41. Gogar, T., Hubacek, O., Sedivy, J.: Deep neural networks for web page information extraction. In: IFIP AIAI (2016)
42. Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y.: Block-based web search. In: SIGIR, pp. 456–463 (2004). Yu et al. 2003
43. Yu, S., et al.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: WWW, pp. 11–18 (2003)
44. Zhu, J., et al.: Simultaneous record detection and attribute labeling in web data extraction. In: KDD, pp. 494–503 (2006)
45. Lample, et al.: Neural architectures for named entity recognition. In: NAACL, pp. 260–270 (2016)
46. Yang, X.: Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: CVPR (2017)
47. Peng, N., Poon, H., Quirk, C., Toutanova, K., Yih, W.-T.: Cross-sentence N-ary relation extraction with graph LSTMs. Trans. Assoc. Comput. Linguist. **5**, 101–115 (2017)
48. Kessi, L., Lebourgeois, F., Garcia, C.: An efficient new PDE-based characters reconstruction after graphics removal. In: ICFHR, pp. 441–446 (2016)
49. Kessi, L., Lebourgeois, F., Garcia, C.: An efficient image registration method based on modified nonlocal-means - application to color business document images. VISAPP (1), pp. 166–173 (2015)
50. Kessi, L., Lebourgeois, F., Garcia, C., Duong, J.: AColDPS - robust and unsupervised automatic color document processing system. In: VISAPP (1), pp. 174–185 (2015)
51. Kessi, L., Lebourgeois, F., Garcia, C.: AColDSS: robust unsupervised automatic color segmentation system for noisy heterogeneous document images. EPS (2015)