Introduction of statistical information in a syntactic analyser for document image recognition

André O. Maroneze, Bertrand Coüasnon, Aurélie Lemaitre

Introduction of statistical information in a syntactic analyzer for document image recognition

André O. Maroneze[a] and Bertrand Coüasnon[a] and Aurélie Lemaitre[b]

[a]INSA - Irisa - UEB, Campus de Beaulieu, 35043 Rennes, France;
[b]University of Rennes 2 - Irisa - UEB, Campus de Villejean, 35043 Rennes, France

ABSTRACT

This paper presents an improvement to document layout analysis systems, offering a possible solution to Sayre's paradox (which states that an element "must be recognized before it can be segmented; and it must be segmented before it can be recognized"). This improvement, based on stochastic parsing, allows integration of statistical information, obtained from recognizers, during syntactic layout analysis. We present how this fusion of numeric and symbolic information in a feedback loop can be applied to syntactic methods to improve document description expressiveness. To limit combinatorial explosion during exploration of solutions, we devised an operator that allows optional activation of the stochastic parsing mechanism. Our evaluation on 1250 handwritten business letters shows this method allows the improvement of global recognition scores.

Keywords: layout analysis, structure recognition, stochastic parsing, content-based analysis, handwritten letters

1. INTRODUCTION

Document analysis systems allow the automatic processing of paper documents. Once a document image is obtained, several processing stages are needed before the information is made available. Skipping the low-level stages such as binarization, noise reduction and component extraction, we can divide the process in two parts: structure analysis (layout and organization) and content recognition (the conversion from an image to the value it represents). An example of structure analysis is presented in figure 1.

Both tasks are complementary, and to obtain best results there must be an interleaving between them, since an element "cannot be segmented before having been recognized and cannot be recognized before having been segmented". This is known as Sayre's paradox[1]; we mention both aspects of this paradox separately as: recognition guiding localization (*recognition* → *localization*) and localization guiding recognition (*localization* → *recognition*). It is therefore crucial that any document analysis system be able to incorporate both kinds of information (structural and content-based) during processing.

In this paper, we are mainly concerned with structure analysis. This process can be improved if it incorporates content-based information obtained from *recognizers*, but such information associates uncertainty and numeric aspects that require special consideration to work adequately with the symbolic nature of structure analysis. Usually, *stochastic parsing* is applied to solve these issues, but it may itself incur a combinatorial explosion of solutions to be explored, which must be taken into account.

We present in this paper a way to integrate such content-based information into the structure analysis. Information can flow in both directions (from content to structure and vice versa) multiple times, creating a feedback loop that increases document description expressiveness and improves recognition rates.

In section 2, we present some related work on document analysis systems and the limitations that motivated our research. In section 3, we present a solution to the different aspects required by the introduction of content-based information; as an example of implementation, we use the DMOS-P (*Description and MOdification of*

Further author information: (Send correspondence to Aurélie Lemaitre)
André O. Maroneze: E-mail: aoliveir@irisa.fr
Bertrand Coüasnon: E-mail: couasnon@irisa.fr
Aurélie Lemaitre: E-mail: aurelie.lemaitre@irisa.fr

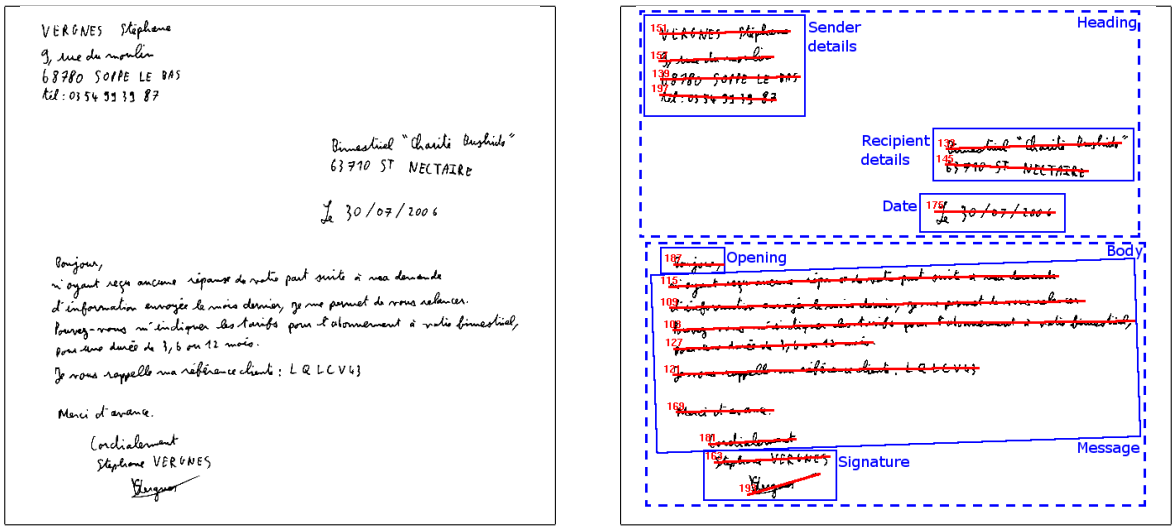


FIG. 1: Example of a document (a handwritten business letter) and a possible result of structure analysis: text lines, labeled blocks (solid rectangles) and superblocks (dashed rectangles).

Segmentation with Perceptive vision) method (presented in subsection 2.4). In section 4, we evaluate our new method on two contexts, one related to handwritten business letters and another related to historical documents. Our evaluation on 1250 handwritten business letters shows improvements in terms of increased expressiveness (allowing a simpler document description) and higher recognition rates. We also present a way to limit the combinatorial explosion due to the stochastic parsing mechanism. Finally, we conclude the paper in section 5.

2. RELATED WORK

Document layout analysis can be performed using different classes of algorithms. Some of them bring out the statistical similarities between several documents of the same class, using this information to infer the layout of new instances; they perform a mainly *statistical* layout analysis. Others emphasize the application of rules to describe the structure, defining *rule-based* representations that can become rather arbitrary[2]; to allow a more precise semantic description for the relations between elements, these methods may incorporate notions based on formal grammars, performing what is usually called *syntactic* (or *grammatical*) layout analysis.

We describe both approaches, statistical and syntactic, along with their shortcomings. We then present the DMOS-P syntactic method, which will be used as example of an application context.

2.1 Statistical methods

Statistical layout analysis methods are based on different formalisms, such as 2D Markov Random Fields[3] and Conditional Random Fields.[4] One key advantage of these methods is the incorporation of noise and uncertainty, often present in document analysis.[2]

Statistical methods usually lack the ability to convey the hierarchical structure of a document; while it is not needed for simple layouts, for more complex documents, such as large tables and mathematical expressions, this can be a major limitation. These methods can deal with noise and local variations, but their inference of the global structure is limited and the information obtained may not be interpretable. In syntactic methods, this global structure is described in document grammars, which lend themselves to human interpretation.

Regarding Sayre's paradox, statistical methods do not allow tight incorporation of the *recognition* $\rightarrow$ *localization* step; for instance, Lemaitre et al.[3] mention the possibility of integrating content recognition to improve the analysis, but this solution, applied at a post-processing stage, does not create a feedback loop. Montreuil et al.[4] mention the integration of textual information to improve the segmentation process. While this does relate to Sayre's paradox, there seems to be no continuous cyclic interaction between both aspects.

2.2 Syntactic (or grammatical) methods

Syntactic methods segment the image in primitives (also called *tokens*) and build a rule tree that describes how to compose these primitives. One of their advantages is the natural expression of recursive structures, such as hierarchical ones. Their rule-based description is often more amenable to human interpretation. If, on the one hand, they do not benefit from an automatic learning process based on evidence, such as the one present in statistical methods, on the other hand they allow background knowledge to be input via the document description grammar.

Syntactic methods usually suffer from a lack of flexibility, having issues with heterogeneous structures. Adding "soft" constraints (usually via numeric attributes that represent costs) can help avoid this situation. For instance, Artières[5] used probabilistic features in his grammar to deal with ambiguous (flexible) situations. However, the use of bottom-up parsing during his process limits the feedback step *localization* $\rightarrow$ *recognition*. A similar situation happens with Fitzgerald et al.[6] on the analysis of handwritten mathematical expressions. Their *fuzzy parsing* technique (also based on a bottom-up parser) can deal with ambiguities, but cannot handle noise.

Both approaches work within their application context (online documents) but do not generalize well to offline contents. The more flexible parsing mechanism is more costly, but the use of bottom-up parsing limits the increase of the search space; however, it also limits its capacity to deal with noise. Highly structured offline documents require a different strategy to be able to deal with the growth in the number of subtrees to be explored, while keeping the ability to reject noise.

2.3 Mixing statistical and syntactic methods

To increase the flexibility of syntactic methods, namely for offline document recognition, some authors have proposed using stochastic grammars. Tateisi and Itoh[7] use a stochastic regular grammar to analyze printed documents. They associate costs with each element, enumerate all solutions in the form of a graph and search the minimal cost path. This is feasible since they use a regular 1D-grammar, which limits the combinatorial explosion of solutions. Another approach, by Mao and Kanungo,[8] uses a Stochastic Context-Free Grammar (SCFG) to process printed bilingual dictionaries using a top-down parser, incorporating uncertainty and noise. Both approaches integrate reasonably well with Sayre's paradox; however, the use of 1D-grammars prevents generalization to less structured documents.

All in all, a generic syntactic document analysis framework should consider: using features to incorporate context or uncertainty (allowing the insertion of numeric attributes); using a 2D-grammar to ensure a good level of expressiveness; dealing with ambiguities; handling noise; and managing the complexity of the solution search space. These requirements allow genericity and maintain a feedback loop between localization and recognition.

By mixing the numeric aspect of statistical methods (for flexibility and incorporation of local variations) with the symbolic aspect of syntactic methods (for complex global structuring and interpretation of the results), we can obtain a satisfying solution to Sayre's paradox. Our proposal deals with the issues resulting from this integration, such as how to symbolically interpret numeric values, and complexity issues due to a larger search space.

2.4 The DMOS-P syntactic method

We consider in this paper an existing grammatical framework, DMOS-P, which incorporates several of the cited requirements but currently lacks the integration of content-based information (*recognition* $\rightarrow$ *localization*).

DMOS-P[9,10] is a generic document recognition method comprising a grammatical language, the *Enhanced Position Formalism* (EPF), and an associated parser. This method has already been applied to a wide variety of document types (historical documents, music scores, mathematical expressions, etc.). For each new document class, one only needs to write an EPF grammar describing it and DMOS-P will compile it to generate the corresponding parser, ensuring separation between domain-specific and general, reusable knowledge.

The DMOS-P parser uses a 2D-grammar and performs a top-down analysis with backtrack, allowing the structure to direct content recognition (*localization* $\rightarrow$ *recognition*). However, like most deterministic multidimensional parsers, it does not handle ambiguities, choosing arbitrarily: the first solution found is always taken, even if there are other solutions available. In fact, it is not possible to rank different acceptable solutions, using

"soft" constraints such as scores. The lack of control of the first solution found leads to unsatisfactory results for ambiguous (despite flexible) grammars, often yielding a non-optimal result. Several authors[2,11,12] recommend using stochastic parsing to solve this problem.

3. PROPOSAL: INCORPORATION OF NUMERIC INFORMATION IN ANY SYNTACTIC ANALYSIS FRAMEWORK

In this paper, we consider content-based information available via statistical recognizers, which apply statistical methods to analyze the input, associating uncertainty (a numeric value) to the recognized elements. During parsing, this uncertainty leads to multiple possible analyses and therefore ambiguities. When dealing with multiple choices, we want the analysis to apply the best result possible; therefore, first we (1) define some preference criteria to rank the possibilities, then we (2) find and choose the best one. We propose the incorporation of these criteria in any syntactic method via the generic notion of *scores* associated to grammar rules.

3.1 Ranking solutions using scores

Based on the idea of rule weights (commonly used in stochastic grammars[13]), we assign a score (a positive value) to each grammar rule and interpret this value as a cost, or penalty, which we try to minimize: if our analysis produces multiple ambiguous results, the one with the lowest penalty is considered the best and thus chosen first. No particular order is defined for results with same score.

We propose the incorporation of these scores via a grammar operator, integrated directly into the syntactic document description. A grammar writer may incorporate as many different scores as he wishes, even from different natures; for instance, uncertainty obtained from recognizers and user-defined distances between elements. We do not constrain the scores to probabilistic values, avoiding normalization issues which would unnecessarily (since our main interest is the notion of *preference*) complicate their usage. Scores (costs) are intuitively combined by addition, which avoids underflow issues. Stochastic parsing usually combines them by multiplication; to reproduce this behavior, we can use logarithmic values if needed.

Once we have established a means of ranking different ambiguous analyses, we can proceed to the search of the best result.

3.2 Obtaining the best solution

Our proposal allows the integration of stochastic analysis in a seamless way: we define a new description grammar operator, `FIND_BEST_FIRST`:

`FIND_BEST_FIRST (R : rule) : rule` - returns the best result for the subtree defined by rule `R`.

This operator takes a grammar rule `R` as parameter and is itself processed as a grammar rule. In other words, it acts as a *decorator* for an existing rule: if the rule produces only one result, then `FIND_BEST_FIRST` will just propagate it (it has no visible effect). If there are multiple results, however, this operator ensures the best one (according to the minimal penalty criterion) will be returned in the first place. Additionally, this operator preserves the semantics of backtracking where it is used: if the analysis fails, it can backtrack to the second best solution, then the third one, and so on.

Our incorporation of stochastic analysis in a limited subset of the parsing tree (via a rule decorator) has the advantage of limiting the extension of the exploration while benefiting from the grammar writer's knowledge about possible ambiguities. For instance, there may be several ambiguous analyses which are equivalent from the grammar writer's point of view, but which would entail an unnecessary increase in the number of solutions to be explored. The operator can then be placed somewhere closer to the leaves.

Our integration of generic scores and an exploration operator allows a grammar writer to easily incorporate stochastic parsing, as long as a useful metric to rank analyses is available. This metric can represent any sort of numeric data, establishing soft constraints that, in combination with syntactic "hard" constraints, allow flexibility and higher expressiveness for the description of documents.

3.3 Combining scores and exploration

One of the uses of the proposed stochastic analysis process is *content-based analysis*: we want to use content information to direct the localization (*recognition* → *localization*). This allows more expressive (and hopefully simpler) document descriptions, since there is more information available to the analysis process. For instance, in the context of mail analysis, if we are interested in finding the subject, instead of looking for "a reasonably short text line, possibly indented, possibly centered, separated from the previous text line by more than one average line height", we can look for "a text line beginning with the word 'subject"'.

In such situations, we might rely entirely on the content-based information (using such text line to separate the heading from the body of a letter, for instance), or we can use both kinds of information, looking for "the first text line beginning with the word 'Subject' that is in the upper half of the document", to reduce the chance of mistakenly choosing a text line from the message body. Note that this last version of the 'subject' rule is still much simpler to describe than the first one. It requires, however, an interleaving between structural and content-based exploration.

The proposed grammar rule decorator `FIND_BEST_FIRST` allows us to alternate between both mechanisms in a way that is transparent to the grammar writer: `FIND_BEST_FIRST` activates the exploration mechanism and combines the scores from possible subtrees. The required effort for a grammar writer consists only in:

- defining suitable metrics (scores) for choosing among ambiguous results;
- defining rules making use of the FIND_BEST_FIRST operator to guide the analysis (placing it as close as possible to the leaves to avoid useless exploration).

Note that the way the `FIND_BEST_FIRST` operator has been introduced, the semantics of the analysis remains unchanged outside its scope. It means the grammar writer does not need to choose between both analyses (structural and content-based), being encouraged to benefit from both.

We now illustrate the application of the proposed mechanism in a concrete example.

3.4 A practical example: analysis of handwritten business letters

We present an example in the context of handwritten business letter recognition: suppose we want to identify the standard zones present in a letter (such as those in figure 1) to aid in further document processing. For handwritten documents such as business letters, which do not have very strict layout constraints, their large structural variability hinders the efficiency of structural rules, so we envisage using content-based analysis as described previously, using the following rule: "the *opening* line of a French letter is a line beginning with the words *Madame* or *Monsieur*" (this is true for almost 90% of all letter openings in our evaluation database, described in subsection 4.1). We will then use this line to separate the heading from the body of the message. To detail our example in a concrete framework, we will use the DMOS-P method in what follows, though the general mechanism is applicable to any syntactic analyzer.

We can encode the previous description of the opening line ("a line beginning with the words *Madame* or *Monsieur*") in a grammar where each token is a handwritten text line*, with a rule that searches all lines for the desired content. Such a grammar is presented in figure 2 (in EPF syntax).

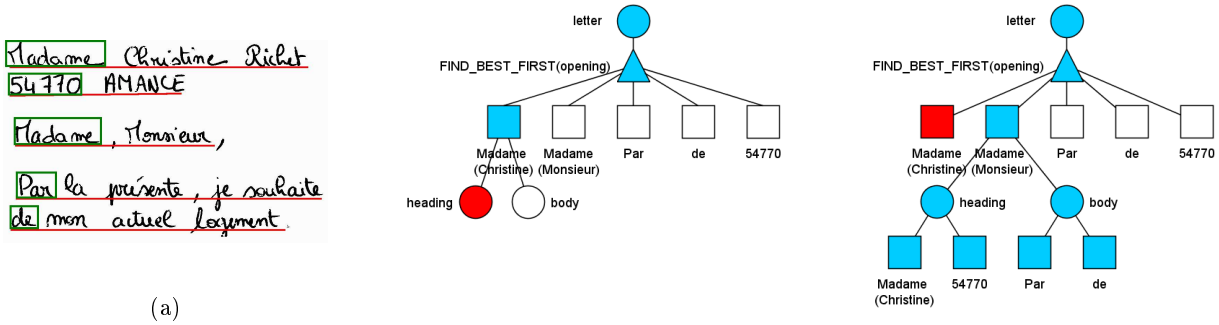
This simple grammar uses a statistical recognizer based on the first word of each line, invoked via the predicate `recognize_first_word`. This predicate has two inputs, a text line and a list of possible words to be recognized, and one output, the associated uncertainty during the recognition process. Ambiguities are caused by the presence of the Prolog-like `member` predicate, which is backtrackable and may succeed with different elements (giving flexibility to our rule).

It is important to define rules that, albeit flexible, prevent incorrect results from being accepted; this is usually done using simple structural constraints. For instance, in the application context of handwritten business letters, there are some documents where the first heading line also begins with *Madame*, misleading the recognizer (see figure 3).

*We consider the "lexical" analysis producing these tokens has already been performed by another method, such as the one used by Lemaitre.[10]

```
1    letter ::=
2        AT(wholePage) &&                  % position operator: where to search in the image
3        FIND_BEST_FIRST(opening) &&       % finds and consumes a token (a text line)
4        AT(aboveOpening) && heading &&    % continues analysis with rule 'heading'
5        AT(belowOpening) && body.         % finishes analysis with rule 'body'
6
7    opening ::=
8        member Line AllLines &&           % any line can be chosen as the opening line
9        recognize_first_word Line ["Madame", "Monsieur"] Uncertainty &&
10       ADD_SCORE(Uncertainty).           % score used by FIND_BEST_FIRST to sort results
```

Figure 2: Example of an EPF document description for handwritten letters. In this Prolog-based syntax, rule names begin with lowercase characters, grammar operators are written entirely in uppercase and variables begin with an uppercase character. Note that `Uncertainty` is an outbound parameter in its first occurrence (line 9) and an inbound parameter in line 10.



(a) (b) (c)

Figure 3: Example of a fragment of a handwritten letter (a) and intermediate analysis trees of this fragment generated by our simple grammar (b and c). The nodes are sorted according to recognition scores, from best to worst: the leftmost node is the first to be explored.

In this example, our analysis begins by positioning the parser with `AT(wholepage)`, which defines the entire document as the active region, then applying the `FIND_BEST_FIRST` operator to the opening line. Several analysis branches (square nodes in figure 3b) are acceptable, and they are ranked according to the uncertainty of the recognizer. We suppose it assigned here the lowest penalties to the two lines beginning with *Madame*, as expected. The order of the remaining lines is not important in this example, since their branches will not be explored.

We can see that our content-based constraint is not enough to identify the correct parse. Indeed, in our example the first line (*Madame Christine*) was chosen as the opening line instead of the third one (*Madame, Monsieur*). Since both lines correspond to what the recognizer expects, they are equally likely to be returned (in practice, the difference between their scores is due to random noise). The integration of structural constraints will allow us to prune this incorrect parse.

After processing the `FIND_BEST_FIRST` operator, our analysis proceeds with `AT(aboveOpening) && heading` (figure 3b). The `AT` operator restricts the analysis to the zone above the opening line (in our case, empty space), then the `heading` rule (omitted here) is applied. This rule tries to find at least one token (text line), but there are none in the active region. The `heading` rule fails, forcing the parsing mechanism to backtrack to the second branch (figure 3c), adopting *Madame, Monsieur* as the opening line. This time, the `heading` rule will succeed and so will `body`. The final result is the expected parse.

A final remark about the mechanism: since the definition of the scores is arbitrary, several different metrics could have been incorporated to compensate for an eventual poor recognizer performance. For instance, since we are interested in identifying regions of similar information which are often structured as contiguous text zones,

we might apply a penalty based on the inter-line distance. Our mechanism would then look for a configuration minimizing total inter-line distance.

With the possibility of seamlessly integrating both structural and content-based information, we increase the expressiveness of document descriptions, allowing them to be stated in a simpler way according to the available information.

4. EVALUATION AND RESULTS

To evaluate our proposal, we considered a particular implementation: a stochastic parsing extension developed for the DMOS-P method. After the development of this extension, three different aspects have been examined: expressiveness of the document description, recognition rate and performance (in terms of number of explored solutions). These aspects have been evaluated using two different document recognition contexts, assessing the genericity of our mechanism: handwritten business letters and historical documents. Grammar expressiveness and recognition rate have been evaluated on the former, while performance has been evaluated on the latter. For each application, we present the context first, then we detail the performed evaluation.

4.1 RIMES : French handwritten business letter database

The RIMES[†] French national evaluation campaign[14] established a database containing thousands of handwritten letters and faxes. These documents can be used for different tasks related to document recognition: layout analysis, writer identification, handwritten text recognition, etc. Images are 300 dpi grayscale scanned pages. All images have been manually annotated with a ground truth. The document in figure 1 is a sample from this database.

The document structure recognition task in RIMES consists in the identification of up to eight different zones in each image and their assignment to one of the following labels: *sender details* (return address), *recipient details* (inside address), *date/place*, *subject*, *opening*, *message body, signature* and *attachment/postscript*. Unlike the example analysis in figure 1, there are no "superlabels" such as *heading* and *body*. The recognition rate for this task is defined as the recall per class: (*number of assigned black pixels*) / (*number of expected black pixels*). Only black pixels are counted to avoid including the background.

Adapting an existing grammar

We adapted a deterministic grammar which is based on structural (geometric) information obtained from text lines: line length, relative positions and distances. Since the standard DMOS-P method does not rank ambiguous results, this grammar uses a deterministic parser and unambiguous rules. To handle variability in the document structure, some rules have several minor variations which match slightly distinct layouts.

To improve recognition rates, we introduced content-based analysis by replacing the rules relating to the opening line with content-based rules similar to the example in section 3.4. The new rules use a recognizer that can identify typical opening expressions in French[‡]. Using a perceptive vision[10] mechanism, we obtain a segmentation of the document in text lines; these lines are fed to the recognizer, which then outputs its uncertainty. Using this uncertainty as a score, and incorporating the `FIND_BEST_FIRST` operator, we modified the grammar, allowing ambiguity concerning the choice of the opening line. Our mechanism then ensures that, if the opening line is correctly recognized, it will be chosen and the analysis will proceed on the rest of the document.

One of the reasons the opening line has been chosen for content-based analysis is due to its ease of recognition: in the vast majority of documents in the evaluation database, a limited number of different expressions is used. If we consider looking for keywords or some form of limited regular expressions, for instance, the same strategy can be applied to several heading lines, such as the ones containing telephone numbers and postal codes.

[†]RIMES: *Reconnaissance et Indexation de données Manuscrites et fac-similES* (recognition and indexing of handwritten documents and faxes).

[‡]A dozen different expressions, such as *Madame/Monsieur*, *Messieurs*, *Bonjour* and *Chers messieurs,* amount to about 95% of all opening lines in the database.

Another reason for the choice of the opening line is that it is recognized quite early in the analysis, and its position determines the effective division between 'heading' and 'body' rules. Improving the recognition rate of this rule avoids propagating analysis errors, considerably improving the overall result.

During our evaluation in this application context, we compared three different analysis strategies:

1. the existing, structured-based analysis, used as reference;
2. the proposed content-based analysis, incorporating our stochastic DMOS-P extension and opening line recognizers, using the modified grammar described previously;
3. a mixed strategy to further improve on recognition rates. This strategy simply combines both previous strategies in the following way: first, we apply the stochastic analysis, obtaining its uncertainty. If this value is below a certain threshold, this means the recognizer is reasonably confident the line corresponds to one of the expected expressions, and therefore it is very likely it is the right one. However, in some cases, all lines are equally bad from the point of view of the recognizer (for instance, if there is no opening line), so the best score roughly corresponds to random noise during the recognition process. In this case, instead of leaving it to chance, we fallback to the structural version, which is more likely to obtain a good result. Normally, we would not have both definitions available in the first place (only the content-based one would have to be developed), but since we already had the structural version, we benefited from it to further improve recognition results.

These three methods allow us to evaluate description grammar expressiveness and recognition rates.

Description grammar expressiveness

During adaptation of the document grammar to incorporate our stochastic operator, we modified the rules describing the opening line. Instead of the 8 complex rules (involving custom position operators and several test conditions) in the structure-based version, we use only 3 simple rules, all related to the same idea: find the best line matching an opening expression and use it. The integration of recognizers and the stochastic mechanism allow simpler rules, expressed more easily using content-based information than geometrical data. This simplification can be attained while improving recognition rates (as will be seen in figure 4). While it is not easy to quantify this simplification in terms of complexity of the description (though we did obtain a smaller set of rules in our example, the number of grammar rules does not necessarily indicate their complexity), we emphasize the fact that our approach only adds to the description expressiveness, since any existing structural methods can still be used without any changes.

Part of the complexity of the grammar description has been transferred to the fine-tuning of recognizers; since they were already employed afterwards during content recognition, our method allows benefiting from the effort spent in their development in an earlier stage, during structure analysis. The result is the continuation of existing methods (which do not incorporate recognizers) and the addition of content-based constructions, enabled by the recognizers, resulting in a net gain of expressiveness.

Improvement in recognition rate

Figure 4 indicates the recall rate for each relevant class (*Signature* and *PS/Att* are not relevant, since they are defined before the opening line and therefore are not affected by our method).

The target class, *Opening*, is indeed the one which benefits the most from content-based analysis: its recall rate improved from 18.6% in the reference version to 6.4% in the combined strategy, an improvement of 66%. Other classes, such as *Body*, *Date/Place* and *Subject*, experience a decrease in their error rates as a consequence of better segmentation between heading and body. Global recognition rates improved from 92.6% to 94.5% (a decrease of 25% of the error rate). According to the results of the second RIMES test phase,[15] our final error rate is better than the ones presented at the contest (presented in table 1). Note that there have been some improvements between the IRISA version submitted for the contest and the one just before the incorporation of our stochastic extension, which explains the difference between these results and those indicated in figure 4.

In terms of number of solutions to explore, our stochastic adaptation in this context does not present a considerable challenge (there are in average 20 text lines per letter, which amounts to roughly the same number of explored branches); to evaluate this aspect, we chose a different context.

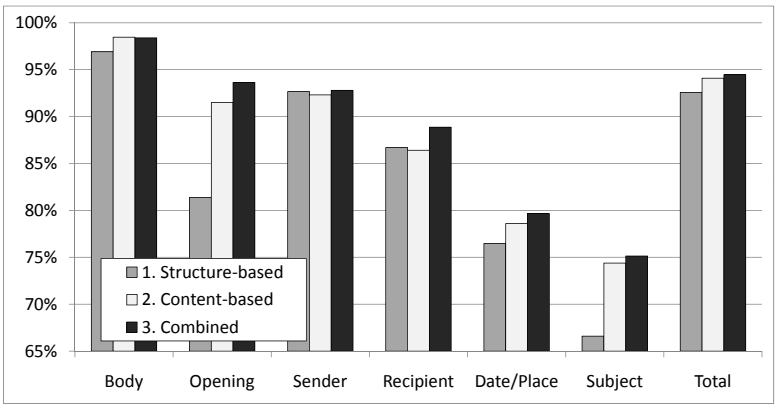


Figure 4: Recall rate for the three handwritten structure analysis strategies: (1) the existing structure-based version, (2) our content-based version using stochastic parsing, and (3) our combined version based on both previous strategies. The *opening* class improved from 81.4% to 93.6% (66% of error reduction), while the total error rate dropped from 92.6% to 5.5% (25% of error reduction). Each strategy has been evaluated on 1250 documents.

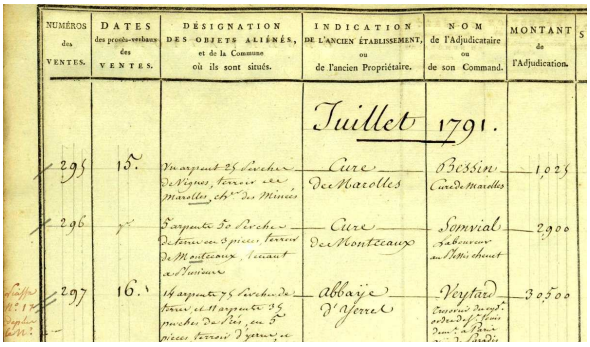
Laboratory	CEP	IRISA	LITIS	CEP/LITIS	Proposed method
Error rate (%)	8.53	8.97	12.62	12.88	5.53

Table 1: Results from the 2009 RIMES evaluation campaign for the layout analysis task, along with the results of our proposed method. For details on the participating teams, we refer the reader to Grosicki et al.[15]

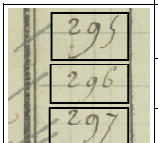
4.2 Historical documents: a performance analysis

To better evaluate the performance of our mechanism (in terms of number of explored solutions) and to assess its genericity, we chose the context of sale registries from the French Revolution, available thanks to a partnership with the *Archives Départementales des Yvelines*. Each registry page contains a table (figure 5a) where the first column represents a transaction number. This value is supposed unique and it is usually the successor of the previous one (except in rare circumstances, for instance *111* and *111-bis*). This constraint will be useful to prune hypotheses later.

Since these numbers are composed of handwritten digits, we can use a recognizer to obtain their values; however, noise and recognition errors must be taken into account. We need to keep not only the first, but the N best alternatives proposed by the recognizer. This list is sorted according to increasing scores (which represent



(a)



Hypotheses (**value**/score)			
395 0,462	**295** 0,464	**595** 0,499	**495** 0,532
296 0,000	**206** 0,983	**266** 0,983	**246** 0,983
2297 0,364	**5297** 0,393	**4297** 0,424	**8297** 0,439

(b)

Figure 5: A page from historical sale registries (a) and a close-up view of its sale numbers (b), with the localized handwritten digits and the list of hypotheses, in order of increasing score (lower is better), produced by the recognizer.

uncertainty), such as in figure 5b. We must also consider the possibility that no hypotheses are correct (for instance, the last line in figure 5b), which prevents us from pruning several inconsistent solutions.

Exponential growth If we consider, for each of the M lines in the document, all N hypotheses proposed by our recognizer, we will have M^N alternatives to explore. Even when applying successor constraints to each number, we still have too many possible analyses, mainly due to error tolerance (we always have at least one valid choice, a "default" value that is always a valid successor). Naively formulating a description grammar to analyze this kind of document will not work: if we apply our stochastic operator `FIND_BEST_FIRST` at the root of this grammar, the computation will not finish in a reasonable amount of time.

Managing complexity We implement local validation using a *sliding window* mechanism to provide some context, limiting the complexity to $M \times W^N$ hypotheses (where W is the window size). This mechanism limits lookahead and lookbehind to W successors before deciding on a particular number. In figure 5b, for instance, we can see that constraining the number in the second line to be greater than the number in the first line, we already obtain the correct values of 295 and 296 for these lines. By setting the window to a reasonable size (between 5 and 10 in our experiments), we avoid local errors while limiting backtrack. The processed images are quite damaged and this damage often affects a whole sequence of numbers; without a sufficiently wide window, we might obtain sequences of internally consistent but globally invalid numbers, such as 1295 - 1296 - 1297 - 298.

We validate one number per window and we do not need to backtrack after validation. Therefore, we can move the `FIND_BEST_FIRST` operator inside a window, avoiding redundant processing. Thanks to our operator's flexibility, this can easily be done (unlike traditional stochastic grammars, which perform extensive exploration throughout the whole tree) and the resulting grammar is fast enough for our needs: with a window size of 8 (to ensure a large safety margin), we need only 28 seconds to analyze 616 lines (about 60 pages) with 10 hypotheses per line.

Other optimizations are considered for future works, such as adapting A*-parsing and lazily exploring sub-optimal branches: since our score function increases monotonically, we never need to fully explore alternative solutions until the score of the current one increases or backtracking is needed. Also, memoization of parsed sub-trees can be applied to avoid recomputing them; parallel exploration of solutions is also possible. All these techniques require substantial modifications to the parsing mechanism and some might incur performance issues for existing deterministic grammars, which prevents them from being directly integrated in the existing mechanism.

5. CONCLUSION

Combining content recognition and structure analysis is a way to improve document recognition systems. Our proposal for seamless integration of symbolic information, obtained via syntactic analysis, and numeric information, acquired from recognizers (or any kind of user-defined metrics), offers a solution to Sayre's paradox: the global structure is captured by the document grammar, allowing *localization* to guide *recognition*, while local variation is captured by statistical recognizers and accepted thanks to a flexible (despite ambiguous) grammar, allowing *recognition* to guide *localization*. These strategies can be interleaved at will, generating a feedback loop that improves recognition rates.

Our proposal allows a grammar writer to use content-based information during structure analysis while preserving the existing deterministic mechanism, avoiding performance penalties outside the scope of the stochastic operator. We have presented the characteristics of a generic mechanism that can be applied to any syntactic layout analysis method, increasing document description expressiveness (via content-based rules, which allow simpler descriptions in some contexts) and using the grammar writer's knowledge to control the combinatorial explosion generated by stochastic parsing. The genericity of the score mechanism avoids (but does not exclude) normalization issues, notably for simple scoring functions where preference can be defined without a complex stochastic framework.

Our evaluation on the DMOS-P method showed an improvement in global recognition rates in the context of handwritten business letters: out of 1250 analyzed images, we obtained a recognition rate of 94,5%, corresponding

to a reduction of 25% of the error rate compared to the structural method. New document grammars can take in account the stochastic mechanism from the ground up and further benefit from the possibilities it offers. The evaluation on more than 60 pages (for a total of 616 lines) of historical documents shows one way to avoid combinatorial explosion of solutions, while assessing the genericity of the mechanism.

There are several directions to pursue development; for instance, to optimize execution time, we might direct the search of the best solution and avoid exploring branches with higher scores until they are effectively needed; to simplify the creation and adjustment of description grammars, we consider the automatic learning of normalization coefficients (via techniques borrowed from natural language stochastic parsing). This helps avoiding issues with heterogeneous trees and metrics, which currently must be manually defined. An inference mechanism would allow a grammar writer to input a series of annotated examples instead of manually defining scores, strengthening the convergence of statistical and syntactic methods.

REFERENCES

[1] Sayre, K., "Machine recognition of handwritten words: A project report," *Pattern Recognition* **5**, 213–228 (1973).

[2] Mao, S., Rosenfeld, A., and Kanungo, T., "Document structure analysis algorithms: a literature survey," in [*Document Recognition and Retrieval (DRR)*], *Proc. SPIE* **5010**, 197–207 (2003).

[3] Lemaitre, M., Grosicki, E., and Preteux, F., "Layout analysis of handwritten letters based on textural and spatial information and a 2D markovian approach," in [*Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*], *IEEE Proceedings* (2008).

[4] Montreuil, F., Grosicki, E., Heutte, L., and Nicolas, S., "Unconstrained handwritten document layout extraction using 2D conditional random fields," in [*Int. Conf. on Document Analysis and Recognition (ICDAR)*], *IEEE Proceedings*, 853–857 (2009).

[5] Artières, T., "Poorly structured handwritten documents segmentation using continuous probabilistic feature grammars," in [*Workshop on Document Layout Interpretation and its Applications*], (2003).

[6] Fitzgerald, J. A., Geiselbrechtinger, F., and Kechadi, T., "Structural analysis of handwritten mathematical expressions through fuzzy parsing," in [*Advances in Computer Science and Technology*], 151–156, ACTA Press (2006).

[7] Tateisi, Y. and Itoh, N., "Using stochastic syntactic analysis for extracting a logical structure from a document image," in [*Int. Conf. on Pattern Recognition (ICPR)*], **2**, 391–394 (1994).

[8] Mao, S. and Kanungo, T., "Stochastic language models for automatic acquisition of lexicons from printed bilingual dictionaries," in [*Workshop on Document Layout Interpretation and its Applications*], (2001).

[9] Coüasnon, B., "DMOS: a generic document recognition method: application to table structure analysis in a general and in a specific way," *Int. Journal on Document Analysis and Recognition* **8**, 111–122 (2006).

[10] Lemaitre, A., Camillerapp, J., and Coüasnon, B., "Interest of perceptive vision for document structure analysis," in [*Human Vision and Electronic Imaging XV*], *Proc. SPIE* **7527** (2010).

[11] Chanda, G. and Dellaer, F., "Grammatical methods in computer vision: an overview," tech. rep., Georgia Institute of Technology (2004).

[12] Nicolas, S., Paquet, T., and Heutte, L., "Un panorama des méthodes syntaxiques pour la segmentation d'images de documents manuscrits," in [*Conf. Int. Francophone sur l'Ecrit et le Document (CIFED)*], 237–242 (2004).

[13] Have, C. T., *Stochastic Definite Clause Grammars*, Master's thesis, University of Copenhagen (2008).

[14] Grosicki, E., Carre, M., Brodin, J.-M., and Geoffrois, E., "RIMES evaluation campaign for handwritten mail processing," *Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)* (2008).

[15] Grosicki, E., Carre, M., Brodin, J.-M., and Geoffrois, E., "Results of the RIMES evaluation campaign for handwritten mail processing," *Int. Conf. on Document Analysis and Recognition (ICDAR)* , 941–945 (2009).