MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

Grammatical Approach for the Physical and Logical Structure of Documents Analysis : Application to Summary Documents

S. TAYEB-BEY, S. SAIDI , H. EMPTOZ
Pattern Recognition & Vision Laboratory Bât 403 - INSA de Lyon
20 Avenue Albert Einstein 69621 Villeurbanne
email : tayebbey@rfv.insa-lyon.fr

Abstract

This paper deals with the use of grammatical formalism to recognize the physical and the logical structures of a composite document. We propose a new system for document recognition and analysis.

The aim of our research is to create a document structuring system by using a two level grammar.

A two level grammar constitute a high level formalism of expression and structuration in document analysis field.

keywords

document analysis, grammatical approach, physical structure, logical structure, two level grammars.

1 Introduction

A document has three principal characteristics :

❶ its content,

❷ its logical structure, i.e. its layout and the logical organization of its elements of information,

❸ its physical structure, i.e. the position of information elements on the pages.

A lot of work [1] has been done in document structuring, showing the growing interest in this field in recent years. As result, several methods have been proposed to solve the problem of structure representation.

Ingold [2] proposed a document description language similar to an attributed grammar. The aim in designing this language has been to top-down analysis several document classes.

Chenevoy [3,4] proposed a general-purpose system for document structure recognition called GRAPHIEN. The system organizes and controls the diverse document recognition processes.

However, the construction of structure models of documents has turned out to be a difficult task, and is often carried out manually because of document diversities.

Grammatical formalism is one of the approaches used in syntactic pattern recognition and it is quite suitable for document analysis. This formalism has also turned out to be a powerful tool in describing a document.

We are interested in structuring documents like summaries; the aim of our research is to create a document structuring system (see figure 2) by using a two level grammar [5]. Work on this system is currently under way.

2 Two Level Grammars

A two level grammar(see figure 1), also W-grammar [6] is a formal system well adapted to the language definition. It is composed of two grammars called metagrammar and hypergrammar. The metagrammar (type≥2) defines the possible domain of values for metavariables. These metavariables appear in the rules of the hypergrammar.

Generally, the metagrammar is a context-free grammar describing the language structures, and then the variables used.

The hypergrammar holds the definition of a contextual grammar, describing the specification of the translation operations expressed in semantic actions.

By applying the principle of uniform replacement in the hyperrules (similar metavariables are replaced by the same value), we obtain a ground instance of the hypergrammar called the protogrammar containing only context-free rules. Note that the protogrammar may potentially be infinite.

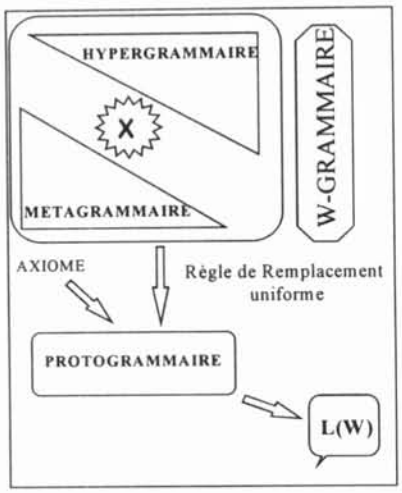

figure 1 : A two level grammar

3 Our System

We propose a new system for document recognition and analysis [7]. For the application part, this system is applied to summaries and produces the corresponding HTML (**H**yper**T**ext **M**arkup **L**anguage) pages.

The analysis which we carry out is done in two steps.

The first is the description and the recognition of a summary document. The second is the summary translation to the HTML language. However, the originality of the document description step lies in the use of the Two Level Grammar.

3.1 Description Step

A learning system concept: the purpose of this step is to infer a grammar of the physical structure called physical grammar , and then to infer a grammar of the logical structure called logical grammar. The result of this step is to construct a model of basic document structure.

Then, in the ***document recognition*** step : the system compares a specific document summary with the model of the learning system. If this document is not recognized as a known summary, the physical and logical grammar will be updated by adding new rules.
The system uses W-grammars in which the physical and logical grammars are given in the metagrammar. Then, the hypergrammar will describe the transformation of the physical and logical structures among other calculus. In our system, we use an operational version of W-grammar (called transparent W-grammar).

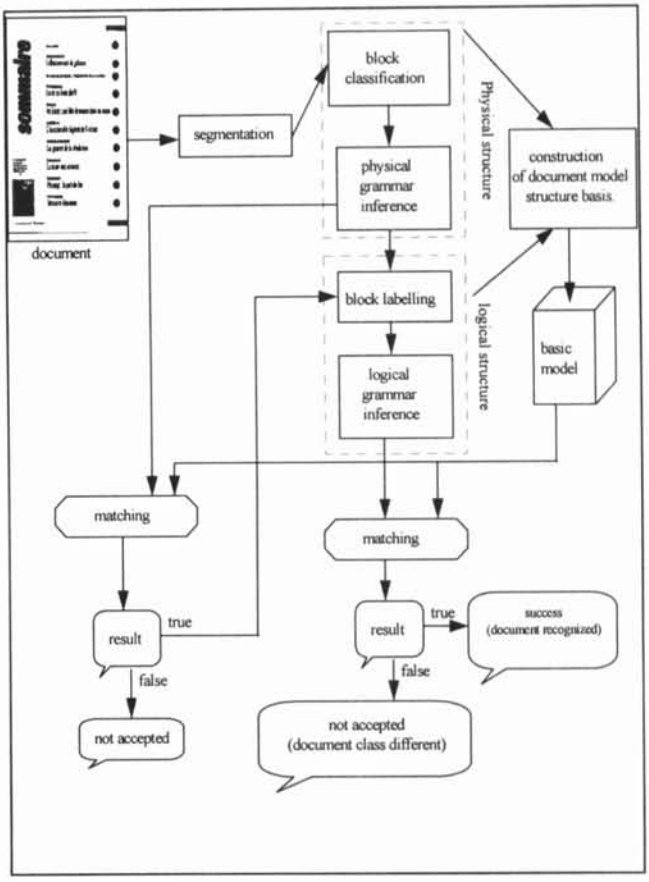

Figure 2 : System Architecture of describing the physical and logical structures.

The general algorithm which allows us to build the basic model is the following :

START
/* in this step, we build the basic model by taking some sample documents. These documents pages are digitilized and stored in an image file */

WHILE not end of file
DO
BEGIN
read one picture
display the picture
segment the picture
extraction of typographic and disposition parameters
classify the blocks (text, graphic, ...)
infer the physical grammar
/*this will represent the physical structure */
label the blocks logically
infer the logical grammar
/* this will represent the logical structure */
IF model does not exist

```
            THEN insert the model in the basic model
            ELSE update the basic model
        ENDIF
    END WHILE

END.
```

Basic model Algorithm

The second step consists to complete and to update the basic model. It can be represented by the following algorithm :

```
START
   /* we deal with documents to recognize, document
        pages are digitilized and stored in a file */

   WHILE not end of file
      DO
       BEGIN
        read the picture
        display the picture
        segment the picture
        extraction of typographic and disposition
       parameters
       classify the blocks (text, graphics, ...)
       infer the physical grammar
        /* this will represent the physical structure */
        IF not matching (physical learning structure and
       physical structure of document to analyze)
         THEN rejected
         ELSE accepted
       ENDIF
        label the blocks logically
         infer the logical grammar
       /* this will represent the logical structure */
   END WHILE

END.
```

Algorithm for document recognition

3.2 Translation step

The document is translated into HTML language by semantic operations (calculus). In our approach, we build a two level grammar where the logical and the physical grammars constitute the metagrammar.
The hypergrammar describes the transformation process of the terms generated by these grammars (the physical structure transformation to logical structure), and the calculus process to generate the HTML text.

4 Conclusion and Perspective

Two level grammars revealed to give many advantages in document structure analysis field.
In order to validate our approach, we apply it to documents like summaries.
The next step is to apply our experience to other documents such as scientific articles.

References

[1] G. Nagy. *A Prototype Document Image Analysis System for Technical Journals.* IEEE Computer Magazine. July 1992.

[2] R. Ingold. *A Document Description Language to Drive Document Analysis.* ICDAR 91. Vol 1. pp. 294-301.

[3] A. Belaid, J. J. Brault and Y. Chenevoy. *Knowledge-Based System for Structured Document Recognition.* In MVA'90 IAPR Workshop on Machine Vision Applications, November 1990.

[4] Y. Chenevoy. *Reconnaissance structurelle de documents imprimés : Etudes et Réalisations.* Ph.D. Thesis.. INRIA-Lorraine. December 1992.

[5] S. Saidi. *Extensions Grammaticales de la Programmation (en) Logique : Application à la Validation des Grammaires Affixes.* Ph.D. Thesis. Ecole Centrale de Lyon. 1992.

[6] A. Van Wijngaarden. *Orthogonal Design and Description of Formal Languages.* Mathematish Centrum Amsterdam, MR 76, 1965.

[7] S. Tayeb-bey. *Approche Grammaticale pour l'Analyse des Structures Physiques et Logique de Documents.* JED'96. Nantes-France. Juillet 1996.