Etiquetage fonctionnel des textes imprimés
Functional labeling of printed texts

Véronique EGLIN, Hubert EMPTOZ
Laboratoire de Reconnaissance de Formes et Vision RFV
INSA de Lyon
20, avenue Albert Einstein 69621 VILLEURBANNE CEDEX
Phone : (33) 04 72 43 60 54 Fax : (33) 04 72 43 80 97
E-mail : eglin@rfv.insa-lyon.fr

Résumé
Cet article présente une approche de l'étiquetage des données textuelles des documents imprimés, basée sur une analyse de texture. Nous abordons ici la caractérisation de la mise en forme typographique des polices et définissons des critères de complexité, de compacité et de relief structural des tracés des textes. L'étiquetage est lié à une recherche d'informations sur le document basée sur le constat que notre perception n'est pas aléatoire mais qu'elle est implicitement liée à la mise en forme matérielle des données. Nous proposons ainsi de référencer et de regrouper les différents types de textes selon leur aspect visuel et l'impression de *texture* qui s'en dégage. A partir de cette caractérisation et de la définition de trois grandes familles génériques et stables correspondant à trois types d'information sur le document (titre, paragraphe, note), nous proposons un étiquetage fonctionnel des blocs de texte. Ce travail s'inscrit dans un projet plus complet de segmentation et de reconnaissance de la structure logique de documents composites.

Abstract
This paper presents a new approach of textual data labeling based on texture analysis. The texture is used here to show the impact of the document making up on the visual exploration. We will show how textural properties are well adapted to typography characterization. In this context, we have defined complexity, compactness and structural relief criteria based on text drawing. The functional labeling is linked to the fact that the information search on a document is not random but directly linked to the document layout. We propose to reference and gather different types of text fonts according to their visual aspect and the visual impress which emerges from the textual data. This characterization allows us to define three kinds of generic families corresponding to three informative classes of texts : the class *title*, the class *paragraph of text (summary, body of the document ...)*, the class of *head or foot notes (or all little and specific and punctual information)*. On the base of this segmentation in three classes, we propose a functional labeling of text blocks. The blocks are obtained by a first structure analysis of the document, which will be quickly presented in this article.

1. Introduction

Les objectifs
La recherche d'information sur les documents faite par le lecteur humain n'est pas aléatoire. Elle dépend implicitement et pour une grande part d'un objectif de recherche ou d'une consigne qui aurait été donnée (recherche d'une information particulière...). Elle est par ailleurs étroitement liée à l'organisation des données sur le document que l'on appelle *mise en forme matérielle*. On peut noter que dans le domaine de l'analyse automatique des documents, les chercheurs s'intéressent de plus en plus aux informations de mise en forme traduisant une intention particulière de l'auteur et permettant ainsi de guider « intelligemment » le lecteur et de lui faciliter la tâche. Si cette mise en forme particulière des données est si importante à la lecture, elle l'est sans doute pour le système qui doit pouvoir sélectionner l'information pertinente sans avoir recours à une analyse linéaire de l'ensemble des données, mais qui, au contraire, doit pouvoir cibler rapidement la région du document dans laquelle l'information est attendue. Cette région peut alors avoir la *fonction* de titre, de sous-titre, de paragraphe de texte ou encore de note d'en-tête ou de pied de page. Pour trouver rapidement la *fonction* de ces données (ou leur nature) et en quelque sorte les régions d'intérêt du document, nous avons choisi de caractériser l'ensemble des données textuelles à partir de l'analyse du tracé des polices, de leur fréquence d'apparition, et de leur graisse. Et c'est une approche par la texture qui va nous permettre cette caractérisation.

En ce sens, le document est plus qu'une simple image de pixels que l'on pourrait traiter indépendamment du message que l'auteur a voulu faire passer au lecteur. Il faut ainsi pouvoir prendre en compte la présence de l'homme aux différents stades du cycle du vie du document (de sa