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

conception à sa lecture), où l'information de *fond* liée au message que l'auteur veut transmettre s'exprime indirectement par une mise en *forme* particulière des données. Certains auteurs, tels que Nagy, Doermann, Tang et Suen dans [7], [4], [11] ont proposé de formaliser ce contexte d'étude en retraduisant l'intervention de l'homme aux différents stades de vie du document.

Les approches liées à la capture d'information

C'est donc à travers la mise en page et les outils typographiques utilisés que le transfert de connaissances peut se réaliser dans les meilleures conditions pour le lecteur. Celui-ci dispose de plusieurs stratégies pour aborder le document : la *lecture complète* (ou lecture mot à mot), l'*inspection* (ou recherche précise dans une région du document) et enfin le *survol* (ou aperçu du document). Sur la figure 1, nous avons reproduit des résultats expérimentaux de mesures oculométriques obtenues sur des observateurs humains à qui a été donnés la consigne de ne s'intéresser qu'à l'information rapidement accessible sans aucun effort de lecture ni de compréhension. Cette tâche a première vue très simple met en jeu des mécanismes de repérage d'informations très complexes, à partir desquels nous avons retenus une hypothèse fondamentale. Cette hypothèse souligne le fait qu'il y a chez l'homme une émergence de l'information par zones d'intérêt, matérialisées sur les images de la figure 1 par des zones de fixations. Plus précisément encore, on s'aperçoit sur un grand nombre de résultats que les fixations sont localisées pour la plupart dans les zones d'images (grandes illustrations ou icônes), de titres et de sous-titres, c'est-à-dire dans les zones que le rédacteur du document aura particulièrement soignées et mises en évidence (taille des caractères, graisses, espaces inter-lignes, couleur...).

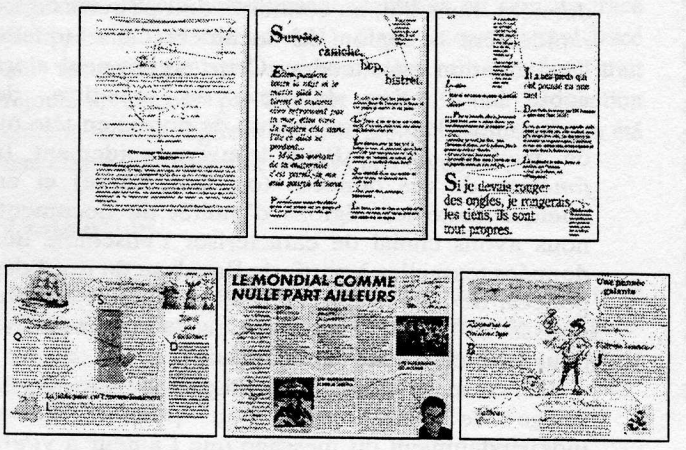


Figure1 Exemples de résultats de mesures oculométriques d'observateurs humains dans un contexte de survol du document.

Le lien très étroit qui existe entre le rédacteur d'un document (celui qui met en forme) et le lecteur est donc fondamental. Désormais, il ne s'agit plus de parcourir aveuglément le document en cherchant l'information pertinente, mais toute la problématique s'oriente à présent vers une recherche *guidée* par des régions présentant un *intérêt* visuel et sémantique.

2 Notre contribution

Actuellement, la plupart des méthodes de classification de fontes (et plus généralement d'analyse de la structure logique d'un document) font appel à des analyses locales du texte souvent basée sur les composantes connexes. Une approche moins courante, et qui nous a semblé intéressante, est de replacer la problématique de la classification des écritures imprimées dans le contexte plus général de caractérisation de *texture*. Nous pouvons considérer le *texte* comme une texture, dans la mesure où l'on définit le caractère comme l'entité élémentaire de texture. Plus précisément, une page de texte peut être considérée comme un ensemble de petits graphismes, les lettres, qui génèrent une impression « macroscopique » de texture. Les caractéristiques visuelles de cette texture dépendent de la disposition des lettres, de leur fréquence d'apparition, de la police, de la graisse, de la présence ou non d'italique, de la langue.

Dans ce qui suit, nous ne nous intéressons pas au sens contenu dans le texte, mais à sa forme, à l'impression visuelle que l'on en a. C'est d'ailleurs, nous semble-t-il, la première étape vers l'analyse du fond, c'est-à-dire de la signification, car les parties importantes sont souvent mises en valeur par les auteurs grâce à des mises en forme, [1]. Généralement, la texture est utilisée dans un contexte de segmentation. Nous l'utilisons ici sur les blocs de texte de style homogène dont les contours ont déjà été localisés. C'est dans ce contexte que nous proposons l'usage de paramètres statistiques, capables de mettre en évidence certaines caractéristiques de la police du texte observé. Ces paramètres nous ont permis de reconsidérer parmi la grande variabilité des styles d'écritures différents, des groupes (ou familles) possédant des caractéristiques génériques et stables. Nous avons considéré dans ces travaux trois grandes familles fonctionnelles : les *titres* (contenant toutes les informations de type titre ou sous-titres), les *paragraphes de textes* (contenant des données de type résumés ou corps de document) et enfin les *notes d'en-tête ou de pied de page*. Dans la suite de l'article, nous allons présenter l'étiquetage fonctionnel des blocs de textes à partir de l'interprétation de paramètre de texture liés à la complexité, la densité et le relief structural des polices.

2.1 Principe général

La première étape d'un processus d'étiquetage des textes consiste à segmenter l'image en régions homogènes (ou en blocs homogènes). Pour cela, nous proposons une approche de structuration physique des documents imprimés dans [5] qui aboutit à un découpage en blocs physiques. D'autres méthodes de segmentation conviendraient également. Citons à titre d'exemple les travaux de Boukined dans [2], de Derrien-Peden dans [3], de Nagy dans [7], de Pavlidis dans [8] et d'Ozaki dans [9]. A ce niveau, on ne s'intéresse qu'aux données textuelles, celles dont l'orientation privilégiée est horizontale. Le schéma de la figure 2 illustre le principe général de l'analyse des textes par la texture.

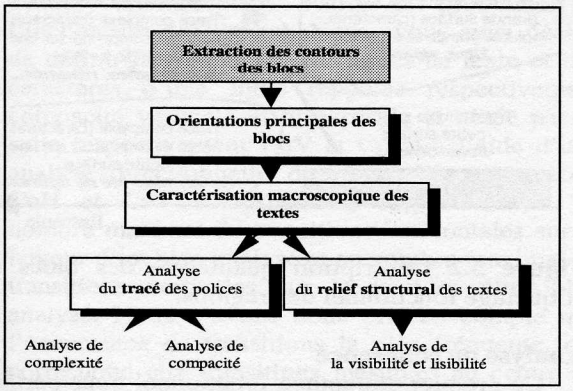


Figure 2 Schéma global des traitements pour l'étiquetage fonctionnel des blocs.

2.2 Directions privilégiées des blocs

La première étape de recherche des orientations privilégiées des blocs de textes est basée sur une analyse par autocorrélation. La fonction d'autocorrélation a souvent été utilisée pour la caractérisation de texture, [9]. En particulier, elle permet de mettre en évidence des périodicités dans une image et traduit de ce fait les orientations principales de certaines textures. Sur la fonction d'autocorrélation, les données relatives à une même direction seront situées sur une même droite, ayant aussi cette direction, et passant par l'origine. Sur la figure 3 nous présentons deux exemples de fonction d'autocorrélation pour deux types de polices. L'orientation préférentielle des blocs est finalement obtenue par le calcul des roses de directions résultant des fonctions d'autocorrélation pour chaque bloc de texte. On ne conserve que les blocs sur le document dont l'orientation préférentielle est horizontale. Dans cet exemple, le deuxième bloc de texte ne serait pas retenu : il possède deux directions privilégiées.

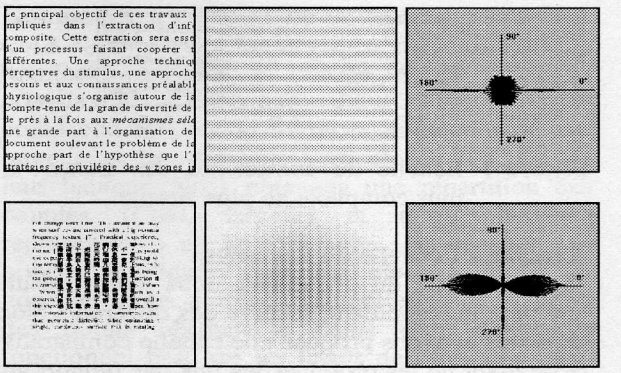


Figure 3 Successivement le bloc origine, sa fonction d'autocorrélation et la rose des directions résultante. Deux exemples sont représentés.

2.3 Primitives macroscopiques des textes

Dans les sections suivantes, nous allons présenter dans le détail les quatre attributs de texture qui sont à l'origine du classement des textes en trois familles. Pour chaque attribut, nous allons procéder selon une même logique : nous définissons d'abord à l'échelle du bloc de texte l'attribut que nous illustrons sur plusieurs exemples de test. Puis, nous appliquons le calcul à l'échelle du document pour obtenir une collection de mesures représentatives de chaque bloc analysé et caractéristiques d'une distribution relative des blocs les uns par rapport aux autres. Ces résultats sont ensuite interprétés par une échelle relative positionnant les blocs les uns par rapport aux autres. Une analyse de cohérence portant sur les labels déduits des résultats permet alors d'affiner le classement. Le label définitif résulte alors de la confrontation des résultats des quatre attributs et permet de donner une description qualitative très complète sur la nature du bloc analysé.

2.4 Analyse de la complexité des tracés

Analyse à l'échelle du bloc de texte

La signification que nous donnons au terme *complexité* est très proche du sens commun : le texte considéré ici comme une *courbe* sera analysé en fonction de son caractère « sinueux » et « tortueux ». On peut facilement comprendre qu'une droite aura une complexité très inférieure à une courbe organisée en une série de boucles telle qu'un mot manuscrit ou à un profil de montagnes. La complexité d'un texte correspond ainsi à l'impression visuelle laissée par le profil des courbes. La complexité peut alors être quantifiée par une mesure d'entropie effectuée sur la courbe (c'est-à-dire sur une portion de texte),[10]. On peut ainsi dire que plus une courbe est complexe, plus la valeur de l'entropie est élevée (en valeur absolue), voir figure 4.

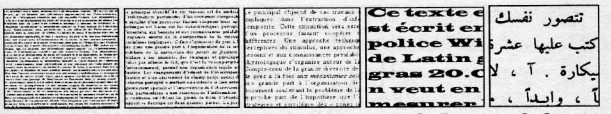


Figure 4 Echelle de complexité des tracés : du plus complexe au plus simple.

Analyse à l'échelle du document

A l'échelle du document, nous avons ensuite comparé la complexité relative des différents blocs de textes et avons proposé une échelle comparative permettant de les ordonner les uns par rapport aux autres. Il va donc falloir prendre en compte un critère supplémentaire : la surface des boîtes englobantes (c'est-à-dire la surface de chacune des portions de textes analysées). Nous avons donc décidé de construire une échelle de *complexité* mettant en relation les valeurs d'entropie de chaque bloc de texte et leur surface, voir Figure 5. Cette échelle traduit pour chaque bloc étudié le couple (Entropie, Surface). Plus l'entropie est faible et la surface importante, plus le bloc contient des caractères généralement gros par rapport aux autres blocs de la page. Il peut s'agir dans ce cas de blocs de titres ou de texte à fort relief structural. Inversement, plus l'entropie est grande et la surface petite, plus le bloc contient des données textuelles de type paragraphe ou résumé. La figure 5.1 résume sur un exemple le résultat de cette analyse.

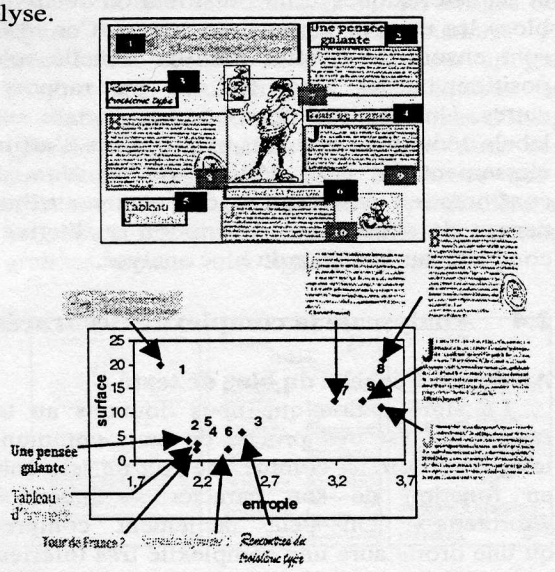


Figure 5.1 Distribution des points représentatifs des couples (Entropie, surface) pour chaque bloc de texte du document.

La figure 5.2 propose une première interprétation de ce graphique en donnant une description qualitative des différentes régions du graphe. Cette interprétation permet de définir pour chaque regroupements de points des *fonctions* (ou rôle) particulières. Ces fonctions correspondent aux familles que nous avons définies dans l'introduction de cette section. Il s'agit des titres ou sous-titres, des paragraphes ou résumés, et des notes d'en-tête ou de pied de page. Sur la figure 5.2, nous avons ainsi fait figurer cet étiquetage des blocs respectant les positions des points représentatifs des couples (Entropie, Surface). La séparation en quatre régions *fonctionnelles* a été obtenue par la localisation de l'isobarycentre des points du graphe (voir figure 5.1) Sur cet exemple, Sur cet exemple, les blocs numérotés 7-8-9-10 ont été étiquetés *paragraphe de texte*, les blocs 2-3-4-5-6 ont été étiquetés *sous-titres* et le bloc 1 a été étiqueté *titre*

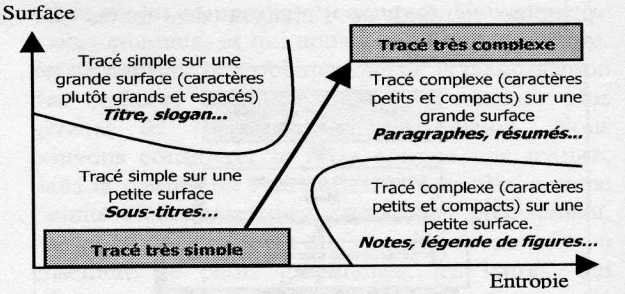


Figure 5.2 Description qualitative des blocs et étiquetage fonctionnel des régions.

Analyse de cohérence

Ce premier étiquetage fonctionnel nous permet de faire une distinction cohérente entre des éléments visuellement très différents (tels que les titres et les paragraphes de texte par exemple). Cependant, cette approche globale par la texture ne permet pas toujours de lever l'ambiguïté qui peut subsister entre les labels *sous-titre* et *note d'en-tête ou de pied de page*, ceux-ci pouvant être confondu dans un bon nombre de cas, notamment parce que leur entropie est généralement faible et leur surface est petite. L'approche que nous préconisons alors consiste à soumettre les blocs ambiguës (les blocs de label *sous-titres* et de label *notes*) à une analyse de cohérence de leur position sur la page du document. Plus précisément, les blocs étiquetés *sous-titres* conserveront leur étiquette s'ils ne sont pas situés sur l'une des extrémités (inférieure ou supérieure) du document. Inversement, les blocs étiquetés *notes* ne pourront pas occupés de position intermédiaire, à certains endroits présumés des sous titres, sauf s'ils sont situés à proximité d'un bloc image, auquel cas ils peuvent être étiquetés *notes* (ou légende de figure). Les notes sont généralement isolées à des emplacements réservés aux bords extrêmes du document.

Conclusion

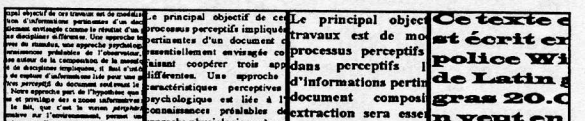
Dans les parties suivantes, nous allons présenter les attributs complémentaires renseignant

sur la densité et la visibilité des polices. Ces attributs sont d'abord considérés individuellement : leur combinaison nous permettra d'obtenir a étiquetage plus précis et également plus réaliste, voir section 2.7. Ainsi, avec quatre paramètres et trois familles de typographies différentes, nous obtenons dans des proportions très satisfaisantes présentées à la section 2.7 un label unique pour le bloc analysé.

2.5 Analyse de la compacité directionnelle des textes

Compacité à l'échelle des blocs indépendants

La structure générale d'un texte est essentiellement caractérisée par ses deux directions préférentielles : horizontale et verticale (nous avons volontairement écarté le problème de l'inclinaison des lignes). Nous avons ainsi décidé de définir la *compacité* des lignes de texte et des caractères d'une ligne (appelée respectivement compacité verticale et horizontale et notée par la suite respectivement CoV et CoH) à l'aide d'une analyse directionnelle du texte. Les paramètres CoH et CoV correspondent respectivement au nombre maximal de transitions horizontales sur la largeur d'un bloc analysé et au nombre maximal de transitions verticales sur la hauteur d'un bloc analysé. Nous n'avons donc pris en compte que l'occurrence de transitions la plus fréquente, qui correspond aux transitions mesurées au centre de chacune des lignes de texte. Ceci nous permet ainsi d'avoir une estimation assez réaliste du nombre de caractères par ligne de texte. Ce nombre est proportionnel, à un facteur près, à la compacité horizontale du texte. Symétriquement, les résultats de compacité verticale nous renseigne sur le nombre approximatif de lignes de textes par bloc. Une estimation statistique portant sur un ensemble de texte latin (de type de ceux que nous analysons) nous a permis d'évaluer le nombre moyen de transitions horizontales par caractère et le nombre moyen de transitions verticales par ligne de texte, voir Figure 6.



1. 2. 3. 4.

Texte #	Transitions *horizontales*	Nbre de car/ligne	Transitions *verticales*	Nbre de lignes/bloc
1	51	33.5	24	15.7
2	38	25	15	9.8
3	28	18.4	12	7.8
4	12	7.8	10	6.5

Figure 6 Estimation de la densité horizontale et verticale de texte de polices différentes

Compacité à l'échelle du document

A l'échelle du document, nous procédons de manière similaire à l'étude menée pour la complexité. Cette fois, nous exprimons la compacité horizontale des blocs en fonction de leur largeur, et la compacité verticale en fonction de leur hauteur. Nous obtenons une répartition des couples (CoH, largeur) et (CoV, hauteur) représentative de la compacité (ou densité) des polices impliqués dans les blocs analysés. A nouveau, nous pouvons interpréter la distribution des points résultats à l'aide d'un étiquetage fonctionnel basé sur les labels de titre, sous-titre, paragraphe (ou résumé) et note d'en-tête ou de pied de page, voir figure 7. L'exemple traité est le même que celui de la figure 5. La numérotation des blocs est identique.

Sur cet exemple, les blocs numérotés 7-8-9-10 ont été étiquetés *paragraphe de texte* pour les deux attributs, les blocs 2-3-4-5-6 ont été étiquetés *titres* pour les deux compacités ; le bloc 1 a été étiqueté plus spécifiquement *titre/accroche* pour la compacité horizontale. Les résultats sont cohérents avec l'analyse de complexité.

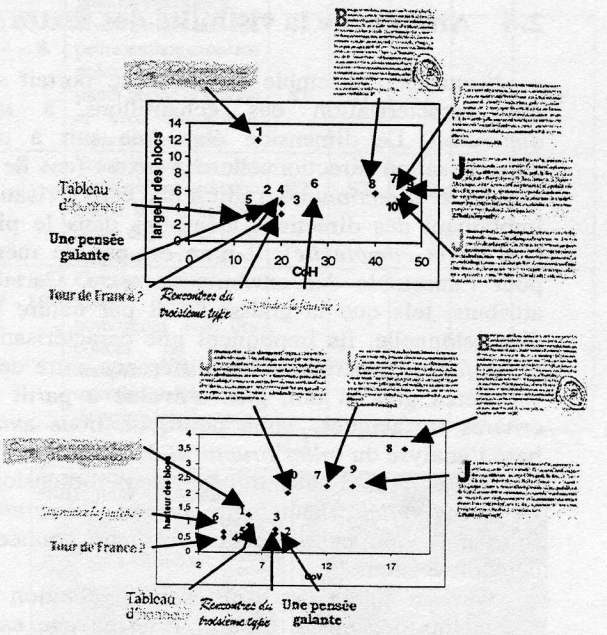
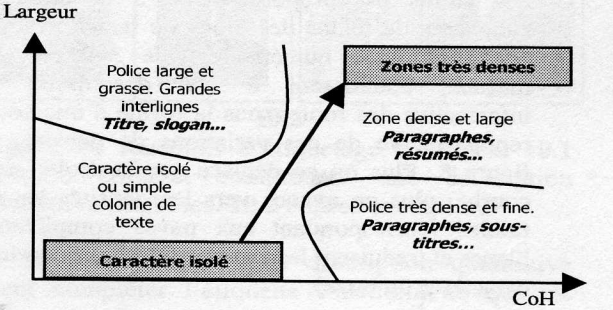


Figure 7.1 Distribution des points représentatifs des couples (CoH, largeur) et (CoV, hauteur) pour chaque bloc numéroté du document



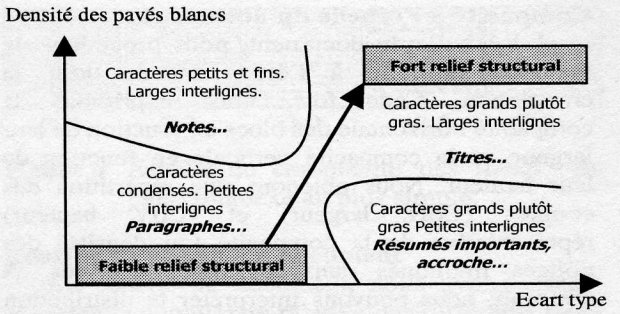


Figure 7.2. Description qualitative et étiquetage fonctionnel des régions, respectivement pour la compacité horizontale et compacité verticale.

Analyse de cohérence

Dans le cas de la compacité, l'analyse de cohérence concerne les mêmes points que pour la complexité. Nous verrons à la figure 10 comment les ambiguïtés d'étiquetage ont ainsi pu être levées. Voyons maintenant le dernier attribut relatif à la graisse des polices : la visibilité.

2.6 Analyse de la visibilité des textes

Jusqu'ici, l'ensemble de l'analyse portait sur une caractérisation des échantillons à *une dimension*. La dimension était liée soit à une caractérisation directionnelle des textes (cas de la *compacité directionnelle*) soit à une caractérisation impliquant des directions aléatoires dans le plan (cas de la *complexité*). Il n'en est pas de même pour l'ensemble des attributs de texte. Certains attributs, tels que la graisse, sont par nature bi-dimensionnelle, ils impliquent une caractérisation de *surface*. En particulier, la différence entre deux caractères gras et non gras s'évalue à partir de critères surfaciques. Voilà pourquoi, nous avons basé l'analyse du *relief structural* des formes sur la prise en compte d'échantillons à deux dimensions. Le terme *relief* traduit cette notion d'*empreinte* scripturale qui, par définition, est une donnée à deux dimensions.

Notre méthode est basée sur l'évaluation de paramètres de densité à partir d'un pavage carré des blocs de texte. Le principe consiste à calculer le niveau de gris moyen des pavés carrés de taille constante recouvrant la surface des blocs. La taille des carrés est proportionnelle à la taille des caractères de textes des blocs de textes avec une résolution à la numérisation de 300 dpi. Ces mesures aboutissent à une description très informative des fontes sous la forme d'une courbe représentative de ces variations de densité, voir figure 8. Plus on se déplace sur la droite sur la courbe, plus on avance vers les densités les plus faibles (correspondant aux pavés complètement blancs et traduisant les espaces interlignes ou inter-caractères du texte).

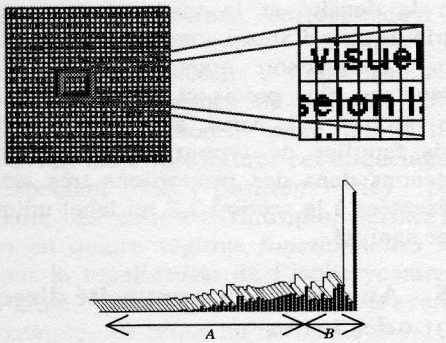
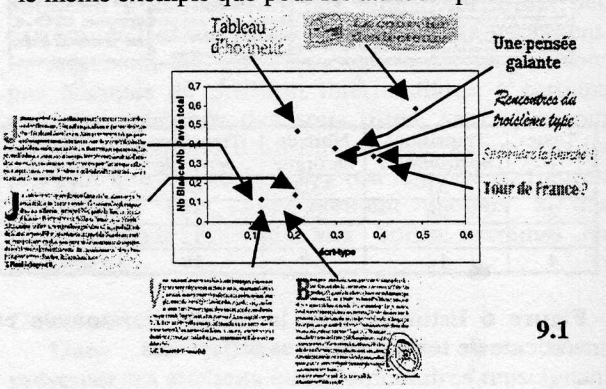
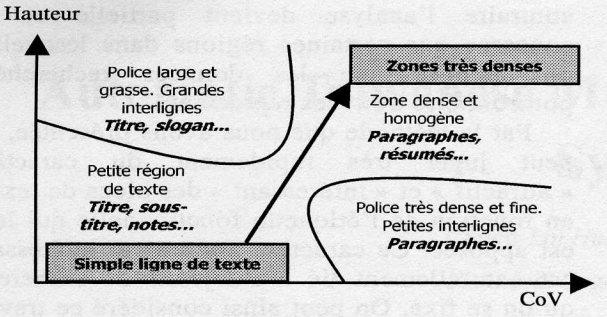


Figure 8 Représentation du pavage carré à partir duquel a été calculée la courbe représentative des variations de densités.

Une manière simple de procéder à l'interprétation de cette courbe consiste à construire une échelle de relief structural des polices à partir du calcul d'un écart-type sur les valeurs de la courbe. Cette valeur figurera en abscisse de l'échelle. La valeur d'écart-type est calculée sur les pavés carrés contenant moins de 5% de pixels blancs (notés *A* à la figure 8). Elle renseigne sur la répartition des densités pour les pavés carrés. Plus l'écart-type est faible, moins les variations de densité sont importantes, traduisant ainsi l'aspect *peu* contrasté de l'image dans la direction relative au pavé considéré. Inversement, plus l'écart-type est important, plus il témoigne d'une forte variation de densité, ce qui se traduit par une *relative importance* de la taille et de la graisse de la police considérée, donc, en somme, par un fort *relief structural*. Comme ordonnée de l'espace de représentation, nous avons considéré un rapport de densité donnant la proportion entre le nombre de pavés carrés contenant plus de 95% de pixels blancs, ils sont notés *B* à la figure 8 (les pavés contenant un, deux ou trois pixels noirs sont considérés comme totalement blancs afin de limiter l'influence d'un bruit impulsionnel), et le nombre total de pavés nécessaires au recouvrement de l'image. Ces définitions nous garantissent ainsi l'indépendance entre les valeurs en abscisse et en ordonnée. La Figure 9 représente les résultats sur le même exemple que pour les attributs précédents.



9.1



9.2

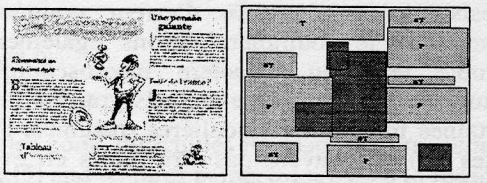
Figure 9 9.1 Distribution des points représentatifs des couples (écart-type, nb Pavés blancs/nb pavés total). **9.2** Description qualitative et étiquetage fonctionnel des régions.

A nouveau, nous pouvons interpréter la distribution des points résultats à l'aide d'un étiquetage fonctionnel, voir figure 9. Sur cet exemple, les blocs numérotés 7-8-9-10 ont été étiquetés *paragraphe de texte*, les blocs 2-3-4-5-6 ont été étiquetés *résumés importants ou accroches* ; enfin, le bloc 1 a été étiqueté *titre*. C'est lui qui possède le plus fort relief structural

2.7 Bilan et étiquetage définitif

Dans cette étude, nous aurions pu utiliser des paramètres statistiques bien plus nombreux (issus par exemple d'une analyse de cooccurrence en niveau de gris). La principale difficulté qui ressortait alors d'une telle approche était de l'ordre de l'interprétation « visuelle » des résultats. Il était alors difficile d'attribuer un degré de *visibilité* ou de *lisibilité* aux polices analysées, ce qui a été rendu possible avec l'analyse que nous venons de proposer. Les paramètres que nous avons retenus ont ainsi l'avantage d'être facilement exploitables car ils correspondent à une réalité de la perception dans un contexte de survol. Ils ont été choisis parce qu'ils sont le reflet du relief structural des textes et permettent ainsi un classement rapide et discriminant. En outre, ces paramètres sont peu nombreux, peu redondants car ils traduisent un aspect à la fois différent et complémentaire des fontes (la complexité, la lisibilité et la densité). Comme nous l'avons présenté, nous pouvons prendre individuellement chaque résultat pour proposer un étiquetage fonctionnel rapide des blocs. Qu'il s'agisse de la lisibilité, de la complexité ou de la compacité, chaque attribut peut être pris comme référence. L'étiquetage définitif va finalement refléter la *complémentarité* des attributs. Ainsi, l'étiquetage de chaque bloc est fonction des quatre paramètres calculés précédemment.

Dans le cadre de cette étude, nous ne pouvons nous trouver que dans une situation où nous avons le choix entre trois labels élémentaires (correspondant aux trois grandes familles définies dans l'introduction de la section2). Ces trois familles sont choisies pour caractériser les résultats de quatre mesures : l'entropie, les compacités et la lisibilité. Chacun de ces paramètres peut conclure à un label différent, mais la combinaison des quatre conclut toujours à un label unique majoritaire. L'étiquetage est donc toujours unique. La figure 10 présente quelques résultats. Pour les notations nous avons utilisé *N* pour note, *P* pour paragraphe, *ST* pour sous-titre, *T* pour titre.



Bloc #	Description qualitative	Etiquette				final
		E	CH	CV	L	
1	Petite région de police large et grasse à grands caractères et grandes interlignes.	T	T	T	T	T
2 – 3 - 4 - 6	Tracé simple sur une petite surface avec une police de caractères plutôt grasse et grande et avec de grands espaces interlignes.	ST	ST	ST	T	ST
5	Tracé simple sur une petite surface avec une police de caractères dense et petite et avec de grandes interlignes.	ST	ST	ST	N	ST
7 - 8 - 9 - 10	Grandes zones de texte denses et homogènes avec des polices de tracé complexe, de petits caractères compacts et de petites interlignes.	P	P	P	P	P

10.1

10.2

Figure 10 Résultats d'étiquetage de blocs. **10.1** Résultat complet comprenant une description qualitative des blocs. **10.2** Autres exemples.

Sur la figure 10.1, la description qualitative vient compléter l'étiquette fonctionnelle qui est

proposée. Dans les exemples que nous avons traités ici les analyses de cohérence successives ont conduit à un étiquetage final réaliste. Cela n'est pas toujours le cas : les erreurs d'étiquetage qui pourraient subsister pourraient provenir de mises en forme de document inattendues (un titre en bas de page avec un faible relief structural, des notes intermédiaires au centre de la page...). La base d'exemples que nous avons testée est constituée d'environ soixante dix documents. Les résultats d'étiquetage des blocs que nous obtenons sont cohérents avec leur fonction dans 98% des cas. Il subsiste néanmoins 2% d'erreur comme nous venons de le souligner. Notons également que compte tenu du caractère macroscopique de l'étude, et du découpage en trois grandes familles, certaines fonctions particulières (résumés, longues notes ...) ne peuvent pas être mises en évidence par des étiquettes spécifiques (autres que celles que nous proposons dans cette étude). A ce stade, c'est la description qualitative qui peut apporter une plus grande précision sur la nature du bloc. Ces informations supplémentaires doivent alors permettre dans un grand nombre de cas de faire une véritable distinction entre par exemple un résumé (de police plus petite ou plus grasse), et un paragraphe de texte au cœur du document. Des connaissances supplémentaires sur le type de document et sur sa charte éditoriale sont alors requises. Nous n'avons pas aborder cette problématique dans cet article.

3. Conclusion et perspectives

Ce travail s'inscrit dans un processus complet de structuration automatique des documents imprimés où l'information est recherchée en fonction de son pouvoir attractif. Dans le contexte dans lequel ce travail a été élaboré, nous avions choisi de simuler ce que nous pourrions appeler un survol du document correspondant à une recherche ne nécessitant pas de lecture complète des données et ne présupposant aucune connaissance a priori sur ce qui doit être trouvé. Ainsi, l'ordre dans lequel les informations sont sélectionnées n'est pas aléatoire. En particulier, les données de type *titre* sont sélectionnées avant les données de type *paragraphe de texte*, elles mêmes avant les blocs de type *notes*. Avec un objectif de recherche différent, on peut privilégier une information d'une autre nature et envisager les traitements adéquats. Ce type de travail est un premier pas vers des applications d'indexation de documents où l'information doit être traitée rapidement et sélectionnée en fonction des objectifs de l'utilisateur. Ceci entraîne notamment que toute l'information du document n'est plus nécessairement traitée uniformément, mais au contraire l'analyse devient partielle et ne concerne que certaines régions dans lesquelles on suppose que les données recherchées pourront être trouvées rapidement.

Par la méthode que nous avons présentée, on peut juger très rapidement du caractère « attractif » et « intéressant » des blocs de textes en fonction de l'étiquette fonctionnelle qui leur est apposée. Ce caractère attractif et intéressant est naturellement lié à l'objectif de recherche qu'on se fixe. On peut ainsi considéré ce travail comme une première étape d'identification des données textuelles des documents à partir de laquelle on peut effectuer des analyses plus fines du contenu (analyse sémantique et lecture complète) permettant la mise en place d'une structuration logique du document. Concernant directement cette méthode, il est important de signaler qu'elle ne présuppose pas de connaissances a priori sur les caractéristiques du texte analysé. Il s'agit d'ailleurs d'une perspective en cours d'exploitation.

4. Références

[1] **BARBARA, M.O., MOJAHID, M., VIVIER, J.** Mise en forme matérielle des textes de consignes et repérage d'informations. *CNED*, Nantes, 1996, pp.229-236.

[2] **BOUKINED, L. , TACONET, B., ZAHOM, A., FAURE, A.** Recherche de la structure physique d'un document imprimé par rectangulation., *RFIA'91,* Villeurbanne, 1991, vol.3, pp.1027-1032.

[3] **DERRIEN-PEDEN, D.** Analyse de la structure macro-typographique d'articles scientifiques. *ICDAR*, Saint-Malo, 1991, pp.311-319.

[4] **DOERMANN, D., ROSENFELD, A., RIVLIN, E.** The function of documents, *ICDAR*, Ulm, 1997, vol.2, pp.1077-1081.

[5] **EGLIN, V., EMPTOZ, H.** Logarithmic Spiral Grid and Gaze Control for the Development of Strategies of Visual Segmentation on a Document. *ICDAR*, Ulm, 1997, pp.689-692.

[6] **EGLIN, V., BRES, S., EMPTOZ, H.** Caractérisation de la mise en forme du texte imprimé par une analyse multirésolution de texture. *CIFED,* Québec, 1998, pp.473-482.

[7] **NAGY, G., VISWANATHAN, M.** Dual Representation of segmented Technical Documents. *ICPR'90,* Atlantic City, 1990, pp.141-151.

[8] **PAVLIDIS, T. , ZHOU, J.** Page segmentation by white streams. *Proceedings of First Int. Conference on Document Analysis and Recognition*, Saint-Malo, 1991, vol.2, pp. 945-953.

[9] **OZAKI, M.** Column Segmentation by white space pattern matching. *ICDAR*, Tokyo, 1995, vol.1, pp.134-138.

[10] **MENDES-FRANCE, M.** Dimension et entropie des courbes régulières. *In "Dimensions non entières et applications"*, G. Cherbit, Paris : Masson, 1987, pp. 329-339.

[11] **TANG, Y.Y., SUEN, C.Y.** Document structures : a survey. *Proceedings of Second Int. Conference on Document Analysis and Recognition*, Montréal (Canada), 1993, vol.1, pp99-102.