Analyse des documents graphiques pour la reconstruction d'objets : une contribution

15 juin 2007

Prologue

Les travaux présentés dans cette Thèse concernent l'analyse des documents, et plus particulièrement l'analyse des documents graphiques. La discipline de l'analyse des documents est historiquement issue de l'analyse d'image de document. Les premières applications d'analyse d'image de document sont apparues dans les années 1960. Elles avaient pour objectif la reconnaissance optique de caractères imprimés pour le traitement en masse de documents (tri postal, traitement de bulletin de salaire, ...). Elles se sont étendues dans les années 1980 à la problématique de l'interprétation d'image de document pour leur rétroconversion. La rétroconversion a pour but la conversion complète d'un document papier en un document électronique compréhensible par l'ordinateur. L'objectif affiché était alors de contribuer à la ré-exploitation des masses de documents papier par les systèmes informatiques. Dans les années 1990, la prolifération des documents électroniques a induit un changement radical dans les activités liées au document (production, publication, diffusion, échange, ...). En effet, durant cette période l'informatique a connu un essor important avec la croissance du parc informatique mondial et des réseaux (locaux et Internet). Cet essor a engendré une invasion massive des documents électroniques. De nouvelles disciplines scientifiques, comme le web sémantique et le web mining, ont alors émergé autour des problématiques de modélisation et d'indexation des bases de documents électroniques. Aujourd'hui, les disciplines de l'analyse d'image de document et celles liées aux documents électroniques tendent à converger. Dans ce contexte, la notion de document (du papier jusqu'au document web) et d'analyse (du traitement d'image vers l'interprétation, l'indexation, et le web mining) sont devenues complexes à identifier. Afin d'éclaircir ces notions, notre chapitre d'introduction (De l'analyse des documents) présente diverses considérations.

La section (Document : histoire et définition) expose quelques considérations de base sur l'historique et la définition d'un document. Par la suite, la section (De l'analyse d'image de document) présente la discipline de l'analyse d'image de document. Cette section présente tout d'abord les différentes thématiques liées à cette discipline (manuscrite, structurée, et graphique), puis les composantes principales des systèmes automatiques d'analyse d'image de document (traitement, contrôle, et connaissances).

La section (Des applications de tri au web mining : un historique) brosse par la suite un historique de l'analyse d'image de document en lien avec le développement du document électronique. Cet historique est présenté progressivement : les applications de tri et de rétroconversion d'image de document, la prolifération des documents électroniques, l'évolution des formats de document et de leurs sémantiques, et enfin le web sémantique et le web mining. La section (Vers l'analyse des documents) présente alors la discipline de l'analyse des documents (ce pluriel est volontaire) comme le rapprochement de la discipline de l'analyse d'image de document et des disciplines liées aux documents électroniques. Différents aspects sont abordés comme les problématiques de recherche, les enjeux techniques et économiques, ... La section (De l'analyse des documents graphiques) présente enfin la problématique spécifique abordée dans cette Thèse.

De l'analyse des documents

Document : histoire et définition

L'origine étymologique ("docere", instruire), du terme "document" signifie "objet porteur de sens" ou encore "support de transmission de connaissances". Celui-ci apparaît dès 9000 ans avant notre ère sous la forme de pierres gravées de différents motifs [Masson 93]. Il a évolué au cours de l'histoire sous différentes formes [Blasselle 97] (tablettes, papyrus, livres, ...) et il a connu deux principales révolutions : l'imprimerie en 1477, et l'informatique (ordinateurs et réseaux) à partir de 1950.

L'imprimerie a démocratisé le document en favorisant sa diffusion. L'informatique en a assuré une diffusion encore plus large, mais elle a surtout transformé la nature du document à tel point que [Roisin 99] indique : "À l'ère des hypermédias, la notion de document devient difficile à identifier[1]". La liste suivante présente quelques caractéristiques extraites de [Labiche 98], [Roisin 99], [Parent 99], [Pédauque 03], et [Breuel 04] définissant la notion de document :

- Son support physique et son format : Un document est défini par son support physique et le format de ce support [Roisin 99] (papier "affiche, lettre", bande magnétique "8mm, 35 mm", fichier informatique "pdf, texte", .).
- Son scénario temporel : Un document est un enchaînement temporel d'objets [Roisin 99] (présentation "transparents", livre "pages", vidéo "images", .).
- Son sens : Le document est avant tout défini comme objet porteur de sens, indissociable du sujet en contexte qui reconstruit ce sens [Pédauque 03].
- Son cycle de vie : Un document évolue dans le temps et répond à un "cycle de vie" [Labiche 98] au cours duquel, il se modifie (correction, ajout, suppression, .), et change de support physique et de format (impression, conversion, scan,.).
- Son contenu et sa structuration : Un document est une concaténation d'objets (image "pixels", mot "caractères", graphique "points", vidéo "images", .) répondant [Roisin 99] à une structure logique (titre, résumé, sections, paragraphes, .) et une structure physique (est inclus, voisin de, .).
- Ses traitements : Un document est défini par des objets et les traitements applicables à ces objets [Roisin 99] (indexation, remplacement, suppression, .).
- Ses méta-données : Un document est défini par des méta-données [Parent 99] qui

[1] Des travaux de définition sont en cours par la communauté française RTP-DOC [Pédauque 03].

décrivent certains aspects de son contenu (type de document, date, auteurs, mot-clés, thématiques, liens sur d'autres documents, .).
- Sa sémantique et son interprétation : Un document est défini par des données appartenant à différentes sémantiques[2] [Breuel 04] selon leurs degrés d'interprétation (images, données textuelles, données vectorielles, .).

À travers cette liste, on peut identifier les principales évolutions du document à l'ère des hypermédias. Tout d'abord la notion de document s'est étendue des documents papier[3] et audio ou vidéo, aux documents multimédias les combinant tous [Roisin 99]. Ensuite, le document est également devenu interprété : les objets (et les relations entre ces objets) qu'il véhicule sont de sémantiques hétérogènes selon leur degré d'interprétation par un ordinateur [Labiche 98]. Enfin, le document est caractérisé par un ensemble de méta-données décrivant son contenu [Parent 99].

De l'analyse d'image de document

Introduction

L'analyse d'image de document concerne le domaine du traitement d'image de document par ordinateur. [Kasturi 02] la présente de la manière suivante : "l'analyse d'image de document correspond aux algorithmes et aux techniques à appliquer aux images des documents pour obtenir une description interprétable par un ordinateur à partir des données pixels"[4]. Elle s'intéresse principalement aux documents papier. Le traitement des documents audio et vidéo concerne le domaine de la vision [Jolion 01] et du traitement des signaux mono-dimensionnels [Kunt 91].

Les premières applications d'analyse d'image de document sont apparues dans les années 1960 [Mori 92]. Elles visaient à la reconnaissance optique de caractères imprimés (OCR)[5]. Ces applications se sont étendues dans les années 1980 [Nagy 00a] à divers aspects : pré-traitement d'images, extraction de tableaux, reconnaissance de signature, ... Aujourd'hui, on distingue trois thématiques principales d'analyse d'image de document [Ceheux 02] : la reconnaissance de l'écriture manuscrite [Vinciarelli 02], l'interprétation de documents structurés [Tang 96] et l'interprétation de documents graphiques [Ablameyko 00]. La figure (1) donne des exemples de documents associés à ces thématiques. Évidemment, ces trois thématiques partagent des axes de recherche communs [Kasturi 02]. Elles sont cependant différentes sur un certain nombre d'aspects [Ceheux 02] résumés dans le tableau (1). Nous détaillons ces différences dans les paragraphes suivants.

[2] Employé dans ce manuscrit comme (par abus de langage) niveau(x) ou contenu(s) de sens.
[3] lettres, livres, journaux, documents administratifs et techniques, photographies, ...
[4] "Document image analysis (DIA) refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data".
[5] Optical Character Recognition

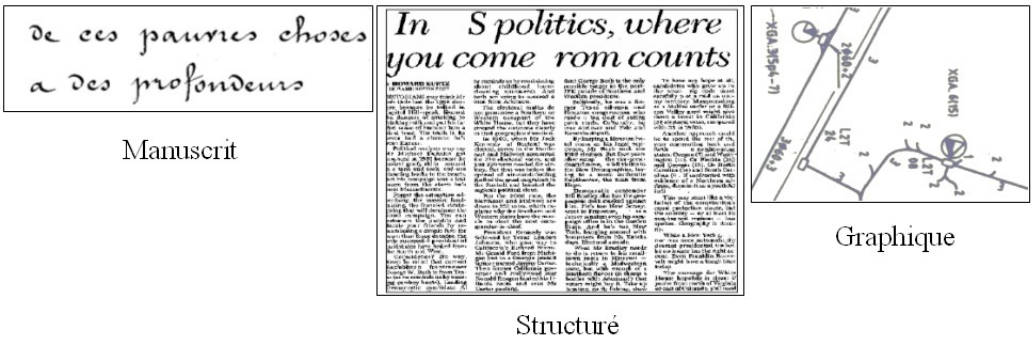


FIG. 1 – Documents manuscrit, structuré, et graphique

	Caractéristiques	**Difficultés**
Manuscrit	segmentation en mots	variabilité des écritures
Structuré	caractères imprimés	structure complexe
Graphique	nomenclature	multi-orienté, inter-connecté

TAB. 1 – Caractéristiques et difficultés des thématiques d'analyse d'image de document

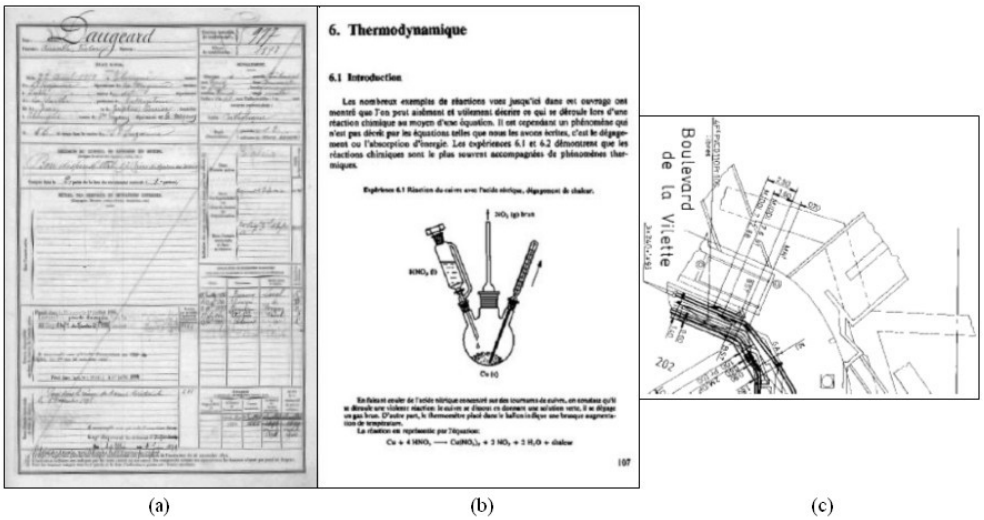
On distingue deux catégories de documents manuscrits : les documents libres (lettres, notes, ...) par opposition aux documents contraints (chèques, formulaires, ...) dont la présentation suit des règles définies. Les mots manuscrits sont en partie segmentés sur les images de document ce qui facilite la tâche de leur localisation. Cependant les caractères restent eux inter-connectés dans les mots [Vinciarelli 02] et présentent de fortes variabilités d'une écriture à une autre [Bensefia 04]. C'est ce dernier aspect qui rend la tâche de reconnaissance de l'écriture complexe. Dans ce contexte, les recherches sur cette thématique ont alors essentiellement abouti en ce qui concerne les documents manuscrits contraints [Bunke 03]. En effet, les applications peuvent alors s'appuyer sur des connaissances a priori fonction du type de document permettant de simplifier le problème de reconnaissance de l'écriture. La reconnaissance automatique de documents manuscrits libres reste un problème de recherche ouvert [Heutte 03].

Les documents structurés englobent diverses catégories de document [Trupin 03] : livres, journaux, formulaires, ... Les applications d'interprétation de documents structurés sont peu confrontées aux problèmes de variabilité et de connexion des caractères rencontrés dans les documents manuscrits [Tang 96]. En effet, ces documents sont composés de caractères imprimés, ce qui en permet une reconnaissance plus simple [Nagy 00b]. Le problème concerne plutôt le traitement de la structure des documents [Mao 03]. En effet si la reconnaissance de la structure physique[6] d'un document semble aujourd'hui un problème maîtrisé, celle de la structure logique[7] constitue une perspective de recherche importante [Mao 03].

[6] Topologie des objets du document : positions, relations de voisinage, ...
[7] Organisation logique des objets du document : titre, section, paragraphe, ...

Les documents graphiques englobent également diverses catégories de documents : cartes, schémas électriques, plans de réseaux, ... Ces documents sont conçus en fonction de chartes graphiques propres à chaque catégorie de document [Ponte 97]. Les applications d'interprétation de documents graphiques [Ablameyko 00] peuvent donc s'appuyer sur ces chartes a priori connues, de façon à guider le processus d'interprétation. Cependant les objets graphiques sont souvent multi-orientés et inter-connectés, ce qui complique fortement leur traitement. Une évaluation récente des systèmes commerciaux existants [Mejbri 02] montre que les solutions actuelles sont peu satisfaisantes : elles nécessitent souvent une étape de post-correction manuelle importante des documents interprétés. Les problèmes de recherche associés à cette thématique, surtout le traitement des parties graphiques fortement connectées, sont donc encore nombreux [Ceheux 02].

Dans les paragraphes précédents, nous avons présenté les principales thématiques de l'analyse d'image de document : la reconnaissance de l'écriture manuscrite [Vinciarelli 02], l'interprétation de documents structurés [Tang 96] et l'interprétation de documents graphiques [Ablameyko 00]. En pratique, il est difficile de les dissocier. En effet, les documents peuvent combiner des parties manuscrites et graphiques, avec des éléments propres aux documents structurés [Kasturi 02] [Nagy 00a]. Les figure (2) (a) (b) (c) donnent des exemples de documents hétérogènes combinant différentes informations. La figure (2) (a) présente une fiche d'incorporation militaire complétée par des parties manuscrites [Camillerapp 04]. La figure (2) (b) présente un manuel scientifique mélangeant des parties textuelles et graphiques [Ingold 02]. Enfin, la figure (2) (c) donne un extrait de plan technique composé essentiellement de parties graphiques, renseignées par des champs textuels [Tombre 02].



(a) (b) (c)

FIG. 2 – Documents hétérogènes : (a) fiche (b) manuel (c) plan

Des systèmes d'analyse d'image de document

On qualifie ainsi un dispositif informatique utilisé pour une application d'analyse d'image de document [Nagy 00a]. Il a pour vocation d'obtenir une description interprétée par un ordinateur des données pixels de l'image du document. La conception d'un tel système relève donc de la discipline du traitement des images [Coster 89] mais se rapproche aussi de celle de l'intelligence artificielle [Haton 91]. Ce rapprochement avec l'intelligence artificielle est également vrai pour des disciplines comme le traitement du signal [Kunt 91] et la vision [Cocquerez 95] [Jolion 01]. Dans les années 1960, le croisement de ces différentes disciplines (traitement du signal, traitement d'image, vision et intelligence artificielle) a donné naissance à celle plus large de la reconnaissance des formes[8] [Tou 74][9]. La conception d'un système d'analyse d'image de document relève donc des trois disciplines connexes que sont : le traitement des images, la reconnaissance des formes, et l'intelligence artificielle.

Il est difficile de proposer un schéma général de l'architecture d'un système d'intelligence artificielle [Haton 91], chaque système étant en effet très spécifique. Nous en présentons deux dans la figure (3) qui nous semblent représentatifs pour les systèmes de supervision de programmes [Thonnat 95] (3) (a) et d'interprétation de documents [Saidali 04] (3) (b). On peut noter des composantes communes sur ces deux schémas : une librairie de traitements (libraries of programs, tools), un système de contrôle (program supervision shell, interpretation) et une base de connaissances (knowledge base, models DTD). Par la suite, nous détaillons ces composantes dans le cadre des systèmes d'analyse d'image de document.

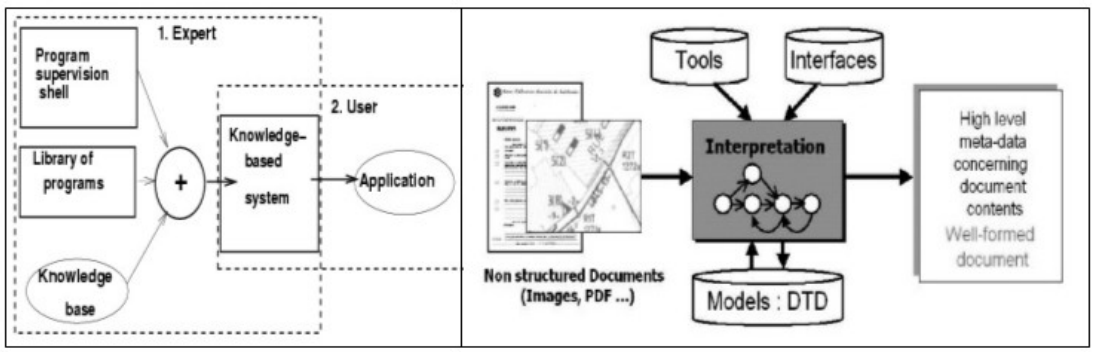


FIG. 3 – (a) supervision de programmes (b) interprétation de documents

[8] pattern recognition

[9] La reconnaissance des formes est considérée comme l'application des méthodes d'intelligence artificielle aux signaux et images, la frontière entre ces deux disciplines est donc assez floue [Milgram 93].

La librairie de traitements constitue la spécificité d'un système [Thonnat 95]. En effet, les traitements utilisés sont différents (par exemple) entre une application de vision et une d'analyse d'image de document [Kunt 00]. Les systèmes de contrôle ainsi que les architectures à base de connaissances peuvent être ré-utilisés entre applications de natures différentes [Ullman 89] [Garneau 02]. En analyse d'image de document les différents traitements de la librairie s'agencent le plus souvent sous forme de chaîne. Une chaîne de traitements classique peut alors être décomposée en deux étapes principales [Kasturi 02] : le traitement d'image et la reconnaissance. Celles-ci sont illustrées sur la figure (4). Le traitement d'images a pour objectifs le pré-traitement des images [Ablameyko 00] et l'extraction de caractéristiques décrivant les formes des images [Loncaric 98]. La reconnaissance exploite les caractéristiques extraites durant la première étape afin de labelliser les objets du document. Il existe deux grandes familles d'approches relatives à ces traitements : statistiques & connexionnistes [Jain 00] et syntaxiques & structurelles [Tombre 96]. Les premières décrivent les objets du document sous la forme de vecteurs de primitives et les deuxièmes sous la forme de graphes.

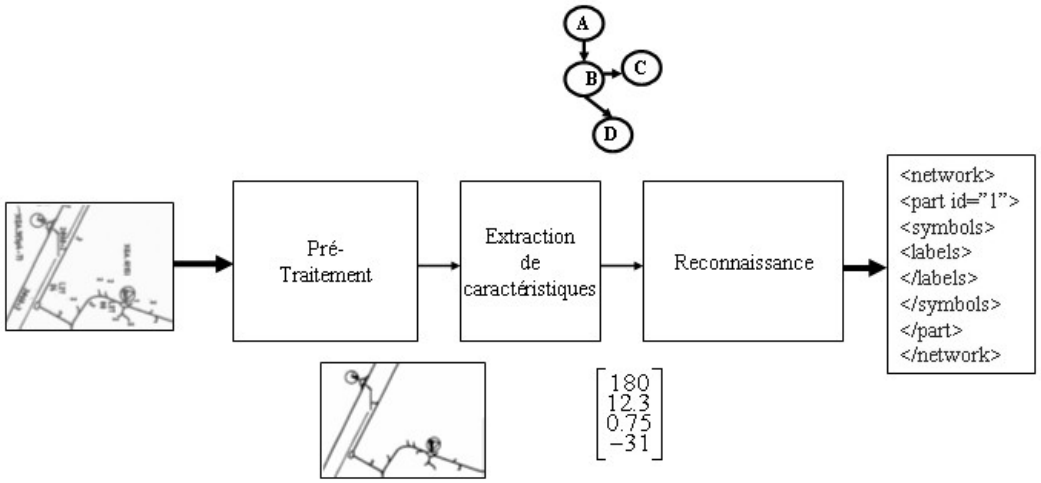


FIG. 4 – Chaîne de traitements en analyse d'image de document

Les systèmes de contrôle supervisent la librairie de traitements [Thonnat 95]. Il peut s'agir de systèmes planifiés a priori ou contextuels[10] [Mullot 00]. Les premiers utilisent un contrôle figé. Ils sont encore largement utilisés en analyse d'image de document [Ogier 00a] mais tendent à devenir obsolètes au profit des systèmes contextuels [Crevier 97]. Ces derniers auto-adaptent leur stratégie de contrôle en fonction du contexte de reconnaissance. Il existe diverses approches : les systèmes à base de tableau noir [Barr 89], les systèmes multi-agents [Ferber 97], les systèmes de pilotage automatique de programmes [Thonnat 95], ...

[10] [Clouard 99] les qualifie de "systèmes opportunistes", [Crevier 97] de "systems based on control issue", nous préférons ici la dénomination de [Mullot 00] de "systèmes contextuels".

Les systèmes capables de déclarer et d'exploiter, ou de produire, des connaissances sont qualifiés de système à base de connaissances (SBC) dans la littérature [Haton 91]. [Charlet 02] présente les connaissances comme définissant la manière dont des données/informations sont exploitées dans les systèmes informatiques. Les connaissances sont représentées au niveau procédural (interne) et déclaratif (externe) dans les systèmes [Haton 91]. Dans les deux cas, cette représentation est fonction du formalisme adopté. Certains formalismes s'apparentent plus à de la représentation de données [Lucas 86] (matrice, pile, ...), d'autres plus complexes à de la représentation des connaissances [Kayser 97] (logique de description, règle, réseau sémantique, objet structuré "frame", ...). Dans les SBC, la base de connaissances regroupe et structure de façon externe (déclarative) les connaissances du système [Kayser 97]. De nombreuses approches existent pour la mise en oeuvre de ces bases : de l'utilisation de bases de données [Ullman 89] à celle de langages de représentation des connaissances [Kayser 97].

Des applications de tri au web mining : un historique

Les deux sections précédentes ont présenté quelques considérations sur la définition du document et introduit l'analyse d'image de document. Dans les années 1990, l'explosion de l'usage du document électronique et sa prolifération ont induit un changement radical dans les activités liées au document (production, publication, diffusion, échange, ...) [Spring 95]. Les disciplines du web sémantique et du web mining [Berendt 02] ont alors émergé autour des problématiques de modélisation et d'indexation des bases de documents électroniques.

Cette section présente un historique conjoint de la discipline de l'analyse d'image de document avec celles liées aux documents électroniques. Il est évidemment difficile de dresser un historique exhaustif de ces disciplines vu le volume de travaux publiés. La liste présentée ci-dessous des revues et des conférences/workshops leur étant dédiées en donne un aperçu. Nous avons donc choisi de limiter notre historique à différents aspects : les applications de tri et de rétroconversion, la révolution du document, l'évolution des formats de document et de leur sémantique, et enfin le web sémantique et le web mining.

- **Les revues :** Computer Vision Graphics and Image Processing, Communications of the ACM, Computer Vision and Image Understanding, International Journal on Document Analysis and Recognition, Pattern Analysis and Applications, Pattern Analysis and Machine Intelligence, Pattern Recognition, Transactions on Systems Man and Cybernetics ...
- **Les conférences (et workshops) :** Conference on Computer Vision and Pattern Recognition, Conference on Document Recognition and Retrieval, Conference on Structural and Syntactical Pattern Recognition, International Conference on Document Analysis and Recognition, International Conference on Pattern Recognition, Workshop on Document Analysis Systems, Workshop on Web Document Analysis, International Workshop on Visual Form, ...

Applications de tri et rétroconversion

Les premières applications d'analyse d'image de document apparues dans les années 1960 [Mori 92] concernaient la reconnaissance optique de caractères imprimés OCR. Elles étaient utilisées pour le traitement en masse des documents administratifs (tri postal, traitement de bulletins de salaires, ...). Elles se sont étendues dans les années 1980 [Nagy 00a] sur divers aspects connexes aux OCR : pré-traitement d'images, reconnaissance de formulaires, ... En effet, dans les années 1980, l'essor technologique et commercial a permis l'extension de l'informatique des domaines réservés (laboratoires scientifiques, armée, industries de haute technologie, hautes administrations, ...) vers les entreprises et particuliers [Campbell 96].

L'analyse d'image de document s'est alors étendue [Tang 91] du problème de la reconnaissance d'objets isolés (caractères) sur les images à celui de leur interprétation pour leur rétroconversion. [Rares 99] présente l'interprétation comme la capacité d'un système à mettre en correspondance des concepts liés et a priori connus (maison - à côté - route, arbre - en bordure - route, personne - dans - maison, ...). [Hilaire 04] présente la rétroconversion dans le cas de l'interprétation d'image de document de la façon suivante : "Étant donné un document papier, la rétroconversion de ce document est le processus qui vise à en trouver une représentation numérique, manipulable par l'ordinateur, la plus proche possible de celle que son concepteur aurait utilisée s'il avait conçu le document devant sa station de travail".

L'objectif des systèmes d'interprétation d'image de document [Tang 91] à travers la rétroconversion était alors de contribuer[11] à la conversion des documents papier en documents électroniques pour leur exploitation au sein des systèmes informatiques [Campbell 96]. Ce passage s'est cristallisé autour du développement de la Gestion Électronique de Documents (GED) [Prax 01]. [Stinckwich 00] la définit de la façon suivante : "le vocable GED signifie gestion électronique de documents de toute nature : du papier aux documents électroniques". La liste qui suit énumère les grandes étapes de la GED telle qu'elle est vue aujourd'hui [Stinckwich 00]. À travers cette liste, on peut voir que la GED est vaste et que l'interprétation d'image de document n'en constitue qu'une étape :

- Acquisition : saisie, numérisation (scanneurs + interprétation) ;
- Composition et présentation : enrichissement, formattage, liens hyper-textuels, règles typographiques, synchronisation des séquences vidéos et audios ;
- Archivage, indexation : fichiers, bases de données ;
- Consultation : requêtes, protocoles clients-serveurs, écran ;
- Diffusion : réseau Internet, Minitel ;
- Production sur support papier ou électronique : CD/ROM, DVD.

[11] 3.5 Milliards de documents techniques papier en circulation aux USA/Canada en 1992 [Filipski 92].

Aujourd'hui le bilan sur les systèmes de rétro-conversion est plutôt contrasté. Les travaux de synthèse récents [Ceheux 02] [Nagy 00a] montrent qu'en l'état actuel des recherches ces systèmes ne peuvent être utilisés que pour des cas restreints ou dédiés. De plus, la conception d'un tel système est une tâche complexe, difficilement adaptable, et coûteuse [Clouard 02]. Par conséquent, les applications probantes de rétroconversion n'ont pu se faire jusqu'alors que dans le cadre de projets ciblés [Grenier 01]. Le tableau (2) synthétise et compare trois de ces projets ayant été appliqués à des bases de documents importantes (plus de dix mille documents). Dans ces conditions, on comprend alors que la majorité des systèmes de GED se soient cantonnés à la numérisation en masse des documents papier [Lalou 01], et occasionnellement à des étapes de prétraitement voire d'OCR de certaines parties de document [Seta 04].

Système	**Type de document**	**Partenaire**	**Base**	**Traitement**
DMOS [Coüasnon 03]	anciens registres d'état civil renseignés de façon manuscrite	Archives Nationales de France	60 000	reconnaissance de la structure physique des documents et des patronymes manuscrits
FACIT [Wille 96]	fiches bibliographiques pour l'indexation d'ouvrages en bibliothèques	Section des Librairies de l'Union Européenne	40 000	reconnaissance de la structure physique des documents et OCR
MARS [Kim 01]	articles de journaux médicaux	USA National Library of Medicine	11 000	reconnaissance de la structure physique et logique des documents, et OCR

TAB. 2 – Quelques systèmes de rétroconversion sur de "larges bases"

La révolution du document : le document électronique

Dans les années 1990, l'informatique a connu un deuxième essor avec l'avènement des réseaux et d'Internet [Segaller 98]. En effet, l'avancée technologique des réseaux locaux et l'accroissement du nombre de machines durant cette période ont motivé les institutions (entreprises, universités, administrations, ...) à mettre en réseau leur parc informatique [Campbell 96]. La GED connaît également son réel essor au cours de cette période. Parallèlement à la construction de ces réseaux locaux, le réseau global Internet et le Web connaissent un développement important [Segaller 98]. La combinaison de ces différents phénomènes a pour conséquence une production et une mise à disposition de plus en plus massive de documents électroniques [Spring 95]. La liste chronologique suivante illustre ces différents aspects :

- **1985-1996 :** Essor d'Ethernet [Segaller 98] : normalisation Ethernet IEEE 1985, Ethernet ATM 1993, Ethernet Gigabit 1996, ...

- **1989-1996 :** Mondialisation d'Internet [Segaller 98] : 1981 - 213 Machines connectées, 1989 - 130 000, 1996 - 9 Millions.
- **1992 :** [Filipski 92] estime qu'environ 1 document technique sur 2 est produit électroniquement aux USA/Canada.
- **1997-1998 :** Une étude du Gartner Group[12] estime qu'en 1998 le nombre d'utilisateurs de systèmes de GED est de 48 Millions contre 4 Millions en 1997.
- **1997-2001 :** Différentes études synthétisées dans [Ouf 01] montrent que le Web double approximativement de taille tous les ans sur cette période. On l'estime en 2001 à 4 Milliards de pages, pour un volume de données avoisinant les 21 Tera octets.
- **1997-fin 2010 :** [Sohm 97] estime que le parc informatique mondial a triplé en dix ans pour atteindre les 750 Millions de machines en 1997. Il estime également une croissance après 1997 d'environ 20% par an jusqu'à fin 2010 pour atteindre un parc d'environ 1,5 Milliards de machines.

Les bases de données constituées par les réseaux locaux et le réseau global Internet ont commencé à fusionner aux alentours de 1995 [Bergman 01] en ce qu'on qualifie aujourd'hui de Web visible et le Web invisible[13] [Ouf 01]. Le Web invisible est constitué de différentes catégories d'information : pages web générées dynamiquement, sites web internes, bases de données locales ouvertes sur Internet, ... [Ouf 01] estime en 2001, à partir de la synthèse de différentes études, que les Web visible et invisible sont composés respectivement de 21 et de 7 440 Tera octets d'information. Cette importante différence de taille s'explique par le fait que le Web invisible est constitué essentiellement de larges bases de données créées avant l'avènement d'Internet et mises à disposition par la suite [Bergman 01]. De même, ce Web invisible croît plus vite que le Web visible grâce à la production électronique régulière des usagers d'ordinateur archivée au sein de leurs systèmes d'informations [Ouf 01]. Ces différents éléments expliquent pourquoi le Web invisible diffère en qualité du Web visible en ce qui concerne l'indexation de l'information (il s'agit d'une information indexée et organisée par les organisations telles que les entreprises, les universités, les administrations, ...) et la qualité de cette information (la publication est contrôlée par ces organisations).

Évolution des formats électroniques de document

Si l'apport sémantique des documents électroniques en informatique par rapport aux documents papier est évident [Spring 95], rien ne garantit que cette sémantique soit suffisante pour une indexation automatique pertinente des bases documentaires électroniques [Berrien 03]. Afin de discuter de cette problématique, nous présentons dans le tableau (3) quelques éléments sur l'évolution des formats électroniques de document extraits de [Murray 96], [Roisin 99], et [Rounds 04].

[12] http ://www4.gartner.com/

[13] "Invisible Web" fut utilisé en 1994 par J. Ellsworth pour désigner l'information invisible aux moteurs de recherche traditionnels, on parle également de "Deep Web" [Bergman 01].

Début	Formats	Description
1985	standardisés et structurés	**Textuel** (Word "1983", TEX"1984" et LATEX"1986", PDF "1993", RTF "1995", ...) **Image** (JPEG "1985", TIFF "1986", GIF "1987", ...) **Graphique** (DXF "1982", PS "1986", Corel Draw "1989", Windows Meta File "1993", PNG "1996", ...) **Donnée** (Lotus *IBM* "1983", Excel "1985", ...)
1990	sémantiques	CGM "1987", OfficeVision *IBM* "1989", HTML "1991" & CSS "1996", ...
2000	applicatifs	OpenDoc *Apple* "1996", OLE *Microsoft* "1997", XML "1998" et XSLT "2000", ...

TAB. 3 – Évolution des formats électroniques de document

Initialement, les données associées aux documents étaient stockées de façon dédiée à chaque système informatique sous format binaire et/ou ASCII. La notion de format telle qu'on la connaît aujourd'hui apparaît au milieu des années 1980 [Campbell 96]. Au cours de cette période, de nombreux formats apparaissent. Le tableau (3) en donne quelques exemples[14]. Ils appartiennent à quatre principales catégories [Rounds 04] : les formats image [Murray 96], les formats graphiques (vecteurs, cercles, rectangles, ...) [Murray 96], les formats textuels [Roisin 99] et les formats de données (tableurs et autres) [Campbell 96]. Le but principal de ces formats était de normaliser l'échange des données entre systèmes informatiques. Certains de ces formats se sont imposés au cours du temps comme standards, alors que d'autres sont restés propriétaires [Murray 96]. La principale caractéristique de ces formats est la structuration des données [Rounds 04] : structuration logique des documents textuels, structuration des données graphiques en couches, structuration des données en tableaux, ... Ces différents formats ont évolué au cours du temps permettant la combinaison de données hétérogènes [Roisin 99] comme des graphiques, du texte, des images, et des données (Word, LATEX, PDF, ...).

Dans les années 1990 une nouvelle génération de formats sémantiques apparaît [Murray 96] [Roisin 99]. Ces formats véhiculent des méta-données caractérisant le contenu des documents [Parent 99]. Le tableau (3) donne quelques-uns de ces formats. Dans les formats graphiques, le format *CGM* est qualifié de format graphique intelligent [Ponte 97]. En effet celui-ci permet, contrairement aux formats graphiques standards, de définir des relations hiérarchiques et symboliques entre les primitives graphiques. Concernant les formats textuels, IBM exploite en 1989 (dans son application OfficeVision) un format basé sur la norme *ODA*[15]. Cette norme permet à la fois de modéliser la structure logique et physique des documents. Citons aussi l'apparition en 1991 du langage[16] *HTML* permettant de décrire le contenu des documents à l'aide de meta-tags, et de relier les documents à l'aide d'hyper-liens.

[14] Ce tableau utilise les dénominations courantes : noms d'extension, noms d'application, ...
[15] Office Document Architecture
[16] Langage au sens de format programmable par un utilisateur [Campbell 96].

Vers la fin des années 1990, de nouvelles spécifications de format apparaissent. Elles sont motivées par l'accroissement de l'hétérogénéité des données véhiculées dans les formats, et par les besoins grandissants de description sémantique de ces données. Ces aspects sont en effet centraux à la problématique de la structuration du Web, via la discipline naissante du web sémantique (voir section suivante) [Dobson 95]. Aussi ces formats ont évolué selon deux voies principales : ils sont devenus adaptatifs [Decker 00], et ils véhiculent des traitements [Stinckwich 00]. [Roisin 99] parle d'applications multimédia, nous employons ici le terme de formats applicatifs pour les désigner. Le tableau (3) liste quelques-uns de ces formats. Cependant on peut distinguer le langage XML des autres formats dans cette liste vu l'engouement et l'ampleur des travaux qu'il suscite [Ceri 00]. Nous détaillons ces différents formats dans les paragraphes qui suivent.

Le format applicatif *OpenDoc* [Apple 97] a été mis en place puis abandonné par IBM entre 1996 et 1998. L'idée était de "tronçonner" les applications classiques (logiciel de traitement de texte, éditeur d'image, tableur, ...) selon leurs composants redondants (correcteur d'orthographe, outil d'impression, éditeur d'équation, ...). Les fichiers OpenDoc véhiculent des données faisant alors appel aux composants dont elles ont besoin. Le format applicatif *OLE* de Microsoft [Microsoft 00] a repris[17] et pérennisé les concepts d'OpenDoc. Dans les développements récents du format OLE, [Microsoft 00] montre dans quelle mesure ce format s'intègre naturellement avec des techniques de data mining (voir section suivante) pour extraire de la sémantique à partir des données contenues dans les fichiers OLE.

L'essor important des formats applicatifs vient des travaux relatifs au langage *XML* [Decker 00]. Ce langage est très centré sur la description de données [Ceri 00]. Il permet la définition de sous-langages [Dui 03], et peut être exploité par des langages de traitement (XSLT, XML-QL, XPATH, ...) offrant des mécanismes de requête et de transformation [Zachary 00]. Il est ainsi possible de coupler ces différents langages de traitement avec XML, afin de définir des documents applicatifs [Villard 00]. Une part importante des travaux sur XML concerne les aspects sémantiques [Boley 03]. Ils servent de fondement à la discipline du web sémantique (voir section suivante) [Decker 00].

Du web sémantique et du web mining

Au regard de l'évolution des formats électroniques de document on peut constater que la sémantique est un problème récurrent depuis le début des années 1990. Ceci soulève la question de l'évaluation de cette sémantique dans les bases documentaires électroniques contenues dans le Web visible et invisible [Ouf 01].

[17] Nous ne détaillons pas les différences entre ces deux formats ici et retournons les lecteurs intéressés à [Alger 94].

En ce qui concerne le Web visible, [Steve 99] estime en 1999 qu'il est composé en volume de données pour 2/3 d'information textuelle structurée essentiellement via le langage HTML[18], et pour 1/3 d'information image. Les moteurs de recherche s'appuient donc [Ouf 01] sur les données textuelles, le peu de structure fournie par le langage HTML [Vijjappu 01], les hyper-liens et les méta-tags[19].

Concernant le Web invisible, on ne peut que spéculer sur la sémantique de ses données. Cependant, diverses initiatives gouvernementales [Cornu 98] [Lawrence 00] [Rounds 04] sur la conservation de l'information numérique laissent supposer que ce Web invisible est composé principalement de formats électroniques standards (voir tableau (3)). En effet, l'utilisation de formats sémantiques pour la production de documents électroniques reste du domaine réservé à la communauté informatique et scientifique. De plus, la production de documents électroniques standards est le plus souvent basée sur des applications WYSIWYG[20], ce qui diminue significativement leur qualité de structuration [Lawrence 00]. Enfin, même si ces documents permettent de véhiculer des données hétérogènes (texte, graphique, ...), force est de constater que l'image reste (et restera) [Breuel 04] un moyen de communication privilégié entre documents de formats différents.

La sémantique du Web (visible et invisible) ne semble donc pas suffisante pour en permettre une indexation pertinente [Berrien 03] [Vijjappu 01]. Diverses disciplines de recherche ont alors émergé autour des problématiques de modélisation et de construction de cette sémantique : le web sémantique [Berners-Lee 01] et le web mining [Wang 00].

Le web sémantique est apparu au milieu des années 1990 [Dobson 95]. [Berners-Lee 01] le présente de la manière suivante : "Le web sémantique est une extension du Web courant dans lequel chaque information est fournie avec sa signification précise, permettant ainsi une meilleure interaction utilisateurs/ordinateurs"[21]. Une part importante des travaux associés au web sémantique a concerné le langage XML et ses aspects sémantiques [Decker 00].

Le web mining s'est développé parallèlement au web sémantique dans le milieu des années 1990 [Etzioni 96]. Il correspond à l'application des techniques issues du data mining [Shapiro 00] aux informations contenues sur le Web [Wang 00]. Il est donc proche du data mining et peut se définir à travers lui. [Radivojevic 03] définit synthétiquement le data mining de la façon suivante : "Le data mining peut être défini comme l'extraction automatique d'informations implicites entre différentes sources de données"[22].

[18] Le volume de données correspondant aux tags HTML a été soustrait.

[19] Les méta-tags sont peu fréquents sur Internet par rapport au nombre de pages en ligne [Ouf 01].

[20] What You See Is What You Get [Campbell 96]

[21] "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation".

[22] "The data mining can be defined as automated extraction of predictive information from different data sources".

Le web sémantique [Berners-Lee 01] concerne donc la modélisation de la sémantique du Web, et le web mining [Wang 00] celle de sa recherche. Ces deux disciplines sont complémentaires, et tendent à se rapprocher aujourd'hui [Berendt 02] dans la discipline plus large du semantic web mining[23,24].

Vers l'analyse des documents

Au cours de ce chapitre nous avons présenté l'évolution conjointe de l'analyse d'image de document et du semantic web mining. Ces disciplines sont en forte corrélation [Smith 04]. Elles tendent à se rapprocher depuis le début des années 2000 [Breuel 04] dans ce que l'on nommera dans ce manuscrit l'analyse des documents (le pluriel est ici volontaire pour en élargir la portée). On entend par documents tous les types de document, papier et/ou électronique [Roisin 99]. Différents états de l'art sur l'image et le web ont été produits récemment par des membres des communautés, de l'analyse d'image de document [Plamondon 00] [Wenyin 04], et du semantic web mining [Walker 00] [Baird 03]. Ces différentes publications sont la preuve du rapprochement de ces disciplines. À travers ces états de l'art on constate que l'analyse des documents est à la croisée de trois disciplines : l'analyse d'image de document [Kasturi 02], l'analyse de document électronique et le semantic web mining [Berendt 02] [Baird 03] [Breuel 04]. En pratique il est évidemment difficile d'établir des frontières strictes entre ces trois disciplines. Cependant, chacune d'entre elles répond à une problématique de recherche propre que nous présentons dans le tableau (4). Nous détaillons dans la suite de cette section ces disciplines et leurs problématiques, ainsi que leurs enjeux économiques associées.

	Analyse d'image de document	**Analyse de document électronique**	**Semantic web mining**
Corpus	documents papier	documents électroniques	Web
Sémantique	image	image, graphique, symbolique, textuel	textuel, métadonnée
Intention	interprétation	de la segmentation à l'interprétation	indexation et recherche automatique
Problématique	interprétation d'images complexes et dégradées	traitement de données sémantiquement hétérogènes	structuration du Web

TAB. 4 – Problématiques des disciplines liées à l'analyse des documents

[23] Nous conserverons dans ce manuscrit l'appellation anglophone.
[24] Workshop on Semantic Web Mining,Workshop on Semantic Web Mining and Reasoning, ...

L'analyse d'image de document avait pour vocation initiale l'interprétation exhaustive de tous les types de documents papier [Tang 91]. La révolution du document dans les années 1990 a bouleversé cette vocation [Spring 95]. En effet, l'usage du document électronique s'est "en partie" imposé sur celui du document papier [Rounds 04]. Cependant, toutes les bases de documents papier dans le monde n'ont pas été rétroconverties [Spring 95], loin s'en faut. Les documents institutionnels (bibliothèques [Lalou 01], cadastres [Ogier 00b], grandes industries [Ponte 97], ...) constituent en effet la majeure partie de ces bases documentaires. Ces documents institutionnels sont, de natures complexes, dégradés, et regroupés en "larges bases" (plusieurs milliers de documents). Ces différents facteurs font que beaucoup des bases documentaires institutionnelles sont encore en attente de solutions de rétroconversion viables [Ceheux 02]. Ces facteurs induisent également des coûts de rétroconversion manuelle trop importants[25] pour que cette solution soit envisagée. Le tableau (5) donne quelques exemples de bases documentaires institutionnelles avec leurs niveaux de rétroconversion. Il montre que les champs d'application de l'analyse d'image de document restent vastes [Ceheux 02], malgré la prolifération des documents électroniques de ces dix dernières années [Spring 95].

Institution	**Base**	**Niveau de rétroconversion**
France Télécom [Grenier 01]	2.4 millions de plans de réseau	$< 5\%$ de plans rétroconvertis manuellement (2001)
Cadastre Français [Ogier 00b]	250 000 planches cadastrales	10% de planches numérisées (1997)
Airbus [Ponte 97]	12 millions de pages/Airbus de documentation technique	numérisée (1997)
Bibliothèque Nationale de France [Lalou 01]	25 000 manuscrits médiévaux	numérisée (fin 2005)

TAB. 5 – Niveaux de rétroconversion de bases documentaires papier institutionnelles

L'analyse de document électronique est au croisement de l'analyse d'image de document et du semantic web mining [Breuel 04]. Elle considère la totalité des données présentes dans les documents électroniques [Walker 00], quelles que soient leurs sémantiques. Elle traite les données de plus faible sémantique afin et de les mettre en correspondance avec les données de plus haute sémantique. [Rigamonti 04] met en évidence différentes faiblesses sémantiques rencontrées dans les documents électroniques : absence de structure physique, falsification de la structure logique, primitives graphiques à base de vecteurs ou de pixels pour l'encodage de caractères, ... Le traitement des données de faible sémantique permet alors d'améliorer les descriptions sémantiques globales des documents pour une meilleure indexation dans les bases de documents électroniques [Baird 03] et/ou dans le Web [Kanungo 01].

[25] 50 millions d'Euro pour les plans de réseau France Télécom [Grenier 01] (tableau (5)).

Les enjeux de l'analyse de document électronique sont importants. En effet, [Ouf 01] estime dans son rapport que les bases de documents électroniques mondiales (Web et systèmes de GED) se composent de 7 461 Tera octets d'information. Au sein de ces documents les données de faibles sémantiques sont présentes en large proportion et plus particulièrement les données image. La liste suivante donne quelques considérations sur cette question :

- [Steve 99] estime que le Web est composé en volume de données pour 2/3 d'information textuelle et pour 1/3 d'information image.
- [Kanungo 01] estime que 42% des images présentes sur le Web contiennent des informations textuelles. Il estime également que 59 % de ces images contiennent des mots non encodés dans les fichiers HTML référençant ces images.
- [Breuel 04] estime que l'image est (et restera) un vecteur de communication privilégié dans les documents électroniques pour ses facilités de représentation, d'édition, de communication, et de conversion.
- [Baird 03] souligne les lacunes importantes en matière de traitement d'image des systèmes de gestion des bases documentaires électroniques. Il "en appelle" à la communauté d'analyse d'image de document afin qu'elle s'oriente sur cette problématique.

Les images présentes dans les documents électroniques sont majoritairement de plus faibles dimensions [Steve 99] et de plus faibles niveaux de dégradation [Kanungo 01] que celles traitées en analyse d'image de document. Les aspects traitement d'image semblent donc "moins complexes" [Walker 00]. En contrepartie les contenus des images semblent plus variables. En effet, les images traitées en analyse d'image de document appartiennent majoritairement à des corpus établis et répondent à des chartes de composition (tableau (5)). À l'opposé, les images traitées en analyse de document électronique sont hétérogènes en ce qui concerne les tailles, les types, et les contenus [Steve 99] [Kanungo 01]. Ces contraintes induisent des capacités d'adaptation accrues des systèmes d'analyse de document électronique [Rigamonti 04].

Le semantic web mining [Berendt 02] a pour objectifs la modélisation de la sémantique du Web et la recherche de cette sémantique. Il aborde donc les aspects externes aux documents (indexation des documents et modélisation de leurs relations sémantiques) [Berendt 02] tandis que l'analyse de document électronique aborde les aspects internes (analyse des contenus) [Rigamonti 04]. Il concerne la construction et l'exploitation des méta-données des documents pour la structuration des bases de documents électroniques et du Web [Fensel 02]. Les enjeux sont importants lorsque l'on considère les volumes de données contenus dans les bases de documents électroniques mondiales [Ouf 01]. Les problématiques abordées par le semantic web mining sont plus proches de l'ingénierie des connaissances [Fensel 02] que de l'analyse de document électronique ou de l'analyse d'image de document. L'évolution actuelle de la GED vers la GEIDE[26] confirme cette démarcation [Pelletier 98].

[26] Gestion Électronique d'Informations et de Documents pour l'Entreprise

Au cours de cette section nous avons présenté les disciplines liées à l'analyse des documents, leurs problématiques et leurs enjeux. L'analyse des documents concerne donc le traitement de bases de documents sémantiquement hétérogènes [Roisin 99]. Ces documents véhiculent des données appartenant à des sémantiques différentes : image, graphique, symbolique, texture, méta-donnée, ... Ils peuvent provenir de différentes sources [Bergman 01] : Web, systèmes de GED, librairies papier et numériques, ... Les buts d'un système d'analyse des documents peuvent donc être variables. Globalement ils visent à transformer une sémantique d'entrée en une sémantique de sortie de plus haut niveau [Saidali 04]. Ces buts dépendent de l'intention de traitement voulue par le concepteur du système. Ces intentions peuvent également être exprimées par les utilisateurs en fonction de la capacité d'un système à dialoguer avec celui-ci et à s'exécuter [Theune 03]. Le tableau (6) illustre quelques catégories d'intentions courantes en analyse des documents à travers quelques exemples de système. On peut constater l'hétérogénéité des données d'entrée et de sortie suivant les différents systèmes [Berrien 03] [Valveny 02]. De même, on peut noter les différences d'intentions des concepteurs/utilisateurs de ces systèmes [Kim 01] [Seguela 01].

Catégories d'intentions	**Contenu du document**	**Descriptions de systèmes**
interprétation	symbolique, textuel, relation sémantique, méta-donnée	**[Kim 01]** : Traitement d'images de documents structurés pour la reconnaissance de caractères, des structures physiques et logiques, et de l'indexation des documents **[Seguela 01]** : Fouille de documents techniques textuels pour la construction de modèles de connaissances
indexation, recherche, navigation	graphique, symbolique, textuel, méta-donnée	**[Berrien 03]** : Analyse lexicale de documents textuels pour l'extraction de mots clés **[Valveny 02]** : Segmentation d'images de documents techniques en composantes connexes et reconnaissance de légendes pour la navigation texte/graphique
reconnaissance	symbolique, textuel	**[Nagy 00b]** : Reconnaissance de caractères imprimés
segmentation	image, graphique	**[Tan 01]** : Segmentation d'images de documents structurés en blocs de texte **[Kise 99]** : Segmentation d'images de documents en structure de voisinage de composantes connexes

TAB. 6 – Catégories d'intention en analyse des documents

De l'analyse des documents graphiques

Cette discipline est la spécialisation de l'analyse des documents aux documents graphiques. On entend donc par documents graphiques tous les types de documents (papier et/ou électroniques), ou partie(s) de documents, contenant des données graphiques quelles que soient leurs sémantiques. En analyse des documents, certaines applications peuvent être considérées comme majoritairement graphiques : l'interprétation de documents techniques [Ablameyko 00], la reconnaissance de symboles [Lladós 02], la reconnaissance d'écritures scripts (en particulier asiatique [Suen 03]), ... D'autres applications ne sont qu'en partie graphiques : le traitement des documents manuscrits anciens [Lebourgeois 04] (reconnaissance de lettrines), l'interprétation de documents structurés [Nagy 00a] (reconnaissance de logos), ... La figure (5) donne quelques exemples d'images graphiques ((a) plan technique [Trupin 01], (b) symbole architectural [Valveny 04], (c) logo [Neumann 02], (d) caractère chinois [Lin 02], (e) lettrine [Lebourgeois 00]), et de représentations de graphiques vectoriels ((f) symbole CAD [Love 01], (g) ClipArt [Fonseca 04]).

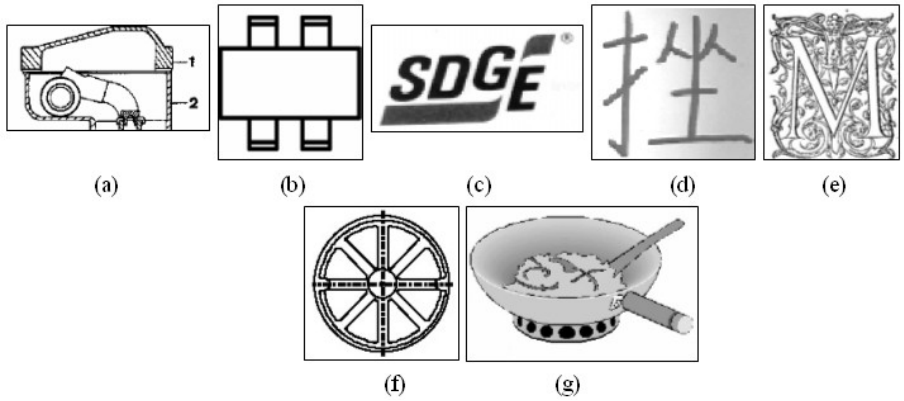


FIG. 5 – (a) plan (b) symbole (c) logo (d) caractère script (e) lettrine (f) symbole CAD (g) ClipArt

Ces exemples (figure (5)) montrent que les graphiques répondent à de fortes relations spatiales. Aussi, les systèmes d'analyse des documents graphiques s'appuient le plus souvent sur les approches structurelles. [Tombre 96] les définit de la façon suivante : "Les approches structurelles sont basées sur l'organisation de caractéristiques bas-niveau dans des structures de haut-niveau"[27]. Les caractéristiques bas-niveau sont communément qualifiées de primitives graphiques dans la littérature [Shih 89]. Celles-ci sont à base de pixels et/ou de vecteurs [Song 03]. Elles sont ensuite agencées dans une structure de plus haut-niveau le plus souvent de type graphe : on parle alors de graphe extrait.

[27] "Structural methods are based on the relational organization of low-level features into higher-level structures".

Le graphe[28] est largement utilisé en informatique pour la représentation et la manipulation de données symboliques [Spinrad 03]. En analyse des documents l'exploitation est le plus souvent basée sur les méthodes à base d'appariement de graphes [Gondran 95] ou de grammaire de graphes [Blostein 96]. Les premières apparient les graphes extraits avec des graphes modèles dans le but de trouver les graphes modèles les plus proches [Gondran 95]. Les secondes appliquent différentes règles afin de transformer les graphes extraits en graphes modèles [Blostein 96]. Dans les deux cas (appariement et grammaire), l'application de ces méthodes permet de labelliser (ou étiqueter) les graphes, et donc de reconnaître les parties graphiques du document [Popel 02] [Mahoney 02].

Lorsque les graphes extraits correspondent exactement aux graphes modèles, le problème de manipulation de graphe est considéré comme un problème de graphes exacts. Malheureusement, dans les applications d'analyse des documents, les graphes extraits sont souvent bruités [Tombre 96]. Le problème de reconnaissance est donc un problème de graphes inexacts. Un des avantages de la représentation graphe réside dans la possibilité de recherche de sous-graphe pour l'extraction de sous-structure dans un document graphique [Lladós 01] [Sánchez 02]. Un sous-graphe correspond à un sous-ensemble de noeuds et d'arcs d'un graphe plus large [Spinrad 03]. Cet avantage est particulièrement adapté aux documents graphiques dans lesquels les parties graphiques sont en interconnection (figure (5)). La figure (6) donne un exemple de résultat de recherche de sous-graphe par appariement [Lladós 01], basé sur des primitives graphiques de type polygone représentant les contours des occlusions.

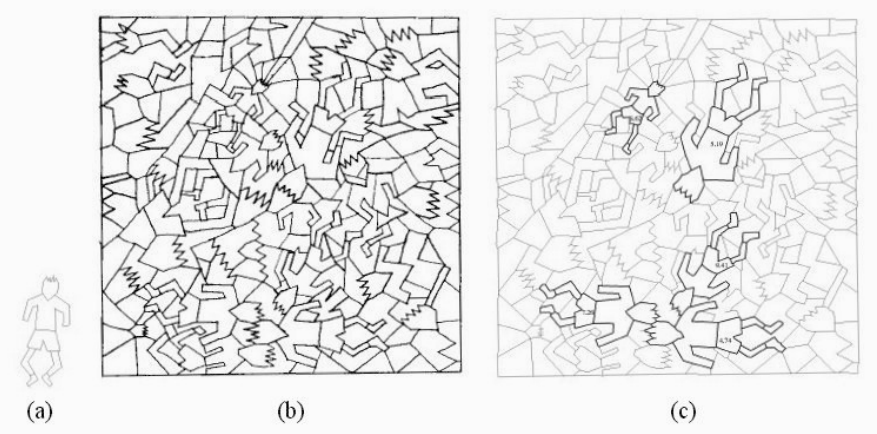


FIG. 6 – (a) modèle (b) image traitée (c) résultat de la recherche de sous-graphe

[28] Nous reportons le lecteur à l'Annexe A pour une introduction sur les graphes.

Introduction générale

Les travaux de cette Thèse concernent l'analyse des documents graphiques. Au cours du chapitre d'introduction nous avons illustré que ces documents répondent à de fortes relations spatiales. Aussi, les systèmes d'analyse des documents graphiques s'appuient le plus souvent sur les approches structurelles. Celles-ci sont basées sur l'organisation de primitives graphiques au sein de structure de plus haut-niveau le plus souvent de type graphe. Dans le cadre des travaux de cette Thèse nous nous sommes plus particulièrement intéressés à l'extraction des primitives graphiques. Plus précisément, nous avons abordé les problématiques de combinaison des méthodes pour l'extraction de ces primitives. Nous proposons pour cela une approche originale par reconstruction d'objets. Ce manuscrit se décompose en deux parties.

La première partie (**??**) présente notre étude bibliographique. Le chapitre (**??**) présente un état de l'art sur les méthodes d'extraction de primitives graphiques. Les avantages et les inconvénients de chacune des méthodes y sont illustrés. Nous y illustrons également les différents niveaux d'extraction selon lesquels se décomposent ces méthodes. L'intérêt de leur combinaison est alors mis en avant, cependant celle-ci soulève le problème de l'échange des primitives graphiques entre les méthodes. Plus largement, cette combinaison soulève le problème de la gestion des connaissances dites graphiques au sein des systèmes. Nous abordons cette problématique au cours du chapitre (**??**). Nous y présentons un état de l'art sur les systèmes à base de connaissances et sur la gestion des connaissances graphiques. Nous avons plus particulièrement centré cet état de l'art sur la représentation des connaissances graphiques (formalismes et modèles). Au travers de cet état de l'art nous illustrons d'abord les fortes relations de composition et de spécialisation existantes au sein des connaissances graphiques. Nous montrons ensuite que les systèmes manipulent des connaissances graphiques proches mais représentées de différentes façons en particulier en ce qui concerne les modèles. Nous en concluons que le problème de l'échange des connaissances graphiques entre opérateurs d'extraction est avant tout un problème d'interopérabilité sur les modèles. Nous argumentons alors que cette interopérabilité ne peut être résolue que par une approche à base de multi-représentation. Nous proposons pour cela de formaliser la combinaison des opérateurs d'extraction comme un processus de reconstruction d'objets.

La deuxième partie (**??**) présente nos contributions. Nous présentons tout d'abord dans le chapitre (**??**) nos différents opérateurs d'extraction de primitives graphiques. Nous avons développé trois classes d'opérateurs basés sur les approches région, contour, et squelette. De cette façon, ces opérateurs permettent de multiples représentations des objets graphiques afin d'être exploités dans le cadre de notre approche de reconstruction d'objets que nous présentons dans le chapitre (**??**) suivant. Dans cette approche les différentes connaissances graphiques sont assimilées à des objets informatiques. Le formalisme objet est choisi car il est particulièrement adapté pour décrire les relations de composition et de spécialisation existantes au sein des connaissances graphiques. Notre formalisme objet permet plus particulièrement la multi-représentation afin d'assurer l'interopérabilité entre les opérateurs. Ces derniers utilisent pour cela des ensembles de spécifications de contraintes déterminant comment les objets graphiques doivent être extraits puis insérés dans la base de connaissances. Nous présentons ensuite les principes généraux de notre approche de reconstruction d'objets. Nous illustrons que celle-ci s'apparente à une combinaison d'opérateurs au cours de laquelle les représentations évoluent (fortement). Elle se formalise alors naturellement sous la forme de graphe biparti. Basé sur ces principes généraux nous présentons alors notre système de reconstruction d'objets. Celui-ci utilise une méthodologie objet pour la formalisation des stratégies de reconstruction. Un système de contrôle, basé sur une approche de type systèmes experts, met alors en oeuvre ces stratégies. Nous présentons la mise en oeuvre de ce système et de notre formalisme dans notre chapitre (**??**) cas d'usage. Nous illustrons alors l'intérêt de notre approche de reconstruction d'objets pour la reconnaissance de symboles sur des images de documents graphiques. Nous montrons également les propriétés de généricité et d'adaptation de notre système.

Finalement, nous concluons et donnons nos perspectives sur ce travail.

Bibliographie

[Ablameyko 00] S. Ablameyko & T.P. Pridmore. Machine interpretation of line drawing images. Springer Verlag Publisher, ISBN : 3-540-76207-8, 2000.

[Alger 94] J. Alger. *OpenDoc vs. OLE When Elephants Fight, Only the Ants Get Hurt.* MacTech Magazine, vol. 10, no. 8, 1994.

[Apple 97] Apple. Opendoc programmer's guide. Editions Paperback, ISBN : 0201479540, 1997.

[Baird 03] H.S. Baird. *Digital Libraries and Document Image Analysis.* In International Conference on Document Analysis and Recognition (ICDAR), volume 1, pages 2–14, 2003.

[Barr 89] A. Barr, P.R. Cohen & E.A. Feigenbaum. The handbook of artificial intelligence, volume 1-4. Addison Wesley Publisher, ISBN : 0201118106, 1989.

[Bensefia 04] A. Bensefia. *Analyse des Documents Manuscrits : Identification et Vérification du Scripteur.* Thèse de Doctorat, Université de Rouen, France, 2004.

[Berendt 02] B. Berendt, A. Hotho & G. Stumme. *Towards Semantic Web Mining.* In International Semantic Web Conference (ISWC), pages 264–278, 2002.

[Bergman 01] M.K. Bergman. *White Paper : The Deep Web : Surfacing Hidden Value.* Journal of Electronic Publishing (JEP), vol. 7, no. 1, 2001.

[Berners-Lee 01] T. Berners-Lee, J. Hendler & O. Lassila. *The Semantic Web.* Scientific American, vol. 279, no. 5, 2001.

[Berrien 03] I. Berrien, F. Laburthe & J.D. Ruvini. *Interaction and Navigation for a Document Database : a Concrete Case Study.* In Workshop on Semantic Web and Databases (SWDD), pages 435–453, 2003.

[Blasselle 97] B. Blasselle. Histoire du livre. Editions Gallimard, ISBN : 2070533638, 1997.

[Blostein 96] D. Blostein, H. Fahmy & A. Grbavec. *Issues in the Practical Use of Graph Rewriting.* In Workshop on Graphics Recognition (GREC), volume 1073 of *Lecture Notes in Computer Science (LNCS)*, pages 38–55, 1996.

[Boley 03] H. Boley. *The Rule Markup Language : RDF-XML Data Model, XML Schema Hierarchy, and XSL Transformations.* In International Conference of Applications of Prolog (INAP), volume 2543 of *Lecture Notes in Computer Science (LNCS)*, pages 5–22, 2003.

[Breuel 04] T.M. Breuel. *The Future of Document Imaging in the Era of Electronic Documents.* Keynote Talk at Workshop on Document Analysis System (DAS), 2004.

[Bunke 03] H. Bunke. *Recognition of Cursive Roman Handwriting - Past, Present and Future.* In International Conference on Document Analysis and Recognition (ICDAR), pages 448–459, 2003.

[Camillerapp 04] J. Camillerapp, L. Pasquer & B. Coüasnon. *Indexation Automatique de Formulaires Anciens par Reconnaissance du Patronyme Manuscrit.* In Congrès Francophone de Reconnaissance de Formes et Intelligence Artificielle (RFIA), pages 1493–1502, 2004.

[Campbell 96] M. Campbell & W. Aspray. Computer. Basic Books, ISBN : 0-465-02990-6, 1996.

[Ceheux 02] G.R. Ceheux. *Stratégie pour l'Interprétation de Documents.* In Assises Nationales du GdR I3, pages 276–287, 2002.

[Ceri 00] S. Ceri, P. Fraternali & S. Paraboschi. *XML : Current Developments and Future Challenges for the Database Community.* In Conference on Extending Database Technology (EDBT), pages 3–17, 2000.

[Charlet 02] J. Charlet. *L'Ingénierie des Connaissances : Développements, Résultats, et Perspectives pour la Gestion des Connaissance Médicales.* Habilitation à Diriger les Recherches, Université Pierre et Marie Currie, France, 2002.

[Clouard 99] R. Clouard, A. Elmoataz, C. Porquet & M. Revenu. *Borg : A Knowledge Based System for Automatic Generation of Image Processing Programs.* Pattern Analysis and Machine Intelligence (PAMI), vol. 21, no. 2, pages 128–144, 1999.

[Clouard 02] R. Clouard, A. Elmoataz & M. Revenu. *Une Méthodologie de Développement d'Applications de Traitement d'Images.* In Congrès Francophone de Reconnaissance de Formes et Intelligence Artificielle (RFIA), pages 1033–1042, 2002.

[Coüasnon 03] B. Coüasnon & I. Leplumey. *A Generic Recognition System for Making Archives Documents Accessible to Public.* In International Conference on Document Analysis and Recognition (ICDAR), volume 1, pages 228– 232, 2003.

[Cocquerez 95] J.P. Cocquerez & S. Philipp. Analyse d'images : Filtrage et segmentation. Editions Masson, ISBN : 2225849234, 1995.

[Cornu 98] J.M. Cornu. Guide de l'information numérique : Comment traiter les données lisibles par machine et les documents numériques. Communautés européennes, ISBN : 92-828-2286-9, 1998.

[Coster 89] M. Coster & J.L. Chermant. Précis d'analyse d'images. Presses CNRS, ISBN : 2-87682-020-X, 1989.

[Crevier 97] D. Crevier & R. Lepage. *Knowledge-Based Image Understanding Systems : A Survey.* Computer Vision and Image Understanding (CVIU), vol. 67, no. 2, pages 161–185, 1997.

[Decker 00] S. Decker, F. Van Harmelen & J. Broekstra. *The Semantic Web - on the Respective Roles of XML and RDF.* IEEE Internet Computing, vol. 4, no. 5, pages 63–74, 2000.

[Dobson 95] S. A. Dobson & V. A. Burrill. *Lightweight Databases.* Computer Networks and ISDN Systems, vol. 27, no. 6, pages 1009–1015, 1995.

[Dui 03] D. Dui & W. Emmerich. *Compatibility of XML Language Versions ?* In Workshop on Software Configuration Management (SCM), volume 2649 of *Lecture Notes in Computer Science (LNCS)*, pages 48–162, 2003.

[Etzioni 96] O. Etzioni. *The World Wide Web : Quagmire or Gold Mine ?* Communications of the ACM, vol. 39, no. 11, pages 65–68, 1996.

[Fensel 02] D. Fensel & al. *Semantic Web Application Areas.* In Workshop on Applications of Natural Language to Information Systems (NLIS), pages 14–28, 2002.

[Ferber 97] J. Ferber. Les systèmes multi-agents, vers une intelligence collective. InterEditions, ISBN : 2-7296-0665-3, 1997.

[Filipski 92] A.J. Filipski & R. Flandrena. *Automated Conversion of Engineering Drawings to CAD Form.* IEEE, vol. 80, no. 7, pages 1195–1209, 1992.

[Fonseca 04] M.J. Fonseca, B. Barroso, P. Ribeiro & J.A. Jorge. *Retrieving Vector Graphics Using Sketches.* In Symposium on Smart Graphics (SG), volume 3031 of *Lecture Notes in Computer Science (LNCS)*, pages 66–76, 2004.

[Garneau 02] T. Garneau & S. Delisle. *Programmation Orientée-Agent : Evaluation Comparative d'Outils et Environnements.* In Journées Francophones pour l'Intelligence Artificielle Distribuée et les Systèmes Multi-Agents (JFIADSMA), 2002.

[Gondran 95] M. Gondran & M. Minoux. Graphes et algorithmes. Editions Eyrolles, 3 edition, ISBN : 0399-4198, 1995.

[Grenier 01] V. Grenier. *Contribution à l'Interprétation de Documents Techniques : une Approche Système.* Thèse de Doctorat, Université de Rouen, France, 2001.

[Haton 91] J.P Haton & al. Le raisonnement en intelligence artificielle. InterEditions, ISBN : 2-7296-0335-2, 1991.

[Heutte 03] L. Heutte. *Analyse et Reconnaissance de l'Ecriture : de Nouvelles Perspectives en Traitement Automatique de Documents Manuscrits.* Habilitation à Diriger les Recherches, Université de Rouen, France, 2003.

[Hilaire 04] X. Hilaire. *Segmentation Robuste de Courbes Discrètes 2D et Applications à la Retroconversion de Documents Techniques.* Thèse de Doctorat, Institut National Polytechnique de Lorraine (INRIA), France, 2004.

[Ingold 02] R. Ingold. *Analyse et Reconnaissance d'Images de Documents.* In Techniques de l'Ingénieur, volume H7020. 2002.

[Jain 00] A.K. Jain, R.P.W. Duin & J. Mao. *Statistical Pattern Recognition : A Review.* Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 1, pages 4–37, 2000.

[Jolion 01] J.M. Jolion. Les systèmes de vision. Editions Hermès, ISBN : 2-7462-0185-2, 2001.

[Kanungo 01] T. Kanungo & C.H. Lee. *What Fraction of Images on the Web Contain Text ?* In Workshop on Web Document Analysis (WDA), pages 43–46, 2001.

[Kasturi 02] R. Kasturi, L. O'Gorman & V. Govindaraju. *Document Image Analysis : A Primer.* Sadhana, vol. 27, no. 1, pages 3–22, 2002.

[Kayser 97] D. Kayser. La représentation des connaissances. Editions Hermès, ISBN : 2-86601-647-5, 1997.

[Kim 01] J. Kim, D.X. Le & G.R. Thoma. *Automated Labeling in Document Images.* In Conference on Document Recognition and Retrieval, volume 4307 of *Proc. SPIE*, pages 111–122, 2001.

[Kise 99] K. Kise, M. Motoi & K. Matsumoto. *On the Application of Voronoi Diagrams to Page Segmentation.* In Workshop on Document Layout Interpretation and its Applications (DLIA), 1999.

[Kunt 91] M. Kunt. Techniques modernes de traitement numérique des signaux, volume 1. Presses Polytechniques et Universitaires Romandes, ISBN : 2-880-74207-2, 1991.

[Kunt 00] M. Kunt. Reconnaissance des formes et analyse de scènes, volume 3. Presses Polytechniques et Universitaires Romandes, ISBN : 2-88074-384-2, 2000.

[Labiche 98] J. Labiche, J. Gardes & E. Trupin. *Cycle de Vie du Document Versus Système d'Interprétation Automatique.* In Colloque International Francophone sur l'Ecrit et le Document (CIFED), pages 443–452, 1998.

[Lalou 01] E. Lalou. *Une Base de Données sur les Manuscrits Enluminés des Bibliothèques.* Bulletin des Bibliothèques de France, vol. 46, no. 4, pages 38–42, 2001.

[Lawrence 00] G.W. Lawrence & al. *Risk Management of Digital Information : A File Format Investigation.* RLG DigiNews, vol. 8, no. 4, 2000.

[Lebourgeois 00] F. Lebourgeois, H. Emptoz, E. Trinth, F. Muge, C. Pinto & I. Granado. *Description du Matériel et Logiciel de Traitement d'Image pour la Numérisation des Collections et leur Interprétation.* Rapport technique WP4.4, Laboratoire Reconnaissance de Formes et Vision, INSA de Lyon, France, 2000.

[Lebourgeois 04] F. Lebourgeois & al. *Documents Images Analysis Solutions for Digital libraries.* In Workshop on Document Image Analysis for Libraries (DIAL), pages 2–24, 2004.

[Lin 02] F. Lin & X. Tang. *Off-Line Handwritten Chinese Character Stroke Extraction.* In International Conference on Pattern Recognition (ICPR), volume 3, pages 249–252, 2002.

[Lladós 01] J. Lladós, E. Marti & J.J. Villuanueva. *Symbol Recognition by Error Subgraph Matching Between Region Adjacency Graphs.* Pattern Analysis and Machine Intelligence (PAMI), vol. 23, no. 10, pages 1137–1143, 2001.

[Lladós 02] J. Lladós, E. Valveny, G. Sánchez & E. Martí. *Symbol Recognition : Current Advances and Perspectives.* In Workshop on Graphics Recognition (GREC), volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 104–127, 2002.

[Loncaric 98] S. Loncaric. *A Survey of Shape Analysis Techniques.* Pattern Recognition (PR), vol. 31, no. 8, pages 983–1001, 1998.

[Love 01] D.M. Love & J.A. Barton. *Drawing Retrieval Using an Automated Coding Technique.* In Conference on Flexible Automation and Intelligent Manufacturing (FAIM), pages 158–166, 2001.

[Lucas 86] M. Lucas. Algorithmique et representation des donnees, volume 2. Editions Masson, ISBN : 2-225-80924-0, 1986.

[Mahoney 02] J. Mahoney & M. Fromherz. *Interpreting Sloppy Stick Figures by Graph Rectification and Constraint-Based Matching.* In Workshop on Graphics Recognition (GREC), volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 222–235, 2002.

[Mao 03] S. Mao, A. Rosenfeld & Tapas Kanungo. *Document Structure Analysis Algorithms : A Literature Survey.* In Conference on Document Recognition and Retrieval, pages 197–207, 2003.

[Masson 93] E. Masson. Vallée des merveilles, un berceau de la pensée religieuse européenne. Editions Faton, ISBN : 2878440110, 1993.

[Mejbri 02] E.F. El Mejbri, H. Grabowski, H. Kunze, R.S. Lossack & A. Michelis. *3D Reconstruction of Paper Based Assembly Drawings : State of the Art and Approach*. In Workshop on Graphics Recognition (GREC), volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 1–12, 2002.

[Microsoft 00] Microsoft. *OLE DB for Data Mining Specification : Version 1.0*. Rapport technique, Microsoft Corporation, 2000.

[Milgram 93] M. Milgram. Reconnaissance des formes, méthodes numériques et connexionnistes. Editions Armand Collin, ISBN : 2200212909, 1993.

[Mori 92] S. Mori, C.Y. Suen & K. Yamamoto. *Historical Review of OCR Research and Development*. IEEE, vol. 80, no. 7, pages 1029–1058, 1992.

[Mullot 00] R. Mullot. *Interprétation de Documents Techniques et Cartographiques : Algorithmes et Systèmes*. Habilitation à Diriger les Recherches, Université de Rouen, France, 2000.

[Murray 96] J.D. Murray & W. Van Ryper. Encyclopedia of graphic file formats. Editions O'Reilly, 2 edition, ISBN : 1565921615, 1996.

[Nagy 00a] G. Nagy. *Twenty Years of Document Image Analysis in PAMI*. Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 1, pages 38–62, 2000.

[Nagy 00b] G. Nagy, T.A. Nartkerb & S.V. Rice. *Optical Character Recognition : An Illustrated Guide to the Frontier*. In Symposium on Electronic Imaging Science and Technology, volume 3967 of *Proc SPIE*, pages 58–69, 2000.

[Neumann 02] J. Neumann, H. Samet & A. Soffer. *Integration of Local and Global Shape Analysis for Logo Classification*. Pattern Recognition Letters (PRL), vol. 23, no. 12, pages 1449–1457, 2002.

[Ogier 00a] J.M. Ogier. *De l'Image au Document Technique, Problèmes d'Interprétation*. Habilitation à Diriger les Recherches, Université de Rouen, France, 2000.

[Ogier 00b] J.M. Ogier, R. Mullot, J. Labiche & Y. Lecourtier. *Semantic Coherency, the Basis of an Image Interpretation Device - Application to The Cadastral Map Interpretation*. Transactions on Systems, Man and Cybernetics, part B : Cybernetics (TSMCB), vol. 30, no. 2, pages 322–338, 2000.

[Ouf 01] R. Ouf. Le dynamisme du world wide web : Taille, croissance, visibilité, distribution et accessibilité de l'information. Rapport de Master, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Université Claude Bernard Lyon 1, France, 2001.

[Parent 99] R. Parent & N. Boulet. Les composantes d'un document electronique : Rapport du groupe de travail sur les métadonnées et les structures logiques. Collection en Ingénierie Documentaire : 2, 1999.

[Pédauque 03] R.T. Pédauque. *Document : Forme, Signe et Relation : une Reformulation du Numérique*. Rapport technique, Réseaux Thématiques Pluridisciplinaires Document (RTP-Doc), 2003.

[Pelletier 98] F. Pelletier. *Introduction à la GEIDE*. Le Magazine de l'Archivage et de la Gestion d'Informations (Mos), 1998.

[Plamondon 00] R. Plamondon & S.N. Srihari. *On-Line and Off-Line Handwriting Recognition : a Comprehensive Survey*. Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 1, pages 63–84, 2000.

[Ponte 97] M. Da Ponte. *ATA Graphics Working Group work : From CGM to Structured Graphics*. Rapport technique, Aerospatiale Matra Airbus, Toulouse, France, 1997.

[Popel 02] D.V. Popel. *Compact Graph Model of Handwritten Images : Integration into Authentification and Recognition*. In Conference on Structural and Syntactical Pattern Recognition (SSPR), volume 2396 of *Lecture Notes in Computer Science (LNCS)*, pages 272–280, 2002.

[Prax 01] J.Y. Prax. La gestion electronique documentaire. Editions Dunod, 2 edition, ISBN : 2-10-007891-7, 2001.

[Radivojevic 03] Z. Radivojevic, M. Cvetanovic & V. Milutinovic. *Data Mining : A Brief Overview and Recent IPSI Research*. Annals of Mathematics, Computing and Teleinformatics (AMCT), vol. 1, no. 1, pages 84–90, 2003.

[Rares 99] A. Rares, M.J.T. Reinders & E.A. Hendriks. *Image Interpretation Systems*. Rapport technique MCCWS 2.1.1.3.C, Delft University of Technology, Nederlands, 1999.

[Rigamonti 04] M. Rigamonti, K. Hadjar, D. Lalanne & R. Ingold. *Xed : un Outil pour l'Extraction et l'Analyse de Documents PDF*. In Colloque International Francophone sur l'Ecrit et le Document (CIFED), pages 85–90, 2004.

[Roisin 99] C. Roisin. *Documents Structurés Multimédias*. Habilitation à Diriger les Recherches, Institut National Polytechnique de Grenoble, France, 1999.

[Rounds 04] S. Rounds & R. Horton. *Electronic Records Management Guidelines*. Rapport technique, State Archives Department of Minnesota, USA, 2004.

[Saidali 04] Y. Saidali, S. Adam, J.M. Ogier, E. Trupin & J. Labiche. *Knowledge Representation and Acquisition for Engineering Document Analysis*. In Wokshop in Graphics Recognition (GREC), volume 3088 of *Lecture Notes in Computer Science (LNCS)*, pages 25–36, 2004.

[Segaller 98] S. Segaller. Nerds : a brief history of the internet. TV Books, ISBN : 1575000881, 1998.

[Seguela 01] P. Seguela. *Construction de Modèle de Connaissances par Analyse Linguistique de Relations Lexicales dans les Documents Techniques*. Thèse de Doctorat, Université de Toulouse III, France, 2001.

[Seta 04] F.J. Seta. *Charte Documentaire de Gallica*. Rapport technique, Département de la coopération Gallica, Bibliothèque Nationale de France (BNF), 2004.

[Shapiro 00] G. Piatetsky Shapiro. *Knowledge Discovery in Databases : 10 Years After*. SIGKDD Explorations Newsletter, vol. 1, no. 2, pages 32–34, 2000.

[Shih 89] C.C. Shih & R. Kasturi. *Extraction of Graphic Primitives from Images of Paper Based Line Drawings*. Machine Vision and Applications, vol. 2, pages 103–113, 1989.

[Smith 04] E.B. Smith & al. *Reports of the DAS02 Working Groups*. International Journal on Document Analysis and Recognition (IJDAR), vol. 36, no. 3, pages 211–217, 2004.

[Sánchez 02] G. Sánchez, J. Lladós & K. Tombre. *An Error-Correction Graph Grammar to Recognize Texture Symbols*. In Workshop on Graphics Recognition (GREC), volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 128–138, 2002.

[Sohm 97] J.C. Sohm. *Où Va la Micro-Informatique ?* In Notes Techniques du CERIG. 1997.

[Song 03] J. Song, M.R. Lyu, M. Cai, & S. Cai. *Graphic Object Recognition from Binary Images : a Survey and Performance Comparison.* http ://appsrv.cse.cuhk.edu.hk/ jqsong/, 2003.

[Spinrad 03] J.P. Spinrad. *Efficient Graph Representations.* In Fields Institute Monographs, volume 19. American Mathematical Society, ISBN : 0-8218-2815-0, 2003.

[Spring 95] M.B. Spring & J.D. Campbell. *The Document Processing Revolution.* In Meeting of the Reference and Information Services Section (RISS), 1995.

[Steve 99] L. Steve & G.C. Lee. *Accessibility and Distribution of Information on the Web.* Nature, vol. 400, pages 107–109, 1999.

[Stinckwich 00] S. Stinckwich. *Aspects Informatiques des Bases Documentaires Hétérogènes et Réparties.* Université de Caen, France, 2000.

[Suen 03] C.Y. Suen, S. Mori, S.H. Kim & C.H. Leung. *Analysis and Recognition of Asian Scripts - the State of the Art.* In International Conference on Document Analysis and Recognition (ICDAR), pages 866– 878, 2003.

[Tan 01] C.L. Tan & Z. Zhang. *Text Block Segmentation Using Pyramid Structure.* In Conference on Document Recognition and Retrieval, volume 4307 of *Proceedings of SPIE*, pages 297–306, 2001.

[Tang 91] Y.Y. Tang, C.Y. Suen, C.D. Yan & M. Cheriet. *Document Analysis and Understanding : a Brief Survey.* In International Conference on Document Analysis and Recognition (ICDAR), pages 17–31, 1991.

[Tang 96] Y.Y. Tang, S.W. Lee & C.Y. Suen. *Automatic Document Processing : A Survey.* Pattern Recognition (PR), vol. 29, no. 12, pages 1931–1952, 1996.

[Theune 03] M. Theune. *Natural Language Generation for Dialogue : System Survey.* Rapport technique TR-CTIT-03-22, CTIT, University of Twente, Netherlands, 2003.

[Thonnat 95] M. Thonnat & S. Moisan. *Knowledge Based System for Program Supervision.* In Workshop on Knowledge-Based systems for the (re)Use of Program Libraries, pages 4–8, 1995.

[Tombre 96] K. Tombre. *Structural and Syntactic Methods in Line Drawing Analysis : To Which Extent Do They Work ?* In Conference on Structural and Syntactical Pattern Recognition (SSPR), volume 1121 of *Lecture Notes in Computer Science (LNCS)*, pages 310–321, 1996.

[Tombre 02] K. Tombre, S. Tabbone, L. Pelissier, B. Lamiroy & P. Dosch. *Text/Graphics Separation Revisited.* In Workshop on Document Analysis Systems (DAS), volume 2423 of *Lecture Notes in Computer Science (LNCS)*, pages 200–211, 2002.

[Tou 74] J.T. Tou & R.C. Gonzalez. Pattern recognition, principles. Addison Wesley Publisher, ISBN : 0-201-07587-3, 1974.

[Trupin 01] E. Trupin, J.M. Ogier, S. Adam & P. Héroux. *Navigation into Technical Documents.* In Workshop on Graphics Recognition (GREC), pages 27–34, 2001.

[Trupin 03] E. Trupin. *De la Reconnaissance Automatique d'Images de Documents.* Habilitation à Diriger les Recherches, Université de Rouen, France, 2003.

[Ullman 89] J.D. Ullman. Principles of database and knowledge base systems, volume 1. Computer Sciences Press, ISBN : 0716781581, 1989.

[Valveny 02] E. Valveny & B. Lamiroy. *Scan to XML : Automatic Generation of Browsable Technical Documents.* In International Conference on Pattern Recognition (ICPR), pages 1051–4651, 2002.

[Valveny 04] E. Valveny & P. Dosch. *Symbol Recognition Contest : A Synthesis.* In Workshop on Graphics Recognition (GREC), volume 3088 of *Lecture Notes in Computer Science (LNCS)*, pages 368–386, 2004.

[Vijjappu 01] L. Vijjappu, A.H. Tan & C.L. Tan. *Web Structure Analysis for Information Mining.* In Workshop on Web Document Analysis (WDA), 2001.

[Villard 00] L. Villard, C. Roisin & N. Layaïda. *An XML-Based Multimedia Document Processing Model for Content Adaptation.* In Conference on Digital Documents and Electronic Publishing (DDEP), volume 2023 of *Lecture Notes in Computer Science (LNCS)*, pages 104–119, 2000.

[Vinciarelli 02] A. Vinciarelli. *A Survey on Off-Line Cursive Word Recognition.* Pattern Recognition (PR), vol. 35, no. 7, pages 1443–1446, 2002.

[Walker 00] F.L. Walker & G.R. Thoma. *Web-Based Document Image Processing.* In Conference on Internet Imaging, volume 3964 of *Proc. SPIE*, pages 268–277, 2000.

[Wang 00] S. Wang, W. Gao, J. Li, T. Huang & H. Xie. *Web Mining and Knowledge Discovery of Usage Patterns.* In Conference on Web Age Information Management (WAIM), 2000.

[Wenyin 04] L. Wenyin. *On-Line Graphics Recognition : State-of-the-Art.* In Workshop on Graphics Recognition (GREC), volume 3088 of *Lecture Notes in Computer Science (LNCS)*, pages 289–302, 2004.

[Wille 96] N.E. Wille. *Fast Automated Conversion with Integrated Tools (FACIT).* Rapport technique 1-5, Roskilde University, Denmark, 1996.

[Zachary 00] I. Zachary & L. Ying. *XML Query Languages in Practice : an Evaluation.* In Conference on Web Age Information Management (WAIM), pages 29–40, 2000.

Table des figures

Liste des tableaux

Table des matières