Représentation de documents par niveaux superposés et compression à base de symboles

Applications à l'archivage et à l'impression

Luc Vincent

Xerox Palo Alto Research Center

Xerox Palo Alto Research Center parc

Plan de la présentation

- Motivations
- Compression a base de symboles
- Representation de documents par niveaux superposees
- Segmentation en niveaux et extraction d'objets graphiques
- Applications:
  - Logiciel *Pagis Pro*
  - *Impression rapide*

Representer des documents comme des images numeriques

- Avantages immediats:
  - Mode de representation peut etre utilise pour des documents scannes comme pour des documents de source electronique
  - Utilisable pour documents anciens et nouveaux
  - Simplicite de la representation: impression, et visualisation aisees
- Limitations:
  - Images de documents peuvent etre des enormes fichiers
  - Manque de structure de ces representations
- Limitations peuvent etre levees par l'utilisation de techniques de segmentation multi-niveau et de compression a base de symboles

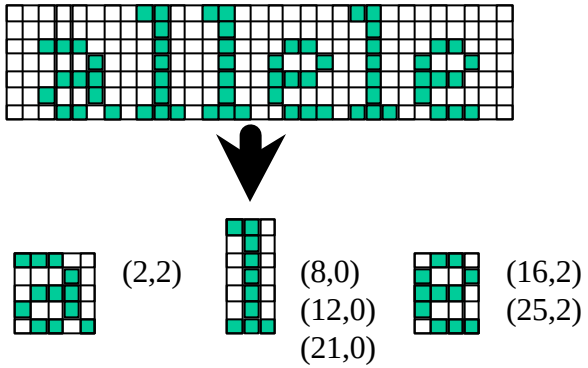
Compression a base de symboles

- Noms dans la litterature: “token-based compression” ou “tokenization”
- Les premieres methodes de compression a base de symboles remontent aux annees 70 a AT&T Bell Labs.
- Compression *avec pertes* qui genere typiquement des fichiers de 3 a 10 fois plus petits que CCITT Groupe 4.
- **Principes**:
  - Identification des symboles (caracteres) qui se repetent dans une image de document
  - Representation: dictionnaire de symboles + liste de positions, codes de maniere appropriee

Compression a base de symboles (cont.)

- Utilisation de techniques de *pattern matching* et de *clustering* pour identifier des classes de formes (tokens)

- Compression:
  - ***Dictionaire de symboles:*** liste des formes presentes
  - ***Liste de positionss:*** là ou chaque symbole apparait



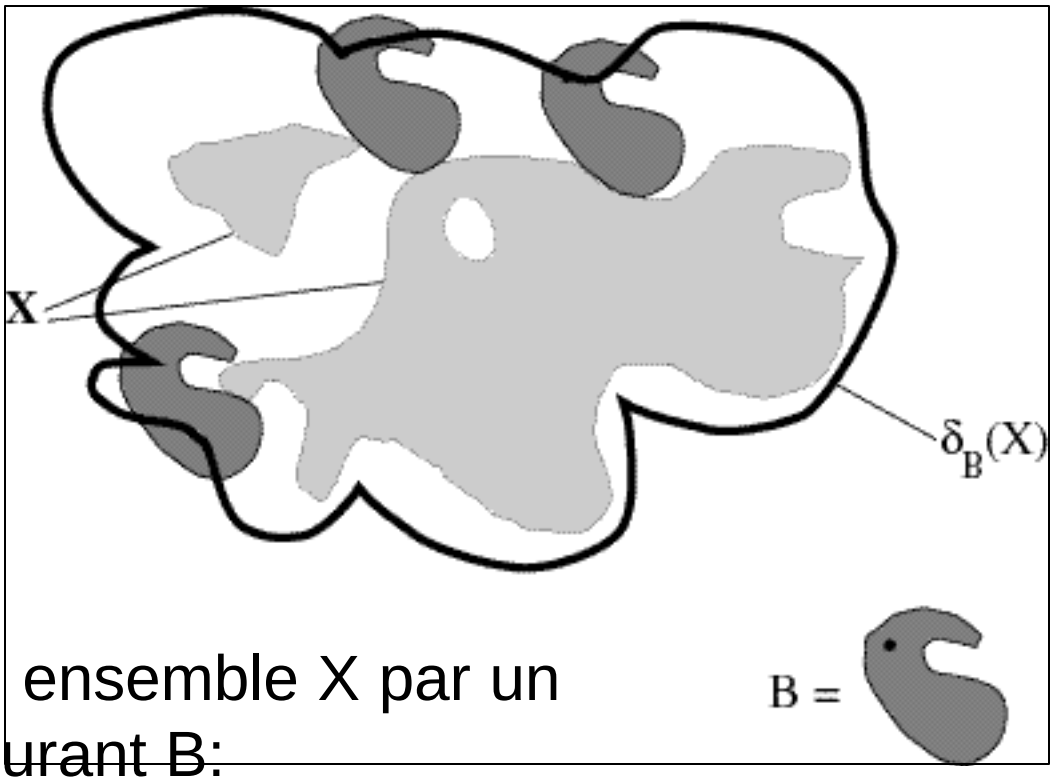
the computations, it also allows us to solve the accuracy problems encountered by most of the algorithms reviewed in Section II-D. First, the labeling of the catchment basins automatically avoids such traps as that of Fig. 7. Now, in order to get perfectly located watershed arcs, the successive geodesic SKIZ involved in the process have to be as good as possible. The first thing to notice is that, according to the discrete distance associated with the underlying grid, the set of pixels which are equidistant to two given connected components may well not be a line, but a very thick area. This is illustrated by Fig. 9. (Recall that the distance between two pixels is equal to the minimal number of grid edges to cross to go from one to the other.) Consequently, some simplistic rules in the computation of the geodesic SKIZ's could result in unwanted thick watershed areas. More precisely, suppose that the plateaus at elevation h are currently being flooded,

Xerox Palo Alto Research Center parc

Mise en correspondance de symboles

- Symboles: composantes connexes de l'image
- Mise en correspondance basee sur la distance de *Hausdorff*:
  - Chaque nouveau symbole ***C*** est compare aux symboles ***S*** deja presents dans le dictionnaire
  - ***C*** est dilate et aligne avec ***S***
  - ***S*** est dilate et aligne avec ***C***
  - Les differences sont analysees. C est soit mis en correspondance avec un symbole du dictionnaire, soit une nouvelle classe est cree
- La methode est optimisee pour eviter les substitutions:
  - Les petites differences sur le pourtour des caracteres sont en general dues a la numerisation et sont ignorees
  - Les petites differences a l'interieur des symboles sont beaucoup plus significatives.

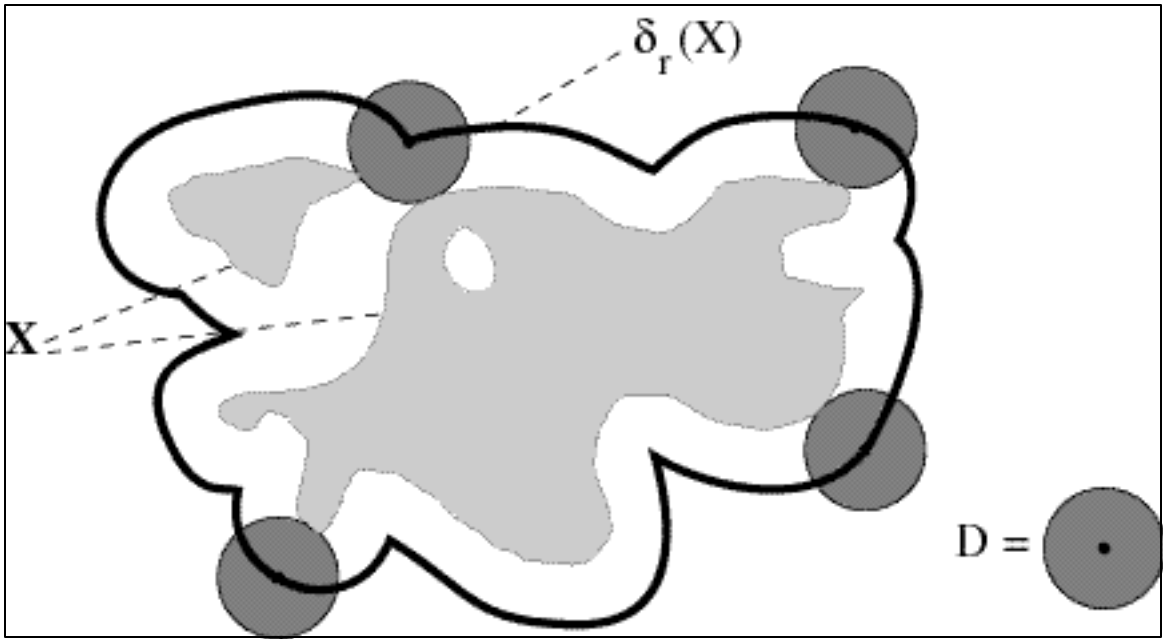
Dilatation et distance de Hausdorff



Dilatation d'un ensemble X par un element structurant B:

$\delta_B(X) = \{y \mid \exists\, b \in B, y + b \in X\}$

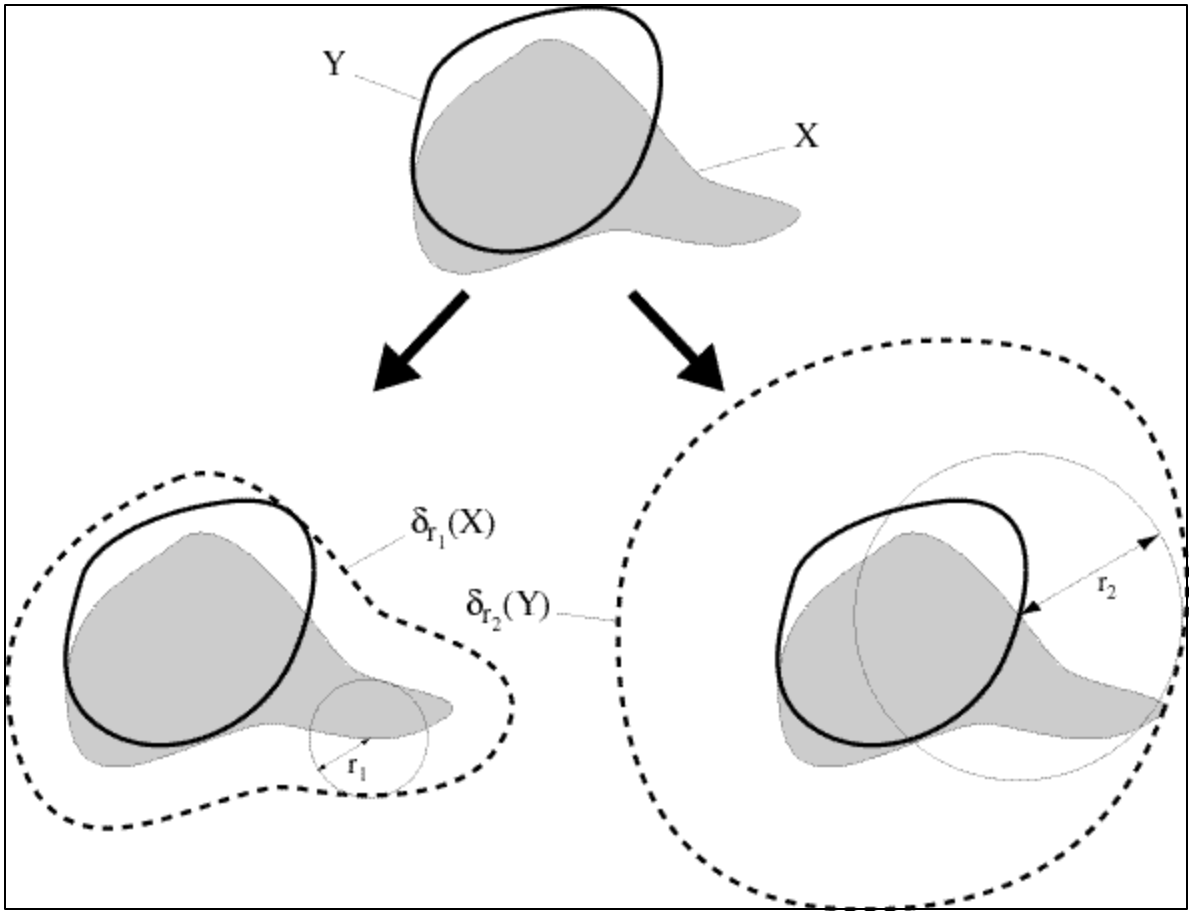
Dilatation par un disque



Dilatation de l'ensemble X par un disque D de rayon r:

$\delta_r(X) = \{y \mid dist\ (y,X) \leq r\}$

Distance de Hausdorff



$d_H(X,Y) = \max(\inf\{r \mid \delta_r(X) \subseteq Y\}, \inf\{r \mid \delta_r(Y) \subseteq X\})$

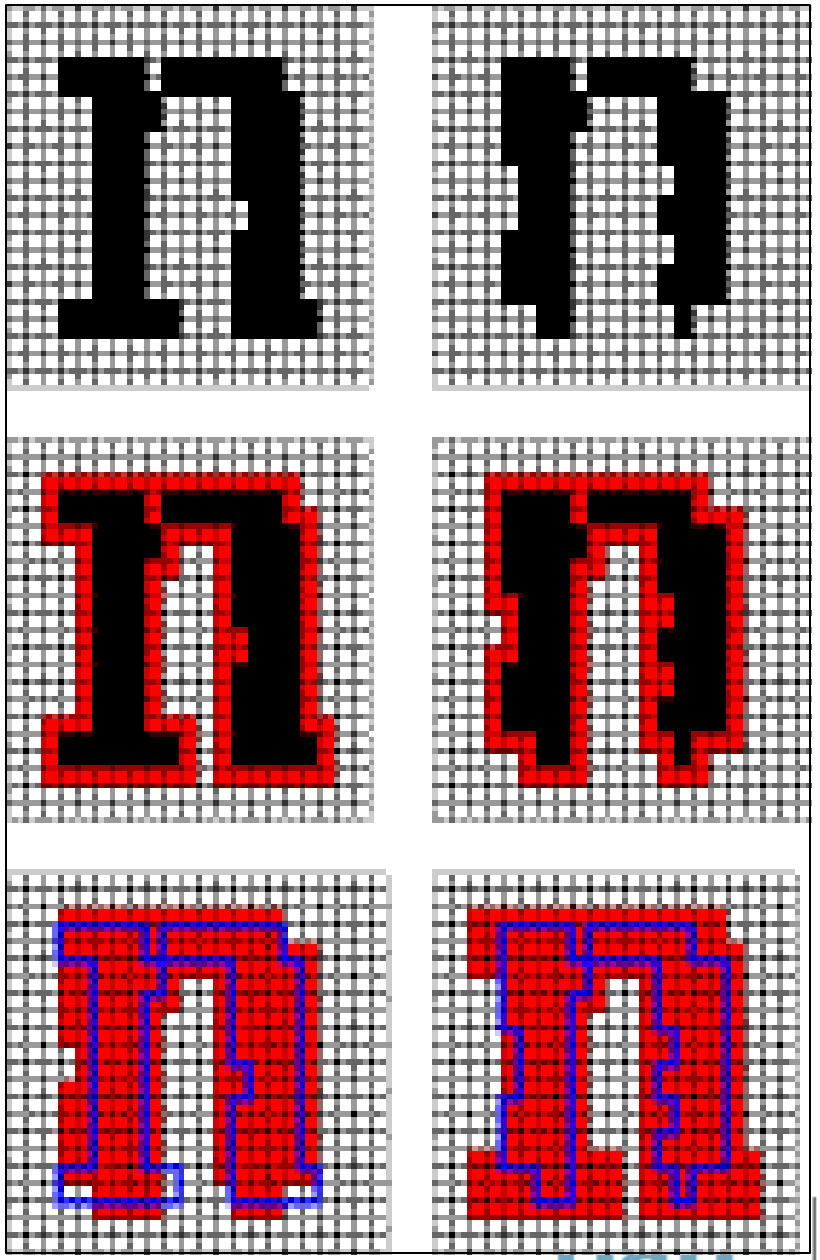
$= \max(r_1, r_2) = r_2$

Xerox Palo Alto Research Center parc

Mise en correspondance de symboles

Dilatation elementaire

Alignement et analyse des pixels peripheriques



Avantages de Hausdorff

- Dilatation elementaire requise peut etre calculee au depart, directement sur l'image originale
- Temps d'execution tres rapide
- Permet d'analyser efficacement les differences entre symboles de forme voisine
- Distance de Hausdorff pure est insuffisante. Une analyse plus fine est requise pour distinguer:
  - Certains serifs et sans-serif
  - Points et virgules
  - Accents
  - Etc.

Exemple d'extraction de symboles

Image binaire originale

Symboles se repetant

Symboles n'apparaissant qu'une fois

Premiere apparition des symboles se repetant

Encapsulation de la representation a base de symboles

- Un ou plusieurs dictionnaires de symboles:
  - Groupement des symboles par classes de hauteur
  - Tri par largeur
  - Compression de chaque classe par CCITT Groupe 4 (fax)
- Des tableaux de positions:
  - Liste de positions (X,Y), avec un numero de symbole associe a chaque position
  - Tri des positions par ordre lexicographique en fonction de X et Y
  - Coordonnees X et Y representees par codage de Huffman differentiel (ou codage arithmetique)
- Representation conforme au standard JBIG2

Representation de documents multi-pages

- Differents types de dictionnaires:
  - Dictionnaires applicables a une page donnee
  - Dictionnaires globaux
  - Dictionnaires applicable a un groupe de pages donne
- Dictionnaires globaux:
  - Prennent en compte la redondance des symboles d'une page a l'autre
  - Permettent d'obtenir des taux de compression tres eleves, jusqu'a 10 fois superieurs a CCITT Group 4

Documents Multi-Pages (cont.)

- Documents peuvent avoir plusieurs dictionnaires globaux:
  - Dictionnaires incrementaux permettent une compression efficace des longs documents
  - Possibilite de mettre plusieurs documents bout a bout sans avoir a decompresser puis recompresser la representation
  - Avantageux pour les longs documents avec changements de police de caractere
- Danger des dictionnaires globaux: necessitent que le decodeur garde en memoire tous les symboles globaux, ce qui n'est pas toujours possible. Remedes:
  - Remise a zero periodique des dictionnaires globaux
  - Association d'un "rayon d'action" a chaque dictionnaire

Avantages des representations a base de symboles

- Taux de compression tres eleve
- Independant de la langue et des types de caracteres
- Pixelisation extremement efficace
  - ➔ Impression et visualisation
- Possibilite d'integrer de l'information textuelle a la representation
  - Une identite peut etre assignee a chaque symbole
  - Traitements speciaux pour caracteres non connexes (i, j, etc.)
  - Possible pour sources electroniques (texte est disponible) et scannees (apres OCR)
  - Representation extremement compacte
- Representation tres efficace de texte couleur

Exemple 1: article scanne

Secure Spread Spectrum Watermarking for Multimedia

Ingemar J. Cox†, Joe Kilian‡, Tom Leighton‡ and Talal Shamoon†*

Abstract

1 Introduction

- Document de 33 pages
- Scanne a 300 ppp (binaire)
- Taille des fichiers:
  - TIFF, groupe 4: 3.5 MB
  - JBIG2 tokens: 0.67 MB

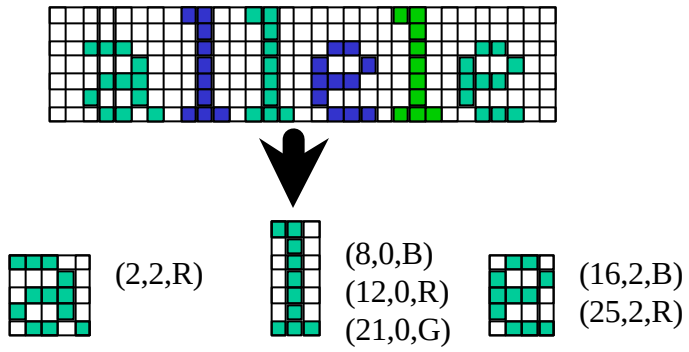
Xerox Palo Alto Research Center parc

Exemple 2: document chinois

- Source: document PostScript, 60 pages, nombreuses figures.
- PostScript:
  - 67 MB, 7MB avec compression gzip
  - Pixelisation prend plusieurs minutes
- Compression a base de symboles:
  - 1MB a 300 ppp, 2MB a 600 ppp
  - Pixelisation: < 10s sur un PC standard (plusieurs centaines de pages par minutes)

Xerox Palo Alto Research Center parc

Symboles couleur



- Caracteres couleur sont en general uniformes (sauf mises en page speciales)
- Graphiques sont en general formes d'une collection de regions de couleur uniforme
- Associe une couleur a chaque instance de symbole:
  - Chaque instance devient: X, Y, ID, couleur
  - Etiquettes couleur peuvent etre representees par codage de Huffman puis codage par plages

    ➔ Uniquement quelques octets par page!

Xerox Palo Alto Research Center parc

Representation multi-niveaux de documents couleur

- Probleme: les documents couleurs (ou a niveaux de gris) scannes ne peuvent pas etre compresses de maniere adequate avec des techniques standard, type JPEG.
- Solutions:
  - *Utilisation de la resolution et nombre de bits/pixel qui est la plus approprie a chaque type de donnee present:*
    - ➔Typiquement: de 300 a 600 ppp pour le texte, et 100 ppp pour les images et photos
  - *Utilisation de techniques de compression differentes pour chaque type de donnee:*
    - ➔Necessite une segmentation prealable de la page
  - *Decomposition du texte et des graphiques en un niveau binaire haute resolution et un niveau couleur basse resolution*
    - ➔Representation et segmentation multi-niveaux

Exemple: page de magazine scannee

FYI

Easy Sale

The last thing you want, after spending hours with a fussy customer choosing the specs of a customized product, is to find that those options are unavailable. There goes the sale—and probably the customer as well.

SalesVision, Client Server Technologies Inc.'s product-configuration software tool, helps to eliminate such situations. Designed for remote sales forces, value-added resellers and telemarketers, SalesVision provides immediate and valid product-configuration information.

The Windows-based product enables a salesperson to enter product options such as make, model, color and style into a laptop computer; it then validates the information and verifies that the order can be filled. Once verification occurs, SalesVision sends the information via modem to the factory for manufacture.

SalesVision can tie directly into a company's credit, accounting and financial records, so remote workers can send or receive data for credit approvals and financing during the product selection and configuration stage. In addition, SalesVision can be customized to access data from a company's marketing and engineering departments, giving salespeople access to explanations of features and analyses of competing products.

Pricing can be obtained by contacting Client Server Technologies in Schaumburg, Ill., at 708-397-7300.

Cost-Effective Conferencing

Mumblers and soft-spoken executives with a tight telecommunications budget don't need to spend an arm and a leg for an audioconferencing unit that will pick up their voices.

Coherent Communications Systems Corp. says its hands-free audio conferencing unit ConferenceMaster Elite will offer exceptional performance on even the poorest of telephone lines for less than the traditional audio-conferencing duplex unit. Coherent, based in Leesburg, Va., uses a new state-of-the-art 16-bit codec for improved sound clarity and 360-degree sound coverage for call participants as far as 15 feet away from the speaker. Automatic gain control allows every participant, even those facing away from the ConferenceMaster Elite, the ability to speak naturally and still be heard clearly from anywhere in a room.

Where conventional conference systems would distort voices or break up a signal when speakers raise their voices or laugh loudly (causing a higher than normal transmission signal level), the ConferenceMaster Elite offers additional signal capacity to handle the call.

ConferenceMaster Elite is a portable, ergonomically designed unit that plugs into a standard analog telephone jack and electrical outlet. ConferenceMaster Elite is available for $995. For more information, call 800-443-0726.

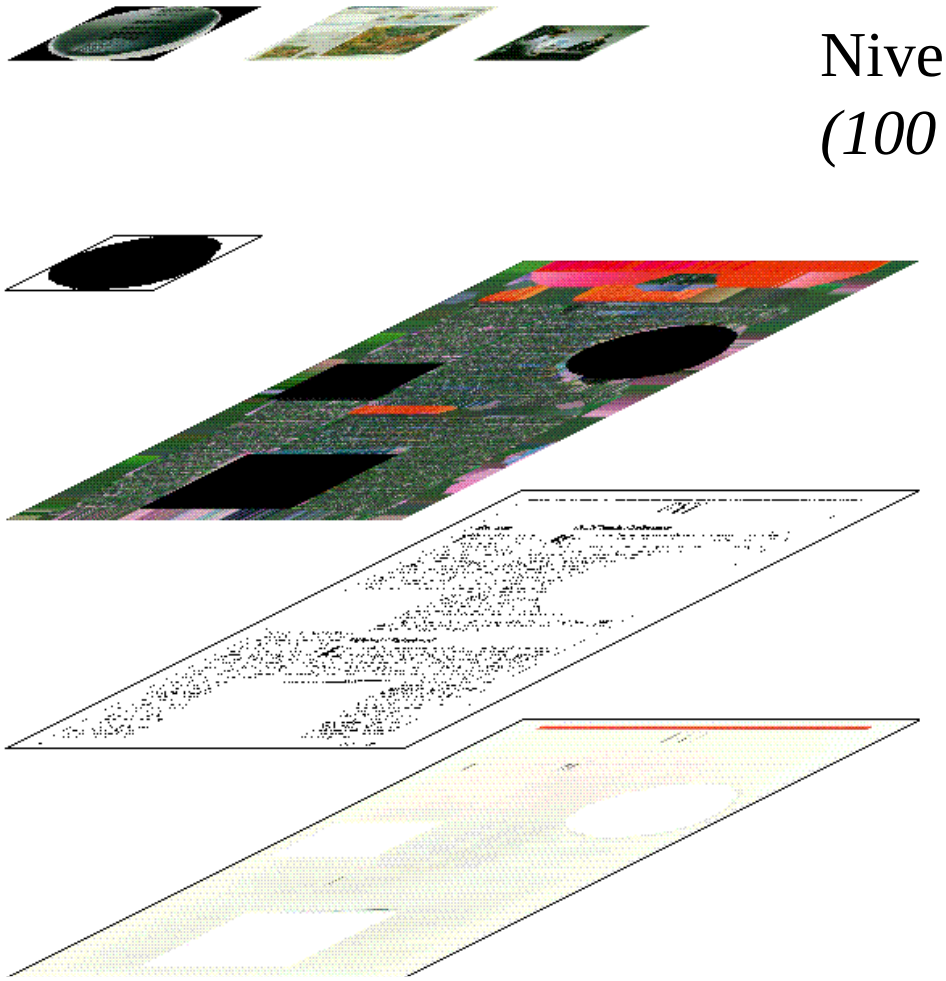
Multimedia Management

Every organizational nut knows that one of the biggest problems with the surge in multimedia is how to store all those CDs, antiquated LPs, video tapes, old home movies, books on tape and more. Corporations are also coming to grips with the organizational quagmire that threatens to overwhelm their multimedia operations. Enter MediaWay Inc.'s family of multimedia database-management and -cataloging applications called MediaDB.

Organizations such as newspapers, real estate brokerages and high-tech marketing firms must deal with the technical challenges of storing and archiving video clips, news clips, sound clips, photos and other multimedia materials that require managing massive amounts of memory-hungry information and unstructured relationships between items. MediaDB, a database management system that can modify multimedia information in an object-oriented environment as easily as it can traditional text and numbers, is designed from the ground up specifically to tackle multimedia organizational challenges.

MediaDB DBMS version 2.0 ranges in price from $2,500 for a five-user license on Novell NetWare to $200,000 for a 2,000-user license on Unix servers. For more information, call MediaWay in Santa Clara, Calif., at 408-748-7400.

Decomposition en niveaux



Niveaux #4 et #5: images
(100 ppp, JPEG et CCITT G4)

Niveau #3: couleur du texte *(50 ppp, JPEG)*

Niveau #2: texte
(300 ppp, JBIG2)

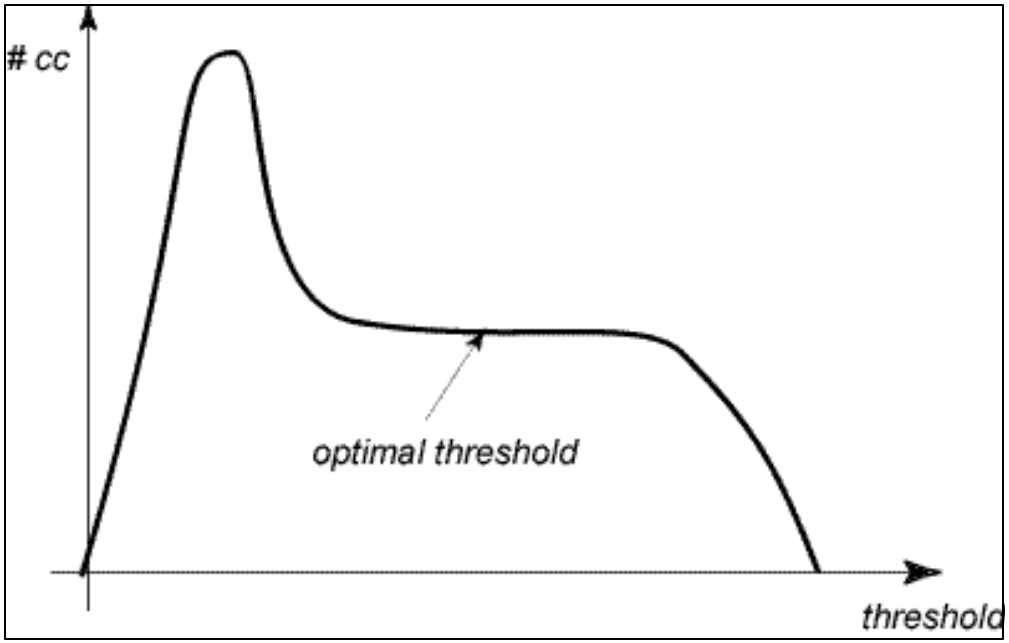
Niveau #1: fond
(75 ppp, JPEG)

Methode de segmentation en niveaux et objets

- Binarisation par seuillage adaptatif
  - Extraction d'un masque binaire du texte et des images
- Niveau du bas (fond) obtenu par reduction de la resolution et "bouchage des trous" laisses par le masque binaire extrait precedemment
- Segmentation du texte et des images dans le masque binaire
- Post-processing: creation et compression des differents niveaux

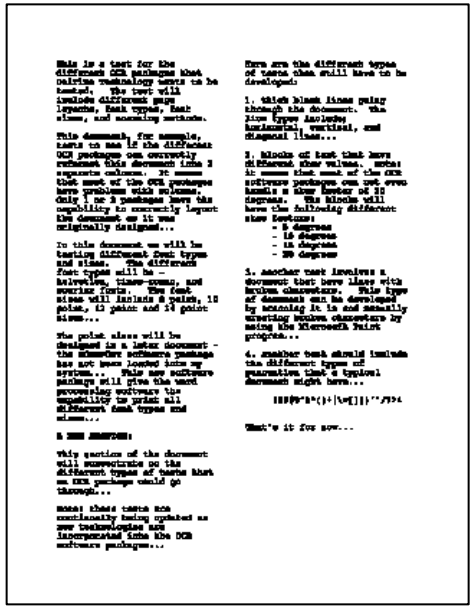
Seuillage adaptatif

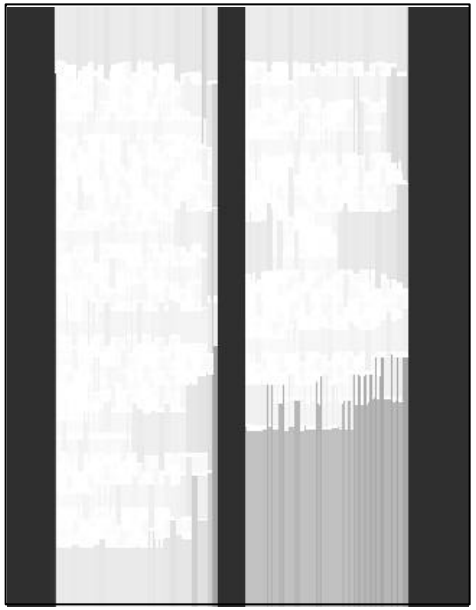
Base sur un histogramme des composantes connexes:



Segmentation texte/image

Base sur la detection et l'analyse hierarchique des "goulets" verticaux et horizontaux de l'image:





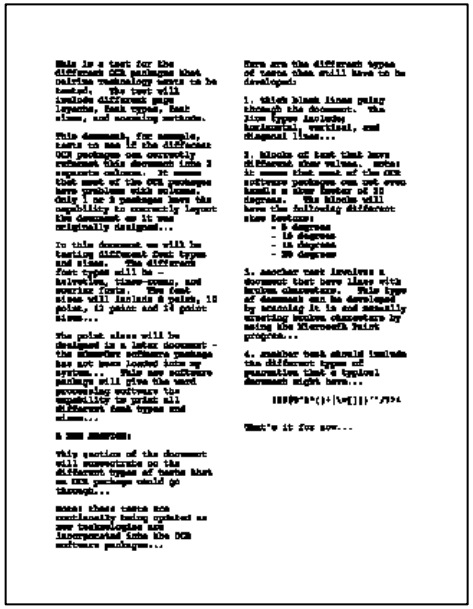
Xerox Palo Alto Research Center parc

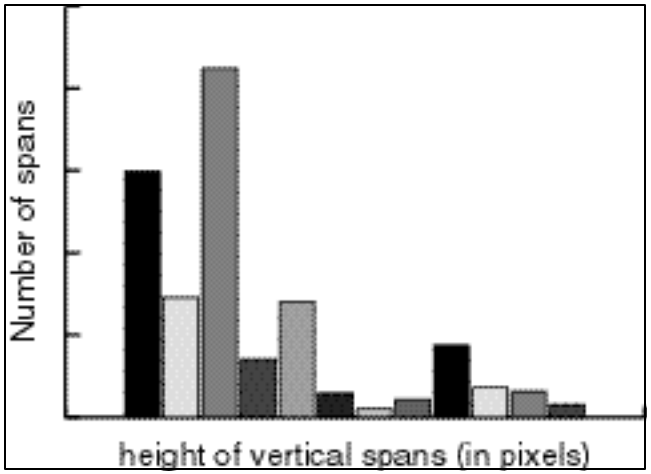
Autres elements cle

- Impossible de traiter efficacement une vaste variete de documents avec des parametres fixes
- La plupart des parametres de la segmentation sont derives d'une analyse globale de l'image
  - Histogrammes des niveaux de gris (couleur)
  - Histogramme des composantes connexes
  - Granulometries verticales et horizontales
  - Etc.

Exemple: estimation (grossiere) de l'espacement entre lignes

Methode: granulometrie verticale par fermeture, a resolution reduite (histogramme de longueur des zones blanches verticales)



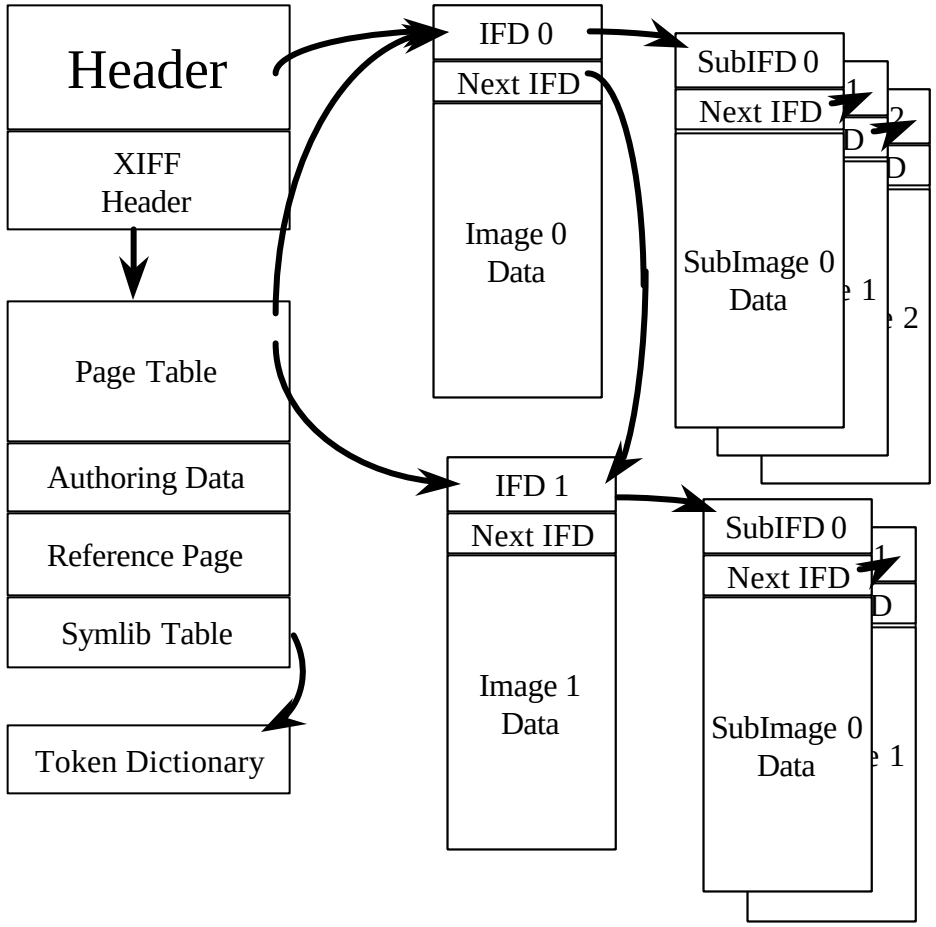


Xerox Palo Alto Research Center
parc

Formats de fichier pour ce type de representation

- DJVU (AT&T/LizardTech, Yann LeCun): non standard, mais sur le point d'etre ouvert (open source)
- XIFF (Scansoft): largement utilise par les utilisateurs de Pagis Pro, mais va etre remplace par TIFF-FX
- TIFF-FX
  - Nouveau standard dont Xerox a ete un des plus ardents supporter
  - Recemment adopte comme un standard par l'organisation IETF
  - En train d'etre adopte par de nombreux partenaires industriels. Nouveau standard de facto?
- Note: PostScript et PDF permettent egalement ce type de representation, mais de maniere sous-optimale

Structure d'un fichier XIFF ou TIFF-FX



Resume des avantages pour le scanning

- Tres hauts taux de compression, mais haute qualite
- Avantageux pour les grosses archives de documents:
  - Plus compacte que d'autres representations
  - Plus efficace pour les systemes distribues ou la bande passante est le facteur limitant
  - Pixelisation tres rapide (bien superieur a PDF, par exemple)
- Ouvre des possibilites pour le scanning et l'archivage personnel de documents

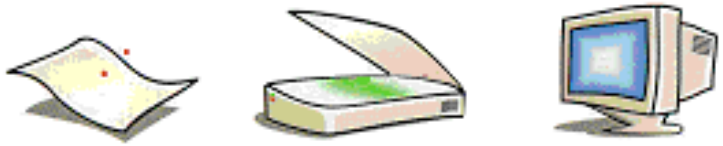
  → Un des elements cle du logiciel Pagis Pro

Avantages pour documents electroniques

- Un driver d'impression peut etre utiliser pour produire des documents TIFF-FX a partir de sources electroniques
- TIFF-FX produit a partir de documents electroniques:
  - ➔ Apparence garantie
  - ➔ Jamais de probleme de polices de caracteres (ce qui n'est pas vrai avec PDF ou PostScript)
  - ➔ Permet d'utiliser un format de fichier uniforme entre documents scannes et electroniques
- Pixelisation est tres efficace et demande tres peu de memoire
  - ➔ A long terme, TIFF-FX peut remplacer ou complementer PostScript ou PCL comme langage d'impression

Logiciel Pagis Pro

Extraits de la page web `http://www.pagis.com`:



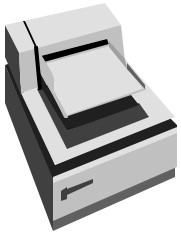
- The Best Way to Scan, Organize and Use Color Documents
- "Together with Office 97... and Pagis Pro... you've got the ultimate business application suite."
  - » 7/97 PC Computing - The Ultimate Office
- Pagis Pro97 is a *fully-featured scanning application* that allows you to scan documents into your Windows desktop. With a color, gray scale or binary scanner, you can easily *scan documents into your PC, then file, copy, print, send or use them* with your application.

Xerox Palo Alto Research Center parc

Pagis Pro (cont.)

- Ne supporte que Windows
- Pagis Pro 1.0 date de 1996. La plus recente version, 4.0/Millenium, date de Avril 2000
- Vendu en lot avec TextBridge Pro OCR
- Technologies sous-jacentes:
  - Segmentation
  - Compression
  - Traitement d'images de documents (amelioration, alignement, rotation, etc.)
  - Indexation par OCR
- Cout: entre $49 et $99, y compris TextBridge Pro
- Supporte tous les scanners les plus courants

Serie d'outils Pagis



Scanner

Telecopie

Email, y compris attachements

Documents web

Appareils photos digitals

- outil de scanning
- Editeur
- moteur d'indexation
- engin de recherche

- Archivage
- Recherche
- Edition
- Annotation
- Email
- Fax
- Impression
- Copie
- Publication sur le web

Xerox Palo Alto Research Center parc

Scenario d'utilisation #1

Scanning et distribution

- Utilisation du Pagis scan tool pour scanner un article et le convertir en un fichier XIFF (typiquement < 150k)
- Utilisation de l'Editeur pour nettoyer et annoter le fichier XIFF (souligner des regions, ajouter des commentaires, etc.)
- Envoi du fichier XIFF file par email
- Destinataire peut utiliser un outil de visualisation gratuit, disponible sur le site Pagis (meme modele que pour PDF)

Scenario d'utilisation #2

Scanning et publication sur le Web

- Scanning d'un document Pagis scan tool, et conversion en fichier XIFF
- Conversion automatique en fichier HTML si necessaire (par le biais du moteur TextBridge)
- Document peut aussi etre publie directement comme un fichier XIFF, et visualise a l'aide d'un plug-in pour Netscape ou Internet Explorer

Scenario d'utilisation #3

Scanning, archivage, recherche, utilisation

- Utilisation du Pagis scan tool pour scanner tout document interessant et a conserver
- Si necessaire, utilisation de l' Editeur pour nettoyer et annoter le fichier XIFF genere
- Le Pagis Update Tool indexe automatiquement tous les fichiers XIFF present (ainsi que tous les formats standard type Word, WordPerfect, PowerPoint, TIFF, etc.)
- Utilisation du Pagis Search Tool pour rechercher des documents dans les archives personnelles ('grep' perfectionne)
- Visualisation, impression, distribution, etc, des documents recuperes

Xerox Palo Alto Research Center parc

Avantages pour l'impression (1)

- Pixelisation extremement efficace:
  - ➔ Operations de base: decompression de JPEG et bit-blits
  - ➔ le moteur d'impression peut etre utilise a sa vitesse maximale avec un controleur bon marche
- Experiences recentes avec un chip StrongARM 233MHz, sous VxWorks, avec du code tres peu optimise:
  - Jusqu'a 700 pages par minutes en noir et blanc, 300 ppp
  - Environ 50 pages couleur par minute, a 300 ppp
- PostScript, en revanche, est un veritable langage et peut prendre un temps arbitraire a imprimer
  - Impression TIFF-FX facilement 10 fois plus rapide en moyenne que l'impression PostScript

Avantages pour l'impression (2)

- Reduction des couts:
  - Controleur bon marche
  - Demande moins de memoire dans l'imprimante
  - Pas besoin de license PostScript!
- Apparence garantie
- Nouveau paradigme: la plupart des calculs sont a present fait par l'ordinateur hote au lieu de l'imprimante
  - Utilisation plus efficace des imprimantes partagees par des groupes d'utilisateurs
  - Ordinateurs plus puissant $\Rightarrow$impression plus rapide

Conclusions

- Taux de compression de 200 ou plus
- Haute qualite
- Format ideal pour la visualisation, l'archivage, l'impression, et le web.
- Structure de niveaux flexible, peut etre adaptee a differents types d'applications
- Pas besoin d'OCR pour obtenir des fichiers de tailles raisonnable a partir de documents scannes
  - Representation basee sur l'image
  - OCR utilise pour l'indexation
  - "Best of both worlds"?