AG BVH

3 décembre 2012

Introduction au Word Spotting

Frédéric Rayar Partha Pratim Roy Jean-Yves Ramel

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

LI Laboratoire d'Informatique EA 2101

POLYTECH TOURS

Introduction

- Préservation et accès au patrimoine culturel
  - Numérisation
  - Transcription (OCR)

- Difficultés à retranscrire les documents historiques

- Alternative : le Repérage de mots (*Word Spotting*)
  - On ne cherche plus à reconnaitre chaque caractère
  - Listes de régions similaires à la requête

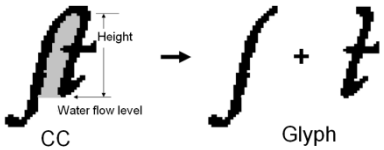
Word Spotting

- Domaine de recherche actif :
  - Documents modernes, anciens
  - Documents manuscrits, imprimés
  - Texte ou graphique simple (logo, ...)
  - Avec ou sans segmentation (lignes, mots)

Approche proposée

- Objectifs
  - Robustesse aux documents historiques
  - Indépendance à l'alphabet
  - Rapidité du temps de réponse à une requête

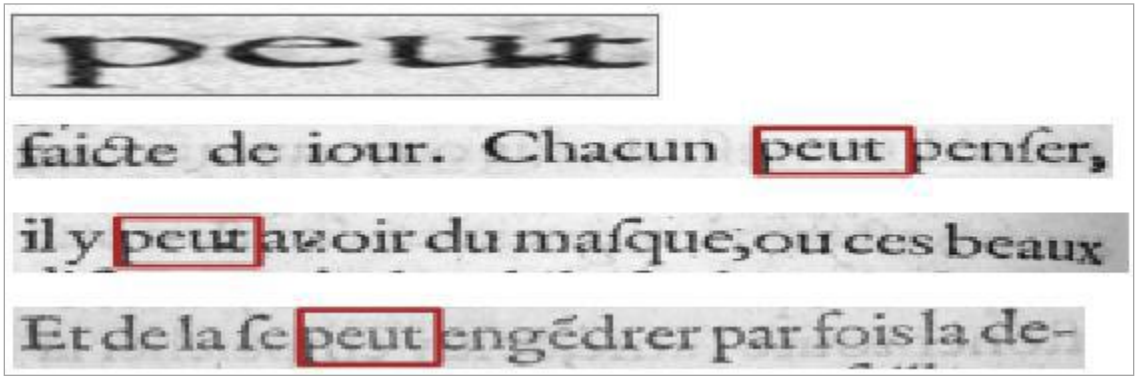
- Principes exploités
  - Segmentation en ligne (AGORA)
  - Considération de primitives de caractères
  - Indexation hors ligne (dictionnaire de formes)
  - Recherche reportée au niveau du texte



Résultats expérimentaux

- Centre d'Etudes Supérieures de la Renaissance

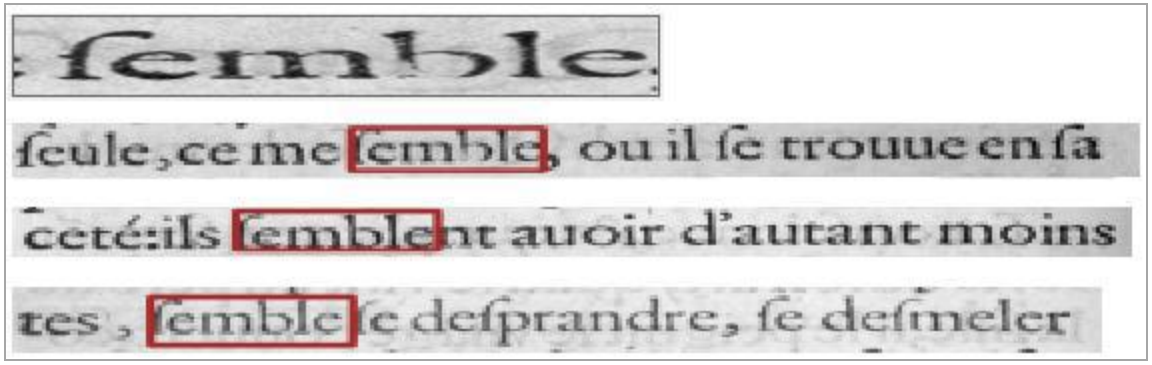
- « *Essais – Livre I* » de Montaigne (1580)


peut

faicte de iour. Chacun peut penſer,

il y peut auoir du maſque, ou ces beaux

Et de la ſe peut engẽdrer par fois la de-


ſemble.

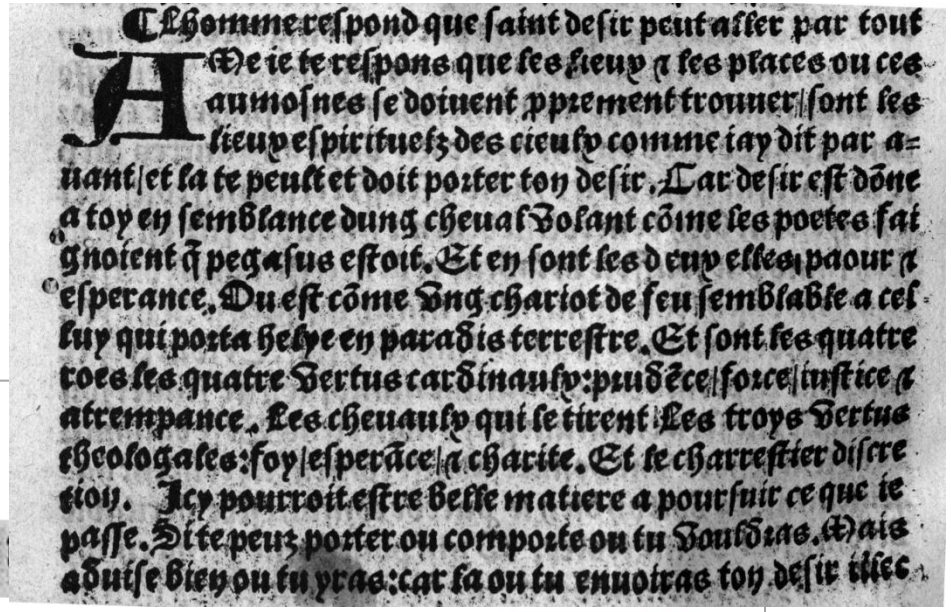
ſeule, ce me ſemble, ou il ſe trouue en ſa

cité: ils ſemblent auoir d'autant moins

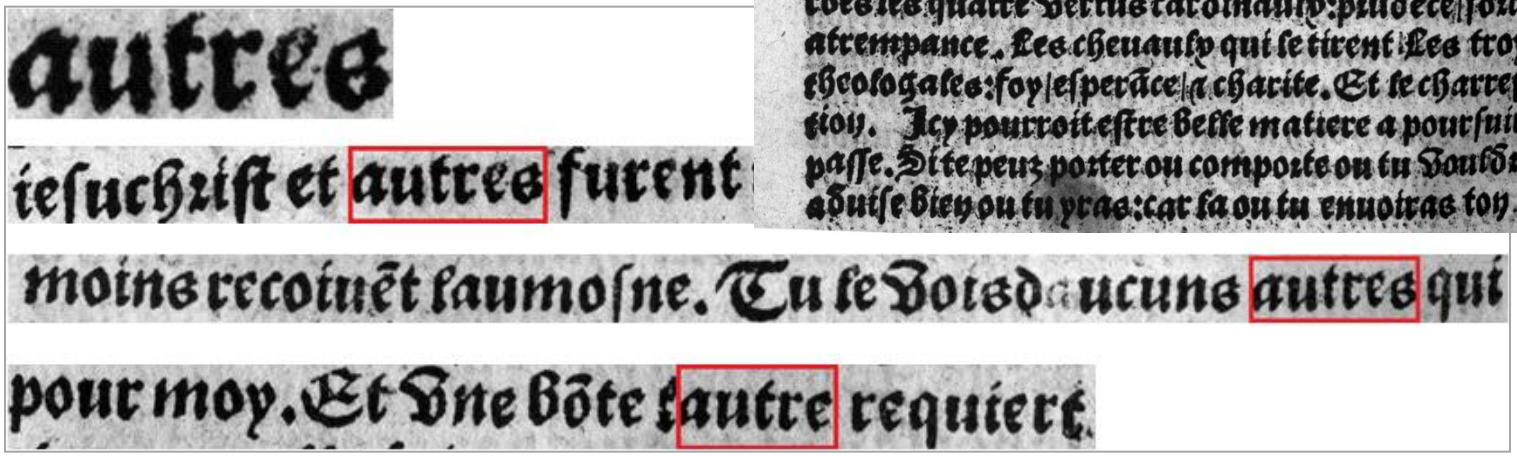
tes, ſemble ſe deſprandre, ſe deſmeler

Résultats expérimentaux

- « La mendicité spirituelle » de Gerson (1501)



L'homme respond que saint desir peut aller par tout
Ame ie te respons que les lieux / les places ou ces
aumosnes se doiuent pprement trouuer / sont les
lieux espirituelz des cieulx comme iay dit par a-
uant / et la te peult et doit porter toy desir. Car desir est dõne
a toy en semblance dung cheual volant cõme les poetes fai
gnoient q̃ pegasus estoit. Et en sont les deux elles / paour /
esperance. Du est cõme vng chariot de feu semblable a cel
luy qui porta helye en paradis terrestre. Et sont les quatre
roes les quatre vertus cardinaulx: prudẽce / force / iustice /
atrempance. Les cheuaulx qui le tirent les trois vertus
theologales: foy / esperãce / charite. Et le charrestier dis cre
tion. Icy pourroit estre belle matiere a poursuir ce que ie
passe. Si te peux porter ou comporte ou tu vouldras. Mais
aduise bien ou tu yras: car la ou tu enuoiras toy desir illec



autres

iesuchrist et autres furent

moins recoiuẽt laumosne. Tu le vois d ucuns autres qui

pour moy. Et vne bõte lautre requiert.

Résultats expérimentaux

- Journaux en (a) hindi et (b) bengali
- Rongo (*Entretiens de Confucius* en japonais) (1533)

(a)

भूगर्भ

भूगर्भ खदानों से कोयला निकालने

रहा है और ओपेन कास्ट से भूगर्भ

के पास भूगर्भ से कोयला निकालने

(b)

বন্‌ধ

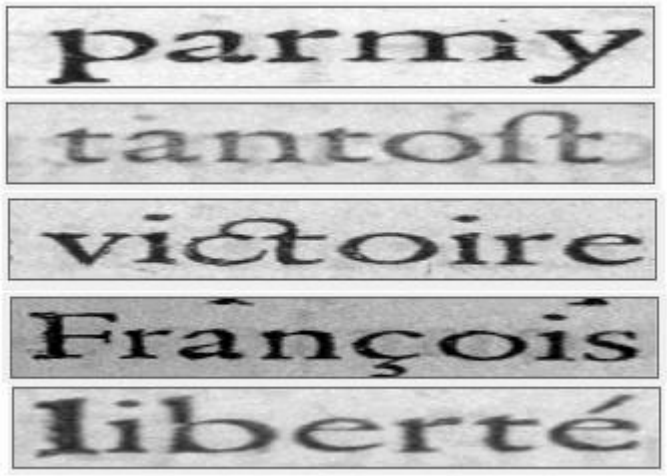
নিজস্ব প্রতিনিধি, কলকাতা: এ রাজ্যে ঘন ঘন বন্‌ধ

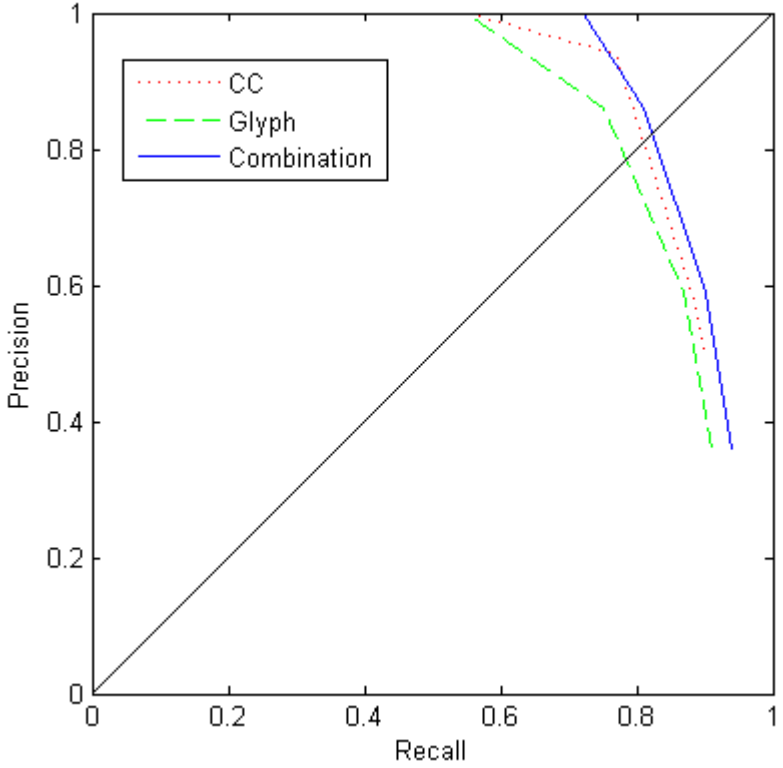
কর্মীর সম্মতি নিয়ে বন্‌ধ ডাকা এবং চিকিৎসক,

বলেন, বন্‌ধের ফলে শিল্পের ক্ষতি হচ্ছে বলে যে প্রচার

Résultats expérimentaux

- « Essais – Livre I » de Montaigne (1580)
  - Vérité terrain fournie par le CESR
  - Sous-ensemble de 78 pages
  - 1579 lignes indexées





- 72% de résultats pertinents avec 100% de précision

Résultats expérimentaux

- PC avec un processeur de type i5 à 2.53 GHz et 4G de RAM
- Tests menés sur le sous-ensemble de 78 pages
- La recherche a été effectuée plusieurs fois, et un temps de réponse moyen a été calculé

Image requête	Temps (s)
(a) 	1,2
(b) 	2,5
(c) 	4,5

- Temps de réponse de 2 secondes en moyenne

Projet ReNom