Améliorer la classification de documents par combinaison de descripteurs visuels et textuels

Olivier Augereau

Gestform
38 Rue François Arago
33700 Mérignac

oaugereau@gestform.com
@oaugereau

Nicholas Journet, Jean-Philippe Domenger

LaBRI
351, cours de la Libération
F-33405 Talence Cedex

CIFED 2014 – 19 Mars 2014

Plan

Introduction

1. Classification texte et image
    1. Sac de mots
    2. Sac de mots visuels
    3. Fusion multimodales
2. Tests sur bases réelles
    1. Résultats
    2. Retours industriels

Conclusion

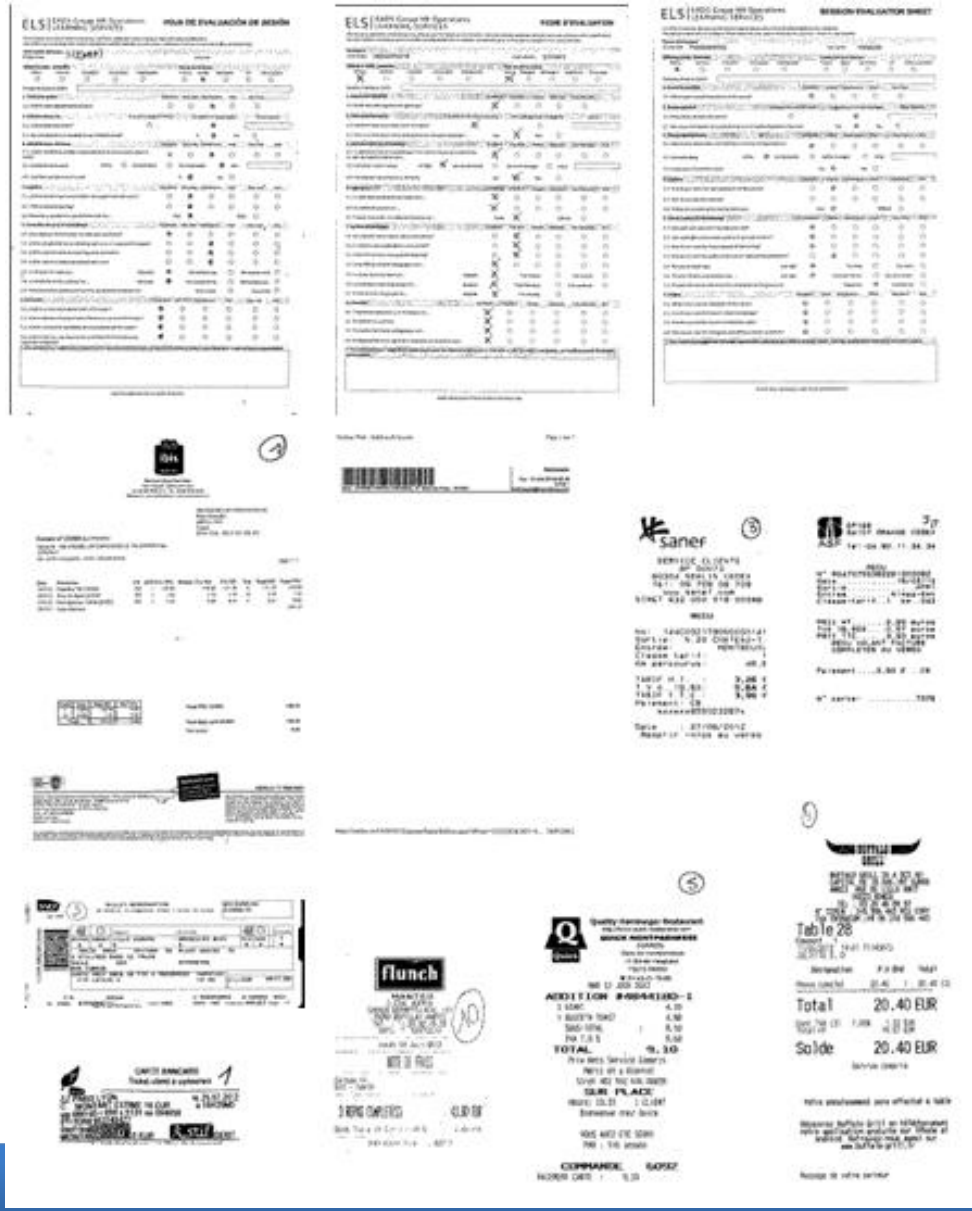
Introduction

Objectifs :

- Classifier des documents dans une base conséquente et variée
- Grande précision

Contexte industriel :

- Déformations géométriques
- Informations manquantes
- Qualités variables

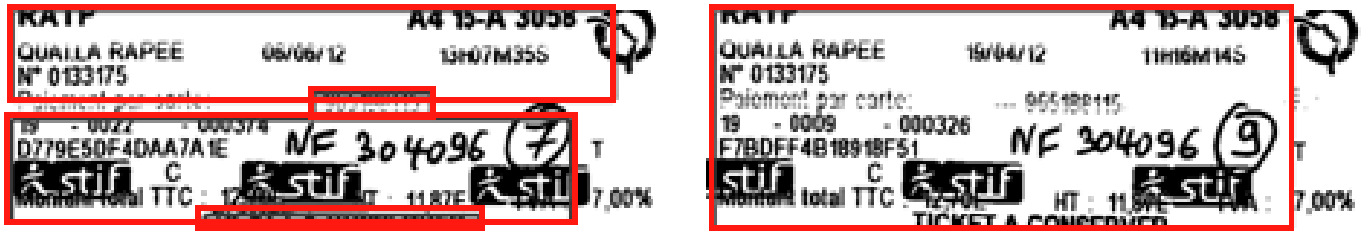


Introduction

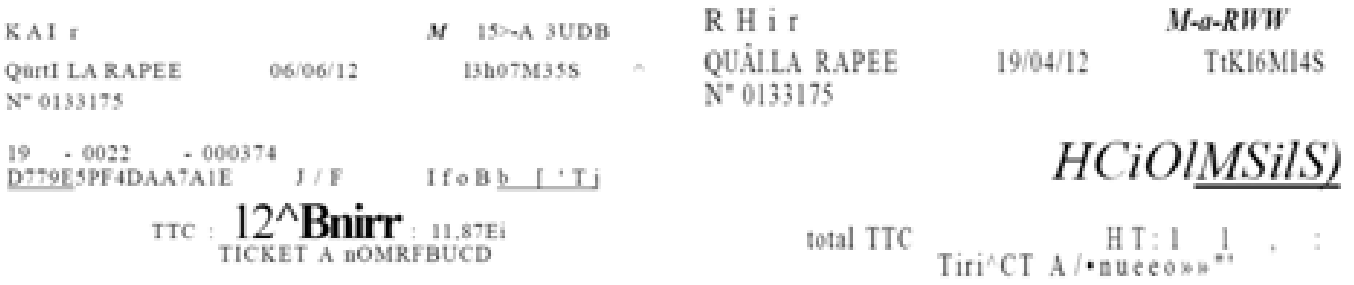
Classification d'image de document : texte, image, structure.

image

structure



texte



Introduction

⇒ Combiner les représentations texte et image [Rusiñol *et al.* 2012]

Texte : n-grams, sacs de mots (BoW), Latent Semantic Analysis (LSA)…
Image : textures, composantes connexes, points d'intérêts, sacs de mots visuels (BoVW), Spatial Pyramid Matching (SPM) …
Combinaison : fusion tardive, précoce, apprentissage, …

Apports :

- tester les BoVW sur les images de documents
- combinaison des BoW et BoVW.

⇒ Amélioration des performances globales par combinaison

Classification texte et image

Sacs de mots

4 étapes :

1. Prétraitement du texte : filtrage, stop words, lemmatisation…

2. Création d'un dictionnaire à partir des K mots les plus fréquents (K=1000)

3. Chaque document est décrit par la fréquence d'apparition des mots qu'il contient qui sont compris dans le dictionnaire => vecteur

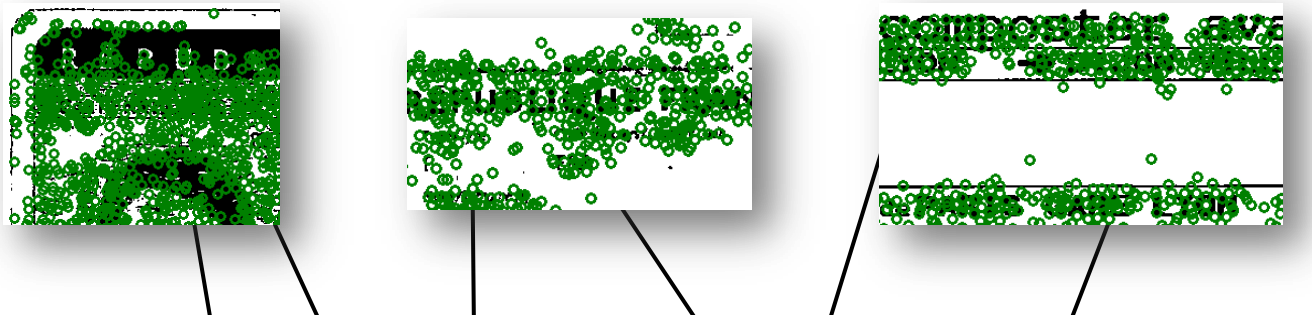
4. Utilisation des SVM pour apprentissage et classification supervisée.
   (SVM à noyau radial ou polynomial > Bayes, Rocchio, C4.5 et k-NN avec distances cosinus [Joachims98])

[Joachims98] Text categorization with support vector machines: Learning with many relevant features

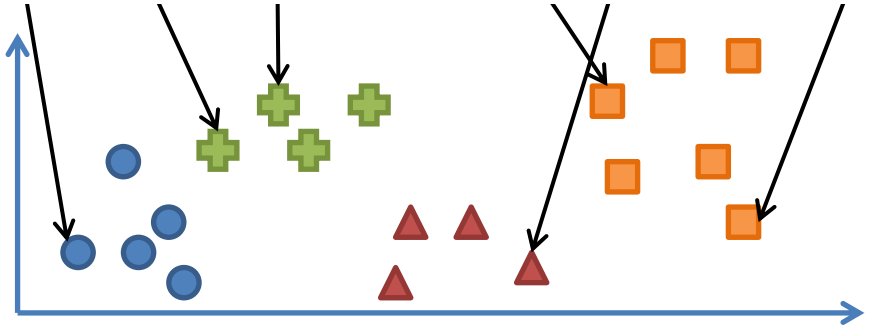
Classification texte et image

Sacs de mots visuels

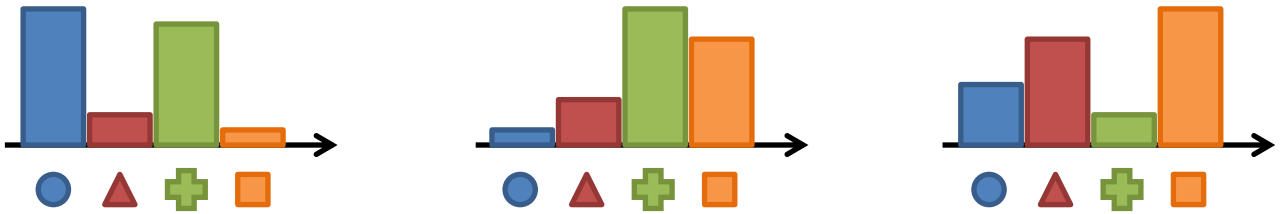
1. Extraction des points d'intérêt



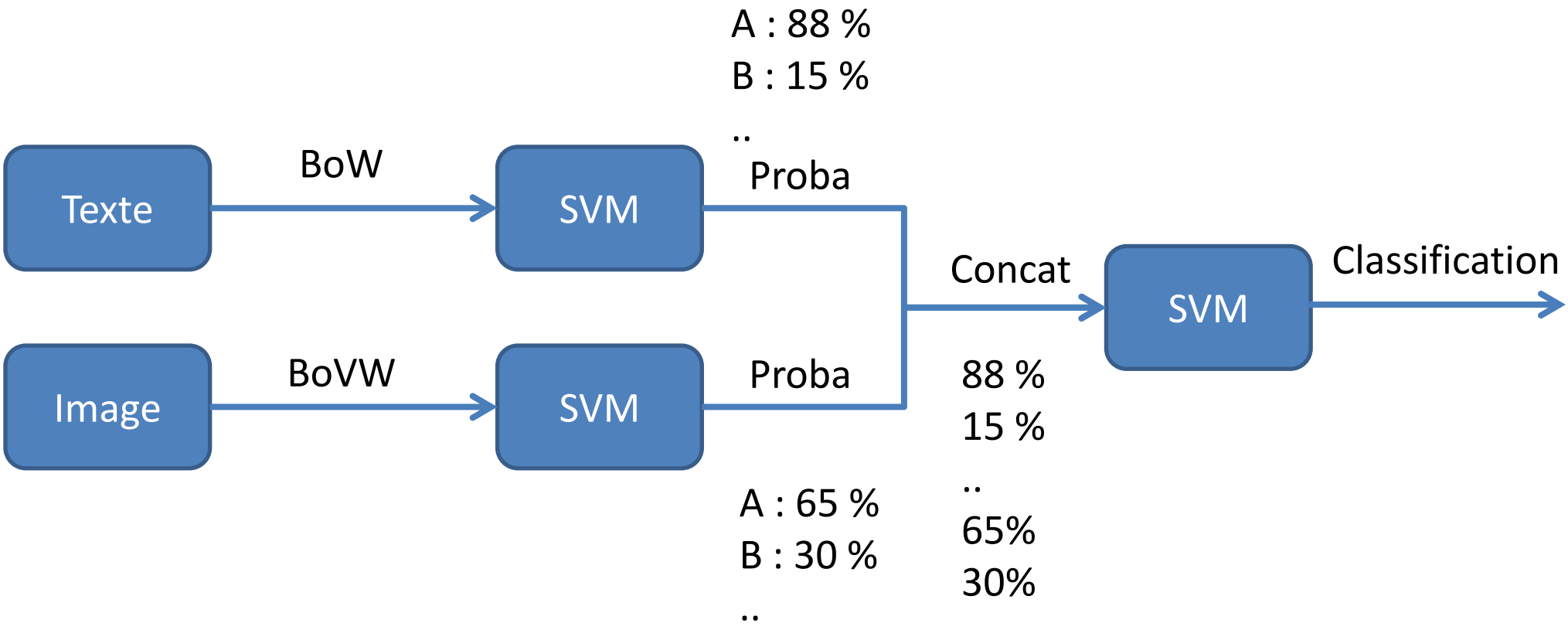
2. Partitionnement et création du dictionnaire visuel



3. Vecteurs de description



Classification texte et image



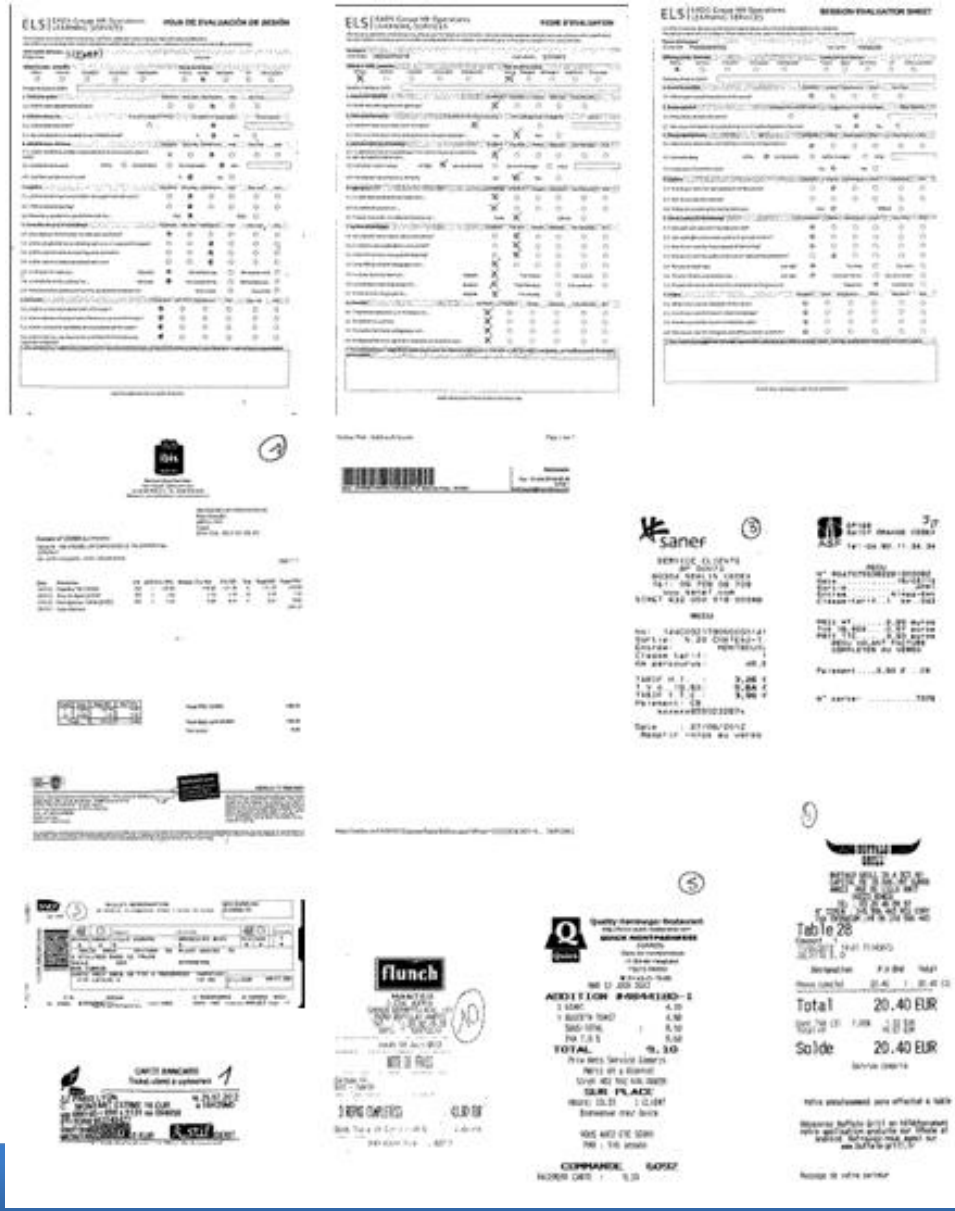
Tests sur bases réelles

Base de test :

- Documents numérisés par Gestform.
- 1985 pages choisies aléatoirement.
- 300 dpi, binarisation par le scanner.
- 12 classes de documents.

Classe	Sncf	ElsEs	ElsFr	ElsUk	Ibis	Lot
Nb doc	194	299	99	430	41	670

Asf	Sanef	Buffalo	Flunch	Quick	Ratp
31	22	13	4	12	110



Tests sur bases réelles

	Rappel	Précision
BoVW	87,1 %	86,9 %
BoW	98,2 %	97,7 %
Fusion Borda-Count	89,9 %	90,2 %
Fusion SVM	**99,4 %**	**98,6 %**

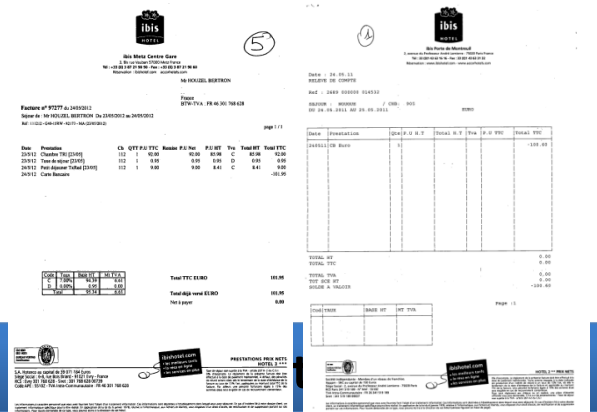
Rangs classifieur 1	Rangs classifieur 2	Rangs Borda
A	C	A
B	A	C
C	B	B

Tests sur bases réelles

		Rappel			Précision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	**0,994**	0,977	0,870	**0,986**

Tests sur bases réelles

		Rappel			Précision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	**0,994**	0,977	0,870	**0,986**



Tests sur bases réelles

		Rappel			Précision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	**0,994**	0,977	0,870	**0,986**

Tests sur bases réelles

- Paramétrage simple.
  - 200 > k (k-means et dictionnaire) > 2000
  - Seuillage de la matrice Hessienne ≈ 500 (doc. binaires)

- BoW, BoVW ou combinaison ?

  - BoW si OCR fiable, documents « propres »
  - BoVW si peu de texte et problème d'OCR
  - Combinaison si documents inconnus ou contenu variable

Conclusion

Conclusion :

Sur une base industrielle (notes de frais, formulaires, documents RH) :

- Les BoVW donnent des résultats intéressants.
- Combinaison BoW et BoVW > BoW.

Perspectives :

- Tester d'autres descripteurs texte/image et combinaisons.
- Rendre public des bases d'images (anonymes) et/ou de descripteurs.

Améliorer la classification de documents par combinaison de descripteurs visuels et textuels

Merci de votre attention

oaugereau@gestform.com
@oaugereau

Présentation à retrouver sur : www.olivier-augereau.com