Traitement automatique des langues pour l'indexation d'images

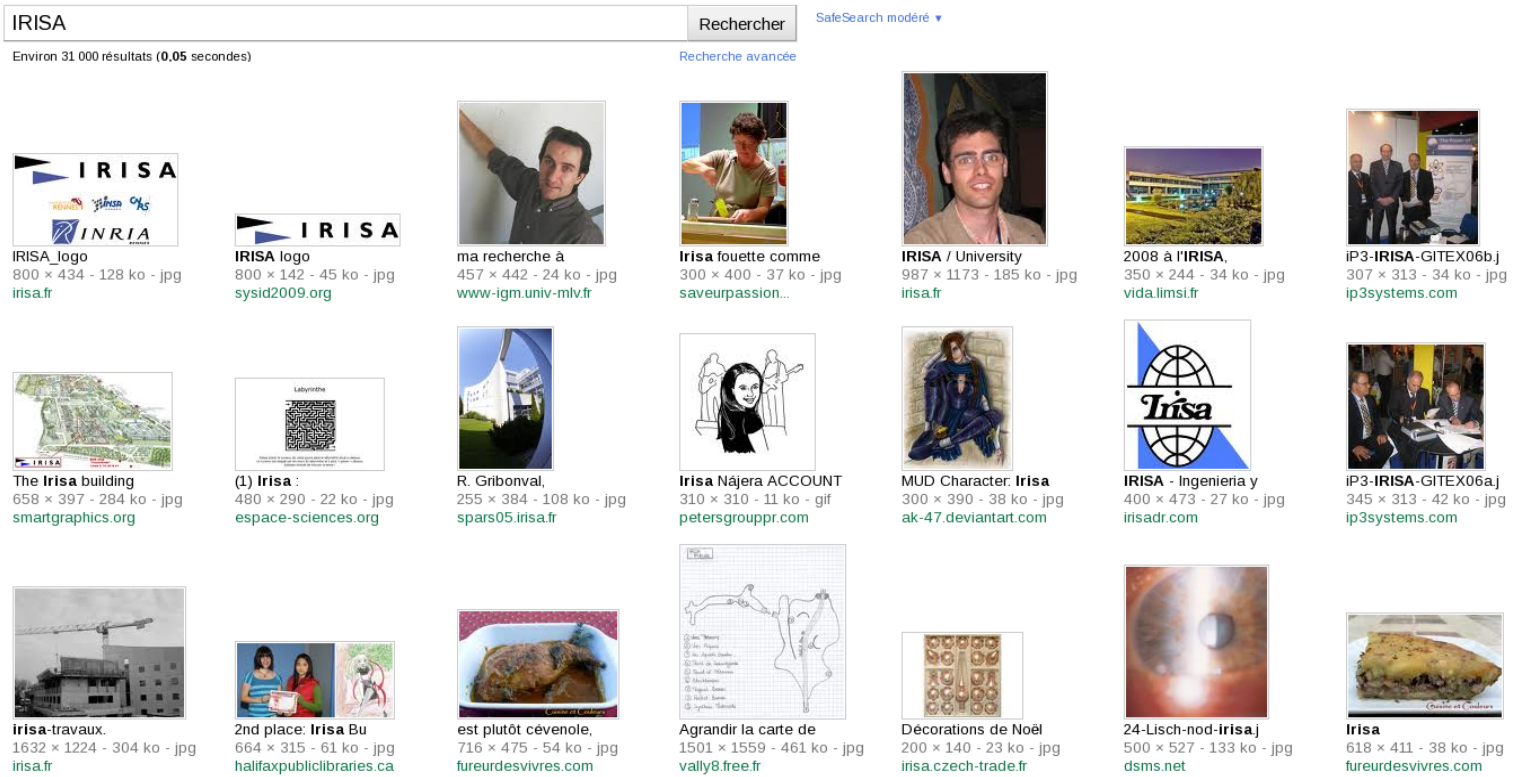
Pierre Tirilly

Équipe-Projet TEXMEX - IRISA

Jury :

Mohand Boughanem	Univ. Paul Sabatier / IRIT	Rapporteur
Philippe Mulhem	CNRS / LIG	Rapporteur
Patrick Gallinari	UPMC / LIP6	Examinateur
Christophe Garcia	FT R&D / Orange Labs	Examinateur
Patrick Gros	INRIA Rennes	Directeur de thèse
Vincent Claveau	CNRS / IRISA	Co-directeur de thèse

Recherche d'images : contexte

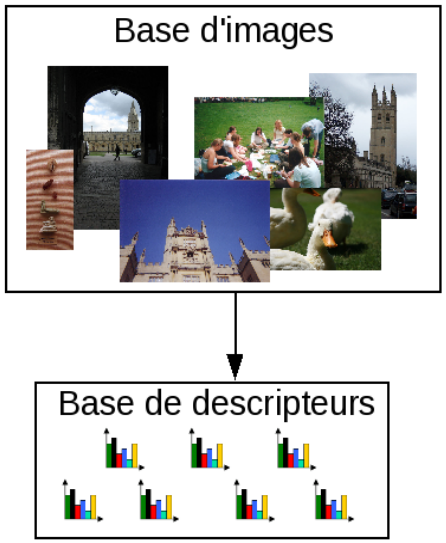


Bases d'images de plus en plus importantes

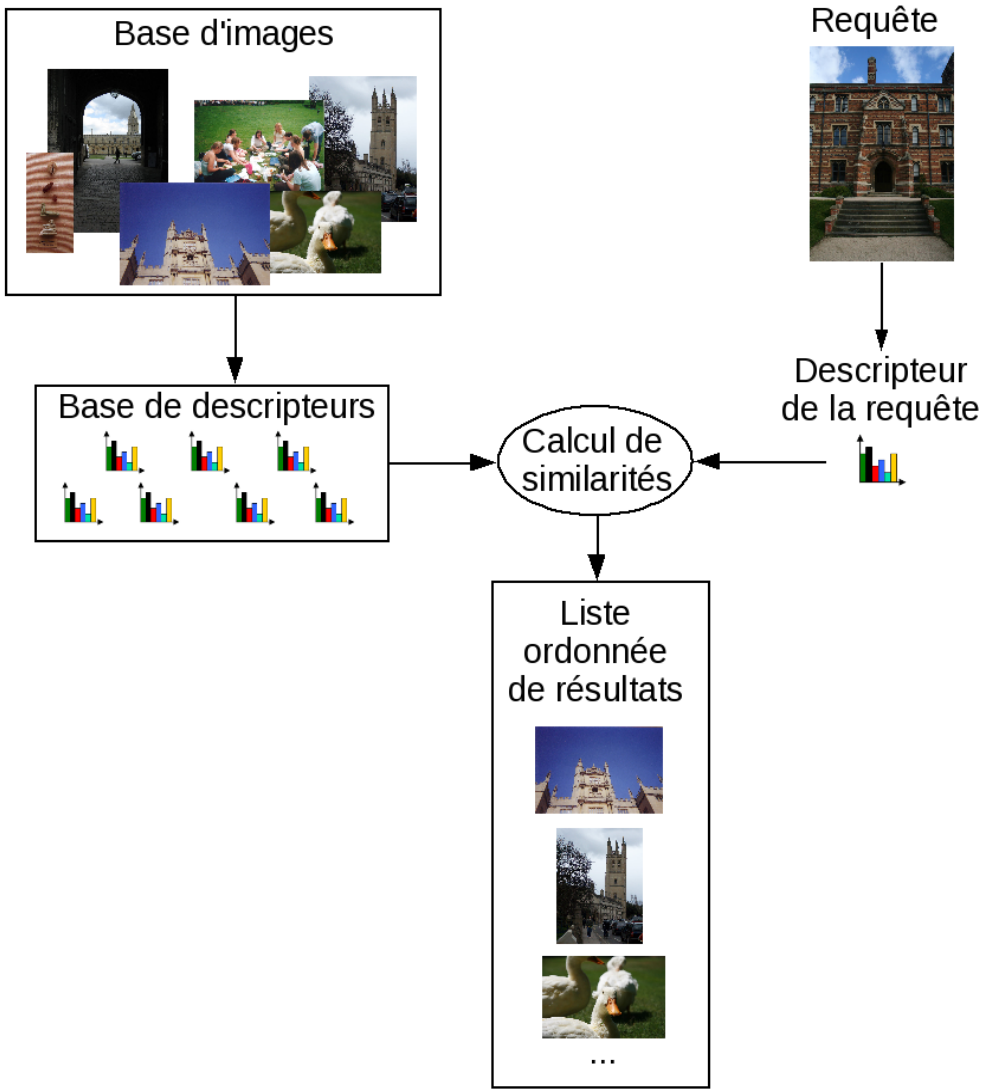
- → Index Google : > 1 milliard
- → FlickR : 4 milliards
- → Facebook : 10 milliards

⇒ **Comment assurer un accès rapide et pertinent à ces images ?**

Recherche d'images : principe



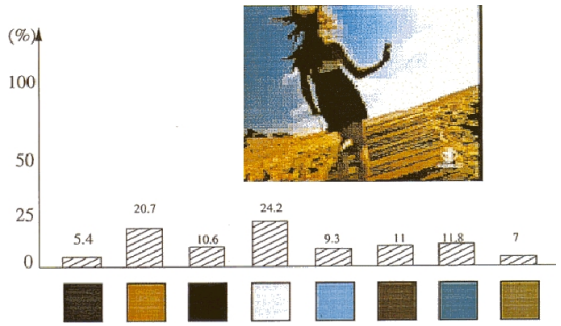
Recherche d'images : principe



Recherche d'images : approches majeures

Recherche par le contenu

- → Description visuelle (couleur, texture, forme) du contenu des images
- → Descripteurs numériques : vecteurs
- → Requêtes : images



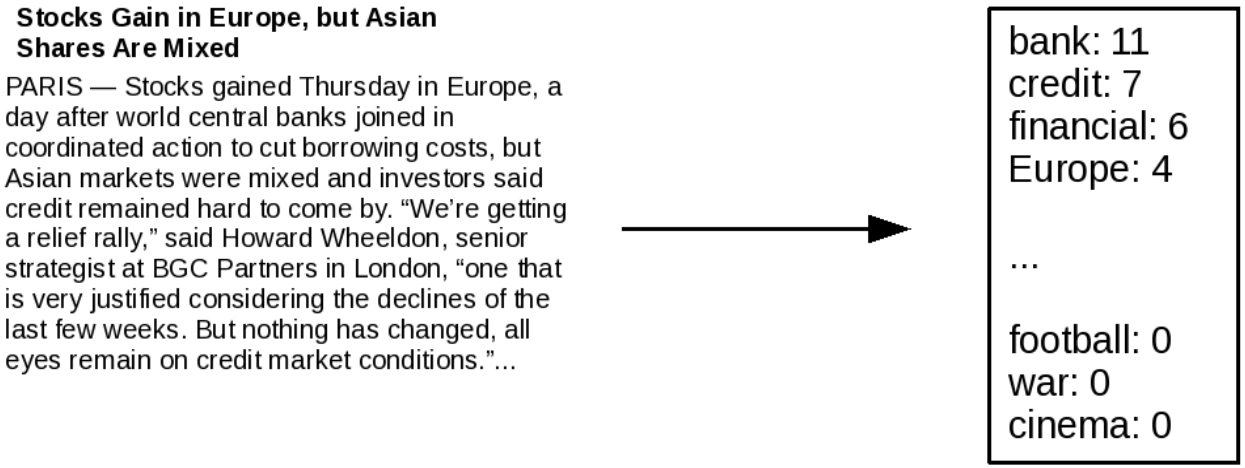
Recherche sémantique d'images

- → Description sémantique (objets, intention...) du contenu des images
- → Descripteurs symboliques : mots-clefs
- → Requêtes : mots-clefs

Recherche d'information (RI) textuelle

→ Principe similaire à la recherche d'images

Description des textes [Salton *et al.*, 1975] :



Systèmes de RI textuelle efficaces

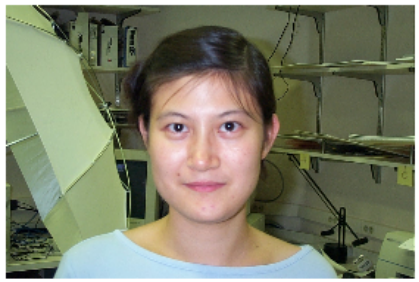
→ Index inversé

→ Manipulation des représentations : interactions avec le traitement automatique des langues (TAL)

Vers une représentation symbolique des images

Stocks Gain in Europe, but Asian Shares Are Mixed

PARIS — Stocks gained Thursday in Europe, a day after world central banks joined in coordinated action to cut borrowing costs, but Asian markets were mixed and investors said credit remained hard to come by. “We’re getting a relief rally,” said Howard Wheeldon, senior strategist at BGC Partners in London, “one that is very justified considering the declines of the last few weeks. But nothing has changed, all eyes remain on credit market conditions.”...



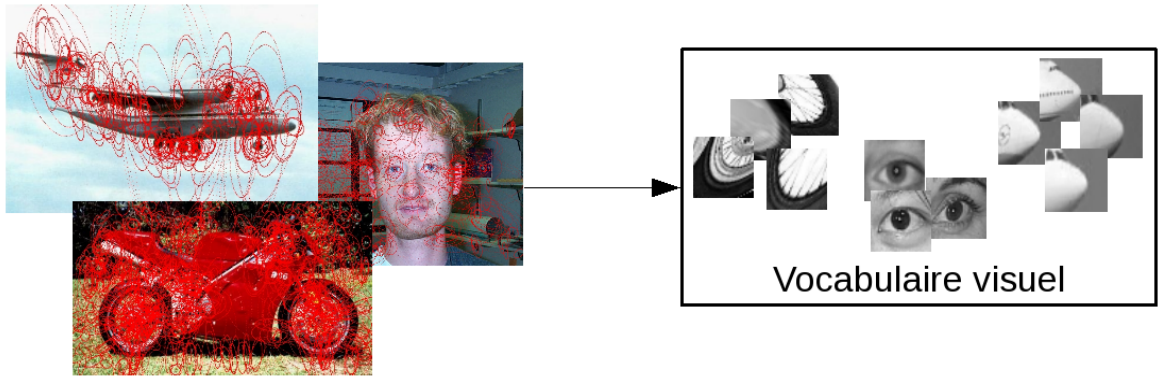
bank: 11
credit: 7
financial: 6
Europe: 4

...

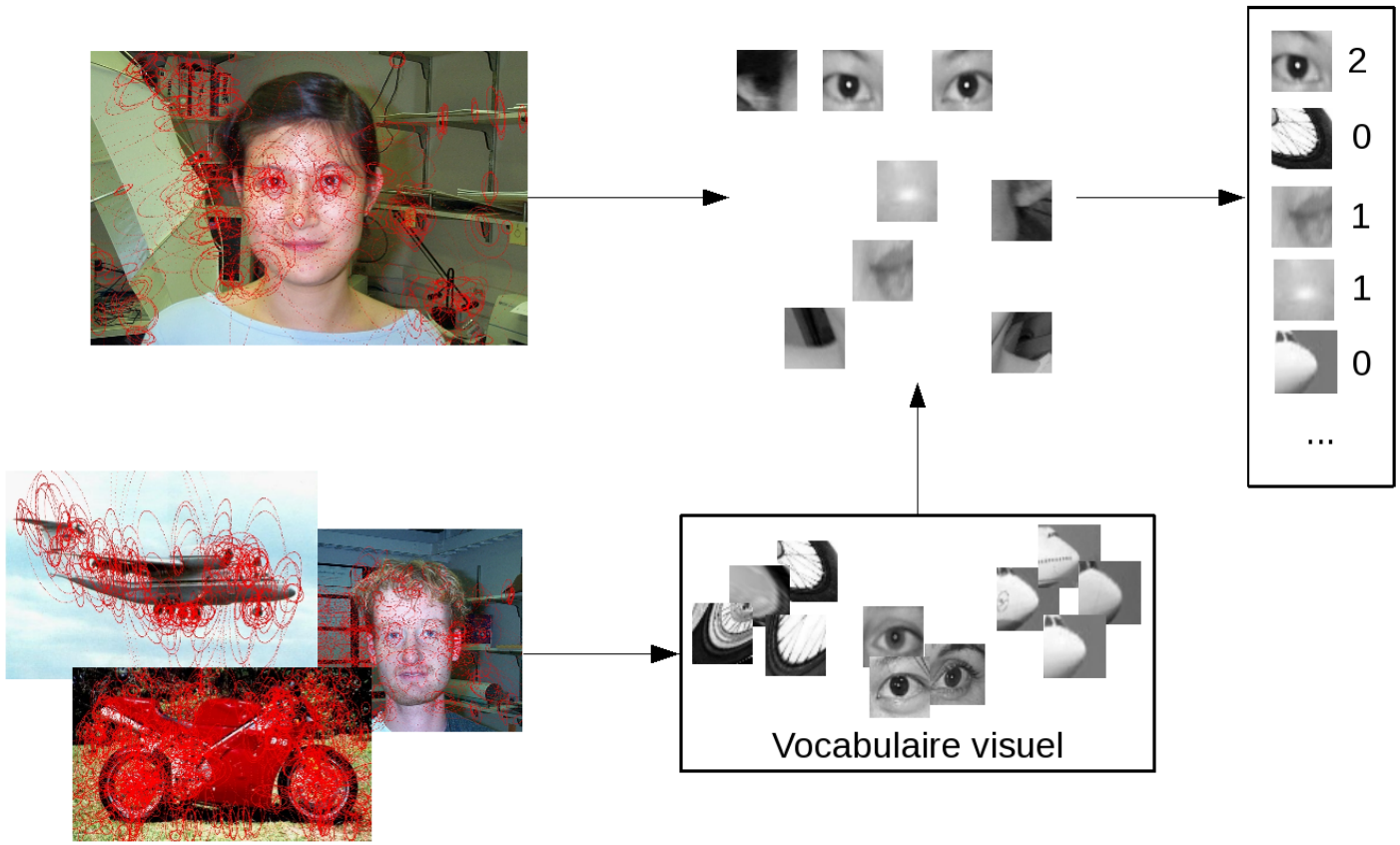
football: 0
war: 0
cinema: 0

?

Construction des mots visuels [Sivic & Zisserman, 2003]



Construction des mots visuels [Sivic & Zisserman, 2003]



Problématiques similaires à celles du texte

Mots visuels et mots textuels

→ Représentation similaire de documents (ensembles de symboles)

→ Problématiques communes

Comment déterminer les mots visuels pertinents ?

→ *Stop-lists* : élimination de mots visuels inutiles

→ Pondérations : associer des scores de pertinence (poids) aux mots visuels

Dépasser l'hypothèse d'indépendance des mots

→ Un objet = un ensemble de mots visuels

→ Outil classique du TAL : les modèles de langues

Axes de travail

Partie I : mesurer la pertinence des mots visuels

Partie II : dépasser l'hypothèse d'indépendance des mots visuels

Partie III : TAL et recherche sémantique d'images

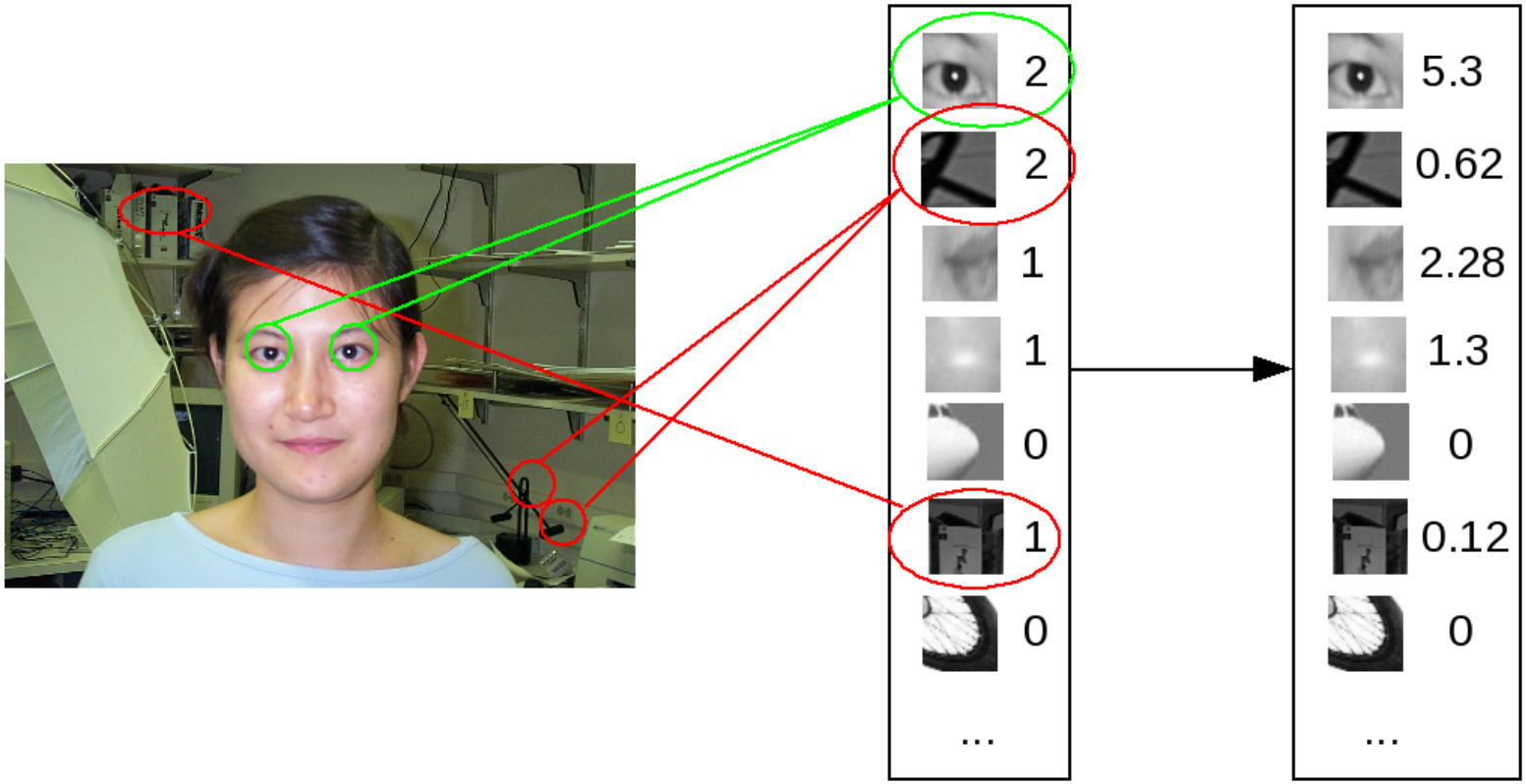
Axes de travail

Partie I : mesurer la pertinence des mots visuels

Partie II : dépasser l'hypothèse d'indépendance des mots visuels

Partie III : TAL et recherche sémantique d'images

Pondérations : principe



→ Pondérations basées sur les propriétés de la quantification [Yang *et al.*, 2007]

→ Pondérations basées sur une segmentation de l'image [Chen *et al.*, 2009]

→ **Approche employée : propriétés statistiques des mots visuels**

Pondérations : généralités

Forme générale des pondérations : $w(i,j) = l(i,j).g(i).n(j)$

Pondération locale $l(i,j)$

→ Poids du mot i dans le document j

→ Forte fréquence dans le document ⇒ poids important [Luhn, 1957]

Pondération globale $g(i)$

→ Poids du mot i dans la collection de documents

→ Mot spécifique à peu de documents ⇒ poids important [Sparck-Jones, 1972]

Facteur de normalisation $n(j)$

→ Fonction de la longueur des documents

→ Fonction de la distance employée

Pondérations locales

Pondération standard

$\rightarrow$ $l_1(i,j) =$ nombre d'occurrences tf_{ij}

Pondération limitant le poids des mots très fréquents

$$\rightarrow l_2(i,j) = \begin{cases} 0 & \text{si } \mathrm{tf}_{ij} = 0 \\ 1 + \log(\mathrm{tf}_{ij}) & \text{sinon} \end{cases}$$

Pondération augmentant le poids des mots très fréquents

$\rightarrow$ $l_3(i,j) = \mathrm{tf}_{ij}^2$

Présence seule

$$\rightarrow l_4(i,j) = \begin{cases} 1 & \text{si } \mathrm{tf}_{ij} > 0 \\ 0 & \text{sinon} \end{cases}$$

Pondérations globales

Poids global constant

$\rightarrow\ g_0(i) = 1$

Fréquence documentaire inverse (IDF)

$\rightarrow\ g_1(i) = \log\left(\frac{N}{\mathrm{df}_i}\right)$

Proposition : Fréquence documentaire inverse et fréquence moyenne

$\rightarrow$ Hypothèse : mots répétés $\Rightarrow$ mots pertinents

$\rightarrow\ g_2(i) = \log\left(\frac{N}{\mathrm{df}_i}\right).\overline{\mathrm{tf}}_i$

Poids global au carré

$\rightarrow\ g_3(i) = g_1(i)^2$

$\rightarrow\ g_4(i) = g_2(i)^2$

Distances et normalisation

$$d_{L_p}(x, y) = \left(\sum_{i=1}^{N} (x_i - y_i)^p \right)^{\frac{1}{p}}$$

Distances de Minkowski : p entier

$\longrightarrow$ Distances classiques utilisées en RI : L_1 et L_2

Distances fractionnelles : $p < 1$

$\longrightarrow$ Bonnes propriétés pour les vecteurs creux [Aggarwal *et al.*, 2001]

Normalisation

$\longrightarrow$ Fonction de la distance employée

Pondérations : protocole expérimental

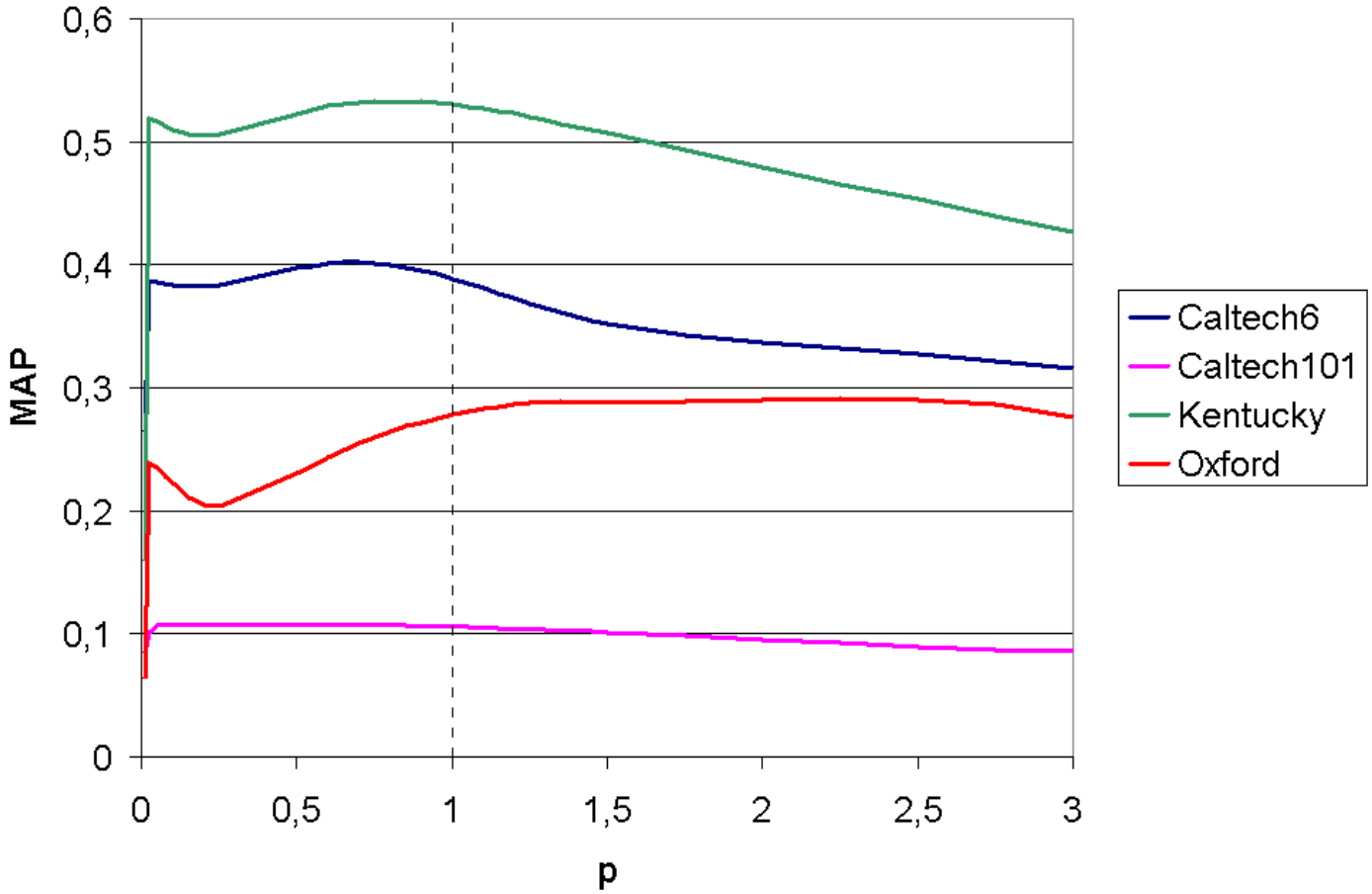
Problèmes traités et données associées :

- → Recherche de scènes identiques
  - → Kentucky [Nister *et al.*, 2006] :
    - 10200 images
    - groupes de 4 images similaires
    - 300 requêtes
  - → Oxford [Chum *et al.*, 2007] :
    - 5000 images
    - nombre d'images pertinentes variable
    - 55 requêtes
- → Recherche d'images catégorisées
  - → Caltech-6 :
    - 5415 images
    - 6 catégories
    - 200 requêtes
  - → Caltech-101 [Fei-Fei *et al.*, 2004] :
    - 8697 images
    - 101 catégories
    - 200 requêtes

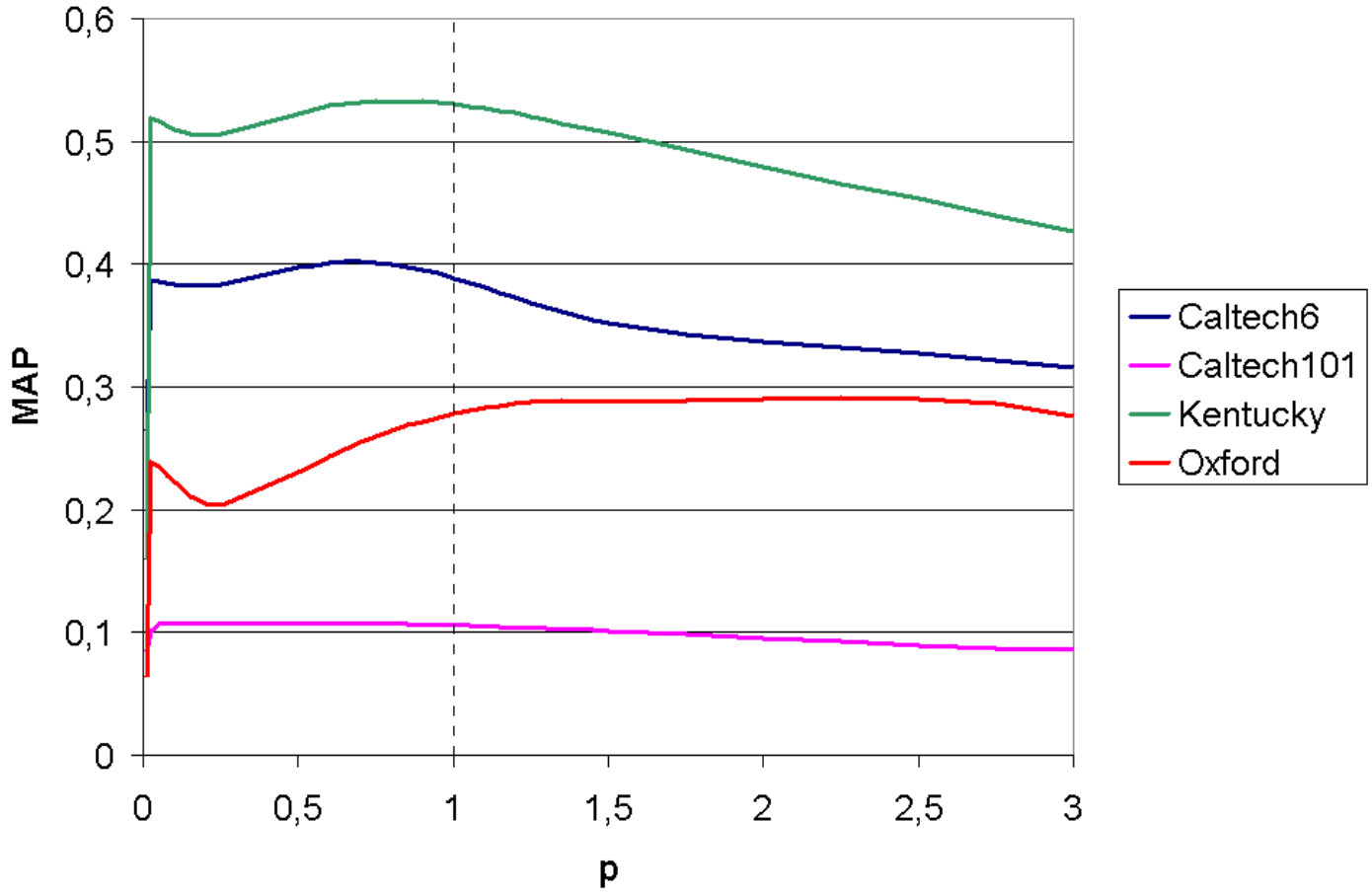
Évaluation :

- → Mesure : *Mean Average Precision* (MAP)

Influence du degré p des distances



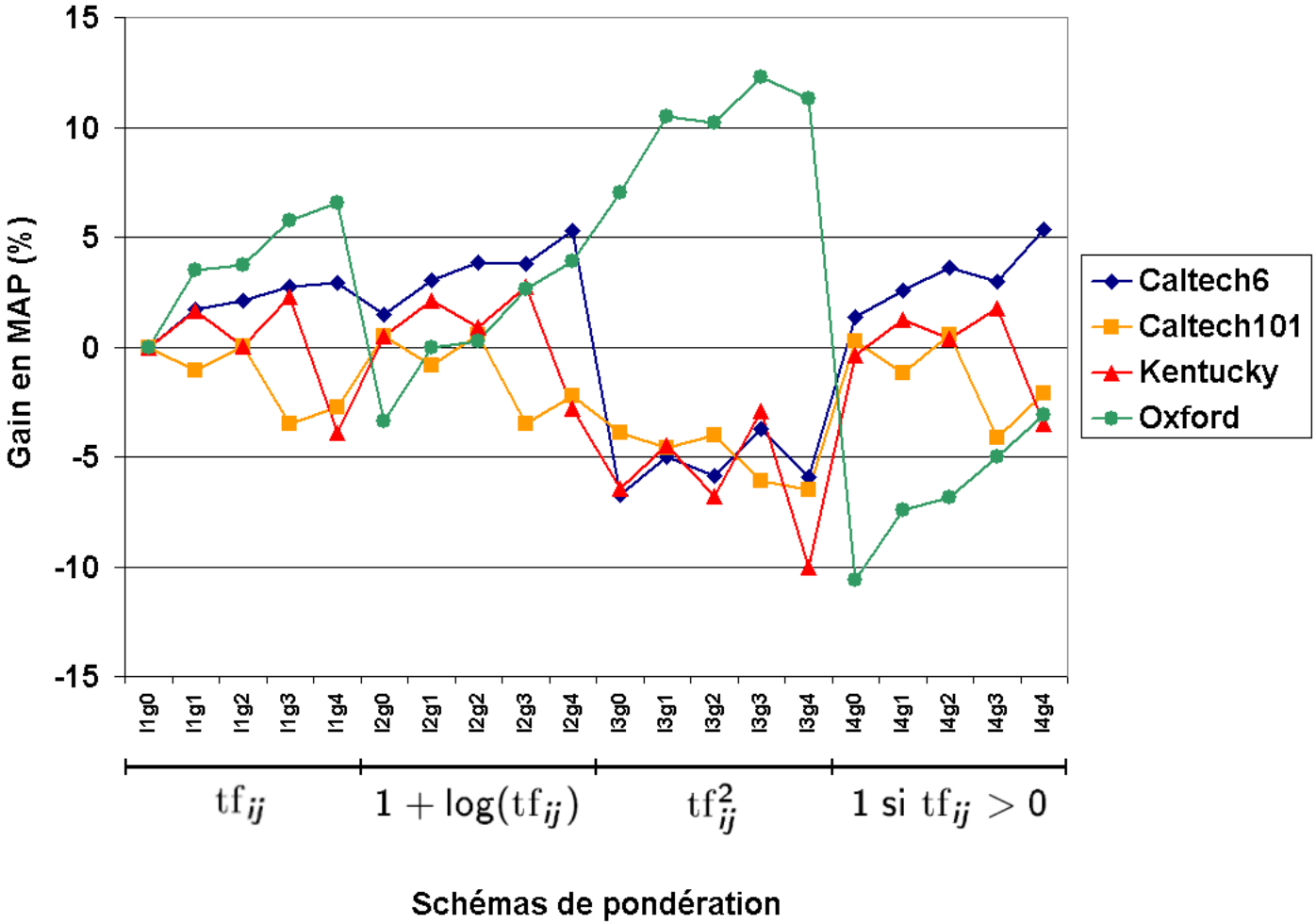
Influence du degré p des distances



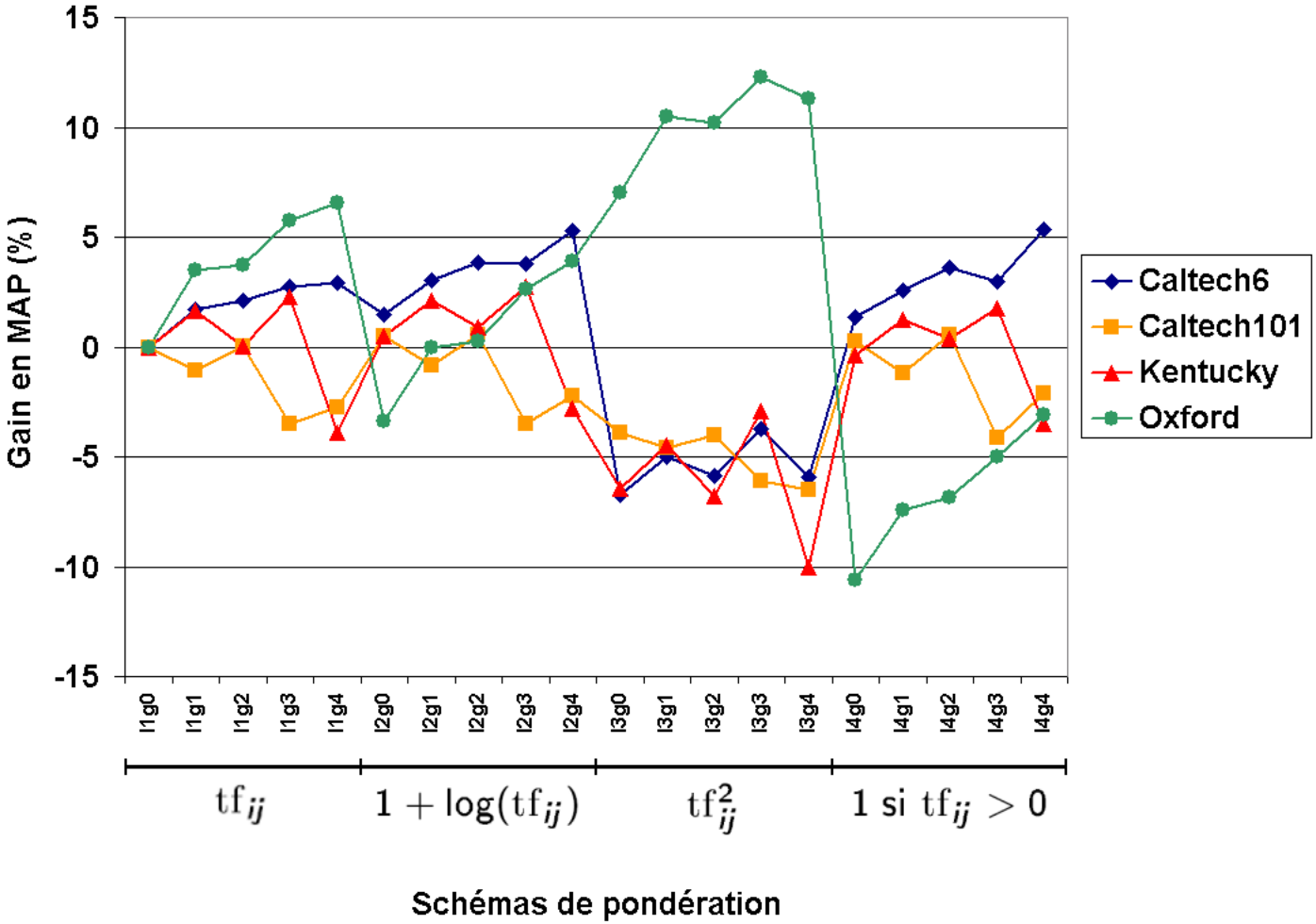
p **sans influence** $\rightarrow$ **données bruitées** [Aggarwal *et al.*, 2001]

$\rightarrow$ Phase de quantification des descripteurs locaux

Influence des pondérations



Influence des pondérations



Pondérations : conclusion

Mise en évidence de propriétés des mots visuels

→ Pas de pondération globalement optimale

→ Influence de la nature du contenu visuel de l'image

Mise en évidence de propriétés des pondérations

→ Relation entre choix des pondérations et choix des distances

Limite des mots visuels

→ Bruit dû à la quantification des vecteurs

[*Distances and weighting schemes for bag of visual words image retrieval*, P. Tirilly, V. Claveau et P. Gros, ACM Conference on Multimedia Information Retrieval MIR'2010]

Axes de travail

Partie I : mesurer la pertinence des mots visuels

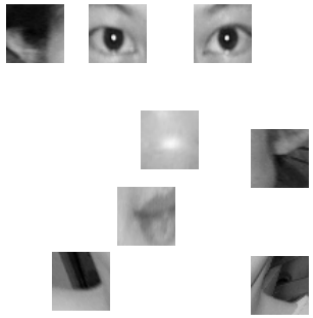
Partie II : dépasser l'hypothèse d'indépendance des mots visuels

Partie III : TAL et recherche sémantique d'images

Indépendance des mots visuels : problématique

Hypothèse d'indépendance des mots inadaptée aux images

→ 1 objet = 1 ensemble de mots visuels



→ Mots visuels non spécifiques aux objets (exemple sur Caltech-6)

Nombre de catégories	1	2	3	4	5	6
Nombre de mots visuels	4	3	11	15	81	6442

Prise en compte de relations géométriques

→ Post-traitement des résultats de recherche [Sivic et Zisserman, 2003]

→ Ensembles de mots visuels [Zheng *et al.*, 2006]

→ **Approche employée : analogie avec la syntaxe des textes**

Modélisation probabiliste de séquences de mots

- $\rightarrow \Pr(w_1 w_2 \ldots w_k) = \prod_{i=1}^{k} \Pr(w_i | w_1 \ldots w_{i-1})$
- $\rightarrow$ Modèle = ensemble de probabilités apprises sur un corpus d'apprentissage

$$
\begin{aligned}
\Pr(\text{Le premier ministre voyage en jet privé}) = \quad & \Pr(\text{Le}) \\
& \times \Pr(\text{premier}|\text{Le}) \\
& \times \Pr(\text{ministre}|\text{Le premier}) \\
& \ldots \\
& \times \Pr(\text{jet}|\text{Le premier ministre voyage en}) \\
& \times \Pr(\text{privé}|\text{Le premier ministre voyage en jet})
\end{aligned}
$$

Modélisation probabiliste de séquences de mots

- $\rightarrow \Pr(w_1 w_2 \ldots w_k) \approx \prod_{i=1}^{k} \Pr(w_i | w_{i-n+1} \ldots w_{i-1})$
- $\rightarrow$ Modèle = ensemble de probabilités apprises sur un corpus d'apprentissage
- $\rightarrow$ Approximation n-grammes (sous-séquences de longueur n)

$$
\begin{aligned}
\Pr(\text{Le premier ministre voyage en jet privé}) \approx \quad & \Pr(\text{Le}) \\
& \times \Pr(\text{premier}|\text{Le}) \\
& \times \Pr(\text{ministre}|\text{premier}) \\
& \ldots \\
& \times \Pr(\text{jet}|\text{en}) \\
& \times \Pr(\text{privé}|\text{jet})
\end{aligned}
$$

Modèles de langues

Modélisation probabiliste de séquences de mots

- $\rightarrow \Pr(w_1 w_2 \ldots w_k) \approx \prod_{i=1}^{k} \Pr(w_i | w_{i-n+1} \ldots w_{i-1})$
- $\rightarrow$ Modèle = ensemble de probabilités apprises sur un corpus d'apprentissage
- $\rightarrow$ Approximation n-grammes (sous-séquences de longueur n)

$$
\begin{aligned}
\Pr(\text{Le premier ministre voyage en jet privé}) \approx \quad & \Pr(\text{Le}) \\
& \times \Pr(\text{premier}|\text{Le}) \\
& \times \Pr(\text{ministre}|\text{premier}) \\
& \ldots \\
& \times \Pr(\text{jet}|\text{en}) \\
& \times \Pr(\text{privé}|\text{jet})
\end{aligned}
$$

Problème

- $\rightarrow$ n-gramme absent du corpus $\rightarrow$ probabilité nulle
- $\rightarrow$ Méthodes de lissage

Modèles de langues

Modélisation probabiliste de séquences de mots

- $\Pr(w_1 w_2 \ldots w_k) \approx \prod_{i=1}^{k} \Pr(w_i | w_{i-n+1} \ldots w_{i-1})$
- Modèle = ensemble de probabilités apprises sur un corpus d'apprentissage
- Approximation n-grammes (sous-séquences de longueur n)

$$\begin{aligned}\Pr(\text{Le premier ministre voyage en jet privé}) \approx \quad & \Pr(\text{Le}) \\ & \times \Pr(\text{premier}|\text{Le}) \\ & \times \Pr(\text{ministre}|\text{premier}) \\ & \ldots \\ & \times \Pr(\text{jet}|\text{en}) \\ & \times \Pr(\text{privé}|\text{jet})\end{aligned}$$

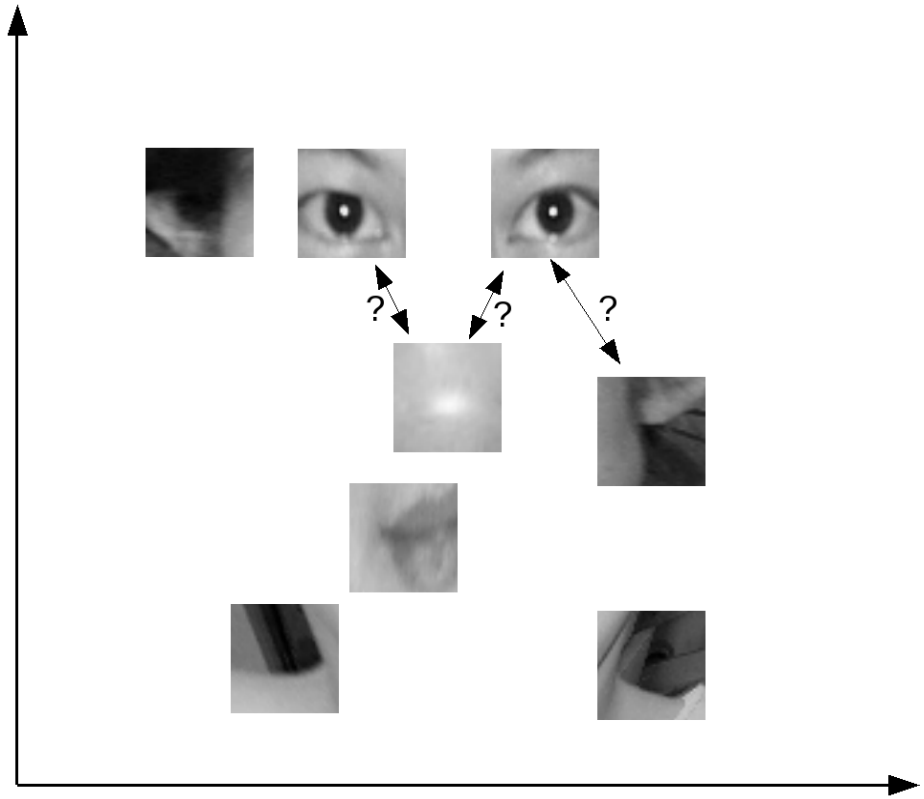
Problème

- n-gramme absent du corpus $\longrightarrow$ probabilité nulle
- Méthodes de lissage

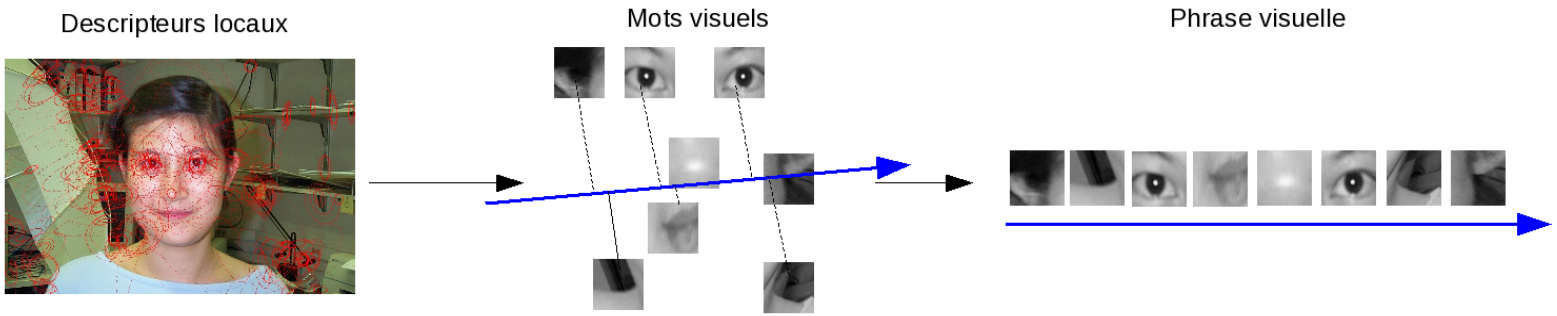
Modèles de langues et classification

- Classification : attribuer une classe parmi X à une séquence $w_1 w_2 \ldots w_k$
- 1 classe $\Leftrightarrow$ 1 modèle de langues
- Classe de $w_1 w_2 \ldots w_k$ $\Leftrightarrow$ modèle maximisant $\Pr(w_1 w_2 \ldots w_k)$

Le premier ministre voyage en jet privé.



Les phrases visuelles



Séquence résultante invariante :

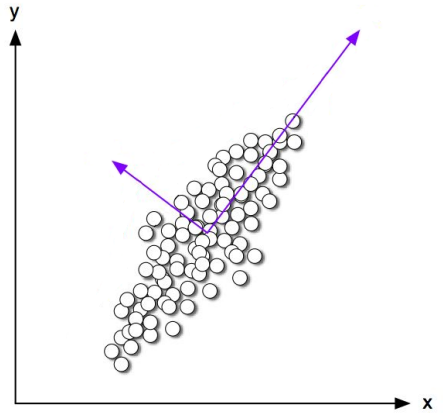
- → aux changements d'échelle
- → aux translations
- → aux rotations (avec un axe adapté)

Choix de l'axe de projection

Axe des abscisses

Choix de l'axe de projection

Axe des abscisses



Axe obtenu par analyse en composantes principales (ACP)

Choix de l'axe de projection

Axe des abscisses

Axe obtenu par analyse en composantes principales (ACP)

Choix de l'axe de projection

Axe des abscisses

Axe obtenu par analyse en composantes principales (ACP)

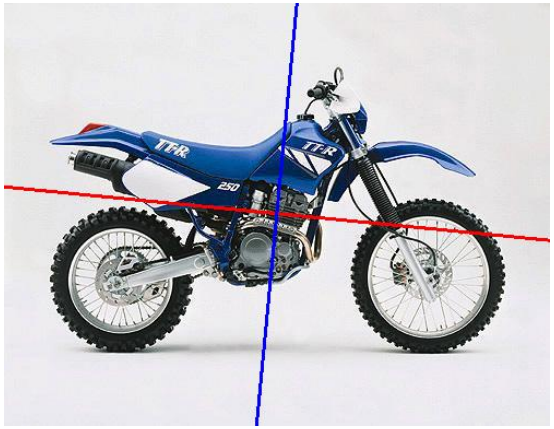
Axe aléatoire

Choix de l'axe de projection

Axe des abscisses

Axe obtenu par analyse en composantes principales (ACP)

Axe aléatoire



Axes multiples

Modèles de langues et mots visuels : protocole expérimental

Données :

→ Caltech-6

→ Caltech-101

Mesure d'évaluation :

→ Taux de classification réussie :

$$P = \frac{\text{nombre d'images correctement classées}}{\text{Nombre total d'images}}$$

Baseline **:**

→ SVM (*Support Vector Machines)* [Csurka *et al.*, 2004]

Catégorisation d'images : résultats sur Caltech-6

Choix de l'axe :

- → Résultats : axe des abscisses > axe obtenu par ACP > axe aléatoire > 2 axes > 10 axes
- → Axe obtenu par ACP : instable
- → Axes multiples : sur-apprentissage

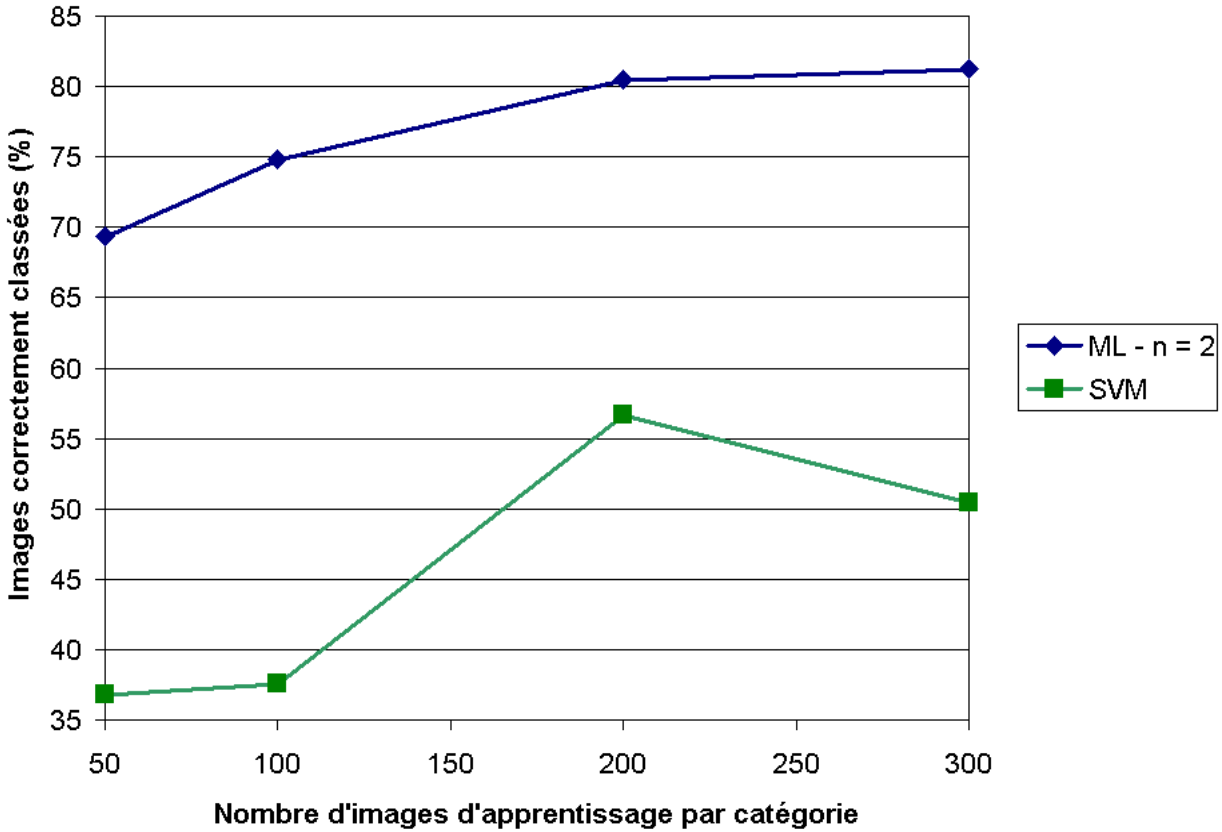
Longueur n des n-grammes

- → Valeur optimale : $n = 2$
- → $n > 4 \Rightarrow$ sur-apprentissage

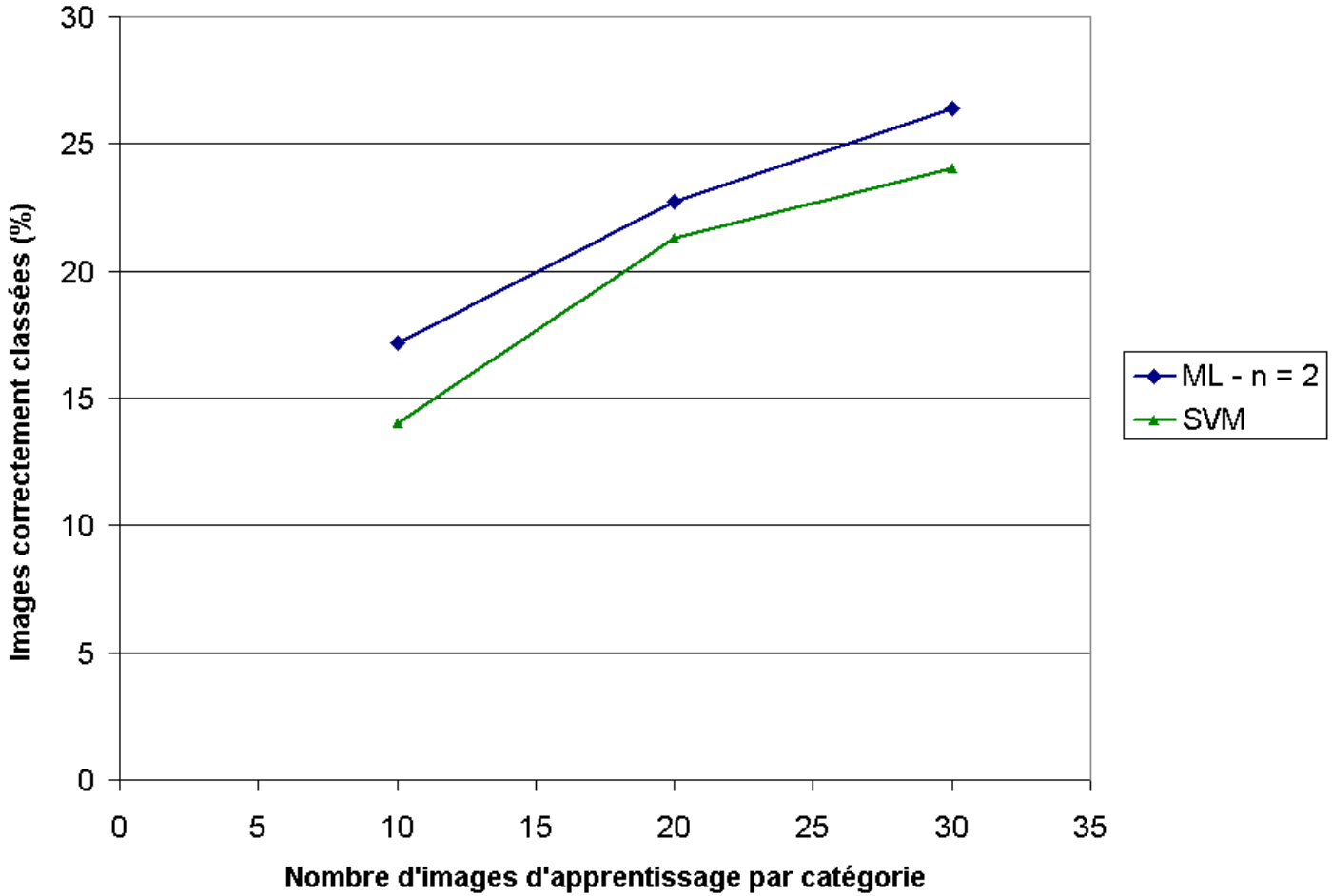
Méthode de lissage

- → Testés : lissage absolu, lissage de Katz, lissage linéaire, lissage de Witten-Bell
- → Optimal : lissage linéaire

Comparaison avec les SVM : résultats sur Caltech-6



Comparaison avec les SVM : résultats sur Caltech-101



Modèles de langues et mots visuels : conclusion

Modèle en phrase visuelle

→ Mise en séquence des mots visuels

Modèles de langues

→ Prise en compte des proximités entre mots visuels

→ Longueur optimale des n-grammes : $n = 2 \Rightarrow$ relations de proximité limitées

Application en classification d'images

→ Amélioration par rapport aux modèles avec indépendance des mots visuels (SVM)

[*Language modeling for bag-of-visual words image categorization*, P. Tirilly, V. Claveau et P. Gros, ACM Conference on Image and Video Retrieval CIVR'2008]

Axes de travail

Partie I : mesurer la pertinence des mots visuels

Partie II : dépasser l'hypothèse d'indépendance des mots visuels

Partie III : TAL et recherche sémantique d'images

Recherche sémantique d'images : approches classiques

Approche classique : annotation par apprentissage supervisé [Barnard *et al.*, 2001]

- → Problème de classification
- → Entrée : descripteurs de bas-niveau
- → Sortie : catégories "sémantiques" (mots-clefs)

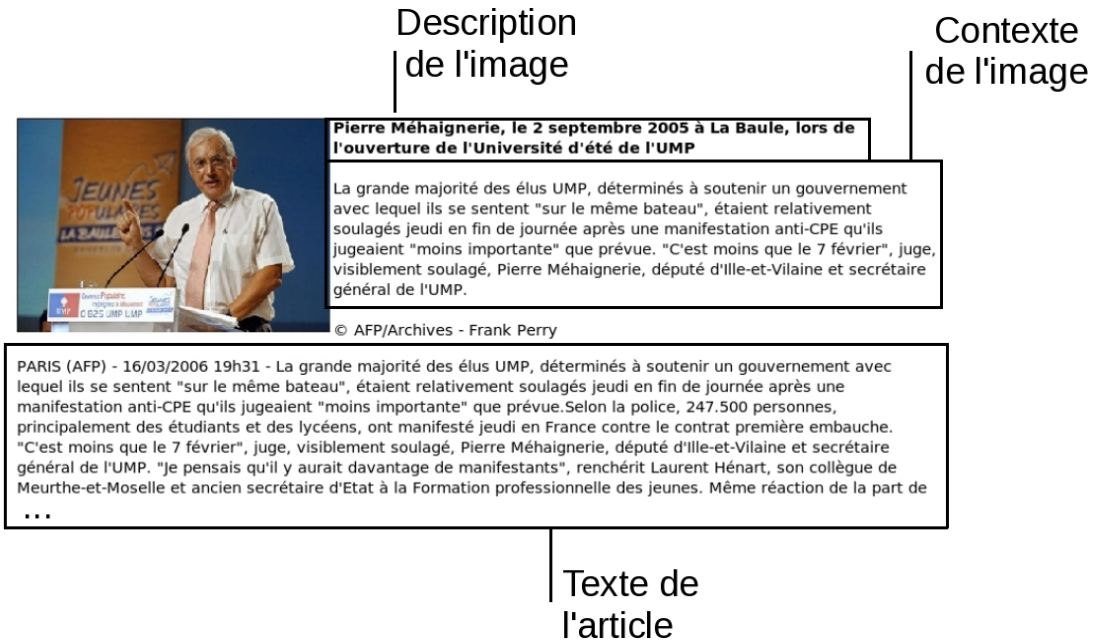
Limites de ces techniques

- → Données artificielles
- → Vocabulaire d'annotation fixe
- → Fossé sémantique

Notre approche

- → Données réelles
- → Utilisation du TAL pour annoter les images à partir de textes

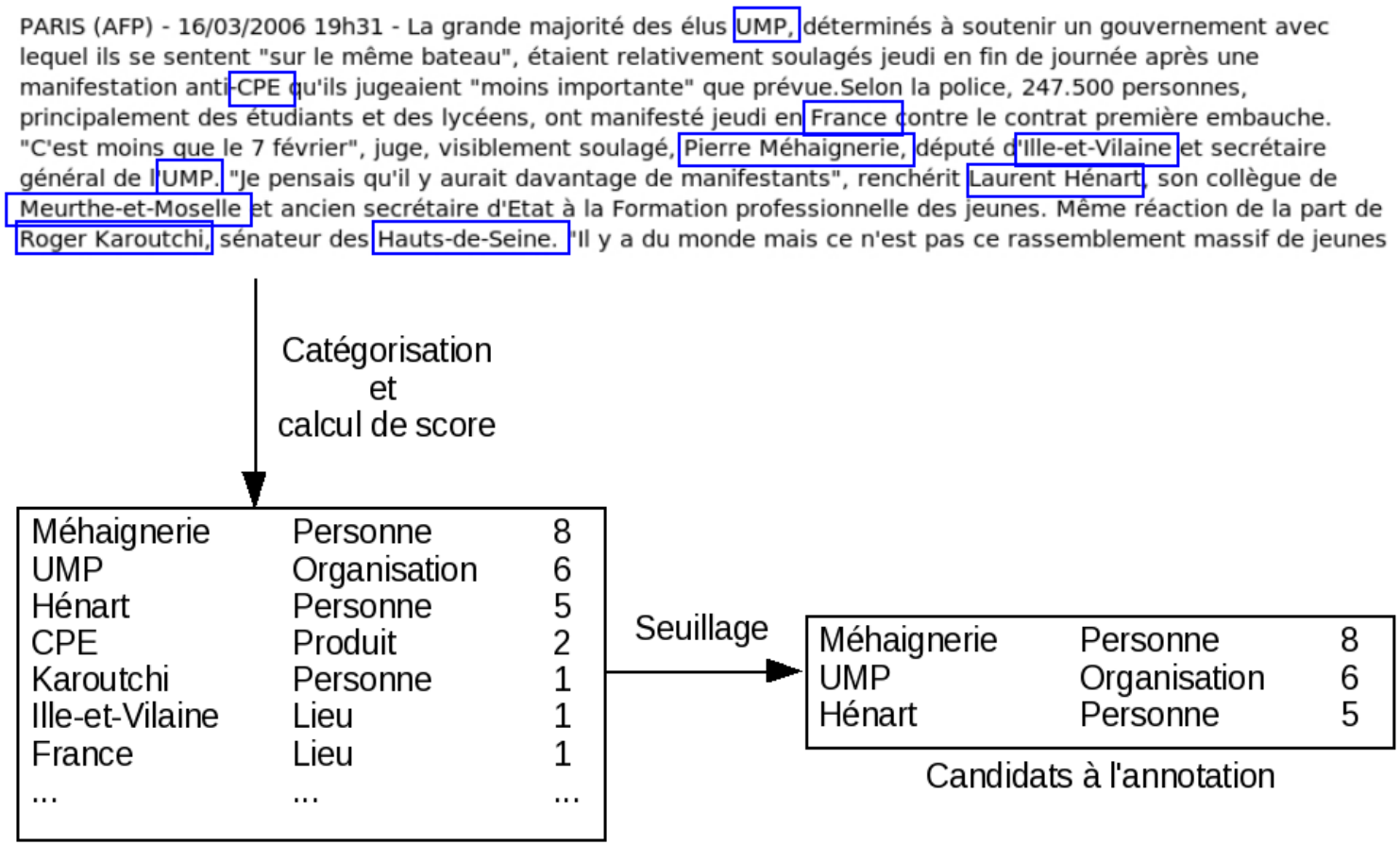
Corpus d'articles de presse illustrés



Pas de corrélation entre les descripteurs visuels classiques et la sémantique associée aux images :

- → Descripteurs : couleur, texture, mots visuels
- → Sémantique : descriptions par des journalistes

Sélection des entités nommées candidates



Détection et catégorisation des entités nommées : Némésis [Fourour, 2002]

Critères de sélection des entités nommées

Calcul des scores :

→ Basé sur les propriétés statistiques des EN

→ Inspiré de la RI textuelle

Scores proposés :

→ Fréquence f seule

→ Fréquence et fréquence documentaire $f-df$

→ Fréquence et fréquence documentaire inverse $f-idf$

Autres critères d'annotation proposés

→ Résultats similaires

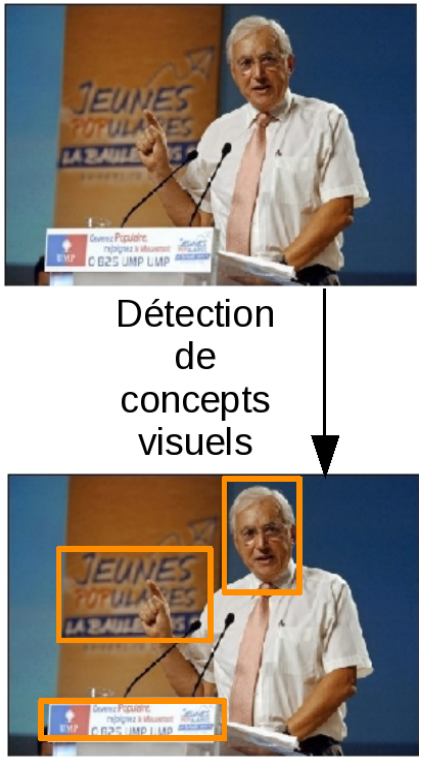
→ Voir manuscrit

Méthode d'annotation

Entités nommées candidates

Méhaignerie	Personne	8
UMP	Organisation	6
Hénart	Personne	5

Méthode d'annotation



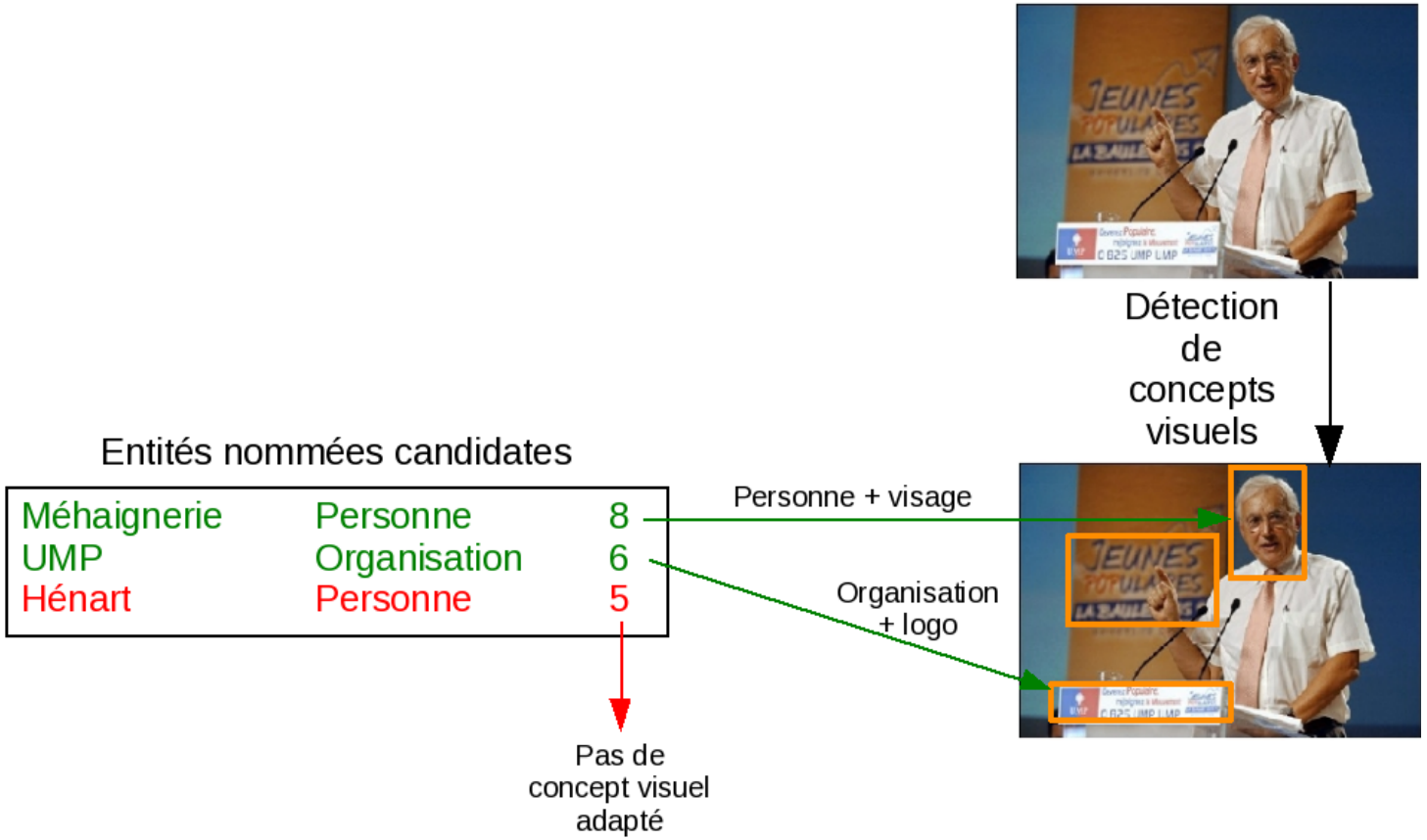
Entités nommées candidates

Méhaignerie	Personne	8
UMP	Organisation	6
Hénart	Personne	5

→ Visages : openCV [Lienhart *et al.*, 2002]

→ Logos : détecteur basé sur les mots visuels et l'apprentissage bayésien

Méthode d'annotation



Données et méthode d'évaluation

Taille du corpus :

→ 27041 articles

→ 42568 images

Vérité-terrain : légendes des images

→ Annotation présente dans la légende = annotation juste

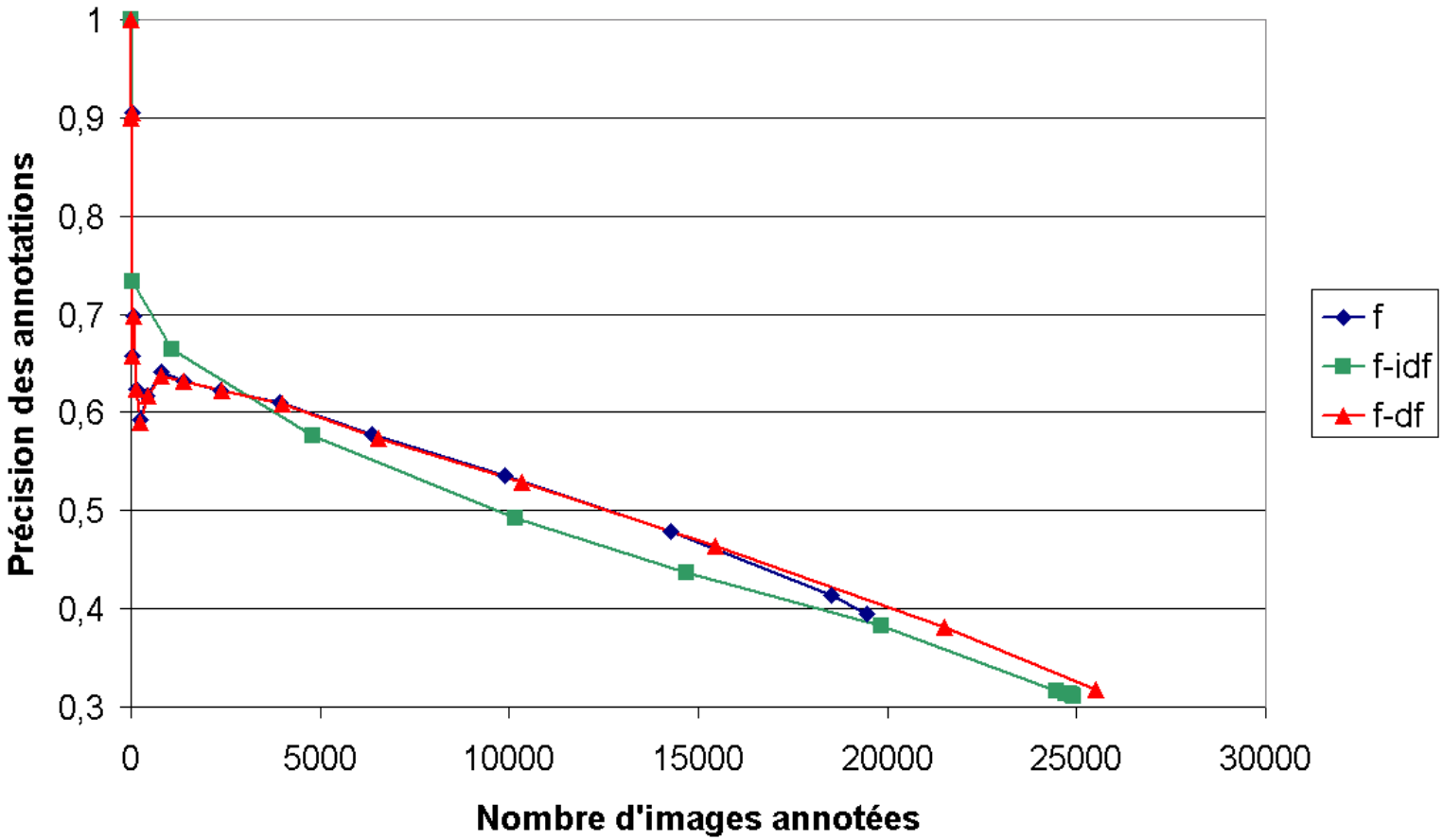
Mesures d'évaluation :

→ Précision $P = \frac{\text{nombre d'annotation justes}}{\text{nombre total d'annotations}}$

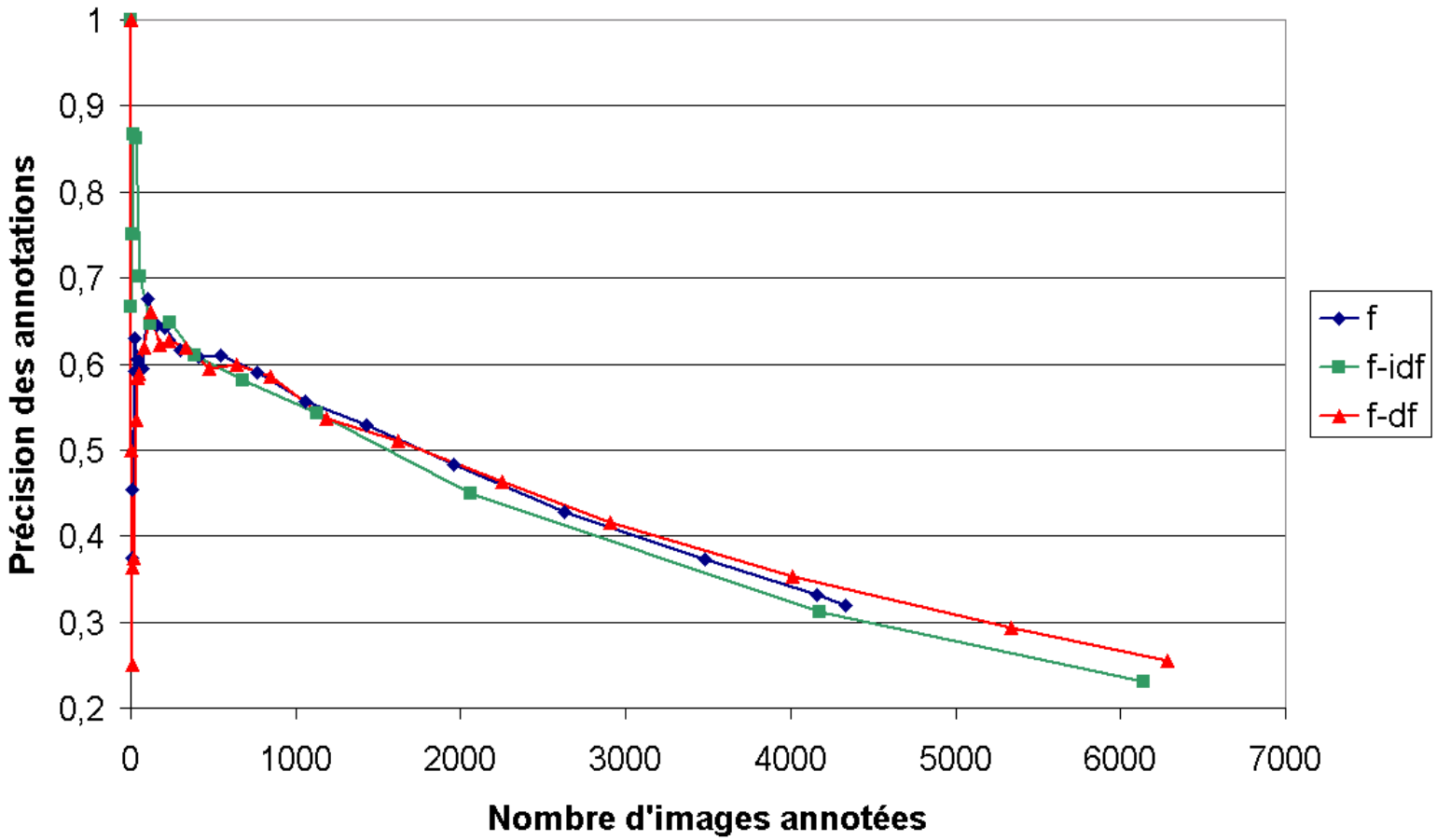
→ Nombre d'images annotées

→ Seuil de sélection variable ⇒ points *(nombre d'images annotées, précision)*

Résultats des annotations pour les visages

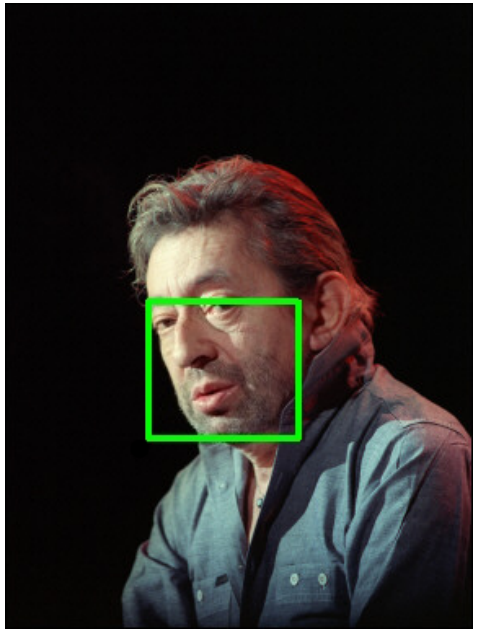


Résultats des annotations pour les logos



Exemples d'annotation

Libération 8
CE 2
SCPL 2
Rotschild 2
Société Civile des
Personnels de Libération 1
Le Monde 1
Comité d'Entreprise 1



Gainsbourg 17
Nelson 2
Melody 2
Birkin 2
Hardy 2

TAL et recherche sémantique d'images : conclusion

Nouvelle approche d'annotation

→ Utilisation des textes accompagnant les images

→ Utilisation de concepts textuels et visuels de haut-niveau

Avantages

→ Pas de phase d'apprentissage

→ Pas de fossé sémantique

→ Vocabulaire d'annotation acquis sur corpus

→ Exécution rapide

→ Précision correcte malgré la simplicité

Inconvénient

→ Difficile à généraliser à d'autres types de concepts visuels et textuels

[*News image annotation on a large parallel text-image corpus*, P. Tirilly, V. Claveau et P. Gros, Language Resources and Evaluation Conference LREC'2010]

Bilan des contributions

Exploitation de méthodes textuelles pour la recherche d'images

Recherche par le contenu : utilisation des mots visuels

- → Mesure de la pertinence des mots visuels
  - → *Stop-lists*
  - → Pondérations
- → Prise en compte de relations géométriques entre mots visuels
  - → Modèle en phrases visuelles
  - → Utilisation des modèles de langues dans un contexte de classification

Recherche sémantique : exploitation des textes accompagnant les images

- → Mesure de corrélations entre descriptions visuelles et textuelles
- → Méthode d'annotation : correspondances entre entités nommées et visages/logos

Outils utiles à ces travaux

- → Mise en place d'un corpus bimodal de données réelles
- → Détecteur de logos

Perspectives à court terme

Mots visuels et pondérations

→ Adaptation des pondérations/de la distance à la requête

Mots visuels et modèles de langues

→ Usage en recherche d'information [Ponte *et al.*, *1998*]

Annotation d'images à l'aide d'entités nommées

→ Autres outils de TAL pour la sélection des entités nommées

→ *Ex : Jacques Chirac est mis en examen : il est accusé d'avoir mis en place des emplois fictifs lorsqu'il était maire de Paris.*

Perspectives à plus long terme

Méthodologies de la RI textuelle pour l'image

- → Évaluation (stabilité des mesures, méthodes de *pooling*. . .)
- → Formalisation des systèmes
  - → Notion de pertinence entre images
  - → Définition des propriétés souhaitées des systèmes [Fang *et al.*, 2004]

Adaptativité des systèmes

- → Choix des descripteurs adapté à la requête
- → Notion de pertinence subjective

TAL et recherche sémantique

- → Vidéo
- → Blogs et réseaux sociaux

Traitement automatique des langues pour l'indexation d'images

Pierre Tirilly

Équipe-Projet TEXMEX - IRISA