Des vocabulaires textuels aux vocabulaires visuels : quels fondements ?

Jean Martinet[1] Omar Rihawi[1]

[1] LIFL/CNRS-UMR 8022 – Université Lille 1

Lille, France

{jean.martinet, omar.rihawi}@lifl.fr

Résumé

L'approche populaire des "sacs de mots visuels" pour la représentation et la recherche de documents visuels consiste à décrire les images (ou trames d'une vidéo) à l'aide d'ensembles de descripteurs, qui correspondent à des caractéristiques de bas niveau discrétisées. La plupart des approches existantes utilisant les mots visuels s'inspirent des travaux en indexation textuelle, en se fondant sur la supposition tacite que les mots visuels peuvent être manipulés de la même façon que les mots textuels. Plus spécifiquement, ces techniques reposent implicitement sur le même postulat de recherche d'information textuelle qui stipule que la distribution des mots d'une langue naturelle suit globalement la loi de Zipf – autrement dit, les mots de la langue naturelle apparaissent dans un corpus avec une fréquence d'occurrence inversement proportionnelle à leur rang. Cependant, notre étude montre que la distribution des mots visuels dépend des caractéristiques de bas niveau et surtout des méthodes de regroupement employées. Nous montrons également que lorsque la distribution des mots visuels est proche de celle des mots textuels, les résultats d'un système de recherche d'images sont améliorés. A notre connaissance, aucune étude n'a encore été menée visant à comparer les distributions des mots textuels et visuels, dans le but d'établir les fondements théoriques des approches des vocabulaires visuels.

Mots clefs

Indexation d'images, Sac de mots visuels, Vocabulaires visuels, Loi de Zipf, Conjecture de Luhn.

1 Introduction

En recherche d'images comme en recherche d'information en général, la qualité de représentation des documents est évidement centrale pour l'efficacité d'un système. Les approches répandues basées sur les sacs de mots visuels proposent de représenter les images par des ensembles de descripteurs, qui correspondent à des caractéristiques de bas niveau discrétisées. Ces caractéristiques sont extraites localement des zones de l'images, générant ainsi un grand nombre de vecteurs de caractéristiques – équivalent des **mots d'un texte**. Une étape de *quantification* permet de réduire ce nombre en regroupant les vecteurs en clusters dans l'espace des caractéristiques, et en associant à chaque vecteur un représentant dans un ensemble discret – équivalent d'un **terme d'indexation**. Les images peuvent ainsi être représentées comme des ensembles, histogrammes ou vecteurs de termes, ces derniers jouant le rôle d'un vocabulaire d'indexation. Cette représentation est inspirée de l'approche des *sacs de mots* en indexation et la recherche d'information textuelle, où les textes sont représentés par des ensembles de termes issus d'un vocabulaire construit à partir du corpus.

Depuis bientôt une décennie, la plupart des systèmes de recherche d'images par le contenu se fondent sur cette représentation désormais classique. De nombreuses techniques classiques d'indexation textuelle telles que la sélection ou la pondération des termes sont directement appliquées aux vocabulaires visuels, ainsi que la mise en correspondance par intersection d'ensembles ou comparaisons de vecteurs. Cependant, les vocabulaires visuels et textuels sont générés par des procédés de natures intrinsèquement différentes, et aucune étude à notre connaissance n'a encore porté sur la validation pour les images des techniques textuelles. L'objectif de ce travail est de comparer les vocabulaires textuels et visuels, afin d'étudier selon certains critères l'applicabilité des techniques textuelles aux vocabulaires visuels.

La suite de cet article est organisée de la manière suivante : la Section 2 présente des généralités sur la recherche d'information textuelle, et les hypothèses implicites sur lesquelles reposent les techniques traditionnelles dans ce domaine. La Section 3 présente le paradigme des sacs de mots visuels, basé sur une analogie texte-image. Cette section détaille également des descripteurs de bas niveau couramment employés et les algorithmes de regroupement que nous choisissons d'étudier. La Section 4 présente notre étude des vocabulaires visuels : les vocabulaires étudiés, les différentes étapes de notre analyse, les conditions expérimentales et les résultats y sont décrits. Nous donnons une conclusion à ce travail et évoquons les travaux futurs dans la Section 5.

2 Recherche d'information textuelle

Les approches de Recherche d'Information (RI) textuelle développées depuis 40 ans se basent sur des mots clés composés de la racine des mots [1]. L'avantage de ces ap-

proches est dû en particulier au fait qu'elles sont efficaces et rapides, comme on le constate avec les moteurs de recherche sur le Web qui sont capables de retrouver rapidement des documents parmi des centaines de millions [2].

2.1 Traitements classiques en RI textuelle

Parmi les modèles existants en RI, le modèle vectoriel [3] a permis d'obtenir de bons résultats à partir de connaissances atomiques telles que les mots clés utilisés classiquement pour indexer des textes. Le contenu d'un document est ainsi exprimé sous la forme d'un ensemble de mots clés jugés représentatifs de ce contenu, tirés du *vocabulaire d'indexation*. Le vocabulaire d'indexation contient tous les termes envisagés pour décrire les documents, et il est construit à partir des mots trouvés dans l'ensemble du corpus, après les opérations et traitements éventuels décrits ci-après.

Filtrage des mots outils et anti-dictionnaire. Seuls les mots porteurs de sens sont habituellement retenus dans le vocabulaire d'indexation : les mots outils de la langue (comme "de", "le", "par" en français), recensés dans une base appelée anti-dictionnaire, ne sont pas conservés dans le vocabulaire d'indexation. De même, les mots rares sont généralement éliminés du vocabulaire d'indexation.

Lemmatisation. Par ailleurs, il est courant de *lemmatiser* (ou *raciniser*) les mots, c'est-à-dire de détecter les variantes orthographiques, comme les pluriels, ou les conjugaisons des verbes, et de ramener les variantes orthographiques d'un même mot à une racine commune, et donc à un même sens [4]. Par exemple, les mots "écrivain", "écrirvez" et "écriture" ont pour racine commune "écri".

Pondération. Les descripteurs d'un document peuvent tous être considérés comme ayant une importance égale, c'est-à-dire aussi représentatifs les uns que les autres de la sémantique de ce document : seule la présence ou l'absence des termes dans les documents importe. Cependant, en réalité, les descripteurs n'ont généralement pas tous la même importance dans les documents. Les descripteurs sont ainsi pondérés selon qu'ils sont jugés plus ou moins représentatifs du contenu d'un document, afin de pouvoir exprimer ces importances relatives et ainsi affiner l'indexation. Les pondérations populaires $TF \times IDF$ [5] et sa variante probabiliste $BM25$ [6] consistent à donner plus d'importance aux termes fréquents dans un document (rôle du TF) et peu fréquents dans le corpus (rôle de l'IDF).

2.2 Loi de Zipf et conjecture de Luhn

Un aspect central des approches de RI textuelle est qu'elles ont pour fondement une caractéristique importante de la distribution des mots d'une langue naturelle. La loi de Zipf lie la probabilité d'occurrence des mots dans un texte en langue naturelle à leur rang lorqu'ils sont rangés par ordre de probabilité décroissante [7]. Cette loi stipule que les occurrences des mots suivent le modèle de distribution suivant :

$$P_n = \frac{1}{n^a} \tag{1}$$

où P_n correspond à la fréquence d'occurrences du mot au n_{ieme} rang et a est une valeur proche de 1. Par exemple, dans un corpus de grande taille en langue anglaise, le terme "the" serait le mot le plus fréquent (il représente généralement environ 7% des occurrences totales). La deuxième place est occupée par le mot "of" qui représente 3,5% des mots, etc. A titre d'exemple, la Figure 1 présente la distribution des mots des articles de Wikipédia en Novembre 2006 sur un repère log-log (bi-logarithmique), ce qui permet de constater la distribution logarithmique.

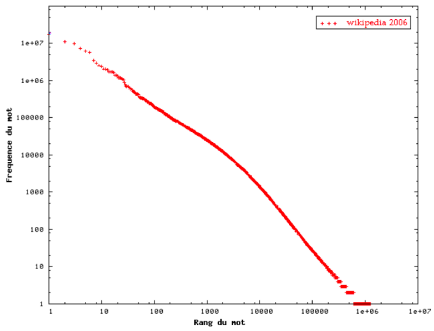

Figure 1 – *Distribution des mots des articles de Wikipédia en Novembre 2006 (Source : Wikipédia).*

Cette distribution logarithmique, interprétée selon la théorie de l'information [8] a notamment permis d'établir les fondements théoriques du filtrage et des stratégies de pondération. En particulier, la sélection des termes discriminants repose largement sur les hypothèses de la conjecture de Luhn [9], fondée elle-même sur la loi de Zipf. Cette conjecture indique une relation entre le rang d'un terme et sa *discriminance* (ou pouvoir de résolution), c'est-à-dire sa capacité à identifier un document pertinent (notion de rappel) combinée à sa capacité à l'isoler des documents non pertinents (notion de précision) – cette relation est illustrée dans la Figure 2. Ainsi, les descripteurs les moins discrimi-

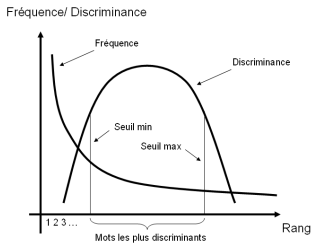

Figure 2 – *Loi de Zipf et conjecture de Luhn.*

nants sont ceux de rangs faibles – très fréquents, et de rangs élevés – très rares ; les descripteurs les plus discriminants sont ceux des rangs médians, et ce sont ces termes qu'il faut sélectionner et conserver dans le vocabulaire d'indexation.

3 Les sacs de mots visuels

A l'instar des documents textuels, les approches des sacs de mots visuels [10, 11, 12, 13] proposent de représenter

les images par des ensembles de descripteurs, qui correspondent à des caractéristiques de bas niveau discrétisées. Ces approches reposent sur une analogie évidente entre les méthodes textuelles et les méthodes visuelles.

3.1 Descripteurs de bas niveau

Parmi les descripteurs de bas niveau, SIFT (Scale-Invariant Feature Transform) [14] et SURF (Speeded-Up Robust Features) [15] sont largement utilisés, et permettent de représenter une image par un ensemble de caractéristiques locales invariantes à l'échelle. Ils intègrent tous deux un détecteur de points d'intérêt et un descripteur pour les points détectés.

SIFT. Le détecteur de SIFT [14] utilise un filtre *Difference of Gaussians* (DoG) pour approximer le *Laplacian of Gaussians* (LoG), afin d'accélérer les traitements. Ce détecteur de points d'intérêt est invariant à l'échelle. Le descripteur calcule un histogramme d'orientations des gradients locaux dans le voisinage du point d'intérêt, et stocke les *bins* dans un vecteur de dimension 128 (8 bins d'orientation pour chacun des 4×4 bins de position). Une particularité de SIFT est de générer un grand nombre de caractéristiques recouvrant l'image par un ensemble complet d'échelles et de localisations. D'une image de résolution 500×500 pixels, il est possible d'extraire environ 2000 caractéristiques stables.

SURF. Le plus récent détecteur-descripteur SURF [15] se base sur une matrice Hessienne pour la détection des points d'intérêt, estimée à l'aide d'images intégrales. Le descripteur calcule la distribution des réponses des ondelettes de Haar dans le voisinage du point d'intérêt, exprimée dans un vecteur de taille 64.

3.2 Discrétisation de l'espace

L'établissement du vocabulaire visuel passe par le regroupement de l'ensemble des caractéristiques extraites de la base d'images. Il s'agit d'une opération de *quantification vectorielle*, qui consiste à représenter chaque caractéristique par le centroïde du groupe (ou *cluster*) qui le contient.

KMeans. L'algorithme KMeans est sans doute le plus utilisé des algorithmes de regroupement. A partir de k centroïdes (généralement initialisés aléatoirement), cet algorithme partitionnel itératif affecte chaque point de l'espace à son plus proche centroïde, puis re-calcule le barycentre de chaque groupe pour déterminer à nouveau les centroïdes. L'algorithme s'arrête lorsqu'un critère de convergence est satisfait, et le partitionnement final détermine le *mapping* entre chaque point de l'espace et l'un des centroïdes.

SOM. Beaucoup moins répandues, les cartes auto-adaptatives (Self-Organizing Maps, SOM) [16, 17] entrent dans la catégorie des réseaux de neurones artificiels, et permettent de faire de la classification non-supervisée. A partir de données dans un espace réel de grande dimension, elles génèrent une cartographie de l'espace en assignant à chaque point un représentant après une étape d'apprentissage. Cette partition des données permet de regrouper des points en cluster.

3.3 Techniques textuelles appliquées aux images

Une fois le vocabulaire visuel généré, les images de la base peuvent être représentées par des ensembles, histogrammes ou vecteurs de termes visuels. De nombreux travaux en représentation et indexation d'images ont adapté les techniques textuelles, telles que le filtrage de mots [10, 18], où les mots visuels jugés peu utiles sont supprimés du vocabulaire d'indexation. Par ailleurs, les termes visuels peuvent être pondérés afin de représenter leur importance. Par exemple, le système SIMPLIcity [19] met en œuvre une pondération $RF \times IPF$ (*Region Frequency* et *Inverse Picture Frequency*), directement inspirée du $TF \times IDF$. Des poids sont attribués aux régions en fonction de leur fréquence dans l'image et dans la collection. Les travaux de [10, 18, 20] adaptent également cette pondération aux images. En allant plus loin dans l'analogie texte-image, plusieurs travaux [21, 18, 22] adaptent les *modèles de langage* aux images, et implémentent des techniques d'expansion de requête [20] (classiques en RI textuelle) en ajoutant automatiquement des termes visuels supplémentaires à la requête dans le but d'améliorer les résultats de la recherche.

4 Etude des vocabulaires visuels

Cette section décrit les expérimentations que nous avons réalisées dans le but de comparer les différents vocabulaires étudiés.

4.1 Distinction terminologique

Nous commençons par proposer la distinction terminologique suivante. Selon l'analogie texte-image, nous appelons *mots visuels* les régions ou zones locales de l'images issues d'une segmentation, d'un découpage en grille, ou dans le voisinage d'un point d'intérêt. Les mots visuels correspondent ainsi aux mots textuels, c'est-à-dire aux occurrences des termes modulées par les variantes orthographiques. Les *termes visuels* correspondant aux termes textuels, c'est-à-dire les éléments du vocabulaire d'indexation. Ainsi, les mots visuels sont des occurrences dans les images des termes visuels, et la variabilité visuelle des régions affectées à un cluster donné peut être vue comme la contrepartie visuelle des variations orthographiques et syntaxiques textuelles.

4.2 Vocabulaires considérés

Les 2 caractéristiques décrites (SIFT et SURF), combinées avec les 2 algorithmes de regroupement KMeans et SOM, permettent la génération de 4 vocabulaires visuels : KMSIFT, KMSURF, SOMSIFT, SOMSURF. Nous proposons de comparer ces vocabulaires selon leurs distributions, et selon les résultats qu'ils permettent d'atteindre dans une tâche de recherche d'images.

4.3 Collections d'images utilisées

Nous avons choisi d'utiliser les bases d'images Caltech-101 et Pascal. Un échantillon de ces collections est donné à la Figure 3. La collection Caltech-101 contient des images classées en 101 catégories, avec environ 50 images par catégorie (variable de 40 à 800). Nous avons sélectionné 10 catégories pour nos expérimentations, résultant en 1345 images.

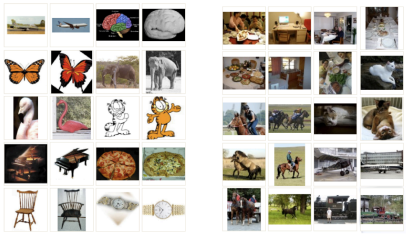

Figure 3 – *Exemple d'images des bases Caltech-101 (gauche) et Pascal (droite).*

La collection Pascal contient des images classées en 20 catégories, et nous avons sélectionné 1000 images uniformément distribuées sur toutes les catégories pour nos expérimentations.
Le temps moyen d'extraction pour ces mots est de 3h13min sur une machine dotée d'un processeur Intel Xeon cadencé à 2.93GHz, avec 4GB de RAM. Le temps requis est sensiblement plus long pour SIFT que pour SURF. On constate que SIFT permet de générer environ 2 fois plus de mots que SURF dans chacune des collections. Cela est dû au détecteur par *différence de gaussiennes* qui extrait un plus grand nombre de points d'intérêt que le détecteur par *matrice hessienne*.
Le Tableau 1 présente le nombre total de mots visuels générés pour chaque collection et chaque descripteur, ainsi que le nombre moyen de mots par image.

Collection (# d'images)	Caltech-101 (1345)		Pascal (1000)	
Descripteur	SIFT	SURF	SIFT	SURF
# total de mots	463K	204K	297K	174K
# mots/image	344	151	297	174

Tableau 1 – *Caractéristiques des mots visuels extraits.*

4.4 Distributions des termes

Nous avons établi des vocabulaires visuels pour les ensembles de mots de chaque collection avec KMeans et SOM. Les vocabulaires obtenus ont une taille standard de 10K termes. Les temps de génération de ces vocabulaires sont très différents : 4,5 jours pour KMeans, et 2,5h pour SOM sur la machine décrite précédemment.
Les Figures 4 et 5 montrent les distributions des différents vocabulaires pour les 2 collections, ainsi que la distribution des mots de Wikipédia en surimpression, identique à la Figure 1. On constate de manière évidente que les vocabulaires KMSIFT et KMSURF ont des distributions plus "plates" que les distributions des vocabulaires SOMSIFT et SOMSURF, cela pour les 2 collections. Cela indique que l'algorithme KMeans génère des vocabulaires plus uniformément distribués que SOM. Autrement dit, cela signifie que les clusters sont de tailles plus égales avec KMeans qu'avec SOM. Cependant, la distribution des vocabulaires engendrés par SOM est très similaire à la distribution des mots de la langue anglaise (pour les articles de Wikipédia en 2006). Cette tendance est identique pour les 2 collections.

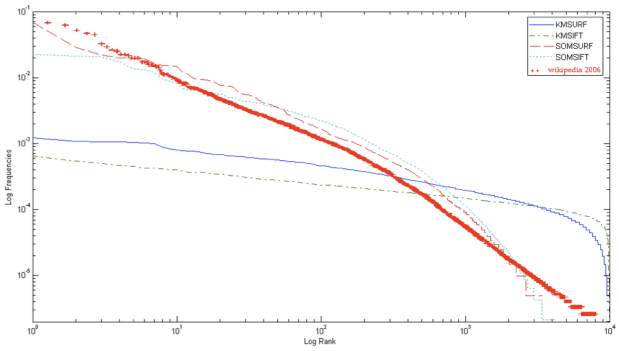

Figure 4 – *Distr. des vocabulaires (Caltech-101).*

Nous avons estimé les pentes des distributions de ces vocabulaires dans le repère log-log. Pour cela, nous avons calculé le coefficient directeur d'une droite de régression selon le critère des moindres carrés, telle que la somme quadratique des erreurs soit minimale. Les pentes estimées sont données dans le Tableau 2. Ces résultats confirment le fait que les vocabulaires générés à l'aide de SOM ont des distributions plus similaires à ceux de la langue naturelle que ceux générés à l'aide de KMeans. Ces résultats indiquent également que le choix des détecteurs et descripteurs utilisés ne semble pas influencer grandement la distribution des vocabulaires.

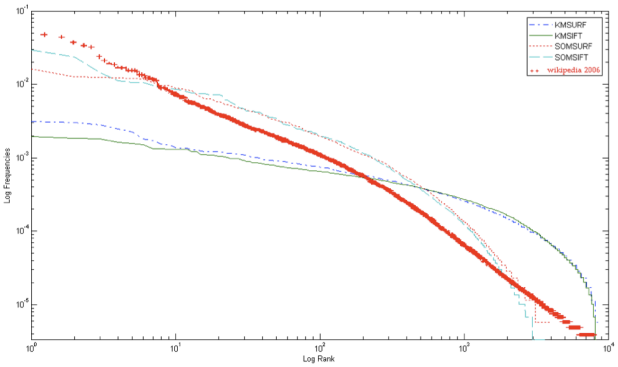

Figure 5 – *Distr. des vocabulaires (Pascal).*

Pour expliquer cela, il faut revenir au processus de regroupement des différents algorithmes : KMeans initialise aléa-

Collection	Caltech-101		Pascal	
Descripteur	SIFT	SURF	SIFT	SURF
KMeans	-7.1°	-6.3°	-8.2°	-9.8°
SOM	-19.0°	-21.5°	-18.7°	-20.1°

Tableau 2 – *Pentes estimées (en degrés) pour les distributions visuelles (pente pour Wikipédia : -24.2°).*

toirement les centroïdes des clusters, et un état de convergence est recherché en déplaçant progressivement les centroïdes. Une situation dans laquelle 2 centroïdes initiaux seraient tous 2 initialement placés dans un groupe de points proches donné (constituant donc un candidat dédié pour un *unique* cluster) est susceptible de se produire. Ainsi, au lieu de produire un unique cluster traduisant la réalité de ce groupe de points, KMeans produirait alors artificiellement 2 clusters mal formés, comme illustré dans la Figure 6. Du fait que le processus de création des clusters est fondamentalement différent pour SOM, cet algorithme ne rencontre pas ce problème.

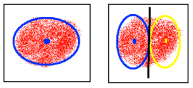

Figure 6 – *Illustration de regroupements correct (à gauche) et incorrect (à droite) pour KMeans, résultants d'initialisation différentes de l'algorithme.*

Dans la suite de cette section, nous sous intéressons aux résultats que ces vocabulaires permettent d'atteindre dans une tâche de recherche d'images.

4.5 Tâche de recherche

La seconde étape de notre étude consiste à évaluer la précision qu'il est possible d'obtenir avec chacun des vocabulaires dans un système de recherche d'images. Dans cette évaluation, un système de recherche d'image a été implémenté sur la base du modèle vectoriel de RI, en mettant en œuvre les techniques de filtrage et de pondération $TF \times IDF$ des termes visuels. Pour les 2 collections, chaque image est utilisée tour à tour comme requête, et les images retournées par le système qui sont dans la même catégorie que la requête sont jugées correctes dans une *vérité terrain*. Les courbes de rappel-précision sont données dans les Figures 7 et 8, pour les collections Caltech-101 et Pascal, respectivement.
On constate dans ces courbes que la précision de recherche semble meilleure pour la collection Caltech-101 que pour la collection Pascal. Cela est dû à la nature des images de la collection : alors que la collection Caltech-101 contient des images ciblées qui présentent généralement un seul objet identifié sur fond uni, les images de la collection Pascal contiennent des images moins ciblées présentant souvent plusieurs objets, avec un fond complexe.
Dans la Figure 7, on peut voir des différences notables des valeurs de précision : le vocabulaire SOMSURF permet

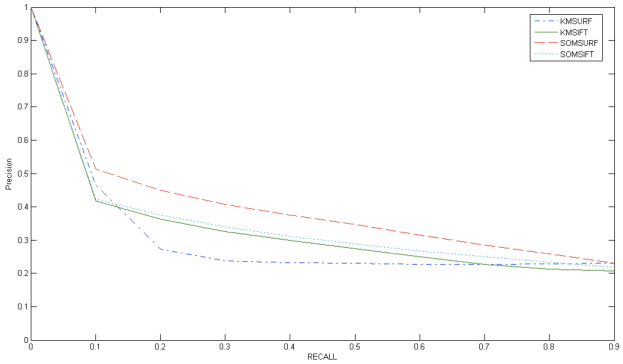

Figure 7 – *Courbes de rappel-précision (Caltech-101).*

d'atteindre une meilleure précision que les autres vocabulaires. Par ailleurs, le vocabulaire KMSURF donne des résultats bien inférieurs aux autres. Cela confirme la faible influence du détecteur-descripteur. La courbe SOMSIFT est légèrement au-dessus de la courbe KMSIFT, mais cette différence est trop faible pour être significative.

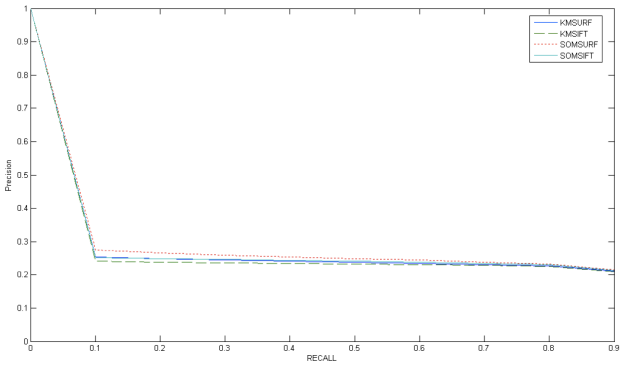

Figure 8 – *Courbes de rappel-précision (Pascal).*

Dans la Figure 8, les résultats sont similaires pour les 4 vocabulaires. A bien plus petite échelle que dans la Figure 7, la courbe SOMSURF est légèrement au-dessus des autres, et la courbe KMSIFT est légèrement en dessous. Mais pour Pascal, cette différence est très faible et non significative. Ici encore, le choix du détecteur-descripteur ne semble pas influencer grandement le processus de création des vocabulaires. Cependant, nous notons que SIFT est moins sensible que SURF à la méthode de regroupement utilisée.
De manière générale, bien que KMeans soit plus populaire que SOM, les résultats dans notre système de recherche d'images des vocabulaires issus de SOM sont meilleurs que ceux de KMeans. D'après notre étude, une justification théorique pourrait être que l'algorithme SOM engendre des vocabulaires dont la distribution est similaire aux mots de la langue naturelle. De ce fait, les hypothèses implicites sur lesquelles reposent les techniques textuelles (postulat de la loi de Zipf, conjecture de Luhn) sont satisfaites dans ce cas pour les images, ce qui serait une validation théorique expliquant ces meilleurs résultats.

5 Conclusion

La plupart des techniques de RI textuelle repose implicitement sur le postulat qui stipule que la distribution des mots d'une langue naturelle suit globalement la loi de Zipf, ainsi que sur la conjecture de Luhn. L'application de ces techniques aux vocabulaires visuels issus des images demande de vérifier la validité de ces postulats.

Dans cet article, nous avons présenté une étude de vocabulaires visuels en comparaison aux vocabulaires textuels, afin d'expliciter les conditions d'applicabilité des techniques textuelles aux images. Notre étude a montré que la distribution des mots visuels dépend grandement des méthodes de regroupement employées. En effet, le choix de la méthode de génération des termes visuels est important, et détermine la distribution des termes du vocabulaire. Nous montrons également que lorsque la distribution des mots visuels est proche de celle des mots textuels, les résultats d'un système de recherche d'images sont améliorés. Ainsi l'algorithme SOM, qui engendre des vocabulaires de distributions similaires à la lange naturelle, permet d'atteindre une meilleure précision de recherche que le pourtant populaire KMeans. A notre connaissance, aucune étude n'a encore été menée visant la validation pour les images des techniques textuelles.

Nos travaux futurs visent à généraliser cette approche comparative aux autres descripteurs de bas niveau courants en représentation d'images. Nous envisageons également d'étudier les conséquences de nos résultats en cas de variation de la taille des vocabulaires, ainsi que de procéder à une comparaison plus formelle des distributions à l'aides d'outils statistiques dédiés. Ce travail s'inscrit dans une démarche globale d'application des techniques textuelles aux images.

Références

[1] R. A. Baeza-Yates et B. A. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] Sergey Brin et Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Dans *In journal of Computer Networks and ISDN Systems (30)*, pages 107–117, Brisbane, 1998.

[3] G. Salton. *The SMART Retrieval System*. Prentice Hall, 1971.

[4] G. Salton et M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[5] G. Salton et C. Buckley. Term-weighting approaches in automatic text retrieval. Dans *Information Processing and Management*, pages 513–523, 1988.

[6] K. Sparck Jones, S. Walker, et S. E. Robertson. A probabilistic model of information retrieval : development and comparative experiments. *Inf. Process. Manage.*, 36 :779–808, November 2000.

[7] George Kingsley Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Addison-Wesley, Cambridge, MA, USA, 1932.

[8] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 :379–423, 623–656, 1948.

[9] H. P Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2) :159–165, 1958.

[10] Josef Sivic et Andrew Zisserman. Video google : A text retrieval approach to object matching in videos. Dans *ICCV*, pages 1470–1477, 2003.

[11] Frédéric Jurie et Bill Triggs. Creating efficient codebooks for visual recognition. Dans *ICCV*, pages 604–610, 2005.

[12] K. Grauman et T. Darrell. The pyramid match kernel : Discriminative classification with sets of image features. Dans *ICCV*, pages 1458–1465, 2005.

[13] Svetlana Lazebnik, Cordelia Schmid, et Jean Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. Dans *CVPR (2)*, pages 2169–2178. IEEE Computer Society, 2006.

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[15] T. Tuytelaars L. V. Gool H. Bay, A. Ess. Surf : Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3), 2008.

[16] Teuvo Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences, vol 30, 2001.

[17] Xinzheng Xu, Wenhua Zeng, et Zuopeng Zhao. A structural adapting self-organizing maps neural network. Dans *ISNN (2)*, pages 913–920, 2007.

[18] P. Tirilly, V. Claveau, et P. Gros. Language modeling for bag-of-visual words image categorization. Dans *CIVR '08*, pages 249–258, 2008.

[19] J.Z. Wang, J.L., et G. Wiederhold. SIMPLIcity : Semantics-sensitive Integrated Matching for picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9) :947–963, 2001.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, et A. Zisserman. Lost in quantization : Improving particular object retrieval in large scale image databases. Dans *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[21] A. Bosch, A. Zisserman, et X. Muñoz. Scene classification via plsa. Dans *ECCV'06*, pages 517–530.

[22] Lei Wu, Yang Hu, Mingjing Li, Nenghai Yu, et Xian-Sheng Hua. Scale-invariant visual language modeling for object categorization. *Multimedia, IEEE Transactions on*, 11(2) :286 –294, feb. 2009.