HAL
open science

Recherche d'images en mobilité - Le système IOTA-EyeSnap

Philippe Mulhem, Jean-Pierre Chevallet, Nicolas Cubaud

To cite this version:

Philippe Mulhem, Jean-Pierre Chevallet, Nicolas Cubaud. Recherche d'images en mobilité - Le système IOTA-EyeSnap. CORIA 2013, 2013, Neuchatel, Suisse. pp.63-72. hal-00953080

HAL Id: hal-00953080

https://inria.hal.science/hal-00953080v1

Submitted on 23 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Recherche d'images en mobilité
Le système IOTA-EyeSnap

Philippe Mulhem[1], **Jean-Pierre Chevallet**[1], **Nicolas Cubaud**[2]

1. UJF-Grenoble;1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France - 2. GlobeVIP, Lyon
{Philippe.Mulhem,Jean-Pierre.Chevallet}@imag.fr, cubaud@globevip.com

RÉSUMÉ. *Cet article décrit le système IOTA-EyeSnap, un système de recherche d'images mobile appliqué à la recherche de photographies de peintures dans le cadre du projet CLICIDE. Ce système repose sur un enchaînement d'approches probabilistes : la première à base de modèle de langue sur des graphes représentant les images, et la seconde à base de modèles bayésiens. Nous détaillons les contraintes et les choix réalisés. Ils sont adaptés à la mobilité dans un cadre muséal. Les différents outils permettant une évaluation a priori et a posteriori sont présentés, et les résultats expérimentaux sont décrits et analysés.*

ABSTRACT. *This article depits the EyeSnap system, dedicated to indexing and retrieval of still images with mobile devices for the CLICIDE project. The system is based on a series of probabilistic matchings : the former uses language models on images, and the latter exploits Bayesian models. The constraints and choices, adopted to a mobility context, are depicted. Several tools, dedicted to assess, a priori and a posteriori, are described, and experimental results are presented.*

MOTS-CLÉS : *Indexation, Recherche, Images, Peinture.*

KEYWORDS: *Indexing, Retrieval, Images, Painting.*

1. Introduction

La recherche d'information visuelle consiste à indexer des images, les requêtes pouvant être du texte ou une image. Dans le cas d'une requête visuelle, le système doit retrouver les images les plus *similaires* à l'image requête en se basant sur son contenu. Ce contenu peut être soit des classes d'objets (ex. des voitures, des personnes), soit des instances (ex. une statue, un tableau). Dans ce dernier cas, il s'agit de reconnaitre uniquement un objet particulier, ou un exemplaire d'un objet existant en plusieurs copies (ex. un produit manufacturé). La recherche d'instances, cadre du travail présenté ici, est plus simple que l'identification de classes d'objets, car on suppose que l'objet à reconnaître n'a intrinsèquement pas de variation visuelle. La variation réside seulement dans la prise de vue, avec des variations de luminosité, de cadrage, et de contexte autour de l'objet. En situation de mobilité, une requête visuelle est plus adaptée qu'une description textuelle. Il *suffit* en effet de prendre un objet en photo, pour lancer une requête à propos de cet objet. Notons que l'image n'est pas le *but* de la recherche mais le *moyen* d'accéder à de l'information relative à l'objet.

Nous avons développé dans le cadre du projet CLICIDE, avec les sociétés GlobeVIP[1] et Ophrys, un système de guidage et d'accès à de l'information à l'intérieur d'un musée, en interrogeant le système à l'aide d'images des œuvres. Le projet CLICIDE vise à permettre à un utilisateur d'accéder, en cours de visite, à des informations sur les œuvres du musée. Il peut également assister à la visite et au guidage à l'intérieur du musée, en suggérant un parcours des œuvres et des salles. Les contraintes fortes du musée (pas de borne WIFI, pas de codes, de marqueurs explicites, etc) ne laissent que l'image comme capteur d'information et de localisation[2].

2. Etat de l'art

L'accès à de l'information en mobilité par l'image s'est développé avant la disponibilité de terminaux portables performants. Par exemple Feiner et al. ont étudié l'usage de la superposition d'informations sur des bâtiments en réalité augmentée pour accéder à de l'information sur le Web (Feiner *et al.*, 1997). Des applications d'informations ajoutées sur l'image affichée du téléphone ont également été imaginées (Haritaoglu, 2001). Pour notre part, nous avons expérimenté avec Snap2Tell (Chevallet *et al.*, 2005, Chevallet *et al.*, 2007) un calcul de caractéristiques simples sur un téléphone (histogrammes de couleurs).

Le système iScope (Zhu *et al.*, 2009) utilise des information de contexte (date, position et auteur de l'image) pour assister la recherche des images similaires à une image requête. Dans ce système, la similarité visuelle est calculée à partir de différentes caractéristiques (histogrammes de couleurs, coefficients d'ondelettes, régions, SIFT, etc). Les images de la base sont organisées en classes (clusters) hiérarchiques,

1. http ://www.globevip.com
2. Le GPS n'est pas disponible à l'intérieur d'un bâtiment.

lors de l'indexation. La recherche se réalise alors uniquement sur l'arbre des centroïdes des clusters. Cette technique permet de réduire les calculs de correspondance pour rendre l'interface de recherche et de navigation plus réactive. Le système PhoneGuide (Bruns *et al.,* 2007) fusionne des informations explicites de localisation (ex. une borne WIFI) avec des informations implicites de contexte (la texture et la couleur des sols). Il a été expérimenté dans le musée de la ville de Weimar. Les travaux plus récent comme Chen (Chen *et al.,* 2011) s'intéressent au temps de réponse du SRI en explorant des structures de stockage des index pour une correspondance *sous linéaire*.

De manière plus générale, les systèmes d'accès par le contenu visuel actuels (Girod *et al.,* 2011) utilisent la technique du *sac de mots visuels* (Csurka *et al.,* 2004). Cette technique s'inspire de la recherche d'information sur le texte. Elle consiste a produire un vocabulaire visuel à partir de l'analyse de la collection des images à indexer. Ce vocabulaire est une abstraction de l'ensemble des caractéristiques (features) calculées sur chaque image. La transformation des images dans un vocabulaire visuel, permet ensuite d'exploiter des modèles de Recherche d'Information. Parmi tous les modèles, nous avons opté pour un modèle de langue, car nous avons pu l'étendre simplement pour en faire un modèle de correspondance sur des graphes (Pham *et al.,* 2011).

Les modèles de langues occupent une place de premier plan dans le domaine de la recherche d'information depuis la fin des années 90 (Ponte *et al.,* 1998). Ces modèles consistent à comparer les distributions probabilistes des documents et celles de la requête posée par un utilisateur. Les éléments clés de ces modèles sont : 1) les variables aléatoires considérées, 2) l'hypothèse de la distribution probabiliste sous-jacente aux documents, 3) le lissage indispensable pour éviter le problème des probabilités nulles, et 4) la formule du calcul de la correspondance entre documents et requêtes. L'hypothèse de distribution multinomiale, proposée par Song et al. (Song *et al.,* 1999), est actuellement la plus utilisée. La correspondance entre documents et requêtes est exprimée par la probabilité $P(Q|D)$, qui estime dans quelle mesure le document D peut produire la requête Q.

Notre approche consiste alors à calculer un vocabulaire global, puis à indexer chaque image à l'aide de ce vocabulaire. Chaque image devient alors un document qui peut être indexé par un Système de Recherche d'Information. Dans la suite nous détaillons le système IOTA-EyeSnap qui met en œuvre la RI d'images en mobilité.

3. IOTA et EyeSnap

3.1. *Architecture*

Les choix spécifiques de conception de l'architecture d'une application mobile impacte la conception de toute l'application informatique. Les problèmes rencontrés sont nombreux et parfois sans solution idéale globale, car il faut équilibrer des paramètres antinomiques (comme l'autonomie en énergie et celle par rapport au serveur), et tenir compte de l'évolution rapide et constante de ces matériels mobiles. Un exemple typique d'équilibre concerne la limitation des communications pour avoir des réponses

rapides, sans réaliser pour autant de traitements locaux coûteux, sacrifiant la faible autonomie des batteries. D'autres difficultés non techniques peuvent survenir, comme le respect de la vie privée, ou le droit de prendre en photographie des oeuvres privées ou des objets manufacturés dans un commerce. Le système IOTA-EyeSnap est composé d'un client mobile (EyeSnap) connecté à un serveur qui exécute le moteur de recherche (IOTA). Le client est actuellement développé sur smartphone avec un système iOS (de type iPhone ou iPad). Le système global est composé des modules suivants :

Interface utilisateur (Front Office) : Elle est l'accès principal au système. Elle doit être simple et directe, orientée vers la fonction principale : se renseigner par l'image. En sortie, elle traite les informations reçue des serveurs et met en forme l'affichage des réponses.

Traitement d'image embarqué : Il gère l'acquisition des images et leur pré-traitements en local. Ces pré-traitements sont : la détection automatique de prise de vue "bougée", le recadrage semi-automatique de l'image, son optimisation et sa réduction en taille afin de l'envoyer au serveur dans un délai très court.

Extracteur de caractéristiques : Son rôle est l'extraction des caractéristiques visuelles des images. L'extraction des caractéristiques visuelles repose sur des outils similaires à l'état de l'art comme le calcul de régions d'intérêt de type SIFT (Lowe, 2004).

Moteur de recherche et d'indexation : C'est un module de RI capable de traiter des images vues comme des pseudo documents avec deux types de termes : les termes du vocabulaire visuel et les parties de graphes de termes pour représenter d'organisation des termes dans une image.

Communication avec le service WEB : Le système de RI IOTA sur le serveur est vu comme un service Web qui accepte des requêtes images dans le protocole standard HTTP et renvoi la réponse en XML.

Gestion des informations associées aux images : Cela concerne le stockage et l'accès aux informations associées aux images, à fournir en réponse à une requête. Ces informations sont des informations de contenu, stockées sur un serveur, ou même un simple lien HTTP. Une information sur la fiabilité de reconnaissance visuelle peut être associée à la réponse pour permettre à l'interface de choisir l'information la plus fiable.

Back Office : Son rôle concerne la gestion des indexations, des informations associées et des droits d'accès aux serveurs.

Le serveur d'information effectue le lien entre une image de la base et une source d'information externe chargée de présenter les informations sur les oeuvres photographiées. Le serveur d'interrogation traite les requêtes en utilisant l'index qui a été

généré précédemment à partir d'un corpus de documents images. Le "back office" permet de lancer des tests à priori et de configurer le serveur d'interrogation pour chaque application : le système est configuré spécialement pour chaque client[3] de la société GlobeVIP.

3.2. *Calcul de correspondance par les modèles probabilistes*

La correspondance du serveur IOTA se base sur un ensemble d'images, et un ensemble d'étiquettes. La fonction $F(i) = e$ signifie que l'étiquette e est celle assignée à l'image i du corpus. Le modèle de recherche détermine, en deux étapes, la probabilité pour une image requête d'être décrite par une étiquette :

1) La première étape consiste en la définition d'un *modèle de langues* sur des caractéristiques visuelles à base de sacs de mots visuels[4], modèles incluant des relations spatiales. L'utilisation des relations spatiales dans les index permet d'augmenter la précision de la description de l'image. La correspondance se base sur un modèle de langue génératif produisant, pour une image requête Q et les images i du corpus, la probabilité $P(Q|i)$. Cette correspondance s'inspire de (Pham *et al.*, 2011) : chaque image du corpus ou requête est représentée par deux distributions probabilistes : une pour les termes, une pour les relations. Le résultat de cette première étape est, pour une requête Q, une liste des images similaires à Q triée selon une valeur de similarité décroissante. Bien que nous connaissions les étiquettes de chacune des images du corpus, une seconde étape de filtrage par étiquette est réalisée.

2) La seconde étape établie une distribution de probabilités sur les étiquettes. Elle détermine la probabilité pour une étiquette e génére la requête Q, par :

$$P(Q|e) = \sum_{i \in I} P(Q|i) * P(i|e).$$

Cette seconde phase considère une traduction des l'étiquettes vers les images. Une fonction binaire $\#(i|e)$ dénote le fait qu'une image i du corpus est associée à une étiquette e : $\#(i|e) = 1$ si $F(i) = e$, et $\#(i|e) = 0$ sinon.

La probabilité $P(i|e)$ est basée sur la fonction $\#(i|e)$, qui indique la probabilité de tirer l'image i parmi l'ensemble des images annotées par e. Elle est calculée par marginalisation sur le nombre d'images annotées par l'étiquette e :

$$p(i|e) = \frac{\#(i|e)}{\sum_{i \in I} \#(i|e)}$$

Cette seconde phase est approximée en se limitant aux n premières images renvoyées par la première étape. Les paramètres, comme les valeurs de lissages pour le modèle de langue sur les graphes, doivent être optimisés à chaque indexation. Le back office qui a été développé permet d'optimiser ces paramètres de manières automatique, en lançant

3. Au sens commercial.
4. Bag of Visual Words (Csurka *et al.*, 2004)

une série d'évaluations des index, soit à partir de requêtes types, si elles existes, soit en prenant successivement chaque image du corpus comme une requête.

3.3. *EyeSnap : le front et back office*

EyeSnap est composé de deux parties principales, le front office qui prend en compte les requêtes d'un utilisateur en mobilité, et le back office qui a pour but de gérer l'indexation, et de tester différentes configurations d'indexation et de recherche.

L'interface de prise de vue présentée en figure 1 permet de rogner l'image de manière semi-automatique : le cadre calculé par quatre transformées de Hough sur le mobile, peut être modifiée manuellement. Ce calcul de cadre permet de supprimer une partie du contexte de l'image pour augmenter le taux de reconnaissance comme nous le montrons dans la section dédiée aux résultats expérimentaux.

Figure 1. *Interface sur le mobile : rognage semi-automatique.*

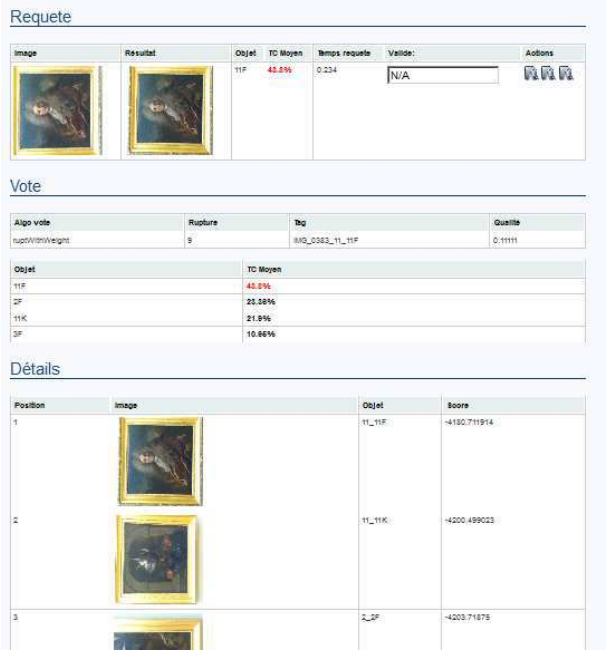
Le back office permet de tester la discrimination des clusters, de lancer des tests et de visualiser des réponses à des images de test comme le montre la figure 2. Plusieurs index peuvent être générés pour un corpus et manipulés pour les évaluer en parallèle. L'objectif est alors de déterminer quelle est la meilleure configuration pour un corpus donné.

4. Expérimentations et résultats

4.1. *Collection de test*

CLICIDE est un projet financé par le Fond Unique Interministériel (FUI), en partenariat avec le Pôle de Compétitivité Imaginove et la Région Rhône-Alpes. Il associe la Société GlobeVip, qui fournit l'infrastructure, la société Ophrys qui est responsable du contenu, et le laboratoire LIG qui fournit la technologie d'analyse des images et

Figure 2. *L'interface du back-office : analyse du résultat d'une requête de test.*



de calcul de correspondance. Le projet se déroule avec le concours du musée de Grenoble.

Nous effectuons nos expérimentations sur la collection permanente du musée, pour laquelle 3425 images correspondant à 512 œuvres ont été prises par les 3 acteurs du projet. Cette collection a été constituée un jour de fermeture du musée, mais avec les mêmes éclairages que lors de l'accueil du public. Le protocole choisi pour la construction de ce corpus est le suivant :

1) Choix des personnes impliquées dans la réalisation du corpus : un ensemble de 4 personnes reliées au projet, mais non spécialistes de recherche d'images, ont créé la collection (corpus et ensemble de requêtes). Cet ensemble est composé d'une femme et de 3 hommes, de 25 à 45 ans.

2) Prise des photographies de tests (requêtes) par les personnes choisies, sans contrainte, dans un premier temps. Ce choix permet de considérer que les photographies de tests sont réalistes, *i.e.* elles ne sont pas faites par des personnes sachant comment les recherches sont effectuées.

3) Prise des photographies du corpus par les personnes choisies, suivant des indications à suivre pour les prises de vues.

Les requêtes (177 images) sont établies sur la collection permanente du musée, en insistant bien sur le fait que les personnes doivent prendre des clichés en temps que visiteur qui chercherait des renseignements sur des œuvres complètes et non des détails de ces œuvres. Certaines de ces images ne correspondaient pas à ce qui était demandé. Après filtrage (photographies de peintures, et photographies non floues) nous avons généré un ensemble de 162 images requêtes "filtrées".

Lors la construction du corpus, chaque personne a été assignée à un sous ensemble de salles du musée. Pour chaque salle, la personne en charge devait effectuer plusieurs clichés de chaque œuvre (prioritairement des peintures), avec un outil mobile similaire à celui qui sera utilisé dans l'application finale pour la recherche. Le protocole suivant est défini sur les prises de vues pour chaque œuvre : a) prendre une photographie du cartouche qui décrit l'œuvre dans le musée ; b) prendre l'œuvre en totalité ; c) prendre des parties de l'œuvre par un découpage en zones avec chevauchement. Le principe de prendre des détails des œuvres permet de prévoir le cas où un utilisateur se focalise sur une partie d'une peinture.

4.2. *Evaluation expérimentale*

Le système doit renvoyer l'identifiant de l'œuvre photographiée par un visiteur avec son mobile. Nous avons effectué une évaluation basée sur un taux de reconnaissance, qui est le pourcentage de requêtes résolues correctement : l'œuvre a été effectivement reconnue. En plus de ces taux globaux, une évaluation par personne, mesure leur cohérence. Un test sur quatre personnes n'est pas significatif, mais cela permet néanmoins d'avoir une idée du comportement du système en terme de stabilité.

Les caractéristiques visuelles extraites sont de type SIFT couleur, en utilisant notre propre code d'extraction de caractéristiques. Un calcul de regroupement (clustering) à base de Kmoyennes a permis de définir un vocabulaire visuel de 2000 dimensions, sur lequel des relations de type "gauche/droite" et "dessus/dessous" sont extraites pour les graphes. La correspondance utilise un modèle de langue à base de lissage de Jelinek-Mercer, et conformément à d'autres expérimentations effectués sur d'autres corpus, nous avons fixé le paramètre λ à 0.7 .

4.3. *Résultats obtenus*

Les résultats obtenus pour les requêtes brutes, par personne et globaux, sur le corpus, sont présentés dans la partie "Requetes brutes" du tableau 1. On remarque que la personne A a effectué des requêtes qui semblent plus simples pour le système. Ces résultats sont de bonne qualité, avec 70% de taux de reconnaissance corrects.

Après étude de ces images requêtes, nous nous sommes rendu compte que le cadrage des images par A était plus "précis" que celui des autres utilisateurs : les murs et le sols apparaissent pas ou peu dans ses clichés. Cela nous a amené à effectuer une seconde évaluation, après cadrage a posteriori des images requêtes de tous les utili-

Tableau 1. *Résultats par personne et global, pour les images brutes et filtrées.*

	Requêtes brutes		Requêtes filtrées	
Personne	Nb. requ.	Taux de reco.	Nb. requ.	Taux de reco.
A	60	85.00%	58	91.38%
B	44	61.36%	44	86.36%
C	43	62.79%	36	86.11%
D	30	63.33%	24	75.00%
TOTAL	177	70.06%	162	86.42%

sateurs. Ce second jeux de test ôte également des images très flous et des clichés de sculptures : les images floues pourraient être détectées automatiquement par l'accéléromètre de l'outil mobile mais pas par analyse d'image car les peintures ont des larges aplats de couleurs qui peuvent suggérer un flou... Les résultats obtenus pour les requêtes filtrées, par personne et globaux, sur le corpus, sont présentés dans partie droite du tableau 1. On constate que le cadrage des images avec élimination des éléments perturbateurs produit une nette amélioration (entre 12% et 25%) des résultats pour tous les utilisateurs et diminue les différences entre utilisateurs (on passe de 24% d'écart entre A et B pour les requêtes brutes à 16% entre A et D pour les requêtes filtrées). Il n'en reste pas moins qu'il existe une variabilité importante entre utilisateurs dans les deux cas de figure, et qu'un système industriel doit minimiser ces variations.

5. Conclusion

Cet article décrit une application mobile commerciale pour la recherche d'images en mobilité, issue du monde de la recherche. Nous avons présenté les particularités et les problèmes liés à la recherche d'images en mobilité. Nous avons également décrit le système IOTA-EyeSnap. Le développement de ce système est réalisé dans le cadre d'une collaboration scientifique entre l'Université Pierre Mendès France de Grenoble (UPMF), le Centre National de la Recherche scientifique (CNRS), d'une part et la société GlobeVIP qui a acquis une licence du logiciel IOTA. Dans le cadre du projet CLICIDE, un effort particulier a porté sur la mise en place d'un protocole de mise en œuvre d'une collection de test afin de vérifier la qualité des résultats obtenues par le système sur la collection permanente du musée de Grenoble. Les bons résultats obtenus vont déboucher, courant 2013 à la mise en service de ce système dans le cadre des visites guidées du Musée de Grenoble.

Remerciements

Nous remercions la Région Rhône Alpes, pour son soutien et son financement dans le cadre du projet CLICIDE, ainsi que le projet LIMA du cluster de recherche Rhône-Alpes ISLE.

6. Bibliographie

Bruns E., Brombach B., Zeidler T., Bimber O., « Enabling Mobile Phones To Support Large-Scale Museum Guidance », *IEEE MultiMedia*, vol. 14, n° 2, p. 16-25, April, 2007.

Chen X., Koskela M., « Mobile visual search from dynamic image databases », *Proceedings of the 17th Scandinavian conference on Image analysis*, SCIA'11, Springer-Verlag, Berlin, Heidelberg, p. 196-205, 2011.

Chevallet J.-P., Lim J.-H., « SnapToTell : Accès ubiquitaire à de l'information multimédia à partir d'un téléphone portable. Application sur une base d'images de Singapour », *CORIA05 COnférence en Recherche Information, Grenoble (France)*, p. 245-260, 9–11 mars, 2005.

Chevallet J.-P., Lim J.-H., Leong M.-K., « Object Identification and Retrieval from Efficient Image Matching Snap2Tell with the STOIC dataset », *Information Processing & Management*, vol. 43, n° 2, p. 515-530, March, 2007.

Csurka G., Dance C. R., Fan L., Willamowski J., Bray C., « Visual categorization with bags of keypoints », *In Workshop on Statistical Learning in Computer Vision, ECCV*, p. 1-22, 2004.

Feiner S., MacIntyre B., Höllerer T., Webster A., « A touring machine : Prototyping 3D mobile augmented reality systems for exploring the urban environment », *Proceedings ISWC '97 (First IEEE Int. Symp. on Wearable Computers)*, vol. 1(4) of *Personal Technologies*, p. 208-217, October 13-14, 1997.

Girod B., Chandrasekhar V., Chen D., Cheung N.-M., Grzeszczuk R., Reznik Y., Takacs G., Tsai S., Vedantham R., « Mobile Visual Search », *Signal Processing Magazine, IEEE*, vol. 28, n° 4, p. 61 -76, july, 2011.

Haritaoglu I., « InfoScope : Link from Real World to Digital Information Space », *in* , G. Abowd, , B. Brumitt, , S. Shafer (eds), *Ubicomp 2001 : Ubiquitous Computing : Third International Conference Atlanta, Georgia, USA,*, vol. 2201/2001 of *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, p. 247-255, September 30 - October 2, 2001.

Lowe D. G., « Distinctive Image Features from Scale-Invariant Keypoints », *Int. J. Comput. Vision*, vol. 60, n° 2, p. 91-110, November, 2004.

Pham T.-T., Mulhem P., Maisonnasse L., Gaussier E., Lim J.-H., « Visual Graph Modeling for Scene Recognition and Mobile Robot Localization », *Multimedia Tools and Applications*p. 1-23, Jan, 2011. Springer Science.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, ACM, New York, NY, USA, p. 275-281, 1998.

Song F., Croft W. B., « A general language model for information retrieval », *Proceedings of the eighth international conference on Information and knowledge management*, CIKM '99, ACM, New York, NY, USA, p. 316-321, 1999.

Zhu C., Li K., Lv Q., Shang L., Dick R. P., « iScope : personalized multi-modality image search for mobile devices », *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, ACM, New York, NY, USA, p. 277-290, 2009.