Étude préliminaire à la recherche de photographies muséales en mobilité

Maxime Portaz, Philippe Mulhem, Jean-Pierre Chevallet

Université Grenoble Alpes / CNRS, LIG UMR 5217, Grenoble, F-38041, France
{Maxime.Portaz, Philippe.Mulhem, Jean-Pierre.Chevallet}@imag.fr

RÉSUMÉ. *Cet article étudie la problématique de l'indexation et de la recherche d'image dans le cadre de visites de musée. Nous nous intéressons en particulier au cas d'utilisation d'outils mobiles "hors ligne" (c'est-à-dire sans connexion à un serveur distant), du point de vue qualité intrinsèque et du point de vue application mobile. Nous décrivons trois approches de référence, et nous étudions leur comportement qualitatif sur une collection de photographies de peintures, prises par des outils mobiles dans le Musée de Grenoble.*

ABSTRACT. *This paper studies the problem of images indexing and retrieval related to museum visits. We especially focus on "offline" use of mobile devices (i.e., without connection to a remote server), from the point view of intrinsic quality and the point of view of mobile potential use. We describe three approaches, and we study their qualitative behavior on a test collection of photographs of paintings taken by mobile devices in the Grenoble museum.*

MOTS-CLÉS : *Recherche d'images, Reconnaissance d'objets, Recherche d'information.*

KEYWORDS: *Image Retrieval, Object recognition, Information Retrieval.*

1. Introduction

Cet article présente une première étape dans la réalisation de recherche d'images par l'exemple, dans un cadre mobile durant des visites de musées. L'idée de base est qu'un visiteur puisse, par un appareil mobile fourni par le musée, obtenir des informations sur les œuvres qu'il photographie. A partir de cette idée générale, l'approche que nous suivons dans cet article est double : nous présentons trois des approches les plus classiques dans le domaine de l'indexation et la recherche de photographies en les étudiant dans une application mobile autonome (sans connexion avec un serveur extérieur), avant de réaliser un ensemble de tests qualitatifs sur des versions de base de ces approches.

L'indexation et la recherche d'images passent par deux étapes. La première, durant l'indexation, est l'extraction de caractéristiques visuelles et la représentation des images à partir des caractéristiques extraites. La deuxième, durant la recherche, est la représentation de l'image requête et par une correspondance entre la requête et les images du corpus afin de renvoyer les images triées suivant leur valeur de similarité avec la requête. Les trois approches que nous considérons ici sont les suivantes : i) des approches à base de représentations continues des images, qui consistent à garder toutes les caractéristiques extraites ; ii) des approches à base sacs de mots visuels, qui synthétisent les caractéristiques extraites des images par un vecteur ; et enfin iii) les approches plus récentes liées à l'utilisation de réseaux de neurones profonds (*Deep Learning*) pour la représentation des images. Ces trois approches sont détaillées dans la section 2 pour la partie indexation et dans la section 3 pour leur utilisation lors de la recherche d'images.

La section 4 portera ensuite sur la description des spécificités matérielles et logicielles des plate-formes mobiles, et sur l'adéquation sur les 3 approches d'indexation et de recherche d'image dans ce cadre. Afin d'avoir une idée plus précise en terme de qualité de ces approches dans le cadre visé, c'est-à-dire les objets visuels dans les musées, nous présentons des résultats préliminaires obtenus avec des systèmes de bases des 3 approches présentées sur des données muséales du musée de Grenoble (Mulhem *et al.*, 2013). Cette étude nous permet de déterminer comment ses approches s'adaptent par défaut à un exemple de données spécifique. Nous concluons ensuite en section 6.

2. Représentation d'image

La recherche d'image passe d'abord par la construction d'une représentation des images pour pouvoir les manipuler et permettre leur comparaison lors de la recherche. De nombreuses méthodes existent dans l'état de l'art pour construire une représentation des images. On peut cependant tirer les grandes lignes suivantes des approches actuelles : a) une première étape repose sur la localisation de régions d'intérêts autour de points remarquables, b) une seconde étape repose sur l'extraction des descripteurs

des régions déterminées précédemment, c) la troisième étape détermine la représentation d'une image complète à partir des caractéristiques extraites.

On retrouve, sous une forme ou une autre, ces étapes dans les trois grandes catégories d'approches pour l'indexation et la recherche d'images que nous considérons : les approches basées sur des sacs des mots visuels (Csurka *et al.*, 2004), celles basées sur des représentations continues (Calonder *et al.*, 2010), et enfin celles à base de réseaux de neurones convolutifs (Simonyan et Zisserman, 2014).

2.1. *SIFT - Sac de mots*

La première méthode, appelée "sac de mots" construit une représentation de l'image en se basant sur un vocabulaire visuel calculé a priori. Ce vocabulaire identifie des "concepts visuels", ou *mots visuels*, qui sont des représentants remarquables d'ensembles de descripteurs visuels. Comme montré sur la figure 1, il est créé grâce à un partitionnement non supervisé d'un grand ensemble de descripteurs extraits d'images, avec l'algorithme des K-Moyennes. Chaque image est alors représentée par le nombre d'occurrence de chaque mot visuel possible, d'une manière très similaire à la *term frequency* en recherche d'information textuelle.

Les descripteurs SIFT, introduits en 2004 par Lowe (Lowe, 2004), sont des descripteurs invariant à l'échelle. Ils ont démontré leurs performances par rapport aux autres descripteurs, que ce soit en reconnaissance d'objets (Mikolajczyk et Schmid, 2005) ou en correspondance d'images (Se *et al.*, 2002). Il sont également utilisés avec des sacs de mots (Csurka *et al.*, 2004).

Pour la représentation d'images à base de SIFT, la détection de régions d'intérêt (avec échelle et orientation) se fait par différence de gaussiennes (Lowe, 1999). A partir de ces régions, les descripteurs SIFT sont calculés. Chaque descripteur (un vecteur à 128 dimensions réelles) est basé sur des différences de niveaux de gris sur huit directions, accordant davantage d'importance au centre de la région d'intérêt. Pour l'étape de représentation d'une image, chaque descripteur SIFT extrait est assigné (par calcul de plus proche voisin) à un mot du vocabulaire visuel. C'est lors de cette assignation que chaque descripteur de l'image est abstrait en un concept visuels, et que l'image entière est représentée par un vecteur du nombre d'occurrences de chaque mot visuel.

Dans cette approche à base de SIFT, le calcul des régions d'intérêt est assez gourmand en temps de calcul : il faut appliquer des flous gaussiens sur l'image d'origine pour trouver les régions les plus intéressantes (i.e., celles qui ont les plus grosses variations de luminosité). Dans cette approche, l'étape d'abstraction possède l'avantage de représenter une image de manière très compacte, au détriment du niveau de détail. En effet, de nombreux descripteurs peuvent être assignés au même concept visuel : si tout se passe bien ils sont très similaires visuellement, mais cela n'est pas toujours vrai. Des améliorations ont été apportées à ce modèle initial, comme les VLAD (Jégou *et al.*, 2010), qui consistent à dépasser une assignation simples à base de plus proche

voisin. Le calcul est le même qu'avec les sacs de mots, à la différence que l'on cumule l'écart entre le descripteur local et le centroïde.

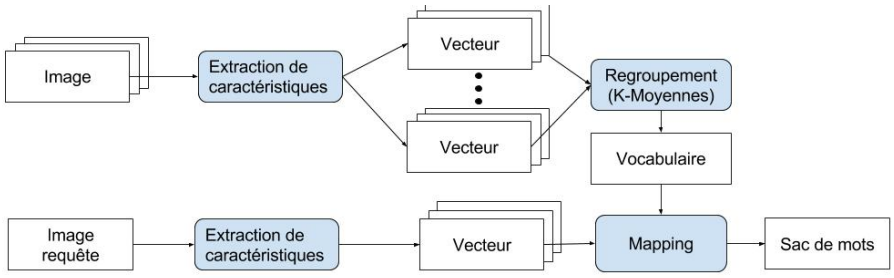

Figure 1. *Construction de vocabulaire et indexation d'image par sac de mots visuels.*

2.2. *ORB - Représentation continue*

Les méthodes à base de représentation continue des images évitent l'utilisation de dictionnaire visuel comme dans les approches à sacs de sacs de mots, mais conservent les descripteurs de l'image afin de les utiliser tels quels lors de la recherche. Pour être utilisables, ces approches nécessitent des descripteurs très rapides à calculer, et qui prennent peu d'espace mémoire. Les descripteurs ORB (Rublee *et al.*, 2011) - Oriented FAST Rotated BRIEF - valident ces deux critères.

Comme dans l'approche par sacs de mots, il faut tout d'abord déterminer des régions d'intérêt des images. Elles sont détectées par un détecteur FAST - *Features from Accelerated Segment Test* - (Rosten et Drummond, 2006). Comme l'approche FAST initiale ne permet pas de déterminer l'information d'orientation qui caractérise une région, l'extension proposée dans (Rublee *et al.*, 2011) décrit les *oFAST - Oriented FAST* -, détermine l'orientation de la région grâce à la méthode "*intensity centroid*" (Rosin, 1999).

Le descripteur BRIEF - *Binary Robust Independent Elementary Features* - (Calonder *et al.*, 2010) est ensuite utilisé pour décrire chaque région d'intérêt. Un descripteur BRIEF applique, en utilisant le centre de la région et des points en périphérie, un test binaire qui mesure le signe des variations d'intensité entre deux points, suivant l'équation (1), où $p(x)$ est l'intensité de l'image p au point x.

$$f(p,x,y) = \begin{cases} 1 & p(x) < p(y) \\ 0 & p(x) \geq p(y) \end{cases} \quad [1]$$

La liste des résultats de ces tests binaires est stocké dans un vecteur de n bits (créé à partir des points comparés), suivant l'équation (2). Pour le descripteur ORB, $n = 256$.

$$u_n(p) = \sum_{1 \leq i \leq n} 2^{i-1} f(p, x_i, y_i) \quad [2]$$

(Rublee *et al.*, 2011) ajoutent une partie d'apprentissage pour déterminer les meilleurs paires de points à sélectionner, en : a) minimisant la corrélation entre les paires pour que chaque paire apporte de l'information, et b) maximisant la variance des pairs, pour qu'elles soient discriminantes.

Calculer les descripteurs ORB est très rapide. La taille du descripteur ORB permet de conserver en mémoire les descripteurs des images du corpus afin d'éviter la perte d'information liée à la création du dictionnaire visuel.

2.3. *Réseaux de neurones convolutifs*

Depuis quelques années, les réseaux de neurones convolutifs profonds (CNN) se sont imposés en vision par ordinateur, et ont radicalement changé les domaines de recherche d'image et d'objets. Les réseaux de neurones profonds sont constitués d'un grand nombre de neurones, organisés en couches successives, qui comportent des couches qui extraient des descripteurs et des couches qui représentent les images, toutes ces couches étant apprises conjointement. Comme montré sur la figure 2, un CNN est composé de couches de convolution et de sous-échantillonnage, dans le but d'extraire des représentations de plus en plus abstraite de l'image. Les dernières couches sont des couches entièrement connectées. L'apprentissage du réseau se fait grâce à la rétro-propagation (LeCun *et al.*, 1989), qui permet d'adapter le poids de chaque neurone pour correspondre le mieux possible à la sortie. Ce qui implique que plus on possède de neurones à entraîner, plus le nombre de données nécessaires à l'apprentissage est importante.

La collection ImageNet, composée de plusieurs millions d'images étiquetées, permet de réaliser l'apprentissage d'un très grand CNN. Cette collection a été utilisée par K. Simonyan et A. Zisserman (Simonyan et Zisserman, 2014). Pour donner une idée de la difficulté d'apprentissage dans un CNN, ils ont eu besoin de 3 semaines de calcul avec GPUs de dernière génération pour calculer les 144 millions de paramètres présents dans un CNN à 19 couches, permettant d'obtenir les meilleurs résultats sur ILSVRC-2012 avec 7,3% d'erreur.

L'avantage des réseaux de neurones profonds se situe dans le fait qu'ils apprennent tous les paramètres nécessaires à la détection d'objet automatique. L'image est passée à travers plusieurs couches de neurones, il n'y a donc pas d'extraction a priori de caractéristiques visuelles par descripteurs figées. Les caractéristiques extraites sont celles qui sont nécessaires au réseau pour calculer la sortie.

3. Recherche d'images

Nous regroupons ici d'un côté les approches à base de sac de mots et de réseaux convolutifs, où chaque image est un vecteur d'entiers ou de réels, et de l'autre l'approche continue qui représente chaque image par un ensemble de vecteurs binaires.

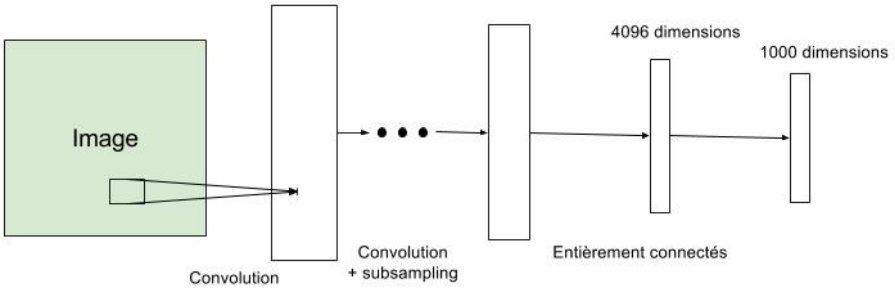

Figure 2. *Organisation d'un réseau convolutif profond.*

3.1. *Un vecteur par image*

Avec la méthode de sac de mots ou l'extraction données grâce au réseau de neurone, une image est représentée par un vecteur de réels ou d'entiers. Étant donnée une image comme requête, nous transformons l'image en sa représentation vectorielle, et nous calculons la distance entre cette image et les images de la base.

Dans le cas des sacs de mots visuels, une implantation utilisant des fichiers inverse, pour accélérer la recherche (Sivic et Zisserman, 2003 ; Jegou *et al.*, 2008), est possible. Pour le représentation extraite du réseaux de neurones, la mesure de similarité classique $s(u, v)$ entre les vecteurs u et v est le cosinus.

3.2. *Un ensemble de vecteurs par image*

Avec la représentation continue de l'image, le calcul de similarité se fait par comparaison directe entre ceux de l'image requête et ceux des images du corpus. Comme décrit en figure 3, pour chaque image nous avons un ensemble de vecteurs, et nous devons comparer chaque vecteur entre eux pour trouver le plus proche. Les implantations utilisant les représentations Local Sensitive Hashing (LSH) (Gionis *et al.*, 1999) permettent d'obtenir très rapidement les plus proches voisins sur des descripteurs binaires. La mesure de similarité est le nombre de descripteurs correspondant entre deux images. Une vérification géométrique peut également être fait pour vérifier que la disposition des descripteur d'une image à l'autre. Ces deux méthodes permettent de trouver l'image la plus proche dans le corpus.

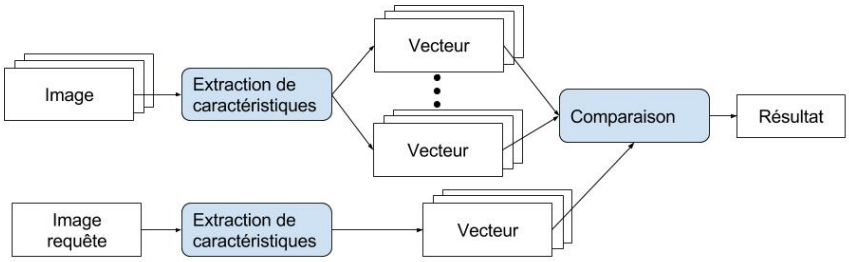

Figure 3. *Traitement d'une requête avec représentation d'image par un ensemble de vecteurs.*

4. Comparaison des approches pour le contexte en mobilité

Comme nous souhaitons à terme porter le système sur mobile, nous nous concentrons sur les points importants dans le cadre d'un développement sur plate-forme mobile : temps de calcul et occupation mémoire.

Comme les calculs nécessaires pour mettre en place la création du dictionnaire visuel dans le cas des sacs de mots, ou l'entraînement du réseau de neurones pour les CNN peuvent être effectués sur serveur, ils sont donc hors de notre étude.

Pour les représentations à base de sacs de mots, une fois l'extraction de descripteurs et le mapping effectués, la recherche d'image est particulièrement efficace, notamment si on utilise une représentation à base de fichier inverse.

Dans le cas de l'approche continue, le pré-traitement consiste uniquement à extraire l'ensemble des descripteurs pour les images de la base. Dans le cas des descripteur ORB, les calculs sont relativement rapide. L'étape de correspondance est, par contre, plus gourmande en temps de calcul, car il faut comparer l'ensemble des descripteurs des images. Notons également que l'ensemble des descripteurs des images du corpus doit être sauvegardé en mémoire, d'où une utilisation importante de celle-ci. Dans le cas d'une application mobile, l'utilisation de LSH est indispensable, et règle ses problèmes.

Pour les réseaux de neurones, une fois l'apprentissage du réseau effectué sur des serveurs, son utilisation est une opération rapide, notamment avec un processeur graphique, même mobile. Une grande quantité de mémoire est cependant nécessaire pour stocker les paramètres du réseau. Le calcul de correspondance est alors très rapide si on utilise des représentations adaptées. Par contre, l'utilisation de fichiers inverses n'existe pas dans l'état de l'art actuel.

La tableau 1 regroupe les points faibles et points forts de chacune des méthodes étudiées suivant les deux axes d'analyse considérés. On constate qu'a priori les approches par sacs de mots sont plus adaptées à une situation en mobilité, de par leur faible besoin de mémoire et de par la rapidité de la recherche dans la base de donnée.

Système	Pré-traitement	Extraction de données	Recherche	Mémoire
BOW-SIFT	-	-	+	+
ORB	+	+	-	-
CNN	-	+/-	+	-

Tableau 1. *Points forts (+) et faibles (-) de chaque approche considérée.*

5. Comparaison Expérimentale

Les expérimentations préliminaires que nous présentons ici ne se préoccupent pas des aspects de traitements liés à la mobilité, mais se focalisent sur l'utilisation d'une implantation-type de chacune des 3 catégories d'approches décrites précédemment. Ces approches sont testées sur une collection de photographies d'œuvres artistiques prises par des outils mobiles. Ce corpus, défini dans (Mulhem *et al.*, 2013), est composé de 3425 images correspondant à 512 œuvres, et de 177 images requêtes. Chaque photographie à été prise par un outil mobile (iPod Touch). Le but étant de simuler la visite d'un musée, dans lequel l'utilisateur peut prendre en photo une œuvre pour obtenir des informations sur celle-ci. Les configurations que nous avons testées sont les suivantes :

– approche sac de mots - SIFT, notée *BOW-SIFT*. Le vocabulaire visuel de 2000 dimensions est créé à partir du corpus. La similarité entre image requête et image du corpus est l'inverse de la distance euclidienne entre les vecteurs. L'implantation utilise les outils par défauts proposés par OpenCV[1].

– approche par représentation continue, notée *ORB*. Le calcul de similarité n'utilise pas le calcul de correspondance rapide basé sur LSH, mais la cardinalité de l'intersection utilisée entre les descripteurs d'une image et ceux de la requête fournit les même résultats qu'une utilisation de LSH. L'implantation utilise les outils par défauts proposés par OpenCV.

– approche par réseau de neurone, notée *CNN*. Nous utilisons le réseaux de neurones généré par (Simonyan et Zisserman, 2014), qui a prouvé sa qualité. Comme ce réseau à appris à reconnaître les 1000 classes du challenge ILSVRC, la dernière couche de neurones (la sortie du réseau) est un vecteur de 1000 réels. Afin avoir une représentation des images moins spécifique à ILSVRC, nous utilisons l'avant dernière couche de ce réseau comme sortie, qui est un vecteur de dimension 4096. Le calcul de similarité utilisé entre image est le cosinus.

Nos résultats sont également comparés, pour Reco@1, à ceux de (Mulhem *et al.*, 2013), basés sur une approche sac de mots avec des SIFT couleur, avec des calculs de correspondance entre requête et documents par modèles de langue. Nous avons utilisé les deux mesures d'évaluation : le taux de reconnaissance à un document, Reco@1, et le taux de reconnaissance à 5 documents, Reco@5. La mesure Reco@5 se place dans

1. http ://opencv.org

le cadre où l'on désirerait appliquer des algorithmes de vote sur la liste des résultats obtenus, comme si nous disposions d'une vidéo ou d'un ensemble de photos pour chaque requête.

system	Reco@1	Reco@5
(Mulhem *et al.*, 2013)	70,06%	/
BOW-SIFT	68.89%*	93.26%*
ORB	54.23%	89.26%
CNN	49.72%	56.50%

Tableau 2. *Résultats obtenus sur la collection (Mulhem* et al.*, 2013) (* : résultats préliminaires, sur une partie de la collection).*

Nous constatons dans la table 2 des taux de reconnaissances pour tous nos tests nettement inférieurs à la référence. En particulier, à notre surprise, les réseaux de neurones profonds ne se comportent pas aussi bien que les deux autres approches. Ceci est sans doute dû à la spécificité des images de la collection (photographies de peintures prises par des appareils mobiles par des amateurs). Rappelons que nous avons utilisé dans ces expérimentations des approches avec leur configurations "de base", sans aucune adaptation. A la vue des résultats obtenus, il résulte que pour toutes ces approches nous devons tout d'abord nous préoccuper des paramètres qui les feront se comporter mieux sur des corpus muséaux avant d'aller plus loin.

6. Conclusion

Nous avons proposé dans cet article une description synthétique des approches classiques en indexation et recherche d'images, à base respectivement de sac de mots, de représentations continues et de réseaux de neurones convolutifs. Nous avons étudié l'adéquation de ces approches dans le cas spécifique pour la recherche d'image mobile "hors ligne". Nous en avons conclut que ces 3 approches sont utilisables en mobilité, moyennant des restrictions.

Les premières expérimentations sur des photographies de peintures de musée que nous rapportons nous poussent à étudier plus en détail les spécificités des données de musée pour obtenir tout d'abord de bons résultats, avant de se focaliser à l'avenir dans un second temps sur les aspects mobiles et l'étude de l'impact de certaines restriction sur la qualité des données qui nous intéressent.

Remerciements

Ce travail a été effectué dans le cadre du projet Guimuteic financé par le Fonds Européen de Développement Régional (FEDER) et de la région Auvergne Rhône-Alpes.

7. Bibliographie

Calonder M., Lepetit V., Strecha C., Fua P., « Brief : Binary robust independent elementary features », *Computer Vision–ECCV 2010*p. 778-792, 2010.

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.

Gionis A., Indyk P., Motwani R. *et al.*, « Similarity search in high dimensions via hashing », *VLDB*, vol. 99, p. 518-529, 1999.

Jegou H., Douze M., Schmid C., « Hamming embedding and weak geometric consistency for large scale image search », *Computer Vision–ECCV 2008*, Springer, p. 304-317, 2008.

Jégou H., Douze M., Schmid C., Pérez P., « Aggregating local descriptors into a compact image representation », *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, IEEE, p. 3304-3311, 2010.

LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D., « Backpropagation applied to handwritten zip code recognition », *Neural computation*, vol. 1, n° 4, p. 541-551, 1989.

Lowe D. G., « Object recognition from local scale-invariant features », *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, Ieee, p. 1150-1157, 1999.

Lowe D. G., « Distinctive image features from scale-invariant keypoints », *International journal of computer vision*, vol. 60, n° 2, p. 91-110, 2004.

Mikolajczyk K., Schmid C., « A performance evaluation of local descriptors », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, n° 10, p. 1615-1630, 2005.

Mulhem P., Chevallet J., Cubaud N., « Recherche d'images en mobilité : le système IOTA-EyeSnap », *CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013.*, p. 63-72, 2013.

Rosin P. L., « Measuring corner properties », *Computer Vision and Image Understanding*, vol. 73, n° 2, p. 291-307, 1999.

Rosten E., Drummond T., « Machine learning for high-speed corner detection », *Computer Vision–ECCV 2006*, Springer, p. 430-443, 2006.

Rublee E., Rabaud V., Konolige K., Bradski G., « ORB : an efficient alternative to SIFT or SURF », *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, p. 2564-2571, 2011.

Se S., Lowe D., Little J., « Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks », *The international Journal of robotics Research*, vol. 21, n° 8, p. 735-758, 2002.

Simonyan K., Zisserman A., « Very deep convolutional networks for large-scale image recognition », *arXiv preprint arXiv :1409.1556*, 2014.

Sivic J., Zisserman A., « Video Google : A text retrieval approach to object matching in videos », *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, IEEE, p. 1470-1477, 2003.