Recherche bibliographie sur la reconnaissance des locations

à partir de l'article "What's make Paris look like Paris ?"

NGUYEN Van Tho, *Promotion 17, Institut de la Francophonie pour l'Informatique*

***Résumé*—Au cours des dernières années, l'émergence de médias géoréférencées, comme des photos géolocalisées, sur l'Internet a ouvert un tout nouveau monde de possibilités pour la recherche et les applications de la reconnaissance de locations. Cet axe de recherche est atteint beaucoup d'attention des chercheurs dans le domaine de traitement d'image et de vision par ordinateur. Ce rapport présente des recherches récentes sur la reconnaissance de la reconnaissance de locations en mettant l'accent sur les défis, les méthodes, les résultats atteints et les échelles de ces recherches.**

***Index Terms*—Traitement d'image, Vision par ordinateur, Reconnaissance des locations, Fouille de données**

I. Introduction

CES dernières années ont vu une explosion des données d'image sur Internet par une grande augmentation des images partagées sur les réseaux sociaux tels que Facebook et Flickr ou sur Google Street View. La reconnaissance des locations devient un besoin indispensable pour la recherche et la récupération d'images et est un domaine de recherche actif. La reconnaissance de locations partage des défis de la reconnaissance d'objets tels que la variation de point de vue, les différence d'échelle d'objet,le désordre du fond et l'occlusion. En plus de ces défis communs, la reconnaissance de locations a les difficultés de la grande taille de données et le manque des bases de données complètes. Plusieurs approches sont proposées pour résoudre ces problèmes. Cependant, nous pouvons voir deux axes principaux : un pour la reconnaissance de locations à l'échelle globale [HE08 ; Gop13] et un pour la reconnaissance de locations à petite échelle (reconnaissance une location précise dans une ville, par exemple) [Tor+13 ; SLK12 ; CS13]. Dans ce travail, nous présentons une petite revue sur les travaux récents de la reconnaissance de locations à l'échelle globale et à petite échelle.

II. Reconnaissance des locations à l'échelle globale

Le but de la reconnaissance à cette échelle est pour répondre l'image requette est à partir de quelle région (ville ou pays). Une des caractéristiques de ce genre de problème est la base de données d'images de tout la planète. On peut résoudre le problème d'identification de location d'image [HE08 ; Gop13] ou on peut résoudre un problème binaire : si l'image est d'une région [Doe+12]. Les défis de la reconnaissance de locations de cette échelle est d'abord la grande demande de calcul. De plus, en réalité, il est difficile de trouver une collection d'images qui peut couvrir tout la planète.

James Hays et al [HE08] utilise un algorithme simple pour estimer la distribution de locations à partir d'une seule image en utilisant l'approche de la correspondance des scènes. Une base de données de 6 millions d'images GPS-marqués d'une collection en ligne Flickr est utilisée pour l'expérimentation. Dans ce travail, la position de l'image estimée est représentée comme une distribution de probabilité sur la surface de la Terre. Pour atteindre leur but, James Hays et al construisent une base de données en sélectionnant les images qui ont à la fois les coordonnées GPS et des mots-clés géographiques. Ceci augmente beaucoup la probabilité de trouver des données géolocalisées précisément et visuellement.

Pour trouver les correspondance des scènes, James Hays et al [HE08] proposent une méthode qui utilise plusieurs types de caractéristiques : petite image à taille 16x16, histogramme en couleur CIE Lab, histogramme de texton, ligne de caractéristique, descripteur gist + couleur et contexte géométrique. Les vecteurs caractéristiques sont pré-calculés pour toutes les images dans la base de données. Pour estimer location de l'image de requête, James Hays et al calcule d'abord le vecteur caractéristique pour cette image. Ce vecteur est ensuite comparé avec tous les vecteurs caractéristiques de base d'apprentissage. Plusieurs méthodes sont utilisées pour l'étape de reconnaissance : 1-NN, KNN sur les coordonnées GPS et clustering en utilisant Meanshift.

Le résultat d'expérimentation de James Hays et al afiché dans la figure 1 montre que la méthode 1-NN et Meanshift rendent le résultat acceptable (avec un taux d'erreur de 25%) à une échelle d'un petit pays. La méthode 1-NN fonctionne bien à petit échelle d'une ville alors que la méthode basant sur la méthode Meanshift est appropriée à l'échelle plus grande.

Utiliser la même base de données d'image que James Hays et al [HE08], Raghuraman Gopalan [Gop13] cependant, utilise une approche de haut en bas. D'abord les domaines sont créés en regroupant les images des locations adjacentes, une hiérachie de domaines sont créée, un example de cette hiérachie est représenté dans la figure 3. Les sous-espaces génératives et discriminatives aux domaines sont ensuite créés. Les informations cross-domain sont ensuite modèlisées en utilisant la géométrie de ces sous-espaces. Les images vont coder selon ces modèles pour inférer leur location.

Un algorithme avec 4 étapes sont utilisé pour l'inférer la location d'image (figure 4). L'auteur expérimente son algorithme avec les differents nombres de caractéristiques. En utilisant toutes les caractéristiques, la performance de programme augmente. Dans cette revue, nous faisons at-

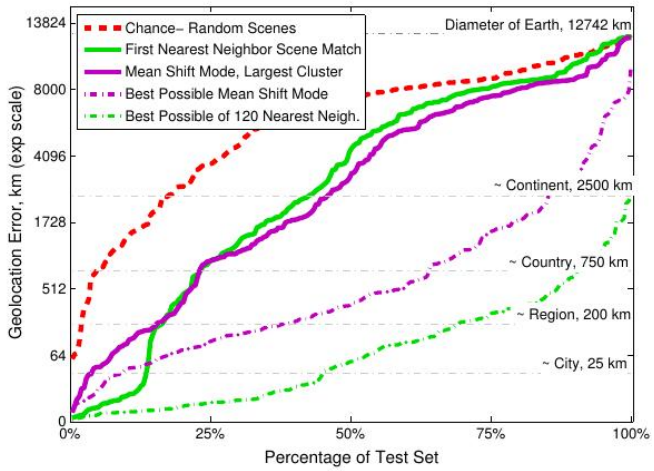

FIGURE 1. Précision de l'estimation de location

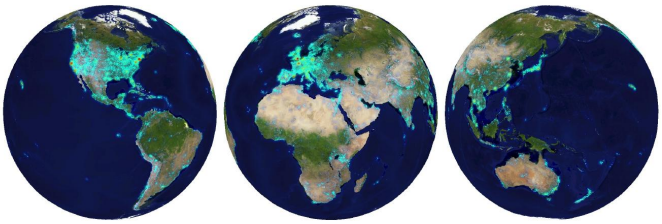

FIGURE 2. Distribution d'images dans la base de données utilisée par James Hays et al [HE08] et Raghuraman Gopalan [Gop13], location d'image est en cyan.

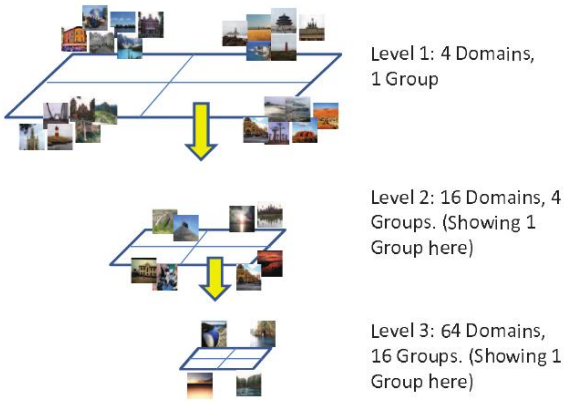

FIGURE 3. Représentation d'une hiérachie des domaines

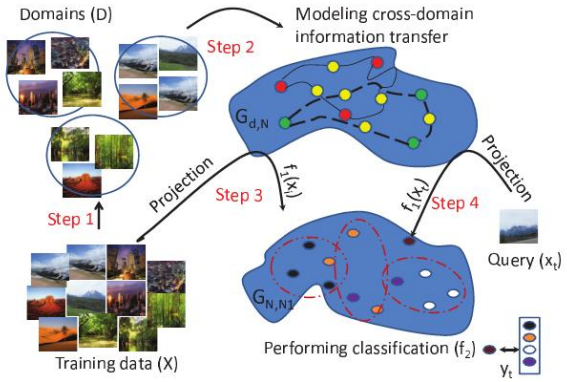

FIGURE 4. Les étapes pour inférer location d'images

tention à la performance de cette approche sur la base de données im2GPS. En fait, la performance de cette approche est meilleure que celle de James Hays et al [HE08]. Ce résultat mets en accent l'importance de la modélisation de données sur le résultat. Cependant, nous constatons que cette approche n'est pas superieur à l'approche de James Hays et al [HE08] en terme de scalabilité.

Le travail intitial de Carl Doersch et al [Doe+12] essaie à chercher automatique des éléments visuels qui caractérisent une location géographique. Leur approche peut appliquer pour déterminer si l'image de requête est de la ville donnée. La base de données utilisée dans leur expérimentation est différence des autres. Ils ont choisi la base de données des images de Google Street Map qui offre certains avantages que les bases de données des résoudre sociaux tel que Flickr.

Pour résoudre ce problème, les auteurs proposent une approche qui utilise la supervision géolocalisée et avec la construction des grappes de bas en haut. On échantillonne aléatoirement un sous-ensemble de 25 000 parcelles à fort contraste pour servir en tant que candidats pour l'ensemencement des grappes. Les parcelles sont représentées en utilisant HOG [Dalal and Triggs 2005]. Les candidates qui ont la plus proportion de ses voisins dans l'ensemble négatif sont supprimés ainsi que les parcelles qui sont presque dupliquées sont rejetées. Pour améliorer le résultat, les auteurs appliquent itérativement un détecteur linaire SVM à chaque élément visuel.

1) utilise k premiers voisins les plus proches dans l'ensemble positif comme des exemples positifs et toutes les parcelles dans l'ensemble négatif comme des exemples négatifs.

2)Dans cette étape, on réitère l'apprentissage SVM en utilisant 5 premières détections comme des exemples positifs.

3)On applique la validation croisée en divisant les deux parties positive et négative de données en 1 sous-ensembles à même taille. À chaque itération de l'apprentissage, on applique les détecteurs formés dans l'étape précédente à un nouveau sous-ensemble de données pour sélectionner les k détections pour rapprendre. Des centaines des détecteurs vont choisir comme les éléments visuels géo-informatives.

Pour évaluer les éléments visuels, Carl Doersch et al testent une nouvelle données dont 50% images de Paris et 50% d'ailleurs en utilisant les 100 premiers détecteurs de Paris. La précision des meilleurs détecteurs était 83%. Pour Prague ce taux était 92%. Alors que les résultats des volontaires étaient 78.5 %. Ce résultat est très prometteur. Cependant, les résultats de l'algorithme sur les villes américaines sont très limites car les styles aux États-Unis manquent des cohérences et des originalités. D'autre limite de cette approche est qu'il est appliqué seulement aux régions urbains, quand on veut l'appliquer sur des scènes naturelles telles que des forêts, des montages ... les éléments visuels sont très peu discriminatifs.

III. Reconnaissance des locations à petit échelle

Étant donné une base de données d'images d'une ville, la tâche de la reconnaissance de locations est de trouver la location exacte de l'image de requête. Les méthodes proposées récemment sont basées sur la méthode "Sac de mots" avec des modifications et des extensions [Tor+13 ; SLK12 ; CS13].

Akihiko Torii et al [Tor+13] propose une méthode pour la reconnaissance des locations avec les structures répétitives. Leur motivation est qu'en réalité, les structures répétitives (qui sont ignorés dans la méthode "Sac de mots visuels") peuvent porter des informations importantes. Ils visent à utiliser ces structures pour améliorer la reconnaissance de location de la méthode Sac de mots visuels. Au lieu d'être supprimés, les structures répétitives sont détectées et ajustées leurs poids.

Pour détecter les répétitives, un graphe $G = (V, E)$, $E = \{(x_i, s_i, d_i)\}$ est construit avec x_i est local invariant features, s_i est la facteur d'échelle et d_i est la descripteur SIFT. Chaque descripteur SIFT est en outre affecté au K = 50 mots visuels les plus proches d'un vocabulaire visuel pré-calculée. Les sommets V_i et V_j sont connexe si 1) $||x_i - x_j|| < c(s_i + s_j)$; c=10, 2) deux caractéristiques ont au moins un mot visuel en commun, 3) la ratio d'échelle σ est entre 0.5 et 1.5. Les groupes des caractéristiques sont construits en trouvant les composantes connexes du graphe. Les auteurs visent à représenter la présence des répétitions plus que le nombre de correspondances d'éléments répétitifs.Une image est représentée par un vecteur :

$$r_d = (r_1, r_2, \ldots, r_V)^T$$

ou le i-ème mot visuel pèse :

$$r_i = \begin{cases} w_{id} \; si \; w_{id} > 0 \; et \; w_{id} < T \\ T \; si \; w_{id} > T \end{cases} \quad (1)$$

La constance T est pour but de représenter l'occurrence (présence / absence) du mot visuel , plutôt que de mesurer le nombre réel d'occurrences. Le poids w_{id} du mot visuel i-ème dans l'image d est obtenue en agrégeant le poids adaptative doux attribué à travers l'image en tenant compte les structures répétitives. Ceci est plus précis et moins ambigu que ceux dans l'état de l'art. Pour vérifier l'algorithme proposée, Akihiko Torii et al [Tor+13] l'expérimentent avec une base de données d'image de Google Street View de la ville Pittsburgh et de la ville San Francisco. Le résultat montre que les structures répétitives améliorent le taux de rappel (figure 5).

Song Cao et Noah Snavely [CS13] mettent en accent la représentation de la données. Leur exploitation montre que une représentation de base de données en graphe améliore la performance de la reconnaissance de location d'une méthode basant sur la méthode Sac de mots visuels. L'approche de ce travail basant directement sur la méthode Sac de mots visuels dont les images sont représentées comme le L2 normalisé de histogramme des mots visuels. En plus des vecteurs de sac de mots, un graphe des images sont aussi utilisé dont chaque noeuds est une image et les sommets représentent les chevauchements entre les images (figure 6).

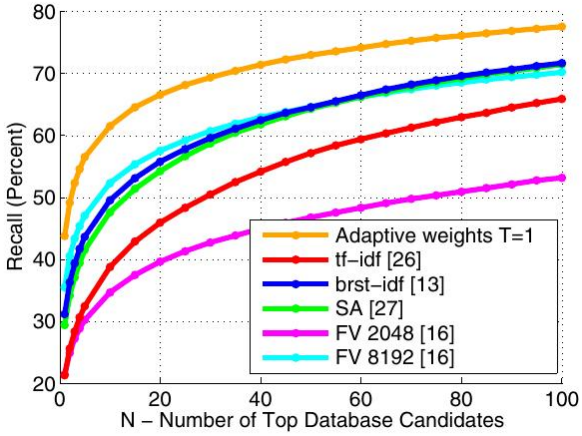

FIGURE 5. Taux de rappel sur la base de données Pittsburgh, résultat de [Tor+13] est en jaune

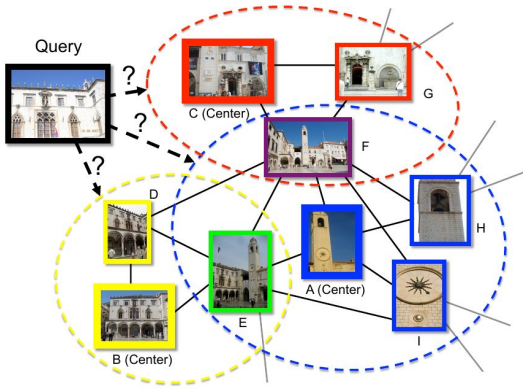

FIGURE 6. Un segment d'un graphe avec 3 clusters définis par les images A, B, C

L'idée principale de cette méthode est d'abord récuper des images similaires (short list) dans la base de données et ensuite mettre en correspondance détaillée jusqu'à une correspondance est trouvée. À étape d'apprentissage, une fonction de similarité est appliquée sur les sous-graphes. Ces sous-graphes sont créés en utilisant un algorithme glouton qui permet d'éviter un problème NP-Complet. À étape de reconnaissance, la distance entre l'image de requête et chaque image de base d'apprentissage, une liste des images similaires est générée. En suite, une vérification géométrique est réalisée avec les meilleurs images dans la liste des images similaires. Les auteurs proposent aussi quelques méthodes pour améliorer la qualité de la liste des images similaires (short list). L'algorithme peut atteindre un taux de précision de 99.5% sur la base de données de Dubrovnik et 99.7% sur la base de données de Rome. Ces taux de précision sont meilleurs que ceux de la méthode Sac de mots qui sont 98.5% et 99.6%.

Cette méthode rend des résultats meilleurs que ceux de la méthode Sac de mots. Cependant, son implémentation est complexe et besoin plus de mémoire.

Torsten Sattler, Bastian Leibe et Leif Kobbelt [SLK12] proposent une recherche active basant sur les deux recherches de 2D à 3D et de 3D à 2D pour les correspondances sup-

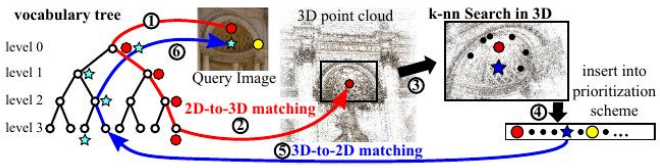

FIGURE 7. Description de la méthode recherche active

plimantaires. Les descripteurs des caractéristiques 2D et des points 3D sont assignés aux mots visuels (figure 7). La pose de caméra est estimée en utilisant l'algorithme RANSAC. Ces correspondances 2D-3D déclenchent un processus de mettre en correspondance 3D-2D. Cette nouvelle approche a la même performance que la méthode basant sur l'arbre mais beaucoup plus rapide. L'idée de la recherche active est qu'après avoir trouvé une correspondance 2D-3D, on cherche activement N_{3D} points plus proches à cette correspondance.

En comparaison le résultat de cette méthode aux ceux dans littérature, on constate qu'elle peut rendre des meilleurs résultats en gardant plus de correspondances qui ont perdu à cause de la quantification. De plus, l'objectif des auteurs sont atteint quand la vitesse leur algorithme est comparable aux autres algorithmes.

IV. CONCLUSION

La reconnaissance de location est un thème de recherche très actif grâce à l'exposition de sources d'images sur Internet. Dans ce travail, nous présentons des recherches sur la reconnaissance de location de ces dernières années pour mettre l'accent sur les résultats atteints et les défis à différentes échelles.

RÉFÉRENCES

[1] Song CAO et Noah SNAVELY. "Graph-Based Discriminative Learning for Location Recognition". In : *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE. 2013, p. 700–707.

[2] Carl DOERSCH et al. "What Makes Paris Look like Paris ?" In : *ACM Transactions on Graphics (SIGGRAPH)* 31.4 (2012).

[3] Raghuraman GOPALAN. "Learning Cross-Domain Information Transfer for Location Recognition and Clustering". In : *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE. 2013, p. 731–738.

[4] James HAYS et Alexei A. EFROS. "im2gps : estimating geographic information from a single image". In : *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2008.

[5] Torsten SATTLER, Bastian LEIBE et Leif KOBBELT. "Improving image-based localization by active correspondence search". In : *Computer Vision–ECCV 2012*. Springer, 2012, p. 752–765.

[6] Akihiko TORII et al. "Visual place recognition with repetitive structures". In : *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE. 2013, p. 883–890.