Localisation Basée Vision : de l'hétérogénéité des approches et des données
Visual Based Localization, a study about approach and data heterogeneity

Nathan Piasco[1,2] Désiré Sidibé[1] Valérie Gouet-Brunet[2] Cédric Demonceaux[1]

[1] Le2i, FRE 2005 CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté
[2] Univ. Paris-Est, LaSTIG MATIS, IGN, ENSG, F-94160 Saint-Mande, France

nathan.piasco@u-bourgogne.fr

Résumé

De nos jours, nous disposons d'une grande diversité de données sur les lieux qui nous entourent. Ces données peuvent être de natures très différentes : une collection d'images, un modèle 3D, un nuage de points colorisés, etc. Lorsque les GPS font défaut, ces informations peuvent être très utiles pour localiser un agent dans son environnement s'il peut lui-même acquérir des informations à partir d'un système de vision. On parle alors de Localisation Basée Vision (LBV). De par la grande hétérogénéité des données acquises et connues sur l'environnement, il existe de nombreux travaux traitant de ce problème. Cet article a pour objet de passer en revue les différentes méthodes récentes pour localiser un système de vision à partir d'une connaissance a priori sur l'environnement dans lequel il se trouve.

Mots Clef

Recherche d'images par contenu, Localisation visuelle, Estimation de pose.

Abstract

We are surrounded by plenty of informations about our environment. From these multiple sources, numerous data could be extracted : set of images, 3D model, coloured point cloud... When classical localisation devices failed (e.g. GPS sensor in cluttered environment) and if the user hold a visual acquisition system, aforementioned data could be used within a localization framework. This is called Visual Based Localization (VBL). Due to numerous data types collected from a scene, VBL regroups a large amount of different methods. This paper present a survey about recent methods that localize a visual acquisition system according to a known environment.

Keywords

CBIR, Visual localization, Pose estimation.

1 Introduction

Localisation basée vision. Les méthodes de localisation basée vision permettent de retrouver la pose (position et orientation 3D) avec laquelle a été prise une image, à partir d'informations visuelles. Ces méthodes s'appuient sur une base de données d'informations géo-localisées préalablement collectée. Au sein de plusieurs communautés de recherche — vision par ordinateur, photogrammétrie, robotique — les contributions scientifiques portant sur l'estimation de la pose par vision n'ont cessé de croître. Ce récent gain d'intérêt s'explique en outre par la mise à disposition de grandes bases de données d'images géo-localisées (terrestres et aériennes), la démocratisation des appareils-photos personnels (présents sur tous les smartphones) et les limites des outils de localisation classique en milieu urbain (situation de "canyon urbain" où le signal GPS est occulté par les immeubles en ville). La localisation basée vision, dénommée par la suite LBV, est utilisée dans divers domaines : la navigation autonome, la construction ou la mise à jour de base de données, les applications de gestion d'albums photos ou la réalité augmentée. Les roboticiens utilisent également la LBV pour résoudre les problématiques liées au scénario dit du "robot kidnappé" ou aux fermetures de boucles en cartographie (SLAM). Quelques exemples de méthodes sont présentés à la figure 1.

Méthodes de localisation. En plus d'une diversité dans ses applications potentielles, les méthodes de LBV présentent une grande hétérogénéité dans les approches et les données utilisées. Une étude de la littérature permet de faire ressortir deux grandes classes de méthodes de LBV :

- Les méthodes basées sur la recherche d'images similaires [15] (cf. figure 1a) : le système reçoit en entrée une requête image et renvoie un groupe d'images similaires présent dans sa base de données. Les images de la base de données étant augmentées d'une information de position, cette classe de méthodes permet d'obtenir une pose approximative de la requête.
- Les méthodes dites "directes" [52] (cf. figure 1b), permettant de retrouver les six degrés de liberté de la caméra ayant acquis l'image présentée en entrée de la méthode. Cette seconde classe de méthodes

(a) Méthode indirecte (b) Méthode directe

FIGURE 1 – **Systèmes de Localisation Basée Vision :** (a) méthode indirect d'après [1] et (b) méthode directe d'après [11]. A partir d'une image requête, les méthodes indirectes retournent un set de données similaires à cette requête (images de gauche), alors que les méthodes direct permet de retrouver la pose exacte d'où à été acquise cette requête (images de droite : les requêtes sont suerposées au modèle 3D de l'envrionnement servant de référence).

permet d'obtenir une pose plus précise, mais nécessite parfois une estimation initiale de la pose, et aussi d'effectuer des pré-traitements plus importants sur la base de données (comme la reconstruction de modèles 3D par *Structure from motion* [19]).

Cet article se focalise sur l'analyse des différents types de données utilisées au sein des méthodes susmentionnées ; leurs études détaillées fera l'objet d'une future contribution.

Données pour la LBV. Communément, les données utilisées dans les méthodes de LBV sont de simples images. Cependant, l'utilisation de ce type de données est limité par deux aspects :

— la nature changeante de l'environnement : elle altère l'apparence des lieux dans le temps, et en fonction des heures de la journée [75], des saisons [2] et des modifications architecturales [1],
— la caractéristique non-globale des systèmes de capture d'images [61] : deux images d'un même lieu prises à des endroits légèrement différents peuvent présenter de grandes différences visuelles.

Pour faire face à ces limitations, des méthodes robustes aux changements d'apparence dans les images ont été développées [59]. D'autre part, pour pallier aux défauts de l'imagerie conventionnelle cités ci-dessus, les méthodes de LBV utilisent une représentation de l'environnement plus complète. Il peut s'agir de données tridimensionnelles associées à l'espace à couvrir pour la localisation, ou de l'ajout d'informations sémantiques à certains lieux.

La suite de l'article est organisée de la façon suivante : la section 2 décrit les méthodes de LBV s'appuyant uniquement sur des images conventionnelles ; la section 3 présente des méthodes exploitant des données géométriques ; la section 4 est consacrée aux méthodes exploitant des informations sémantiques ; quant à la section 5, elle traite des méthodes employées pour comparer des données de natures différentes ; enfin la section 6 conclut sur cet état de l'art.

2 Informations visuelles

La véritable première méthode de LBV pour la navigation s'appuyait sur une base de données d'images géolocalisées [47]. Robertson et Cipolla désiraient localiser des façades de bâtiments ; la base de données utilisée pour leur travaux reflète ainsi l'application visée. Cependant d'autres applications peuvent être envisagées, qui nécessitent chacune une base de donnée adaptée.

2.1 Applications de la LBV

Suivant la problématique considérée par les auteurs, les images formant la base de donnée d'un système de LBV peuvent être de natures différentes. Notamment, une grande partie des travaux de recherche se projettent dans le cas où l'on souhaite connaître la pose d'une caméra dans un environnent urbain. On peut différencier ici trois types de méthodes : la localisation en intérieur [27], la localisation pour les piétons (ou les robots [13]) dans une ville [47, 54, 10, 74], la localisation de véhicules évoluant sur des routes en zone urbaine ou suburbaine [33, 35, 44]. Les bases de données exploitées présentent alors des différences significatives en fonction de l'application visée. D'autres méthodes se focalisent sur la localisation de véhicules aériens et utilisent des clichés de type satellitaire comme référence [68].

2.2 Couverture de la base de données

De part leur nature, les applications précédemment évoquées ne couvrent pas la même zone de localisation. Une application développée pour l'estimation de position d'une voiture devra nécessairement être capable de localiser le véhicule dans une zone plus large qu'une application qui aurait pour objectif le guidage visuel d'un piéton dans une ville. Ainsi on trouve des méthodes qui permettent de donner une position approximative d'une photographie à une échelle mondiale [70, 17], et au contraire d'autres systèmes qui permettent d'obtenir une position plus précise au détriment de l'étendue de la zone de couverture [57]. La couverture d'une zone par des images classiques peut être aug-

mentée en utilisant des appareils de capture adaptés. Dans les travaux de [1, 72, 73], les auteurs construisent une base de données en utilisant des images sphériques permettant une capture omnidirectionnelle de l'environnement dans lequel on souhaite se localiser. Les images aériennes permettent de couvrir de grandes étendues mais sont moins facilement exploitables dans le cas d'application de localisation en milieu urbain : elles présentent de fortes différences de points de vue avec les clichés pris au niveau du sol [64, 29].

2.3 Homogénéité de la base de données

Quelque soit le type d'images utilisé, on peut distinguer deux types de base de données : les bases de données homogènes (prises avec le même appareil dans un intervalle de temps proche) et les bases de données hétérogènes où les images ont été prises à des instants espacés dans le temps, par des systèmes d'acquisition ou des opérateurs différents. Les bases de données homogènes peuvent être générées au travers de système comme *google street view*[1] ou le système Stéréopolis de l'IGN [38], et permettent d'effectuer des traitements systématiques sur les images récupérées [61, 32]. D'autre part, l'utilisation de bases de données hétérogènes permet une mise à jour plus flexible de la base de données de l'environnement et d'introduire une robustesse aux changements d'apparence [46, 15], étant donnée que le système est construit sur un ensemble de données présentant une certaine variabilité. Dans les travaux de [1], les auteurs introduisent volontairement des images de différentes sources pour permettre à leur système de réaliser la localisation sur le long terme. D'autres travaux présentent une approche similaire, afin de localiser une image quelque soit la saison de l'année à laquelle elle a été prise [36].

3 Informations géométriques

La préoccupation principale de la LBV utilisant l'imagerie classique est de développer des méthodes robustes au changement d'apparence inhérent à ce type de donnée. L'utilisation d'informations géométriques supplémentaires associées aux images permet, en partie, de contourner cette difficulté. Nous allons voir dans cette partie comment l'ajout d'informations spatiales permet d'améliorer les résultats des systèmes de LBV.

3.1 Informations géométriques faibles

Dans [61, 10], les auteurs utilisent une information simple décrivant les principaux plan 3D présents dans une image. Cette information leur permet de modifier les images de la base de données, soit pour les rectifier [10], soit afin de générer davantage d'images pour augmenter la zone de couverture du système de localisation [61].

Cham *et al.* [9] utilisent une carte d'emprises au sol des bâtiments dans leur système de LBV. Pour une image donnée, ils procèdent à l'extraction des angles formés par les façades d'immeubles présents dans l'image pour retrouver sa position dans la carte. Les travaux de [3] utilisent d'une manière similaire un modèle 2.5D de la ville de Graz pour initialiser la pose d'un smartphone pour une application de réalité augmentée.

Baatz *et al.* [6] introduisent l'utilisation d'un modèle numérique de terrain dans le cadre d'une application de LBV dans le massif rocheux des Alpes.

Toutes ces approches permettent de considérer la géométrie de la scène et ainsi de ne plus être sensible aux variations d'illumination (cycle jour/nuit, couverture nuageuse, ombres, etc.).

3.2 *Structure from Motion*

Les récentes avancées dans le domaine de la création de nuage de points 3D par *Structure from Motion* (*SfM*) ont inspiré les recherches en LVB. De nombreuses méthodes [25, 30, 52, 22, 19, 18, 34, 60, 5] effectuent un prétraitement sur une base de données d'images géoréférencées afin de reconstruire un modèle 3D concis (nuage de points clairsemés) de la zone de localisation (cf. figure 1a). De même qu'évoqué dans la section 2.3, les modèles reconstruits peuvent être issus de collections d'images homogènes [22, 20] comme hétérogènes [19, 50]. Dans [30], les auteurs présentent une méthode particulière où à la fois la requête et la base de données sont représentées sous forme de nuage de points 3D générés par une séquence vidéo. Les méthodes se basant sur un modèle reconstruit par *SfM* offrent de meilleurs résultats que celles se basant essentiellement sur de la recherche d'images similaires [53]. En effet, des règles de consistance géométrique et spatiale peuvent être utilisées pour retrouver la véritable position de la requête [25].

3.3 Géométrie tri-dimensionnelle

Plutôt que par une opération coûteuse de *SfM*, la géométrie d'une scène peut être directement capturée par des capteurs de profondeur (caméra stéréo, caméra de profondeur, laser, lidar, etc.). Les méthodes de LBV présentées dans cette section utilisent ce type de capteurs.

Imagerie RGB-Profondeur. Les données brutes des capteurs de profondeur sont souvent utilisés pour rajouter un canal d'information au contenu visuel. Les travaux de [33, 67, 37] utilisent une carte de disparité obtenue par stéréo-vision. [55, 16, 14] développent leur application de LBV en utilisant une caméra de profondeur fonctionnant par projection de motifs structurés dans l'infrarouge. Une caméra similaire est utilisée dans [24] pour permettre la localisation dans le noir complet de la caméra.

Modèle 3D. Les dernières méthodes présentées dans cette section s'appuient sur un modèle 3D complet comme base de données pour la localisation. Pour la tâche de localisation en intérieur, [55] utilisent un modèle 3D construit à partir d'images de profondeurs. Les travaux décrits dans [4, 43] font appel à des modèles 3D de quartiers de villes ou de monuments. L'approche de [3] permet de localiser pré-

1. https ://www.google.fr/intl/fr/streetview/

cisément des peintures réalistes par rapport à plusieurs modèles.

4 Informations sémantiques

Bien qu'assez discriminantes, les méthodes de LBV se basant sur des informations géométriques de la scène restent coûteuses à mettre en œuvre et nécessitent l'utilisation de données plus complexe à acquérir [48]. Cependant un autre type d'information à fort pouvoir discriminant, et s'avérant très compact à large échelle, peut également être utilisé pour la LBV : les informations sémantiques. Deux approches peuvent être considérées pour extraire une représentation sémantique d'une donnée : la segmentation et la catégorisation. La segmentation est une méthode locale consistant à reconnaître à l'intérieur de la donnée des sous-ensembles à signification sémantique (e.g. détection d'objets dans une image). La catégorisation considère la donnée dans son ensemble et lui adresser une classe ou un contexte (e.g. interprétation de scène à partir d'une photographie).

4.1 Segmentation

La notion d'information sémantique peut varier d'une définition à une autre : dans [12] l'approche sémantique considérée consiste à extraire des plans présents dans une scène, alors que dans [49] les auteurs segmentent un nuage de point afin de reconnaître des objets d'un degré d'interprétation sémantique supérieur comme des tables et des chaises. Néanmoins les points communs à ces méthodes sont leur organisation hiérarchique et leur représentation sous forme de graphe. Cette représentation de l'environnement permet de s'affranchir des problématiques de changements d'apparence et offre une représentation compacte. La segmentation sémantique est utilisée dans [30] pour affiner l'espace de recherche lors d'une tâche de localisation à grande échelle. Qu *et al.* [45] proposent une méthode de localisation s'appuyant sur la détection de panneaux et marquages routiers en milieu urbain. Les travaux de [3] montrent également un cas d'application concret de la segmentation sémantique pour la tâche de LBV.

4.2 Catégorisation

La catégorisation de scène est un autre emploi possible des informations sémantiques pour la tâche de LBV [71]. Classifier une image parmi un nombre prédéfini de classes permet d'obtenir un sous-ensemble de la base de donnée dans lequel chercher la pose précise de la requête. Torralba *et al.* [62] introduisent cette notion, qui sera reprise dans les travaux de [37]. Dans [58], les auteurs proposent d'utiliser la classification du réseau de neurones *ImageNet* [23] pour réduire l'espace de recherche.

5 Association hétérogène

Les méthodes de LBV impliquent obligatoirement deux éléments : une requête et une base de données où localiser la requête. Nous avons vu jusqu'à maintenant la diversité en **type de données** de ces deux éléments. Cette dernière partie est consacrée aux associations possibles entre le type de données de la requête, et le type de données de la base. On distinguera ici deux catégories : d'une part la localisation *cross-domain* qui fait référence à l'association de données de domaines différents, et d'autre part la localisation *cross-data* qui fait référence à l'association de données de natures différentes.

5.1 Localisation *cross-domain*

Ils existent deux cas d'usage que l'on peut associer au *cross-domain* dans la LBV. Une série d'articles traitent la problématique de la localisation dite *cross-view* : localiser une image prise au niveau du sol à l'aide d'une base de donnée d'images aériennes [28, 29, 64]. Cette association de domaines (domaine aérien/domaine terrestre) est motivée par le fait que par nature, les bases de données d'imagerie aérienne sont très souvent associées à une couverture bien plus vaste que celle des bases de données terrestres.

Les travaux de Russell *et al.* [48], suivis de ceux de Aubry *et al.* [4] considèrent la mise en relation de peintures représentant des sites architecturaux avec des modèles réels de ces sites. Les auteurs présentent des méthodes permettant de déterminer le point de vue qu'un artiste a voulu représenter au travers de son œuvre. Ces travaux, motivés cette fois-ci par des recherches archéologiques, sont l'exemple typique de localisation *cross-domain* où l'aspect visuel de la requête par rapport à la base de donnée change drastiquement [56]. La localisation *cross-domain* est aussi associée aux données anciennes ou culturelles que l'on souhaite comparer à des contenus géographiques actuels, pour leur localisation et plus généralement pour leur valorisation hors des murs [7].

5.2 Localisation *cross-data*

A l'instar de la localisation *cross-domain*, les applications de localisation *cross-data* sont bien plus nombreuses. Les méthodes basées sur les modèles reconstruits par *SfM* (section 3.2) comparent une image 2D au modèle 3D reconstruit [19, 50, 34]. Cette recherche de similarité 2D-3D est rendue possible par l'extraction de descripteurs communs aux deux espaces de représentation.

Les bases de données augmentées d'informations géométriques présentées dans la section 3.3 sont souvent confrontées à un cliché simple [3, 6, 22]. Cette comparaison entre différents types de données peut être réalisée au travers de méthodes d'apprentissages [20, 65], ou en "projetant" les données dans un même espace permettant la comparaison. Ainsi on retrouve dans beaucoup de travaux des méthodes consistant à synthétiser des vues 2D à partir de modèles 3D [4, 6, 61, 32, 19, 43]. Il existe également des méthodes globales permettant d'aligner une image [39] ou un nuage de points *SfM* [40] à un modèle 3D de façon optimale.

Pour conclure cette section, nous avons regroupé dans le tableau 1 les méthodes de LBV combinant différents types

TABLE 1 – Résumé des différentes associations de données présentes dans les méthodes de LBV.

Base de donnée / **Requête**	Collection d'images Section 2	Géométrie faible Section 3.1	Nuage de points (*SfM*) Section 3.2	Modèle 3D Section 3.3
Une image	Recherche par similarité[3]	[9, 3, 6, 61, 10]	[25, 30, 51, 52, 22, 19, 18, 34, 5]	[4, 48, 22, 66, 39]
Plusieurs images[4]	[72, 70]	**X**	**X**	**X**
Image + Profondeur	**X**	**X**	**X**	[55, 16, 63, 14]
Un Modèle *SfM*	**X**	**X**	[30]	[40]

de données.

6 Discussion et Conclusion

6.1 Tendances actuelles

Les méthodes indirectes de LBV sont aujourd'hui dominées par les approches exploitant les réseaux de neurones à convolution comme descripteurs globaux d'images [15, 46]. Les descripteurs des images sont obtenus par l'agrégation dans un même vecteurs de l'ensemble des poids associés à une certaine couche du réseau [59]. Ces réseaux, initialement utilisés pour l'interprétation d'images [23], permettent d'obtenir les meilleurs score de *Mean Average Precision (mAP)* sur les data-sets classiques de recherche d'images représentant des milieux urbains (Oxford et Paris data-sets [41, 42]).

On assiste actuellement à une recrudescence du nombre de contributions traitant de méthodes directes. En effet, ces méthodes permettent d'obtenir une pose plus précise de la requête, pouvant être ensuite utilisée par une multitude d'applications (robotique, véhicules autonomes, réalité augmentée, etc.). Elles bénéficient également de l'essor des données géométriques, de plus en plus facile d'accès, comme l'atteste les récents data-sets de grandes envergures comprenant des images couplées à des acquisitions lasers [31, 69]. On retiendra deux nouvelles familles de méthodes direct de LBV prometteuses : celle s'appuyant sur des réseaux de neurones [22, 21] et celle utilisant des forêts de régression [55, 8]. Cependant, les meilleurs résultats sur les data-sets les plus couramment utilisés (Dubrovnik [26] et Cambridge [22]) sont toujours obtenus par des méthodes «classiques» d'appariement de descripteurs locaux [52, 11].

6.2 Conclusion et orientation future

La localisation basée vision est au cœur d'une dynamique de recherches nouvelles depuis plus d'une dizaine d'années. Nous avons dressé un panorama de la diversité des applications et des données en lien avec la LBV. En considérant dans un premier temps les applications de LBV utilisant simplement de l'imagerie classique, nous avons défini les paramètres importants associés à la base de données des systèmes de localisation. Dans un second temps nous avons évoqué des méthodes de LBV s'appuyant sur des informations géométriques pour pallier les défauts d'une représentation essentiellement visuelle de l'environnement. Nous avons enfin montré que l'ajout d'informations sémantiques permet une représentation abstraite, compacte et robuste de l'environnent, ouvrant la voie à des méthodes de raisonnement discriminantes et capable de passer à l'échelle, pour la tâche de LBV.

La combinaison des approches décrites ouvre la voie à de nouvelles applications (*cross-domain localization*) tandis que la combinaison des données (*cross-data localization*) permet d'augmenter la précision et la rapidité des méthodes existantes. Les données visuelles augmentées d'informations géométriques sont de plus en plus courantes [31, 38], tout comme les applications les utilisant (projet Tango [2]). D'autre part, l'émergence des réseaux de neurones et leur performance dans le domaine de la classification [23] permet de générer rapidement de l'information sémantique. Le domaine de localisation visuelle pourra certainement bénéficier de l'association de ces avancées scientifiques.

3. Toutes les méthodes de recherche d'image par similarité qui confrontent une image RGB à en ensemble d'images RGB. Les trop nombreuses références concernant ce type de méthode n'ont pas été insérées dans le tableau dans un soucis de clarté.

4. Plusieurs travaux considèrent un problème de LBV où l'agent fournit au système de localisation une série d'images d'un même lieu plutôt qu'une image isolée.

Références

[1] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD : CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307. 2, 3

[2] Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., and Romera, E. (2016). Fusion and Binarization of CNN Features for Robust Topological Localization across Seasons. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4656–4663. 2

[3] Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., and Lepetit, V. (2015). Instant Outdoor Localization and SLAM Initialization from 2.5D Maps. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 21(11) :1309–1318. 3, 4, 5

[4] Aubry, M., Russell, B. C., and Sivic, J. (2014). Painting-to-3D model alignment via discriminative visual elements. *ACM Transactions on Graphics (ToG)*, 33(2) :1–14. 3, 4, 5

[5] Azzi, C., Asmar, D., Fakih, A., and Zelek, J. (2016). Filtering 3D Keypoints Using GIST For Accurate Image-Based

2. https ://get.google.com/tango/

Localization. In *British Machine Vision Conference (BMVC)*, number 2, pages 1–12. 3, 5

[6] Baatz, G., Saurer, O., Köser, K., and Pollefeys, M. (2012). Large Scale Visual Geo-Localization of Images in Mountainous Terrain. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 7573, pages 517–530. 3, 4, 5

[7] Bhowmik, N., Weng, L., Gouet-Brunet, V., and Soheilian, B. (2017). Cross-domain Image Localization by Adaptive Feature Fusion. In *Joint Urban Remote Sensing Event (JURSE)*. 4

[8] Cavallari, T., Golodetz, S., Lord, N. A., Valentin, J., Di Stefano, L., and Torr, P. H. S. (2017). On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. *arXiv preprint*. 5

[9] Cham, T. J., Ciptadi, A., Tan, W. C., Pham, M. T., and Chia, L. T. (2010). Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 366–373. 3, 5

[10] Chen, D. M., Baatz, G., Köser, K., Tsai, S. S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., and Grzeszczuk, R. (2011). City-scale landmark identification on mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 737–744. 2, 3, 5

[11] Feng, Y., Fan, L., and Wu, Y. (2016). Fast Localization in Large-Scale Environments Using Supervised Indexing of Binary Features. *IEEE Transactions on Image Processing (ToIP)*, 25(1) :343–358. 2, 5

[12] Fernandez-Moral, E., Mayol-Cuevas, W., Arevalo, V., and Gonzalez-Jimenez, J. (2013). Fast place recognition with plane-based maps. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2719–2724. 4

[13] Garcia-Fidalgo, E. and Ortiz, A. (2015). Vision-based topological mapping and localization methods : A survey. *Robotics and Autonomous Systems (RAS)*, 64 :1–20. 2

[14] Glocker, B., Shotton, J., Criminisi, A., and Izadi, S. (2015). Real-time RGB-D camera relocalization via randomized ferns for keyframe encoding. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 21(5) :571–583. 3, 5

[15] Gordo, A., Almazan, J., Revaud, J., and Larlus, D. (2016). Deep Image Retrieval : Learning Global Representations for Image Search. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 241–257. 1, 3, 5

[16] Guzman-Rivera, A., Pushmeet, K., Glocker, B., Shotton, J., Sharp, T., Fitzgibbon, A., and Izadi, S. (2014). Multi-Output Learning for Camera Relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6. 3, 5

[17] Hays, J. and Efros, A. A. (2008). IM2GPS : Estimating Geographic Information From a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 05. 2

[18] Heisterklaus, I., Qian, N., and Miller, A. (2014). Image-based pose estimation using a compact 3D model. In *IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, pages 327–330. 3, 5

[19] Irschara, A., Zach, C., Frahm, J.-m., and Bischof, H. (2009). From Structure-from-Motion Point Clouds to Fast Location Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3, 4, 5

[20] Kendall, A. and Cipolla, R. (2016). Modelling Uncertainty in Deep Learning for Camera Relocalization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 3, 4

[21] Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

[22] Kendall, A., Grimes, M., and Cipolla, R. (2015). PoseNet : A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3, 4, 5

[23] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105. 4, 5

[24] Li, R., Liu, Q., Gui, J., Gu, D., and Hu, H. (2016). Nighttime indoor relocalization using depth image with Convolutional Neural Networks. *Proceedings of the IEEE International Conference on Automation and Computing (ICAC)*, pages 261–266. 3

[25] Li, Y., Snavely, N., Huttenlocher, D., and Fua, P. (2012). Worldwide Pose Estimation Using 3D Point Clouds. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 15–29. 3, 5

[26] Li, Y., Snavely, N., and Huttenlocher, D. P. (2010). Location Recognition using Prioritized Feature Matching. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 791–804. 5

[27] Liang, J. Z., Corso, N., Turner, E., and Zakhor, A. (2013). Image Based Localization in Indoor Environments. In *Computing for Geospatial Research and Application*. 2

[28] Lin, T.-Y., Belongie, S., and Hays, J. (2013). Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898. 4

[29] Lin, T.-Y., Cui, Y., Belongie, S., and Hays, J. (2015). Learning Deep Representations for Ground-to- Aerial Geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number JUNE, pages 5007–5015. 3, 4

[30] Lu, G., Yan, Y., Ren, L., Song, J., Sebe, N., and Kambhamettu, C. (2015). Localize Me Anywhere , Anytime : A Multi-task Point-Retrieval Approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2434–2442. 3, 4, 5

[31] Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016). 1 year, 1000 km : The Oxford RobotCar dataset. *The International Journal of Robotics Research (IJRR)*, page 0278364916679498. 5

[32] Majdik, A. L., Albers-Schoenberg, Y., and Scaramuzza, D. (2013). MAV urban localization from Google street view data. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3979–3986. 3, 4

[33] McManus, C., Upcroft, B., and Newman, P. (2014). Scene Signatures : Localised and Point-less Features for Localisation. In *Robotics Science and Systems (RSS)*. 2, 3

[34] Middelberg, S., Sattler, T., Untzelmann, O., and Kobbelt, L. (2014). Scalable 6-DOF localization on mobile devices. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 8690 LNCS(PART 2) :268–283. 3, 4, 5

[35] Milford, M. J., Lowry, S., Shirazi, S., Pepperell, E., Shen, C., Lin, G., Liu, F., Cadena, C., and Reid, I. (2015). Sequence Searching with Deep-learnt Depth for Condition- and Viewpoint- invariant Route-based Place Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 18–25. 2

[36] Neubert, P., Sünderhauf, N., and Protzel, P. (2015). Superpixel-based appearance change prediction for long-term navigation across seasons. *Robotics and Autonomous Systems (RAS)*, 69(1) :15–27. 3

[37] Ni, K., Kannan, A., Criminisi, A., and Winn, J. (2009). Epitomic location recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(12) :2158–2167. 3, 4

[38] Paparoditis, N., Papelard, J.-P., Cannelle, B., Devaux, A., Soheilian, B., David, N., and Houzay, E. (2012). Stereopolis II : A multi-purpose and multi-sensor 3D mobile mapping system for street visualisation and 3D metrology. *Revue française de photogrammétrie et de télédétection*, 200(1) :69–79. 3, 5

[39] Paudel, D. P., Habed, A., Demonceaux, C., and Vasseur, P. (2015a). LMI-based 2D-3D registration : From uncalibrated images to Euclidean scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4494–4502. 4, 5

[40] Paudel, D. P., Habed, A., Demonceaux, C., and Vasseur, P. (2015b). Robust and Optimal Sum-of-Squares-Based Point-to-Plane Registration of Image Sets and Structured Scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2048–2056. 4, 5

[41] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

[42] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in Quantization : {I}mproving Particular Object Retrieval in Large Scale Image Databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

[43] Poglitsch, C., Arth, C., Schmalstieg, D., and Ventura, J. (2015). A particle filter approach to outdoor localization using image-based rendering. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 132–135. 3, 4

[44] Qu, X., Soheilian, B., Habets, E., and Paparoditis, N. (2016). Evaluation of SIFT and SURF for vision based localization. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41(July) :685–692. 2

[45] Qu, X., Soheilian, B., and Paparoditis, N. (2015). Vehicle localization using mono-camera and geo-referenced traffic signs. *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2015-Augus :605–610. 4

[46] Radenovic, F., Tolias, G., and Chum, O. (2016). CNN Image Retrieval Learns from BoW : Unsupervised Fine-Tuning with Hard Examples. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 3–20. 3, 5

[47] Robertson, D. and Cipolla, R. (2004). An Image-Based System for Urban Navigation. In *British Machine Vision Conference (BMVC)*. 2

[48] Russell, B. C., Sivic, J., Ponce, J., and Dessales, H. (2011). Automatic alignment of paintings and photographs depicting a 3D scene. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4, 5

[49] Salas-Moreno, R. F., Newcombe, R. A., Strasdat, H., Kelly, P. H. J., and Davison, A. J. (2013). SLAM++ : Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359. 4

[50] Sattler, T., Leibe, B., and Kobbelt, L. (2011). Fast image-based localization using direct 2D-to-3D matching. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 667–674. 3, 4

[51] Sattler, T., Leibe, B., and Kobbelt, L. (2012a). Improving image-based localization by active correspondence search. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 7572 LNCS, pages 752–765. 5

[52] Sattler, T., Leibe, B., and Kobbelt, L. (2016). Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, X(1). 1, 3, 5

[53] Sattler, T., Weyand, T., Leibe, B., and Kobbelt, L. (2012b). Image Retrieval for Image-Based Localization Revisited. In *British Machine Vision Conference (BMVC)*, pages 76.1–76.12. 3

[54] Schindler, G., Brown, M., and Szeliski, R. (2007). City-Scale Location Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

[55] Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., and Fitzgibbon, A. (2013). Scene coordinate regression forests for camera relocalization in RGB-D images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937. 3, 5

[56] Shrivastava, A., Malisiewicz, T., Gupta, A., and Efros, A. A. (2011). Data-driven visual similarity for cross-domain image matching. *ACM Transactions on Graphics (ToG)*, 30(6) :1. 4

[57] Song, Y., Chen, X., Wang, X., Zhang, Y., and Li, J. (2016). 6-DOF Image Localization From Massive Geo-Tagged Reference Images. *IEEE Transactions on Multimedia (ToM)*, 18(8) :1542–1554. 2

[58] Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. J. (2015a). On the performance of ConvNet features for place recognition. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2015-Decem :4297–4304. 4

[59] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. J. (2015b). Place Recognition with ConvNet Landmarks : Viewpoint-Robust, Condition-Robust, Training-Free. In *Robotics Science and Systems (RSS)*. 2, 5

[60] Svarm, L., Enqvist, O., Kahl, F., and Oskarsson, M. (2016). City-Scale Localization for Cameras with Known Vertical Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8828(c) :1–1. 3

[61] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., and Pajdla, T. (2015). 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3, 4, 5

[62] Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 273–280. 4

[63] Valentin, J., Fitzgibbon, A., Nießner, M., Shotton, J., and Torr, P. H. S. (2015). Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4400–4408. 5

[64] Vo, N. N. and Hays, J. (2016). Localizing and Orienting Street Views Using Overhead Imagery. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 494–509. 3, 4

[65] Walch, F. (2016). Deep Learning for Image-Based Localization. *arXiv preprint*. 4

[66] Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., and Cremers, D. (2016). Image-based Localization with Spatial LSTMs. *arXiv preprint*. 5

[67] Wan, W., Liu, Z., Di, K., Wang, B., and Zhou, J. (2014). A Cross-Site Visual Localization Method for Yutu Rover. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-4(May) :279–284. 3

[68] Wan, X., Liu, J., Yan, H., and Morgan, G. L. K. (2016). Illumination-invariant image matching for autonomous UAV localisation based on optical sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119 :198–213. 2

[69] Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., and Urtasun, R. (2016). TorontoCity : Seeing the World with a Million Eyes. *arXiv preprint*. 5

[70] Weyand, T., Kostrikov, I., and Philbin, J. (2016). PlaNet - Photo Geolocation with Convolutional Neural Networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 37–55. 2, 5

[71] Wu, J., Christensen, H. I., and Rehg, J. M. (2009). Visual place categorization : Problem, dataset, and algorithm. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4763–4770. 4

[72] Zamir, A. R. and Shah, M. (2010). Accurate image localization based on google maps street view. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 6314 LNCS, pages 255–268. 3, 5

[73] Zamir, A. R. and Shah, M. (2014). Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8) :1546–1558. 3

[74] Zhang, W. and Kosecka, J. (2006). Image Based Localization in Urban Environments. In *3D Data Processing, Visualization and Transmission (3DPVT)*. 2

[75] Zhou, H., Sattler, T., and Jacobs, D. W. (2016). Evaluating Local Features for Day-Night Matching. In *Proceedings of the IEEE European Conference on Computer Vision Workshop (ECCVW)*. 2