A survey on Visual-Based Localization: On the benefit of heterogeneous data

Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, Valérie Gouet-Brunet

A Survey on Visual-Based Localization: On the Benefit of Heterogeneous Data

Nathan Piasco[a,b], Désiré Sidibé[a], Cédric Demonceaux[a], Valérie Gouet-Brunet[b]

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

Abstract

We are surrounded by plenty of information about our environment. From these multiple sources, numerous data could be extracted: set of images, 3D model, coloured points cloud... When classical localization devices failed (*e.g.* GPS sensor in cluttered environments), aforementioned data could be used within a localization framework. This is called Visual Based Localization (VBL). Due to numerous data types that can be collected from a scene, VBL encompasses a large amount of different methods. This paper presents a survey about recent methods that localize a visual acquisition system according to a known environment. We start by categorizing VBL methods into two distinct families: indirect and direct localization systems. As the localization environment is almost always dynamic, we pay special attention to methods designed to handle appearances changes occurring in a scene. Thereafter, we highlight methods exploiting heterogeneous types of data. Finally, we conclude the paper with a discussion on promising trends that could permit to a localization system to reach high precision pose estimation within an area as large as possible.

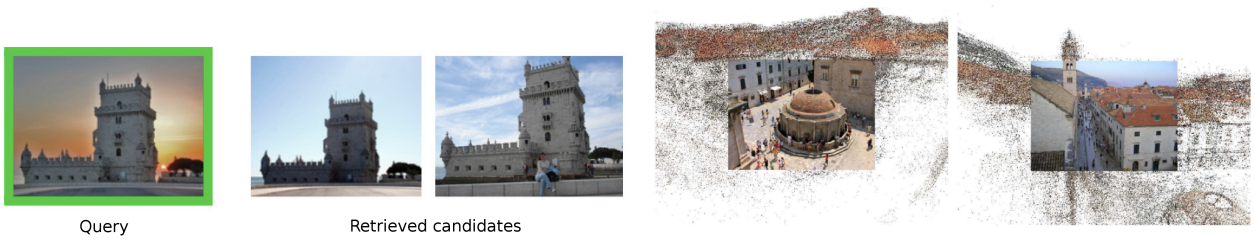
Keywords: Image-based localization, Visual geo-localization, Camera Relocalisation, Pose estimation

1. Introduction

Visual-Based Localization (VBL) consists of retrieving the pose (position + orientation) of a visual query material within a known space representation. For instance, recovering the pose of a camera that took a given photography according to a set of geo-localized images or a 3D model is a simple illustration of such a localization system. VBL has been an increasingly dynamic research subject in the last decade. This recent gain of interest is due to the provision of large geo-localized images database, the multiplication of embedded visual acquisition system (*e.g.* camera on smart-phone) and the limitation of usual localization system in urban environment (*e.g.* GPS signal failure in cluttered environment). Aforementioned localization problem is involved in several present-day practical applications, such as GPS-like localization system, indoor or outdoor navigation, 3D reconstruction, models and databases update, consumer photography—"Where did I take these photos?"—and augmented reality. Visual-based localization is also used in robotics to solve SLAM loop-closure problem or kidnapped robot scenario.

There is no standardized designation for VBL, methods name vary from one paper to another. In this survey, we include in Visual-based Localization methods categorized as: Image-based localization [11], Visual localization [201], Structure-based localization [172], Visual geo-localization [222, 21], Camera Relocalisation [183], Image-based pose estimation [66], and all the possible rearrangement of these terms.

Topics addressed. Along the whole survey, we will focus on urban VBL as it represents the most studied end-user application in literature. This can be explained by the fact that most of the related applications take place in non-rural environment. As an illustration, VBL as GPS pedestrian localization system should be used when buildings (so inside a city) disrupt the satellite signal. Most of the augmented reality applications are also designed for indoor or urbanized environment. Similar reasoning can be employed with robotic applications. Nowadays principal concerns about robots are related to human assistance or supervision and autonomous vehicles. Those services occur in indoor and outdoor man-made areas, therefore the robot localization should be studied for these sites. The other aspect that invites researchers to focus on urban environment is that large datasets are mainly describing cities or road networks, because they are the most reachable places. With the exception of airborne and satellite imageries, that are abundant all over the globe but these data restrict the range of possible uses.



(a) Indirect method (b) Direct method

Figure 1: **Examples of Visual Based Localization systems:** (a) indirect method from [160]: according to an input query (left images), indirect methods recover a set of similar data (right images). (b) direct method from [49]: direct methods recover the exact pose of the input query (the two figures represents superposed images according to a reference 3D model).

As well as VBL presents a heterogeneity about its end-user applications, methods and data involved in the process of localizing an image are various. These methods are divided into two categories: **indirect methods** [2, 160] (see figure 1a) that cast the localization task as an image retrieval problem and provide a coarse pose information about the query location and **direct methods** [80, 176] (see figure 1b) that directly regress the 6 Degrees of Freedom (DoF) pose of the visual acquisition system. Indirect methods used in VBL slightly differ from classical vision object-retrieval algorithm [185] on two points: the images in the query and the database represent scenes rather than objects (*e.g.* street view panorama, buildings images, indoor scenes) and the performance of such system is evaluated according to the precision rate rather than the recall rate (i.e. a perfect VBL system should recover in its top ranked candidates documents that display the exact location of the query). Unlike indirect methods, direct methods aim to recover instantly the pose of the query data. Where Structure from Motion (SfM) or SLAM techniques provide a *relative pose* of a sequence of data, VBL tackles the problem of retrieving the *absolute pose* of a query data according to a known representation. Nevertheless, this representation could have been built thanks to SfM or SLAM mapping module. Figure 1 presents representatives of the two different methods studied in this survey.

When designing a VBL system, the type of method is not the only parameter to consider. As pointed out in [111], robustness to environment appearance changes over time is a main concern. Data involved in the process of localization also define specifications of the system, like area covered by the VBL method or precision of the regressed pose. Data types are various in VBL: visual data, geometric information (provided by RGB-D camera, LIDAR, etc.) and semantic clues. Combination of different data in VBL aims to overcome limitation of images-only based method.

Related works. VBL is a well studied topics, and many contributions propose overview of this domain. The closest survey to ours is the paper of Brejcha and Čadík [21]. It presents many works on VBL and classify them depending on the environment for which the particular method was developed. Conversely, we focus our study on systems built for city-scale localization as it concerns the most VBL applications. Moreover, we propose a comprehensive description of the two types of methods used in VBL, and highlight the benefits of the use of heterogeneous data in the context of localization in challenging scenarios. Zamir *et al.* [222] gather recent articles to draw a large panorama of VBL, corroborating the growing importance of this domain in current research. This assumption is comforted regarding the many tutorials (CVPR2014, CVPR2015 and CVPR2017) and workshop (CVPR2015) about the Visual Localization problem in high impact international conferences.

Visual Place Recognition is a roboticist problem, defined in the general sense in [111] as the visual ability of a human, an animal or a robot to recognize an already visited place. It is a main concern for navigation, especially when we consider topological mapping [54]. Despite the fact that Visual Place Recognition shares huge similarity with the issues addressed in this survey, the two problems differ on three major points. On the one hand, visual-based localization and visual place recognition purposes differ; where Place Recognition decides if a given place have already been seen, VBL produces a pose of the visual acquisition system. This explains the difference in their respective pipeline. Visual place recognition is composed of three main components (the data processing module, the mapping module and the belief generation modules) while visual-based localization does not consider the mapping module. On the other hand, the study presented here aims to consider VBL in a more general context. Communities and applications of the reviewed methods belong to the Computer Vision community [174], as well as the Robotic [54] and the Photogrammetric com-

munities [212]. Finally, in this survey, we consider *heterogeneous* visual data without restriction, including: raw colour and grey-scale images, depth images, point cloud and 3D models, as well as semantic information extracted from aforementioned data.

However, we advise reader to refer to the recent surveys related to Visual Place Recognition [111, 54, 86] in order to capture a global panorama of existing approaches involved in localization process with visual data.

The rest of the paper is organized in two parts: on the one hand we study two different methods, introducing in Section 2 several data representations used in VBL followed by description of indirect (Section 3) and direct (Section 4) systems; the second part focus on data involved in VBL, with in Section 5 an analysis of the problem of challenging association across data variability and in Section 6 an overview of the different type of database and query used in VBL; finally, in Section 7, trends and representative methods are discussed and Section 8 concludes this work.

2. Data Representation

Table 1: **Features in VBL:** Synthetic overview of features used in Visual-Based Localization. Columns Det. and Desc. stand for detector and descriptor respectively and describe the capability of the feature.

Name	Feature type	Det.	Desc.	Used in VBL
Pseudo Corners detector [129]	Point	✓	✗	[129, 130]
Hessian-affine [126]	Point	✓	✗	[71, 2, 100, 173, 4]
FALoG [216]	Point	✓	✗	[49]
SIFT [109]	Point	✓	✓	[188, 181, 103, 223, 224, 140, 220]
RootSIFT [3]	Point	✗	✓	[125, 4, 173, 199, 200]
SURF [18]	Point	✓	✓	[44, 100, 190, 158, 188, 205]
ORB [169]	Point	✓	✓	[61]
BRIEF [24]	Point	✗	✓	[90, 87]
BRISK [94]	Point	✓	✓	[49, 125, 133]
Learned features	Point	✗	✓	[27, 151, 87]
Lines [206]	Geometric	✓	✗	[65, 7, 130, 161]
Contours [25]	Geometric	✓	✗	[162, 170]
VLD [108]	Geometric	✗	✓	[115]
PGM [100]	Geometric	✗	✓	[100]
GIST [141]	Global	—	✓	[65, 170, 135, 11]
Tiny images	Global	—	✓	[65, 56, 43]
Histogram	Global	—	✓	[65, 138]
Fourier Transform	Global	—	✓	[212]
Convolutional Neural Network (CNN)	Global	—	✓	Refer to table 2
HOG [46]	Patch	✗	✓	[184, 10, 121, 130]
RPN [164]	Patch	✓	✗	[60]
Edge boxes [230]	Patch	✓	✗	[193, 142, 220]
MSER [119]	Blob	✓	✗	[82, 140]
Planar surface	3D	✓	✗	[50]
Normal vector	3D	✓	✗	[99, 50]
Spherical function [179]	3D	✗	✓	[112]
VCLH [33]	Semantic	✓	✓	[33]
Skyline	Semantic	✓	✓	[180, 203, 37]
PointRay [15]	Semantic	✓	✓	[15]
Objects	Semantic	✗	✓	[112, 171, 6, 159]
CNN ImageNet [91]	Semantic	—	✓	[192]

What is the best manner for representing visual data? This central question, present in various Computer Vision, Robotic and Photogrammetric images-indexing applications, leads up to numerous answers. The data representation, termed features, should incorporate as much as possible discriminant information from the

initial visual document and be fast to compute and compare. We present in this section representations used in VBL. Table 1 summarizes the following presented features.

2.1. Local Features

Local features are widely used in VBL and more generally in Computer Vision. Their description occurs at pixel level among a local neighbourhood of several points in the image. The description through local features is two-step: firstly detect salient region (the extraction phase) and then characterize them according to their neighbourhood (the description phase).

Point features. Several criteria are taken into account for the selection of point features: scale, orientation and illumination invariance, as well as computational cost and descriptor vector dimension. A comprehensive list of local feature descriptors used in topological mapping in robotics can be found in Garcia-Fidalgo and Ortiz [54] survey. Krajník *et al.* [87] explore in-deep many combination of detector/descriptor for the specific task of images matching across seasons. The most used point feature in VBL remains the Hessian-affine detector [126] combined with SIFT [109] descriptor. Important contribution from Arandjelović and Zisserman [2] introduces RootSIFT which presents better results in matching step with minor overhead in computational load. SURF descriptors [18], light version of SIFT, are employed when real-time performance are required [44, 158, 191]. Interesting work from Feng *et al.* [49] combines rapidity and precision by using binary BRISK descriptor [94].

Learned local features is a well studied topic [27, 151]. Features are described through CNN trained for the task of similar features association [221]. Although not democratized for the task of VBL, we advice reader to refer to the recent comparison of hand-crafted and learned feature proposed by Schönberger *et al.* [182]. This paper shows, amongst others, that traditional local features perform the best in some scenario related to VBL.

Geometric features. Visual data can be described by primitive geometric shapes. Despite the fact that geometric features are less compliant than point features, they include semantically meaningful information. For example, vertical lines are convenient descriptor in urban environment to represent buildings [7, 130, 161]. On the basis of this observation, Hays and Efros [65] introduce line extraction in combination with others descriptors to describe images. Contour extraction have also been employed by Russell *et al.* [170] to recover the pose of an image in a site of archaeological excavations. Considering 3D data, several works use three-dimensional geometric features like normal vectors [99] or planar surfaces [50].

Point features with geometric relations. The lack of geometric consistency across the whole image is a shortcoming associated with point features. Various contributions propose to overcome this limitation by adding local geometric information directly on the point descriptor [180, 70] or with the geometric association of numerous points [108, 100]. SIFT features contain scale and orientation information, that have been originally used in [70] through the Weak Geometry Consistency framework. Following the same idea, Saurer *et al.* [180] encode features relative pose in the image to perform geometric verification at matching time. Liu and Marlet [108] introduce a geometric descriptor called Virtual Line Descriptor (VLD) by connecting two local features with each other. The subsequent lines are used to reinforce the robustness of the matching process in VBL scenario [115]. Li *et al.* [100] propose a different pairwise geometric descriptor (PGM), showing great results on both urban and landscape scenes.

2.2. Global Features

Another description approach considers the image as a whole and produces one signature with high dimensionality. Compared to local descriptors, global features are considered less robust in viewpoint changes, occlusion and local variations in the image. However, they are computationally less intensive to extract and capture a comprehensive description of the visual data. With the recent emergence of CNN, a new class of very efficient global descriptor have been created.

Hand-crafted features. GIST descriptor introduced by Oliva and Torralba [141] is the most used hand-crafted global descriptor in VBL [170, 11, 65]. The raw image can serve as a descriptor, with systematic resizing in order to obtain thumbnail [65, 43] (potentially augmented with depth information [56]). Simple descriptor computed through an histogram upon various criteria (colour, texture [65] or depth [138]) also provides a fast global information. Taking the image as a whole in a different representation space that is more discriminant for similarity research can also be considered as global description. For instance, Fourier Transform (FT) is used by Wan *et al.* [212].

Learned features. Democratization of CNN in computer vision domain leads to state-of-the-art techniques in image retrieval for urban scenes [1, 60, 83, 160]. Descriptors created through CNN are global features collected by grouping weights of a given layer into a single vector [13]. Table 2 presents main CNN architectures involved in VBL. Firstly, CNN have been used for the task of localization without special training, exploiting inherent domain transfer capability of neural network. It is also interesting to notice that the most discriminative descriptors for the task of image-retrieval are extracted from mid-level convolutional layers instead of fully-connected layers [13, 192].

In recent works [1, 160, 60], authors tackle the problem of fine tuning a pre-trained network for the specific task of similar images association. Arandjelović *et al.* [1] introduce a weakly supervised triplet ranking loss, feeding the network with positive and negative examples before applying the back-propagation. Positive examples are candidates related to a query image and negative examples are non-relevant candidates. Works in [160, 60] use a different approach: two [160] or three [60] identical networks receive negative and/or positive examples at each iteration, and the back-propagation is operated on the different networks. By sharing the weights between the different networks and by applying a relevant common loss, the system can learn a data representation suitable for similarity research. Authors insist on the need of having a clear and large database. Therefore, works from [160, 60] introduce novel methods to automatically reject wrong images, cluster the data into similarity groups and associated positive images with hard negative samples (*e.g.* very dissimilar images) for training. The Time Machine functionality of Google Street-View is used in [1] to gather a large database composed of the same places at different periods of time. Data synthesis by view rendering is also performed for gathering large amount of data [75, 186]. Pre-treatment on training data are more generally used with learning-based methods [82, 26].

Further explanations regarding CNN and specially aggregation methods performed inside the network can be found in the next section §3.1.

2.3. Hybrid Features

We have decided to call hybrid features two different kinds of approaches: the first one consider features that cannot be considered neither as local nor global (i.e. patch or blob) and the second one are features that combine several types of descriptors.

Patch features. Patch features consider region of interest in the image, it can be interpreted as a compromise between local and global features. The patch could be manually extracted (with a fixed grid on an image, or a sliding window [46]) or automatically chosen in according to image saliency [119]. The discriminative HOG [46] descriptor has been used in VBL for capturing architectural cues of building and landmarks [184, 10, 121, 130]. In the work of [140, 82], MSER blob detector by Matas *et al.* [119] is used to extract visual information. Sünderhauf *et al.* [193] present promising works where the feature patches are automatically extracted with an edge boxes detector [230]. Another CNN approach is introduced to perform VBL in [60], authors use a custom region proposal network [164] to extract regions of interest.

Combined features. Image features are often combined to provide a complementary description of the scene [96]. Hays and Efros [65] combine up to 5 global and local descriptors to qualify images. Azzi *et al.* [11] use features in a cascade scheme to first narrow the search scope with global feature GIST and then select the good candidates with local features SIFT. Morago *et al.* [130] use a combination of local and patch features to describe repetitive shapes. Patch detector coupled with global descriptors are a common use of multiple features, as illustrated in [82, 60, 193, 220]. Recent work from Bhowmik *et al.* [19] study a new approach for pairing various descriptors in order to increase the result of the retrieval step depending on the targeted dataset.

2.4. Semantic Features

Previously presented data descriptors belong to an abstract class of features that focus on the raw data extracted from the acquisition sensor. On the contrary, semantic representation aims to categorize the data meaningfully. Works in [33] introduce a semantic line-based descriptor. The vertical lines are extracted using Canny filtering and coded into VCLH (Vertical Corner Line Hypothesis) to represent building corners. Skyline features introduced by Saurer *et al.* [180] have been used to describe mountain panoramas. Dehaze segmentation is used to extract the skyline that is thereafter encoded in a curve bin descriptor. More recently, Bansal and Daniilidis [15] use *PointRay* to represent building corners. We refer reader to §6.3 for more investigation about semantic representation in VBL.

3. Indirect VBL Methods

The aim of **indirect** VLB methods is to retrieve a set of data presents in the database that are similar to an input query. This is a problem related to Content Based Image Retrieval [32, 228]. As the visual data used in VBL are augmented with geospatial information (*e.g.* a geotag associated to an image), retrieving documents comparable to the input provides an information on the possible location of the query. This image-retrieval-like problem is two-step: description of the visual data (for both the query and the database, see Section 2) and similarity association across the description vectors previously extracted.

We explore three key steps present in indirect VBL: features aggregation, similarity research and candidates re-ranking.

3.1. Features Aggregation

The similarity search can be computationally expensive when the data is described by a large descriptor, i.e. a vector of high dimension. Particularly, local features (§2.1) are prompt to produce a large number of descriptors for each single data. Features aggregation is then performed in order to reduce the dimensionality of the descriptor vector. In VBL, the aggregation process emphasize specific features that are more beneficial for the localization task.

Quantization. Quantization methods have been widely adopted in image-retrieval domain since pioneer contribution of Sivic and Zisserman [185]. They consider the problem of object retrieval in an image described through local features in the same manner as text document research. Words equivalent in image domain becomes local features and a dictionary is build upon a large set of features extracted from visual documents' database. These features are clustered to reduce the size of the dictionary; clusters' centroids are then called visual words. For each visual word in the dictionary, and inverted file is maintained to efficiently retrieve all the data that present this specific visual word. The bag of features (BoF) associates a vector of the dimension of the dictionary containing the visual word frequency of a specific visual document. With this representation, data similarity can be efficiently computed by a simple inner product of their respective visual word frequency vector.

Feature to visual word assignment. BoF original scheme [185] proceeds to a hard assignment from the extracted feature to the nearest visual word in the dictionary. However, depending on where the feature lies inside the Voronoi cell created in the clustering step, hard assignment can deteriorate the representation of the visual document. Soft assignment [156] methods have been considered by associating the feature according to a linear combination of the k nearest visual words. Hamming embedding (HE), introduced by Jégou *et al.* [70], subdivides Voronoi cells and associates to each feature a binary signature to refine its position in the visual vocabulary. This method leads to excellent result in term of accuracy and rapidity and is still used in state-of-the-art VBL [4]. Inspired by Fisher Vectors (FV) formulation [154], Jégou *et al.* [73] introduce Vector of Locally Aggregated Descriptors (VLAD) representation for image-based retrieval. The difference between feature and its closest visual word is assigned to the final descriptor, instead of the visual word itself. The underlying idea behind VLAD representation have inspired various VBL methods [82, 199, 1, 220]. For instance, Kim *et al.* [82] introduce PBVLAD method to locally fuse SIFT features detected inside a MSER blob. Novel features aggregation methods have been recently presented in [74, 147].

Weighting scheme. The weighting step is supposed to emphasize discriminative features regarding the similarity comparison. Original method by [185] used *tf-idf* weighting, relying on the occurrence frequency of the features in the database. Jégou *et al.* [71] handle the problem of intra and inter burstiness of visual words (i.e. the fact that a feature is more likely to appear in an image if it has already been detected once) by adapting the weight of the visual words before (inter-burstiness) and during (intra-burstiness) the query process. Torii *et al.* [200] tackle the problem of visual burstiness introduced by repetitive structures (abundant in urban environment) and introduce meta-features encompassing several similar descriptors (comparable both in their descriptor vector and their spatial location in the image). Such improvement permits a dense extraction of local features in images, bringing superior result in urban environment VBL [158, 199]. Another work from Morago *et al.* [130] that exploits the redundancy present in buildings facades. Recently, Arandjelović and Zisserman [4] improve *tf-idf* scheme by considering the descriptors' density in feature space. With their DisLoc weighing, 7% of the less discriminative visual words can be removed from the database without impacting the performances

of the similarity computation. Mousavian *et al.* [132] introduce semantic knowledge in the local feature weighting process, reducing the impact of features associated with non-relevant elements for localization (i.e. elements that are likely to change or disappear, such as trees and cars).

Table 2: **CNN image descriptor in VBL:** Details of CNN architectures used in VBL for the purpose of data representation. Abbreviations present in the table refer to: Regions of Interest (ROI), Landmark Distribution Descriptors (LDD), Maximum Activations of Convolutions (MAC), Regional MAC (R-MAC), Contextual Reweighting Network (CRN). † Specify if the network is feeded with the whole image or with fragments. ‡ Aggregation method within the CNN architecture. * Off-the-shelf means that the network was initially trained for scene classification [91] or over computer vision tasks.

Reference	Input†	Aggregation‡	Training*	Commentary
[13, 192, 220]	Whole image	No	Off-the-shelf	Use fully connected layers as image features
[93]	Whole image	No	Off-the-shelf	Aggregating of multiple features to form a panorama descriptor
[193, 220]	Image ROI	No	Off-the-shelf	Data reduction with Gaussian random projection
[142]	Image ROI	No	Off-the-shelf	The multiple extracted features are associated into a LDD
[12]	Whole image	SPoC	Off-the-shelf	Feature dimensionality reduction with PCA
[163]	Image sub-part (regular grid)	MAC	Off-the-shelf	Extraction of multiple features at different scales on the image
[229]	Image ROI	Partial Mean	Off-the-shelf	Two-stage features and patches mean pooling
[198]	Whole image	R-MAC	Off-the-shelf	Feature dimensionality reduction with PCA
[60]	Whole image	R-MAC	Fine tuned	ROI extracted on the convolutional layer side rather than on the image
[160]	Whole image	MAC/R-MAC	Fine tuned	Fine tuning with two siamese shared-weight networks
[1]	Whole image	NetVLAD	Fine tuned	Network specially fine-tuned for the task of image retrieval
[68]	Whole image	NetVLAD	Fine tuned	Evaluation of panorama representation through CNN
[83]	Whole image	NetVLAD	Fine tuned	Use of a CRN to emphasize discriminative features

Aggregation in CNN. In CNN based methods, feature aggregation is also a subject of study. There are two different aspects of aggregation in CNN domain: gathering of features extracted from networks and intra-pooling of deep features within the CNN.

Multiple features aggregation. CNN descriptor can be combined with local or patch detectors, in order to obtain sparse representation of the data (see table 2 for examples). In this case, features extracted from the image can be gathered into a single descriptor, like in the BoF framework. VLAD embedding is employed in [220] and in [142] patches are sorted according to their relative position in the image and aggregated in a Landmark Distribution Descriptor (LDD) to improve the subsequent similarity search. Zhi *et al.* [229] exploit the intensity response of each patches to discard the descriptors with the lowest intensity. In [68], authors create panorama features by aggregating multiple image representations (extracted from a CNN) in a memory vector.

Pooling of deep feature maps. The meaningful representation of an image through neural network is achieved by extracting responses of convolutional layers [13, 192]. A convolutional layer can be seen as a feature bloc, composed of several activation maps of the same size. Considering the raw response of such a layer results in a high dimensionality feature vector. In order to capture a more discriminative image representation, several activation map pooling methods are applied. Table 2 presents the different convolutional layer aggregation scheme employed in VBL. Maximum Activation of Convolutions (MAC) [163] reduce the feature size by aggregating the maximum of each activation maps into unidimensional vector. Sum-Pooled Convolutional features (SPoC) [12] shows superior results compared to MAC aggregation. Instead of computing the maximum over all the activation maps, authors simply sum the responses for each map. Regional Maximum Activation of Convolutions (R-MAC) [198] is an improvement of the precedent MAC method, consisting of the computation of the maximum of

activation over regions of various sizes on the activation map. Gordo *et al.* [60] achieve state-of-the-art performances by combining R-MAC representation with a custom Region Proposal Network (RPN) that autonomously detects regions on the activation map to compute the max-pooling. An entirely trainable aggregation layer, called NetVLAD, have been proposed in [1]. Authors design a differentiable architecture that aim to mimic VLAD aggregation scheme. In combination with an adapted training framework, this architecture seems to be the best suited for VBL tasks. Kim *et al.* [83] use the NetVLAD aggregation layer coupled with an Contextual Reweighing Network (CRN) to downgrade irrelevant features according to their local neighbourhood, without the use of any manually annotated data.

3.2. Similarity Research

Comparison between descriptors is a trivial operation: it consists in a simple distance computation (with $L2$ norm as usually used metric) between vectors. However, when the number of descriptors is very large, a brute-force approach cannot be considered and similarity search algorithms are employed. We describe below these approaches.

Pre-processing. Dimension reduction of descriptor is often performed to reduce matching time and memory footprint. The most used technique remains the Principal Component Analysis (PCA). PCA is applied on hight dimension vector, *e.g.* weights extracted from CNN layers ([1, 60]). PCA has also been used to reduce the size of local features aggregated vectors [82, 199] or global descriptors [138]. Gaussian Random Projection is applied in [193, 142] and in a different work, binary locality-sensitive hashing [192] is used instead. To reinforce data consistency, whitening could be applied to final features before the similarity search [69, 59, 198, 1, 60, 160].

Nearest Neighbour Search. In some works, when the amount of data to compare remain acceptable, brute-force retrieval (or exact nearest neighbours retrieval) procedure can be employed to retrieve the closest neighbours. This is the case when a single vector is used to describe a document, i.e. where global descriptors are used (§2.2). Global descriptor from CNN trained for the task of image description [13, 192, 160, 60, 1] produce a global feature vector that is afterwards ranked against each vectors in the database according to its cosine distance. Other techniques based on local or hybrid features [223, 224, 193] perform brute-force comparison, limiting the number of features that can be handled.

Exact nearest neighbour search becomes impracticable when the amount and/or dimensionality of the features are too large. Authors then turn to approximate nearest neighbour search to trade efficiency for rapidity, thus accepting some errors in the retrieved neighbours. Approximate matching involve hashing methods [57] and quantization frameworks [140, 155, 72]. Interested readers may see [213] for more details.

Several nearest neighbour search algorithms are implemented in the `FLANN` library [134], and in the new Facebook `FAISS` library [76].

Machine Learning Methods. Learning the distribution of the extracted features is an alternative to aforementioned nearest neighbour search methods.

SVM classifier is used in numerous works [184, 26, 121, 10] to cast the similarity research as a classification task. Cao and Snavely [26] initially cluster the database according to the resemblance of the images. On top of this graph of similar images, they trained SVM for each cluster and at query time oppose the input image to all classifiers. By selecting the data associated to the SVM reaching the higher score of classification, this approach permits to quickly retrieve a pool of similar images. In [121, 10] authors train linear classifiers on HOG descriptors to robustly retrieve similar images that present extreme appearances changes. Aubry *et al.* [10] take the advantages of Linear Discriminant Analysis (LDA) data representation in order to avoid expensive SVM training (like hard negative mining used in [184, 82]). Similarly, Kim *et al.* [82] train SVM classifier to predict the robustness of extracted descriptors. This improves the matching process and reduces the number of features to compare against the database.

Lu *et al.* [112] introduce a Multi-Task Learning (MTL) layout designed for features similarity association. Works from Torralba *et al.* [202] and Ni *et al.* [138] present VBL methods that are able to localize an input query among a set of predefined places. Authors embedded the recognition process into probabilistic framework, Gaussian Mixture Model (GMM) in [202] and epitome in [138], trained upon images representing different areas. Such paradigms allow an easy integration of additional features (such as depth information [138]).

Other Matching Methods. Stumm *et al.* [190] introduce an innovative method based on graph matching. The visual vocabulary abstraction is employed and augmented with a graph of covisibility of the visual words in images. The graph is constructed as follows: nodes represent visual word detected in images and edges are created between two nodes if they are seen together in a same image. This formulation permits integrating geometric relations between the extracted features. Authors use a graph kernel for the similarity comparison among the query graph and the database [189, 191]. Notice that graph-based approaches are often employed when scenes are described by spatially organized semantic clues such as office furnitures [171] or street equipments [6].

Area correlation algorithms is another approach for computing data similarity. Simple forms of correlation like Sum of Squared Difference (SSD) or Sum of Absolute Difference (SAD) have been used in VBL to compare images [157, 127]. Wan *et al.* [212] use PC (Phase Correlation) on images described with FT (Fourier Transform) in order to be robust to shadow artefacts. In the work of [43], authors compare shadow invariant grey-scale images with Zero Mean Normalized Cross-Correlation (ZNCC).

3.3. Candidates re-ranking

Data can be processed after the similarity research to improve the final result. Post-processing methods are widely used to re-rank the candidate list, improving relevance of retrieved data [36].

Specific VBL re-ranking. Unlike conventional methods of object-retrieval, indirect VBL can benefit from geo-localization information associated to the documents present in the database. As discussed earlier, this information can be used to construct structured graph for the similarity search process [201, 26] or exploited to re-rank the candidates list [223, 224, 173]. Zamir and Shah [223] introduce this geographic re-ranking after a classical image-retrieval algorithm to quickly remove irrelevant candidates. Authors go one step further in [224] and embed the matching process within a Generalized Minimum Clique Graphs scheme to retrieve consistent candidates according to the GPS tag associated to the visual data. Sattler *et al.* [173] generalize the problem of visual burstiness introduced by [71] to a geographic level, introducing the concept of geometric burstiness. They improve the relevance of the ranked list of candidates using position and popularity meta-information of database images.

Innovative contribution from Torii *et al.* [201] refine the query location with a linear interpolation in the feature space domain of the closest database images. The database is arranged with a graph representation, where images represent the nodes and the edges encode spatial relation, i.e. images that are close to each other (according to their GPS-tag) are connected. Firstly, a set of putative candidates are retrieved with a conventional quantization method, then the discrete feature space of the candidate is extended into a continuous space by linear interpolation according to their position in the graph. The exact position of the query is then guessed according to linear combination of GPS information of the database images. Although promising, this method relies on complete panorama images, limiting its range of applications.

Generic re-ranking. Query expansion is a post-process that re-query the database after a first retrieval step to increase the recall rate [41, 40, 197]. However, increasing the recall rate is not the main concern of VBL indirect method. Indeed, as exposed in the introduction, a perfect VBL indirect system should retrieve at first position the closest visual document present in the database. However, more suitable top ranked candidates in the list of retrieved data could benefit to a subsequent pose estimation step [188]. The VBL system presented by Cao and Snavely [26] increase the diversity of retrieved images by introducing a probabilistic re-ranking on the assumption that the first ranked candidate is not a good one and by maximizing the probability that the second one is. Last but not least, geometric consistency check is often used to reject wrong matching. Relative pose between the query and the database candidates is computed by considering homography or multiple-view transformation (see more details in the following §4.1), and candidates that produce the most consistent pose are ranked up. Philbin *et al.* [155] democratize the use of spatial verification by introducing prior on the pose of the photography by assuming a top-oriented view. Authors perform spatial check hierarchically to get more flexibility between time computation and retrieval precision. The geometric transformation between the query and the candidate is usually computed with minimal algorithm embedded in random consensus, like RANSAC [52]. There exists multiple alternatives to the classical RANSAC algorithm. PROSAC by [39], used in [48], prioritize specific features during the random selection step. We can also enumerate LO-RANSAC used in [155] and AC-RANSAC in [159, 158]. Novel method F-SORT presented by Chan *et al.* [34] show outstanding result both in term on matching quality and computation efficiency. Notice that these algorithms, beside improving the relevance of the retrieved candidates, can give information about the relative pose of the query. That is why numerous direct methods, presented in the next Section 4, rely also on these techniques.

4. Direct VBL methods

At this point, we introduce the notion of **direct** VBL methods that instantly recover the exact 6 DoF pose of the query according to a known reference. Compared to indirect approaches, direct methods provide a more accurate query pose to the detriment of the area coverage. Indeed, direct VBL requires a smaller database and in some case a coarse prior information about the query pose. From this class of methods, we consider the three following approaches:

Direct VBL with prior: these methods are built upon the assumption that we get a prior information about the query pose. The pose prior can be obtained through localization sensor (GPS [35, 7, 157], magnetic compass [225, 194]) or by using an indirect VBL system [201, 188, 177].

Features to points matching: this class of methods performs the global localization of the query by establishing correspondences between two-dimensional features extracted from an image and three-dimensional point cloud model of the environment (see figure 1b).

Pose regression approaches: the last considered family of algorithms are methods that learn to directly regress from an input visual data to its corresponding pose. Standard regression techniques [183] and CNN architecture [80] are employed to perform this task.

4.1. Direct VBL with prior

Many applications in Computer Vision and Robotics require an initial pose estimation of the visual data acquisition system: augmented reality [7], visual odometry [146], SLAM [128] or visual servoing [116], to name a few. Coarse estimation provided from standard geo-localization system (*e.g.* GPS) are not accurate enough for such applications, and other processing are required to initialize the system with a suitable pose.

Methods overview. Arth *et al.* [7] introduce a method to estimate a fine pose of a mobile camera to initialize AR applications or SLAM systems. Given a coarse prior pose of the camera (obtained by GPS and compass embedded in a smart-phone), authors refine the global pose by matching extracted geometric features to buildings outlines. Similarly, Russell *et al.* [170] investigate techniques to retrieve the pose of realistic painted or drawn piece of art according to recent photographies. Given a coarse pose prior, the query location is refined by establishing edges correspondences with the real model. Poglitsch *et al.* [157] introduce a particle filter to perform localization. The particles are randomly generated over a 3D model from a coarse position information from a GPS sensor. Widely spread in robotic community, particle filters have also been used to refine a coarse pose of a mobile robot in known ground 3D space [117] or an aerial map [38, 23].

Similar to work described in [168, 177], Song *et al.* [188] present a typical cascade scheme to estimate the 6 DoF pose of a given image. The authors perform a first step indirect method to retrieve a set of potential similar candidates, and then refine the pose with relative pose computation algorithms.

An aerial localization of an unmanned aerial vehicle (UAV) from down-looking camera images is presented in [212]. Authors estimate a fine pose by registering the embedded camera image on a satellite images. A coarse pose information is needed for reducing the search scope. This solution is also validated for VBL on foreign planet of the solar system. Over works present VBL for lunar rover in extremely challenging condition [211]. In order to perform a fine pose estimation under hard conditions (lunar panorama with few discriminative visual elements), the authors have to use a reliable prior pose of the robot given by IMU and wheel odometers.

Pose Computation. Numerous techniques can be applied to recover the exact 6 DoF of a given query. If the reference data are geo-localized images displaying primarily planar surfaces, homography retrieval can be employed [53]. The relative pose from the query and the reference images can also be regressed with multi-view algorithms [64]. Nistér's 5-points algorithm [139] or the 8-points algorithm by Hartley [63] are used in many VBL scenario [158]. Recently CNN have been used to warp an affine or thin-plate-spline transformation between two images [166], or to estimate the relative transformation between two images [122]. Although not precise as classical methods, the presented network is able to deal with drastically different images in appearance. If numerous reference images are available, more complete methods are used involving trifocals geometry [64] like in [188]. Kneip and Furgale [84] introduce `OpenGV` library, a modern `C++` tool to compute relative and absolute pose with various algorithms.

Pose refinement. Depending on the available data, heavier processing can be applied to refine the query pose. Bundle adjustment is the widely used technique when dealing with 3D structures or point cloud obtained from images. Works from [125, 211, 53] apply bundle adjustment to refine the first guess pose from their method. Local bundle adjustment are used when real-time performances are targeted [101, 158]. Another famous refinement method is the Iterative Closest Points algorithm (ICP), used in VBL context in [170, 180, 130]. Paudel *et al.* [148] provide algorithms to obtain an optimal alignment between images and point cloud data, or between point cloud and 3D model [149].

4.2. Features to Points matching

A widely represented family of direct method aims to regress the pose of a camera based on the analysis of a 3D point cloud reconstructed by SfM algorithms. The principle of these methods is to establish 2D features to 3D points correspondences (F2P). In a first step, three-dimensional representation of the environment is built thanks to many images. Triangulated points within this structure are associated to the local features (most of the time SIFT vectors [109]) extracted from all the images where the considered point is visible. At query time, local features from the image to localize are matched against the set of pre-computed 3D points. Finally, the features to points correspondences permit a 6 DoF pose estimation of the acquisition system.

These methods share a lot of similarity with indirect methods described in Section 3 and they present the same two-step pipeline: data description and data similarity association. Yet the use of a geometrically structured database introduces interesting elements not exploitable in a classical image-retrieval scheme [178].

Methods overview. Between methods based on prior information and F2P methods, works from Arth *et al.* [8] present a system that recover the pose of a smart-phone camera by confronting an image to a subset of 3D points that should be visible in the query according to a prior pose information. Irschara *et al.* [67] introduce the first F2P method based on SfM environment representation. Authors perform scalable VBL by registering the point cloud into synthetic visual documents covering the entire model. Latter improvement by Li *et al.* [101] reverse the conventional process by searching from the point cloud correspondences in the image (P2F), instead of matching features from the image to points. This formulation causes an overhead in computation but is correctly handled by considering a compressed version of the SfM model and by implementing end-conditions and rejection cases in their algorithm.

Sattler *et al.* [174] consider the original features to points correspondences scheme by [67] and introduce a Vocabulary-based Prioritized Search (VPS) inspired by BoF matching method. Subsequent works by the same authors [176] augment the VPS framework with the points to features matching P2F [101]. Li *et al.* [102] show that the class of methods introduced in [67, 101] can deal with large environment. Authors augment the P2F matching with hypothesis of co-occurrence of 3D points present in a close neighbourhood. Based on similar spatial observation, Sattler *et al.* [172] consider visibility graph to reject wrong matchings. Heisterklaus *et al.* [66] introduce MPEG compression for visual document in order to speed-up the system. In the work described in [48], authors use the descriptor redundancy associated to 3D points to train random ferns on the top of each points. F2P matching time requirement is by the fact greatly reduced.

Works from [125, 113] tackle the problem of VBL embedded in a mobile device with limited memory storage and computational power. To achieve real-time performances, authors in [125] produce a very light 3D model to track the mobile camera in an urban environment. They send at regular interval key-frames to a server that is in charge of computing the global pose of the camera regarding a pre-produced point cloud. Aligning a light relative point cloud reconstructed with SfM to a bigger one have also been investigated in [112]. Svarm *et al.* [194] consider the problem of VBL with F2P matching as a combinatorial optimization problem and design a fast outliers rejection scheme. This promising work have been improved through [225] contribution.

State of the art VBL methods based on SfM are dominated by techniques combining previously mentioned improvements [176]. Recent work by Feng *et al.* [49] reduce drastically the computational power requirement by considering fast point extractor and binary descriptors combined with an efficient similarity research. Authors show an order of magnitude in time reduction without any pose estimation performances deterioration.

Features to Points pose estimation. F2P provides correspondences between 2D pixels and 3D colourized points. Defined by Hartley and Zisserman [64], perspective-n-point (PnP) formulation is the most common tool to recover the absolute camera pose according to the point cloud reconstructed by SfM.

Embedded in a random consensus scheme (see §3.3), six correspondences between the image and the 3D model are sufficient to retrieve the pose, if we have no information about the intrinsic parameters of

the camera [48, 101, 101, 66]. This formulation is known as P6P and can be solved with Direct Linear Transformation (DLT [64]).

In particular cases, three correspondences between the image and the model are sufficient (P3P pose computation problem). Especially, the pose estimation problem can be reduced to a P3P formulation if the intrinsic parameters of the camera are known [67, 125], or if 3 or more DoF are fixed [225, 158]. In those particular cases, P3P solver introduced by Kneip *et al.* [85] is mostly used to recover the pose.

4.3. Direct pose regression

The last class of reviewed direct methods cast VBL as a pose regression problem. Two different kinds of regressors are employed in the literature: regression forest and CNN.

Regression forest. In the initial works by Shotton *et al.* [183], authors encode, thanks to RGB-D data, the global position of each pixel associated to a known environment in a regression forest. At query time, a handful of pixels from a depth camera frame are processed into the regression forest. The multiple pose hypothesis obtained for each pixel is then optimized in a random consensus to regress the camera position and orientation. This method is fast and precise and can be used on texture-less data. However, the depth information associated to each pixel is needed and the authors have to train a specific forest for each 3D scene. This initial method have been improved in [62], where authors take in consideration several candidates for the final pose regression obtained by trained predictors. Valentin *et al.* [204] introduce mixture of Gaussian to represent the uncertainty associated with the regression forest prediction and significantly improve the 6 DoF estimation by embedding this information within the full camera pose regression step. The regression forest have been replaced by Neural Network (NN) in [118], bringing slightly better result at the cost of computational overhead. Meng *et al.* [123] consider only RGB images at query time. The loss in precision is compensated by a post pose refinement step based on nearest neighbours search with sparse extracted SIFT features.

Inspired by works presented above, Glocker *et al.* [58] design a system based on regression ferns to quickly associate an RGB-D image to a binary feature. Ferns produce descriptor according to randomly initialized binary rules, and a look up table is maintained to directly associated image signature with 3D pose in the scene. Presented system is less precise that the one presented by Shotton *et al.* [183] but has the advantages of not relying on a heavy pre-processing step (i.e. the spawning of the regression forest). Along the same line, Cavallari *et al.* [31] propose a new method based on pre-trained regression forest. This method permit to recover the pose of a RGB-D camera without prior knowledge about the 3D scene, more precisely than Glocker *et al.* [58].

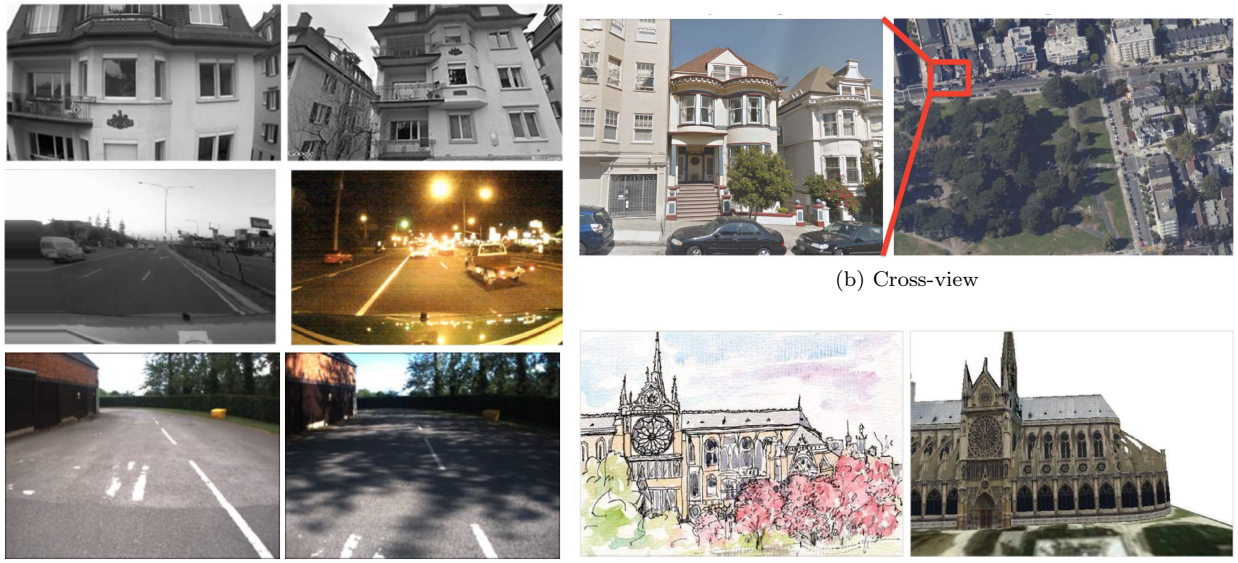
CNN regressors. Introduced in 2015 by Kendall *et al.* [80], PoseNet consists of the fine-tuning of a CNN for the task of localization. The network is trained upon a set of paired image/pose and automatically regress the 6 DoF pose of a camera that acquired a colour image. The pose obtained through this method is not as accurate as the pose obtained with "classical" direct methods [49, 176] but provide great tolerance to changes in scale and appearance. Compared to regression forest [204], CNN seems more appropriate to handle large environment and does not rely on depth information.

Recent improvement have been proposed by the original authors [78] to integrate an uncertainty estimation in the regression process. Liu *et al.* [107] integrate this CNN architecture with only depth map information for recovering the pose of a camera in complete obscurity. The work by Walch et al. [209, 210] present a combination of a PoseNet [80] with a Long Short-Term memory (LSTM) units plugged at the output of the network in order to encode stronger spatial information from the image. This combination slightly improves the precision of the system. Jia *et al.* [75] highlight the limited number of training example available for CNN pose regressors. Though the learning transfer seems to be efficient [80], authors propose a new method to gather supplementary image/pose pairs for the fine-tuning step. They generate artificial images from a dense point cloud model obtained by SfM thanks to a rendering software. Computer graphics shaders effects are added on some rendered views for simulating various illuminations. Contreras and Mayol-Cuevas [42] exploit this CNN architecture in order to create a fixed size map that can be improved by adding new trajectories. Authors were able to reduce the original size of the CNN by factor of three while maintaining similar localization performances on indoor scenes. Recent contribution [79] investigate new loss-functions for the training phase of the CNN, mimicking the philosophy used in multi-view geometry standard systems [64].

Differently, recent work from Weyand *et al.* [217] consider the localization problem as a classification task. They perform a worldwide training on 126M images categorized into 26k places across the globe. According to a given image a CNN, named PlaNET, estimates a map of probable location for the query. Localization

of multiples photos taken from a common album can be performed by augmenting the original network with a LSTM layer. Vo *et al.* [208] push further the study of such a neural network and conclude that the features extracted from layers of PlaNET are more discriminative to determine the location of an input image that the CNN classifier itself. By extracting features instead of using a classification algorithm, their contribution is closer to the original world-wide localization method IM2GPS [65].

5. Data with Dissimilar Appearances



(a) Appearance changes (b) Cross-view (c) Cross-domain

Figure 2: **Illustration of appearance changes present in VBL system:** (a) Visual dissimilarity between the query (left) and the closest image in the database (right). Cause of the change, from top to bottom: viewpoint differences in a system for MAV localization in urban environment from [115], extreme illumination changes from [127], shadow interferences from [43]. (b) Cross-view localization system from [105]: left represent the ground-level query image and right the bird's eye view of the same scene (red rectangle). (c) Cross-domain VBL system from [10]: on the left the query painting and on the right the corresponding pose according to a 3D model.

As pointed out by Lowry *et al.* [111], permanent changes occurring in our environment is a huge concern in vision domain. In VBL, to the difference of SLAM based navigation methods [54, 111], the environment representation (i.e. the database) is most of the time acquired at a single date and query can be opposed to the system years after. To take into account local changes of the environment the database needs to be updated. Depending on the size of the covered area, database update can be a costly operation. Thus, an ideal VBL system should be able to handle minor visual changes from various sources: daily and season cycle, difference in viewpoint or modifications of the local geometry of the scene. In this section we review selected VBL papers that tackle the problem of visual changes in the environment. We dedicate the second part of the section to localization methods that consider extreme appearance changes between the query and the database, namely: cross-view and cross-domain VBL systems.

5.1. Appearance changes

Viewpoint changes. Common visual acquisition systems capture a part of the environment lying inside the frustum of the sensor. Indeed, camera are oriented-device and due to the complex geometry of our surrounding environment, viewpoint changes in visual data impact drastically the appearance of the same scene [77]. To handle those changes, local descriptors described in §2.1 have been widely used. By describing partial areas of the whole scene, local features are naturally robust to a certain amount of changes introduced by difference in viewpoint, small occlusion or scene modification. Wan *et al.* [211] treat extreme viewpoints changing (when the camera are facing each other) in repetitive lunar environment. To achieve VBL in such conditions they match the ground part of the image (which is subject to large affine distortion) with a fully affine invariant feature [131].

Image rectification [53] is also employed in VBL to minimize appearance changes introduced by different viewpoints. With strong assumption on the environment where the localization is performed (*e.g.* such as Manhattan world assumption [135, 33]), images rectification ensure that facing direction of all visual data will be barely the same. With the hypothesis of an urban scene, vanishing points can be extracted [95, 64, 53] and images rectified to display front facing buildings [165, 35, 130, 7, 33].

Other approaches consists of filling the database with additional data to cover all the possible viewpoints for a given environment. Milford *et al.* [127] generate translated view on a database road-circuit for preventing miss-matches if the car, carrying the acquisition system, is moving on a different traffic way than the one used to collect the database. Notice the use of a CNN depth estimator from mono image in order to produce consistence synthetic shifted-views. Work from [67, 10, 199] increase the number of documents in the database by automatic data generation to ensure that whatever the viewpoint of an incoming query, a document displaying a similar view can be retrieved. Majdik *et al.* [115] perform air-ground matching of picture taken by a Micro Air Vehicle (MAV) against street view images (top of figure 2a). The main challenge outlined in this paper is the large difference in angle viewpoint. Authors generate artificial view from both the database and the query image to handle the affine transformation introduced by altitude differences (inspired by the work of [131]).

Illumination invariance & long-term localization. Lowry *et al.* [111, Section VII] explore exhaustively Visual SLAM methods that perform strong illumination invariance place recognition (*e.g.* SeqSLAM [128, 152, 153] or FAB-MAP [44, 45, 150]). Illumination perturbation are caused by three main phenomena: weather conditions and illumination changes across season, daily cycle and finally shadow casting (see figure 2a for illustration). In [110], authors present an invariant-free image representation in order to overcome aforementioned perturbation in visual domain. In a more robotic-oriented-scenario, Mühlfellner *et al.* [133] investigate map invariance representation when multiple instances of the same environment are available.

Seasons & Weather. Illumination changes are usually handled at the beginning of the VBL pipeline, during the data description step. Local features robust to illumination, like SIFT or SURF, consider gradient quantity in order to be invariant to pixel intensity variation caused by different illumination conditions [81]. However, Valgren and Lilienthal [205] have shown that these representations are not well suited for similarity association across season cycles. GRIEF local descriptor [87] (derivatives of BRIEF [24]) or ORB feature [61] show better results for this task. The use of heterogeneous databases (i.e. composed of data acquired by different supports, see §6.1) constrain the system to be robust to disparate illumination conditions [1]. Works described in [90] model seasonal-like cycle in a probabilistic framework in order to downgrade features that are not likely to appear during a given period of time. Rosen *et al.* [167] also propose a model to take in account the features persistence, decreasing the probability of encountering a feature that have been met for the first time a long time ago. On the other hand, learned descriptors show good performances if trained for the specific inter-season matching task [27].

Nocturnal illumination. In some application, especially for vehicle localization, VBL has to be performed during a complete day, including overnight [121, 127] (see middle example of figure 2a). Dense descriptors' extraction used in [199] exhibit promising result for daytime to overnight images matching. At first glance, artificial lights ubiquitous in urban scene can be considered as sources of disruption. However, Nelson *et al.* [136] focus on this particular clues to perform localization across only night road images.

Shadows. Some researches focus on the specific perturbation introduced by shadow casting over images. Wan *et al.* [212] outline that satellite and overhead images can change drastically in appearance depending on the relative position of the sun during the day. Authors show that Fourier transforms can be used to create shadow-invariant image representation. Corke *et al.* [43] implement the shadow suppression method presented in [51] to localize street images with important depth artefacts projected by trees or buildings. This method still remains very sensor-dependent.

Dynamic scene. As mentioned previously, methods based on local descriptors are prompt to handle local changes in images due to dynamic modifications of the environment (*e.g.* vegetation growing, buildings construction or annihilation, presence of pedestrians or vehicles, partial occlusions, etc.). Several investigations have been led for designing robust descriptors to local geometric changes. Kim *et al.* [82] train SVM classifiers to discriminate strong and weak local features for the VBL task. The method shows promising results

where features are more often selected when they are attached to persistent objects, such as facades, and dismissed when they represent ephemeral or changing elements, such as people or trees. Based on similar observation, Mousavian *et al.* [132] introduce down-weighting of irrelevant features according to their semantic class. Learning approaches have also been investigated in other works. Arandjelović *et al.* [1] train a CNN for global description upon images from the Google Street View Time Machine to get diverse representation of the same scene captured over a period of ten years. From this kind of representation of the environment, persistent clues can be efficiently extracted [137]. Similarly, Kumar [93] proposes a CNN approach for place recognition across seasons.

5.2. Cross-appearance localization

Subsequent part focus on methods that reach an extreme with change-invariance consideration by creating cross-appearance algorithms for VBL. We distinguish between two main categories of applications: cross-view VBL, where authors localize a ground-view image against database of aerial images, and cross-domain VBL, where the purpose is to localize an image of a certain nature within a database of different nature.

Cross-view. Cross-view localization, also denoted as ultra-wide baseline matching [16], consider the problem of ground level localization from aerial-level set of photo shoots (see figure 2b for an illustration of data association targeted by cross-view systems). Cross-view VBL is motivated by the fact that satellite photographies are rich sources of information, available almost all over the globe. However, finding similarity between data acquired at a ground level and data captured with flying devices is a hard task due to the extreme change in viewpoint. A series of works consider cross-view localization [104, 218, 30, 207, 195]. In [218, 207], authors investigate the use of a CNN to automatically associate ground level images taken from street view service with fine-grained overhead images. Vo and Hays [207] compare several CNN architectures and conclude that triplet trained network provides the most suitable descriptors for cross-view matching. Rotation invariance between ground and overhead images is also studied through auxiliary loss and special training.

In [17, 16, 105], authors use bird's eye imagery to localize ground level snapshots. Bansal *et al.* [17] method relies on ground level images rectification, like methods focused on viewpoint changes (refer to §5.1).

Cross-domain. Another field of research where the data association is very challenging is the cross-domain localization (an example of cross-domain VBL is presented in figure 2c). Russell *et al.* [170] work, followed by Aubry *et al.* [10] contribution, focus on the task of retrieving the pose of an old hand-drafted document (a sketch or a painting) according to a known realistic representation. In [10], hard training of HOG-based descriptors are used to capture the global shape of the architectural scene displayed in the documents, in the same manner as [184]. Results are impressive, but the used descriptor is not robust to viewpoint changes. Cross-domain techniques are also used to recover the pose of ancient photographies and to confront them with current data [14, 19].

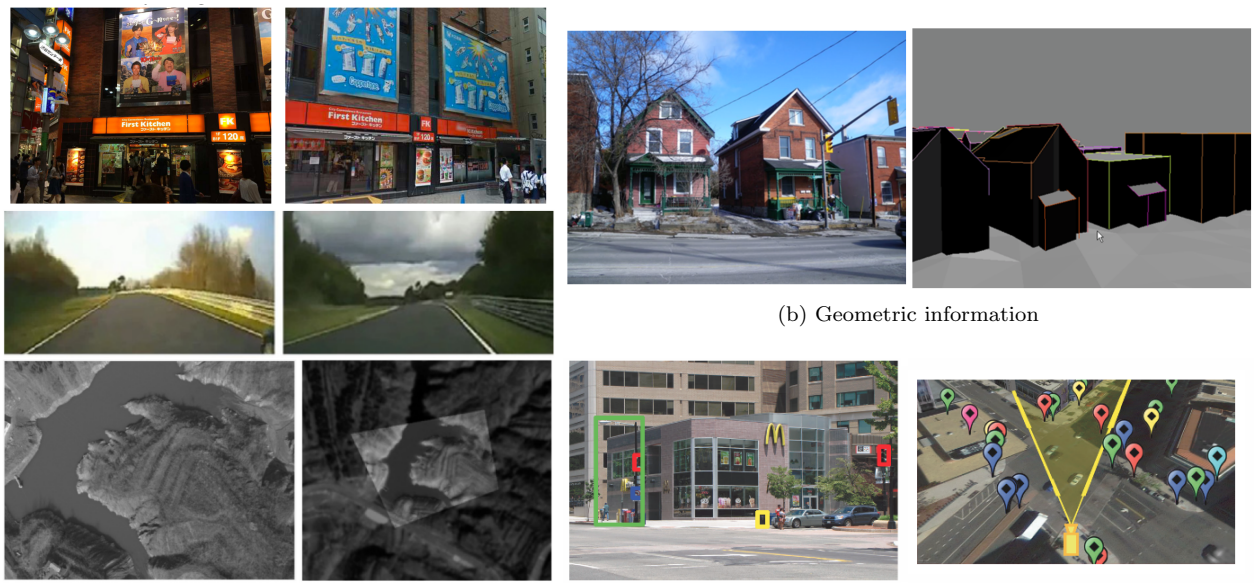
6. Data heterogeneity

Originally, images were the dedicated data to VBL system [165]. Still, conventional images for the task of localization have limitations, as mentioned in the previous section (see Section 5). The use of other type of data, such as geometric and semantic information, can circumvent these limitations. This section aims to exhaustively present the three different kinds of data used in VBL: optical, geometric and semantic.

6.1. Optical information

Robertson and Cipolla [165] present the first VBL method based on a geo-referenced database of images. Authors work aimed to localize buildings facades; therefore this database reflects the end-user application and was composed of images of front facing buildings. However, depending on the targeted application, the appearance of the database is likely to change. Figure 3a illustrates the diversity of imagery used for VBL.

Application. Most of the VBL systems are designed to retrieve the location of a camera within an urban environment, where a GPS signal cannot be properly received. Urban VBL systems are categorized in four classes: indoor VBL [103], VBL for pedestrians (or robots [54]) inside a city [165, 181, 35, 226], vehicles VBL on road traffic scenes for urban or suburban navigation [121, 127, 158, 23] and aerial vehicles localization system [212]. The appearance of the images present in the database changes drastically giving the purpose of the visual localization system.



(a) Optical information (b) Geometric information (c) Semantic information

Figure 3: **Illustration of the data heterogeneity in VBL:** (a) Image-based VBL systems, left represent the query image and right the closest corresponding image in the database. Images appearance differ depending on the final application, from top to bottom: street-level localization in urban environment from [199], car localization on road from [128], air-plane localization with aerial imagery from [212]. (b) Localization system built upon a DEM from [120]: left represent the query image and right the closest corresponding pose according to the 3D model. (c) Localization system with semantic information integration from [6]: on the left the query image with segmented objects (colour boxes) and on the right the retrieved pose (yellow camera) from a map with semantic objects location (coloured landmarks).

Coverage. By definition, the previously described applications do not cover the same area. For instance, a system design for car pose estimation should be able to localize a vehicle in a larger area than a pedestrian VBL system should do. Thus, there exists systems designed for world-wide localization estimation [65, 217, 208] as well as more spatially focused system [188]. Database coverage can be extended by using wide angle or omnidirectional cameras. Works presented in [1, 68, 93, 162, 201, 223, 224] use databases composed of spherical panoramas. On the other hand, aerial images are likely to cover large area but restrict VBL applications [212].

Database Consistency. Regardless of the type of images used, we distinguish between two kinds of databases: homogeneous and heterogeneous. Homogeneous databases are composed of images gathered through the same optical acquisition system in a restricted time interval. Google street view[1] or *IGN Stéréopolis* platform [143] give access to this kind of database. Homogeneous databases are suitable for applications that perform systematic processing on the data [115, 199]. Conversely, heterogeneous databases are made of images collected by various people with different cameras at inconsistent periods. These databases are often constructed through on-line collaborative platforms, like flickr[2], by downloading images associated to specific tags. Heterogeneous databases have the advantage of introducing visual variations in the VBL system and therefore improves robustness to appearance changes [160, 60] (see Section 5). Additionally, heterogeneous databases are easier to augment or update compared to homogeneous ones.

6.2. Geometric information

As presented in the outlines of this section, adding geometric information aims to overcome the limitations of only-based-optical VBL.

Weak Geometry. In [199, 35], authors introduce weak geometric clues that describe principal 3D planes present in the scene. This information is then used to modify existing images in the database: for rectification purpose [35] or to generate more images in order to cover a larger area [199]. Cham *et al.* [33] use a 2D

[1] https://www.google.fr/intl/usa/streetview/
[2] https://www.flickr.com/

buildings outline map for VBL. From a given image, authors extract buildings corner and match them according to the map. Along the same line, VBL method from [7] relies on a 2.5D map of Gratz (schematic buildings outlines boxes from OpenStreetMap[3]). 2D map is also used as geo-reference in the work from [22] (extended version in [23]) where authors produce, thanks to a stereo-camera, a path of a vehicle that is afterwards matched against the map. The matching process is embedded in a probabilistic framework to handle large environment. Saurer *et al.* [180] introduce the use of a Digital Elevation Model (DEM) to perform localization in mountainous terrain [162, 203, 37]. Bansal and Daniilidis [15] extend this idea in urban localization to perform purely geometric VBL with images as query input and DEM of a city as database. These purely geometric descriptions, also used in [120, 38, 162, 161] (see figure 3b), permit localization independently of the illumination conditions (compared to optically dependant methods, see §5.1).

3D geometry. Previous section §4.2 emphasizes the growing importance of colourized point clouds obtained by SfM in VBL. Because such geometric models are built upon images collections, reconstructed point clouds also lie on two categories: homogeneous [80, 78] and heterogeneous [67, 174] models (see §6.1). The addition of geometric relation by SfM improves retrieval performances [178] and permits precise pose estimation of the query, on the contrary of methods based on vanilla images collections.

However, SfM reconstruction is a costly operation. Rather than preprocessing the visual data to recover the geometry of the scene, appropriate sensors permit a direct capture of a 3D scene (*e.g.* stereo camera, depth camera, laser, lidar, etc.). Raw data from depth sensor are often used to add supplementary information channel to the VBL system. Works from [138, 121, 211] use disparity map from stereo camera. Several authors [183, 62, 58] used active depth camera that project infra-red pattern to estimate depth. Similar technology is used in [97] to perform VBL in complete obscurity. Consistent 3D models are also used to perform VBL task. For indoor localization, works from [183, 145] use textured model reconstructed from RGB-D sensor [183] or hand-crafted with dedicated software [145]. City-scale models are used by [10, 157, 146, 144, 29] to perform outdoor VBL.

Table 3: **Data heterogeneity in VBL.** Summary of the various types of data used in VBL. First row indicates the type of data present in the database and the first column the type of the query confronted to this database. References in **bold** denote works using semantic interpretation of the data (see §6.3). † Image-retrieval based methods that compare a input image to a set of images. Numerous references not displayed to readability reason. ‡ Several works consider VBL with as an input a set of images from the same scene rather than only one image input.

Database / **Query**	Images collection §6.1	Weak geometry §6.2	SfM §4.2	3D model §6.2
Image	Similarity search† **[5, 55, 132, 202, 138, 217]**	[180, 33, 35, 199] **[9, 6, 7, 30, 215]**	[8, 48, 66, 67, 49, 101, 102, 112, 113, 125, 174, 172, 176, 194, 225]	[10, 117, 170]
Several images‡	[217, 223]	[23]	[92]	**X**
Image + Depth	**X**	**[38]**	**X**	[31, 56, 58, 62, 183, 204, 186] **[50, 171]**
SfM	**X**	**X**	[125] **[112]**	**X**

6.3. Semantic information

Robustness and precision brought by geometric information has a significant cost in term on data acquisition, processing power and storage needs. Nevertheless, there is a good alternative and discriminant data representation: the semantic information. Semantic representation has received a growing interest in research community [106], in particular for robot navigation purpose [86]. Semantic-based methods are notably efficient to perform robust VBL to all kind of appearance changes. Indeed, by capturing a meaningful information about the scene composition, this approach is by definition invariant to local changes in appearance and in geometry. Semantic information used in VBL are classified between two classes: segmentation and categorization. Segmentation involves local methods that recognize within a data sub-parts with a semantic meaning (*e.g.* object detection in an image). On the other hand, categorization can be seen as global descriptors that associated semantic labels to a given data (*e.g.* scenes interpretation [47]).

[3]https://www.openstreetmap.org

Segmentation. The world "semantic" can be understood in different ways: Fernandez-Moral *et al.* [50] semantic approach consists of extracting planar surfaces in a scene, while in [171] authors segment a point cloud to extract objects with higher semantic meaning, like chair or table. Semantic approaches encode the data with a graph where nodes represent objects (plan in [50], furnitures in [171]) and edges spatial relations between objects. Graph representation offers a compressed data representation capable to handle local changes and minor measurement errors [189]. Semantic segmentation is used in [112] to narrow the search scope and in [6, 30, 38] to directly recover the pose of the query (illustration on figure 3c). Works described in [5, 132] consider the re-weighting of extracted local features in image according to the semantic class of the pixel obtained by image segmentation. Using this information, authors reduce the influence of local features that are not semantically robust for VBL, like vegetation or cars. In a same manner, Arth *et al.* [7] present concrete application of semantic segmentation of an image to reinforce hypothesis about building facades segmentation (the image is segmented with a SVM classifier). On the other hand, several methods rely on annotated map [9, 215] or Geographic Information System (GIS) [6, 30, 159] to guide the localization.

Categorization. Scene categorization [219] is a different manner to exploit semantic clues for VBL. High level semantic features have been popularized with the augmentation of labelled data and the accessibility of high computation power devices (GPU, Cloud Computing). ImageNet challenge introduced in 2009 by [47] permits the emergence of fast and robust classification methods, like the one described in [91]. Image classification produces a sub-sample of semantically identical images associated to a class. In VBL, classification can be used to decimate the database in order to proceed in a subsequent step to a more precise pose search. This method is successfully applied in [192], where the used CNN has a dual-purpose: narrowing the search scope by semantic labelling and producing a global descriptor by weight aggregation (see § 2.2). In [202, 138], classical learning methods like Gaussian Mixture Models (GMM) or epitome are employed for associating images to a finite number of possible locations. Recent work from [55] use semantic categorization in order to establish transition probabilities from a given type of environment to another one. Authors embed this framework in the SeqSLAM algorithm, improving the global system accuracy. Finally, works presented in [65, 217] consider the problem of classification of images at a world-wide level.

6.4. Cross-data localization

We have presented in this section three different kinds of information that can be used for VBL: optical, geometric and semantic. These types of data are commonly used together to improve localization. In this part, we consider the scenario where all types of data are not available at query time, for instance if the database uses more complete representation of the environment than the query input. It is a common scenario because some data are more difficult to acquire or required specific sensors (*e.g.* geometric information). In this case, methods have to deal with asymmetric representation of the environment in term of data type. We denote this problem cross-data VBL and classify the methods founded in the literature in two categories: methods using a common description regardless of the type of data and methods projecting one type of data within another data representation space.

Common description. Features to Points (F2P) VBL (see §4.2) oppose 2D images to 3D point cloud. In fact, all the features are exclusively extracted from images. On the other hand, semantic abstraction permit cross-data comparison by considering semantic object extracted from various types of data: images to 2D building outline map [33], images to map [6, 159, 30, 23], RGB-D data to DEM [38], etc. Referring to a similar physical entity, not necessary semantic, is also a manner to link information from various types of data. Images to DEM correspondences is performed in [15] based on a method relying on purely geometric clues extracted both in the image and the model. Recent work from [186, 98] use joint descriptors to merge RGB and depth data into a single feature.

Data projection. Another widely used method for combining data of different types consists of projecting one of the engaged data into the representation space of the other. For instance, lot of methods consider the challenging problem of registering photographies upon 3D models [180, 80, 7, 145, 146, 144]. Similarity comparison is performed thanks to synthetic images generated from the 3D models [170, 117, 10, 157]. Notice the synthesis of skyline profiles from DEM in the work of Saurer *et al.* [180]. Special attention is paid to placement of the artificial cameras that generate fake 2D views [67, 56, 199]: Aubry *et al.* [10] generate cameras over a regular grid to cover a maximal area. A pruning is then applied to only keep the most discriminant views.

To conclude this section, the different types of data used in VBL methods are summarized in table 3.

7. Discussion

This section aims to highlight common usage and emerging trends in VBL. As VBL panorama is wide and varied, we first propose a review of recent datasets and evaluation metrics used for comparing different approaches.

7.1. Datasets

Current datasets. Because of the important difference between direct and indirect methods, there exists two kinds of datasets used in VBL: unsorted list of images and spatially consistent datasets (that can be composed of point cloud or fine geo-referenced images). Table 4 presents numerous datasets used in VBL. Notice the growing number of publicly available datasets starring complete 3D scans of large cities [dataset][124, 114, 214]. As mentioned earlier, long-term localization in changing environment is an hot topic in robotic research. We observe therefore appearance of several datasets featuring multiple acquisitions of the same place over long periode of time [114, 28, 88, 90].

Table 4: **Currently used datasets in VBL.** Depending on the method to be evaluated (*e.g.* direct or indirect), different datasets are used. Data information (pose, type and homogeneity) concern the documents composing the database and not the one used at query time. Data homogeneity refer to the definition given in paragraph 6.1. RGB-D refer to data recorded with depth-cameras and RGB-S to information collected with standard cameras coupled with laser-scan depth estimation. † 6 DoF available in [177].

Name	Application domain	Data Pose Info.	Data Type	Data Homog.
INRIA Holidays [dataset][70]	Scene retrieval	No	RGB	No
Oxford Buildings [dataset][155]	Landmark retrieval	No	RGB	No
Paris [dataset][156]	Landmark retrieval	No	RGB	No
World Cities Dataset [dataset][196]	Image retrieval	GPS	RGB	No
Pittsburgh 250k [dataset][200]	Image retrieval	GPS	RGB	Yes
San Francisco Landmark [dataset][35]	Landmark retrieval	GPS†	RGB	Yes
Pittsburgh Street View [dataset][224]	Image retrieval	GPS + Compass	RGB	Yes
Tokyo 24-7 dataset [dataset][199]	Image retrieval	GPS + Compass	RGB	No
Nordland train dataset	Inter-season matching	GPS	RGB	No
Stromovka dataset [dataset][88]	Inter-season matching	Inter-season pair	RGB	No
CH1 dataset [dataset][180]	Localization in mountain	GPS	RGB	No
CH2 dataset [dataset][180]	Localization in mountain	GPS	RGB	Yes
GeoPose3K [dataset][20]	Localization in mountain	6 DoF Pose	RGB	No
Cambridge Dataset [dataset][80]	Camera localization	6 DoF Pose	SfM	Yes
Rome16K [dataset][101]	Camera localization	6 DoF Pose	SfM	No
Dubrovnik6K [dataset][101]	Camera localization	6 DoF Pose	SfM	No
Aachen [dataset][175]	Camera localization	6 DoF Pose	SfM	No
Notre Dame dataset [dataset][187]	Camera localization	6 DoF Pose	SfM	No
7 scenes [dataset][183]	Multi-purpose (indoor)	6 DoF Pose	RGB-D	Yes
Witham Wharf dataset [dataset][89]	Multi-purpose (indoor)	6 DoF Pose	RGB-D	No
North Campus dataset [dataset][28]	Multi-purpose	6 DoF Pose	RGB-S	No
Oxford Robotcar [dataset][114]	Multi-purpose	6 DoF Pose	RGB-S	No
TorontoCity dataset [dataset][214]	Multi-purpose	6 DoF Pose	RGB-S	Yes
KITTI dataset [dataset][124]	Multi-purpose	6 DoF Pose	RGB-S	Yes

Evaluation Metrics. Authors use various types of performances criteria in order to compare indirect methods. The recall @k, or recall @k%, is the most discriminative metric for VBL evaluation. It represents the percentage of queries that present a good match within the k or k% top ranked images. Usually k is set to 10 or 1%. Classical object-retrieval metric can be used, like RoC curves (precision against recall), mAP (the mean of average precision value) or the simple recall rate. If images in the database are augmented with GPS tag, authors often decide that a query is correctly localized if one among k retrieved candidates lies inside a tolerance radius (usually 10m, depending on the dataset). Result visualization is obtained by plotting a variable number k of candidates against the fraction of correctly localized images.

Concerning direct methods, authors often simply compute the mean of absolute position and orientation error relative to the available ground truth. Another criterion can be extracted from the inlier count obtained against a robust geometric verification. A query is considered as successfully matched if enough inliers are found after the application of an iterative RANSAC-like algorithm. However, such a metric does not ensure that the data is well localized according to the model [172].

7.2. Runtime consideration

Real-time performances and embedded architectures are constraints mainly present in the robotic community. In VBL, such criteria are not always taken into account. This can be explained by the fact that recovering the localization of an input query is a one-shoot action; i.e. it has to be performed only once compared to tracking systems [116] or SLAM algorithms [54]. Furthermore, more and more embedded systems rely on deported architecture or cloud computer, resolving the problem of low computational power of portable devices [125]. Yet, some authors manage to reduce the computational cost of their system [183, 58, 113]: for instance Feng *et al.* [49] introduce a light version of F2P method and works from [217, 80, 42] embed their localization system in a compact CNN architecture loadable on a smart-phone.

7.3. Trends in VBL

A quantitative comparison between all VBL systems is impossible due to the diversity in both methods and applications. Nevertheless, we refer reader to recent papers that quantitatively compare specific types of state-of-the-art methods. Concerning indirect methods, following recent contributions [160, 60] show comprehensive comparisons. Direct F2P methods based on SfM are carefully compared on three papers [49, 176, 194] and Kendall and Cipolla [79] propose a detailed comparison between SfM-direct methods (§4.2) and CNN-direct methods (§4.3). In [21], authors propose an overview of both indirect and direct VBL methods by reporting results of various works in a common table. Finally, Sattler *et al.* [177] present the first comparison between indirect and direct methods based on images collection for the task of query accurate localization. In the following, we propose our qualitative analysis of VBL panorama.

From indirect to direct methods. As discussed earlier, there is a trade-off between the area covered and the precision reached by the VBL system; the survey of Brejcha and Čadík [21] provides a complete overview of this problem. Indirect methods prioritize the space coverage, city scale [60] to word-wide [217], whereas direct methods focus on precision and exact 6 DoF estimation [49]. This survey describes VBL methods in a chronological order, that is why we present indirect methods before direct ones. However, during the last decade, research focus seems to have turned to direct methods. This can be explained by the drastic increase of applications using precise VBL for both professional (*e.g.* robotics [115]) and individual (*e.g.* augmented reality [7]) purposes.

The growing importance of geometric data. Limitations of methods using only images have been discussed in Section 5, and the growing accessibility of geometric data promotes the development of systems based on depth information [143]. Furthermore, geometric data facilitates the final pose estimation, which also explains the rapid development of direct methods [92, 148, 149]. When available, depth information can directly improve the result of VBL methods [138, 56, 183, 199, 31]. Point clouds remain the favourite type of geometric data in VBL [176], nevertheless less complete but more compact models have shown promising result for some applications [162, 15, 38, 199].

Emergence of semantic localization. Less used in VBL, semantic data offer promising results [6, 30, 38]. In addition to being generic regarding the original "raw" scene representation, semantic abstraction permits a discriminative and robust description of the scene. Moreover, the use of semantic graph representations can handle huge amount of data. Loss in area coverage granted by the use of direct methods and complex data can be balanced by semantic information. Indeed, in [6, 112] authors initially narrow the research scope with extracted semantic clues.

Data combination and cascade schemes. Getting both wide area coverage and high precision of the query pose is the current challenge of VBL. Cascade scheme, that can be seen as a combination of indirect and direct methods, are certainly a good alternative to achieve this objective [177]. Indeed, firstly reducing the amount of data and in a second step recovering the exact pose of the query is a well studied topic [168, 11, 188, 123, 177]. This architecture facilitates the use of heterogeneous data [112] in a common framework. Combination of various types of data benefits to the task of VBL by exploiting all the available sources of interest present at a given location [99].

7.4. Benefit of heterogeneous data

All along this survey we emphasize the growing importance of multiple types of data (optical, geometric and semantic information) for the task of VBL. As discussed in section 6, using more sophisticate data aims to overcome shortcoming (presented in section 5) of only-optical based systems. Geometric and depth information permit a total abstraction to the pixel intensity, naturally providing to the system a robustness to illumination variability that can occur in the scene. Local changes in scene appearance and geometry can also be handle with the use of a semantic representation. Optical data, though, contain extremely specific information and are much more easier to collect compared to semantic and geometric data. That is why these three aforementioned data should be considered as complementary information. Based on these observations, there is a real benefit of using heterogeneous data to achieve better results in VBL.

7.5. VBL and machine learning

VBL benefits from the recent progress in machine learning. Indirect methods are now dominated by CNN approaches [160, 60] (see §2.2). Image description obtained with convolutional networks gathers all the characteristics needed for the task of scene retrieval. New network architecture has been proposed to resolve the direct VBL problem (presented in §4.3). Despite the simplicity and the robustness of these methods, state-of-the-art direct localization results are still obtained through point to feature approaches [210]. However, active researches are pursuing in this promising direction [107, 75, 79]. The last VBL sub-domain improved by CNN is the semantic scene segmentation and categorization used in some localization methods [171, 6, 7, 38]. Conventional methods, like Deformable Parts Model (DPM) [6], SVM [7] or Automatic Labelling Environment (ALE) [38], should be quickly replaced by CNN [193, 227].

8. Conclusion

This survey is an attempt to describe the large panorama of VBL. Two families of methods have been comprehensively reviewed to highlight the current capabilities of existing localization systems and to address the remaining challenges in the domain. The principal concern about the future of VBL can be summarized in this question: "How should we combine precision and large area coverage in a single system?". This essential issue may be solved thanks to the massive emergence of novel data. Complete 3D models of large cities are more and more made publicly available, especially with the democratization of self-driving vehicles. The adaptation to larger area of methods based on exhaustive geometric data, until now restricted to indoor applications, is therefore a promising avenue of research. Semantic interpretation, a well studied topic, can also support VBL scale up while improving the robustness of the localization.

Finally, respective advances both in Visual-Based Localization and Visual Place Recognition can be beneficial to each other. In particular, graph-based methods widely used in Place Recognition are under-represented in VBL, though reaching excellent results. However, the gap between those two research domains progressively reduces, considering the increasing number of co-works between the two original communities (Computer Vision for VBL and Robotics for Visual Place Recognition).

Acknowledgements

We would like to acknowledge the French ANR project pLaTINUM (ANR-15-CE23-0010) for its financial support.

References

[1] Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2017. NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) XX (X), 5297–5307. 5, 6, 7, 8, 14, 15, 16

[2] Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (April), 2911–2918. 2, 3, 4

[3] Arandjelović, R., Zisserman, A., 2013. All About VLAD. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1578–1585. 3

[4] Arandjelović, R., Zisserman, A., 2014. DisLocation : Scalable descriptor. In: Proceedings of the Asian Conference on Computer Vision (ACCV). 3, 6

[5] Arandjelović, R., Zisserman, A., 2014. Visual vocabulary with a semantic twist. In: Proceedings of the Asian Conference on Computer Vision (ACCV). Vol. 9003. pp. 178–195. 17, 18

[6] Ardeshir, S., Zamir, A. R., Torroella, A., Shah, M., 2014. GIS-assisted object detection and geospatial localization. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 8694 LNCS. pp. 602–617. 3, 9, 16, 17, 18, 20, 21

[7] Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., 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, 10, 14, 17, 18, 20, 21

[8] Arth, C., Wagner, D., Klopschitz, M., Irschara, A., Schmalstieg, D., 2009. Wide area localization on mobile phones. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 73–82. 11, 17

[9] Atanasov, N., Zhu, M., Daniilidis, K., Pappas, G. J., 2016. Localization from semantic observations via the matrix permanent. The International Journal of Robotics Research (IJRR) 35 (1-3), 73–99. 17, 18

[10] Aubry, M., Russell, B. C., Sivic, J., 2014. Painting-to-3D model alignment via discriminative visual elements. ACM Transactions on Graphics (ToG) 33 (2), 1–14. 3, 5, 8, 13, 14, 15, 17, 18

[11] Azzi, C., Asmar, D., Fakih, A., Zelek, J., 2016. Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization. In: British Machine Vision Conference (BMVC). No. 2. pp. 1–12. 1, 3, 4, 5, 20

[12] Babenko, A., Lempitsky, V., 2015. Aggregating local deep features for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 1269–1277. 7

[13] Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V., 2014. Neural Codes for Image Retrieval. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). pp. 584–599. 5, 7, 8

[14] Bae, S., Agarwala, A., Durand, F., 2010. Computational rephotography. ACM Transactions on Graphics (ToG) 29 (3), 1–15. 15

[15] Bansal, M., Daniilidis, K., 2014. Geometric Urban Geo-Localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3, 5, 17, 18, 20

[16] Bansal, M., Daniilidis, K., Sawhney, H. S., 2012. Ultra-wide baseline facade matching for geo-localization. Proceedings of the IEEE European Conference on Computer Vision (ECCV) 7583 LNCS (PART 1), 175–186. 15

[17] Bansal, M., Sawhney, H. S., Cheng, H., Daniilidis, K., 2011. Geo-localization of street views with aerial image databases. Proceedings of the ACM International Conference on Multimedia (MM), 1125. 15

[18] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding (CVIU) 110 (3), 346–359. 3, 4

[19] Bhowmik, N., Weng, L., Gouet-Brunet, V., Soheilian, B., 2017. Cross-domain Image Localization by Adaptive Feature Fusion. In: Joint Urban Remote Sensing Event (JURSE). 5, 15

[20] Brejcha, J., Cadik, M., 2017. GeoPose3K: Mountain Landscape Dataset for Camera Pose Estimation in Outdoor Environments. Image and Vision Computing. 19

[21] Brejcha, J., Čadík, M., 2017. State-of-the-art in visual geo-localization. Pattern Analysis and Applications. 1, 2, 20

[22] Brubaker, M. A., Geiger, A., Urtasun, R., 2013. Lost! leveraging the crowd for probabilistic visual self-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3057–3064. 17

[23] Brubaker, M. A., Geiger, A., Urtasun, R., 2016. Map-based probabilistic visual self-localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38 (4), 652–665. 10, 15, 17, 18

[24] Calonder, M., Lepetit, V., Strecha, C., Fua, P., 2010. BRIEF: Binary robust independent elementary features. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 6314 LNCS. pp. 778–792. 3, 14

[25] Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (6), 679–698. 3

[26] Cao, S., Snavely, N., 2015. Graph-Based Discriminative Learning for Location Recognition. International Journal of Computer Vision (IJCV) 112 (2), 239–254. 5, 8, 9

[27] Carlevaris-Bianco, N., Eustice, R. M., 2014. Learning visual feature descriptors for dynamic lighting conditions. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 2769–2776. 3, 4, 14

[28] Carlevaris-Bianco, N., Ushani, A. K., Eustice, R. M., 2016. University of Michigan North Campus long-term vision and lidar dataset. The International Journal of Robotics Research (IJRR) 35 (9), 1023–1035. 19

[29] Caselitz, T., Steder, B., Ruhnke, M., Burgard, W., 2016. Matching Geometry for Long-term Monocular Camera Localization. Proceedings of the IEEE International Conference of Robotics and Automation Workshop (ICRAW). 17

[30] Castaldo, F., Zamir, A. R., Angst, R., Palmieri, F., Savarese, S., 2015. Semantic Cross-View Matching. Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW) 2016-Febru, 1044–1052. 15, 17, 18, 20

[31] Cavallari, T., Golodetz, S., Lord, N. A., Valentin, J., Di Stefano, L., Torr, P. H. S., 2017. On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 12, 17, 20

[32] Celik, C., Bilge, H. S., 2017. Content based image retrieval with sparse representations and local feature descriptors: A comparative study. Pattern Recognition 68, 1–13. 6

[33] Cham, T. J., Ciptadi, A., Tan, W. C., Pham, M. T., 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). pp. 366–373. 3, 5, 14, 16, 17, 18

[34] Chan, J., Lee, J. A., Kemao, Q., 2016. F-SORT : An Alternative for Faster Geometric Verification. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 1–15. 9

[35] Chen, D. M., Baatz, G., Köser, K., Tsai, S. S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R., 2011. City-scale landmark identification on mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 737–744. 10, 14, 15, 16, 17, 19

[36] Chen, Y., Li, X., Dick, A., Hill, R., 2013. Ranking consistency for image matching and object retrieval. Pattern Recognition 47 (3), 1349–1360. 9

[37] Chen, Y., Qian, G., Gunda, K., Gupta, H., Shafique, K., 2015. Camera geolocation from mountain images. In: Proceedings of the International Conference on Information Fusion (Fusion). pp. 1587–1596. 3, 17

[38] Christie, G., Warnell, G., Kochersberger, K., 2016. Semantics for UGV Registration in GPS-denied Environments. arXiv preprint. 10, 17, 18, 20, 21

[39] Chum, O., Matas, J., 2005. Matching with PROSAC-progressive sample consensus. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. IEEE, pp. 220–226. 9

[40] Chum, O., Mikul, A., Perdoch, M., Matas, J., 2011. Total Recall II : Query Expansion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 9

[41] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A., 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 9

[42] Contreras, L., Mayol-Cuevas, W., 2017. Towards CNN Map Compression for camera relocalisation. arXiv preprint. 12, 20

[43] Corke, P., Paul, R., Churchill, W., Newman, P., 2013. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 2085–2092. 3, 4, 9, 13, 14

[44] Cummins, M., Newman, P., 2008. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. The International Journal of Robotics Research (IJRR) 27, 647–665. 3, 4, 14

[45] Cummins, M., Newman, P., 2010. Accelerating FAB-MAP with concentration inequalities. IEEE Transactions on Robotics (ToR) 26 (6), 1042–1050. 14

[46] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. IEEE, pp. 886–893. 3, 5

[47] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 248–255. 17, 18

[48] Donoser, M., Schmalstieg, D., 2014. Discriminative feature-to-point matching in image-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 516–523. 9, 11, 12, 17

[49] Feng, Y., Fan, L., 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, 3, 4, 11, 12, 17, 20

[50] Fernandez-Moral, E., Mayol-Cuevas, W., Arevalo, V., Gonzalez-Jimenez, J., 2013. Fast place recognition with plane-based maps. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 2719–2724. 3, 4, 17, 18

[51] Finlayson, G. D., Hordley, S. D., Lu, C., Drew, M. S., 2006. On the removal of shadows from images. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28 (1), 59–68. 14

[52] Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24 (6), 381–395. 9

[53] Förstner, W., Wrobel, B. P., 2016. Photogrammetric Computer Vision. Springer. 10, 11, 14

[54] Garcia-Fidalgo, E., Ortiz, A., 2015. Vision-based topological mapping and localization methods: A survey. Robotics and Autonomous Systems (RAS) 64, 1–20. 2, 3, 4, 13, 15, 20

[55] Garg, S., Jacobson, A., Kumar, S., Milford, M. J., 2017. Improving Condition and Environment-Invariant Place Recognition with Semantic Place Categorization. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). 17, 18

[56] Gee, A. P., Mayol-Cuevas, W., 2012. 6D Relocalisation for RGBD Cameras Using Synthetic View Regression. In: British Machine Vision Conference (BMVC). pp. 1–11. 3, 4, 17, 18, 20

[57] Gionis, A., Indyk, P., Motwani, R., 1999. Similarity Search in High Dimensions via Hashing. In: Proceedings of the 25th VLDB Conference. pp. 518–529. 8

[58] Glocker, B., Shotton, J., Criminisi, A., 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. 12, 17, 20

[59] Gong, Y., Wang, L., Guo, R., Lazebnik, S., 2014. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). pp. 1–17. 8

[60] Gordo, A., Almazán, J., Revaud, J., Larlus, D., 2017. End-to-End Learning of Deep Visual Representations for Image Retrieval. International Journal of Computer Vision (IJCV) 124 (2), 237–254. 3, 5, 7, 8, 16, 20, 21

[61] Griffith, S., Pradalier, C., 2017. Survey Registration For LongTerm Natural Environment Monitoring. Journal of Field Robotics. 3, 14

[62] Guzman-Rivera, A., Pushmeet, K., Glocker, B., Shotton, J., Sharp, T., Fitzgibbon, A., Izadi, S., 2014. Multi-Output Learning for Camera Relocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–6. 12, 17

[63] Hartley, R., 1997. In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 19 (6), 580–593. 10

[64] Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press. 10, 11, 12, 14

[65] Hays, J., 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). Vol. 05. 3, 4, 5, 13, 16, 18

[66] Heisterklaus, I., Qian, N., Miller, A., 2014. Image-based pose estimation using a compact 3D model. In: IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin). pp. 327–330. 1, 11, 12, 17

[67] Irschara, A., Zach, C., Frahm, J.-m., 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). 11, 12, 14, 17, 18

[68] Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O., 2017. Panorama to panorama matching for location recognition. In: Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR). 7, 16

[69] Jégou, H., Chum, O., 2012. Negative evidences and co-occurrences in image retrieval : the benefit of PCA and whitening. In: Proceedings of the IEEE European conference on computer vision (ECCV). 8

[70] Jégou, H., Douze, M., Schmid, C., 2008. Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). No. October. pp. 304–317. 4, 6, 19

[71] Jégou, H., Douze, M., Schmid, C., 2009. On the burstiness of visual elements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1169–1176. 3, 6, 9

[72] Jégou, H., Douze, M., Schmid, C., 2011. Product Quantization for Nearest Neighbor Search Herve. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33 (1), 117–128. 8

[73] Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Schmid, C., Pérez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 1–12. 6

[74] Jégou, H., Zisserman, A., 2014. Triangulation embedding and democratic aggregation for image search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6

[75] Jia, D., Su, Y., Li, C., 2016. Deep Convolutional Neural Network for 6-DOF Image Localization. arXiv preprint (413113), 1790–1798. 5, 12, 21

[76] Johnson, J., Douze, M., Jégou, H., 2017. Billion-scale similarity search with GPUs. arXiv preprint. 8

[77] Karakasis, E. G., Amanatiadis, A., Gasteratos, A., Chatzichristofis, S. A., 2015. Image moment invariants as local features for content based image retrieval using the Bag-of-Visual-Words model. Pattern Recognition Letters 55, 22–27. 13

[78] Kendall, A., Cipolla, R., 2016. Modelling Uncertainty in Deep Learning for Camera Relocalization. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 12, 17

[79] Kendall, A., 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). 12, 20, 21

[80] Kendall, A., Grimes, M., 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). 2, 10, 12, 17, 18, 19, 20

[81] Kim, B., Yoo, H., Sohn, K., 2013. Exact order based feature descriptor for illumination robust image matching. Pattern Recognition 46 (12), 3268–3278. 14

[82] Kim, H. J., Dunn, E., Frahm, J.-M., 2015. Predicting good features for image geo-localization using per-bundle VLAD. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 1170–1178. 3, 5, 6, 8, 14

[83] Kim, H. J., Dunn, E., Frahm, J.-M., 2017. Learned Contextual Feature Reweighting for Image Geo-Localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 5, 7, 8

[84] Kneip, L., Furgale, P., 2014. OpenGV: A unified and generalized approach to real-time calibrated geometric vision. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1–8. 10

[85] Kneip, L., Scaramuzza, D., Siegwart, R., 2011. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2969–2976. 12

[86] Kostavelis, I., Gasteratos, A., 2015. Semantic mapping for mobile robotics tasks: A survey. Robotics and Autonomous Systems (RAS) 66, 86–103. 3, 17

[87] Krajník, T., Cristoforis, P., Kusumam, K., Neubert, P., Duckett, T., 2017. Image features for visual teach-and-repeat navigation in changing environments. Robotics and Autonomous Systems (RAS) 88 (November), 127–141. 3, 4, 14

[88] Krajník, T., Faigl, J., Vonásek, V., Košnar, K., Kulich, M., Preucil, L., 2010. Simple yet stable bearing-only navigation. Journal of Field Robotics 27 (5), 511–533. 19

[89] Krajník, T., Fentanes, J. P., Mozos, O. M., Duckett, T., Ekekrantz, J., Hanheide, M., 2014. Long-Term Topological Localisation for Service Robots in Dynamic Environments using Spectral Maps. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). No. Iros. pp. 4537–4542. 19

[90] Krajník, T., Fentanes, J. P., Santos, J. M., Duckett, T., 2017. FreMEn: Frequency Map Enhancement for Long-Term Mobile Robot Autonomy in Changing Environments. IEEE Transactions on Robotics (ToR), 1–14. 3, 14, 19

[91] Krizhevsky, A., Sutskever, I., Hinton, G. E., 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 1097–1105. 3, 7, 18

[92] Kroeger, T., Van Gool, L., 2014. Video registration to SfM models. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 8693 LNCS. pp. 1–16. 17, 20

[93] Kumar, D., 2016. Deep Learning Based Place Recognition for Challenging Environments. arXiv preprint. 7, 15, 16

[94] Leutenegger, S., Chli, M., Siegwart, R., 2011. BRISK: Binary robust invariant scalable keypoints. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 2548–2555. 3, 4

[95] Lezama, J., Von Gioi, R., Randall, G., Morel, J.-M., 2014. Finding vanishing points via point alignments in image primal and dual domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 509–515. 14

[96] Li, K., Zou, C., Bu, S., Liang, Y., Zhang, J., Gong, M., 2017. Multi-modal Feature Fusion for Geographic Image Annotation. Pattern Recognition (Ke Li). 5

[97] Li, R., Liu, Q., Gui, J., Gu, D., Hu, H., 2016. Night-time indoor relocalization using depth image with Convolutional Neural Networks. Proceedings of the IEEE International Conference on Automation and Computing (ICAC), 261–266. 17

[98] Li, R., Liu, Q., Gui, J., Gu, D., Hu, H., 2017. Indoor Relocalization in Challenging Environments With Dual-Stream Convolutional Neural Networks. IEEE Transactions on Automation Science and Engineering, 1–12. 18

[99] Li, S., Calway, A., 2016. Absolute pose estimation using multiple forms of correspondences from RGB-D frames. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Vol. 2016-June. pp. 4756–4761. 3, 4, 20

[100] Li, X., Larson, M., Hanjalic, A., 2015. Pairwise geometric matching for large-scale object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 07-12-June. pp. 5153–5161. 3, 4

[101] Li, Y., Snavely, N., Huttenlocher, D. P., 2010. Location Recognition using Prioritized Feature Matching. Proceedings of the IEEE European Conference on Computer Vision (ECCV), 791–804. 11, 12, 17, 19

[102] Li, Y., Snavely, N., Huttenlocher, D. P., Fua, P., 2012. Worldwide Pose Estimation Using 3D Point Clouds. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). pp. 15–29. 11, 17

[103] Liang, J. Z., Corso, N., Turner, E., Zakhor, A., 2013. Image Based Localization in Indoor Environments. In: Computing for Geospatial Research and Application. 3, 15

[104] Lin, T.-Y., Belongie, S., Hays, J., 2013. Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 891–898. 15

[105] Lin, T.-Y., Cui, Y., Belongie, S., Hays, J., 2015. Learning Deep Representations for Ground-to-Aerial Geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. JUNE. pp. 5007–5015. 13, 15

[106] Liu, Q., Li, R., Hu, H., Gu, D., 2016. Extracting Semantic Information from Visual Data: A Survey. Robotics 5 (1), 8. 17

[107] Liu, Z., Duan, L.-Y., Chen, J., Huang, T., 2016. Depth-Based Local Feature Selection for Mobile Visual Search. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). 12, 21

[108] Liu, Z., Marlet, R., 2012. Virtual line descriptor and semi-local matching method for reliable feature correspondence. In: British Machine Vision Conference (BMVC). pp. 11–16. 3, 4

[109] Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60 (2), 91–110. 3, 4, 11

[110] Lowry, S., Milford, M. J., 2016. Supervised and Unsupervised Linear Learning Techniques for Visual Place Recognition in Changing Environments. IEEE Transactions on Robotics (ToR) 32 (3), 600–613. 14

[111] Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., Milford, M. J., 2016. Visual Place Recognition: A Survey. IEEE Transactions on Robotics (ToR) 32 (1), 1–19. 2, 3, 13, 14

[112] Lu, G., Yan, Y., Ren, L., Song, J., Sebe, N., Kambhamettu, C., 2015. Localize Me Anywhere, Anytime: A Multi-task Point-Retrieval Approach. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2434–2442. 3, 8, 11, 17, 18, 20

[113] Lynen, S., Sattler, T., Bosse, M., Hesch, J. A., Pollefeys, M., Siegwart, R., 2015. Get out of my lab: Large-scale, real-time visual-intertial localization. Robotics Science and Systems (RSS), 37–46. 11, 17, 20

[114] Maddern, W., Pascoe, G., Linegar, C., Newman, P., 2016. 1 year, 1000 km: The Oxford RobotCar dataset. The International Journal of Robotics Research (IJRR), 0278364916679498. 19

[115] Majdik, A. L., Albers-Schoenberg, Y., Scaramuzza, D., 2013. MAV urban localization from Google street view data. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 3979–3986. 3, 4, 13, 14, 16, 20

[116] Marchand, E., Uchiyama, H., Spindler, F., 2016. Pose Estimation for Augmented Reality : a Hands-On Survey. IEEE Transactions on Visualization and Computer Graphics (ToVCG) 22 (12), 2633–2651. 10, 20

[117] Mason, J., Ricco, S., Parr, R., 2011. Textured Occupancy Grids for Monocular Localization Without Features. In: Proceedings of the IEEE International Conference of Robotics and Automation (ICRA). pp. 5800–5806. 10, 17, 18

[118] Massiceti, D., Krull, A., Brachmann, E., Rother, C., Torr, P. H. S., 2017. Random Forests versus Neural Networks - What's Best for Camera Relocalization? In: Proceedings of the IEEE International Conference of Robotics and Automation (ICRA). 12

[119] Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22 (10), 761–767. 3, 5

[120] Matei, B. C., Vander Valk, N., Zhu, Z., Cheng, H., Sawhney, H. S., 2013. Image to LIDAR matching for geotagging in urban environments. In: Proceedings of IEEE Workshop on Applications of Computer Vision (WACV). pp. 413–420. 16, 17

[121] McManus, C., Upcroft, B., Newman, P., 2014. Scene Signatures : Localised and Point-less Features for Localisation. In: Robotics Science and Systems (RSS). 3, 5, 8, 14, 15, 17

[122] Melekhov, I., Kannala, J., Rahtu, E., 2017. Relative Camera Pose Estimation Using Convolutional Neural Networks. arXiv preprint, 1–12. 10

[123] Meng, L., Chen, J., Tung, F., Little, J. J., de Silva, C. W., 2016. Exploiting Random RGB and Sparse Features for Camera Pose Estimation. In: British Machine Vision Conference (BMVC). pp. 1–12. 12, 20

[124] Menze, M., Geiger, A., 2015. Object Scene Flow for Autonomous Vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 19

[125] Middelberg, S., Sattler, T., Untzelmann, O., 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, 11, 12, 17, 20

[126] Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. International Journal of Computer Vision (IJCV) 60 (1), 63–86. 3, 4

[127] Milford, M. J., Lowry, S., Shirazi, S., Pepperell, E., Shen, C., Lin, G., Liu, F., Cadena, C., 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). pp. 18–25. 9, 13, 14, 15

[128] Milford, M. J., Wyeth, G. F., 2012. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 1643–1649. 10, 14, 16

[129] Morago, B., Bui, G., Duan, Y., 2015. An Ensemble Approach to Image Matching Using Contextual Features. IEEE Transactions on Image Processing (ToIP) 24 (11), 4474–4487. 3

[130] Morago, B., Bui, G., Duan, Y., 2016. 2D Matching Using Repetitive and Salient Features in Architectural Images. IEEE Transactions on Image Processing (ToIP) 7149 (c), 1–12. 3, 4, 5, 6, 11, 14

[131] Morel, J.-M., Yu, G., 2009. ASIFT: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences 2 (2), 438–469. 13, 14

[132] Mousavian, A., Kosecká, J., Lien, J. M., 2015. Semantically guided location recognition for outdoors scenes. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Vol. 2015-June. pp. 4882–4889. 7, 15, 17, 18

[133] Mühlfellner, P., Bürki, M., Bosse, M., Derendarz, W., Philippsen, R., Furgale, P., 2015. Summary Maps for Lifelong Visual Localization. Journal of Field Robotics 23 (0), 245–267. 3, 14

[134] Muja, M., Lowe, D. G., 2009. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In: Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP). pp. 1–10. 8

[135] Murillo, A. C., Singh, G., Kosecká, J., Guerrero, J. J., 2013. Localization in urban environments using a panoramic gist descriptor. IEEE Transactions on Robotics (ToR) 29 (1), 146–160. 3, 14

[136] Nelson, P., Churchill, W., Posner, I., Newman, P., 2015. From Dusk till Dawn: Localisation at Night using Artificial Light Sources. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 5245–5252. 14

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

[138] Ni, K., Kannan, A., Criminisi, A., Winn, J., 2009. Epitomic location recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31 (12), 2158–2167. 3, 4, 8, 17, 18, 20

[139] Nistér, D., 2004. An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 26 (6), 756–770. 10

[140] Nistér, D., Stewénius, H., 2006. Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 2. pp. 2161–2168. 3, 5, 8

[141] Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision (IJCV) 42 (3), 145–175. 3, 4

[142] Panphattarasap, P., Calway, A., 2016. Visual place recognition using landmark distribution descriptors. In: Proceedings of the Asian Conference on Computer Vision (ACCV). 3, 7, 8

[143] Paparoditis, N., Papelard, J.-P., Cannelle, B., Devaux, A., Soheilian, B., David, N., 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. 16, 20

[144] Pascoe, G., Maddern, W., Newman, P., 2015. Direct Visual Localisation and Calibration for Road Vehicles in Changing City Environments. Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2016-Febru, 98–105. 17, 18

[145] Pascoe, G., Maddern, W., Newman, P., 2015. Robust Direct Visual Localisation using Normalised Information Distance. British Machine Vision Conference (BMVC), 1–13. 17, 18

[146] Pascoe, G., Maddern, W., Stewart, A. D., Newman, P., 2015. FARLAP : Fast Robust Localisation using Appearance Priors. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 6366–6373. 10, 17, 18

[147] Passalis, N., Tefas, A., 2017. Neural Bag-of-Features learning. Pattern Recognition 64 (August 2016), 277–294. 6

[148] Paudel, D. P., Habed, A., Demonceaux, C., Vasseur, P., 2015. LMI-based 2D-3D registration: From uncalibrated images to Euclidean scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4494–4502. 11, 20

[149] Paudel, D. P., Habed, A., Demonceaux, C., Vasseur, P., 2015. 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). pp. 2048–2056. 11, 20

[150] Paul, R., Newman, P., 2010. FAB-MAP 3D: Topological mapping with spatial and visual appearance. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 2649–2656. 14

[151] Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronnin, F., Schmid, C., 2015. Local convolutional features with unsupervised training for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 91–99. 3, 4

[152] Pepperell, E., Corke, P., Milford, M. J., 2014. All - Environment Visual Place Recognition with SMART. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 14

[153] Pepperell, E., Corke, P., Milford, M. J., 2016. Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. The International Journal of Robotics Research (IJRR) 35 (9), 1057–1179. 14

[154] Perronnin, F., Liu, Y., 2010. Large-Scale Image Retrieval with Compressed Fisher Vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6

[155] Philbin, J., Chum, O., Isard, M., Sivic, J., 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). 8, 9, 19

[156] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6, 19

[157] Poglitsch, C., Arth, C., Schmalstieg, D., 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). pp. 132–135. 9, 10, 17, 18

[158] Qu, X., Soheilian, B., Habets, E., 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. 3, 4, 6, 9, 10, 11, 12, 15

[159] Qu, X., Soheilian, B., 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. 3, 9, 18

[160] Radenović, F., Tolias, G., 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). Vol. 9905. pp. 3–20. 2, 5, 7, 8, 16, 20, 21

[161] Ramalingam, S., Bouaziz, S., Sturm, P., 2011. Pose Estimation Using Both Points and Lines for Geolocation. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 3, 4, 17

[162] Ramalingam, S., Bouaziz, S., Sturm, P., Brand, M., 2010. SKYLINE2GPS: Localization in urban canyons using omni-skylines. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 3816–3823. 3, 16, 17, 20

[163] Razavian, A. S., Sullivan, J., Carlsson, S., Maki, A., 2014. Visual Instance Retrieval with Deep Convolutional Networks. arXiv preprint 4 (3), 251–258. 7

[164] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Annual Conference on Neural Information Processing Systems (NIPS). pp. 91–99. 3, 5

[165] Robertson, D., Cipolla, R., 2004. An Image-Based System for Urban Navigation. In: British Machine Vision Conference (BMVC). 14, 15

[166] Rocco, I., Arandjelović, R., Sivic, J., 2017. Convolutional neural network architecture for geometric matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 10

[167] Rosen, D. M., Mason, J., Leonard, J. J., 2016. Towards Lifelong Feature-Based Mapping in Semi-Static Environments. In: Proceedings of the IEEE International Conference of Robotics and Automation (ICRA). pp. 1–8. 14

[168] Rubio, A., Villamizar, M., Ferraz, L., Penata-Sanchez, A., Ramisa, A., Simo-Serra, E., Sanfeliu, A., Moreno-Noguer, F., 2015. Efficient Monocular Pose Estimation for Complex 3D Models. In: Proceedings of the IEEE International Conference of Robotics and Automation (ICRA). Vol. 2. pp. 1397–1402. 10, 20

[169] Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 3

[170] Russell, B. C., Sivic, J., Ponce, J., Dessales, H., 2011. Automatic alignment of paintings and photographs depicting a 3D scene. In: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW). 3, 4, 10, 11, 15, 17, 18

[171] Salas-Moreno, R. F., Newcombe, R. A., Strasdat, H., Kelly, P. H. J., 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). pp. 1352–1359. 3, 9, 17, 18, 21

[172] Sattler, T., Havlena, M., Radenović, F., Schindler, K., Pollefeys, M., 2015. Hyperpoints and fine vocabularies for large-scale location recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 2102–2106. 1, 11, 17, 19

[173] Sattler, T., Havlena, M., Schindler, K., Pollefeys, M., 2016. Large-Scale Location Recognition and the Geometric Burstiness Problem. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3, 9

[174] Sattler, T., Leibe, B., Kobbelt, L., 2011. Fast image-based localization using direct 2D-to-3D matching. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 667–674. 2, 11, 17

[175] Sattler, T., Leibe, B., Kobbelt, L., 2012. Improving image-based localization by active correspondence search. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 7572 LNCS. pp. 752–765. 19

[176] Sattler, T., Leibe, B., Kobbelt, L., 2016. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) X (1). 2, 11, 12, 17, 20

[177] Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., Pajdla, T., 2017. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 10, 19, 20

[178] Sattler, T., Weyand, T., Leibe, B., Kobbelt, L., 2012. Image Retrieval for Image-Based Localization Revisited. In: British Machine Vision Conference (BMVC). pp. 76.1–76.12. 11, 17

[179] Saupe, D., Vranić, D. V., 2001. 3D model retrieval with spherical harmonics and moments. In: Joint Pattern Recognition Symposium. Springer, pp. 392–397. 3

[180] Saurer, O., Baatz, G., Köser, K., Ladicky, L., Pollefeys, M., 2016. Image Based Geo-localization in the Alps. International Journal of Computer Vision (IJCV) 116 (3), 213–225. 3, 4, 5, 11, 17, 18, 19

[181] Schindler, G., Brown, M., Szeliski, R., 2007. City-Scale Location Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3, 15

[182] Schönberger, J. L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative Evaluation of Hand-Crafted and Learned Local Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. January. 4

[183] Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., 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), 2930–2937. 1, 10, 12, 17, 19, 20

[184] Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A. A., 2011. Data-driven visual similarity for cross-domain image matching. ACM Transactions on Graphics (ToG) 30 (6), 1. 3, 5, 8, 15

[185] Sivic, J., Zisserman, A., 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1470–1477. 2, 6

[186] Sizikova, E., Singh, V. K., Georgescu, B., Halber, M., Ma, K., Chen, T., 2016. Enhancing Place Recognition using Joint Intensity - Depth Analysis and Synthetic Data. Proceedings of the IEEE European Conference on Computer Vision Workshop (ECCVW), 1–8. 5, 17, 18

[187] Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. In: ACM Transactions on Graphics (TOG). Vol. 25. ACM, pp. 835–846. 19

[188] Song, Y., Chen, X., Wang, X., Zhang, Y., Li, J., 2016. 6-DOF Image Localization From Massive Geo-Tagged Reference Images. IEEE Transactions on Multimedia (ToM) 18 (8), 1542–1554. 3, 9, 10, 16, 20

[189] Stumm, E., 2015. Building Location Models for Visual Place Recognition. Ph.D. thesis. 9, 18

[190] Stumm, E., Mei, C., Lacroix, S., 2015. Location graphs for visual place recognition. In: Proceedings of the IEEE International Conference of Robotics and Automation (ICRA). No. May. 3, 9

[191] Stumm, E., Mei, C., Lacroix, S., Nieto, J., Hutter, M., Siegwart, R., 2016. Robust Visual Place Recognition with Graph Kernels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4535–4544. 4, 9

[192] Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M. J., 2015. On the performance of ConvNet features for place recognition. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). Vol. 2015-Decem. pp. 4297–4304. 3, 5, 7, 8, 18

[193] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M. J., 2015. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. In: Robotics Science and Systems (RSS). 3, 5, 7, 8, 21

[194] Svarm, L., Enqvist, O., Kahl, F., 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. 10, 11, 17, 20

[195] Tian, Y., Chen, C., Shah, M., 2017. Cross-View Image Matching for Geo-localization in Urban Environments. arXiv preprint. 15

[196] Tolias, G., Avrithis, Y., 2011. Speeded-up, Relaxed Spatial Matching. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain. 19

[197] Tolias, G., Jégou, H., 2014. Visual query expansion with or without geometry: Refining local descriptors by feature aggregation. Pattern Recognition 47 (10), 3466–3476. 9

[198] Tolias, G., Sicre, R., Jégou, H., 2016. Particular object retrieval with integral max-pooling of CNN activations. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 1–11. 7, 8

[199] Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T., 2015. 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3, 6, 8, 14, 16, 17, 18, 19, 20

[200] Torii, A., Sivic, J., Okutomi, M., Pajdla, T., 2015. Visual Place Recognition with Repetitive Structures. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 37 (11), 2346–2359. 3, 6, 19

[201] Torii, A., Sivic, J., Pajdla, T., 2011. Visual localization by linear combination of image descriptors. In: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW). 1, 9, 10, 16

[202] Torralba, A., Murphy, K. P., Freeman, W. T., Rubin, M. A., 2003. Context-based vision system for place and object recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 3. pp. 273–280. 8, 17, 18

[203] Tzeng, E., Zhai, A., Clements, M., Townshend, R., Zakhor, A., 2013. User-driven geolocation of untagged desert imagery using digital elevation models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 237–244. 3, 17

[204] Valentin, J., Fitzgibbon, A., Nießner, M., Shotton, J., 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). pp. 4400–4408. 12, 17

[205] Valgren, C., Lilienthal, A. J., 2010. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. Robotics and Autonomous Systems (RAS) 58 (2), 149–156. 3, 14

[206] VC, Hough, P., 1962. Method and means for recognizing complex patterns. 3

[207] Vo, N. N., Hays, J., 2016. Localizing and Orienting Street Views Using Overhead Imagery. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 9905. pp. 494–509. 15

[208] Vo, N. N., Jacobs, N., Hays, J., 2017. Revisiting IM2GPS in the Deep Learning Era. arXiv preprint (1). 13, 16

[209] Walch, F., 2016. Deep Learning for Image-Based Localization. Ph.D. thesis, Technical University of Munich. 12

[210] Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D., 2017. Image-based Localization with Spatial LSTMs. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 12, 21

[211] Wan, W., Liu, Z., Di, K., Wang, B., 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. 10, 11, 13, 17

[212] Wan, X., Liu, J., Yan, H., 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. 3, 4, 9, 10, 14, 15, 16

[213] Wang, J., Zhang, T., Song, J., Sebe, N., Shen, H. T., 2017. A Survey on Learning to Hash. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 13 (9). 8

[214] Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R., 2016. TorontoCity: Seeing the World with a Million Eyes. arXiv preprint. 19

[215] Wang, S., Fidler, S., Urtasun, R., 2015. Lost shopping! monocular localization in large indoor spaces. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 2695–2703. 17, 18

[216] Wang, Z., Fan, B., Wu, F., 2013. FRIF: Fast Robust Invariant Feature. In: British Machine Vision Conference (BMVC). 3

[217] Weyand, T., Kostrikov, I., Philbin, J., 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 9905. pp. 37–55. 12, 16, 17, 18, 20

[218] Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 3961–3969. 15

[219] Wu, J., Christensen, H. I., Rehg, J. M., 2009. Visual place categorization: Problem, dataset, and algorithm. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 4763–4770. 18

[220] Yan, K., Wang, Y., Liang, D., Huang, T., Tian, Y., 2016. CNN vs. SIFT for Image Retrieval. In: Proceedings of the ACM International Conference on Multimedia (MM). pp. 407–411. 3, 5, 6, 7

[221] Yi, K. M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned Invariant Feature Transform. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 9905. pp. 467–483. 4

[222] Zamir, A. R., Hakeem, A., Van Gool, L., Shah, M., Szeliski, R., 2016. Large-scale visual geo-localization. Advances in computer vision and pattern recognition. 1, 2

[223] Zamir, A. R., Shah, M., 2010. Accurate image localization based on google maps street view. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 6314 LNCS. pp. 255–268. 3, 8, 9, 16, 17

[224] Zamir, A. R., 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, 8, 9, 16, 19

[225] Zeisl, B., Sattler, T., Pollefeys, M., 2015. Camera Pose Voting for Large-Scale Image-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2704–2712. 10, 11, 12, 17

[226] Zhang, W., Kosecká, J., 2006. Image Based Localization in Urban Environments. In: 3D Data Processing, Visualization and Transmission (3DPVT). 15

[227] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2016. Pyramid Scene Parsing Network. arXiv preprint. 21

[228] Zheng, L., Yang, Y., Tian, Q., 2017. SIFT Meets CNN: A Decade Survey of Instance Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 14 (8). 6

[229] Zhi, T., Duan, L.-Y., Wang, Y., Huang, T., 2016. Two-Stage Pooling of Deep Convolutional Features for Image Retrieval. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). 7

[230] Zitnick, C. L., Dollár, P., 2014. Edge boxes: Locating object proposals from edges. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Springer, pp. 391–405. 3, 5