ARL-TR-8938 • Apr 2020

Outdoor Visual Localization: A Survey

by Sarah Leung and E Jared Shamwell

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8938 ● APR 2020

Outdoor Visual Localization: A Survey

by Sarah Leung
General Technical Services, LLC

E Jared Shamwell
Sensors and Electron Devices Directorate, CCDC Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2020	Technical Report	July 2019-September 2019

4. TITLE AND SUBTITLE
Outdoor Visual Localization: A Survey

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Sarah Leung and E Jared Shamwell

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Combat Capabilities Development Command Army Research Laboratory
ATTN: FCDD-RLS-SI
2800 Powder Mill Road
Adelphi, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8938

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Primary author's email: <sarah.leung.ctr@mail.mil>.

14. ABSTRACT
GPS-denied localization for the Soldier is paramount to military mission success. Sensors such as cameras and inertial measurement units are already ubiquitous in the aid of navigation. However, state estimation utilizing such sensors produces drift that currently must be squashed using absolute position measurements from unreliable GPS signals. To mitigate this uncertainty, we consider visual localization to provide an alternative global state estimate without the need for any current infrastructure. In this report, we investigate the current methodologies and capabilities, and review state-of-the-art approaches for visual localization and place recognition.

15. SUBJECT TERMS
visual localization, visual geo-localization, visual place recognition, scene matching, GPS-denied navigation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sarah Leung
Unclassified	Unclassified	Unclassified	UU	54	19b. TELEPHONE NUMBER (Include area code) 301-394-5557

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

Military operations for the Soldier, whether dismount or vehicle-borne, are currently highly dependent upon satellite-based navigation, including the use of GPS signals, which provide location and time with relatively high accuracy and precision. The typical commercially available GPS on smartphones claims accuracy of $5\,\mathrm{m}$ in open sky areas, with reduced accuracy around tall buildings.[1] Real Time Kinematic differential ranging techniques for the Global Navigation Satellite System (GNSS), which utilize base station and rover pairs with carrier signals and transmitted error corrections, can provide accuracy on the order of $1\,\mathrm{cm} + 1\,\mathrm{ppm}$,[2] thus trivially solving the *localization* problem for many outdoor applications. However, GNSS signals are weak by design, balancing energy allocation for the constant multivariate broadcast from solar powered satellites, which open them up to vulnerabilities of jamming and spoofing.[3–5] Jammers, being inexpensive and commercially available (albeit illegal), have been shown to disrupt civil services in many cases.[5] GPS spoofing is an additional threat to trajectory tracking, as instead of blocking the signal, it deceives the receiver with false location or time. Thus, there is a critical need and growing research to mitigate the unreliability of GPS signals for localization scenarios by developing alternative or auxiliary methods for accurate position recovery.

1.2 Visual Localization

Given the limited power and weight capacity of the dismount Soldier, we consider the use of low cost, size, weight, and power (C-SWaP) sensors. It is typical for pose or position estimation to be performed via algorithms for the ubiquitous camera. The process of determining the pose or odometry of a camera-equipped body using vision is called visual odometry (VO). Another commonly used navigation sensor is the inertial measurement unit (IMU), which contains accelerometers, gyroscopes, and optionally magnetometers, measuring linear acceleration, rotational rates, and compass direction, respectively. Commercial smartphones often utilize small form microelectromechanical systems-based IMUs. Both camera and IMU are typically contained on handheld devices, such as smartphones, as well as on small robotic platforms, such as the micro aerial vehicle. Localization pipelines that use both vision and IMU data are called visual-inertial odometry (VIO).

Despite VO and VIO algorithms performing well on numerous computer vision benchmarks,[6] they are not robust to failures stemming from environmental artifacts. The algorithms often produce drift that grows with time in the absence of correction from a global sensor (e.g., GPS) or loop closure from a simultaneous localization and mapping (SLAM) system. In a SLAM system, the loop closure component is usually contained in an offline module, decoupled from the online pose estimation. In order to close the loop on the map, the system must perform a task called *visual place recognition*, in which it determines if the current image matches a location previously visited, after which the map and trajectory are adjusted accordingly.

In this work, we consider not only the place recognition component, but also the full *visual localization* approaches of determining the 6-degrees-of-freedom (DOF) pose of a query image, given a representation of the known environment, available as a database of images or a 3D model. We provide an overview of the state-of-the-art in outdoor visual localization for GPS-denied environments of various scales, across changing appearances, with consideration for multiple sensory and ancillary modalities.

1.3 Overview

There is much overlap between the fields of place recognition, visual localization, and SLAM, which are well studied in various surveys. Most similar to this report is the work of Brejcha and Čadík,[7] which provides an overview of visual geolocalization. The recent survey of Piasco et al.[8] provides a more comprehensive review focused on all components of city-scale visual-based localization, which is the most widely studied scale in the area of visual localization. The work of Lowry et al.[9] gives an overview of place recognition, detailing its main modules of image processing, mapping, and belief generation. The mapping component of place recognition, in which the system decides which information to add or remove from the world representation, is not studied in this survey. Place recognition and visual localization are also vital in SLAM for the detection and closing of loops in a trajectory. We do not discuss the details of SLAM in this report, but refer the reader to existing works[10,11] for a comprehensive look at the visual-based approaches to SLAM.

The organization of the remainder of this report is as follows: Section 2 discusses sensory modalities and ancillary environmental data modalities used in visual lo-

calization; Section 3 reviews techniques for scene matching between a query and database images as used for image retrieval in visual localization; Section 4 discusses pose estimation techniques for matched images; Section 5 outlines classifications of visual localization approaches and highlights specific works utilizing different types of approaches; Section 6 discusses cross-domain visual localization for a wide demonstration of applications and capabilities; Section 7 describes metrics and datasets for evaluation and benchmarking of approaches; and Section 8 provides a discussion of the findings and conclusions for future work needed to adapt the field for military relevance.

2. Sensor and Ancillary Modalities

2.1 Visual

2.1.1 Electro-Optical

The most ubiquitous sensor used in visual localization is the electro-optical camera. This is due to its low C-SWaP, as well its wide commercial availability. For this reason, it is the most commonly studied modality for place recognition and visual localization. However, there are inherent weaknesses to using purely visual imagery. Visual imagery is affected by artifacts, such as exposure, noise, and motion blur, as well as time-of-day, weather, and seasonal changes. Many methods attempt to achieve robustness against such changing conditions by augmenting data with random transformations, as well as training algorithms across multiple environmental conditions.

While most commercial cameras capture red, green, blue (RGB) images, many conventional image descriptors, and the VO algorithms that employ them, require grayscale images. However, some whole image descriptors and learning-based methods have been using RGB images. RGB imagery is often unreliable during the daytime in the presence of illumination changes from sunlight and weather, but has demonstrated consistency at nighttime.[12] Color information can also be used to identify and remove shadows, which can cause scene matching failures.

2.1.2 Stereo / Depth

Some localization systems utilize stereo cameras, because depth can be estimated with a known stereo baseline and known camera calibration. Monocular VO and depth estimation suffer from scale ambiguity. However, the use of stereo allows for

recovery of true scale. Wide stereo baselines provide discernment of greater depths, but have a larger blind range, in which objects are too near and cannot be seen by both cameras. There are also depth cameras, which can be units containing stereo cameras whose data are processed and outputted as a depth image, or red, green, blue, and depth (RGB-D) cameras. RGB-D cameras produce dense depth registered with visual imagery, and can exploit 3D data to improve place recognition.[9] RGB-D data is also widely used in SLAM to produce dense point cloud models with color information also useful for semantic segmentation.

2.1.3 Infrared (IR) / Thermal

It is critical in military operations to have sensors that can see regardless of time of day, illumination, weather, and environmental conditions, such as fog and dust. For this reason, we consider infrared (IR) imagers, which capture wavelengths beyond the visual spectrum, encompassing the near infrared (NIR), short-wave infrared (SWIR), and long-wave infrared (LWIR) ranges. Moving away from the focus of improving single modality performance, there is a desire to fuse additional sensor modalities to improve performance and robustness to sensor failures and changing environmental conditions. Using IR sensors can bypass some of the issues present in visual imagery, discussed above, that often cause localization failures. It has been shown that a combined visible-IR representation provides promising place recognition results.[13] We refer the reader to Section 6.3 for a more detailed description of cross-spectral work.

2.2 Environment Map / Model

Place recognition and visual localization require prior knowledge of an environment. In many approaches, a database of geotagged images is used. For more accurate pose estimation, some image databases have been processed to build 3D reconstructions. The points in the databases typically include associated feature descriptors for matching. Many datasets for visual localization are catalogued in Section 7.

2.2.1 2D Maps

Several methods utilize 2D data, including satellite/aerial imagery[14,15] and land attribute cover maps.[14] Land cover data for the United States, provided through the National Land Cover Database*,[16–18] is publicly available from 2001 to present,

*https://www.usgs.gov/centers/eros/science/national-land-cover-database

at 5-year-update intervals. It is a more general land cover database, with 16 land cover classes, at 30-m resolution. Additional US land cover data more focused on habitat identification is available*, including more classes of land cover. Figure 1 illustrates a land cover map and example classes correponding to aerial data of the same geographic location.

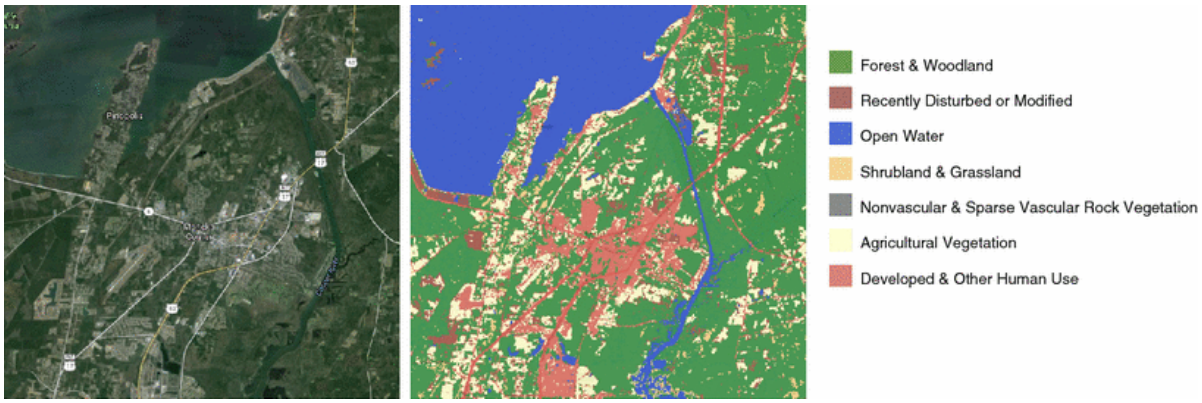

Fig. 1 2D map data used in visual localization, including aerial data (left) and land cover data (right). Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Pattern Analysis and Applications, State-of-the-art in visual geo-localization. Brejcha, Jan and Čadík, Martin, 2017

2.2.2 3D Models

City-scale (urban) approaches typically rely upon sparse 3D models, represented as a point cloud. These models can be generated using light detection and ranging (LIDAR) data, or reconstructed from multi-view stereo (MVS)[19] or Structure-from-Motion (SfM).[20] Recent benchmarking datasets have generated SfM models from database images using the open-source COLMAP.[21] SfM constructs a 3D model from many overlapping images, and thus not all images require geotags for the model to be geo-registered. Other 3D models used in urban environments may be voxel-based or coarse computer-aided design (CAD) models.[22]

For natural (mountainous) landscapes, approaches usually utilize a digital elevation model (DEM), digital surface model (DSM), or digital terrain model (DTM). These models are typically available publicly through government databases. The National Elevation Dataset†[23] provided DEMs for the United States. However, it has recently

*http://gapanalysis.usgs.gov/gaplandcover/
†https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map

been updated with the ongoing 3D Elevation Program*, started in 2016, to produce high resolution elevation data using LIDAR, combined with the bare earth elevation data, to be completed within a decade. The multitude of point clouds currently available contain more than 12 trillion LIDAR points.

3. Scene Matching

In order to perform visual scene matching, the query and database images require data representation, in the form of descriptive and distinctive features. These features, often highly dimensioned, may require aggregation and quantization to allow for efficient detection, description, and matching.

3.1 Features

Data representation in an image conventionally consists of features, which are extracted using interest point detectors, to produce a set of stable regions under various viewing conditions. For detected regions, descriptors encode the distinctive content of the feature for matching. Some features were originally designed as joint detectors and descriptors, but are typically decoupled for the purpose of *image retrieval*, which is the search for image similarity. See Table 1 for a list of commonly used feature detectors and descriptors in place recognition and visual localization.

Local features are detected and described at the pixel and sub-pixel level, and are typically more robust than global features against changes in appearance and occlusions. Introduced two decades ago, the Scale-Invariant Feature Transform (SIFT)[24] is still the most widely used and a top performing local feature descriptor. Although SIFT is both a detector and a descriptor, it is often used as a desciptor with the Hessian-affine[25] point detector. The Principal Components Analysis-SIFT (PCA-SIFT)[26] descriptor reduces the SIFT dimension from 128 to 36, thus lowering its discriminative ability, in order to expedite the matching process. RootSIFT[27] improves the matching performance of SIFT with minor overhead cost. For real-time applications, Speeded Up Robust Features (SURF)[28] have been employed as a light alternative to SIFT. Binary features, such as Binary Robust Independent Elementary Features (BRIEF)[29] and Binary Robust Invariant Scalable Keypoints (BRISK),[30] accelerate matching speed using the Hamming distance instead of Euclidean distance.

The survey of Mikolajczyk and Schmid[31] provides an evaluation of early traditional

*https://www.usgs.gov/core-science-systems/ngp/3dep

local descriptors. With the introduction of convolutional neural networks (CNNs) to the field of image description, it has been necessary to evaluate learned descriptors against traditional descriptors, which has been performed by Schönberger et al.,[32] while Balntas et al.[33] introduce a benchmark for evaluating these two classes of local descriptors. Several learned descriptors, such as TFeat[34] and Patch-CKN,[35] seek to be a drop-in for the default SIFT descriptor, used with a traditional interest point detector. Other learned features, such as DeepDesc,[36] Learned Invariant Feature Transform (LIFT),[37] SuperPoint,[38] and D2-Net[39] provide the full detection

Table 1 Feature detectors and descriptors

Name	Type	Learned	Detector	Descriptor
Hessian-affine[25]	Local		x	
FAST[40]	Local		x	
SIFT[24]	Local		x	x
PCA-SIFT[26]	Local			x
RootSIFT[27]	Local			x
SURF[28]	Local		x	x
BRIEF[29]	Local		x	x
BRISK[30]	Local		x	x
ORB[41]	Local		x	x
FREAK[42]	Local			x
TFeat[34]	Local	x		x
Patch-CKN[35]	Local	x		x
DeepDesc[36]	Local	x	x	x
LIFT[37]	Local	x	x	x
SuperPoint[38]	Local	x	x	x
D2-Net[39]	Local	x	x	x
Color histogram[43]	Global			x
Gist[44]	Global			x
BRIEF-Gist[45]	Global			x
WI-SURF[46]	Global			x
HOG[47]	Patch			x
DoG[48]	Blob		x	
MSER[49]	Blob		x	
Lines[50]	Semantic		x	
Edges[51]	Semantic		x	
Skyline[52]	Semantic		x	x
PointRay[53]	Semantic		x	x

and description pipeline.

Global features do not require detection, but describe the whole image with a single signature of high dimensionality, which tends to be less computationally intensive than describing multiple local salient regions in an image. Gist[44] is a commonly used global image descriptor using averaged Gabor filters at various orientations and frequencies to represent the "gist" of a scene. Color histograms[43] are another global descriptor well used in image retrieval.

Global features can be used on image patches, typically extracted on a grid or with a sliding window. Patches can also be automatically selected using saliency with an attention mechanism. Histogram of Oriented Gradients (HOG)[47] is a patch descriptor used for describing architectural features in landmarks. Maximally Stable Extremal Regions (MSER) and difference of Gaussians (DoG) are blob detectors that extract regions of interest in an image. Global features can also be combined with or formed from local features extracted on a grid. Sünderhauf and Protzel[45] use a global descriptor based on the local BRIEF descriptor to perform loop closure. Badino et al.[46] use a whole image SURF (WI-SURF) for localization.

Global features can be learned for place recognition using the highly accessible and versatile CNN.[54] Many of these learned global descriptors are used in end-to-end pipelines that will be discussed in greater detail in the following sections.

Semantic features use higher-level representation to add meaning to data. These can be geometric shapes such as lines,[50] which are good for man-made structures such as buildings, or contours from edge detection.[51] For methods like horizon matching, semantic features such as skylines[52] are often used. For urban geo-localization, Bansal and Daniilidis introduced PointRay,[53] pairing points with direction vectors to represent building corners.

3.2 Quantization

The Bag-of-Words (BoW) approach was originally used for modeling documents with the use of word histograms. It was introduced to the image retrieval community in the work of Sivic and Zisserman.[55] DBoW2[56] is an open-source method utilizing a Bag-of-Binary-Words for fast place recognition, and has been used in SLAM methods for loop closure. In the image domain, the BoW model, also called Bag-of-Features, represents the image similarly to a document as a bag of visual words,

and is made usable with the introduction of features such as SIFT. Local features are quantized to visual words, which are encoded into a vocabulary, a pre-trained codebook. This can be done as a hard quantization, or a soft quantization, where each feature can be quantized to multiple visual words.

The visual words are assigned specific weights, emphasizing discriminative features for matching. Feature weighting commonly uses term frequency-inverse document frequency, which is the product of the term frequency, which is defined by the occurrence count of the word within the image, and the inverse document frequency, which is a global measure of the word in the vocabulary. Visual burstiness[57] is a phenomenom in word weighting that occurs in the presence of repetitive structures in an image, which corrupts visual similarity measures. Jégou et al.[57] proposed effective solutions adjusting the term frequency during the detection phase. Torii et al.[58] utilize burstiness as a distinguishing feature in buildings and modify the similarity measurements accordingly.

Jégou et al.[59] introduced the Hamming embedding, providing a more efficient and discriminative encoding by subdividing Voronoi cells, and assigning binary signatures to features. The Fisher Vector[60] uses a Gaussion mixture model (GMM) to model the distribution of features extracted in an image, trained using the maximum likelihood estimation. The encoding aggregates the first and second order statistics of the features under the GMM. Jégou et al.[61] introduced a new method for aggregating local descriptors into a compact image representation called Vector of Locally Aggregated Descriptors (VLAD), which is capable of searching a 10 million image dataset in about $50\,\mathrm{ms}$. Rather than assigning the closest visual word to the feature, VLAD uses the difference between them. Many image retrieval techniques have employed the VLAD descriptor, including DenseVLAD[62] and the learning-based NetVLAD,[63] which will be detailed in Section 5.2.1.

The codebook is clustered using an approach such as k-means, and partitioned into Voronoi cells with the visual words at the centroids. For large codebooks, approximate methods are critical for scalability. The hierarchical k-means (HKM)[64] algorithm first partitions the space into a few clusters, and then recursively partitions each cluster into smaller clusters, resulting in a significally lower complexity than a flat k-means clustering. The approximate k-means[65] algorithm indexes the centroids using a random forest, and has been shown to yield lower quantization error

than the HKM, leading to higher performance.

3.3 Matching

Matching of descriptors is a search for the closest distance, using Euclidean (L2) distance for floating point descriptors and Hamming distance for binary descriptors. For small scale applications, a simple nearest neighbor (NN) search can be used. Approximate nearest neighbor (ANN) search methods have been utilized, including Principal Components Analysis (PCA) for dimension reduction and binary encoding for hashing-based methods. Muja and Lowe[66] present scalable NN methods for high dimensional spaces, released in a popular library called Fast Library for Approximate Nearest Neighbors.

For more efficient storage and retrieval, large codebooks for BoW approaches often employ the inverted index, which makes use of the sparsity of the visual vocabulary. The inverted index entries are the visual words, which have attached inverted lists, posting the image IDs and binary features.

Several machine learning methods have been used in place of NN and ANN methods. Linear and Support Vector Machine classifiers have been used to treat matching as a classification task, or to predict descriptor robustness for improved matching with the removal of poorly discriminant features.

Re-ranking methods improve image retrieval performance by post-processing the candidates to remove irrelevant ones. Lowe's ratio test[48] uses the ratio of the distances of the nearest neighbor and second-nearest neighbor, noting that a simple threshold on distance to closest feature is invalid due to differences in descriptor discriminative ability. For the task of visual geo-localization in particular, it is practical to consider geographical consistency. Zamir and Shah[67] utilize geographic re-ranking in geo-localization to quickly remove geographically inconsistent candidates. One shortfall of the BoW model is the disregard of geometry, and thus many researchers choose to use spatial re-ranking to enforce geometric consistency in candidates. One widely used method for geometric re-ranking is Random Sample Consensus (RANSAC),[68] which calculates affine transformations for correspondences to reject outliers that are not consistent with the consensus transformation. RANSAC provides robust estimation but suffers from inefficiency. Hough voting in parameter space is another spatial verification technique that is more efficient

than RANSAC at the cost of lower accuracy. Many RANSAC variants have been created to mitigate the efficiency shortcomings. Philbin et al.[65] utilize spatial verification with LO-RANSAC,[69] or locally optimized RANSAC, using subclasses of 3- to 5-DOF affine transformations.

As an alternative or addition to re-ranking, Jégou et al.[59] introduce a weak geometric constraint to verify consistency of the angle and scale of matching descriptors, integrated into the BoW inverted indexing via additional scores computed from histograms. Toft et al.[70] introduce a semantic match consistency check for scoring correspondences and rejecting poor matches, using the semantic labels associated with the feature points. The semantic consistency scores are used to bias the RANSAC sampling toward semantically consistent correspondences, increasing robustness and efficiency.

3.4 CNN Methods

A variety of CNN-based methods have been used for image retrieval and demonstrated state-of-the-art performance surpassing the traditional approaches. The methods include approaches that follow the traditional pipeline, but replace the conventional features with learned features, as well as models trained end-to-end to produce encoded global descriptors. Models trained for large scale image classification, such as AlexNet,[71] can be fine-tuned for the task of place recognition.

Fully connected layers can be used as global feature descriptors, due to global receptive field, achieving fair retrieval accuracy using Euclidean distance. Lower-level convolutional filters can be extracted as intermediate local descriptors, due to smaller receptive fields, often densely applied over the whole image, and are more robust to image transformations including occlusions.

Features can be aggregated using pooling, through both sum-pooling and max-pooling. Sum-Pooled Convolutional[72] features sum the responses for each activation map. Maximum Activations of Convolutions (MAC)[73] compute a global descriptor, using the maximum value of each intermediate feature map, concatenated within a convolutional layer, in a single pass. Regional Maximum Activations of Convolutions (R-MAC)[73] improve upon MAC by computing the maximum activations over multiple-sized regions in the image. Gordo et al.[74] use R-MAC with a Region Proposal Network to select the max-pooled regions on the activation map.

NetVLAD[63] is an end-to-end trainable aggregation layer modeled after the VLAD descriptor.

Figure 2 compares traditional and CNN-based pipelines. We refer the reader to the decade survey of Zheng et al.[75] for a comprehensive comparison of traditional and CNN-based techniques in the broader field of content-based image retrieval.

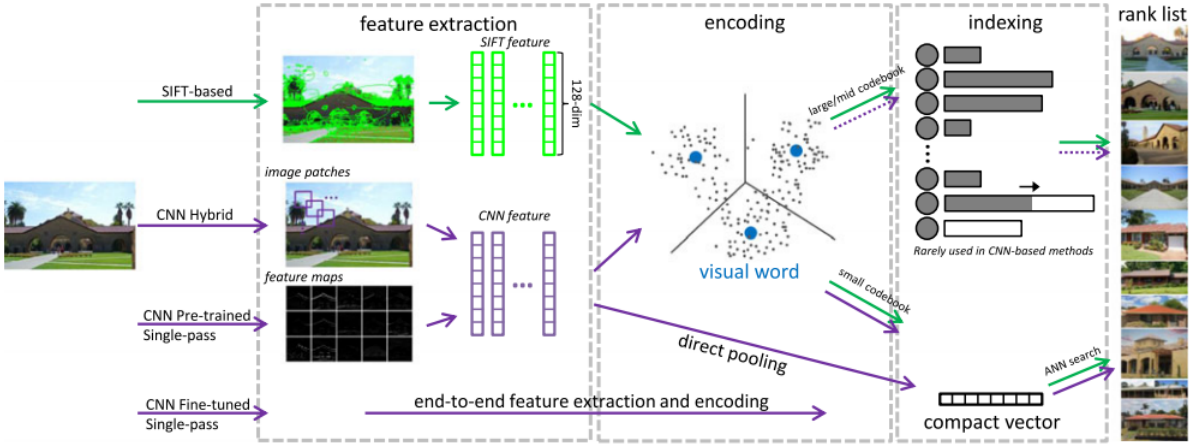

Fig. 2 Comparison of traditional (SIFT-based) and CNN-based image retrieval pipelines. ©2017 IEEE. Reprinted, with permission, from Zheng L, Yang Y, Tian Q. SIFT meets CNN: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;40(5):1224–1244

4. Pose Estimation

4.1 2D-2D

2D image retrieval–based localization techniques require extra post-processing to obtain 6-DOF camera poses. The baseline approach is to simply use the position or pose associated with the selected nearest neighbor. Spatial re-ranking is also a method for retrieving pose, based on inliers or geometric burstiness. Sattler et al.[76] propose performing small-scale SfM on the fly for the local neighborhood of the query image, and obtaining the global pose by registering the local SfM reconstruction with the geotags from the database images. Pose recovery from 3D structure is discussed in the following section.

For predominantly planar problems, a simple homography can be used. For more generic cases, many multi-view geometry algorithms[77] have been utilized. Nistér's

five-point algorithm[78] is an efficient and robust solution to the relative pose problem with five correspondences, evaluated against the traditional eight-point algorithm,[79] which provides a unique solution to the relative pose between two uncalibrated images. Rocco et al.[80] used CNNs to estimate an affine transformation, followed by a thin-plate spline transformation between two images. Recently, many deep learning VO algorithms have been introduced to estimate relative pose between images, but without stereo or depth, cannot recover metric scale.

4.2 2D-3D

For 6-DOF pose estimation using 3D structure, in the form of a reconstructed SfM point cloud, with associated feature descriptors, Features to Points (F2P) with established 2D-3D correspondences is employed. Hartney and Zisserman outlined the widely used Perspective-n-Point (PnP)[77] formulation for obtaining absolute camera pose relative to an SfM model. Without known camera intrinsincs, pose can be recovered with a minimum of six F2P correspondences, called the P6P formulation, using a RANSAC scheme and Direct Linear Transformation.[77] If camera intrinsics are known, the problem can be reduced to a P3P problem solvable with three correspondences.[81]

Pose refinement with 3D point clouds has widely been achieved using bundle adjustment or the Iterative Closest Point[82] algorithm.

4.3 Pose Regression

Some methods approach visual localization as a pose regression problem. In such cases, machine learning and CNN approaches can be used. For localization using depth data, a regression forest or GMM may be utilized, but needs to be trained for each known environment.

PoseNet[83] uses a CNN for camera relocalization. The network is trained on paired image-pose data to automatically regress the 6-DOF pose of a camera from a color image, without the need for depth. The network is able to relocalize within $2\,\mathrm{m}$ and 3° in real-time, with a $5\text{-}\mathrm{ms}$ runtime. The authors introduce a Bayesian variant[84] that leverages the uncertainty measure to estimate localization error. The authors again improve performance with the use of novel loss functions based on geometry and scene reprojection.[85] Walch et al.[86] use long short-term memory (LSTM) units on CNNs for structured feature dimensionality reduction and demonstrate improved

performance over PoseNet.

5. Visual Localization

Visual localization, also called visual geo-localization, can be categorized by a few different factors. Following Brejcha and Čadík,[7] we consider the first factor of interest to be the scale and environment of the approach. Some work is geared toward working on the global scale, while others focus on a smaller scale, in a city-scale or natural environment. The second factor is the algorithmic approach, which typically falls under 2D image retrieval or 3D structure based, as detailed in Section 3. In the following sections, we review the state-of-the-art approaches under these two types of classifications.

5.1 Scale and Environment Categories

For this work to be militarily relevant, it is necessary for visual localization to be accomplished at multiple scales and environments. Many military operations may take place in areas outside of large civilian populations. It is important to be able to localize using both man-made structures, abundant in urban environments, and natural scenes. Missions may be long-term, and it is thus also necessary to consider both global and city-scale localization.

5.1.1 City-Scale

City-scale visual localization is the most studied scale categorization of the field. For this reason, there are also the most publicly available datasets designed for benchmarking city-scale visualization. Because of the smaller scale of the problem, high accuracy can often be achieved, particularly with the use of 3D structure. Many of these urban approaches will be discussed in greater detail in Sections 5.2.1 and 5.2.2.

Horizon matching has been used in city-scale urban settings, often with upward facing omni-directional cameras. Meguro et al.[87] used an omni-directional IR camera to compare skylines to a DSM with extracted edges. The use of an IR camera instead of a color camera reduces the effects of light exposures that cause errors in the extraction of the skyline. Another approach is SKYLINE2GPS,[22] which uses omni-skylines extracted from color images, and requires a coarse 3D CAD model of the city. Sky segmentation is used for images with clear skies in the daytime, with an alternative algorithm not requiring sky detection for more complex conditions.

5.1.2 Natural Environments

Several methods exist to tackle the problem of visual localization in natural (mountainous) landscape environments. These approaches typically require the use of a digital land model (DEM, DSM, or DTM). It is typical when working with mountainous terrain to utilize skyline or horizon lines as distinct features for matching. Baboud et al.[88] introduced a method for automatic photo-to-terrain alignment. Edges are extracted from query images, and silhouettes from the 3D terrain model, and aligned to produce a registered image. The approach achieved 86% correctly aligned at 0.2° orientation accuracy, at $10\,\mathrm{min}$ per image. The method requires known camera field-of-view and a viewpoint position estimate on the order of 100-m accuracy. Baatz et al.[89] performed geo-localization in mountainous terrain, under the assumption of small roll and low camera height. The method was able to achieve 88% top 1 correct within 1-km radius, at $2\,\mathrm{s}$ per image. Continued work[52] improved sky segmentation, demonstrating image-based geo-localization in the Swiss Alps, covering an area of 40,000 km^2.

There is a sparsity of data and algorithms to tackle other natural environments, such as desert or forest scenes. Some urban and suburban benchmarks contain vegetation, but these countryside and park scenes prove to be more challenging for visual localization, due to dynamic objects and changing appearances. Methods that utilize land cover attribute maps also make determined efforts to localize in natural areas that are removed from dense populations.

5.1.3 Global

Global geolocation estimates image location (typically 3-DOF GPS-like position) as a probability distribution over the Earth's surface. Global geolocation approaches have been studied, with models trained on large datasets containing millions of crowd-sourced geotagged images. These approaches typically retrieve global location on the order of 1000-km accuracy. An early state-of-the-art approach for geolocation was IM2GPS,[90] which utilized a dataset with 6 million geotagged images, the majority of which represent more highly populated areas. Figure 3 shows the global image density of the database used in IM2GPS. The authors evaluate a selection of conventional local and global features for scene matching and perform mean-shift clustering of matched image geolocations.

A more recent learning-based approach called PlaNet[91] used an adaptive partition-

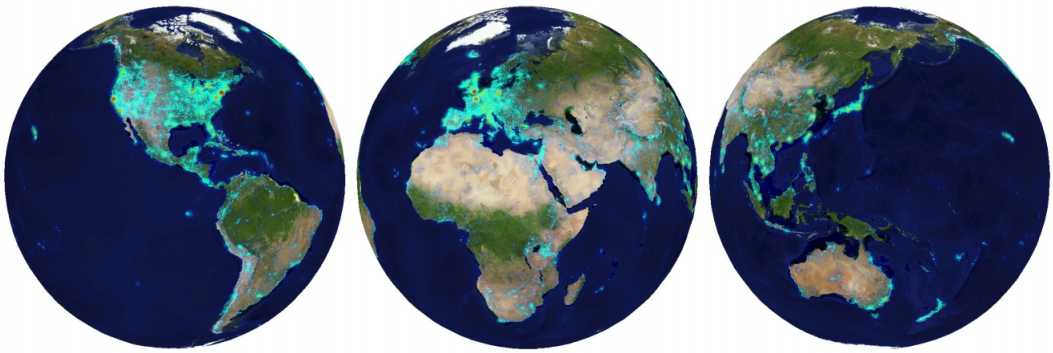

Fig. 3 IM2GPS image distribution over the Earth. ©2008 IEEE. Reprinted, with permission, from Hays J, Efros AA. IM2GPS: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2008; p. 1–8.

ing of the world into cells and LSTM networks to exploit temporal coherence in images. The approaches perform best in images with man-made or natural landmarks, with median localization error growing in city and natural scenes without distinctive landmarks. Vo et al.[92] revisit IM2GPS in the age of deep learning by treating geolocation as a retrieval rather than classification problem, and estimate density of the retrieved nearest neighbors. The authors achieve similar performance gain to PlaNet with a fraction of the database.

5.2 Approach Categories

Most visual localization approaches fall under two algorithmic categories, which have traditionally been investigated independently. The first is the 2D image-based approach and the second is the 3D structure-based approach. Both of these approaches utilize the techniques discussed in Sections 3 and 4. Figure 4 gives a visual comparison of the 2D and 3D visual localization pipelines.

5.2.1 Image-Based Approaches

Image-based approaches to visual localization use image retrieval, with 2D-2D matches from a database of images annotated with 3-DOF location or 6-DOF pose. Image retrieval approaches are widely used in computer vision tasks outside of place recognition and visual localization, such as image classification and scene recognition, and thus have been widely studied at large scale. When used for place recognition, these approaches are typically more robust to appearance changes, and easier to maintain at large scale.

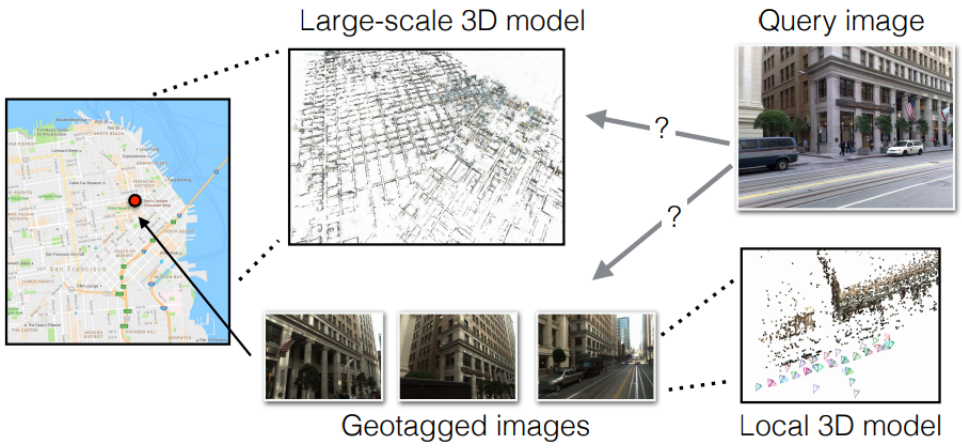

Fig. 4 Illustration of 2D and 3D visual localization. ©2017 IEEE. Reprinted, with permission, from Sattler T, Torii A, Sivic J, Pollefeys M, Taira H, Okutomi M, Pajdla Tomas. Are large-scale 3D models really necessary for accurate visual localization? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 1637–1646.

DisLocation[93] is a state-of-the-art approach that uses the BoW quantization with the Hamming embedding, adapted to prioritize feature distinctiveness by examining the density of the Hamming space. Zamir and Shah[94] perform camera localization using Google Street View images, with nearest neighbor tree search and a *Confidence of Localization* parameter to determine the reliability of the localization estimate. DenseVLAD[62] also utilizes a combination of synthetic views with the VLAD scheme to aggregate RootSIFT[27] descriptors, densely sampled on a regular grid of the image, for a compact image representation. These methods evaluated on Google Street View images demonstrate potential for world-wide scalability. NetVLAD[63] uses a CNN to learn descriptors, which are aggregated into the VLAD descriptor. The authors developed a weakly supervised training procedure using the Google Street Time Machine to obtain imagery of places from various times. NetVLAD demonstrated performance gain over both traditional approaches using RootSIFT and end-to-end learned approaches such as AlexNet.

5.2.2 Structure-Based Approaches

A second approach to visual localization is structure-based, leveraging a 3D model of the environment typically reconstructed using SfM. In many approaches, 2D-3D matching is used to directly regress a 6-DOF pose of the query image, bypassing the image retrieval step used in 2D image-based approaches. Structure-based ap-

proaches typically produce higher accuracy than image retrieval approaches. However, maintaining a large model uses excessive memory, and searching it requires computation capabilities that do not scale well with model size.

Sattler et al.[95] use direct 2D-3D matching with visual word quantization and inverted files, prioritizing cheaper correspondences. Li et al.[96] exploit both the co-occurrence and co-visibility to demonstrate state-of-the-art performance on several large datasets. *Co-occurrence* relationships are illustrated within an image in the frequent appearance of two features spatially close in a place, and in the rare appearance of two features at different times of day. *Co-visibility* considers bidirectional matching of both 2D-3D and 3D-2D correspondences. Active Search[97,98] is a method that improves performance through actively searching for additional matches using both 2D-3D and 3D-2D matching, the latter of which is more efficient due to matching a single point against features. After a 2D-3D correspondence is found, a 3D-2D search is initiated for the neighboring 3D points.

Several approaches assume known camera height and gravitational direction, using IMUs or accelerometers from smartphones under zero or fixed velocity assumptions, in which pitch and roll can be recovered. Camera Pose Voting[99] is one such method that refines pose using a Hough voting scheme and RANSAC with a 3-point solver. City-Scale Localization[100] is another related method that uses fast outlier rejection, and 3-point or 4-point solvers to perform point-to-cone registration, with reprojection error of points propagated as cones in 3D.

Structure-based approaches have used SfM models, but to address the growing availability and superior quality of LIDAR-based point clouds, Nadeem et al.[101] proposed a method to directly match 2D image descriptors with 3D point cloud descriptors for 6-DOF camera localization. The authors trained a descriptor matcher network and demonstrated competitive performance on indoor and outdoor datasets. However, despite its intended usage with LIDAR point clouds, the datasets for training and evaluation use point clouds developed using photogrammetry.

5.2.3 Hybrid Approaches

Some researchers have been using a hybrid between the image and structure based approaches. Irschara et al.[102] used vocabulary tree-based indexing of features to retrieve relevant places in a 3D model, synthetic views to cover image views not evident in the reference database, and a compressed scene representation to increase

signal-to-noise ratio of vocabulary tree queries. Sattler et al.[103] revisited image retrieval for visual localization and analyzed the gap between retrieval-based and direct 2D-3D matching approaches. The authors identified the false positive matches as the main source of the performance gap, and demonstrated that a selective voting scheme can improve performance even when compared with direct matching methods.

Sattler et al. introduced Hyperpoints,[104] which searches for locally unique matches, called hyperpoints, containing no co-visible points, and uses an inverted index with a fine visual vocabulary of 16M words. Sattler et al.[76] later sought to evaluate whether memory-heavy 3D SfM models are necessary for high performance on city-scale visual localization. The authors demonstrated that it was sufficient to use local SfM models inspired by single-photo SfM[105] reconstruction to achieve state-of-the-art performance similar to fully structure-based methods, and provided the first datasets intended for direct comparison of 2D and 3D approaches.

Recently, hierarchical and semantic approaches have achieved state-of-the-art performance in challenging city-scale visual localization benchmarks, winning the 2019 CVPR Long-Term Visual Localization Challenge*, and are thus detailed in Sections 5.2.4 and 5.2.5.

5.2.4 Hierarchical Approaches

Hierarchical approaches use both image-based and structure-based methods advantageously in a two-step process. The pipelines start with a coarse image retrieval to reduce the 3D model size, and finish with a fine local feature matching and pose regression. Sarlin et al. introduced HF-NET,[106] a learned global and local descriptor network, and the authors performed global retrieval to obtain location hypotheses, only matching local features within the retrieved candidate places, as illustrated in Fig. 5. The method is able to achieve 96% daytime and 49% nighttime localization performance on the RobotCar Seasons dataset,[107] at the $(5\,\mathrm{m},\ 10^{\circ})$ error threshold. Germain et al.[108] utilized a similar hierarchical approach with sparse-to-dense matching, using sparse features from the database images and dense aggregated intermediate features, called hypercolumns, from the query images. The authors achieve similar high performance on the benchmark for visual localization under changing conditions.

*https://www.visuallocalization.net/workshop/cvpr/2019/

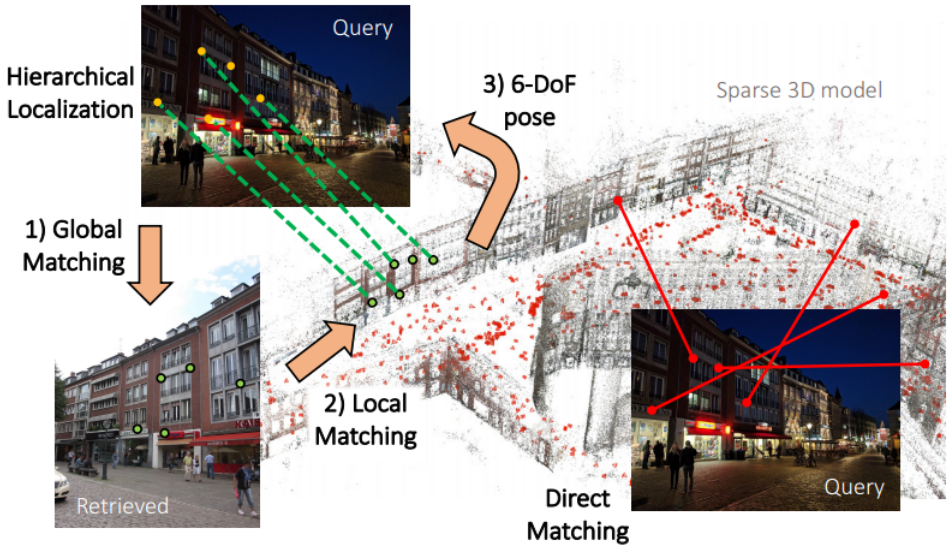

Fig. 5 Hierarchical Localization. ©2019 IEEE. Reprinted, with permission, from Sarlin PE, Cadena C, Siegwart R, Dymczyk M. From coarse to fine: Robust hierarchical localization at large scale. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019; p. 12716–12725

5.2.5 Semantic Approaches

Semantic approaches use semantic segmentation to bring a higher level representation to features for improved visual scene understanding. Semantic mappings are inherently invariant to appearance changes and thus can provide more robustness to visual localization. These earlier approaches[109–112] focused on using semantics to determine salient information and discard ambiguous information in order to improve place recognition. Other methods utilize semantic objects such as lane markings found on roads to guide localization in more focused applications.

More recently, Schönberger et al.[113] performed semantic visual localization using semantic scene completion and a Bag-of-Semantic-Words for quantization. The approach learned 3D descriptors, but required the use of depth alongside the query images and their semantic segmentation, and still struggled with the common failure point of repetitive structures. Toft et al.[70] introduced a semantic match consistency check looking at semantic inliers for feature correspondences to improve long-term visual localization performance. However, the method requires camera height and gravity direction from ground truth. Shi et al.[114] take semantic visual localization one step further by using a sparse semantic 3D map created from semantic segmen-

tations of the database images. Semantic segmentation is used to remove dynamic objects (e.g., pedestrians and vehicles) from the map to produce a more sparse and static representation of the environment. The method is able to achieve 97% daytime and 47% nighttime localization performance on the RobotCar Seasons dataset,[107] at the $(5\,\mathrm{m}, 10^\circ)$ error threshold.

5.2.6 Sequence-Based Approaches

It has been shown that using a sequence of query images rather than a single image improves performance of visual localization. Newman et al.[115] used sequences of images in a visual-laser SLAM approach with traditional descriptors and additional algorithms to handle self-similar foliage and architecture. SeqSLAM[116] is a sequence-based approach that does not use features, but instead uses global image descriptors with local best matching. SeqSLAM demonstrated significant improvement over the early state-of-the-art feature-based SLAM algorithm, FAB-MAP.[117] Recent deep VO approaches[118,119] have also used CNNs and LSTMs with sequences of images for camera localization.

6. Cross-Domain Localization

It is vital for visual localization algorithms to be robust across a variety of appearance changes.

6.1 Cross-Time

The world is by nature a conglomeration of transient places. Some of these are long-term changes, but even over the course of the day, the appearance of a landmark, particularly in the visible domain, will change drastically due to lighting and weather conditions, as well as dynamic objects such as pedestrians and vehicles. Over the course of a year, seasonal changes will be prominent, exhibiting major weather and precipitation changes, as well as vegetation growth. Finally, over the long-term, changes in both natural and man-made landmarks will be evident, as new objects crop up and existing objects change. Examples of these cross-time environment changes within a single place are shown in Fig. 6.

For this reason, many recent datasets focus on benchmarking robust visual localization, across time-of-day and seasonal changes. Google Street View Time Machine tracks the appearance of a location over several years' time. DEMs and land cover maps are also updated every few years to track changes in land.

Fig. 6 Visual localization in changing urban conditions, including day-night, weather, and seasonal changes over time. ©2018 IEEE. Reprinted, with permission, from Sattler T et al. Benchmarking 6DOF outdoor visual localization in changing conditions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; p. 8601–8610

6.2 Cross-View

Robustness to viewpoint changes is also important for visual localization, as many images are taken with different cameras at different angles. To this end, many publicly available datasets contain small changes in viewpoint angles. However, it has been noted that most of the available ground-level imagery, typically used as query images in visual localization, only cover the more densely populated areas of the world. Satellite imagery, however, covers much of the world, from an aerial perspective. There has been some work recently which seeks to perform ground-to-aerial localization, querying ground-level imagery against databases of aerial-view (orthogonal or oblique) imagery.

The work of Lin et al.[14] utilized ground-level, aerial, and land cover attribute imagery during training for cross-view geolocalization. In later work,[15] the authors performed ground-to-aerial geolocalization by querying ground-level images from

a large oblique ($45°$) aerial-view database, providing a dataset of 78,000 aligned cross-view image pairs. Viswanathan et al.[120] focused on localizing an unmanned ground vehicle capturing ground-level images in remote areas with few features, and analyzed the performance of various feature descriptors on the cross-view matching. Shan et al.[121] addressed the geo-registration of MVS models using ground-to-aerial matching. Workman et al.[122] used learned features rather than hand-crafted features, represented in a semantic feature space. Vo and Hays[123] introduced several CNN architectures for cross-view matching. Zhai et al.[124] learned semantic features in ground and aerial imagery. Tian et al.[125] used a Faster R-CNN[126] to detect buildings in query and database images to improve cross-view localization. CVM-Net[127] used a Siamese architecture and NetVLAD[63] global descriptors, demonstrating significant improvement over many earlier state-of-the-art approaches.

6.3 Cross-Spectrum

Li et al.[128] discussed state-of-the-art image fusion, including cross-spectral fusion. Sappa et al.[129] performed a comparative study of wavelet-based image fusion of visible and infrared images. Maddern et al.[13] investigated methods for fusing visual and long-wave infrared (LWIR) thermal imagery for robust night and day place recognition. Rather than fusing the imagery itself, the authors combined the BoW representations of the visual and thermal modalities, and demonstrated that the joint representation yields the best performance in place recognition. Aguilera et al.[130] used CNNs to learn cross-spectral similar measures for matching across images in the visible and IR spectra.

Ricuarte et al.[131] evaluated state-of-the-art feature descriptors on LWIR imagery and compared its performance with the visual domain, in the presence of scale, rotation, blur, and noise. The authors report that SIFT is among the top performer, but results are otherwise inconclusive. Johansson et al.[132] performed further evaluation, adding viewpoint changes and downsampling of feature detectors and descriptors on IR imagery in a systematic approach with standard metrics.[31] Aguilera et al.[133] introduced a feature descriptor (LGHD) for matching across nonlinear intensity variations, including changes in illumination, modality, and spectrum, as well as providing a visual-LWIR dataset. Bonardi et al.[134] also designed a feature descriptor (PHROG) for working across visual modalities, and evaluated performance on datasets including RGB, NIR, SWIR, and LWIR imagery. Firmenichy et al.[135] introduced a gradient direction invariant SIFT (GDISIFT), providing a multi-spectral

interest point descriptor for visible-NIR registration.

Vidas and Sridharan[136] introduced the first monocular SLAM approach employing the thermal-IR modality. They evaluated the algorithms using handheld and bicycle-mounted sequences at human walking speed. Maddern et al.[12] presented applications of localization and mapping using illumination invariant imaging (similar to IR), demonstrating 6-DOF localization over a 24-h day. Color images were converted into an illumination invariant color space to reduce the effects of sunlight and shadows. Borges and Vidas[137] proposed a monocular VO approach using a thermal camera.

Very few datasets exist for benchmarking IR localization algorithms, when compared to visual benchmarks detailed in Section 7. Most IR datasets are geared toward other computer vision tasks such as pedestrian detection or object tracking. Many of the datasets that are taken from a moving camera do not provide ground truth camera pose data. There are a few small IR datasets that could potentially be used for place recognition and visual localization. The Barcelona[133] dataset contains 44 visual-LWIR registered image pairs. The Multi-spectral SIFT[138] dataset contains several hundred scenes in color and NIR imagery, for the purpose of scene category recognition. CVC-13[139] and CVC-15[140,141] Multimodal Stereo Datasets contain hundreds of visual-LWIR image pairs capturing outdoor urban scenes.

7. Evaluation

7.1 Datasets

Most benchmarking datasets contain scenes from urban and suburban environments, many focused on cities with distinctive manmade landmarks. A few datasets are geared toward natural, mountainous terrains (e.g., in the Alps). Methods that seek to perform global geolocation utilize unordered geotagged images sourced from the internet via sites such as Flickr. Generally, visual localization algorithms are evaluated on more densely populated areas, where more images are found. Methods that seek to prove potential with less populated regions of the world often utilize aerial or satellite imagery, or landcover and terrain models as described in Sections 2.2.1 and 2.2.2.

A recent benchmark for outdoor visual localization in changing conditions[107] contains subsets of previous public datasets,[103,142,143] with added ground truth 6-DOF

poses for all query images and reference 3D SfM models. This dataset contains a wide variety of urban, suburban, and park images, including day and night images, seasonal changes, precipitation, and a range of vegetation. The complete dataset contains over 32K database images, 88K query images, and 10M 3D points from 53M features. Table 2 illustrates some of the state-of-the-art results for city-scale visual localization, comparing 2D image-based, 3D structure-based, hierarchical, and semantic localization approaches using this benchmark. Table 3 details characteristics and performance from an encompassing representative of visual localization and place recognition approaches, excluding those already listed in Table 2. Table 4 provides a list of datasets used in outdoor visual localization, arranged by data type, application, and environment.

7.2 Metrics

For strict place recognition or image retrieval, where success is characterized by the matches found by the system, evaluation is typically based on the metrics of precision and recall. Correct matches are *true positives (TP)*, incorrect matches are *false positives (FP)*, and actual matches not recognized by the system are *false negatives (FN)*. Precision is then defined as the percentage of the retrieved matches that are true positive matches. Recall is defined as the percentage of all actual matches that are retrieved as true positive matches.

Table 2 Results (%) from a selection of city-scale visual localization approaches on the RobotCar Seasons and CMU Seasons datasets[107]

			RobotCar Seasons		**CMU Seasons**		
			Day	**Night**	**Urban**	**Suburban**	**Park**
		m	0.25 / 0.5 / 5.0	0.25 / 0.5 / 5.0	0.25 / 0.5 / 5.0	0.25 / 0.5 / 5.0	0.25 / 0.5 / 5.0
		deg	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10
Name	**Type**	**Runtime**					
DenseVLAD[62]	2D	0.3 s	7.6 / 31.2 / 91.2	1.0 / 4.4 / 22.7	22.2 / 48.7 / 92.8	9.9 / 26.6 / 85.2	10.3 / 27.0 / 77.0
NetVLAD[63]	2D	0.1 s	6.4 / 26.3 / 90.9	0.3 / 2.3 / 15.9	17.4 / 40.3 / 93.2	7.7 / 21.0 / 80.5	5.6 / 15.7 / 65.8
Active Search[97,98]	3D	0.6 s	35.6 / 67.9 / 90.4	0.9 / 2.1 / 4.3	55.2 / 60.3 / 65.1	20.7 / 25.9 / 29.9	12.7 / 16.3 / 20.8
City-Scale Loc.[100]	3D	50 s	45.3 / 73.5 / 90.1	0.6 / 2.6 / 7.2	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1	7.0 / 9.6 / 17.0
HF-NET[106]	Hier.	0.05 s	53.8 / 80.4 / 96.0	11.2 / 27.1 / 49.1	**91.6 / 96.4 / 99.1**	**84.7 / 91.5 / 98.6**	**69.3 / 77.8 / 90.5**
Germain et al.[108]	Hier.	0.4 s	45.7 / 78.0 / 95.1	**22.3 / 61.8 / 94.5**	65.7 / 82.7 / 91.0	66.5 / 82.6 / 92.9	54.3 / 71.6 / 84.1
Toft et al.[70]	Sem.	N/A	50.6 / 79.8 / 95.1	7.6 / 21.5 / 45.4	75.2 / 82.1 / 87.7	44.6 / 53.9 / 63.5	30.4 / 37.8 / 48.0
Shi et al.[114]	Sem.	N/A	**54.5 / 81.6 / 96.7**	12.3 / 28.5 / 46.5	88.8 / 93.6 / 96.3	78.0 / 83.8 / 89.2	63.6 / 70.3 / 77.3

Table 3 Results and characteristics from a selection of visual localization approaches

Name	Application	Environment/Scale	Metric	Performance	Runtime
IM2GPS[90]	Global geolocation	Global	Localized within Region threshold (200 km)	15%	N/A
PlaNet[91]	Global geolocation	Global	Localized within Region threshold (200 km)	38%	N/A
Vo et al.[92]	Global geolocation	Global	Localized within Region threshold (200 km)	44%	N/A
Meguro et al.[87]	Horizon matching	Urban	Mean error	1.3 m	N/A
SKYLINE2GPS[22]	Horizon matching	Urban	Mean error	2.8 m	N/A
Baboud et al.[88]	Horizon matching	Mountains	Correctly aligned at 0.2° accuracy	86%	2 min
Baatz et al.[89]	Horizon matching	Mountains	Top 1 correct within 1 km radius	88%	10 s
Lin et al.[14]	Cross-view	Urban/Suburban	Recall @ Top 1%	17%	N/A
Where-CNN[15]	Cross-view	Urban/Suburban	Recall @ Top 1%	22%	N/A
Workman et al.[122]	Cross-view	Urban/Suburban	Recall @ Top 1%	34%	N/A
Vo and Hays[123]	Cross-view	Urban/Suburban	Recall @ Top 1%	64%	N/A
Zhai et al.[124]	Cross-view	Urban/Suburban	Recall @ Top 1%	43%	N/A
CVM-Net[127]	Cross-view	Urban/Suburban	Recall @ Top 1%	91%	N/A
Zamir and Shah[94]	Image-based	Urban/Suburban	Localized within 100 m	38%	N/A
Zamir and Shah[67]	Image-based	Urban/Suburban	Localized within 100 m	47%	N/A
DisLocation[93]	Image-based	Urban/Suburban	Localized within 5 m	55%	N/A
Li et al.[96]	Structure-based	Urban/Suburban	Recall	68%	~5 s
Camera Pose Voting[99]	Structure-based	Urban/Suburban	Localized within 5 m	47%	4 s
Irschara et al.[102]	Hybrid	Urban/Suburban	Recall @ Top 10	88%	0.3 s
Hyperpoints[104]	Hybrid	Urban/Suburban	Localized within 5 m	58%	~5 s
PoseNet[83]	Pose regression	Urban	Median localization error	1.66 m, 4.86°	~0.005 s
Baysian PoseNet[84]	Pose regression	Urban	Median localization error	1.74 m, 4.06°	~0.005 s
Kendall et al.[85]	Pose regression	Urban	Median localization error	0.99 m, 1.06°	~0.005 s
Walch et al.[86]	Pose regression	Urban	Median localization error	0.99 m, 3.65°	N/A
FAB-MAP[117]	SLAM	Urban/Suburban	Recall @ 100% precision	11%	N/A
SeqSLAM[116]	SLAM	Urban/Suburban	Recall @ 100% precision	60%	N/A

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

Precision and recall are not typically evaluated in isolation, but are related on a precision-recall curve. Precision and recall can be combined into a single measure (e.g., precision at 80% recall). Because false positive matches can cause catastrophic failure in mapping and localization, place recognition approaches traditionally prioritize achieving 100% recall, and use precision at 100% recall as the key metric.

For visual localization, recall is often used as the most discriminative metric. It is typical for an image retrieval system to choose the top-k ($1 \leq k \leq 10$) ranked candidates and evaluate whether any of the candidates lie within a tolerance radius for localization. The percentage of queries localized within the threshold is a measure of the recall of the system. More recently, benchmarks have used tiered thresholds of joint position and orientation error, e.g. $(0.25\,\text{m}, 2^\circ)$/$(0.5\,\text{m}, 5^\circ)$/$(5.0\,\text{m}, 10^\circ)$ as high/medium/coarse accuracy metrics.[107] Other methods directly evaluate absolute position and orientation error against ground truth.

8. Conclusion

It has been shown that the bulk of research has been performed in urban, city-scale environments, with a wealth of distinctive landmarks from manmade buildings. However, few focus on vegetation-heavy locations. Even on datasets that include suburban and park environments with more self-similar imagery, it has consistently been obvious that these places with more natural landscapes cause more failures of even the most robust algorithms. More research needs to be focused on overcoming these challenges, which may entail the need to also collect more data in vegetation-heavy environments. Because the Soldier must often rely on sensor modalities beyond the visual spectrum (e.g., IR), there is a great need for additional datasets with IR imagery, especially paired with registered visual imagery.

Milford et al.[157] discussed how small a feature can be to be distinguishable, in an attempt to answer the question of what amount and quality of information are needed

*https://roboticvision.atlassian.net/wiki/spaces/PUB/pages/14188617/

Table 4 Datasets for outdoor visual localization

Name	Application	Environment/Scale	Ground Truth	Data Type
INRIA Holidays[59]	Image retrieval	Global	No	RGB
YFCC100M[144]	Image retrieval	Global	GPS	RGB
World Cities[145]	Image retrieval	Urban (Global)	GPS	RGB
Oxford5K[65]	Landmark recognition	Urban	No	RGB
Paris[146]	Landmark recognition	Urban	No	RGB
San Francisco Landmark[147]	Landmark recognition	Urban	GPS	RGB
Pittsburgh250K[58]	Place recognition	Urban	GPS	RGB
VPRiCE 2015*	Cross-time	Urban/Suburban	No	RGB
SPED[148]	Cross-time	Urban/Suburban	GPS	RGB
Tokyo 24/7[62]	Cross-time	Urban	GPS + Compass	RGB
Alderley Day/Night[116]	Cross-time	Suburban	Location	RGB
Nordland[149]	Cross-time	Train route	GPS	RGB
CMU Visual Localization[143]	Localization	Urban/Suburban	GPS	RGB
Google Street View[67]	Localization	Multi-city	GPS + Compass	RGB
CH1+CH2[52]	Localization	Mountains	GPS	RGB (+ DEM)
Alps100K[150]	Localization	Mountains	GPS	RGB (+ DEM)
GeoPose3K[151]	Localization	Mountains	6-DOF Pose	RGB (+ DEM)
CVUSA[122]	Cross-view	Urban	GPS	RGB + Aerial
GTCrossView[123]	Cross-view	Urban	6-DOF Pose	RGB + Aerial
Panorama[124]	Cross-view	Urban	6-DOF Pose	RGB + Aerial
Toronto City[152]	Cross-view	Urban	6-DOF Pose	RGB + Aerial + Laser
NCLT[153]	Cross-time	Urban/Suburban	6-DOF Pose	RGB + Laser
Oxford RobotCar[142]	Cross-time	Urban	6-DOF Pose	RGB + Laser
KITTI[6]	Odometry	Urban/Suburban	6-DOF Pose	RGB + Laser
Landmarks[96]	Landmark recognition	Urban	No	SfM
Landmarks 3D[154]	Landmark recognition	Urban	No	SfM
Vienna[102]	Landmark recognition	Urban	No	SfM
Aachen[103]	Localization	Urban/Suburban	No	SfM
Dubrovnik6K[155]	Localization	Urban	Location	SfM
Rome16K[155]	Localization	Urban	Location	SfM
Quad6K[156]	Localization	Urban	GPS	SfM
Cambridge[83]	Localization	Urban	6-DOF Pose	SfM
San Francisco[76,96]	Localization	Urban	6-DOF Pose	SfM
Aachen Day-Night[107]	Cross-time	Urban/Suburban	6-DOF Pose	SfM
RobotCar Seasons[107]	Cross-time	Urban	6-DOF Pose	SfM
CMU Seasons[107]	Cross-time	Urban/Suburban	6-DOF Pose	SfM

for localization. Similarly, Latif et al.[158] reduced image size down to 5x4 to determine whether meaningful results can still be obtained at low resolution. This demonstrates that even distant features in an image may be usable as landmarks for recognition. With the use of a sequence of images, even more data is informative, taking advantage of both the spatial and temporal coherence.

There is still a tradeoff between scalability and accuracy, as shown in the performance and capability of 2D and 3D visual localization techniques. Image retrieval techniques operate on very large scale image databases capable of covering the global environment. Cross-view matching techniques offer the ability of global geo-localization due to the coverage of satellite imagery over the Earth. City-scale techniques using SfM models cover significantly smaller area, but can achieve near GPS-level accuracy, recovering orientation in addition to location. In particular, hierarchical and semantic approaches have shown state-of-the-art performance on benchmarks focused on long-term visual localization in changing environments, demonstrating high performance even in nighttime conditions and high-vegetation park scenes.

CNN-based approaches can have runtimes on the order of milliseconds, and are often not dependent on factors such as database size and inlier/outlier counts for RANSAC. Many of the discussed visual localization approaches do not perform in real-time. However, with applications of place recognition in tasks such as loop closure for SLAM run in a decoupled manner in parallel with the primary online localization task, real-time performance may not be necessary. Visual localization can be used as an offline alternative to GPS, providing absolute pose updates at low frame rate. Many approaches have demonstrated robustness to changing viewpoints and changing conditions in a variety of outdoor environments which are applicable for military operations.

9. References

1. Van Diggelen F, Enge P. The world's first GPS MOOC and worldwide laboratory using smartphones. In: Proc. of the 28th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2015); 2015; p. 361–369.

2. Jeffrey C. An introduction to GNSS: GPS, GLONASS, Galileo and other global navigation satellite systems. NovAtel; 2010.

3. Cheng XJ, Cao KJ, Xu JN, Li B. Analysis on forgery patterns for GPS civil spoofing signals. In: International Conference on Computer Sciences and Convergence Information Technology; 2009; p. 353–356.

4. Jafarnia-Jahromi A, Broumandan A, Nielsen J, Lachapelle G. GPS vulnerability to spoofing threats and a review of antispoofing techniques. International Journal of Navigation and Observation. 2012;2012.

5. Coffed J. The threat of GPS jamming: The risk to an information utility. Report of EXELIS. 2014;6–10.

6. Geiger A, Lenz P, Urtasun R. (are we ready for autonomous driving? the kitti vision benchmark suite). In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012; p. 3354–3361.

7. Brejcha J,Čadík M. State-of-the-art in visual geo-localization. Pattern Analysis and Applications. 2017;20(3):613–637.

8. Piasco N, Sidibé D, Demonceaux C, Gouet-Brunet V. A survey on visual-based localization: On the benefit of heterogeneous data. Pattern Recognition. 2018;74:90–109.

9. Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P, Milford MJ. Visual place recognition: a survey. IEEE Transactions on Robotics. 2015;32(1):1–19.

10. Garcia-Fidalgo E, Ortiz A. Vision-based topological mapping and localization methods: a survey. Robotics and Autonomous Systems. 2015;64:1–20.

11. Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha JM. Visual simultaneous localization and mapping: a survey. Artificial Intelligence Review. 2015;43(1):55–81.

12. Maddern W, Stewart A, McManus C, Upcroft B, Churchill W, Newman P. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In: Proc. of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA); 2014; Vol. 2; p. 3.

13. Maddern W, Vidas S. Towards robust night and day place recognition using visible and thermal imaging. In: RSS 2012 Workshop: Beyond laser and vision: Alternative sensing techniques for robotic perception; 2012; p. 1–6.

14. Lin TY, Belongie S, Hays J. Cross-view image geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2013; p. 891–898.

15. Lin TY, Cui Y, Belongie S, Hays J. Learning deep representations for ground-to-aerial geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; p. 5007–5015.

16. Homer CH, Fry JA, Barnes CA. The national land cover database. US Geological Survey Fact Sheet. 2012;3020(4):1–4.

17. Fry JA et al. Completion of the 2006 national land cover database for the conterminous United States. PE&RS, Photogrammetric Engineering & Remote Sensing. 2011;77(9):858–864.

18. Homer C, Dewitz J, Yang L, Jin S, Danielson P, Xian G, Coulston J, Herold N, Wickham J, Megown K. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing. 2015;81(5):345–354.

19. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR); 2006; Vol. 1; p. 519–528.

20. Westoby MJ, Brasington J, Glasser NF, Hambrey MJ, Reynolds JM. 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology. 2012;179:300–314.

21. Schönberger JL, Frahm JM. Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; p. 4104–4113.

22. Ramalingam S, Bouaziz S, Sturm P, Brand M. SKYLINE2GPS: Localization in urban canyons using omni-skylines. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2010; p. 3816–3823.

23. Gesch D, Oimoen M, Greenlee S, Nelson C, Steuck M, Tyler D. The national elevation dataset. Photogrammetric engineering and remote sensing. 2002;68(1):5–32.

24. Lowe DG et al. Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision (ICCV);1999; Vol. 99; p. 1150–1157.

25. Mikolajczyk K, Schmid C. Scale & affine invariant interest point detectors. International Journal of Computer Vision. 2004;60(1):63–86.

26. Ke Y et al. PCA-SIFT: A more distinctive representation for local image descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2004; p. 506–513.

27. Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012; p. 2911–2918.

28. Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. In: European Conference on Computer Vision (ECCV); 2006; p. 404–417.

29. Calonder M, Lepetit V, Strecha C, Fua P. BRIEF: Binary robust independent elementary features. In: European Conference on Computer Vision (ECCV); 2010; p. 778–792.

30. Leutenegger S, Chli M, Siegwart R. BRISK: Binary robust invariant scalable keypoints. In: IEEE International Conference on Computer Vision (ICCV); 2011; p. 2548–2555.

31. Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005;27(10):1615–1630.

32. Schönberger JL, Hardmeier H, Sattler T, Pollefeys M. Comparative evaluation of hand-crafted and learned local features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 1482–1491.

33. Balntas V, Lenc K, Vedaldi A, Mikolajczyk K. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 5173–5182.

34. Balntas V, Riba E, Ponsa D, Mikolajczyk K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: British Machine Vision Conference (BMVC); 2016; Vol. 1; p. 3.

35. Paulin M, Douze M, Harchaoui Z, Mairal J, Perronin F, Schmid C. Local convolutional features with unsupervised training for image retrieval. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 91–99.

36. Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P, Moreno-Noguer F. Discriminative learning of deep convolutional feature point descriptors. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 118–126.

37. Yi KM, Trulls E, Lepetit V, Fua P. LIFT: Learned invariant feature transform. In: European Conference on Computer Vision (ECCV); 2016; p. 467–483.

38. DeTone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2018; p. 224–236.

39. Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, Sattler T. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019; p. 8092–8101.

40. Rosten E, Drummond T. Machine learning for high-speed corner detection. In: European Conference on Computer Vision (ECCV); 2006; p. 430–443.

41. Rublee E, Rabaud V, Konolige K, Bradski GR. ORB: An efficient alternative to SIFT or SURF.. In: IEEE International Conference on Computer Vision (ICCV); 2011; Vol. 11; p. 2.

42. Alahi A, Ortiz R, Vandergheynst P. FREAK: Fast retina keypoint. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012; p. 510–517.

43. Swain MJ, Ballard DH. Color indexing. International Journal of Computer Vision. 1991;7(1):11–32.

44. Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision. 2001;42(3):145–175.

45. Sünderhauf N, Protzel P. Brief-gist-closing the loop by simple means. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2011; p. 1234–1241.

46. Badino H, Huber D, Kanade T. Real-time topometric localization. In: IEEE International Conference on Robotics and Automation (ICRA); 2012; p. 1635–1642.

47. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2005; p. 886–893.

48. Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004;60(2):91–110.

49. Matas J, Chum O. Robust wide-baseline stereo from maximally stable extremal regions.

50. Hough PV. Method and means for recognizing complex patterns. 1962 US Patent 3,069,654.

51. Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986;(6):679–698.

52. Saurer O, et al. Image based geo-localization in the Alps. International Jour-nal of Computer Vision. 2016;116(3):213–225.

53. Bansal M, Daniilidis K. Geometric urban geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014; p. 3978–3985.

54. Sünderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M. On the performance of ConvNet features for place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2015; p. 4297–4304.

55. Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision (ICCV); 2003; p. 1470.

56. Gálvez-López D, Tardos JD. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics. 2012;28(5):1188–1197.

57. Jégou H, Douze M, Schmid C. On the burstiness of visual elements. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009; p. 1169–1176.

58. Torii A, Sivic J, Pajdla T, Okutomi M. Visual place recognition with repetitive structures. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2013; p. 883–890.

59. Jegou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search. In: European Conference on Computer Vision (ECCV); 2008; p. 304–317.

60. Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2007; p. 1–8.

61. Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2010; p. 3304–3311.

62. Torii A, Arandjelović R, Sivic J, Okutomi M, Pajdla T. 24/7 place recognition by view synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015; p. 1808–1817.

63. Arandjelović R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; p. 5297–5307.

64. Nister D, Stewenius H. Scalable recognition with a vocabulary tree. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR); 2006; Vol. 2; p. 2161–2168.

65. Philbin J, Chum O, Isard M, Sivic J, Zisserman A. Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2007; p. 1–8.

66. Muja M, Lowe DG. Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(11):2227–2240.

67. Zamir AR, Shah M. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(8):1546–1558.

68. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM. 1981;24(6):381–395.

69. Chum O, Matas J, Kittler J. Locally optimized RANSAC. In: Joint Pattern Recognition Symposium; 2003; p. 236–243.

70. Toft C, Stenborg E, Hammarstrand L, Brynte L, Pollefeys M, Sattler T, Kahl F. Semantic match consistency for long-term visual localization. In: European Conference on Computer Vision (ECCV); 2018; p. 383–399.

71. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS); 2012; p. 1097–1105.

72. Babenko A, Lempitsky V. Aggregating local deep features for image retrieval. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 1269–1277.

73. Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations (ICLR); 2016.

74. Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision (ECCV); 2016. p. 241–257.

75. Zheng L, Yang Y, Tian Q. SIFT meets CNN: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;40(5):1224–1244.

76. Sattler T, Torii A, Sivic J, Pollefeys M, Taira H, Okutomi M, Pajdla T. Are large-scale 3D models really necessary for accurate visual localization?

In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 1637–1646.

77. Hartley R, Zisserman A. Multiple view geometry in computer vision. Cambridge University Press; 2003.

78. Nistér D. An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004;26(6):0756–777.

79. Hartley RI. In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997;19(6):580–593.

80. Rocco I, Arandjelović R, Sivic J. Convolutional neural network architecture for geometric matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 6148–6157.

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

82. Besl PJ, McKay ND. Method for registration of 3-D shapes. In: Sensor fusion IV: control paradigms and data structures; 1992; Vol. 1611; p. 586–606.

83. Kendall A, Grimes M, Cipolla R. Posenet: A convolutional network for real-time 6-DOF camera relocalization. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 2938–2946.

84. Kendall A, Cipolla R. Modelling uncertainty in deep learning for camera relocalization. In: IEEE International Conference on Robotics and Automation (ICRA); 2016; p. 4762–4769.

85. Kendall A, Cipolla R. Geometric loss functions for camera pose regression with deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 5974–5983.

86. Walch F, Hazirbas C, Leal-Taixe L, Sattler T, Hilsenbeck S, Cremers D. Image-based localization using LSTMs for structured feature correlation. In: IEEE International Conference on Computer Vision (ICCV); 2017; p. 627–637.

87. Meguro Ji, Murata T, Nishimura H, Amano Y, Hasizume T, Takiguchi Ji. Development of positioning technique using omni-directional IR camera and aerial survey data. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics; 2007; p. 1–6.

88. Baboud L, Čadík M, Eisemann E, Seidel HP. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011; p. 41–48.

89. Baatz G, Saurer O, Köser K, Pollefeys M. Large scale visual geo-localization of images in mountainous terrain. In: European Conference on Computer Vision (ECCV); 2012; p. 517–530.

90. Hays J, Efros AA. IM2GPS: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2008; p. 1–8.

91. Weyand T, Kostrikov I, Philbin J. PlaNet-photo geolocation with convolutional neural networks. In: European Conference on Computer Vision (ECCV); 2016; p. 37–55.

92. Vo N, Jacobs N, Hays J. Revisiting IM2GPS in the deep learning era. In: IEEE International Conference on Computer Vision (ICCV); 2017; p. 2621–2630.

93. Arandjelović R, Zisserman A. DisLocation: Scalable descriptor distinctiveness for location recognition. In: Asian Conference on Computer Vision (ACCV); 2014; p. 188–204.

94. Zamir AR, Shah M. Accurate image localization based on google maps street view. In: European Conference on Computer Vision (ECCV); 2010; p. 255–268.

95. Sattler T, Leibe B, Kobbelt L. Fast image-based localization using direct 2D-to-3D matching. In: IEEE International Conference on Computer Vision (ICCV); 2011; p. 667–674.

96. Li Y, Snavely N, Huttenlocher D, Fua P. Worldwide pose estimation using 3D point clouds. In: European Conference on Computer Vision (ECCV); 2012; p. 15–29.

97. Sattler T, Leibe B, Kobbelt L. Improving image-based localization by active correspondence search. In: European Conference on Computer Vision (ECCV); 2012; p. 752–765.

98. Sattler T, Leibe B, Kobbelt L. Efficient & effective prioritized matching for large-scale image-based localization. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016;39(9):1744–1756.

99. Zeisl B, Sattler T, Pollefeys M. Camera pose voting for large-scale image-based localization. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 2704–2712.

100. Svärm L, Enqvist O, Kahl F, Oskarsson M. City-scale localization for cameras with known vertical direction. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016;39(7):1455–1461.

101. Nadeem U, Jalwana MA, Bennamoun M, Togneri R, Sohel F. Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Clouds. In: International Conference on Neural Information Processing; 2019; p. 222–234.

102. Irschara A, Zach C, Frahm JM, Bischof H. From structure-from-motion point clouds to fast location recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009; p. 2599–2606.

103. Sattler T, Weyand T, Leibe B, Kobbelt L. Image Retrieval for Image-Based Localization Revisited.. In: British Machine Vision Conference (BMVC); 2012; Vol. 1; p. 4.

104. Sattler T, Havlena M, Radenovic F, Schindler K, Pollefeys M. Hyperpoints and fine vocabularies for large-scale location recognition. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 2102–2110.

105. Schönberger JL, Radenovic F, Chum O, Frahm JM. From single image query to detailed 3D reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; p. 5126–5134.

106. Sarlin PE, Cadena C, Siegwart R, Dymczyk M. From coarse to fine: Robust hierarchical localization at large scale. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019; p. 12716–12725.

107. Sattler T et al. Benchmarking 6DOF outdoor visual localization in changing conditions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; p. 8601–8610.

108. Germain H, Bourmaud G, Lepetit V. Sparse-To-Dense Hypercolumn Matching for Long-Term Visual Localization. In: International Conference on 3D Vision (3DV); 2019.

109. Knopp J, Sivic J, Pajdla T. Avoiding confusing features in place recognition. In: European Conference on Computer Vision (ECCV); 2010; p. 748–761.

110. Arandjelović R, Zisserman A. Visual vocabulary with a semantic twist. In: Asian Conference on Computer Vision (ACCV); 2014; p. 178–195.

111. Kobyshev N, Riemenschneider H, Van Gool L. Matching features correctly through semantic understanding. In: International Conference on 3D Vision (3DV); 2014; Vol. 1; p. 472–479.

112. Mousavian A, Košecká J, Lien JM. Semantically guided location recognition for outdoors scenes. In: IEEE International Conference on Robotics and Automation (ICRA); 2015; p. 4882–4889.

113. Schönberger JL, Pollefeys M, Geiger A, Sattler T. Semantic visual localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; p. 6896–6906.

114. Shi T, Shen S, Gao X, Zhu L. Visual Localization Using Sparse Semantic 3D Map. In: IEEE International Conference on Image Processing (ICIP); 2019.

115. Newman P, Cole D, Ho K. Outdoor SLAM using visual appearance and laser ranging. In: IEEE International Conference on Robotics and Automation (ICRA); 2006; p. 1180–1187.

116. Milford MJ, Wyeth GF. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: IEEE International Conference on Robotics and Automation (ICRA); 2012; p. 1643–1649.

117. Cummins M, Newman P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research. 2008;27(6):647–665.

118. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 1851–1858.

119. Wang S, Clark R, Wen H, Trigoni N. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: IEEE International Conference on Robotics and Automation (ICRA); 2017; p. 2043–2050.

120. Viswanathan A, Pires BR, Huber D. Vision based robot localization by ground to satellite matching in GPS-denied situations. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2014; p. 192–198.

121. Shan Q, Wu C, Curless B, Furukawa Y, Hernandez C, Seitz SM. Accurate geo-registration by ground-to-aerial image matching. In: International Conference on 3D Vision (3DV); 2014; Vol. 1; p. 525–532.

122. Workman S, Souvenir R, Jacobs N. Wide-area image geolocalization with aerial reference imagery. In: IEEE International Conference on Computer Vision (ICCV); 2015; p. 3961–3969.

123. Vo NN, Hays J. Localizing and orienting street views using overhead imagery. In: European Conference on Computer Vision (ECCV); 2016; p. 494–509.

124. Zhai M, Bessinger Z, Workman S, Jacobs, N. Predicting ground-level scene layout from aerial imagery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 867–875.

125. Tian Y, Chen C, Shah M. Cross-view image matching for geo-localization in urban environments. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; p. 3608–3616.

126. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS); 2015; p. 91–99.

127. Hu S, Feng M, Nguyen RM, Hee Lee G. CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; p. 7258–7267.

128. Li S, Kang X, Fang L, Hu J, Yin H. Pixel-level image fusion: A survey of the state of the art. Information Fusion. 2017;33:100–112.

129. Sappa A, Carvajal J, Aguilera C, Oliveira M, Romero D, Vintimilla B. Wavelet-based visible and infrared image fusion: a comparative study. Sensors. 2016;16(6):861.

130. Aguilera CA, Aguilera FJ, Sappa AD, Aguilera C, Toledo R. Learning cross-spectral similarity measures with deep convolutional neural networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2016; p. 1–9.

131. Ricaurte P, Chilán C, Aguilera-Carrasco C, Vintimilla B, Sappa A. Feature point descriptors: Infrared and visible spectra. Sensors. 2014;14(2):3690–3701.

132. Johansson J, Solli M, Maki A. An evaluation of local feature detectors and descriptors for infrared images. In: European Conference on Computer Vision (ECCV); 2016; p. 711–723.

133. Aguilera CA, Sappa AD, Toledo R. LGHD: A feature descriptor for matching across non-linear intensity variations. In: IEEE International Conference on Image Processing (ICIP); 2015; p. 178–181.

134. Bonardi F, Ainouz S, Boutteau R, Dupuis Y, Savatier X, Vasseur P. PHROG: a multimodal feature for place recognition. Sensors. 2017;17(5):1167.

135. Firmenichy D, Brown M, Süsstrunk S. Multispectral interest points for RGB-NIR image registration. In: IEEE International Conference on Image Processing; 2011; p. 181–184.

136. Vidas S, Sridharan S. Hand-held monocular SLAM in thermal-infrared. In: International Conference on Control Automation Robotics & Vision (ICARCV); 2012; p. 859–864.

137. Borges PVK, Vidas S. Practical infrared visual odometry. IEEE Transactions on Intelligent Transportation Systems. 2016;17(8):2205–2213.

138. Brown M, Süsstrunk S. Multi-spectral SIFT for scene category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011; p. 177–184.

139. Campo FB, Ruiz FL, Sappa AD. Multimodal stereo vision system: 3D data extraction and algorithm evaluation. IEEE Journal of Selected Topics in Signal Processing. 2012;6(5):437–446.

140. Barrera F, Lumbreras F, Sappa AD. Multispectral piecewise planar stereo using Manhattan-world assumption. Pattern Recognition Letters. 2013;34(1):52–61.

141. Aguilera C, Barrera F, Lumbreras F, Sappa AD, Toledo R. Multispectral image feature points. Sensors. 2012;12(9):12661–12672.

142. Maddern W, Pascoe G, Linegar C, Newman P. 1 year, 1000 km: The Oxford RobotCar dataset. The International Journal of Robotics Research. 2017;36(1):3–15.

143. Badino H, Huber D, Kanade T. Visual topometric localization. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV); 2011; p. 794–799.

144. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ. YFCC100M: The new data in multimedia research. Communications of the ACM. 2016;59(2):64–73.

145. Tolias G, Avrithis Y. Speeded-up, relaxed spatial matching. In: International Conference on Computer Vision (ICCV); 2011; p. 1653–1660.

146. Philbin J, Chum O, Isard M, Sivic J, Zisserman A. Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2008; p. 1–8.

147. Chen DM et al. City-scale landmark identification on mobile devices. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011; p. 737–744.

148. Chen Z, Jacobson A, Sünderhauf N, Upcroft B, Liu L, Shen C, Reid I, Milford M. Deep learning features at scale for visual place recognition. In: IEEE International Conference on Robotics and Automation (ICRA); 2017; p. 3223–3230.

149. Sünderhauf N, Neubert P, Protzel P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In: Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), 2013; p. 2013.

150. Čadík M, Vašíček J, Hradiš M, Radenović F, Chum O. Camera elevation estimation from a single mountain landscape photograph. In: British Machine Vision Conference (BMVC); 2015.

151. Brejcha J,Čadík M. GeoPose3K: Mountain landscape dataset for cam-era pose estimation in outdoor environments. Image and Vision Computing. 2017;66:1–14.

152. Wang S, Bai M, Mattyus G, Chu H, Luo W, Yang B, Liang J, Cheverie J, Fidler S, Urtasun R. TorontoCity: Seeing the world with a million eyes. In: IEEE International Conference on Computer Vision (ICCV); 2017.

153. Carlevaris-Bianco N, Ushani AK, Eustice RM. University of Michigan North Campus long-term vision and lidar dataset. The International Journal of Robotics Research. 2016;35(9):1023–1035.

154. Hao Q, Cai R, Li Z, Zhang L, Pang Y, Wu F. 3D visual phrases for landmark recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012; p. 3594–3601.

155. Li Y, Snavely N, Huttenlocher DP. Location recognition using prioritized feature matching. In: European Conference on Computer Vision (ECCV); 2010; p. 791–804.

156. Crandall D, Owens A, Snavely N, Huttenlocher D. Discrete-continuous optimization for large-scale structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011; p. 3001–3008.

157. Milford M. Vision-based place recognition: how low can you go? The International Journal of Robotics Research. 2013;32(7):766–789.

158. Latif Y, Huang G, Leonard JJ, Neira J. An Online Sparsity-Cognizant Loop-Closure Algorithm for Visual Navigation. In: Robotics: Science and Systems; 2014.

List of Symbols, Abbreviations, and Acronyms

ANN	approximate nearest neighbor
BoW	Bag-of-Words
BRIEF	Binary Robust Independent Elementary Features
BRISK	Binary Robust Invariant Scalable Keypoints
CAD	computer-aided design
CNN	convolutional neural network
COTS	commercial off-the-shelf
C-SWaP	cost, size, weight, and power
DEM	digital elevation model
DOF	degrees-of-freedom
DoG	difference of Gaussians
DSM	digital surface model
DTM	digital terrain model
F2P	Features to Points
FAST	Features from Accelerated Segment Test
FREAK	Fast Retina Keypoint
GMM	Gaussion mixture model
GNSS	Global Navigation Satellite System
HKM	hierarchical k-means
HOG	Histogram of Oriented Gradients
IMU	inertial measurement unit
IR	infrared

LIDAR	light detection and ranging
LIFT	Learned Invariant Feature Transform
LSTM	long short-term memory
LWIR	long-wave infrared
MAC	Maximum Activations of Convolutions
MSER	Maximally Stable Extremal Regions
MVS	multi-view stereo
NIR	near infrared
NN	nearest neighbor
ORB	Oriented FAST and Rotated BRIEF
PCA	Principal Components Analysis
PCA-SIFT	Principal Components Analysis-SIFT
PnP	Perspective-n-Point
RANSAC	Random Sample Consensus
RGB	red, green, blue
RGB-D	red, green, blue, and depth
R-MAC	Regional Maximum Activations of Convolutions
SfM	Structure-from-Motion
SIFT	Scale-Invariant Feature Transform
SLAM	simultaneous localization and mapping
SURF	Speeded Up Robust Features
SWIR	short-wave infrared
VIO	visual-inertial odometry
VLAD	Vector of Locally Aggregated Descriptors
VO	visual odometry

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	CCDC ARL FCDD RLD CL TECH LIB
7 (PDF)	CCDC ARL FCDD RLS RL C MORRIS W NOTHWANG A MATTHIS J CONROY E SHAMWELL J BRODY S LEUNG