Survey on Visual-Based Localization

Andrew Zhou
Princeton University
Princeton, NJ 08544
ajzhou@princeton.edu

Abstract

Visual-Based Localization (VBL) is the computer vision task of retrieving the pose of a camera given a query image that the camera has captured. VBL has received increased attention as a research subject due to heavy reliance of important applications, such as augmented reality, on location knowledge. This paper overviews feature representations, evaluation metrics, and datasets commonly used in VBL literature and surveys recent state-of-the-art methods in VBL, while examining the trade-off between scale and precision in VBL methods.

1. Introduction

The human brain is adept at inferring location and viewpoint from visual information. Looking at a picture of the Eiffel Towel, we can instantly recognize that the picture was taken in Paris and predict what viewpoint the picture was taken from. *Visual-Based Localization (VBL)* is the computer vision task of retrieving the pose, referring to position and possibly orientation, of a camera given an image the camera has taken. An example of a VBL task is to geo-tag outdoor images in a photo gallery without using GPS. Various other names are used in place of VBL in literature. In this survey, as in [21], VBL covers terminologies such as: Image-based Localization, Visual Localization, Structure-Based Localization, Visual Geo-Localization, Camera Relocalisation, Image-Based Pose Estimation, and all other rearrangements of these terms.

VBL has been receiving increasing research attention over the past decade. Important applications, such as augmented reality, indoor navigation, self-driving vehicles, 3D reconstruction, and more, are heavily reliant on location knowledge. GPS-like localization systems alone cannot provide the level of accuracy that the aforementioned applications demand. GPS-like localization systems perform especially poorly in crowded urban environments. Furthermore, the supply of large geo-localized image database and the proliferation of embedded visual acquisition system, such as smart-phone camera, in recent years reshape what is possible to achieve for VBL methods.

Despite the increasing attention VBL has received as a research subject, there is a lack of surveys on this particular computer vision task. This paper serves to fill that void.

This survey focuses on systems designed for city-scale localization as it concerns the most VBL applications. In section 2, we look at other existing VBL surveys and how they differ from our paper in terms of topics covered.

Central to all computer vision tasks is visual data representation. There are various data types such as point, patch, and geometric in VBL. Picking the right data representation, or combination of representations, can help overcome challenges in VBL like viewpoint and illumination changes and make systems more robust to environment appearance changes over time. We take a deeper look at different VBL data representations in section 3.

We examine various VBL datasets in sectin 4 and evaluation metrics in section 6. Then, we discuss current state-of-the-art VBL methods in section 4 and categorize them in terms of indrect and direct localization methods, which excel in scale and precision, respectively.

The trade-off between scale and precision is central to the VBL narrative, currently. In section 7, we examine future trends that could potentially lead to harmony between scale and precision while improving performance in both areas simultaneously.

2. Related Work

We compare and contrast our work to existing VBL surveys and briefly discuss a different task, *Visual Place Recognition*, which has methods that intersect with ones in VBL.

2.1. Existing Surveys

[6] presents several papers on VBL and classify them depending on the environment for which the particular method acts on. They divide environment criterion into three classes:

- *Global:* unrestricted visual-based localization at the planet scale
- *City-Scale:* Visual-based localization in urban environments
- *Natural:* Visual-based localization in non-urban environments

VBL systems for different types of environment demand different data representations and act on different datasets. In contrary to [6], our paper primarily focuses on city-scale.

[29] creates a selection of recent articles that help paint the larger landscape of VBL. They organize the articles in terms of three categories: data-driven geo-localization, semantic reasoning based geo-localization, and geometric matching based geo-localization. We present VBL methods in section 4 in a similar manner.

[21] is a survey that is most similar to ours. They similarly focus on city-scale VBL methods. However, we survey different and more recent articles and highlight VBL systems that generalize well on the mobile platform.

2.2. Visual Place Recognition

Visual Place Recognition is a roboticist problem that captures the visual ability of a human or robot to recognize and already visited place [18].Visual Place Recognition and VBL differ in goals. While Visual Place Recognition is interested in deciding if a given place have already been seen, VBL produces an output of position and possibly orientation of the visual acquisition system. Nevertheless, methods in Visual Place Recognition share similarity with methods in VBL. Studying approaches in Visual Place Recognition can give a better panorama of methodolgies involved in localization process with visual data.

3. Data Representation

We represent visual data with features. Features should incorporate as much discriminant information as possible and should be fast to compute and compare. We group visual data representation into local, global, and hybrid feature types.

3.1. Global Features

Description of global features considers the image as a whole and produces a high dimensional output. Advances in Convolutional Neural Networks (CNN) greatly increases efficiency of computing global descriptor.

3.1.1 Hand-crafted features

The most common hand-crafted global descriptor is GIST, described in [20]. GIST uses a set of perceptual dimensions, such as roughness, openness and ruggedness, that represent the dominant spatial structure of a scene and estimate these dimensions reliably using spectral and coarsely localized information.

Specific information regarding object shape is not required for localization problems. Modeling a holistic representation of the scene can also inform about its associated camera pose.

3.1.2 Learned features

With the rise of deep learning, learned feature becomes increasingly popular in localization research. For example, [2] demonstrates that deep learning has led to state-of-the-art techniques in image retrieval for urban scenes using learned features. Extracting learned features often involves extracting the output of a specific convolutional layer in a CNN trained for image classfication. [4] shows that CNN-generated descriptors used as global features are lightweight and can be computed efficiently.

3.2. Local Features

Description of local features occur at pixel level among a local neighborhood of points in an image. Local features have been shown to be well suited to matching and recognition as well as to many other applications as they are robust to occlusion, background clutter and other content changes. The difficulty is to obtain invariance to viewing conditions.

3.2.1 Point features

Selection of points features are dependent on scale, orientation, illumination invariance, computational cost, and descriptor vector dimension. The most used point feature in VBL is the Hessian-affine [19] detector combined with SIFT [17] descriptor.

3.2.2 Geometric features

Geometric features refer to primitive geometric shapes and include semantically meaningful information. For example, [2] shows that vertical lines can act as descriptors to represent buildings in urban environments. Geometric data is often used in conjunction with 3D data. For example, geometric features such as normal vectors and planar surfaces are often used for tracking and localization in augmented reality.

3.2.3 Point features with geometric relations

A limitation to point features in an image is the lack of geometric consistency. [16] proposes a solution to this limitation through geometric association of points. Geometric relation between point features can help eliminate false

matches and outliers in image retrieval methods that use feature-based similarity association algorithms.

3.3. Hybrid Features

Hybrid features consist of features that are neither local nor global and features that combine several types of descriptors.

3.3.1 Patch features

Patch features consider regions of interests in an image. For example, [23] uses a region proposal network to extract regions of interest. To generate region proposals, they slide a small network over the convolutional feature map output by the last shared convolutional layer. This small network takes as input an $n \times n$ spatial window of the input convolutional feature map. Each sliding window is then mapped to a lower-dimensional feature.

3.3.2 Combined features

[5] uses an approach to pair various descriptors in order to increase the result of the retrieval step depending on the targeted dataset.

4. Datasets

This section examines datasets that are used to measure and compare existing and novel approaches to VBL. There are two main types of datasets: Image-based datasets and stucture from motion (SfM) datasets.

For image-based datasets, we discuss the Google Maps Street View dataset, the IM2GPS dataset, the YFCC100M dataset, the San Francisco Landmark dataset, and the Alps100K dataset.

For SfM datasets, we discuss the Rome16K dataset, the Dubrovnik6K dataset, the Quad dataset, the Landmark 3D dataset, and the Cambridge Landmarks dataset.

4.1. Image-based datasets

The Google Maps Street View dataset in [30] contains 102K images taken from the Google Maps Street View site. The images capture scenes from Pittsburgh, PA and Orlando, FL. The dataset contains full 360 degree panoramic images with distance of about $12m$ between consecutive locations.

The San Francisco Landmark dataset in [8] is a database containing 1.7 million images of buildings in San Francisco with ground truth labels, geo-tags, and calibration data. The dataset also includes a difficult query set of 803 cell phone images taken with a variety of different camera phones. The generation process utilizes vehicle-mounted cameras with wide-angle lenses to capture spherical panoramic images. For all visible buildings in each panorama, a set of overlapping perspective images is generated. The dataset is created with the intention of facilitating further research in landmark recognition with mobile devices.

YFCC100M in [25] is a dataset with 100 million images and distributed via Amazon AWS so that it is public accessible. Of the 100 million images in YFCC100M, there are exactly $48,366,323$ photos and $103,506$ videos that have been annotated with a geographic coordinates, either manually by the user or automatically via GPS. The annotated images and videos are crowd-sourced and cover major cities including London, Paris, Tokyo, New York, San Francisco, and Hong Kong. Overall, the dataset spans 249 different territories in the world.

The IM2GPS dataset in [11] contains roughly 6 million geo-tagged images downloaded from Flickr. The dataset is generated by searching images with both GPS coordinates and geographic keywords, which has a higher likelihood of being accurately geo-located and visually useful data than images with one or none of the two attributes. The dataset averages 0.0435 pictures per square kilometer of Earth's land area, but distribution skews heavily towards places where people live or travel. Query images are likely to come from the same places the database images cover.

The Alps100K dataset in [7] contains nearly 100K annotated outdoor images from mountain environments. Annotations include GPS coordinates, elevation, and EXIF if available. The images exhibit high variation in elevation, ranging from 0 to $4782m$, as well as in landscape appearance. The images also span all seasons of the year. The database is generated by first creating a list of all hills and mountain peaks located in the seven Alpine countries, then querying the list on Flickr.

4.2. SfM datasets

The Rome16K and Dubrovnik6K datasets in [15] contain 3D reconstructed models of some of the most notable landmarks in Rome and Dubrovnik. The suffix of each dataset indicates how many images are used to generate the corresponding SfM models. Specifically, the Rome dataset has 3D models for 69 different sites and have a total of $4,312,820$ 3D points generated from images taken by distinct cameras. The Dubrovnik dataset only has 3D model for 1 landmark but contains $2,208,646$ million 3D points.

The Landmark3D dataset in [10] provides a collection of web images and 3D reconstructed models for research on landmark recognition. It serves as a useful benchmark for evaluating and comparing different methods meaning to operate on 3D models. The dataset is evolving and currently contains 3D models for 25 landmarks generated by operating SfM on $45,180$ web images. Each 3D reconstructed model is generated from about 1.4K to 2K images. A estimate of 2.7 million total 3D points are included in the 3D

models. In addition, there are also 58 million SIFT features registered to the 3D models, which can be used to evaluate 2D-to-3D localization methods.

The Cambridge Landmarks dataset in [13] contains 12K images with full 6-DoF camera poses generated using SfM. The dataset provides data to train and test pose regression algorithms in a large scale outdoor urban setting. The collected data contain urban clutter such as pedestrians and vehicles and have varying lighting and weather conditions because the images capture scenes at different points in time. Train and test images are taken from distinct walking paths and not sampled from the same trajectory making the regression challenging.

The Quad dataset in [9] contains a 3D reconstructed model of the Arts Quad at Cornell University. This dataset also contains $6,514$ images of the Arts Quad. Of the $6,514$ images, $5,000$ of them are recorded by an iPhone 3G camera and geo-tagged, while 348 images have precise GPS coordinates measured using service-quality differential GPS that can be used for ground truth during training and evaluation.

5. Evaluation Metrics

In this section, we examine three evaluation metrics, commonly employed in VBL literature: Percentage of localization error, top-k candidates, precision/recall.

5.1. Percentage of localization error

The percentage of localization error metric measures the number of queries that are localized with error of a certain threshold. It is a straightforward way to illustrate how accurate a particular localization method is by showing the number of query images in different error ranges. Plotting localization error, referring to distance bewteen estimated location abd ground truth, against percentage of images that are localized within the same error range is a popular visualization to perform. Comparing performance of different localization methods is intuitive using this metric.

5.2. Top-k candidates

When a localization method returns an ordered list of candidate locations for each query image, the top-k candidates metric counts how many query images are localized correctly within k top-ranked candidates. Normally, k is set to 10 or 1%. If database images contain geo-tags, it is common practice to decide that a query image is correctly localized if 1 out of k candidates lies inside a tolerance radius circle. Plotting a variable number of k candidates against the fraction of correctly localized images is a popular visualization for this evaluation metric.

5.3. Precision/Recall

Precision and recall are standard metrics used for evaluation of classification and retrieval methods. The standard formula for precision is

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

and the standard formula for Recall is:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

.

Image retrieval is a problem commonly solved as part of a VBL problem. Precision is a particularly useful metric that indicates the percentage of query images that are correctly localized within a certain tolerance.

6. Methods

We present VBL methods using two broad categories: indirect localization and direct localization. A trade-off between scale and precision exist between the two categories. Indirect localization methods take advantage of internet-scale image databases and provide a coarse camera pose estimation by solving an image retrieval problem. On the other hand, direct localization methods regress the precise 6-DoF pose, but often at the sacrifice of area coverage. The are two main subcategories under direct localization methods: feature-based and learning-based methods.

6.1. Indirect Methods

Indirect methods generally cast the localization task as an image retrieval problem. It provides a coarse camera pose estimation for the query image, but runs efficiently and has large area coverage. Generally, there are three steps involved in indirect localization methods. The first step is generating a numeric description, or descriptor, for the query image and for each database image which contains spatial information, such as GPS coordinates. Then, in the second step, indirect methods perform similarity association between query image and database images through descriptors. The third step involves post-processing and selecting the best-matched candidate.

[30] presents a classical indirect method that performs localization using a structured data set of about 100K images downloaded from the Google Maps Street View site. The system finds the camera location of a query image with a precision comparable to hand-held GPS devices. For feature descriptor, they use SIFT. They index the SIFT descriptors of the detected SIFT interest points in the database images using a tree. To localize a query image, they compute the SIFT descriptors for the query image and query the aforementioned tree. They use a novel GPS-tag-based pruning method to remove less reliable descriptors. Then, they

utilize a smoothing step with an associated voting scheme to allow each query descriptor to vote for the location its nearest neighbor belongs to. A parameter called Confidence of Localization, which is based on the Kurtosis of the distribution of votes, is defined to determine how reliable the localization of a particular image is.

In addition to localizing a single query image, [30] also proposes a novel approach to simultaneously localize groups of images in a hierarchical manner. First, they attempt to localize each image in the group individually. Then, the rest of the images in the group are matched against images in the neighboring area of the found first match. The final location of the group is determined based on the Confidence of Localization parameter. The proposed image group localization method can deal with very unclear queries which are not capable of being localized individually.

The approach used by [30] can efficiently localize the camera pose using a captured image by querying a descriptor tree built during training time. This method can scale smoothly with more input training images, annotated with GPS coordinates. However, this system only obtains a precision comparable to hand-held GPS devices.

[9] proposes an indirect method that takes advantage of significantly larger datasets. Rather than using merely 100K images downloaded from the Google Maps Street View site, Crandall et al. study the indirect localization problem by classifying landmarks using a much larger dataset of 30 million images. Given a query image, they localize its camera by classifying the query image as a known landmark, then retrieving the relevant coordinates. The dataset and categories of landmarks are formed automatically using geotagged photos from Flickr by looking for peaks in the spatial geo-tag distribution corresponding to frequently photographed landmarks. To handle such a large scale dataset, they use a multi-class SVM classifier using SIFT-based bag-of-word features to classify query images. Through extensive experiments, they demonstrate that their system has a classification accuracy comparable to that of humans on the same task. They also find that using a structured SVM to classify the stream of photos taken by a single camera, rather than classifying individual photos, yields dramatic improvement in the classification rate. Streams of photos provide temporal context and is just one kind of potential contextual information that the system can extract from photo sharing sites. When image-based classification results are combined with text features from tagging, classification accuracy can be hundreds of times the random guessing baseline.

The results of [9] demonstrate the potential for indirect localization using Internet-scale image collections and large labeled datasets. The drastic improvement in accuracy created by incorporating other inputs foreshadow our conclusion that to balance scale and precision, VBL need to consider additional forms of inputs. The aforementioned two systems are extremely scalable but only consider indirect localization of outdoor images.

With localization of indoor images in mind, [14] detail a marker-based indoor navigation system. An alternative method to the proposed system is to establish indoor locations through signal strength of radio frequency (RF) bands of the IR echo distance as in [28]. Such RFID-based systems characterize locations by measuring signal strength using an RF sensor. Extending upon the idea, the proposed system characterizes a location through a marker, color information, and prior knowledge. Each marker has a black and white colored square with a unique pattern. Topographical information of the indoor environment is established using prior knowledge of location and is represented in the location model implemented using a hierarchical tree structure. Kim et al. use a wearable mobile PC with camera to obtain an image sequence, which is then sent to remote PCs for processing. The remote PCs perform marker detection, image sequence matching, and location recognition. To improve performance, they use an adaptive thresholding method to detect markers under illumination changes and use the location model to reduce execution time during image sequence matching. The system achieves an average location recognition success rate of 89%. Compared to RFID-based systems, the proposed system is a more economical solution and is not limited by signal propagation and multiple reflections.

Contrary to most indirect methods, the system proposed in [14] does not scale well. Their system functions only when pre-constructed markers are placed in indoor environments. Furthermore, such a system requires constant communication between a mobile device and remote servers. Latency problems can seriously affect performance. Nevertheless, the system proposed by Kim et al. is one of the first attempts to solve the indoor navigation problem.

6.2. Direct Methods

Unlike indirect methods, direct localization methods compute the precise 6-DoF camera pose, but often times sacrifice efficiency and area coverage. Direct methods retrieve the absolute camera pose of a query image according to a known representation, which is pre-generated by mapping modules such as SfM or SLAM.

We first take a look at methods that generate a 3D representation of space using SfM. Then, we look at systems that computes absolute pose through feature-based 2D-to-3D matching. Lastly, we examine end-to-end learning-based localization systems.

6.2.1 Structure from motion

SfM is a time-consuming task. [27] proposes a method to reduce time consumption of SfM to near linear. Time complexity of incremental SfM is often known as $O(n^4)$ with respect to the number of cameras. Wu proposes a novel bundle adjustment strategy that provides a good balance between speed and accuracy. He shows that SfM requires only $O(n)$ time on most major steps through extensive experiments. The proposed method maintains accuracy by regularly re-triangulating feature matches that initially failed to triangulate.

Not only time consuming, SfM is also computationally expensive. [1] is a classical paper that discusses 3D scene reconstruction using SfM techniques. Agarwal et al. reconstruct 3D scenes from extremely large collections of photographs found by searching a given city (e.g., Rome) on Internet photo sharing sites, such as Flickr. 3D reconstruction is computationally expensive. Hence, the system focuses on maximizing parallelism at each stage in the pipeline and minimizing serialization bottlenecks. They experiment with a variety of alternative algorithms at each state of the pipeline and report on which ones work best in a parallel computing environment. The paper achieves reconstructing cities consisting of 150K images in less than a day on a cluster of 500 compute cores.

6.2.2 Feature-based

To localize after generating a 3D model, one approach is to find correspondence between 2D features of a RGB query image and 3D points in the reconstructed model. [24] proposes such a system designed for 2D-to-3D matching. They use a direct matching framework based on visual vocabulary quantization and a prioritized correspondence search. In the pipeline, the system first computes the 2D features of an input RGB image. Then, it associates the features individually with visual words pre-computed during training time. Since the 3D points in the reconstructed model are generated from RGB images, the 3D points are also associated with the visual words. The system performs linear search to find the best correspondence between the input features and the points associated with the visual words. At last, the system computes an absolute pose using Random sample consensus (RANSAC).

Extending direct localization to "worldwide" scale, [15] addresses the problem of determining where a photo was taken by estimating a full 6-DoF-plus-intrinsics camera pose with respect to a very large geo-registered 3D point cloud. Their method directly establishes correspondence between 2D features of a query image and 3D points in a point cloud created by running SfM on over 2 million images. Their dataset has over 800K reconstructed images and more than 70 million 3D points, covering hundreds of distinct places around the globe. They propose two new techniques. The first is the use of statistical information about the co-occurrence of 3D model points in images to yield an improved RANSAC scheme. The second is a bidirectional matching algorithm between 3D model points and image features.

Implementing a direct localization system that is more invariant to appearance changes than previous work, [3] describes a system that matches a given architectural drawing, painting, or historical photograph to a 3D model of the corresponding site. The task is difficult as the appearance and scene structure in 2D depictions can be very different from the appearance and geometry of the 3D model. Factors such as specific rendering style, drawing error, age, lighting or change of seasons associated with the input image create immense challenges for the task. In addition, a hard search problem exists. The number of possible alignments of the input image to a set of 3D models from different architectural sites is very large. They develop a compact representation of complex 3D scenes to address this hard search problem. 3D models of several scenes are represented by a set of discriminative visual elements that are automatically learned from rendered views. Through experiments, the paper shows that visual elements learned in a discriminative fashion can reliably find its match despite large variations in rendering style and structural changes of the scene. The proposed approach can identify the correct architectural site as well as recover an approximate viewpoint of historical photographs and paintings with respect to the 3D model of the site.

6.2.3 Learning-based

A classical deep learning pose regression paper, [13] presents a robust and real-time monocular 6-DoF relocalization system. Their proposed system trains a CNN to regress the 6-DoF camera pose from a query RGB image in an end-to-end manner. No additional engineering or graph optimization is needed. The system operates for both indoors and outdoors in real time, and only takes about $5ms$ per frame to compute. Furthermore, it obtains approximately $2m$ and 3-degree accuracy for large scale outdoor scenes and $0.5m$ and 5-degree accuracy indoors. They demonstrate their concept using an efficient 23 layer deep CNN and show that CNNs can be used to solve complicated out of image plane regression problems. Specifically, they leverage transfer learning from large scale classification data. The system, called PoseNet, localizes from high level features and is robust to difficult lighting, motion blur, and different camera intrinsics where point based SIFT registration fails. At last, they show how the pose feature that is produced generalizes to other scenes allowing them to regress pose with additional training examples.

PoseNet can generate a more precise output than what hand-held GPS devices can achieve. It also runs efficiently at test time because of its end-to-end nature. The bottleneck for PoseNet is training data. For every new place, there needs to be a complete coverage of the area using images that are labeled with precise camera poses. For the system to function, the distance gap between images needs to be small. Generating training images through crowd-sourcing is often times insufficient under such a constraint. For areas that are not covered by training images, the system would simply produce a false output, because learning does not help the neural network to localize unknown areas.

Most learning-based localization systems do not fine-tune CNN models, however [22] shows that fine-tuning can improve localization results. Specifically, [22] proposes an unsupervised fine-tuning of CNN for image retrieval and uses this fine-tuning method to achieve state-of-the-art localization results for Oxford buildings and Paris datasets. The system bypasses the dense manual annotations necessary for CNNs by using a SfM generated 3D model to guide the selection of training data for CNN fine-tuning. They show that both hard positives and hard negative examples enhance the performance in image retrieval task by enhancing the derived image representation. Compared to previous supervised approaches, the variability in training data selected from 3D reconstructions shows superior performance.

Beyond fine-tuning CNNs for localization, there are also Learning-based localization methods that utilize deep learning networks besides CNNs. [26] proposes a combined CNN and LSTM architecture to regress camera pose for both indoor and outdoor scenes. The purpose of CNN is to learn suitable feature representations for localization that are robust against motion blur and illumination changes. LSTM, on the other hand, operates on the CNN output to capture contextual information. The combined architecture leads to improvement in localization performance by enlarging the receptive field of each pixel. Walch et al. present a new large-scale indoor dataset with ground truth generated by a laser scanner. Experimental results show that the system can localize images in hard conditions such as in the presence of mostly textureless surfaces

7. Future Trends

We predict four future trends in the research area of VBL: greater availability of geometric data, more focus on direct localization methods, increasing usage of deep learning systems in VBL, and growing interest in using additional inputs to balance and improve VBL precision and scale.

7.1. Availability of geometric data

As shown in [27] and [1], SfM is becoming increasingly efficient and less computationally expensive. As geotagged images on Internet photo-sharing platforms continue to grow, the number of 3D models for landmarks worldwide would proliferate. Furthermore, with the introduction of mobile phones equipped with depth-sensing camera such as the Google Tango phone, we are expecting to see more use of RGB-D data in VBL. RGB-D data can significantly reduce the complexity of reconstructing 3D models because of the added depth dimension.

7.2. Focus on direct methods

A greater availability of geometric data would inevitably lead to a greater focus on improving direct localization methods. Direct localization methods would also receive greater attention than indirect localization methods because of the importance of precision. Applications such as augmented reality or indoor navigation demands high precision. An error range greater than $10m$, commonly seen in localization results of indirect methods, is unacceptable. The trade-off between precision and scale would also begin to fade as availability of geometric data increases, putting direct localization methods in a more favorable light.

7.3. Deep learning in VBL

CNNs and other deep-learning networks are improving in performance for most computer vision tasks, and VBL is no exception to this trend. After the introduction of [13], a growing amount of literature in VBL now use a learning-based localization approach. Like learned features, learning-based localization systems are efficient and lightweight. They produce precise results without expensive computations or heavy time consumption. The only bottleneck for learning-based approaches is training data, which also translates into a problem of low area coverage. However, with the increasing availability of geometric data, there could soon be answers to the training data generation problem.

7.4. Additional inputs

A important dichotomy exists in VBL, which is the trade-off between scale and precision. In this survey, we have examined indirect localization methods and direction localization methods, which excel in scale and precision, respectively. The increasing availability of geometric data may bring a balance to scale and precision. However, in the mean time, it is important to consider using additional forms of input to achieve such a balance. For example, [12] combines inertial measurements, camera sensor data, and depth data for localization. Furthermore, if mobile phones have access to accurate heading information, for instance,

pose estimation can be drastically simpler by disregarding candidates with the wrong orientation in a direct localization system that utilizes 2D-to-3D matching.

8. Conclusion

VBL is a growing research area that can benefit important applications such as augmented reality and indoor navigation. We first overview data representations, evaluation metrics, and datasets commonly seen in VBL literature. Then, we examine specific VBL methods that fall under two broad categories, indirect and direct, which excel at scale and precision, respectively. After discussing current state-of-the-art VBL methods, we analyze future trends in VBL that would lead to a balance and improvement in scale and precision.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. *ICCV*, 2009.

[2] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit. Instant outdoor localization and slam initialization from 2.5d maps. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 21(11):13091318, 2015.

[3] M. Aubry, B. Russell, and J. Sivic. Visual geo-localization of non-photographic depictions via 2d3d alignment. *Large-Scale Visual Geo-Localization*, 21(11):255–275, 2016.

[4] A. Babenko and V. Lempitsky. Aggregating deep convolutional features for image retrieval. *ICCV*, 2015.

[5] N. Bhowmik, L. Weng, V. Gouet-Brunet, and B. Soheilian. Cross-domain image localization by adaptive feature fusion. *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[6] J. Brejcha and M. adk. State-of-the-art in visual geo-localization. *Pattern Analysis and Applications*, 1(2):20, 2017.

[7] M. Cadik, J. Vasicek, M. Hradis, F. Radenovic, and O. Chum. Camera elevation estimation from a single mountain landscape photograph. *British Machine Vision Conference*, 2015.

[8] D. Chen, G. Baaz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Back, M. Pollefeys, B.Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. *IEEE*, pages 737–744, 2011.

[9] D. Crandall, Y. Li, and S. Lee. Recognizing landmarks in large-scale social image collections. *Pattern Analysis and Applications*, 1(2):20, 2015.

[10] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu. 3d visual phrases for landmark recognition. *IEEE*, 2012.

[11] J. Hays and A. Efros. Im2gps: Estimating geographic information from a single image. *IEEE*, 2008.

[12] E. S. Jones and S. Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approah. 2010.

[13] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. *ICCV*, 2015.

[14] J. Kim and H. Jun. Vision-based location positioning using augmented reality for indoor navigation. *IEEE Transactions on Consumer Electronics*, 54(3):954–962, 2008.

[15] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. *ECCV*, 2012.

[16] Z. Liu and R. Marlet. Virtual line descriptor and semi-local matching method for reliable feature correspondence. *British Machine Vision Conference (BMVC)*, page 1116, 2012.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, page 91110., 2004.

[18] S. Lowry, N. Snderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions and Robotics (ToR)*, pages 600–613, 2016.

[19] K. Mikolajczyk and C. Schmid. Visual place recognition: A survey. *International Journal of Computer Vision (IJCV)*, page 6386, 2004.

[20] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelop. 2001.

[21] N. Piasco, D. Sidib, C. Demonceaux, and V. Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogenous data. 2017.

[22] F. Radenovic, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. *ECCV*, 2016.

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2016.

[24] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. *ICCV*, 2011.

[25] B. Thomee, G. F. D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[26] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization with spatial lstms. 2016.

[27] C. Wu. Towards linear-time incremental structure from motion. *International Conference on 3D Vision*, 2013.

[28] Z. Xiang, S. Song, J. Chen, H. Wang, J. Huang, and X. Gao. A wireless lan-based indoor positioning technology. *IBM Journals Research and Development*, 48(5/6):617–626, 2004.

[29] A. R. Zamir, A. Hakeem, L. V. Gool, M. Shah, and R. Szeliski. Large-scale visual geo-localization. *Advances in computer vision and pattern recognitoin*, 1(2), 2016.

[30] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. *ECCV*, 2010.