Semantic Signatures for Urban Visual Localization

Li Weng*†
*School of Automation
Hangzhou Dianzi University
310018 Hangzhou, China

Bahman Soheilian and Valérie Gouet-Brunet
†LASTIG MATIS
Univ. Paris-Est, IGN, ENSG
94160 Saint-Mande, France

***Abstract*—Visual localization is a useful alternative to standard localization techniques. In a typical scenario, features are extracted from images captured by cameras and compared with geo-referenced databases. Location information is then inferred from the matching results. Conventional schemes mainly use low-level visual features. They offer good accuracy but suffer from scalability issues. In order to assist localization in large urban areas, this work explores a different path by utilizing high-level semantic information. It is found that object information in a street view can facilitate localization. A novel descriptor scheme called "semantic signature" is proposed to summarize this information. A semantic signature consists of type and angle information of visible objects at a spatial location. Several metrics and protocols are proposed for signature comparison and retrieval. They illustrate different trade-offs between accuracy and complexity. Extensive simulation results confirm the potential of the proposed scheme in large-scale applications.**

I. Introduction

Visual localization [1], [2] is an alternative to conventional signal-based positioning solutions. It can be used for automatic navigation [3] and location-related multimedia applications [4], [5], such as landmark recognition and augmented reality. The goal of visual localization is to infer where an image is taken by matching it with a database of geo-referenced information. The problem is typically modelled as an image feature retrieval scenario, and solved by (approximate) nearest neighbor search. More specifically, features are extracted from a query image and compared with features in a database; the location is inferred from e.g. the best matches. Depending on the required accuracy, there are mainly two tasks: 1) place recognition and 2) camera pose estimation. The former estimates the zone where the image was acquired; the latter estimates the pose up to six degrees of freedom (6-DOF). Conventional schemes typically achieve these tasks using low-level visual features, such as SIFT [6]. Related research mainly focuses on accuracy and efficiency. Various efforts have been devoted to database indexing and query strategies [7], [8], [9], [10]. They offer good accuracy but suffer from scalability issues due to large amounts of data.

This work studies a novel approach for visual localization. Instead of low-level features, high-level *semantic features* are exploited. Semantic features are related to what humans see in the environment. For example, in dense urban areas, one can typically see buildings, cars, trees, etc. It is found that such information can facilitate localization too. Compared with conventional visual features, semantic features have several advantages. First, they can be encoded in a compact way and require much less storage. Second, they can be efficiently obtained from geographic information systems (databases), such as OpenStreetMap. In particular, we focus on *semantic objects* which are static and widely available in urban areas. It is assumed that such objects can be detected by object detection algorithms from street-view images. Once they are detected, localization can be achieved by retrieving locations with similarly distributed objects from a database.

In our application scenario, the goal is to achieve urban localization using street-view images captured by a mobile device. Given a query image, a complete work flow consists of a series of steps: 1) perform feature detection; 2) narrow down the search scope using semantic features; 3) optionally, retrieve similar images or low-level visual features; 4) perform place recognition or pose estimation. This paper covers the second step, which focuses on the representation, indexing, and matching of semantic information.

Specifically, we propose to use type and angle information of some street objects (called semantic objects) for localization. This approach has not been widely considered. It can be used alone or as a complement to other localization techniques to improve accuracy and efficiency. We model our problem as string matching and solve it with a retrieval framework. Compared with conventional approaches, the advantages of the proposed scheme include: 1) small database size; 2) large coverage area; 3) fast retrieval speed. As a localization method, our approach achieves coarse localization; as a complement to other localization methods, it can effectively reduce the search scope by filtering out irrelevant regions.

The contribution of this paper is multi-fold. First, localization by semantic information is relatively new. Second, a novel descriptor "semantic signature" is proposed to summarize semantic objects. Third, suitable metrics and protocols are proposed for signature matching and retrieval. Promising results from extensive experiments indicate the potential of our approach in large-scale applications.

II. Related work

Existing work on visual localization can be mainly divided into two categories: feature point retrieval and image retrieval. In the former approach, place recognition and camera pose estimation are solved by point-based matching and voting.

This work was funded by the European THINGS2DO project (KET ENIAC JU grant agreement No 621221).

For example, Schindler et al. propose a city-scale place recognition scheme [7]. They use a vocabulary tree to index SIFT features with improved strategies for tree construction and traversal. Irschara et al. [11] also use a vocabulary tree for sparse place recognition using 3D point-clouds. They not only use real views, but also generate synthetic views to extend localization capability. Li et al. [8] address city-scale place recognition and focus on query efficiency. They prioritize certain database features according to a set covering criterion, and use a randomized neighborhood graph for data indexing and approximate nearest neighbor search. Zamir and Shah [12] use Google street-view images for place recognition. They distinguish single image localization and image group localization, and derive corresponding voting and post-processing schemes for refined matching. Chen et al. [13] study the localization of mobile phone images using street-view databases. They propose to enhance the matching by aggregating the query results from two datasets with different viewing angles. Sattler et al. [9] propose to accelerate 2D-to-3D matching by associating 3D points with visual words and prioritizing certain words. Li et al. [10] consider world-wide image pose estimation. They propose a co-occurrence prior based RANSAC and bidirectional matching to maintain efficiency and accuracy.

The other category of localization techniques is based on image retrieval. Conventionally, this is only used for place recognition [5], [14] or geolocation [15]. For example, Zamir and Shah [16] propose multiple nearest neighbor feature matching with generalized graphs. Arandjelovic and Zisserman [17] propose an improved bag-of-features model. Torii et al. [18] apply the VLAD descriptor [19] to synthesis views. Arandjelovic et al. [20] extend VLAD with a deep neural network architecture. Iscan et al. [21] propose to aggregate descriptors from panoramic views. Since 3D models can be built from 2D images with structure-from-motion techniques [22], it is possible to directly estimate 6-DOF with an image database. Recently, Song et al. [23] propose to estimate 6-DOF after image retrieval. Sattler et al. [24] show experimentally that image retrieval approaches are perhaps more suitable for large-scale applications.

Our approach is different from existing work, because we use semantic information above visual information. A relevant idea can be found in [25], where Ardeshir et al. use existing knowledge of objects to assist object detection. While they show the potential of semantic objects in localization, we perform more extensive study in this paper. We also find that edit distance works better than their histogram based metric. Another related scheme is [26], where Arth et al. use a different kind of semantic information. They perform re-localization by extracting straight line segments of buildings from a query image and comparing with a database. While our work focuses on objects, it can also extend to include other semantic features. On the other hand, our approach can also be used as an initial step together with some existing work.

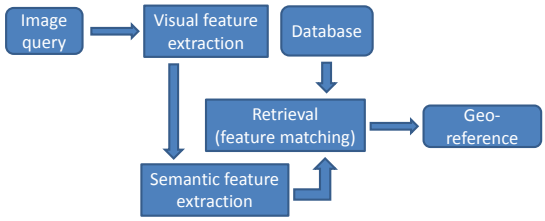


Fig. 1. The application scenario.

III. The Proposed Scheme

The target application is localization in urban environments. In a typical scenario, a user has a mobile device that captures images of the surrounding area. The goal is to tell the user's location according to these images. In a retrieval-based approach, it is tackled by extracting information from the images and comparing with a geo-referenced database. Figure 1 illustrates the application scenario. A critical question here is what kind of information to extract from images. In this work, the focus is semantic information, which is high-level information based on human perception. In our context, it is about what people see from images. For example, people can tell their location by describing their surroundings. The same principle can be applied to localization. Since the images taken by the mobile device are typically street views, the semantic information contains objects such as buildings, streets, the sky, the ground, cars, humans, etc. It is found that some of these objects are useful for localization. In general, *semantic objects* with the following properties are of particular interest:

- Permanent – the object does not move;
- Informative – the object is distinguishable from others;
- Widely available – the object is distributed in the scene.

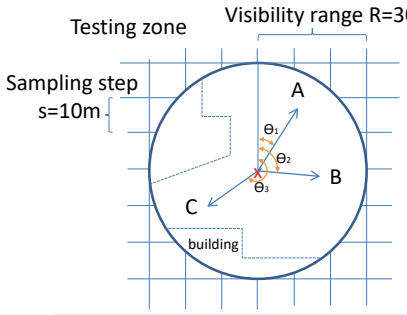
Additionally, the objects should have unambiguous locations and be suitable for object detection algorithms. In this paper, we assume that detecting such objects is feasible and focus on retrieval aspects.

A. Semantic signatures

Once semantic objects have been detected, they are encoded into a compact representation, which we call *semantic signature*. A semantic signature describes some properties of the corresponding objects. It is required to be compact and easily indexable. In this work, we compose a semantic signature by:

- Object type – the category (class) of an object;
- Object angle – the relative angle of an object.

Specifically, the object type is a label, denoted by t; the object angle is measured according to the north and a view point, denoted by a. Given a view point coordinate (x, y) and a visibility range R, each location can be associated with a semantic signature, which is related to the semantic objects that can be seen from that location. In our implementation, semantic objects are identified by a panoramic sweep in a clockwise order. A semantic signature is the concatenation of two parts: $s = \{s^{(1)}|s^{(2)}\}$, where $s^{(1)} = t_1|\cdots|t_n$ represents the type



	Semantic signature database		
ID	**Type**	**Angle**	**Location**
...	...	...	...
i	ABC	$\Theta_1\Theta_2\Theta_3$	x,y

Fig. 2. The generation of semantic signatures.

sequence of the corresponding objects, $s^{(2)} = a_1|\cdots|a_n$ represents the corresponding angle sequence, and n is the number of visible semantic objects within R. Figure 2 illustrates the generation of semantic signatures. Ideally, each signature is unique, so that localization can be achieved by matching a query signature with a signature database. A database of semantic signatures can be built from existing data sources, such as geographic information systems.

In addition, it is required by one of our signature comparison metrics that the north is known when generating a signature. This is not unrealistic, because nowadays mobile devices are typically equipped with a compass. The centroid of an object is used for representing its location. In order to have a stable angle sequence, it is necessary to quantize angle values. We use 4-bit quantization, i.e., each angle value is quantized by 16 levels (22.5° per step).

B. Signature comparison

Given two semantic signatures, an important question is how to compare them. Since localization is achieved by signature search and retrieval, a similarity metric is needed. Since a signature has two parts, for simplicity it is preferable to use a metric that is compatible with both parts. This is possible if the two parts are considered as two general sequences. In this work, we consider the following metrics: 1) Jaccard distance; 2) Histogram distance; 3) Edit distance.

Denote two ordered sequences as $\mathbf{x} = x_1 \cdots x_n$, $\mathbf{y} = y_1 \cdots y_m$. The Jaccard distance [27] is defined as

$$1 - \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|} \ . \tag{1}$$

The histogram distance is defined as

$$1 - \sum_c \frac{\min\{|\mathbf{x}_c|, |\mathbf{y}_c|\}}{\max\{|\mathbf{x}_c|, |\mathbf{y}_c|\}} \ . \tag{2}$$

where c represents an object class. This metric was used in [25], so it is a good candidate for performance comparison.

The edit distance [28] is defined by the recurrence

$$d_{i0} = i \ , \ d_{0j} = j \ , \ 1 \le i \le m, 1 \le j \le n$$

$$d_{ij} = \begin{cases} d_{i-1,j-1} & \text{if } x_j = y_i \\ \min \begin{cases} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1} + 1 \end{cases} & \text{if } x_j \neq y_i \end{cases}$$

and normalized by $\max(m, n)$. This metric requires that the north is known when generating signatures.

Given two sequences of symbols, these metrics compare the value or the order of the symbols, but they exhibit different levels of "strictness". The Jaccard distance only considers the occurrence and completely ignores the order; the histogram distance also ignores the order but counts the frequency of symbols; the edit distance takes into account both the order and the frequency. By selecting different metrics, different trade-offs between robustness and discrimination power can be achieved. A coarse metric is useful for rough and quick matching, while a fine-grained metric is useful for refined matching. On the other hand, the computation cost is also different. The more complex the metric, the more computation.

C. Retrieval scheme

The localization problem is solved by a retrieval-based framework. With a query image, the following steps apply:

1) A query signature is computed from the query image;
2) Similar signatures are retrieved from a database according to the query signature;
3) The best t matches are returned.

After the best matches are identified, post-processing schemes may follow depending on the specific application. In this paper, the focus is to find the best matches in an accurate and efficient way. Since a semantic signature has two parts – type and angle, we propose to use "metric fusion".

In this scheme, a similarity score is first computed from each part of the signature. Then a weighted sum of the two scores is computed. Signatures are ranked according to the total score. Denote two signatures as $s_a = \{s_a^{(1)}|s_a^{(2)}\}$ and $s_b = \{s_b^{(1)}|s_b^{(2)}\}$. The distance d is defined as

$$d = \alpha \cdot d_1 + \beta \cdot d_2 \tag{3}$$

$$= \alpha \cdot D_1(s_a^{(1)}, s_b^{(1)}) + \beta \cdot D_2(s_a^{(2)}, s_b^{(2)}) \tag{4}$$

where α and $\beta = 1 - \alpha$ are weight factors, D_1 and D_2 are the chosen similarity metrics. When α or β is zero, the scheme reduces to single metric based ranking.

D. Prerequisite and post-processing

The proposed scheme utilizes two properties of semantic objects – type and angle. In general, an object recognition algorithm is needed to provide such information. In case an object recognition algorithm is not available, the type information might be provided by a human user (because the semantic objects are easy to recognize) as a query, which is an alternative way to use the proposed scheme. Experiment

TABLE I
SEMANTIC OBJECTS.

ID	Name	Number	Symbol
1	Alignment tree	1752696	B
2	Water fountain	6713	C
3	Street light	2299639	D
4	Indicator	36333	E
5	Traffic light	102240	G
6	Bike station	14397	H
7	Automatic WC	8006	I
8	Autolib (car) station	4421	J
9	Taxi station	2537	K
10	Public chair	135748	L
11	Bus stop	32320	M

TABLE II
DATABASE PROPERTIES.

(a) basic properties

Visibility range	30 meters
No. of signatures	312134
Mean signature length	14 objects
Covered area	79 km^2
Data storage	38.7 MB

(b) signature group size

	by type	by angle
count	140296	204891
mean	2.2	1.5
std	121.5	11.1
min	1	1
25%	1	1
50%	1	1
75%	1	1
max	29958	1240

results later show that even if angle information is missing, type information can individually facilitate localization.

The general goal of the proposed scheme is to provide a list of potential locations according to a query signature. The most straight-forward way is to take the best match as the answer, i.e., $t = 1$. When $t > 1$, some analysis can be performed with the candidate locations. If extra information is available, such as street-view images or 3D models at the candidate locations, then one may perform 2D-to-2D [23], [29] or 2D-to-3D [9] matching using the query image. However, since these operations are expensive in computation and data storage, it is desirable to restrain them in a small scale. Therefore, it is important that the proposed scheme returns "good" candidates in a short list. This is confirmed by the experiment results.

IV. EXPERIMENTS

The proposed scheme has been extensively evaluated with a city-scale dataset. The dataset, the evaluation framework, and the results are presented in this section.

A. The dataset

Our dataset is about Paris. It consists of approximately $300,000$ semantic signatures that cover most of the city (79km^2). These signatures are built from 11 categories of objects, as listed in Table I. These objects are found from Open Data Paris[1] with known coordinates. The signature database is constructed by sampling the Paris region with a step of $s = 10$ meters. At each sampling point (cell), a semantic signature is created to summarize objects within 30 meters, i.e., the visibility range R is set to 30. Some basic properties of the database are listed in Table IIa. Each database record contains a location (represented by an $s^2 = 100\text{m}^2$ cell) and its signature. If database records are grouped by the signature using only type information, then the number of groups is approximately 45% of the number of signatures (see Table IIb), i.e., on average less than three cells have the same signature. It is expected that each signature group contains only one cell. Some more statistics about the signature groups are listed in Table IIb. It is true that most signature groups ($\geq 75\%$) have only one cell. This is crucial to effective localization. Note that there are also rare cases when it is almost impossible to find the correct location. For example, there are 29958 cells with the same signature type "DDD", which means three street lights. This implies that our proposed solution works in a probabilistic sense. Nevertheless, the localization power can be improved when type and angle information is combined. The last column of Table IIb shows that angle information is even more discriminative than type information. It is also worth noting that the overall file storage only takes 38.7 MB (without optimization) to cover a large area. This is an extremely small cost for city-scale localization. Conventional low-level feature based approaches, e.g. [10], [29], at least require several GB even for a small scene.

In this work, it suffices to use a linear scan scheme for signature retrieval, thanks to the compactness of signatures. Since a signature is encoded by symbols of small alphabets, more efficient indexing is possible if necessary.

B. The evaluation framework

The signature database can be matched with query signatures obtained by various means. In order to evaluate the retrieval aspects of the proposed scheme, we skip object detection. A query set is formed by randomly selecting 10,000 locations and the associated signatures from the database. Each signature in the query set is used for querying the database. The average performance for all queries is noted. We mainly consider two benchmarks:

- Cumulative distribution of distance errors;
- Recall rate of correct locations.

The first benchmark measures the average distance from the ground truth location to a candidate location. The best results among top t candidates is noted. In our experiments, we set $t = 100$. The second benchmark examines the rank of the ground truth location among all candidates, emphasizing the capability as a filtering tool. It can be considered as a special retrieval scenario with only one relevant answer per query. They will be explained with more details later.

In practice, object detection is not perfect. Therefore, we propose to simulate errors in object detection. Each query signature is first randomly distorted before matching with the database. We consider a medium level of distortion, corresponding to 7 occurrences of random distortion, including miss detection, false detection, and false classification. Since the average signature length is 14, each time up to more than

[1]Open Data Paris (https://opendata.paris.fr) hosts a collection of more than 200 public datasets provided by the city of Paris and its partners.

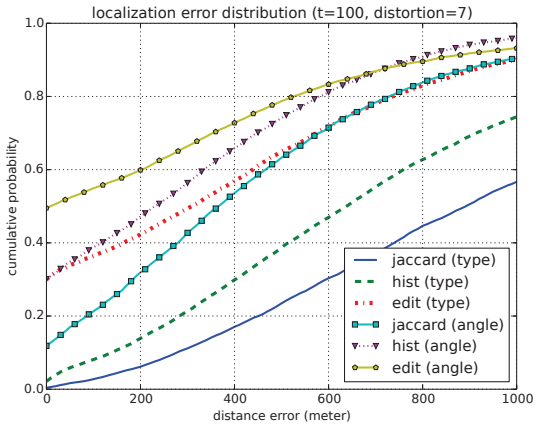


Fig. 3. Localization error (single metric, medium distortion).

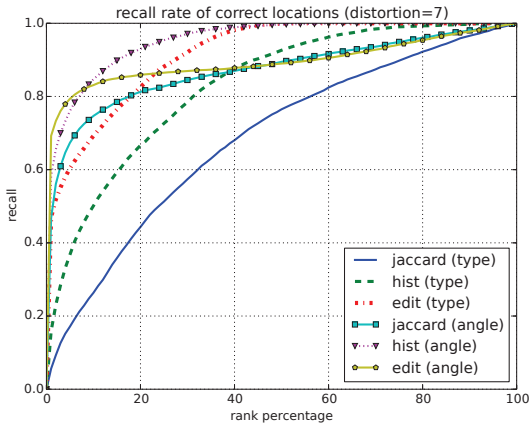


Fig. 4. Location recall (single metric, medium distortion).

50% objects in a signature are distorted. In addition to type distortion, angle noise is always applied following a normal distribution with the standard deviation equal to 5 and the maximum value clipped to 30.

C. Experiment results

In this section, we evaluate the proposed scheme in terms of the two benchmarks defined in Sect. IV-B. Due to space limit, only part of the results are presented. The two signature parts, type and angle, are separately tested first, followed by the metric fusion scheme.

We first examine the effectiveness of the signature scheme. Figure 3 shows the localization error when only one part of a signature is used. There are six curves corresponding to the three metrics and the two signature components. For each point (x, y) on a curve, it means for $y \times 100\%$ queries, the distance error is not larger than x. In general the cumulative probability increases with the localization error. A higher curve means better performance. We observe that both type and angle information can be used for localization, but the smaller error shows that angle generally works better. Among the three metrics, edit distance is the best, followed by histogram distance and Jaccard distance. In the best case, more than 50% queries result in the correct locations or have errors within 10 meters, which is close to GPS accuracy. In the worst case, most queries are located with errors less than 1 km.

Next, we consider the recall rate. In Fig. 4, a point (x, y) means for $y \times 100\%$ queries, the corresponding ground truth rank is not lower than $x\%$. Ideally, we expect the recall to be as high as possible. The results show that given a query, our proposed method can effectively filter out irrelevant regions. The good settings can keep the ground truth rank within top 20% for 80% of queries. That means only the top 20% database candidates need to be considered in general. It is also noted that metrics with worse performance for higher ranks sometimes give better recalls for lower ranks. For example, the histogram distance with angle information gives higher recalls when ranks lower than 10% are considered.

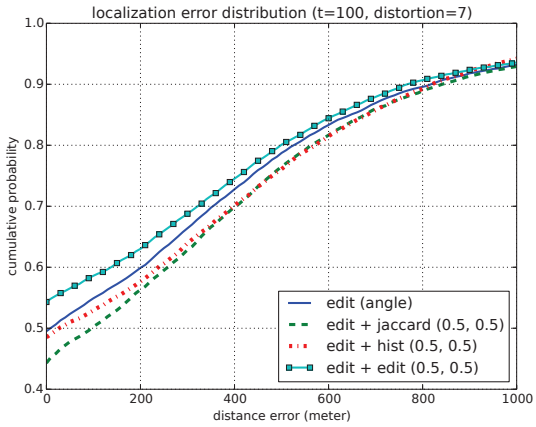


Fig. 5. Localization error (metric fusion, medium distortion).

Figure 5 shows the distribution of localization errors for metric fusion. Since edit distance performs best as a single metric, we fix it for type information and try different metrics for angle information. For example, the legend "edit + jaccard (0.5, 0.5)" means that edit distance is used for type, and jaccard distance is used for angle; the weight factors are empirically set to 0.5 and 0.5. The figure confirms that combining type and angle information indeed brings performance improvement. The initial probability increases from 0.5 to 0.55. It is obvious that "edit + edit" is the best combination. On the other hand, note that using edit distance and angle information alone even outperforms some combinations. That is because metric fusion suffers from more noise than a single metric approach. In some cases, the gain by metric fusion is cancelled out by the stronger noise. The corresponding recall is shown in Fig. 6. Compared with Fig. 4, the advantage of metric fusion is clear for higher ranks where curves are relatively close; for lower ranks, "edit + edit" continues to outperform "edit (angle)", but "edit + hist" performs worse than "hist (angle)" as a trade-off for a slight improvement at higher ranks. We conclude that in general metric fusion is beneficial.

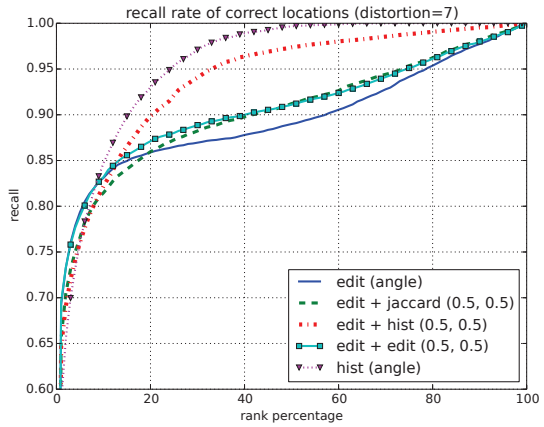


Fig. 6. Location recall (metric fusion, medium distortion).

V. Conclusion and Discussion

In this work, we propose to use semantic information for urban localization. We focus on special objects that can be seen from street views, such as trees, street lights, bus stops, etc. These semantic objects can be obtained from public data sources. They are encoded as semantic signatures. The localization problem is solved by signature matching. Given a query signature, similar signatures are retrieved from a database. The query location is inferred from the best matches' geo-reference. A semantic signature consists of two parts, a type sequence and an angle sequence. We select a few metrics for sequence matching and find that edit distance shows promising results. In order to aggregate both type and angle information, a metric fusion framework is proposed for signature retrieval. Simulation shows that the proposed technique ideally achieves close-to-GPS accuracy.

This paper focuses on retrieval. A number of existing semantic objects are used. While object detection is not covered here, the main message of this self-contained work is that if a sufficient amount of semantic objects exist, then satisfactory localization is possible even in a large scale.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[2] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[3] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-DOF localization in large-scale environments," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1043–1050.

[4] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *International Conference on World Wide Web (WWW)*, 2009, pp. 761–770.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[7] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–7.

[8] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 791–804.

[9] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *International Conference on Computer Vision (ICCV)*, Nov 2011, pp. 667–674.

[10] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3D point clouds," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 15–29.

[11] A. Irschara, C. Zach, J. M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 2599–2606.

[12] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 255–268.

[13] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 737–744.

[14] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 154:1–154:10, Dec. 2011.

[15] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[16] A. R. Zamir and M. Shah, "Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1546–1558, Aug 2014.

[17] R. Arandjelović and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *Asian Conference on Computer Vision (ACCV)*, 2014, pp. 188–204.

[18] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1808–1817.

[19] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 3304–3311.

[20] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[21] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Panorama to panorama matching for location recognition," in *ACM International Conference on Multimedia Retrieval*, 2017, pp. 392–396.

[22] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *International Conference on Computer Vision (ICCV)*, Sept 2009, pp. 72–79.

[23] Y. Song, X. Chen, X. Wang, Y. Zhang, and J. Li, "6-DOF image localization from massive geo-tagged reference images," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1542–1554, Aug 2016.

[24] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla, "Are large-scale 3D models really necessary for accurate visual localization?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6175–6184.

[25] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *Eupropean Conference on Computer Vision (ECCV)*, 2014, pp. 602–617.

[26] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit, "Instant outdoor localization and SLAM initialization from 2.5D maps," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 11, pp. 1309–1318, 2015.

[27] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[28] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[29] N. Bhowmik, L. Weng, V. Gouet-Brunet, and B. Soheilian, "Cross-domain image localization by adaptive feature fusion," in *Joint Urban Remote Sensing Event*, 2017, pp. 1–4.