Experiments in Place Recognition using Gist Panoramas *

A. C. Murillo
DIIS-I3A, University of Zaragoza, Spain.
acm@unizar.es

J. Kosecka
Dept. of Computer Science, GMU, Fairfax, USA.
kosecka@cs.gmu.edu

Abstract

In this paper we investigate large scale view based localization in urban areas using panoramic images. The presented approach utilizes global gist descriptor computed for portions of panoramic images and a simple similarity measure between two panoramas, which is robust to changes in vehicle orientation, while traversing the same areas in different directions. The global gist feature [14] has been demonstrated previously to be a very effective conventional image descriptor, capturing the basic structure of different types of scenes in a very compact way. We present an extensive experimental validation of our panoramic gist approach on a large scale Street View data set of panoramic images for place recognition or topological localization.

1. Introduction

Recent interest in large scale acquisition of visual imagery of large urban areas gives rise to many novel applications and also needs to develop automated methods for organizing, geo-registering and annotating the data. In this paper we investigate qualitative view based localization in large urban areas using panoramic views. Given a database of reference panoramas covering a large urban area, and a new query view, we would like to find its nearest reference view in the database. This problem has been typically addressed in the last years using image representations in terms of scale invariant features [15, 5, 3, 8] computed either over individual views or the entire panoramas. The large scale issues have been addressed using various forms of visual vocabularies and inverted file index. Some approaches, e.g. [8], additionally use GPS measurements to constrain the image similarity evaluation. However, GPS measurements are not always accurate or available and they do not provide orientation information. Therefore, we intend to investigate approaches based on vision only, where indeed GPS restrictions can easily be included if available.

Scale-invariant features have been shown to be very effective and to achieve good performance rates in view based localization. As image similarity measure, the most commonly used models adopted the so called bags of features models, where in initial voting the spatial relationships between features were not considered. It has been also shown [12, 5] that the performance of these approaches depends on the size of the visual vocabulary. These factors are likely to be emphasized further with the increasing scale of the problem and large variations in considered visual appearance.

In this paper we investigate alternative image based localization using global gist descriptor proposed by [14] adapted to omni-directional views. We postulate that to achieve accurate view based localization at large scales in the absence of other sensors such as GPS, the data has to be organized based on simple computable notions of visual similarity. The gist descriptor introduced in [13] has been shown to suitably model semantically meaningful and visually similar scenes. Although the descriptor is not very discriminative, the attractive feature of this representation is that it is very compact, fast to compute and that roughly encodes spatial information. The gist descriptor enables us to obtain a small number of visually similar clusters of alike locations, e.g., large open areas, suburban areas, large high-rise buildings etc. Each of these visually similar areas could then be endowed with its own visual vocabulary.

We demonstrate the performance of this representation for large scale place recognition in urban areas, focusing on two tasks: 1) given gist representation of a conventional view, we adapt it for panoramas and show how effective it is to find the nearest panorama in a reference database; 2) given gist representation of each panorama in a sequence, we endow it with a topological model and evaluate the effectiveness of the gist panorama for topological mapping.

Figure 1. Street view data set of panoramic views used, *dataset-L*.

*This work was supported by projects DPI2006-07928, DPI2009-14664.

2. Related work

Our work is closely related to recent trends in large scale vision based localization and topological mapping using omni-directional views. We briefly review some representative related works from each of them.

Large Scale Location Recognition. Several approaches attempted recently to tackle the problem of location recognition in large scale. These advances are mostly facilitated by the design and matching of scale and viewpoint invariant features [9] and large scale nearest neighbour methods [12] or classification methods which use large number of features [8]. In the feature based setting, the central problem of location recognition is the following: given a current query view, find the nearest view in the database of reference images. In [3] the scalability and speed were achieved by using a version of vocabulary tree along with inverted file index, used to retrieve the top k-views with the closest visual word signatures. In [8] vocabulary trees and randomized trees are used to retrieve the top k-views, which were limited only to the views which were in the vicinity of the current location determined by GPS.

In this work we would like to explore the efficiency and suitability of the gist descriptor in retrieving the closest views of street scenes. We adjust the gist representation to be suitably adapted to deal with panoramic images, enabling the full 360^o panoramic image localization, assuming a discrete set of orientations of the vehicle.

Topological localization. Localization using omnidirectional images has been shown very effective in the context of topological localization. The wide field of view allows the representation of the appearance of a certain place (topological location) with a minimal amount of reference views. Besides, omnidirectional images provide bigger intersection areas between views, and therefore facilitate the feature correspondence search. Topological maps, or visual memories, based on different types of omnidirectional vision sensors have been used already for several years for environment representation and navigation in robotics framework [6, 10, 17] with very good results. We find works for both offline and online topological map building [7, 19], topological localization based on local features indoors (room recognition) [21] or outdoors [20] or hierarchical approaches to integrate global and local features similarity evaluation [11].

Works on place or scene recognition using conventional images such as [4, 16, 2] are also related to our location recognition goals.

3. Panorama Gist

Street panoramas. We create one panoramic image by warping the radially undistorted perspective images onto the sphere assuming one virtual optical center. One virtual optical center is a reasonable assumption considering that the structure around the sensor is very far compared to the discrepancy between optical centers of all the cameras. The sphere is backprojected into a quadrangular prism to get a piecewise perspective panoramic image, see Fig. 2. Our panorama is composed then of four perspective images covering in total 360^o horizontally and 127^o vertically. We do not use the top camera as there is not much information. The panorama is then represented by 4 views (front, left, back and right) which we will refer to as F, L, B, R in the later text. We discard the bottom part of all views, to convert them into squared views, discarding the areas of the images that captured parts of the car acquiring the panoramas.

Figure 2. Panorama acquisition device. (a) Point Grey *LadyBug* camera. (b) A panoramic piecewise perspective image as an outer surface of the prism.

3.1. The gist descriptor

The gist descriptor [13, 14] is a global descriptor of an image that represents the dominant spatial structures of the scene in the image. Each image is represented by a 320 dimensional vector (per color band). The feature vector corresponds to the mean response to steerable filters at different scales and orientations computed over 4 x 4 sub-windows. For a more intuitive idea of what this descriptor encloses, see the clustering results and average images obtained when clustering gist values from a big set of outdoor images, shown later in Figs. 4, 5. The advantage of this descriptor is that it is very compact and fast to compute. The gist has been shown effective in holistic classifications of scenes into categories containing tall buildings, streets, open areas, highways and mountains etc [18] and has been used effectively for retrieving nearest neighbors from large scale image databases.

In order to obtain the gist descriptor for the entire panorama, we estimate the standard gist descriptor for each of the 4 views and subset of reference views. We then quantize the space of all descriptors to build a vocabulary of gist words and later on represent each view by an index of the closest *gist word* obtained in our gist vocabulary. The final gist descriptor includes then 4 indexes of closest visual words in the gist vocabulary.

3.2. Gist vocabulary

The motivation for quantizing the space of gist descriptors is two fold. First due to the nature of the gist descriptor

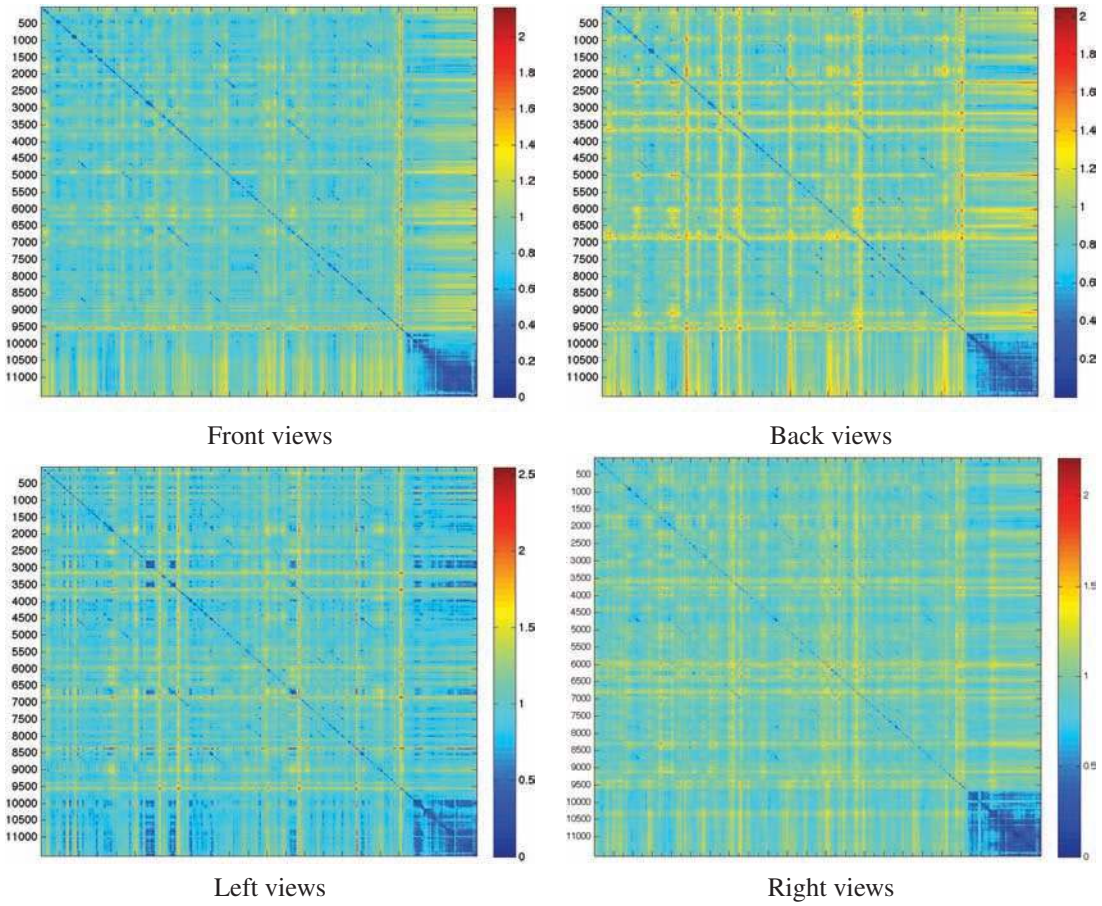

Figure 3. Similarity matrix of the gist in single views.

we would like to demonstrate that the clusters in the gist descriptors space correspond to visually similar views of urban environments. Second, we would like to increase the efficiency of the gist descriptor matching stage, where instead of finding exact nearest neighbour, we seek the nearest cluster of the gist in each view, limiting the number of reference views which will be considered in the later stage.

To demonstrate the visual similarity between neighboring panoramas, as well as other similar areas, Fig. 3 shows four plots that represent the affinity matrices between gist signatures of all views from each type (front, left, back, right). Dark blue areas represent the higher visual similarity (smaller gist distances). We can observe that certain areas naturally form clusters demonstrated by blocks in the gist affinity matrix. The clearest division occurs for the last set of panoramas, indexes 9700 to the end, where the trajectory (see Fig. 1) goes out of downtown area to enter a highway area. Highway areas have very different structure than any downtown area, so they get a quite distinct gist descriptor.

We could also observe in the affinity matrices when we are traversing intersection areas, such as frames around number 500, where gist in front and back views are very similar for a while, but the lateral views get higher distance between the same consecutive areas. Besides, it can be observed when there are areas which are re-visited: in different sets of rows we see multiple block diagonal structures of similar panoramas with the same structure as the main diagonal in the images (e.g. around frame 3500 with 6500).

A subset of the panoramas, which we will call reference panoramas, is used for building the visual vocabulary. We run a simple k-means algorithm on all gist descriptors from all 4 views of reference panoramas. The k-centroid gist values obtained with this clustering technique are the most representative views we find in our dataset. We will call them the k *words* of our gist *vocabulary*, following the terminology typically used in object recognition, where visual vocabularies are build from scale invariant features [12].

Fig. 4 shows a set of elements in some of the clusters computed (in our experiments we use a $k = 35$). There, we can see how views with similar basic structure are classified together. Fig. 5 presents an average image of all views that belong to the same cluster for one of the panorama reference sets used. Notice that qualitatively different features of urban areas are revealed by individual clusters.

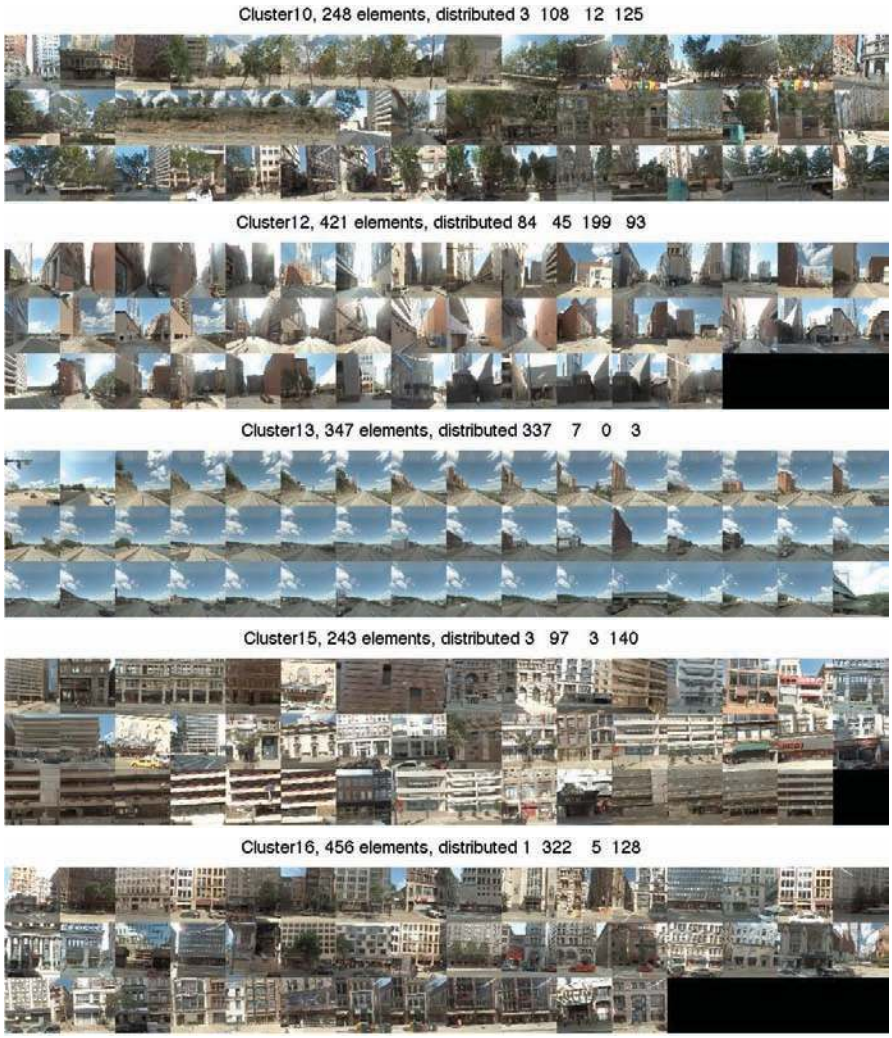

Figure 4. Clustering of the reference panorama view gists into a 35-words vocabulary: a few views from some of the clusters. Each image includes on top the number of views (elements) that fall into that cluster and how many are distributed in each type of view $[F, L, B, R]$.

Figure 5. Average view in each of the gist-vocabulary clusters built from $\approx$ 1000 reference panoramas (4000 views).

3.3. Panorama Gist matching

Once we have computed the view-gist vocabulary, we can efficiently compare panoramas with the following approach. For each query panorama:

a) Compute the gist of each of the 4 views in the panorama and obtain 4 gist descriptors for the panorama $\mathbf{g} = [g_f, g_l, g_b, g_r]$.

b) Estimate which word of the reference view-gist vocabulary is closer to each of the view-gist: we get 4 word indexes $[w_i, w_j, w_k, w_l]$ that represent the structure of the query panorama.

c) We look for the reference panoramas which share as many visual words as possible with the query view, and these words are in the same relative positions (i.e, if we find one reference panorama with four words in common, we only keep panoramas with 4 words in common). To verify the relative position, we check the relative offset we would

need to align the panoramas. Since the gist representation is not rotationally invariant, the relative offsets at the level of visual words reflects the fact that we are at the same location but with different orientation. For example, if we have a query panorama with words $[w_1, w_2, w_3, w_8]$ and a reference panorama words $[w_5, w_1, w_2, w_3]$, both panoramas have three words in common with same "relative position": words 1, 2 and 3 from the vocabulary, with an offset of -1.

d) We compute the Euclidean distance between the four aligned views of each panorama. The distance from the query panorama Q and each of the selected reference panoramas properly aligned R is computed as: $distG(Q, R) = \|(\mathbf{g}_Q, \mathbf{g}_R)\|$.

e) Finally, a reference panorama is considered "similar enough" and accepted as candidate match if we find at least two corresponding words to align the panoramas and the $distG(Q, R)$ is below a threshold. This threshold is established relative to the average distance within cluster elements of the gist vocabulary. Otherwise, the reference panorama is discarded for this query. The accepted panoramas are the candidate matching set. We keep the top N candidates, if there are more than N, sorted by $distG(Q, R)$. In the following section we analyze in detail the performance obtained for different configurations of these parameters.

4. Experimental Results

We demonstrate the performance of our approach in a large dataset of 12,000 street view panoramas, *dataset-L*. This data set is composed of a 13 mile long run in urban area and can be seen in Fig. 1. First we compute the gist-vocabulary from the reference panorama set, as shown in section 3.2. To extract the explained gist descriptor, we have used the code available on the web [1]. We show the performance on the following experiments using two different sets of reference panoramas: one composed by every 4^{th} frame from the dataset sequence (around 3000 reference panoramas, let's name it *refset-3000*), and another one composed by every 10^{th} frame from the sequence (around 1000 reference panoramas, let's name it *refset-1000*).

We evaluate the performance of our representation on two tasks: panorama matching and place recognition, both using the gist-based similarity evaluation proposed in previous section 3.3.

4.1. Panorama matching.

The goal is to identify the reference panorama that is more similar to a certain query panorama. We run the similarity evaluation for all the panoramas not used to build the gist vocabulary (reference panoramas). Fig. 6 shows a few test panoramas and the reference panoramas that were selected as most similar. The number of matched panoramas is variable, depending how many have passed the similarity evaluation test (see section 3.3). The last two row examples are four panoramas that did not get any match. This happens because of the restriction that we impose on having more than one gist-word in common with the matched panoramas. We have observed that in some cases, a failure can be originated by the fact of view-gist words wrongly estimated for a particular gist. We are currently investigating soft assignment strategies to avoid these effects.

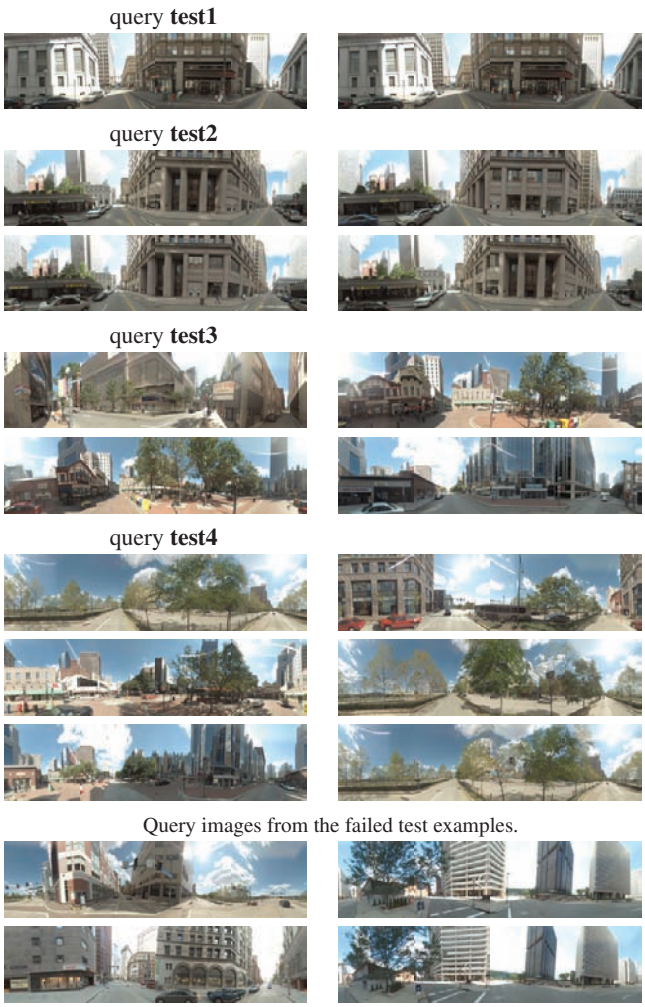

Figure 6. Examples of the panoramas matched to a particular query (panorama on the top-left in each block). Last two rows are the query images that failed to be matched.

Fig. 7 shows the same three tests than previous figure, but on an overview of the whole scene/trajectory. The two first examples are clearly correct matches, first case with just one selected match that is from the correct location, and second case with several selected matches, all of them correct. However, test 3 corresponds to a test that fails to recognize the correct place: it gets too many candidates accepted far from the real location, so if we select a small N

[1] http://people.csail.mit.edu/torralba/code/spatialenvelope/

test 1 test 2 test 3 test 4

Figure 7. Examples of the matched panoramas (yellow circles) considered similar enough, to a particular query panorama (red pin). Same four tests than Fig. 6.

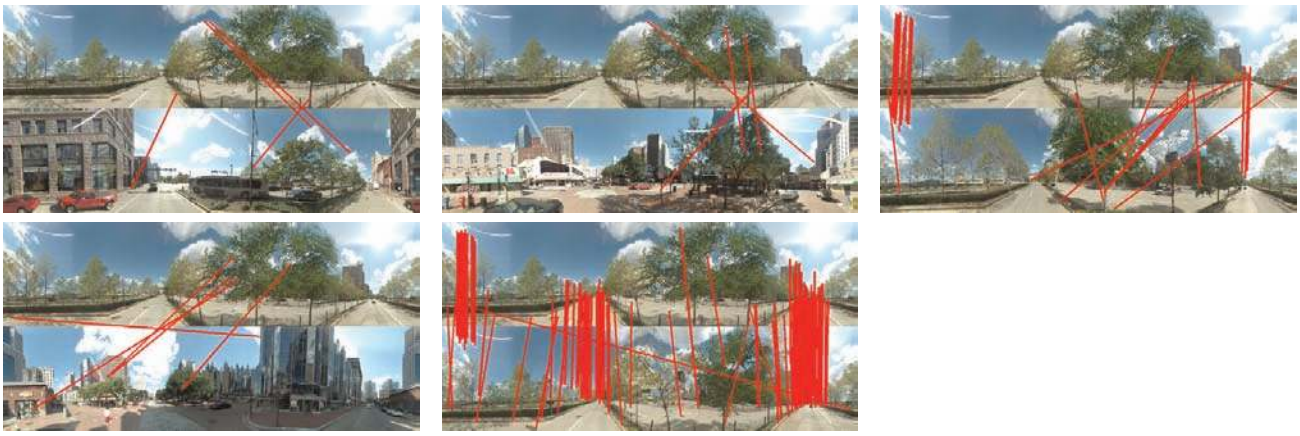

Figure 8. Test 4. Example of local feature matching between a query (top panorama) and the top-5 candidates to select the best one.

for the top N matches selected, we do not include any correct one. Test 4 presents another situation where we also obtain several matches accepted, some of them correct and some from an incorrect place.Many times the correct solution could be established running local feature matching with the panoramas in the accepted candidate set. Fig. 8 shows an example of a simple nearest-neighbour matching with SURF features [1] with the top-5 candidates in test 4, where we could clearly detect than the last candidate is the correct one according to the local feature matches.

The average performance of the following visual localization experiments is presented in Table 1, where we can observe several interesting issues detailed next.

Experiment 1. Vocabulary & panoramas from *dataset-L*. First we run several tests using only the main dataset available, *dataset-L*. The left part of the table presents the results when using the smaller set of reference images (every 10^{th} panorama, *refset-1000*), and the middle part presents the results using a bigger reference set of panoramas (every 4^{th} panorama, *refset-3000*). Not surprisingly, with smaller reference set results are slightly better since it suffers less from hard cluster assignment in the visual vocabulary. This is reflected by the fact that having less reference panoramas also implies having less chances of discarding the right reference panorama.

Experiment 2. Vocabulary from *dataset-L* & panoramas from *dataset-S*. In this second part of experiments, we try to go a bit further, and check how general the kind of gist-words that we have obtained are when we work on completely separated datasets (with the only common characteristic of being urban areas). So these tests have used the same vocabulary as before, computed from panoramas from *dataset-L*, but the reference and test panoramas used are from a different dataset, *dataset-S*. This set is from a completely separated urban area, composed of around 600 panoramas from a trajectory of approximately 600 meters. Right part of Table 1 shows the matching ratios with this configuration.

Each row from the same Table shows the ratio of tests with the closest match to a reference panorama at a certain

Table 1. Results for place recognition based only on gist comparisons, using a gist-vocabulary computed from *dataset-L* panorama views.

	Experiment 1														**Experiment 2**		
	dataset-L, refset-1000							*dataset-L, refset-3000*							*dataset-S*		
	top 5	top 10	top 20	top 30	top 40	top 50	top 60	top 5	top 10	top 20	top 30	top 40	top 50	top 60	top 5	top 10	top 20
distG $<$ 5m	0.58	0.61	0.62	0.63	0.63	0.65	0.65	0.52	0.57	0.61	0.62	0.63	0.63	0.64	0.55	0.59	0.60
distG $<$ 10m	0.68	0.71	0.73	0.75	0.76	0.77	0.78	0.59	0.63	0.67	0.68	0.69	0.70	0.70	0.65	0.69	0.69
distG $<$ 20m	0.71	0.74	0.76	0.77	0.78	0.79	0.80	0.63	0.68	0.71	0.73	0.73	0.74	0.75	0.77	0.79	0.80
distG $<$ 40m	0.73	0.76	0.78	0.80	0.80	0.81	0.82	0.66	0.71	0.74	0.76	0.77	0.78	0.78	0.86	0.87	0.87
distG $<$ 60m	0.75	0.78	0.80	0.81	0.82	0.83	0.84	0.68	0.73	0.76	0.78	0.79	0.80	0.80	0.92	0.93	0.93
No Match				0.0008							0.001					0.002	
Place detected	0.71	0.73	0.74	0.74	0.74	0.75	0.76	0.7	0.74	0.76	0.77	0.77	0.78	0.78	0.73	0.75	0.76

distance (under 5 meters, under 10 meters, etc.); row *No Match* shows how many query views did not pass the matching criteria. Each column presents the same results but if we consider only the best 5, 10, 20, etc. matches (sorted by smaller panorama-gist distance). Final row of each table explains the visual localization results taking into account the segmentation into places of the reference panorama set, detailed in next subsection.

We can notice than a considerable amount of tests only achieve a correct localization within more than 50 meters. As mentioned before, this localization would be easily made more accurate by running a local feature matching, e.g. using [9, 1], between the query panorama properly aligned to the reference matched panoramas.

4.2. Place recognition using panorama gist.

We can organize automatically the reference panoramas using the same panorama similarity evaluation used for panorama matching. The reference set of panoramas can be automatically segmented into clusters, sections or places with a simple online approach:

- If the panorama-gist, 4-sorted gists, is similar enough to previous panorama, the same place is still assigned.

- Else a new place is initialized. We could try to merge it with a previously visited cluster, but gist only is not enough to distinguish if we are re-visiting an area or we are just in a similarly structured area, we would need to use odometry or extra-features to confirm that we have a "re-visit" candidate.

Experiment 1. Fig. 9 shows how the reference set of panoramas gets clustered into different sections or areas. We can use this division to evaluate if the visual localization performed in previous subsection 4.1 is correct: place recognition is achieved if there is any panorama in the candidate set selected that belongs to the same place than the test panorama. As ground truth, a test panorama is considered to be at the same place than the closest reference panorama (according to additional GPS image tags).

Last row of table 1, *Place detected*, shows the ratio of tests where the place was correctly identified. The results are promising, since the number of correct real places could be easily still improved: the automatic segmentation of the reference frames into places is still too simple, since we should work on detecting re-visited areas, to tag them as the same place. Right now the following situation is likely to be considered wrong, and it would be corrected if we would filter the re-visited places in the topological map: the query belongs to the second time we traversed a certain street, but the selected reference image belongs to the first time we traversed this same street, so they have been assigned to different sections or places although they should be the same.

Figure 9. **Experiment 1**, *dataset-L*. Top: aerial image of dataset trajectory. Bottom: online clustering of *refset-1000* panoramas.

Experiment 2. In this second experiment, we obtain pretty similar results than in previous experiment 1 to organize this new dataset into a set of reference panoramas segmented into places, as well as to localize a set of test panoramas with regard to those places. Fig. 10 shows the segmentation into places of the reference panoramas in this experiment over the aerial picture of the area. Right part of

Figure 10. **Experiment 2,** *dataset-S*. Segmentation of reference panoramas into places (topological map).

Table 1 shows the statistics for place recognition results in this experiment. It is expected than the ratios if we accept localization within 60 meters are a bit higher because this dataset covers a much smaller area than experiment 1.

5. Conclusions and Future Work

In this paper we have presented an approach for panorama similarity evaluation based on gist features. The goal of our work is to contribute on visual localization techniques focused on large scale omnidirectional image datasets. We have tested the presented ideas in a large scale dataset composed of panoramas of an urban area. The results achieved are very promising, with good ratios of correct panorama localization and nice segmentation of the reference panoramas into clusters or places of a reference topological map. The proposal still has several open issues that would easily improve the current performance, such as taking into account re-visited places in the topological map, to avoid false negatives in place recognition and more complex similarity measures. Other tasks yet to be explored are to determine the minimum amount of reference information necessary or to use as reference panorama-gist the average values of each topological map place.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. of ECCV*, 2006, http://www.vision.ee.ethz.ch/ surf/.

[2] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(4), 2008.

[3] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *Int. J. of Robotics Research*, 27(6):647–665, 2008.

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of IEEE CVPR*, pages 524–531, 2005.

[5] F. Fraundorfer, C. Engels, and D. Nistér. Topological mapping, localization and navigation using image collections. In *In Proc. of IEEE/RSJ IROS*, pages 3872–3877, 2007.

[6] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Trans. on Robotics and Automation*, 16(6):890–898, 2000.

[7] T. Goedemé, M.Nuttin, T. Tuytelaars, and L. Van Gool. Omnidirectional vision based topological navigation. *Int. J. of Computer Vision*, 74(3):219–236, 2007.

[8] A. Kumar, J.-P. Tardif, R. Anati, and K. Daniilidis. Experiments on visual loop closing using vocabulary trees. *CVPR Workshop*, 0:1–8, 2008.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004, http://www.cs.ubc.ca/ lowe/keypoints/.

[10] E. Menegatti, T. Maeda, and H. Ishiguro. Image-based memory for robot navigation using properties of the omnidirectional images. *Robotics and Autonomous Systems*, 47(4):251–267, 2004.

[11] A. C. Murillo, C. Sagüés, J. J. Guerrero, T. Goedemé, T. Tuytelaars, and L. Van Gool. From omnidirectional images to hierarchical localization. *Robotics and Autonomous Systems*, 55(5):372–382, 2007.

[12] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE CVPR*, pages 2161–2168, 2006.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. of Computer Vision*, 42(3):145–175, May 2001.

[14] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. In *Visual Perception, Progress in Brain Research*, volume 155. Elsevier, 2006.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of IEEE CVPR*, 2007.

[16] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *Proc. of IEEE CVPR*, pages 1–7, 2007.

[17] A. Tapus and R. Siegwart. Incremental robot mapping with fingerprints of places. *In Proc. of IEEE/RSJ IROS*, pages 2429–2434, 2005.

[18] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proc. of IEEE ICCV*, page 273, 2003.

[19] C. Valgren, T. Duckett, and A. J. Lilienthal. Incremental spectral clustering and its application to topological mapping. In *Proc. of IEEE ICRA*, pages 4283–4288, 2007.

[20] C. Valgren and A. J. Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *European Conf. on Mobile Robots*, pages 253–258, 2007.

[21] Z. Zivkovic, O. Booij, and B. Krose. From images to rooms. *Robotic and Autonomous Systems*, 55(5):411–418, 2007.