Gist vocabularies in omnidirectional images for appearance based mapping and localization

A. C. Murillo*, P. Campos*, J. Kosecka† and J. J. Guerrero*
* DIIS-I3A, University of Zaragoza, Spain † Dept. of Computer Science, GMU, Fairfax, USA
Email: {acm, jguerrer}@unizar.es Email: kosecka@cs.gmu.edu

***Abstract*— Appearance based topological localization and mapping is a subject of great interest for autonomous robotics systems. The need for mapping larger environments and describing them at different levels of abstraction requires alternatives to purely metric representations of the environment and additional ability to deal with large amounts of data efficiently. The key component of image based localization is the search for the closest view in the image database representing the environment. In this work we study the efficiency and potential of global gist descriptor [1] adapted to catadioptric systems and present a new hierarchical approach for topological mapping and localization.**

Our method relies on the omni-gist descriptor and integrates local feature matching to refine the candidates at places where loop-closure detection can occur. Three different omnidirectional image datasets from real scenarios are used to demonstrate the performance of this method, providing comparable results for appearance based localization than previous approaches based on local features. The storage and computation efficiency of global descriptors notably improves the efficiency of the system.

I. Introduction

Visual databases of urban environments are becoming widespread in many fields due to wider availability and lower cost of cameras, higher computational resources available in mobile devices or higher bandwidths in communications that easily allow us to share visual information online. Therefore, effective and efficient ways of searching this type of databases has become a key issue in solutions to many problems and tasks mainly from the fields of computer vision, machine learning or more particular robotics problems. Efficiency is required typically due to large amounts of data used, often real-time requirements either from autonomous systems or because systems and services that interact with humans need a response as fast as possible to have a good usability.

This work is focused on dealing with databases of omnidirectional images, in particular, images acquired with catadioptric systems (mirror based). Since our area of interest is in omnidirectional vision systems, which is particularly advantageous for autonomous systems and robotics, we pay special attention to basic robotics tasks related with processing visual databases: appearance based mapping and localization.

In the last years, there have been several impressive results demonstrating content based image retrieval using local scale-invariant features, using various approximate nearest neighbor methods (e.g. k-d trees, vocabulary trees) and inverted file index. Most of them use local features as image representation [2, 3, 4]. Alternative works which use global descriptors, such as gist descriptor [1] have shown impressive image retrieval results on internet datasets [5, 6]. The goal in this work is to explore how global descriptors, instead of or in combination with local features, could improve or contribute to the effective processing of omnidirectional image datasets, in particular for appearance based mapping and localization.

A. Contributions

First, we propose how to extract the gist descriptor in catadioptric images and how to evaluate the similarity between images according to this descriptor, detailed in Section III. In the experimental comparison we demonstrate the effect of unwarping of the omnidirectional image in the context of the proposed descriptor and show that for our goals the performance and efficiency is clearly higher without unwarping the images. Finally, based on these proposals, new approach for appearance based mapping and localization based on the omnidirectional image gist descriptor is presented, see Section IV. Extensive experiments on three different datasets demonstrate the good performance of the proposals in Section V.

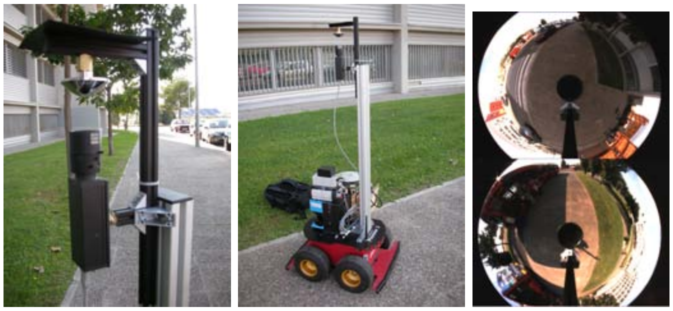

(a) (b) (c)

Fig. 1. Catadioptric vision system (a) mounted on a robot (b) and sample omidirectional images acquired with it (c).

II. Related Work

Appearance based topological mapping and localization have been extensively studied, both with conventional and omnidirectional vision systems [7, 8]. The key component of vision based localization is to find the closest view in the database representing the environment. This problem is closely related to the more general problem of image based retrieval in large databases [5]. The issues of scale have been also tackled in the context of visual loop closing problem in localization in large scale environments using local features [9]. Current issues of interest for the appearance based place recognition, localization and mapping problems are mainly related to efficiency and processing of large scale datasets, as well as being able to perform long-term mapping with models/maps robust to strong changes occurring along big time intervals.

A. Offline and online topological mapping.

Related works to topological mapping approaches can be grouped in two types, offline and online approaches. On one hand, offline approaches allow to optimize the image clustering after the whole data set has been acquired, e.g., using graph cut techniques as proposed in [10]. This type presents the advantage of allowing an optimization over the whole dataset, but has the disadvantage of being offline and usually computationally more expensive. On the other hand, online approaches build a topological map of the environment by instantiating a new cluster each time the algorithm detects a significant change in the image acquired with regard to previous one. Many different criteria have been studied to define what a significant change is. In some cases the partitions are made for each small subset along the image sequence, while other times the partitioning involves a complex decision process [11, 12]. Online approaches present extra difficulties but are also more efficient and have useful advantages for robotics since they build the topological map online.

B. Local and global features for appearance based localization and mapping

Regarding the image features used, most recent advances in appearance based localization are based on local feature constructions, including geometric constraints [10] and efficient approximate nearest neighbour methods and quantization schemes, e.g. k-means clustering, vocabulary trees and inverted file structures [2, 3, 4]. Recent proposals have shown scalability to datasets covering very large appearance variations in large areas [9] and in different seasons, with the corresponding landscape appearance changes [13]. Alternative successful proposals in the computer vision community are based on global image descriptors, in particular the gist descriptor [1] that have been shown to work nicely for scene recognition, scene alignment or duplicate detection in very large datasets, typically web-data sets [14, 5, 6]. The use of global descriptors has well-known disadvantages, mainly lower invariant properties and worse robustness to big occlusions, but they also present advantages, mainly a more compact and efficient representation of each image, allowing great enhancements in storage and computation efficiency. In spite of these results, we find very few recent works on robotic applications (e.g., mapping, localization) exploring the advantages of these types of global descriptors and furthermore applying them to catadioptric vision systems. Some previous works show results using global descriptors, different types of histograms, to efficiently find similar omnidirectional images to a certain query [8, 15, 16] and the work presented in [17] shows promising directions on applying gist global feature to omnidirectional images, in that case multi-camera systems. More detailed study of applications of these descriptors for visual loop closing can be found in [18]. This work builds on the recent results on panorama-gist based place recognition using the quantized gist descriptors [17] and extends that proposal, by i) defining how to apply it to omnidirectional images from catadioptric cameras, rather than panoramas from multi-camera systems, ii) proposing a more powerful similarity measure between raw omnidirectional images applied in new appearance based mapping and localization algorithms.

III. Omnidirectional image gist

The gist descriptor [1, 19] is a global descriptor of an image that represents the dominant spatial structures of the scene captured 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.

The type of images we are focused at in this work are those acquired with a catadioptric vision system, like the one shown in Fig. 1. This section details how we propose to extract the gist descriptor [1], originally designed for conventional and squared images, in this type of image.

A. Computing the gist of a catadioptric image: omni-gist.

Similarly to the approach followed in [17] for multi-camera panoramic vision systems, we segment the omnidirectional image in four symmetric pieces, rotating them to a canonical orientation and computing the gist descriptor in each of these pieces. An important practical problem that appears with catadioptric images is that we have to ”mask” parts of the image, where there are artifacts produced mostly because of the reflection of parts of the catadioptric system in its own mirror. Fig. 2 a) shows an example of the parts of an omnidirectional image that we should not take into account for the processing (marked in green), Fig. 2 b) presents how we segment the image into four parts to compute the gist. Fig. 2 c) contains an example of the four sectors extracted for a sample from each dataset used in this work.

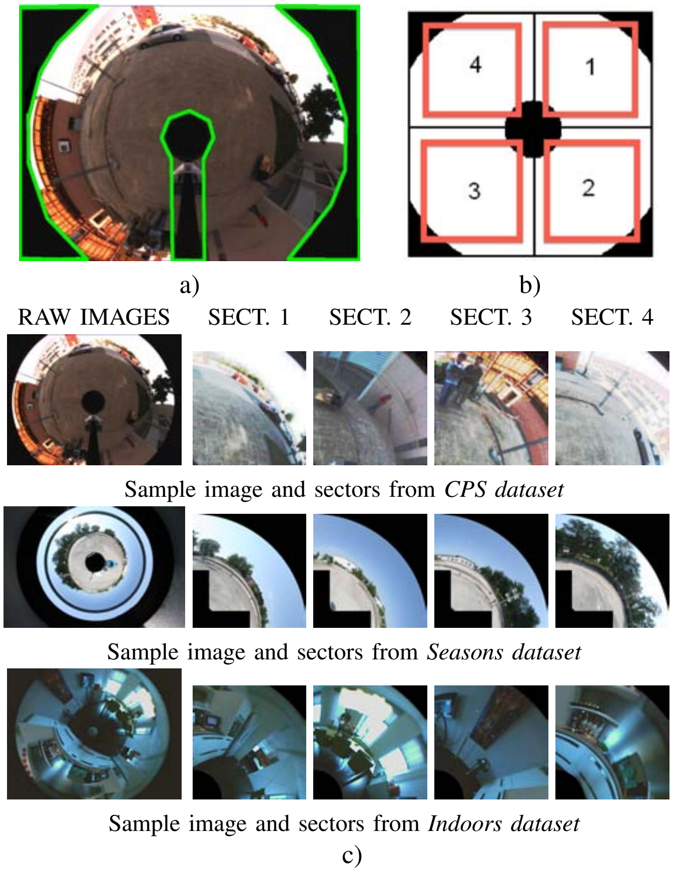
The omnidirectional image gist then *omni-gist* for short, consists of a vector of four conventional gist descriptors computed over the 4 sections we have divided the raw omnidirectional image: $\mathbf{g}_{omni} = [g_1 \;\; g_2 \;\; g_3 \;\; g_4]$.

B. Omni-gist similarity measure

The omni-gist captures the basic structure of the scene contained in the image, together with the spatial relationship between those four parts, since the relative position between them is essential information of the scene. To efficiently arrange, sort and compare images based on this descriptor, we need to define a similarity evaluation process. Our proposal is based on the following steps.

a) Omni-gist alignment and distance: Given two omni-gist descriptors, g_{omni} and g'_{omni}, the distance $d_g(g_{omni}, g'_{omni})$ considered between them is the minimum distance that can be obtained from the four possible alignments of the four sectors of the image and is defined as:

$$\begin{aligned} d_g(g_{omni}, g'_{omni}) = \quad & \min(\, ||g_{1234} - g'_{1234}||, ||g_{2341} - g'_{1234}||, \\ & ||g_{3412} - g'_{1234}||, ||g_{4123} - g'_{1234}||\,). \end{aligned} \tag{1}$$



a) b)

RAW IMAGES SECT. 1 SECT. 2 SECT. 3 SECT. 4

Sample image and sectors from *CPS dataset*

Sample image and sectors from *Seasons dataset*

Sample image and sectors from *Indoors dataset*

c)

Fig. 2. a) Omnidirectional image with areas we should not process in green. b) Scheme of the image partition to get symmetric parts from each quadrant. c) Sample images and partitions from different datasets used in this work.

where $g_{omni} = [g_1 \;\; g_2 \;\; g_3 \;\; g_4]$, $g'_{omni} = [g'_1 \;\; g'_2 \;\; g'_3 \;\; g'_4]$, and g_{2341} is the cyclic permutation corresponding to the omni-gist descriptor $g_{omni} = [g_2 \;\; g_3 \;\; g_4 \;\; g_1]$. Considering the best of the four permutations, each corresponding to the rotation of 90^o, implicitly makes an assumption that if the places are revisited, the heading of the vehicle is related by the multiple of 90^o to the headings of images stored in the database. This assumption is enough if we have known restrictions on the camera motion, e.g., mounted on a vehicle that only moves along the direction of the roads/ways in urban environments. Otherwise, this discrete and unique way of dividing the image in 4 pieces can become a drawback for the approach if it is not done carefully. Two images taken in the same place but rotated 45^o can produce gist descriptors very different to each other. This problem can be addressed by storing for each reference image different omni-gist values corresponding to different partitions rotated angles different than 90^o.

As the distance between the composite gist descriptors we use the mean distance between individual quadrants

$$||g_{1234} - g'_{1234}|| = \frac{1}{4}(||g_1, g'_1||, ||g_2, g'_2||, ||g_3, g'_3||, ||g_4, g'_4||). \quad (2)$$

b) Gist vocabulary: Mainly for efficiency reasons we also explore a similarity measure based on the quantized version of gist descriptors. In order to obtain it we compute a gist vocabulary $\mathbf{V}_{gist}$ by clustering all single gist descriptors estimated on each sector of each image into n clusters. As the result of this stage we obtain:

- The vocabulary $\mathbf{V}_{gist} = \{w_1, ..., w_n\}$ composed of n clusters.
- The centroid of each cluster i and its associated representative gist value $\bar{w}_i$.
- The distance matrix D_w, between all gist cluster centroids in the vocabulary $D_w(i,j) = \|\bar{w}_i - \bar{w}_j\|$

In our experiments we use a standard k-means algorithm for clustering. The number of clusters, is selected through a semi automatic iterative process. We try to get as few clusters as possible to keep the process as efficient as possible. Starting with a small number of clusters, we run k-means and evaluate the in-cluster compactness, and keep increasing the number of clusters until the compactness stops increasing over a worth-threshold.

Fig. 3 shows a couple of examples of the set of images that get clustered in the vocabulary for one of the datasets used later in the experiments, with n=18. Each group of snapshots corresponds to one of the gist clusters.

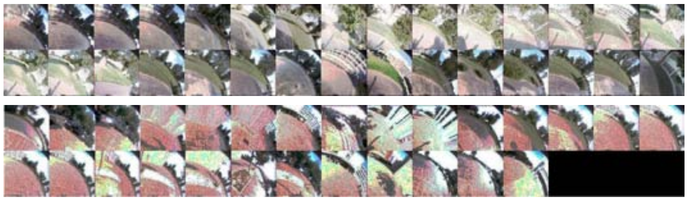


Fig. 3. Two of the clusters/words obtained in the vocabulary extraction with the CPS dataset used in the experiments.

When working with this vector quantized (VQ) vocabulary, the similarity measure between two omni-directional images $d_g^{VQ}(g_{omni}, g'_{omni})$ is defined based on the distance between the closest cluster centers the original gist descriptors belong to. Each of the four gist values composing the omni-gist descriptor is assigned the nearest gist word from $\mathbf{V}_{gist}$

$$\begin{aligned} g_{omni} &= [g_1 \;\; g_2 \;\; g_3 \;\; g_4] -> [w_a \;\; w_b \;\; w_c \;\; w_d] \\ g'_{omni} &= [g'_1 \;\; g'_2 \;\; g'_3 \;\; g'_4] -> [w_e \;\; w_f \;\; w_g \;\; w_h] \end{aligned}$$

Then, the distance between two catadioptric images is computed as the average distance between the centroids of the words assigned using the best possible alignment as explained above in a).

$$d_g^{VQ}(g_{omni}, g'_{omni}) = d_g([\bar{w}_a, \bar{w}_b, \bar{w}_c, \bar{w}_d], [\bar{w}_e, \bar{w}_f, \bar{w}_g, \bar{w}_h]) \quad (3)$$

Note that in the process of building the vocabulary, we compute the distance matrix between all gist centroids D_w, what provides further efficiency for the similarity evaluation process using d_g^{VQ}.

C. Effects of unwarping

Before we proceed with the description of the experiments on topological mapping and localization, we examine the effects of the unwarping on the proposed similarity measures.

A typical doubt when processing the omni-directional images is whether to unwarping or not the raw image. Thanks to a conversion from polar to cartesian coordinates and to standard interpolation approximations in the lower resolutions areas (center of the image), we can obtain an unwarped image as shown in Fig.4.

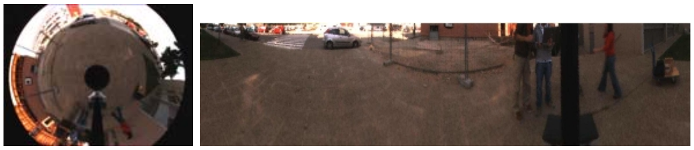


Fig. 4. A raw omnidirectional image (left) and result of unwarpping it (right).

From the following results, we can conclude that for our goals, it is not beneficial to unwarp the raw omnidirectional image. We examine this in the context of image based localization experiments, where given a database of reference images S_{train} and corresponding set of test images S_{test} we compare the localization performance using gist descriptors computed for unwarped or original/raw omni-directional images. As ground truth in this localization experiment we consider the GPS tags of each image, and require that distance between query and reference image selected should be under a threshold. This threshold is of 10 m and reference images are separated 5 m from each other. More details on the dataset can be found in the experimental section V.

Table I shows the percentage of correct localizations obtained if we keep the original image or unwarp it. Variations of this test, with different reference images, criteria and distances showed the same effects and differences between the use of raw or unwarped omnidirectional images. We can observe that the performance increases if we extract the gist in the raw images. Moreover, difference in performance is higher when we run more difficult tests, i.e., we have less reference images and they are further apart. The gist vocabulary $\mathbf{V}_{gist}$ obtained is different with raw than unwarped images, and the vocabulary from unwarped images turned out to be less meaningful, because the unwarpping process makes the area close to the camera/robot occupy more pixels, therefore it affects more the gist descriptor values. Since in this case the closer part is floor/ground and is less discriminative than scene structures away from the camera, the gist clustering on unwarped images turns out to be less significant. Gist computation over unwarped images takes into account with the same importance structures of important elements of the scene, further from camera, and textures on the floor close to the camera. This could be not so important indoors, where scene elements are closer to the camera, but from our tests it appears to be a key issue outdoors.

TABLE I

CORRECT LOCALIZATION WITH RAW OR UNWARPED OMNI-IMAGES.

# images in $TrainSet$	*Raw images*	*Unwarped images*
251 (every 5^{th})	97.6 %	97.6 %
126 (every 10^{th})	92.02 %	85.2 %
83 (every 15^{th})	91.22 %	82.07 %

Since the performance is even better for original, not unwarped, views and the efficiency is clearly improved as well if we avoid the costly unwarping process, we will just work on the raw images for the rest of the paper.

IV. HIERARCHICAL APPEARANCE BASED LOCALIZATION AND MAPPING

Given the proposed similarity measure between two omni-directional images, we now describe a method for topological mapping and localization algorithms. Hierarchical approaches that go from coarse to finer steps for image similarity evaluation have been successfully designed in many previous works on visual localization, such as [11, 20]. We also propose a hierarchy, not only on the localization approach but also on the mapping. Both algorithms consist, in brief, of two parts. In the first stage we use the omni-directional gist descriptor and its associated similarity for both topological mapping and location recognition. In case the decisions can no longer be established with sufficient confidence and we have several candidate locations available, we refine the search using more accurate local feature based image similarity measures. We demonstrate in the following section the proposed similarity measure and associated topological mapping and localization techniques on three different datasets.

A. Topological map building.

The appearance based topological mapping consists of arranging a set of omnidirectional images into similar areas or clusters connected to each other. To build a topological map of the environment online, we propose a hierarchical process built on the same basis of previous online approaches [11, 12]: first, to compare consecutive images and keep them in the same cluster if similarity is high enough (under t_{high}), otherwise consider the option of revisiting particular location (loop closure) of a previously initialized cluster and if neither of these cases occurs, initialize a new cluster. In order not to obtain too fragmented topological maps due to large changes of visual appearance in nearby locations, we represent individual locations by set of representative sub-clusters, with elements that are particularly similar to each other (under t_{low} threshold distance). A more intuitive idea of the advantages of this sub-clustering results can be seen at the experimental results in next section.

In brief, the proposed mapping process starts initializing the centroid of current sub-cluster with first image. Then for every new image, the distance d_g between current image gist and current sub-cluster centroid image is computed. If this distance is below t_{low} threshold, we add this new image to current sub-cluster and update the average omni-gist value of the corresponding cluster and sub-cluster. Otherwise, in case the current omni-gist is similar enough, distance below t_{high}, with any previously visited cluster centroid, we have a loop detection candidate and then we will refine this possible loop-closure hypothesis by local feature matching. We use SURF features [21] to match current view and selected candidates from previous cluster centroid images, and follow simple voting strategy to establish the presence of previously visited cluster. In case none of these cases occurs, we just initialize a new cluster. The following Algorithm 1 details a bit more this process.

```
Input: Sequence of omni-images
Result: Topological map
/* obtain sequence's clusters and sub-clusters */
/* Start map building */
for i= 1 to num_images do
    $g_{omni}(i)$ = computeOmniGist($Im_i$);
    if $||g_{omni}(i), g_{omni}(currentSubCluster)|| < t_{low}$ then
        /* Still in same sub-cluster */
        ReEstimateCentroid($currentCluster$);
        ReEstimateCentroid($currentSubCluster$);
    else
        /* Re-visit or New Cluster? */
        Revisit=0;
        /* Sort clusters according to distance
           from centroid to current image gists */
        sortedClusters=sort($allClusterCentroids$, $d_g$);
        for nCluster in sortedClusters do
            if $||g_{omni}(i), g_{omni}(nClusterCentroid)|| < t_{high}$ &
            $SURFmatches(i, nClusterCentroid) > minM$
            then
                Revisit=$nCluster$;
                ReEstimateCentroid($currentCluster$);
                StartNewSubCluster($numberOfSubClusters + 1$);
            end
        end
        if Revisit==0 then
            StartNewCluster($numberOfClusters + 1$);
        end
    end
end
```

Algorithm 1: Online topological mapping based on a hierarchy of omni-gist descriptors and local features.

B. Similar image search: appearance based localization.

Another important task we propose how to solve in this work is to localize the robot/camera in its environment, given a set of reference images. We present a general method for appearance based localization and analyze later in the experiments how the localization performance can vary if we do take into account a topological map built from the reference images or if we just use the set of reference images as a visual memory. The appearance based localization simply tries to select the most similar image to the current view with regard to the set of reference views representing the visual map. As in previous hierarchical approaches [16], our method first selects a set of candidates, pruning the possible selected locations set and then select more carefully the most likely candidate. These are the steps of this process:

• We first select the top 30% nearest candidate views from the reference image set, according to the omni-gist descriptors similarity measure, d_g^{VQ} described in previous Section III-A, while providing the best alignment of each candidate with the query view. We have observed in our experiments, that this initial set of candidates contains at least a correct one for around 99% of the tests.

• The set of initial candidates is re-ranked using more accurate distance measure that estimates the distance between the actual omni-gist values, d_g, as opposed to quantized values associated with assigned visual words, and from here we select the top-k candidates.

• Finally, we compute local feature matches between the query and the top-k candidates and select as most likely the candidate with the largest amount of matches.

V. Experiments on mapping and localization

We have tested our proposed algorithms for topological mapping and localization based on omni-gist descriptors with three different datasets (see a sample view from each at previous Fig. 2). Two of them are outdoors (CPS, Seasons), one indoors. The goal of these experiments is to evaluate the performance of our proposal and compare it to other approaches that only rely on local features. For the outdoor experiments, we show the topological maps obtained with the method described in Sec. IV-A and comment their differences with regard to previous works using the same datasets [22, 12]. For all experiments, we present the localization results comparing three methods:

• VQ_{omni} denotes the proposed localization method using the quantized omni-gist descriptor and local features for final refinement as described in Sec. IV-B.

• E_{omni} method does not use the quantization stage and computes the exact (E) distances between omni-gist descriptors from query image and the whole set of reference images.

• V_{surf} is a well known strategy for image similarity evaluation based on local features, bag of features, using SURF [21] and the implementation provided by Vedaldi[1] of a simplified version of the proposal from [2].

Regarding other practical issues relevant to these experiments, when computing the d_g (eq. 1,2), we have evaluated the performance computing the distance between individual quadrants using L1 and L2 norms (Euclidean), with comparable results for both. Besides, computing the average distance from only the three lower quadrants also increases the performance in the experiments, probably due to a higher robustness to noisy parts and occlusions in certain image quadrants. The evaluation of the localization accuracy is done using two different criteria:

• *GPS-dist* criterion accepts the localization selected if it is under a certain distance from the query, according to the GPS tags of each image.

• *Topological-dist* criterion uses the topological map of the environment and considers a localization correct if the image selected as most similar belongs to the same cluster than the query image. This measure can also be useful to validate a

[1] http://www.vlfeat.org/∼vedaldi/code/bag/bag.html

topological map and check if the localization results are similar to those obtained based on GPS data.

A. Experiment 1: CPS

This first experiment uses the CPS dataset [22], that has been acquired with the vision system shown in Fig. 1. It consists of 1260 catadioptric images acquired during an outdoor trajectory of around 500 m, with several loops. An overview of the dataset is shown in Fig. 5 and Fig. 2c) includes a sample raw image from this dataset and its four sectors (each sector used is reduced to 256×256 pixels, where the gist descriptor is computed).

Reference image set in this dataset consists of 126 images, equally spread over the trajectory, every 10^{th} image, approximately every 5 m. For testing we use another 126 images of the same data, equally spread and taking the image in between each reference image, as separated as possible, so the closest reference image for each test is located at around 2.5 m.

Online Mapping. Fig. 5 shows the resulting topological map obtained for the reference image set of this experiment dataset. In comparison with the map obtained in previous works [22] on the CPS dataset, estimating an offline clustering based on local feature matches only, the current proposal result looks more homogeneous and re-visited areas are better detected. Currently all re-visited areas are detected, and the accuracy of this detections, if we take into account how many of the images where correctly assigned to the re-visited cluster is around 90%.

Each reference image is represented with a colored dot of the cluster where it has been assigned, and the images that represent cluster or sub-cluster centroids have a $+$ or o respectively. The fact of sub-dividing the clusters into sub-clusters when particularly high intra-cluster similarity is detected, can be a good hint to automatically establish how many reference images is good to keep as ”way-points” to facilitate the use of this map for robot path planning and navigation. Later in the localization tests, we can observe how the cluster or sub-cluster centroids can be used as the only information kept to represent the environment without losing important information.

Localization. The V_{gist} extracted from the reference images of this dataset was of 18 words ($n = 18$). The number of words is established after an iterative process to evaluate which n gives better intra-cluster (word) similarity. The topological map used to evaluate the topological-dist criteria is the map obtained in mapping experiments above. The threshold to accept the localization based on GPS data is of 10 m, while reference images are separated 5 m from each other.

Table II shows a summary of the localization results obtained with each of the three approaches and different criteria previously described. Last column, average search time, presents the average time, executing all methods in Matlab and in the same computer, for searching the most similar image with each approach (just the search of the most similar image). Note that results obtained using omni-gist descriptors as basic similarity evaluation step are of comparable or better quality to those obtained using only local features. Besides, quantized gist descriptors present reduced computation and storage requirements. Fig. 6 presents a more visual representation of the localization results obtained in this experiment for the proposed approach VQ_{omni} ranking the final top-3 candidates with SURF matches. Red lines represent incorrect localization

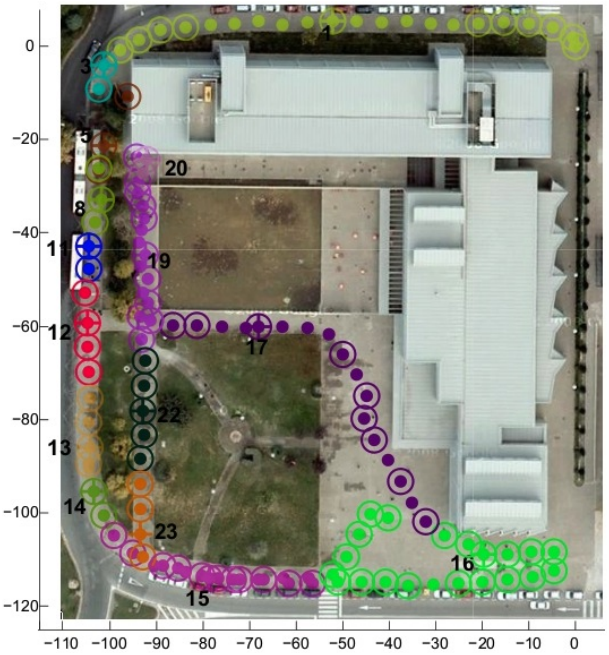


Fig. 5. Experiment 1. Topological map obtained with our online approach based on omni-gist and SURF descriptors.

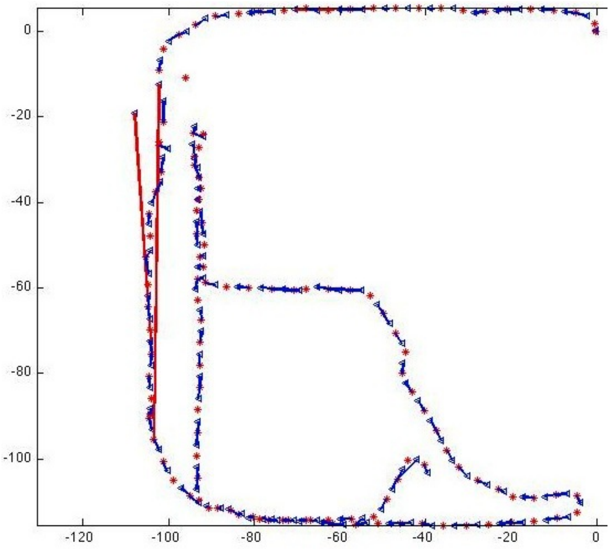


Fig. 6. Experiment 1. Appearance based localization with VQ_{omni} (3). Reference images are e marked with $*$ (red), and query images are marked with a $<$ (blue). Lines represent which reference image is selected as closest to each query image. Blue lines are accepted by the GPS-dist (10 m) while red lines are considered errors.

results according to the GPS-dist, so we can appreciate very few (2) errors in this experiment.

Another important issue to be analyzed is if we could obtain similar results with less reference images, and how to select the most representative images. Indeed, once we have the topological mapping of the environment it could be a good starting point. Table III presents the localization results obtained if we use all reference images, only the map cluster centroid images or only the map sub-cluster centroid images. Column #im. points the number of images used in each case. These results confirm the usefulness of the sub-cluster subdivision, since using only their centroid image provides similar results that using all reference images, so this construction provides an automatic way of selecting which reference images are really necessary for a complete representation of the environment.

TABLE II

EXPERIMENT 1: CORRECT LOCALIZATION ACCORDING TO DIFFERENT CRITERIA. CPS DATASET, 126 IMAGE TRAINSET (SEPARATED $\approx$ 5 M)

Localization approach (k)*	GPS dist.	Topological dist.	Average search time $\star$
VQ_{omni} (0)	92.06 %	96.52 %	0.037 s.
VQ_{omni} (3)	97.62 %	97.46 %	0.037 + 3 × S_t s.
VQ_{omni} (5)	98.41 %	96.64 %	0.037 + 5 × S_t s.
E_{omni} (0)	92.06 %	96.58 %	0.179 s.
E_{omni} (3)	97.62 %	97.50 %	0.179 + 3 × S_t s.
E_{omni} (5)	99.21 %	97.52 %	0.179 + 5 × S_t s.
V_{surf}	96.03 %	95.65 %	0.45 s.

* (k) = # of top candidates evaluated using SURF for a final decision.
$\star$ S_t = time to match SURF features in two images.

TABLE III

EXPERIMENT 1 & 2: APPEARANCE BASED LOCALIZATION USING WHOLE OR REDUCED (ONLY CENTROIDS) REFERENCE IMAGE SET.

Localization approach	CPS set			Seasons set C		
& reference images used	Topo-dist	search t.	#im.	Topo-dist	search t.	#im.
VQ_{omni} (0) - all ref. im.	96.52 %	0.04 s.	126	88.57 %	0.04 s.	149
— sub-cluster centroid im.	95.69 %	0.03 s.	93	85.10 %	0.035 s.	120
— cluster centroid im.	42.62 %	0.01 s.	15	45.51 %	0.01 s.	21
E_{omni} (0) - all ref. im.	96.58 %	0.11 s.	126	90.71 %	0.15 s.	149
— sub-cluster centroid im.	95.76 %	0.10 s.	93	88.65 %	0.12 s.	120
— cluster centroid im.	80.00 %	0.02 s.	15	70.20 %	0.03 s.	21

B. Experiment 2: Seasons

This second experiment uses the longest sequence available, round C, from Seasons dataset [13]. We are only using the outdoor parts of this sequence, 945 views, and taking as reference set 149 equally spread images along this trajectory, separated approximately 6 or 7 meters from each other. Test images are another 149 image set selected from the dataset images as separated as possible from any reference image (in this case each test has the closest reference image at 3 or 3.5 m). An overview of the dataset is shown in Fig. 7 and Fig. 2c) includes a raw image of this dataset and the four sectors it is divided to estimate the omni-gist (each sector is 179 × 179 pixels).

Online mapping The topological map obtained for the reference image set with our proposal in this experiment is shown in Fig. 7. We can also compare the topological map obtained in this experiment to previous results obtained with the same dataset (see Fig. 4 from the results presented in [12]). Authors from this previous work were using all reference images from the trajectory and building the map using local features from the images. It is hard to compare the quality in this case, what we can conclude is that the arrangement obtained with the current proposal looks more compact and regarding the important issue of re-visit detection, it succeeds in similar areas (see parts with a green circle) and has trouble also in similar places (marked with red circles) than the local feature based mapping results. In particular, the area around $[60, 60]$ coordinates, seems to have more random movements with rotations different from multiples of 90^o, so this is probably making the system fail due to the unique discretization of the omnidirectional image into four sectors. It is likely that storing multiple segmentations for each reference image would improve the re-visit detection performance in those areas.

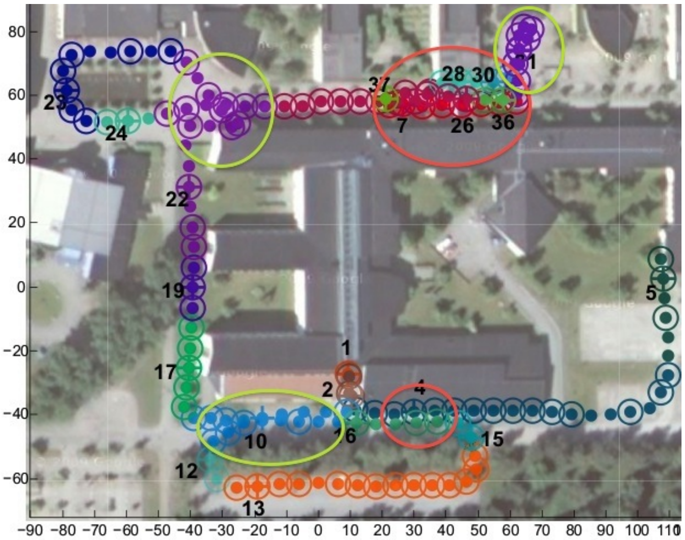


Fig. 7. Experiment 2. Topological map obtained with our online approach based on omni-gist and SURF descriptors.

Localization The V_{gist} extracted for this dataset was of 22 words ($n = 22$). The topological map used to run the topo-dist criteria is the map obtained in previous mapping experiments, with 21 clusters. The threshold to accept the localization based on GPS data is of 10 m. Note that in this experiment reference images are separated 6 or 7 meters from each other.

Table IV shows similar summary than in previous Experiment 1 for the localization results with different approaches and different acceptance criteria. Again we can observe the same behaviour: comparable quality in the localization and interesting efficiency gains with the quantized approach based on omni-gist similarity VQ_{omni}.

The results for this experiment using as reference images only the cluster or sub-cluster centroids are shown in previous Table III, together with previous experiment results. Again we can observe that to achieve similar performance than keeping all reference images (every 6 meters), we should use the sub-cluster centroids. Also similar graphical summary of localization results than the one presented in previous experiment, is shown in Fig. 8 for this experiment.

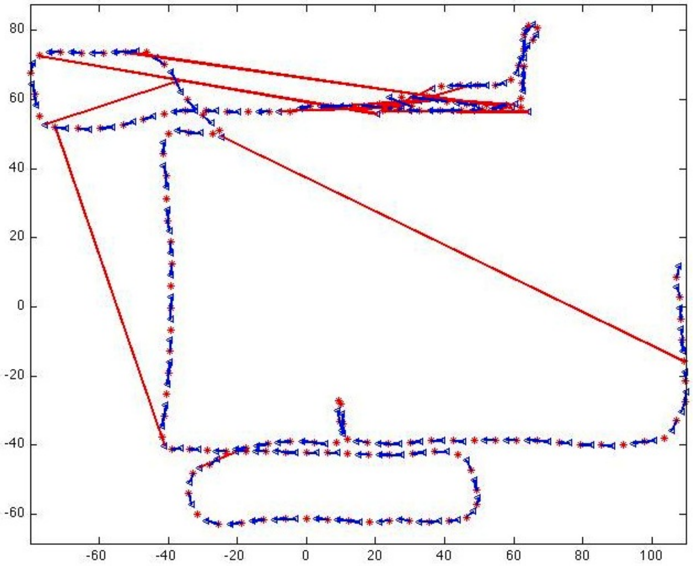


Fig. 8. Experiment 2. Appearance based localization with VQ_{omni} (3). Reference images are e marked with $*$ (red), and query images are marked with a $<$ (blue). Lines represent which reference image is selected as closest to each query image. Blue lines are accepted by the GPS-dist (10 m) while red lines are considered errors.

TABLE IV

EXPERIMENT 2. CORRECT LOCALIZATION ACCORDING TO DIFFERENT CRITERIA. SEASONS DATASET, 149 IMAGE TRAINSET (EVERY $\approx 6, 7$ M)

Localization approach (k)*	GPS dist	Topological dist	Average search time $\star$
VQ_{omni} (0)	83.89%	88.57%	0.04 s.
VQ_{omni} (3)	91.28%	92.91%	0.04+ 3 × S_t s.
VQ_{omni} (5)	93.96%	94.33%	0.04+ 5 × S_t s.
E_{omni} (0)	88.59%	90.71%	0.18
E_{omni} (3)	94.63%	93.43%	0.18 + 3 × S_t s.
E_{omni} (5)	96.64%	94.85%	0.18+ 5 × S_t s.
V_{surf}	87.92 %	81.21 %	0.52 s.

* (k) = # of top candidates evaluated using SURF for a final decision.
$\star$ S_t = time to match SURF features in two images.

C. Experiment 3: Indoors

For better completion of the evaluation of the proposal, we have run a last experiment on a indoor dataset, provided in [23], to get an overview of the performance this method can offer at indoor settings. In this case, instead of GPS tags, we use the corresponding improved odometry information provided as ground truth with the dataset[2] together with different comments regarding the different runs of the dataset. Reference information is extracted from run 1 of the dataset, a clean trajectory without occlusions, to be organized into a topological map that represents the different rooms of the environment, see Fig. 9 (a), in this case manually obtained to avoid clustering errors in this localization test. We use 87 reference images, extracted every 0.5 meters along this run. Test images are from another run, number 5, corresponding to a run with much more dynamic changes, such as open doors or people moving around. We use 209 images from this run for testing localization, extracted every 0.25 meters along this run.

Localization In this case the gist-vocabulary, V_g has been obtained as a 14-word vocabulary. Similarly to previous experiments, Table V shows the summary of the localization results for the test images of this experiment. In this case we include only the results of the methods based on omni-gist distances without using SURF features at all in the localization process. The threshold to accept a localization as correct, Odometry-dist, is established at 2 meters. We obtain good results with a tight threshold, reference images are separated 0.5 m from each other, and test images are from a different run and trajectory than reference images. The main difference we observe in this results with regard to previous experiments is that the brute force gist comparison with exact distances E_{omni} gives slightly better results this time. This can be due to the fact that this experiment is run indoors, then the dataset images contain significant changes, even between consecutive reference images, due to the proximity of the objects and obstacles and the dynamic changes, therefore the quantization of the omni-gist is not as good or representative as in outdoor settings. This is maybe the reason as well to obtain lower results with the local feature quantization V_{surf}, and makes that the vocabulary obtained is not representative enough to obtain optimal results.

TABLE V

EXPERIMENT 3. CORRECT LOCALIZATION INDOORS. ODOMETRY DISTANCE THRESHOLD FOR ACCEPTANCE = 2 M.

Localization approach	Odometry dist	Topological dist	Average search time
VQ_{omni} (0)	87.56 %	90.91%	0.04 s.
E_{omni} (0)	95.21 %	96.17%	0.11 s.
V_{surf}	80.38%	80.38 %	0.5 s.

VI. CONCLUSIONS

This work analyzes how to adapt the gist descriptor to catadioptric images, using a 4-gist descriptor for each image, omni-gist. We define a similarity measure to compare omnidirectional images based on the omni-gist and present a new approach for topological mapping and appearance based localization. Both methods rely on the omni-gist descriptor and integrate local feature matching to refine decisions such as loop-closure detection. We show that for the current goals, the typical unwarping of the catadioptric images is not beneficial. Besides, the extensive experimental results, with three different

[2] http://staff.science.uva.nl/∼zivkovic/FS2HSC/dataset.html

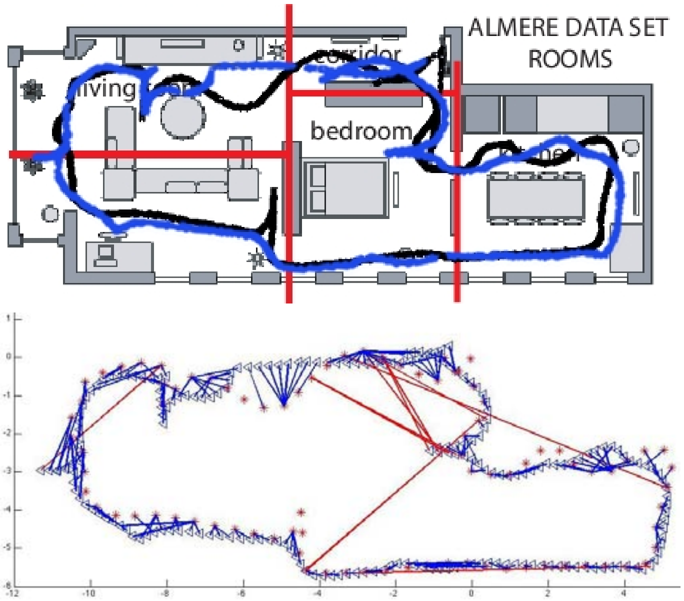


Fig. 9. Experiment 3. Localization with indoors dataset. Top image shows the environment scheme and the room-clustering (red lines are cluster limits) and the odometry from runs 1(black) and 5 (blue). Bottom image shows the localization results summary with the E_{omni} (0) approach.

omnidirectional image datasets from real scenarios, are used to demonstrate the performance of these methods, providing comparable results both for appearance based mapping or localization than previous approaches based on local features only, and demonstrating the efficiency improvements that the use of the proposed methods can provide.

Acknowledgments

This work was supported by spanish project VISPA, DPI2009-14664-C02-01 and US National Science Foundation Grant No. IIS-0347774.

References

[1] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. of Computer Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[2] D. Nistér and H. Stewénius, “Scalable recognition with a vocabulary tree,” in *Proc. of IEEE CVPR*, 2006, pp. 2161–2168.

[3] J. Sivic and A. Zisserman, “Video Google: Efficient visual search of videos,” in *Toward Category-Level Object Recognition*, ser. LNCS. Springer, 2006, vol. 4170, pp. 127–144.

[4] 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.

[5] A. Torralba, R. Fergus, and Y. Weiss, “Small codes and large databases for recognition,” in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[6] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, “Evaluation of gist descriptors for web-scale image search,” in *International Conference on Image and Video Retrieval*, july 2009.

[7] J. Santos-Victor, R. Vassallo, and H. Schneebeli, “Topological maps for visual navigation,” in *Int. Conf. on Computer Vision Systems*. Springer-Verlag, 1999, pp. 21–36.

[8] I. Ulrich and I. Nourbakhsh, “Appearance-based place recognition for topological localization,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2000, pp. 1023–1029.

[9] M. Cummins and P. Newman, “Highly scalable appearance-only slam - fab-map 2.0,” in *Robotics Science and Systems (RSS)*, Seattle, USA, June 2009.

[10] Z. Zivkovic, O. Booij, and B. Krose, “From images to rooms,” *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 411–418, 2007.

[11] T. Goedemé, M.Nuttin, T. Tuytelaars, and L. Van Gool, “Omnidirectional vision based topological navigation,” *Int. J. of Computer Vision*, vol. 74, no. 3, pp. 219–236, 2007.

[12] C. Valgren and A. J. Lilienthal, “Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2008, pp. 1856–1861.

[13] ——, “Sift, surf & seasons: Appearance-based long-term localization in outdoor environments,” *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[14] B. C. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman, “Object recognition by scene alignment,” in *Advances in Neural Information Processing Systems*, 2007.

[15] J. Košecká and F. Li, “Vision based topological Markov localization,” in *IEEE Int. Conf. on Robotics and Automation*, 2004, pp. 1481–1486.

[16] 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*, vol. 55, no. 5, pp. 372–382, 2007.

[17] A. C. Murillo and J. Kosecka, “Experiments in place recognition using gist panoramas,” in *9th IEEE Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (OMNIVIS), with Int. Conf. on Computer Vision*, 2009, pp. 2196–2203.

[18] G. Singh and J. Kosecka, “Visual loop closing using gist descriptors in manhattan world,” in *Omnidirectional Robot Vision workshop, with IEEE Int. Conf. on Robotics and Automation*, 2010.

[19] A. Oliva and A. Torralba, “Building the gist of a scene: The role of global image features in recognition,” *Visual Perception, Progress in Brain Research*, vol. 155, 2006.

[20] J. Wang, H. Zha, and R. Cipolla, “Coarse-to-fine vision-based localization by indexing scale-invariant features.” *IEEE Trans Syst Man Cybern B Cybern*, vol. 36, no. 2, pp. 413–422, April 2006.

[21] 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/.

[22] A. C. Murillo, P. Abad, J. J. Guerrero, and C. Sagüés, “Improving topological maps for safer and robust navigation,” in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2009, pp. 3609–3614.

[23] *Data set from FS2HSC - IEEE/RSJ IROS 2006 Workshop*, http://staff.science.uva.nl/∼zivkovic/FS2HSC/dataset.html.