Abstract—In the context of vision-based topological navigation, detecting loop closures requires to compare the robot’s current camera image to a large number of images stored in the map. For efficient image comparisons, we apply distance functions to global image-descriptors, i.e. low-dimensional descriptors derived from the entire panoramic images. To identify promising combinations of descriptors and distance functions, we formulate the loop-closure detection as a binary classification problem and analyze the resulting receiver operator characteristics (ROC). The results of comparing a wide range of descriptors and distance functions reveal that reliable loop-closure detection is possible with a single 16- to 128-dimensional image-descriptor based on gray-value histograms or Fourier descriptors and that all considered distance functions have a comparable performance.

I. MOTIVATION

For map-building applications of mobile robots it is essential to correctly recognize places which have already been visited. This problem is referred to as loop-closure problem [1], [2], [3]. Correctly detecting loop closures is not only essential for SLAM and other position-based navigation methods (the estimate of the robot’s position might drift from the true position) but also for position-less topological navigation methods. Both position-based and position-less navigation strategies have to rely on external sensor cues to detect loop closures. In case loop closures are not detected correctly, the resulting map will get inconsistent, and navigation algorithms can fail.

In this paper, we propose a parsimonious appearance-based method to detect loop closures. We assume a topological representation (with or without additional position information) of space which stores panoramic images. Detecting loop closures is then a matter of comparing the robot’s current image with all the images stored in the map. For this reason, visual loop-closure detection is closely related to image retrieval (review: [4]). With their full 360° azimuthal field of view, panoramic images are well suited for vision-based loop-closure detection since images taken at identical positions in space but with different orientations of the robot contain the same visible image information. If operating on cylindrical images obtained from unfolding the original camera image (Fig. III), only the horizontal positions of the imaged features are shifted depending on the robot’s orientation (assuming that the robot moves in the plane and rotates around its vertical axis). Instead of comparing images pixel-by-pixel, we transform each cylindrical image into a single descriptor with a low dimensionality and use these descriptors to compute the image dissimilarities. By choosing appropriate image transformations, this approach does not only allow for more efficient image comparisons but also computes descriptors which are invariant against changes of the robot’s orientation. Hence, two images taken at identical positions but with different robot orientations can be recognized as identical without being aligned w.r.t. a common reference direction as required for pixel-by-pixel comparisons. If the descriptors are stored in a topological map, they have to be computed only at the time the corresponding place node is added to the topological map. In the following steps, they can be reused for image comparisons.

The methods proposed in this paper will be used for navigation of cleaning robots, where correctly detecting loop closures is a prerequisite to avoid repeated coverage and uncleaned areas. In conjunction with an extension of the navigation strategy described in [5], the loop-closure detection methods will be integrated into a framework capable of completely covering entire rooms or apartments by combining multiple segments of parallel lanes. Our navigation strategy [5] builds up a dense topological map of the environment storing snapshots acquired at regular distances and along parallel lanes (Fig. III). This future application strongly influences the methods considered in this paper: (i) we focus on methods with a low computational complexity, which can be used on the on-board computers of autonomous cleaning robots, and (ii) the methods should accurately identify the closest snapshot image stored in the dense topological map instead of identifying nearby locations.

The remainder of the paper is structured as follows: Related work from the field of omnidirectional visual robot navigation is discussed in section II. Section III describes our approaches to loop-closure detection in detail. Experiments and data evaluation methods are outlined in section IV, and the results are discussed in section V. Section VI summarizes the paper and points out future working directions.

II. RELATED WORK

Topological maps represent places by the sensory information obtained at those places [1]. In the case of vision-based mapping and localization, [6] propose a categorization into representing places (i) by the raw image data, (ii) by a set of local feature descriptors computed at points of interest, and
In the following, we will briefly discuss these three approaches.

In case the entire image without further processing is used to characterize places, pixel-by-pixel comparisons are required to derive a measure of the image dissimilarity. Hence, the computational complexity of the image comparison is strongly influenced by the size of the images. As common image distance functions (review: [7]) are not invariant under rotations of the robot, the images have to be aligned w.r.t. a common reference direction. Although this can be solved by applying a visual compass method (e.g. [8]), considerable additional computational effort is required.

Characterizing places by the raw image data is frequently used for topological navigation methods which derive spatial relations based on local visual homing methods (e.g. [9], [10], [5]; review of homing methods: [11]).

For characterizing places by a set of visible features, points of interest have to be detected and local feature descriptors have to be computed (e.g. SIFT [12] or SURF [13]; review: [14]). Place recognition can be done by establishing correspondences between two images. The drawback of this method is that a large number of features is required to accurately recognize places and that the dimensionality of the feature descriptors is rather large (64- or 128-dimensional in the case of SIFT and SURF) making feature matching computationally demanding. Such a definition of places is often used for localization (e.g. [15], [16]) and trajectory-based SLAM (e.g. [17], [18]).

If places are characterized by a global image-descriptor, the entire image has to be transformed to a lower-dimensional representation, which is then stored in the topological map. As we apply image processing operations with low computational complexity and as we compute image descriptors with a low dimensionality, the computation time and the additional storage capacity required to compute and store the image descriptors are negligible. In an ideal case, the resulting image descriptor exhibits the following four properties: (i) the descriptor changes depending on the robot’s position in space, (ii) the dissimilarity of two descriptors increases with increasing spatial distance between the places of image acquisition, (iii) the descriptor is invariant under rotations of the robot, and (iv) the descriptor is robust against environmental changes such as dynamic scene changes or changes of the illumination. In the context of topological navigation, image descriptors such as color histograms [19], [20], [21], gray-value histograms [6], color statistics [20], eigenspace representations (rotation dependent [22], [23] and rotation invariant [24]) and absolute Fourier coefficients [25] have been used. Invariance against rotations of the robot is usually achieved because these transformations derive image descriptions independent of exact pixel positions.

III. APPEARANCE-BASED LOOP CLOSURE DETECTION

For detecting loop closures, we pursue the following approach (Fig. III): The robot’s current camera image $C_p$ is unfolded to a cylindrical image $I_p$, which is used to compute a lower-dimensional image descriptor $p$ by applying a signature function $p = s(I_p)$. In a second step, the current descriptor $p$ is compared to each of the descriptors $q_i$ stored in the topological map by applying a distance function $d(p, q_i)$ yielding a scalar measure $\ell$ of the image dissimilarity. Based on $\ell$ and a threshold $\ell_{\text{th}}$, a decision is made whether the images $I_p$ and $I_i$ (and therefore the positions of their acquisition) are identical or not. Hence, we formulate the problem of loop-closure detection as a binary classification problem.

A. Signature Functions

As signature functions $s(I)$ we use the following functions computing rotation-invariant image descriptors based on gray-value histograms (hist and chist), on image statistics (mom3, mom4, and cog), and on Fourier coefficients (afc).

The signature functions

$$s_{\text{hist}}(I) = h(I) = (h_0, h_1, \ldots, h_{b-1})^T$$

and

$$s_{\text{chist}}(I) = c(I) = (c_0, c_1, \ldots, c_{b-1})^T$$

with $c_k = \sum_{i=0}^{b-1} h_i$

use the histogram $h(I)$ and the cumulative histogram $c(I)$ of relative gray-value frequencies (i.e. $\sum_{k=0}^{b-1} h_k = 1$) as descriptors [26]; both types of histograms contain $b$ bins. The $\text{mom}3$ and $\text{mom}4$ descriptors computed by the functions

$$s_{\text{mom}3}(I) = (m_1(I), m_2(I), m_3(I))^T$$

and

$$s_{\text{mom}4}(I) = (m_1(I), m_2(I), m_3(I), m_4(I))^T$$

use the first three and four statistical moments $m_k$ (mean, variance, skewness, and kurtosis) to represent the image [27].

The signature function

$$s_{\text{cog}}(I) = \left\| \frac{1}{n} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} I(x, y) \left( \frac{\hat{x}}{\hat{y}} \right) \right\|$$

with

$$n = \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} I(x, y)$$

and

$$\left( \frac{\hat{x}}{\hat{y}} \right) = y \left( \cos \left( \frac{2 \pi x}{w} \right), \sin \left( \frac{2 \pi x}{w} \right) \right)$$

computes the length of the image’s center of gravity (CoG) position vector as one-dimensional descriptor. If the length of the CoG vector were computed in Cartesian image coordinates $(x, y)$, it would depend on the robot’s orientation. To achieve rotational invariance, the CoG vector is computed w.r.t. an auxiliary polar representation of the image with the angular coordinate depending on the image column $x$ and the radius depending on $y$. For the purpose of vector averaging, this representation is expressed in Cartesian coordinates $(\hat{x}, \hat{y})$.

Image descriptors based on absolute Fourier coefficients were first proposed by [25]. Instead of applying a 1D-FFT to each row of the image $I$, we average over the columns of the image $I$ in order to obtain a 1D panorama $I$, which is
then transformed:

\[ s_{\text{safe}}(I) = (f_0, f_1, \ldots, f_{b-1})^\top \]
with

\[ f_k = \frac{1}{w} \sum_{x=0}^{w-1} I(x) \exp \left( -2\pi ki \frac{x}{w} \right). \]

The absolute values of the Fourier coefficients are used as elements of the descriptor in order to achieve rotation invariance by eliminating phase information. As most of the image information is contained in the lower-order coefficients, the descriptor vector is formed by the first \( b \) Fourier coefficients only [25].

B. Distance Functions

The dissimilarity of two image descriptors \( p \) and \( q \) is computed by applying a distance function \( d(p, q) \). For this purpose, we apply the norm functions

\[ d_{\text{max}}(p, q) = L_{\infty}(p, q) = \sum_{k=0}^{b-1} |p_k - q_k|, \]

\[ d_{\text{eucl}}(p, q) = L_2(p, q) = \sqrt{\sum_{k=0}^{b-1} (p_k - q_k)^2}, \]

\[ d_{\text{man}}(p, q) = L_1(p, q) = \sum_{k=0}^{b-1} |p_k - q_k|, \]

\[ d_{\text{max}}(p, q) = L_{\infty}(p, q) = \max_{k=0}^{b-1} |p_k - q_k|. \]

For histogram-based descriptors, we additionally use the histogram-specific distance functions summarized in [26]. As our objective is to develop methods which are in line with the computational power of on-board computers of autonomous cleaning robots (Sec. VI), we do not consider comparison functions with an iterative optimization process, such as the earth mover’s distance [26]. The Kullback-Leibler divergence

\[ d_{\text{kl}}(p, q) = \sum_{k=0}^{b-1} p_k \log \frac{p_k}{q_k} \]
as well as the Jeffrey divergence (a symmetric variant of the Kullback-Leibler divergence)

\[ d_{\text{jeff}}(p, q) = \sum_{k=0}^{b-1} \left( p_k \log \frac{p_k}{m_k} + q_k \log \frac{q_k}{m_k} \right) \]

with

\[ m_k = \frac{p_k + q_k}{2} \]
both measure the difference between probability distributions. The distance function

\[ d_{\text{chisq}}(p, q) = \sum_{k=0}^{b-1} \frac{(p_k - m_k)^2}{m_k} \]
is based on \( \chi^2 \) statistics, and thus computes the likelihood that the image descriptors are histograms of different distributions. In contrast to these functions, the quadratic-form distance

\[ d_{\text{qf}}(p, q) = \sqrt{(p - q)^\top A (p - q)} \]
also compares non-corresponding bins. Cross-bin matching is realized by the dissimilarity matrix \( A \) with entries

\[ a_{k_1, k_2} = 1 - \frac{d_{\text{qf}}(p, q)}{d_{\text{max}}} \]
depending on the ground distance (i.e. the dissimilarity or distance between bin representatives).

To decide whether two images are identical, the computed distance value \( \ell = d(p, q) \) can be used for binary classification w.r.t. the decision threshold \( \ell_t \).

C. Dimensionality of the Image Descriptors

The proposed descriptors have fixed dimensionalities (\( b = 1 \) for \text{cog}, \( b = 3 \) for \text{mom3}, and \( b = 4 \) for \text{mom4}) or dimensionalities \( b \) equal to the number of bins (\text{hist} and \text{chist}) and the number of Fourier coefficients (\text{afc}). The dimensionality of the computed image descriptor \( p \) influences the computational requirements for storing and comparing descriptors as well as their discriminability. Finding an appropriate descriptor dimension is a trade-off because the first aspect favors low-dimensional descriptors whereas a larger number of dimensions should facilitate more reliable loop-closure detection.

In order to increase the dimension of the image descriptors, we follow the approach of [6] and subdivide the cylindrical image \( I \) into \( r \) subpanoramas \( S_j \) (\( 0 \leq j \leq r - 1 \)) of equal height (in the original camera image, each subpanorama \( S_j \) corresponds to a concentric ring, Fig. III). For each of these subimages, an image descriptor \( p_j = s(S_j) \) is computed, and the resulting descriptors are combined to the descriptor \( p = (p_0, p_1, \ldots, p_{r-1})^\top \). Hence, the overall dimension of the resulting descriptor \( p \) is the product of
the subdescriptor’s dimensionality \( b \) and of the number of subimages \( r \).

For \( r > 1 \), we then compute the dissimilarity \( \ell \) of two image descriptors \( p \) and \( q \) by summing up the dissimilarities of their subdescriptors:

\[
d(p, q) = \sum_{j=0}^{r-1} d(p_j, q_j).
\]

### IV. Experiments and Evaluation

We performed database experiments in order to run a large number of tests for parameter optimization and to identify promising approaches. In our case, image databases consist of camera images taken with a real robot setup at a regular grid with a grid distance of 10 cm and therefore resemble the dense and regular structure of the topological map built by our navigation strategy [5]. The databases used (Tab. I and Fig. 2) were collected in several apartments under real illumination conditions. Since the collection of larger databases took several hours, these databases contain considerable illumination changes caused by changes of the daylight and of the weather conditions. For the \( \text{living} \) databases, we additionally collected images at identical positions under natural illumination during day and under constant artificial illumination during night. Cross-database experiments, i.e. pairing images from a \( \text{day} \) database with images from the corresponding \( \text{night} \) database (or vice versa), allow to test the methods under very strong changes of the illumination. For our test application, such drastic changes are unlikely to occur as we expect the robot to clean the accessible workspace within a maximum of 1 hour (also limited by the capacity of the robot’s batteries).

For each database, we compared each image \( I_p \) against each image \( I_q \) taken from the same database or —in case of cross-database experiments— taken from the corresponding cross-database. Thus, a total of \( (n_x \times n_y)^2 \) of image distance values were computed for each database. All images had a size of \( 461 \times 64 \) pixels. For both types of experiments, the second image \( I_q \) was randomly rotated and disturbed by zero-mean Gaussian noise with \( \sigma = 0.05 \). This procedure was repeated for several combinations of signature functions \( s \) and distance functions \( d \). Additionally, we systematically varied the number of subimages \( r \in \{1, 2, 4, 8, 16, 32\} \) and —for the histogram-based and Fourier-based image descriptors— the subdescriptor dimensionality \( b \in \{1, 2, 4, 8, 16, 32\} \). These parameter combinations resulted in descriptors with overall dimensions ranging from 1 to 1024.

Since we use image databases, the true outcome of the classification is known for our experiments, and we can compute the true-positive rate and the false-positive rate depending on the classification threshold \( \ell \). To avoid the assignment of a specific threshold \( \ell \) at the current state of our work, we used receiver operating characteristics (ROC) to analyze classification performance (details: [28]). ROC curves are obtained by plotting the true-positive rate versus the false-positive rate for any possible threshold \( \ell \). To compare and evaluate our experiments based on a scalar performance measure, we used the area under the ROC curve (AUC). The AUC yields 1.0 for a perfect classification and 0.5 for random classification.

As we aim to develop methods with an optimal performance over many different workspaces, we did not compute database-specific AUC values, but rather computed the AUC values after pooling the distance values obtained for several databases: The \( \text{day} \) group pools a total of 540,419 distance values obtained for the databases collected during day (i.e. \( \text{kitchen, living1/2/3/4day, moeller1/2, and roeben} \)). In the \( \text{day/night} \) group, a total of 108,252 distance values obtained for the cross-databases (i.e. \( \text{living1/2/3/4day and living1/2/3/4night} \)) are subsumed.

### V. Results and Discussion

To evaluate and interpret the data obtained by the experiments described in Sec. IV, we first identify promising combinations of signature and distance functions. In a second step, we then analyze the influence of the descriptor’s dimensionality on the classification performance for these methods.

For the first evaluation step (Tabs. I and II), we analyze the best AUC values obtained for each combination of signature...
ture and distance functions independent of the descriptor’s dimensionality. For all tested descriptors, the different distance functions yield almost identical AUC values. Therefore, the influence of the distance functions on the classification performance is not further discussed in the remainder of this section. For the day experiments (Tab. III), perfect classification (i.e. AUC values of 1.0) is achieved by the \textit{afc} and the \textit{cog} descriptors. The histogram-based descriptors \textit{hist} and \textit{chist} allow for very accurate classification with AUC values close to 1.0. The \textit{mom3} descriptor yields AUC values around 0.98, whereas the results for the \textit{mom4} descriptor show a considerably worse performance with AUC values around 0.81.

Tab. III shows the results for the day/night cross-database experiments. For all combinations of descriptors and distance functions, the AUC values reveal a strong decrease of the classification performance. With an AUC value of approximately 0.60, the best performance is achieved by the \textit{hist}, \textit{chist}, and the \textit{afc} descriptors. For this data set, the descriptors based on statistical signature functions (\textit{mom3}, \textit{mom4}, and \textit{cog}) only achieve AUC values close to 0.5. As for the day experiment, the results do not show considerable differences between the distance functions. For these cross-database experiments, one has to keep in mind that this data set contains changes of illumination much stronger than the illumination changes occurring during a single cleaning run of the robot, which we expect to take up to 1 hour. To this end, the results shown in Tab. III should be considered as the worst-case performance of the proposed methods.

The first step of data evaluation reveals (i) that the descriptors \textit{hist}, \textit{chist}, and \textit{afc} perform best and show almost identical performance, and (ii) that the choice of the distance function does not have a considerable influence on the performance. As our objective is to develop parsimonious methods for loop-closure detection, we only consider the \textit{hist} and \textit{afc} descriptors in combination with the maximum norm \(d_{\text{max}}\) for further evaluation. In Fig. 3, the AUC values are plotted depending on the descriptor’s overall dimension \(r \cdot b\) (please note the logarithmic scale with base 2 of the \(x\)-axes and the different scales on the \(y\)-axes). The shaded areas depict the ranges of AUC values between the maximum (best) and minimum AUC values. As identical overall dimensions can be obtained for several combinations of subpanoramas \(r\) and subdescriptor dimensions \(b\), the different markers used in the figures code the possible choices of \(b\) (see the caption of Fig. 3 for a complete list). For the \textit{hist} descriptor and the day experiments (Fig. 3a), the obtained performance ranges from chance level to almost perfect classification. The best performance with an AUC value of 0.996 is achieved with a 128-dimensional descriptor; for higher-dimensional descriptors, the performance decreases. For an overall dimension in the range from 4 to 128 dimensions, the best AUC values were all obtained for histograms with \(b = 4 \) bins \((\circ)\). Hence, the number of subpanoramas \(r\) ranges from 1 to 32. With only 1 dimension, the \textit{afc} descriptor (Fig. 3b) yields an AUC value of 0.981. For 2 and more dimensions, this combination allows for a perfect classification independent of \(r\) and \(b\).

For the day/night experiments, the \textit{hist} descriptor (Fig. 3c) achieves AUC values between chance level and 0.604. As for the day experiments, the best AUC value is obtained with an overall dimensionality of 128. Again, the AUC value decreases for higher-dimensional descriptors. For up to 128-dimensional descriptors, histograms with \(b = 2\) \((\square)\) and \(b = 4\) \((\circ)\) bins yield the best AUC values. The best AUC values of the \textit{afc} descriptor (Fig. 3d) are at a constant performance level of 0.61 for 32 and higher-dimensional descriptors. For 16- and 32-dimensional descriptors, the best AUC values were obtained with \(b = 4\) \((\circ)\) and \(b = 8\) \((\times)\) Fourier coefficients, respectively. In both cases, \(r = 4\) subpanoramas were used. Compared to the \textit{hist} descriptor, the variance of AUC values obtained for the \textit{afc} descriptor is considerably smaller.

From these results, we conclude that reliable loop-closure detection can be achieved with \textit{hist} and \textit{afc} descriptors and an overall dimensionality ranging from 16 to 128. For both descriptors, the best results were obtained with a subdescriptor dimensionality of \(b = 4\) bins or Fourier coefficients. However, the choice of the parameter combinations of \(r\) and \(b\) has a larger influence on the resulting performance for the \textit{hist} descriptors. For this purpose, we will in future work focus on the Fourier-based \textit{afc} descriptors where the influence of \(r\) and \(b\) is much smaller.

VI. CONCLUSIONS AND OUTLOOK

In this paper, we approached the problem of vision-based loop-closure detection by comparing low-dimensional image descriptors derived from entire panoramic images. The experiments revealed that such global image-descriptors based on gray-value histograms \textit{(hist)} and on absolute Fourier coefficients \textit{(afc)} achieve a very good classification performance. In both cases, the best results were obtained for descriptors with an overall dimensionality between 16 and 128. Although the performances of both descriptors are comparable, we favor the \textit{afc} descriptor because the choice of the values for the parameters \(r\) (number of subpanoramas)
Fig. 3. Detailed analysis of the AUC values depending on the descriptor’s dimensionality. The x-axis represents the descriptor’s overall dimensionality (i.e. \( r \cdot b \)) in a logarithmic scale with base 2; the y-axis carries the AUC values. In all cases, the maximum norm (\( d_{\text{max}} \)) was used as distance function. The range between the maximum (best) and minimum (worst) AUC values is depicted by the shaded area. The number of bins \( b \) is coded by the marker: (a) \( b = 1 \), (C) \( b = 4 \), (b) \( b = 8 \), (d) \( b = 16 \), and (e) \( b = 32 \). Due to numerical effects, some AUC values are slightly smaller than 0.5.

and \( b \) (number of Fourier coefficients) seems to have less influence on the classification performance. Additionally, the resulting performance is independent of the distance functions used to compare the global-image descriptors. We therefore suggest to use standard norm functions since they are computationally cheaper than the other tested distance functions. In comparison to approaches using a set of several local feature descriptors to characterize places (Sec. II), the dimensionality of our global image-descriptors is comparable to the dimensionality of a single local feature descriptor.

The next step in our work is to conduct real robot experiments which then require to choose a decision threshold \( \ell_t \). Furthermore, we will integrate the proposed methods into our framework for vision-based control of autonomous cleaning robots.

REFERENCES


The final version of this article is copyrighted by the IEEE and can be downloaded at:
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6088548

Parsimonious Loop-Closure Detection based on Global Image-Descriptors of Panoramic Images
L. Gerstmayr-Hillen, O. Schützer, M. Krzykowski, and R. Möller
June 20–23, 2011, Tallinn (Estonia), DOI: 10.1109/ICAR.2011.6088548