Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera

Ming Liu, *Student Member, IEEE*, and Roland Siegwart, *Fellow, IEEE*

***Abstract*—Scene recognition problems for mobile robots have been extensively studied. This is important for tasks such as visual topological mapping. Usually, sophisticated key-point-based descriptors are used, which can be computationally expensive. In this paper, we describe a lightweight novel scene recognition method using an adaptive descriptor, which is based on color features and geometric information that are extracted from an uncalibrated omnidirectional camera. The proposed method enables a mobile robot to perform online registration of new scenes onto a topological representation automatically and solve the localization problem to topological regions simultaneously, all in real time. We adopt a Dirichlet process mixture model (DPMM) to describe the online inference process. It is based on an approximation of conditional probabilities of the new measurements given incrementally estimated reference models. It enables online inference speeds of up to 50 Hz for a normal CPU. We compare it with state-of-the-art key-point descriptors and show the advantage of the proposed algorithm in terms of performance and computational efficiency. A real-world experiment is carried out with a mobile robot equipped with an omnidirectional camera. Finally, we show the results on extended datasets.**

***Index Terms*—Graphic model, non-parametric learning, omnidirectional camera, scene recognition, topological segmentation.**

I. Introduction

A. Motivation

In this paper, we propose a lightweight descriptor for omnidirectional vision. It enables a mobile robot to incrementally build a topological map that is based on image appearances and localize itself at scenes simultaneously. Generally, a model of the surrounding environment is needed for robotic missions. Metric and topological maps are the two fundamental types of environment representations. A metric map describes the surrounding environment in a precise and measurable way, usually by defining free and occupied space with occupancy grids [1] or a set of positions/poses of features [2], [3]. Usually, raw-range sensor measurements are applied to construct such a precise model. Although metric maps are able to incorporate redundant information for precise mapping, they are typically not capable of handling the data in an efficient way.

In order to efficiently represent the environment, topological mapping is widely applied for several vision-based applications [4]–[6], since it contains sufficient information and excludes overly detailed metrics, which may be computationally expensive.

Topological mapping and scene recognition techniques are efficient ways to model an environment with sparse information. It facilitates humans' cognition and recognition of their surroundings as well. When people describe where they are, they normally use unique labels of the places such as "my office," "the first part of the corridor," etc.

According to the psychological research [7], region-based topological structures are mostly used by humans when such information is learned or recognized. It relies highly on their ability to learn egocentric positions that are based on visual hints. In most cases, the information that humans use is simple. Several studies have shown that color information can affect the perception of humans in terms of spatial dimensions [8] and scene cognition [9], [10]. These intuitive observations can be extended to similar tasks for mobile robots.

The ability to visually detect scene changes and recognize existing places is essential to mobile robots. Moreover, since robots may have multiple tasks at the same time, it is preferable if these detection and recognition methods are online, which implies the need for minimum computational and memory cost in real time.

However, for most existing techniques that deal with scene recognition, major computational time goes into the feature extraction due to the complexity of the feature detector and describer, e.g., SIFT [11], SURF [12]. Instead of computing complex robust but computationally expensive descriptors, we would like to focus on matching simple and lightweight descriptors, by forming the extraction and matching process as a statistical modeling procedure.

B. Contributions

In our previous work [13], we proposed a lightweight framework for the scene recognition problem using an omnidirectional camera as the only sensor. It focused on the nonparametric modeling of color-based features that are extracted from the panoramic images, without detailed tests for performance and

Manuscript received July 9, 2012; revised May 14, 2013; accepted June 26, 2013. Date of publication August 2, 2013; date of current version April 1, 2014. This paper was recommended for publication by Associate Editor P. Jensfelt and Editor D. Fox upon evaluation of the reviewers' comments. This work was supported in part by the EU FP7 project NIFTi under Contract 247870 and in part by the EU project Robots@home under Grant IST-6-045350.

M. Liu is with the Autonomous Systems Lab, ETH Zurich, Zurich 8092, Switzerland and also with the The Hong Kong University of Science and Technology, Hong Kong (e-mail: ming.liu@mavt.ethz.ch).

R. Siegwart is with the Autonomous Systems Lab, ETH Zurich, Zurich 8092, Switzerland (e-mail: rsiegwart@ethz.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2013.2272250

analysis. Based on these existing results, the following contributions and evaluations are discussed in this paper.

1) Detailed explanation and discussion of the differences between two matching algorithms: The heuristic naive matching algorithm that is based on distance in the feature space and the statistical method, namely, Dirichlet process modeled fast adaptive color tags (*DP-FACT*). By investigating illustrated cross-comparison matrices, we show the decoupled feature space that is used by *DP-FACT* leads to a more flexible algorithm and reliable results.
2) Further evaluations are carried out, e.g., the linearity of the inference time against the number of topological nodes, time cost for different descriptors, and different types of CPUs, by which we show the potential to apply the proposed framework on platforms with unfavorable hardware. We also discuss the reason why the proposed statistical model performs better than the naive matching algorithm.
3) Extended experiments on a widely cited dataset are introduced. The results show the generalization capability and lightweight computational complexity of the proposed *DP-FACT* algorithm.

C. Outline

The remainder of this paper is organized as follows. We first introduce the state of the art by referring to several aspects of related works. After introducing the formation of *FACT* descriptors in Section III, we recall some basic concepts related to the DPMM and introduce how we use it for the topological mapping problem in Section IV. In Section V, we describe an approximation method for online reference of the DPMM, followed by results of real-time experiments given in Section VI. The conclusions and future steps of this study are discussed at the end of the paper.

II. Previous Work

A. Perception and Descriptors

Most of the existing place recognition systems assume a finite set of place labels. The target problem is to determine the labels for each image frame. These classifier-based approaches [14] are limited to applications in predefined or known environments. One of the mainstream techniques for visual scene recognition is based on object detections [15]–[18]. A representative scenario of these methods is to first detect known objects in the scene and then maximize the posterior of the place label given these recognized objects. Methods that are based on similar concepts [19]–[21] use key-point-based features for complete scenes, by which feature retrieval techniques, such as visual vocabulary, are usually used for large-scale applications [22]. These methods are very robust when the objects are correctly detected. Nevertheless, the state-of-the-art object detection methods [11], [12] are usually computationally expensive. They are likely to be unfeasible on computers with limited resources, even with optimizations [23], [24], let alone if the robot had simultaneous tasks besides place recognition.

Several lightweight key-point descriptors were developed as well [25], [26] and widely applied in scene recognition problems [27]–[29]. Unfortunately, most existing applications either deal with categorization of a finite number of known places or are limited to offline inferences.

Besides the key-point-based approaches, descriptors using the transformation/inference of whole images [30]–[35] are also popular. Amongst the most similar to our previous contributions [13], [36] is the "fingerprint of a place" [37], [38]. Both fingerprint and *FACT* use segments from unwrapped panoramic images. The difference is that both [37] and [38] used a laser range finder to help the matching of the descriptors, and *FACT* only used color information from the segments. This means that the proposed algorithm deals with noisier data and alleviates the extra hardware constraints.

B. Sensors

As far as sensors are concerned, omnidirectional vision has been shown to be one of the most suitable sensors for scene recognition and visual topological mapping tasks because of its 360° field of view [39], [40]. Another reason for choosing omnidirectional vision is that when the camera is mounted perpendicularly to the plane of motion, the vertical lines of the scene are mapped onto radial lines on the images. This means that the vertical lines are well preserved after the transformation [36]. Several other approaches utilized this feature as well, e.g., [37] and [41]. The proposed algorithm does not require a full calibration of the omnidirectional camera. It only conducts a detection function of the center coordinate of the panoramic image, which supports the unwrapping operation.

C. Hierarchical Bayesian Modeling

Regarding inference approaches, hierarchical Bayesian methods have been widely used. For example, Vasudevan *et al.* first built a hierarchical probabilistic representation of the environment and then used Naive Bayesian to carry out model inference [17]. Furthermore, hierarchical probabilistic methods based on statistical techniques are very successful in text mining and biological information processing [42], [43].

Theoretical advances in hierarchical probabilistic models, such as LDA [43] and HDP [42], provide good support for the proposed algorithm. The DPMM enables countable infinite clusters for the measurements, which can be used to represent the process of state detection and recognition. Fei-Fei and Perona [44] proposed a key-point-based approach using this framework to cluster natural scenes. Nevertheless, this study deals with indoor environments by using lighter descriptors. In general, these related works only consider one type of observation, e.g., object names or text semantics. In this study, we adopt the classical mixture model and then fit it into multiple types of observations. At the same time, we allow infinite increment of the number of labels. Furthermore, the model is to be learned, updated, and inferred in real time online.

D. Clustering

In order to automate the classification and recognition process, an unsupervised learning algorithm is required. Sophisticated clustering algorithms usually rely, at least in part, on

iterative calculations such as K-means, spectral clustering [45], or affinity propagation [46]. An example of an online reasoning approach is the Chow–Liu tree-based segmentation for static data [47] and change-point detection for sequential data [48], [49]. Recent research showed that semisupervised clustering is also an efficient way in topological structural analysis [50]. For extreme cases, the synchronization of multisensor data needs to be managed [51], [52] or spatial and temporal hints must be jointly considered [53]. In this paper, a naive online change-point detection algorithm is implemented.

E. Recognition and Inference

Recognition is at the core of most robotic applications. For example, robot topological mapping requires detection and recognition of loop closure; semantic mapping usually requires recognition of objects, and human–machine interfaces require recognition of human behaviors. Researchers targeting these core problems attempt to seek the best algorithms to represent this perception process efficiently.

Concerning inference approaches, hierarchical probabilistic methods based on statistical techniques have achieved great success in text classification and biological information processing [42], [43]. In this study, we adapt the classical mixture model to allow multiple types of observations. At the same time, we allow infinite increment of the number of labels. Furthermore, the model is derived and inferred online.

In most of the related works, change-point detection is the basis for segmenting a data sequence [49], [54], [55]. In this study, since we are targeting a lightweight method, the change-point detection is not feasible when using multiple hypothesis methods, e.g., particle filtering [49]. Instead, we use a nonparametric statistical test to evaluate the labeling for each frame separately. This may cause instability in the output label, but it alleviates the requirement to store all previous measurement data in the sequence.

F. Topological Mapping

Several works deal with the extraction of topological regions from metric maps. Intuitively, a topological map is defined as a graph structure, which is composed of *nodes* and links among them. In general, there are two types of topological maps, depending on what a *node* represents. Most existing works consider *nodes* in a topological map as waypoints [56]–[61] in the metric map. The rest consider *nodes* as interesting regions [62]–[64], namely region-based topological maps. Conceptually, both types describe environments by topology representations, typically by using graph structures. However, they have different purposes in terms of robotic tasks. The first type is generally built for facilitating navigation [60], [65] using local navigation [66], [67] methods between nodes, or using topological pose graph [68]. The second type is to enable robots to share a common understanding of the environment with humans, such as in [64], [69], for service purposes. The proposed method considers topological mapping in twofold way. First, it aims to find a way similar to how humans would model the world. Second, at the same time, it helps several aspects of robot navigation, e.g., loop closing or node formation.

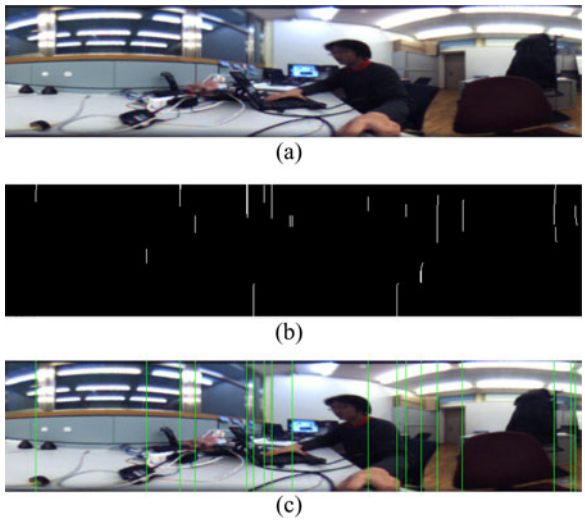


Fig. 1. Extraction of vertical segments. (a) Unwrapped panoramic image. (b) Output of vertical edges detection. (c) Segmentation result.

G. Further References for Color Features

As for color features, besides the fingerprint of place [37], a detailed report on the state of art can be found in [70]. Generally speaking, a color feature is a weak descriptor, as it can be easily affected by lighting conditions. It is the main reason for using a statistical method to minimize the uncertainty.

III. Descriptor

In this section, we first introduce how the *FACT* descriptor is extracted from omnidirectional camera images. Then, we show how *FACT* can be represented, based on absolute measurements and statistical ways, e.g., histograms. In addition, we present a naive approach to match two *FACT* descriptors and depict a distance matrix over frames of an indoor dataset.

A. Segmentation of the Panorama

The proposed descriptor is based on color features and segmentation of the panoramic image. Since the vertical lines from the environment are preserved during a typical motion of mobile robots in a 2-D plane, the dominant vertical lines are used to segment the panoramic image. The examples of extraction results are shown in Fig. 1.

After unwrapping the raw panoramic image [see Fig. 1(a)], we apply, in sequence, Sobel filtering (only along the x direction), Otsu thresholding [71], and morphological operators to extract the most dominant vertical lines. Fig. 1(b) shows the result of the vertical extraction process. Note that only half of the unwrapped image is shown here because of the width limitation.

The dominant vertical lines are chosen based on their length. The lines with length above the average are retained. Morphological operators are used to fuse the lines, which are too close

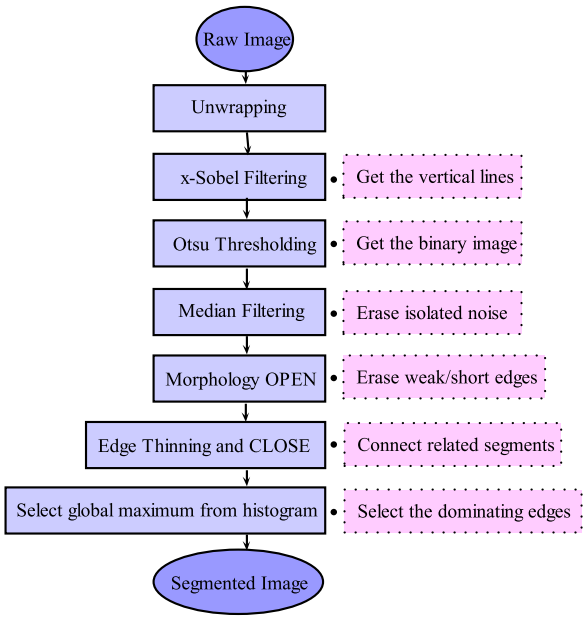


Fig. 2. Segmentation process.

to each other, as a single line. The detailed processing phases are shown in Fig. 2.

As observed in Fig. 1(c), the vertical lines partition the panoramic image into multiple regions. In the next section, we will explain how to extract color-based lightweight descriptors, which we name *Tag*s, from these regions.

B. FACT Descriptor

In this section, we describe the components that construct a *FACT* descriptor. The color representation in RGB color space is not suitable, since it is sensitive to illumination changes, which may be caused by translation and rotation of the omnidirectional camera, as well as different times of the day. Alternatively, we use the YUV color space. The Euclidean distance between two color sites is shown as

$$\begin{cases} U_i = 0.7 * R_i - 0.6 * G_i - 0.1 * B_i \\ V_i = 0.9 * B_i - 0.3 * R_i - 0.6 * G_i \\ \text{Dis}_{1,2} = \sqrt{(U_1 - U_2)^2 + (V_1 - V_2)^2}. \end{cases} \quad (1)$$

Different environment lights may cause white-balance changes. In order to automatically adjust the white-balance, we implement a PD controller to adjust camera parameters. The camera parameters are tuned according to sampled UV values from a reflective white paper, which is stuck in the field of view.

C. Construction of the FACT Descriptor

We extract the descriptor based on the segmented unwrapped image explained in the previous section. In this study, we chose the YUV color space, where the Y signal represents the overall brightness of the pixel and U–V are the two chromatic components. The benefit of using this color space is that we only need two elements (i.e., U and V) to represent a color, regardless of its brightness.

For each region between two vertical lines, the average color value in the U–V space is extracted. Compared with other keypoint-based or edge-based descriptors, an obvious advantage of our approach is that the similarity between features in the U–V space can be simply measured in terms of a 2-D Euclidean distance. The descriptor is formed by the U–V color information and the width W (in pixels) of the region, which is delimited between two vertical edges. Instead of taking each pixel in every region into account, we directly use the average of U–V value that was calculated for each region. U_i and V_i indicate the color information of region i.

One primitive idea is that even if the width of each region may change during the translation of the camera, the projected area in the real world can be well determined in a local neighborhood, as long as the segmentation stays consistent. In this case, the average value of a certain region in the color space remains constant. On the other hand, we must avoid false positive matches that are caused by color similarity of regions. For example, the difference between a green cup and a green cabinet may be very small in color space, but the geometric features of these two are distinguishable. Therefore, we employ the width of the corresponding region W_i as the third dimension of our descriptor. By testing the ratio of the corresponding regions' width, the descriptor can get more reliable results. Let N be the number of regions segmented from the unwrapped image[1]. The dimension of the *FACT* descriptor of a scene is $3 \times N$. A sample descriptor D is shown in (2). Each column in the descriptor is named a *Tag*

$$D = \begin{pmatrix} U_1 & U_2 & & U_N \\ V_1 & V_2 & \dots & V_N \\ W_1 & W_2 & & W_N \end{pmatrix}. \quad (2)$$

D. Statistical Representation of FACT

DP-FACT [13] grants the *FACT* descriptor statistical meanings. *DP-FACT* uses two multinomial distributions, i.e., *DP-FACT*$_t := \{w_t, g_t\}$, to show the statistical distributions of *Tag*'s over discrete feature spaces. Considering a serialized discretization of the U–V color space into m bins, and the number of *Tag*s for a given image at timestamp t is N, we derive the distribution of the color component g_t as follows:

$$g_t \sim p(\bar{r}; N, m, \bar{h}) = \frac{N!}{r_1! r_2! \cdots r_m!} h_1^{r_1} h_2^{r_2} \cdots h_m^{r_m} \quad (3)$$

where the variable $\bar{r}$ is a vector with m integers for which $0 \le r_i \le N$ and $\sum_i r_i = N$. $N > 0$ and $m > 2$ are integers, and $\bar{h}$ is a vector with elements $0 \le h_i \le 1, \sum_i h_i = 1$.

Similarly, the width component w_t can also be defined for each image t, such that

$$w_t \sim p(\bar{r'}; N, n, \bar{h'}) = \frac{N!}{r'_1! r'_2! \cdots r'_n!} h'^{r'_1}_1 h'^{r'_2}_2 \cdots h'^{r'_n}_n \quad (4)$$

[1]According to our experiment, N is usually smaller than 100 and greater than 20 for typical indoor environments.

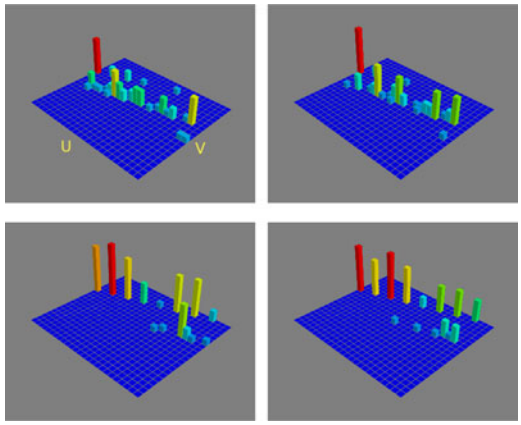


Fig. 3. Example of histograms over discretized UV space extracted from four images. The color bars indicate the number of hits onto the discretized cell of UV space. The dimension indices are marked on the first image.

where the vectors $\bar{r}'$ and $\bar{h}'$ are similarly defined as $\bar{r}$ and $\bar{h}$, respectively. The parameter n is the cardinality of the discretization for the width of *Tag*s.

In practice, normalized histograms of the number of *Tag*s over the discretized feature space can be used and considered as samples from the feature distributions. Four examples of g_t's are illustrated in Fig. 3, where the color of the bars indicates the height, i.e., the relative number of collected features for each bin.

Intuitively, the distributions in the same row are similar, namely that the difference between rows are greater than that within the same row. The quantitative representation of the differences is analyzed in Section V. Note that *DP-FACT* facilitates modeling and inference of the sensor measurements, since multinomial distributions are adopted to represent statistical characteristics of each image frame directly.

Before we jump into the DPMM-based topological mapping, let us first have a look at a naive matching algorithm. We will then conclude this section by examining the drawbacks of this naive matching algorithm.

E. Naive Matching Algorithm

1) Descriptor Matching Based on Euclidean Distance: Because color descriptors are very weak, the fundamental part of our method is the matching. In our previous paper [36], we demonstrated a naive matching algorithm using a three-step strategy as follows:

1) *Test 1*: Tag matching in the U–V Color Space;
2) *Test*2: Tag matching in geometric space;
3) *Test 3*: Descriptor matching.

The matching process is summarized in Algorithm 1. It includes a *Tag*-level comparison, as depicted in Fig. 4. From Algorithm 1, it can be seen that the Euclidean distance is used to measure the difference between two color descriptors. We could imagine that observation noise, such as misdetection of vertical separation lines, can easily damage the result, because it relies on accurate absolute measurements.

Algorithm 1: Nodes Matching and Node Identification

Input:
Input panorama image: $Im(n)$
Existing Node and *FACT* tags: $Featurebase(m)$
Threshold of positive matching result: TH_{global}
Matching local threshold: TH_{local}

Output:
Matching Distance: $Distance(n)$
Feature for current frame: $Feature(n)$
List of nodes: $Nodelist$
List of *FACT*s: $FACTlist$

1 Extract *FACT* Tags $\rightarrow Feature(n)$;
2 **for** ***each*** *Node m in* $Nodelist$ **do**
3 **for** ***each*** *Tag k in* $Feature(n)$ **do**
4 **for** ***each*** *Tag j in* $Featurebase(m)$ **do**
5 **if** *n matches the feature in j* **then**
6 $Distance(k)(j) = \sqrt{\triangle U^2 + \triangle V^2}$;
7 $Distance(k) = \mathbf{Min}\big(Distance(k)(j)\big)$;
8 $s = argmin_{s \in j} Distance$;
9 **if** $Distance(k) < TH_{local}$ ***and*** *the width fit the geometric constraint* ***and*** *s is not in the current matched list* **then**
10 Add s to the matched list;
11 $Bel(n|Nodelist) = \frac{length\ of\ the\ matched\ list}{length\ of\ Featurebase(m)}$;
12 **if** $Bel(n|Nodelist) < TH_{global}$ **then**
13 Update $Nodelist$ and $FACTlist$;
14 m++;
15 Goto line 1 until stop condition;

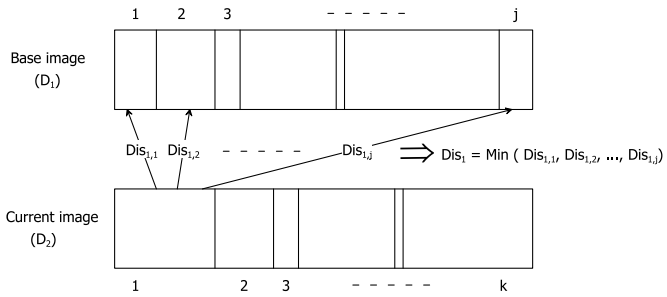


Fig. 4. Schematic diagram of Test 1 during the matching process.

2) Pairwise Distance of Color Features: In order to test the saliency of the proposed lightweight color feature, i.e., the last two dimensions in YUV color space, we cross compare the belief that frame n belongs to the topological node defined by frame m.

The test is carried out on a dataset taken in a typical indoor office environment. The dataset is also used for further evaluations in Sections V-C and VI. It is comprised of a sequence of 920 images, captured on a differential driven mobile robot, with an average frame rate of around 7 fps.

The pairwise beliefs in the sequence, indicated by line 11 in Algorithm 1, construct a distance matrix as shown in Fig. 5. It intuitively shows the distinctions among different frames using

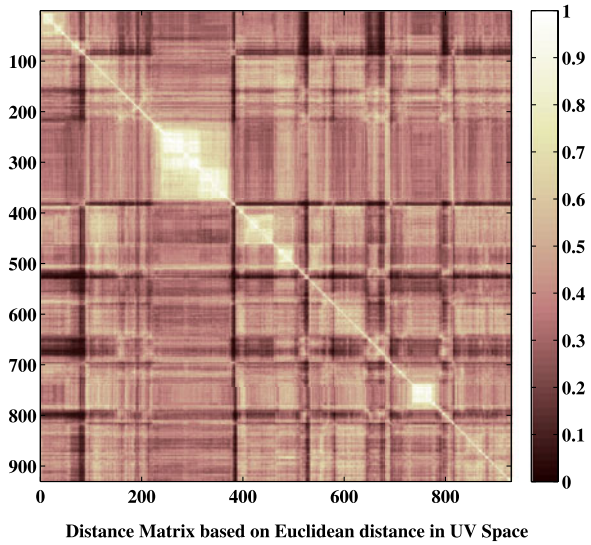


Fig. 5. Distance matrix for the belief pairs.

color-based appearances. We can see that the adjacent frames from the image sequences show higher beliefs to be clustered together, which are illustrated with lighter color blobs. At the same time, there are certain possibilities for those nonadjacent frames to be classified as the same scene as well, in the case that the robot may have returned to a previously visited place.

3) Drawbacks of the Naive Matching Algorithm: The naive matching algorithm in Algorithm 1 has been studied and compared with different perspectives [66], [72]–[74]. According to our further study, the major disadvantages of the naive matching approach are as follows.

1) The matching step was a point estimator, which did not consider probability and multihypotheses.
2) The false positive ratio of scene changing detection was high; therefore, it required an offline refinement.
3) The cardinality of the control parameters was big. Five parameters needed to be adjusted.

In order to overcome these shortcomings, we need to refactorize it as a probability-based framework. We present the alternative DPMM, which we propose for topological mapping in the next section.

IV. Model of Topological Mapping

Topological mapping and scene recognition are two sides of the same coin. They both reflect the process of detecting changes and relocalizing in an existing topological environment model. The optimized DPMM for topological mapping that uses the proposed lightweight color descriptor is shown in Fig. 6. The parameters are depicted in rectangles and random variables are in circles. As a convention, we use a plate representation for repeated components. The model depicts two conditional independent processes given measurements g_t and w_t. The two processes are explained as follows.

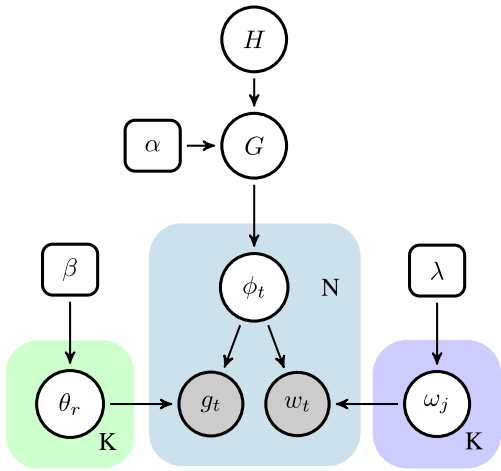


Fig. 6. System model.

A. Chinese Restaurant Process (CRP)

The Dirichlet process G is formulated with a base distribution H and a concentration parameter α. The base distribution is the mean of G and the concentration parameter α acts as an inverse variance. The distribution G is comprised of point masses, and samples from G will be repetitively drawn, considering the case of an infinite sequence. Additionally, ϕ_t is an indicator of the cluster identity to which the current image belongs.

Therefore, ϕ_t is the target variable of inference. If the process is considered a partition problem, a CRP model is commonly used. A CRP model can use priors that are obtained from a stick-breaking process [75]. By integrating over G, the next sample for cluster identity is described by

$$\phi_t|\phi_{1:t-1} \sim \frac{\sum_{n=1}^{t-1} \delta_{\phi_n} + \alpha H}{t-1+\alpha}$$

where δ_{ϕ_n} is an indicator of the mass-point function located at the nth frame, which is labeled as ϕ_n. It implies that the more we see a certain cluster of data, the higher a prior that data from such cluster may be observed again. The problem is then converted to an estimator of the posterior

$$P(\phi_t|\phi_{\setminus t}, G, \boldsymbol{g}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\omega}; \alpha, \beta, \lambda)$$

where $\phi_{\setminus t}$ is the full set of indicators excluding the current one up to time t. Variables $\boldsymbol{g}$ and $\boldsymbol{w}$ are the observed feature measurements, following the definitions of (3) and (4). $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the sensor models for color and geometric information, with priors β and λ. Conventionally, random variables and parameter sets are shown in bold text.

B. Sensor Data Perception

The observable random variables from the model are two multinomial distributions g_t and w_t, which are associated with two histograms by accumulating the number of features that hit their own discretized space. Taking w_t for instance, it is a multinomial distribution that represents a single histogram of different width of *Tag*s in one *FACT* feature. The dimensions of samples g_t and w_t are D_{uv} and D_w, respectively, indicating

the dimensions of the discrete UV space and width space. The number of samples is represented by N, which is equal to the number of sequential frames during the experiment.

By only considering w_t, for example, as it is a multinomial distribution, w_t is subject to a Dirichlet distribution prior ω_j. Assuming there are K different scenes, ω will be a matrix of dimensions $K \times Z$. w_t's of dimension Z are drawn from ω. Z is the number of possible histograms given the maximum number of *Tag*s in a frame, which is a large number. Since we use an approximation method for the inference in Section V, the precise expression of Z is not necessary. Note that because θ and ω are discrete, $P(\theta_{t1} = \theta_{t2}) \neq 0$,$P(\omega_{t1} = \omega_{t2}) \neq 0$, for different time stamps $t1$ and $t2$.

In summary, on one hand, observations are inherently determined by its label ϕ_t, as defined previously; on the other hand, we can also consider the observations g_t and w_t as samples from a sensor model θ_r and ω_j for cluster k, respectively. The sensor model priors are given by β and λ. So far, we have built a model of two subprocesses, namely a perception process and a sensoring process. They serve as the basis for building a data-driven inference model for the recognition problem.

C. Model Inference

As a summary of the proposed model

$$\begin{aligned} G &\sim \text{Dir}(\alpha H) \\ \phi_t|G &\sim G \\ g_t &\sim F(\phi_t, \theta_{\phi_t}) \\ w_t &\sim Q(\phi_t, \omega_{\phi_t}) \end{aligned}$$

where F and Q represent the generation processes of the measurements from the base models, according to the label ϕ_t.

The joint probability can be written directly as

$$p(\phi\, G\, \theta\, \omega\, \boldsymbol{g}\, \boldsymbol{w}; \beta, \lambda) = \prod_{r=1}^{K} p(\theta_r; \beta) \prod_{j=1}^{K} p(\omega_j; \lambda) \times \prod_{t=1}^{N} p(G; H, \alpha) p(\phi_t|G) p(g_t|\theta_{\phi_t}) p(w_t|\omega_{\phi_t}).$$

In order to factorize it into independent components, we integrate the joint probability over ω, θ, and G

$$\begin{aligned} p(\phi\, \boldsymbol{g}\, \boldsymbol{w}; \beta, \lambda) &= \int_\omega \int_\theta \int_G p(\phi\, G\, \theta\, \omega\, g\, w; \beta, \lambda) dG\, d\theta\, d\omega \\ &= \int_\omega \prod_{j=1}^{K} p(\omega_j; \lambda) \prod_{t=1}^{N} p(w_t|\omega_{\phi_t}) d\omega \\ &\quad \times \int_\theta \prod_{r=1}^{K} p(\theta_j; \beta) \prod_{t=1}^{N} p(g_t|\theta_{\phi_t}) d\theta \\ &\quad \times \int_G \int_H \prod_{t=1}^{N} p(\phi_t|G) p(G; H\alpha) dH dG. \end{aligned} \tag{5}$$

The last component is an expectation on G, i.e., $E_G\left[p(\phi_1\, \phi_2\, \phi_3\, \phi_4 \cdots \phi_N|G)\right]$. According to the characteristics of the Dirichlet process, it is proportional to the product $\prod_{t=1}^{N} p(\phi_t|\phi_{\setminus t}) \propto p(\phi_t|\phi_{\setminus t})$. Therefore

$$\int_G \int_H \prod_{t=1}^{N} p(\phi_t|G) p(G; H\alpha) dH dG \propto \frac{\sum_{t=1}^{N-1} \delta_{\phi_t} + \alpha \delta_{\phi_{\bar{k}}}}{N - 1 + \alpha} \tag{6}$$

where δ_{ϕ_n} is a mass point function located at ϕ_n. $\bar{k}$ is the indicator of a new cluster.

The first two parts can be treated in a similar manner. Taking the first part for instance, using n_v^j to represent the number of frames, whose width histogram is the vth element in ω_j within cluster j, we obtain

$$\begin{aligned} &\int_\omega \prod_{j=1}^{K} p(\omega_j; \lambda) \prod_{t=1}^{N} p(w_t|\omega_{\phi_t}) d\omega \\ &= \prod_{j=1}^{K} \int_{\omega_j} p(\omega_j; \lambda) \prod_{t=1}^{N} p(w_t|\omega_{\phi_t}) d\omega_j \\ &= \prod_{j=1}^{K} \int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v - 1} \prod_{v=1}^{Z} \omega_{j,v}^{n_v^j} d\omega_j \\ &= \prod_{j=1}^{K} \int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v + n_v^j - 1} d\omega_j \end{aligned} \tag{7}$$

since the integral of the Dirichlet distribution equals unity

$$\int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v + n_v^j)}{\prod_{v=1}^{Z} \Gamma(\lambda_v + n_v^j)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v + n_v^j - 1} d\omega_j = 1. \tag{8}$$

Equation (7) can be continued as

$$\begin{aligned} &\int_\omega \prod_{j=1}^{K} p(\omega_j; \lambda) \prod_{t=1}^{N} p(w_t|\omega_{\phi_t}) d\omega \\ &= \prod_{j=1}^{K} \int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \frac{\prod_{v=1}^{Z} \Gamma(\lambda_v + n_v^j)}{\Gamma(\sum_{v=1}^{Z} \lambda_v + n_v^j)}. \end{aligned} \tag{9}$$

It is similar for the integration over θ. The joint probability is then represented using

$$\begin{aligned} &p(\phi\, \boldsymbol{g}\, \boldsymbol{w}; \beta, \lambda) \\ &\quad \propto \prod_{j=1}^{K} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \frac{\prod_{v=1}^{Z} \Gamma(\lambda_v + n_v^j)}{\Gamma(\sum_{v=1}^{Z} \lambda_v + n_v^j)} \\ &\quad \prod_{j=1}^{K} \frac{\Gamma(\sum_{v=1}^{Y} \beta_v)}{\prod_{v=1}^{Y} \Gamma(\beta_v)} \frac{\prod_{v=1}^{Y} \Gamma(\beta_v + n_v^j)}{\Gamma(\sum_{v=1}^{Y} \beta_v + n_v^j)} \\ &\quad \times \left(\frac{\sum_{t=1}^{N-1} \delta_{\phi_t} + \alpha \delta_{\phi_{\bar{k}}}}{N - 1 + \alpha} \right). \end{aligned} \tag{10}$$

When we consider a collapsed Gibbs sampling process on the cluster indicator ϕ_t at time t, we have

$$p(\phi_t|\phi_{\setminus t}\, \boldsymbol{g}\, \boldsymbol{w}; \beta, \lambda) \propto p(\phi_t\, \phi_{\setminus t}\, \boldsymbol{g}\, \boldsymbol{w}; \beta, \lambda). \tag{11}$$

However, the large size of Z makes the direct inference impossible. In general, sampling methods [76] are used to estimate the posterior, but they are usually computational expensive. Here, we propose a real-time approximated solution. The first two parts are indications of the relation between the reference distribution of ω_k and β_k, and the current measure of frame i. Using $\xi()$ and $\mu()$ to represent these two relations, we could rewrite (11) as

$$p(\phi_t = k|\phi_{\setminus t}\boldsymbol{g}\,\boldsymbol{w}) \propto \frac{\Gamma(\lambda_p + n_p^k)}{\Gamma(\sum_{v=1}^{Z}\lambda_v + n_v^k)}\frac{\Gamma(\beta_q + c_q^k)}{\Gamma(\sum_{u=1}^{Y}\beta_u + c_u^k)}\left(\frac{\sum_{t=1}^{N-1}\delta_k + \alpha\delta_{\phi_{\bar{k}}}}{N-1+\alpha}\right) = \xi(w_t|\omega_{\phi_t})\mu(g_t|\beta_{\phi_t})p(\phi_t|\phi_{\setminus t}). \tag{12}$$

In the next section, we approximate both conditional probabilities $\xi(\cdot|\cdot)$ and $\mu(\cdot|\cdot)$ based on a common nonparametric statistical test: χ^2 test. This leads to the improved approach for matching two *DP-FACT* features.

V. Matching of Dirichlet Process Modeled Fast Adaptive Color Tags

Most existing methods for DPMM use offline inference, mainly because the inference is time consuming. The Monte Carlo Markov chain (MCMC) sampling method [77] is considered as the standard approach [76]. In order to solve the inference problem in real time in an online manner, the inference of the conditional probabilities is to be approximated directly. When it is possible, it relieves the need to calculate the joint probability. Recall that the equation of the posterior of the place labeling, depicted in (12), includes three parts. The last part is a representation of a prior CRP based on the previous observed labels. It can be calculated directly from the history of measurements. The first two parts are similar. Typically, they are estimated by sampling methods. A closer look at them will reveal that they calculate the gamma function of the count of a certain observation over all the possibilities. In other words, they represent the probability of a certain histogram showing up in a sequence of observations. Therefore, it is a measure of the similarity of the current observation to all the predefined models. As a result, no sampling methods are needed to estimate this measure if we can approximate the underlying similarity between the current observation and the reference models. This is the basic idea of our online inference method.

A. Nonparametric Test

Since both observation and existing models are inherently histograms, the similarity between them can be estimated by nonparametric statistical methods. Here, we introduce our approximation of 11 using the χ^2 test.

The χ^2 test is formalized as follows [78]:

$$\chi^2(m,n) = \sum_{t=1}^{r}\frac{(n_t - N\hat{p}_t)^2}{N\hat{p}_t} + \sum_{t=1}^{r}\frac{(m_t - M\hat{p}_t)^2}{M\hat{p}_t} \tag{13}$$

where $\hat{p}_t = \frac{n_t+m_t}{N+M}$, $N = \sum_{t=1}^{r} n_t$, $M = \sum_{t=1}^{r} m_t$, r is the dimension of both histograms, and n_t and m_t are the number of hits at the bin t. The converging condition is $\sum_{t=1}^{r} p_t = 1$ according to the definition. For the bins where both histograms have 0 measure, the calculation is skipped.

According to (10), the observed distribution is determined by both the history of observations and the Dirichlet prior. However, the χ^2 test only provides an estimation of the probability of the current observation referring to the base distribution. It can be further inferred as a statistical count of occurrences while considering the history of observations. In order to compensate the lack of information of the Dirichlet prior, we define a weighting factor ρ to adjust the influence of both measures, i.e., the measure in the color space and geometry space. The estimator of the target label is therefore approximated as

$$p(\phi_t = k|\phi_{\setminus t}, \boldsymbol{g}\,\boldsymbol{w}) \equiv p(\phi_t|\phi_{\setminus t})\cdot\xi(w_t|\omega_{\phi_t})\cdot\mu(g_t|\beta_{\phi_t}) \propto \left(\frac{\sum_{t=1}^{N-1}\delta_{\phi_t} + \alpha\delta_{\phi_{\bar{k}}}}{N-1+\alpha}\right)e^{-\rho\chi^2(w_t,\omega_k)-(1-\rho)\chi^2(g_t,\theta_k)} \tag{14}$$

where $\rho \in [0,1]$. If $\rho = 1$, then the estimator (14) considers only the geometry measure, and vice versa. Using the form of (12), the two targeting conditional probabilities are formalized as follows:

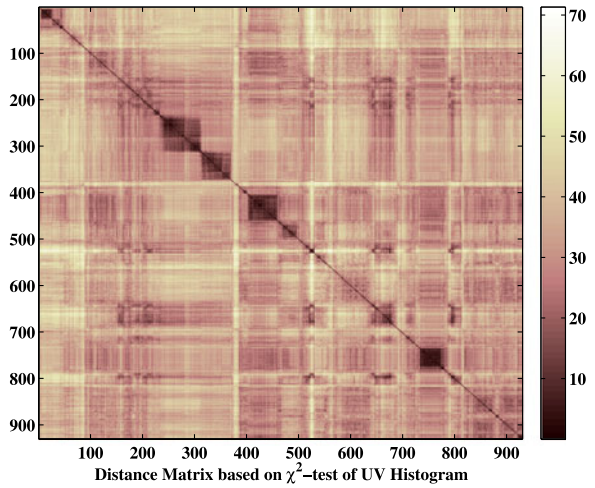
$$\xi(w_t|\omega_{\phi_t}) \propto e^{-\rho\chi^2(w_t,\omega_{\phi_t})}$$
$$\mu(g_t|\beta_{\phi_t}) \propto e^{-(1-\rho)\chi^2(g_t,\theta_{\phi_t})}. \tag{15}$$

B. Model Update

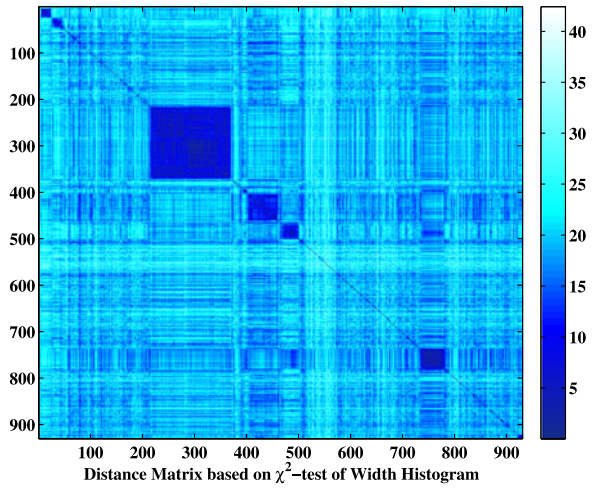
Despite the fast calculation, the nonparametric statistic that was introduced in (14) has an obvious disadvantage. It can be seen that the nonparametric test is a point estimation without considering previous information. In order to remedy this disadvantage, a model update algorithm is developed. Unlike (12), where the previous information is represented by the counts of occurrences n_p^k and c_q^k, we require a method to take the history of data into account. This means that the reference models ω_k and θ_k need to be able to fuse information from all the existing measurements. Instead of saving all the previous observations, we propose an iterative method to fuse the current measurements with existing models as follows:

$$\theta_k^{t+1} = \frac{n_k^t}{n_k^t+1}\theta_k^t + \frac{1}{n_k^t+1}g_t$$
$$\omega_k^{t+1} = \frac{n_k^t}{n_k^t+1}\omega_k^t + \frac{1}{n_k^t+1}w_t \tag{16}$$

where n_k^t is the number of frames that have been clustered as label k by time t. Therefore, the update process in (16) is a weighted mean by combining the existing knowledge and the new observation at each time step. The advantages of this model update algorithm is obvious. On one hand, it can be calculated online with low requirements on computational and spatial costs. On the other hand, it reflects the history of data in the updated model directly.



(a)

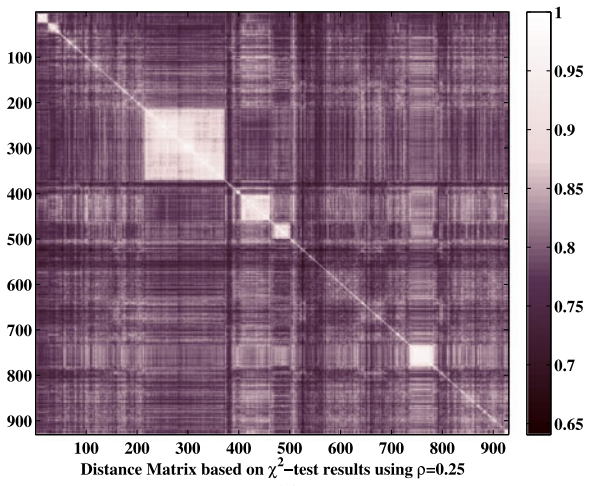


(b)

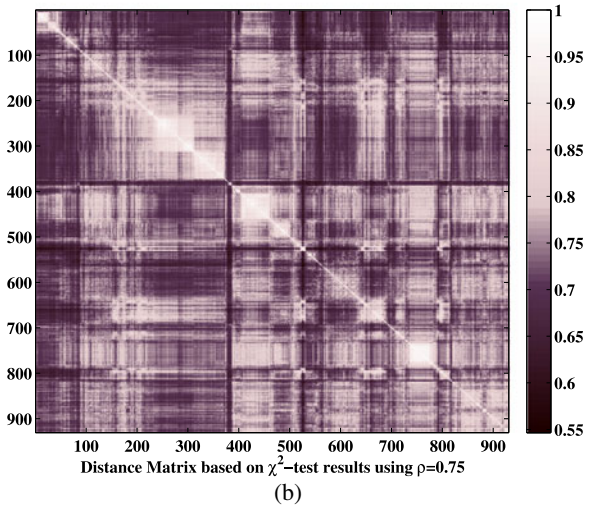
Fig. 7. Pairwise distance matrix in UV-color and width space.

C. Pairwise Distance for DP-FACT

Following the discussion in Section III-E2, we analyze the distance matrices of the χ^2 test results and the compound posterior. In order to keep the consistency, the same dataset as that are mentioned in Section III-E2 is used. The result of the distance matrix in color space is depicted in Fig. 7(a), and that in width space is shown in Fig. 7(b). An intuitive observation is that color features are able to partition the whole sequence into more segments compared with width information. The χ^2 test leads to more distinctive separations than the results of the Euclidean distance. Besides, it is interesting to see that the χ^2 test result for histograms of *Tag*-width can also indicate the similarity between adjacent frames. This means that the fusion of these two parts of information can determine the recognition results more reliably by introducing multiple constraints. As previously discussed, adjustment to the parameter ρ leads to changes in the posterior. Compared with Fig. 7, we show in Fig. 8 that a higher ρ value, which increases the significance of color features, leads to a greater number of potential change points, since the color features are more salient than width features.



(a)



(b)

Fig. 8. Distance matrix using combined features for different ρ values.

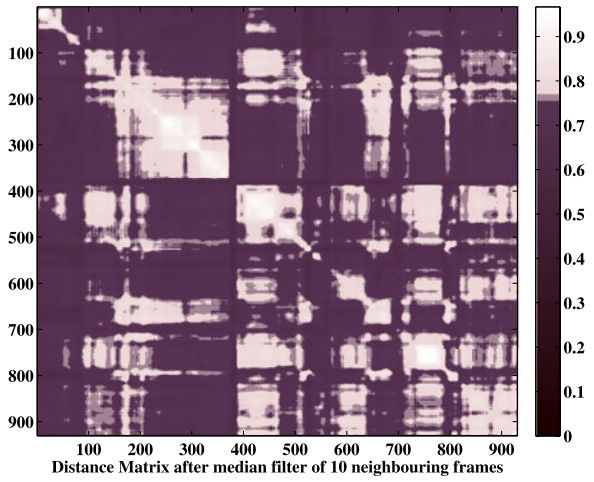


Fig. 9. Resulting distance matrix using a median filtering.

The result obtained by introducing a median filter is shown in Fig. 9. Note that we change the color map of the figure intentionally for better visibility. The off-diagonal light blobs show that the corresponding frames are similar in appearance. However, this pairwise result does not imply the number of topological

nodes, since the node models keep evolving regarding (16), whenever a new positive reading is detected. Here, we simply use these plots to reveal the distinctiveness and feasibility of the proposed *DP-FACT* features. The result of a real-time experiment including all the components of (12) is shown in the next section.

D. Discussion on the χ^2 Test and Naive Matching

In this section, we show the theoretical insight into why the χ^2 test provides more distinct results. A typical result can be seen by comparing Figs. 5 and 7(a). For two feature vectors $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, a summation form of Euclidean distances is represented by

$$d_e = \sqrt{(\boldsymbol{x_1}-\boldsymbol{x_2})^T(\boldsymbol{x_1}-\boldsymbol{x_2})} = \sqrt{(\boldsymbol{x_1}-\boldsymbol{x_2})^T I_{kk}(\boldsymbol{x_1}-\boldsymbol{x_2})} \tag{17}$$

where I_{kk} is an identity matrix with the same dimension as the feature. It can be interpreted that the covariance of the features is not considered.

On the other hand, considering the χ^2 test that is introduced in (13), it has a limiting χ^2 distribution with $N+M-1$ degrees of freedom. Under the null hypothesis, it has the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$ as follows:

$$\boldsymbol{\mu} = (N+M)\times(\hat{p_1},\hat{p_2},\ldots,\hat{p_r})^T \tag{18}$$

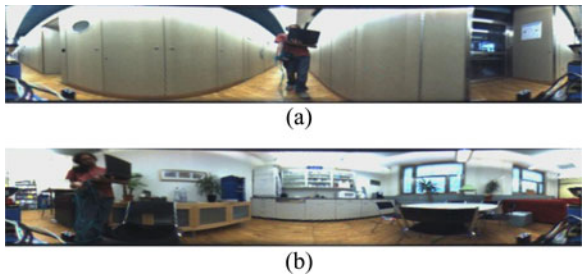
$$\boldsymbol{V} = (N+M)\begin{pmatrix} \hat{p_1}(1-p_1) & -\hat{p_1}\hat{p_2} & \cdots & -\hat{p_1}\hat{p_r} \\ -\hat{p_2}\hat{p_1} & \hat{p_1}(1-\hat{p_1}) & \cdots & -\hat{p_2}\hat{p_r} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{p_r}\hat{p_1} & -\hat{p_k}\hat{p_2} & \cdots & \hat{p_r}(1-\hat{p_r}) \end{pmatrix} \tag{19}$$

using the notation of (13). This additional information enables that all the *Tag*s in a *FACT* descriptor are jointly considered with respect to the frequency of hits on each discrete bin of the feature space.

E. Discussion on Data Modeling

We address the major differences between the proposed *DP-FACT* and the naive matching method introduced by Algorithm 1 as follows.

1) The matching process is no longer a point estimator with the DPMM. With multiple hypotheses, the final output is a maximum-a-posterior (MAP) result instead.
2) New nodes are incrementally generated online, and the false positives are, due to noise, with low priority with regard to the CRP process. It is an important feature for applications on a larger scale.
3) Last but not the least, the designed DPMM takes into consideration that the observations from sensors are conditionally independent when the label is given. This can also be inferred from (14). This shows that even with more than two observations, it is possible to fuse the information related to a certain label (place, scene) efficiently.



(a) (b)

Fig. 10. Sample images: (a) corridor and (b) coffee room.

VI. Experiments

The results of experiments are introduced in this section. Our approach is compared with key-point-based methods in terms of labeling accuracy, performance, and inference complexity. Two samples of the unwrapped images are shown in Fig. 10.

A. Comparison in Accuracy

As described in [29], the SIFT feature demonstrates a superior accuracy in scene transition detection and recognition accuracy compared with CENTRIST and the texture-based method. In this paper, we compare the proposed *DP-FACT* with SIFT, as well as a newly developed lightweight key-point descriptor BRISK [79]. These two descriptors represent the most sophisticated and novel state-of-the-art binary descriptors, respectively. It has been reported [79] that the BRISK feature is around 15 times faster than SURF [12] features considering feature extraction and matching. Meanwhile, the BRISK feature has similar performances to SURF, or even better.

As for the key-point-based methods (SIFT and BRISK), we use the unwrapped images as inputs. The algorithm is designed as follows. First, key-point feature extraction is performed on the input images. Then, the current image is matched with reference images, which have been observed in the past, to get the most similar reference. If the ratio of #[2] positively matched features and # features extracted from reference images is above a given threshold (in this case 70%), then we label the current image with the same label as the best matched reference; otherwise, the current image is given a new label by taking it as a newly added reference image.

The test result is shown in Fig. 11. In order to ease the comparison, the figures are aligned in time series. The first two plots on the top are the raw output of *DP-FACT* and the result after median filtering over the past five frames. Note that further offline smoothing of the labeling can be implemented as well [36], [80], which can potentially provide more precise results. The results of key-point-based methods after the same median filtering are given in the third plot. Specifically, the vertical axis of the top three subfigures in Fig. 11 indicates the online recognized labels.

The results indicated by the top three plots in Fig. 11 show that the proposed *DP-FACT* framework leads to more stable outputs than key-point-based methods. On the contrary, key-point-based methods have a high false positive ratio on the transition

[2]# means "the number of."

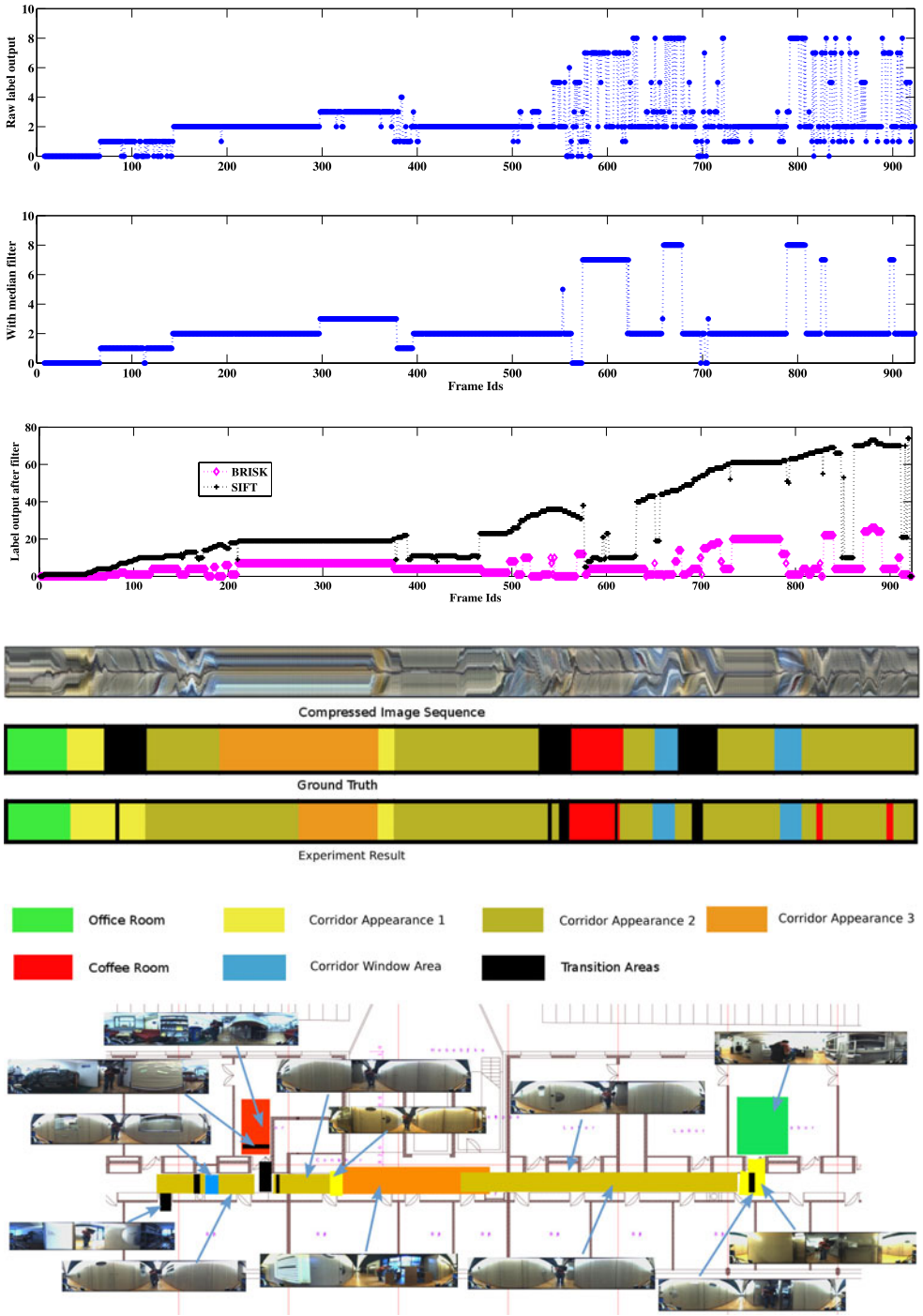


Fig. 11. Experimental results. (From top to bottom) The raw labeling output of *DP-FACT*; the result of *DP-FACT* after the median filter of five frames; the result of key-point-based approaches (SIFT, BRISK) after median filtering; the image sequence in a compressed layout; the manually labeled ground truth; the result of *DP-FACT*; label explanations; and an overlaid sketch of the test environment by detected scene appearances.

detection, because the labeling is dominated by massive changes of key points, even in the same scene. As a result, a high scene change rate was observed and the number of scenes detected from the sequence is much higher.

The "compressed image sequence" stripe shows a squeezed summary of the whole image sequence, from which the scene changes can be intuitively observed. The "experimental result" shows the filtered output of *DP-FACT*.

The "transition areas" indicate that the robot is closely passing a doorway or turning corners, where the scene recognition does not make much sense and is therefore not considered in the statistical results. The corresponding behavior of the algorithm is that the output label can hardly be stable even after median filtering, which can be readily detected and labeled.

The "ground truth" is manually labeled by only observing the input video sequences. This means that the images with

TABLE I
ACCURACY OF SCENE RECOGNITION

Method	Recognition Accuracy
SIFT	73.3%
BRISK	66.7%
DP-FACT	89.4%

a similar appearance are considered to be obtained from the same scene. Compared with "experimental result," we could infer that the change point detection is more practical than the key-point-based approaches. An overlaid 2-D sketch of the target environment against the experimental results of *DP-FACT* is depicted at the bottom of Fig. 11. The image sequence starts from the right side of the map. Sample images from different scenes are illustrated around the sketch, which shows the differences in appearance at various scenes. Although with some misclassifications, *DP-FACT* shows more reliable and feasible results for the scene recognition.

As part of the quantitative comparison, Table I shows the accuracy of scene recognition. Because the transition detection for key-point-based methods is vague, which leads to frequent false positives, the scene recognition results for key-point-based methods are calculated by considering nonrepeated labels in the same scene as a group. Since *FACT* requires an offline filtering, the comparison is not included. We can see that *DP-FACT* has the best recognition accuracy, though color is a relatively "weaker" feature than key-point descriptors.

Two possible reasons why key-point-based methods do not perform as well as *DP-DACT* can be considered. First, the distortion of the uncalibrated omnidirectional images causes nonuniform resolution of the unwrapped images, which makes the key-point-based feature extraction unstable, especially when the key points are at different distances. Second, *DP-FACT* is structured only in the horizontal direction, by which the information is summarized in one dimension. However, key points can only be possibly detected anywhere on the whole 2-D surface of the image. This consistency of feature construction maximizes the difference between any two labels and more importantly minimizes the influence of unexpected randomness.

B. Evaluation of Time Cost

The evaluation of time cost is shown in Fig. 12. Because the number of nodes rises during the test, we see that the overall time rises slightly as well. Compared with the time cost of common sampling methods, the gray area in Fig. 12 indicates that the expected inference time of the proposed estimation is less than 5 ms.

We make a further study of the relation between the inference time and the complexity of the model. Fig. 13 depicts a regression result of the inference time over the number of nodes, which is substantially linear. This result implies that the potential of the proposed method can be extended to large-scale environments without jeopardizing the capability of performing in real time.

Let us recall the test in Fig. 11. In addition to the superior recognition accuracy, *DP-FACT* shows faster performance.

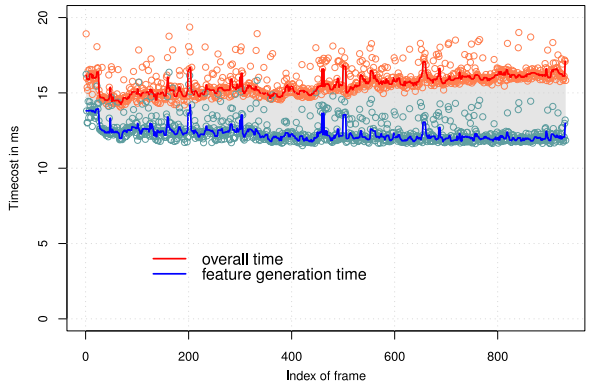


Fig. 12. Time cost of *DP-FACT* over frames. The lines are filtered results out of raw measurements (in circles). The gray area indicates the inference time.

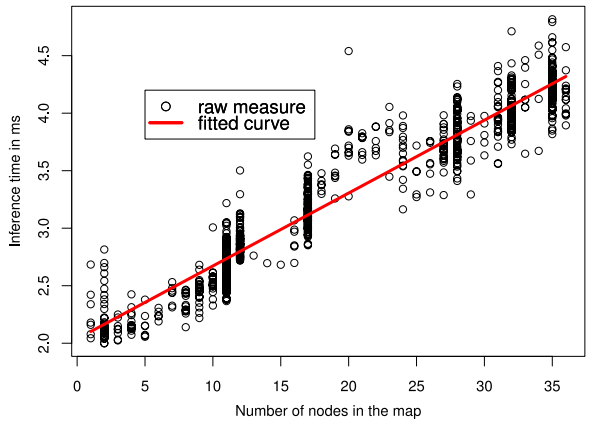


Fig. 13. Inference time versus the number of nodes.

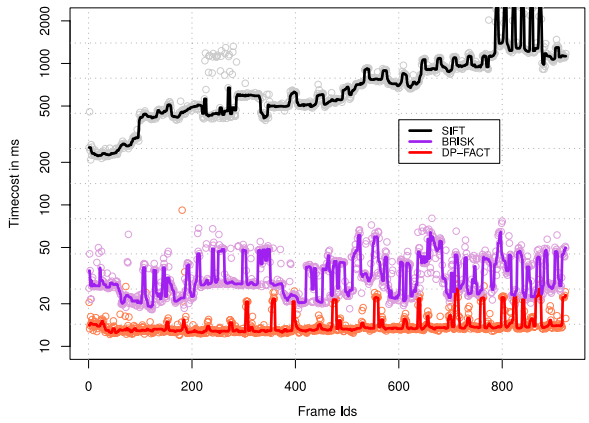


Fig. 14. Time cost comparison.

Fig. 14 illustrates the computational cost of each approach. Further details about implementations are given in Table II.

Our aim is to develop a scene recognition algorithm, which can be implemented online with limited computational resources. We evaluated the algorithm on three different types of CPUs in order to show that our method is suitable for different applications. The result is shown in Fig. 15. We see that even for early CPUs, the algorithm can still reach around 17 Hz (cycle time = 58.6 ms).

TABLE III
TEST RESULT ON COLD DATASETS

Dataset	#Image	#SemNode	#Node	#Tran	#Tag	Total Time (ms)				Inference Time (ms)			
						Mean	Max	Min	StdVar	Mean	Max	Min	StdVar
Freiburg PathA	1459	5	10	17	25.3	15.03	69.72	7.43	2.23	4.59	14.03	3.14	0.78
Ljubljana PathA	1871	6	9	22	19.1	13.34	38.22	7.06	1.23	3.74	12.92	2.82	0.74
Saarbrücken PathA	2941	8	14	19	20.5	14.39	25.90	7.48	1.17	4.49	12.30	3.09	0.73
Saarbrücken PathB	1306	5	9	11	21.6	14.64	32.29	7.49	1.78	4.07	8.24	3.07	0.72

TABLE II
IMPLEMENTATION DETAILS OF THE THREE COMPARED APPROACHES

Method	Language	Comments
SIFT	C/C++, GLSL	GPU Implementation by [23]
BRISK	C/C++, ASM	Without rotation invariant
DP-FACT	C/C++	Based on OpenCV 2.3

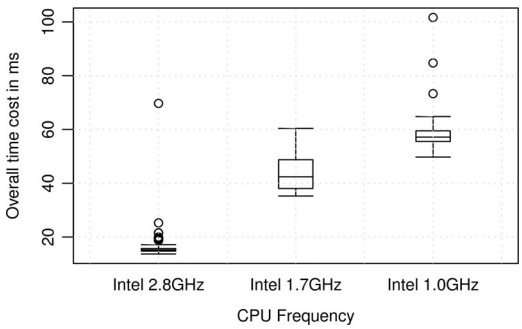


Fig. 15. Performance on different CPUs, using single core.

C. Further Experiments on a Public Dataset

In order to further validate the results, we apply the proposed *DP-FACT* onto a widely cited dataset called COLD [81]. It is a collection of indoor omnidirectional images from Freiburg, Ljublijana and Saarbrücken. Since the white balance for those images could not be properly adjusted in an online manner, we use only the ones captured in cloudy weather in order to minimize the influence on standard color. Statistical results are illustrated in Table III. The definition of each column is explained as follows:

Dataset :	name of the test dataset;
#Image :	#images included in the dataset;
#SemNode :	#provided labeled semantic nodes;
#Node :	#detected nodes using $DP\text{-}FACT$;
#Tran :	detected #scene transitions;
#Tag :	average #detected Tags per image;
Total Time :	statistics of the total time cost per image;
Inference Time :	statistics of the inference time per image.

Table III shows that *DP-FACT* can segment the environment with low computational time. The Freiburg dataset results in higher total time cost. This is because there are in general more *Tag*s detected, and the additional time cost is due to the descriptor construction. Note that the number of detected nodes (i.e., #Nodes) is usually higher than the number of manually labeled semantic nodes, because, for most cases, the manually labeled semantic areas, such as "a corridor," may contain multiple appearances. Taking Fig. 16 as an example, where the standard

Fig. 16. Segmentation result based on *DP-FACT* for Freiburg Path A [81].

trajectory for Path A is used [81], we see that the corridor labeled ② is actually segmented into five regions with respect to door positions, etc. Different colors are used to indicate the segmentation results. The statistical results in the last two columns show that the computational time is stably low. Importantly, the rise of #Image does not significantly increase the inference time, by which the generalization capability for larger datasets is revealed.

VII. CONCLUSION AND FUTURE WORK

In this paper, a lightweight color-based framework for scene recognition and topological modeling of indoor environments using omnidirectional cameras has been presented. We proposed using a DPMM to manage new scene registration and recognition simultaneously. The results of the experiment showed the advantage of the proposed framework in terms of online computation ability and better recognition performance than keypoint-based methods. This study also showed that the inference of a DPMM can be approximated by reasoning the conditional probability directly. We envision that similar concepts can be

adopted to solve other inference problems with large target space as well. It is also possible to use such models for a data modeling problem with multiple observations.

The proposed *FACT* descriptor only deals with indoor environments, where vertical lines are preserved in the field of view of unwrapped panoramic images obtained by omnidirectional cameras. Therefore, the results do not imply that the extended applications for a semistructured environment are easily feasible. Not withstanding this limitation, this study does suggest that color-based features can be integrated with a real-time online scene recognition and topological mapping robotics system, with relatively good performance. It can be imagined that the combination of key-point-based and color-based methods will help to solve this problem at a hybrid level, without limiting the target environment. Regarding the loop-closing problem, the proposed framework can help in the selection of target poses to be matched, with low computational cost. The conducted results will be shown in our future work.

REFERENCES

[1] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, Nara, Japan, Nov. 2007, pp. 225–234.

[3] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.

[4] A. Murillo, C. Sagüés, J. Guerrero, T. Goedemé, T. Tuytelaars, and L. Van Gool, "From omnidirectional images to hierarchical localization," *Robot. Autonom. Syst.*, vol. 55, no. 5, pp. 372–382, 2007.

[5] M. Liu, C. Pradalier, F. Pomerleau, and R. Siegwart, "Scale-only visual homing from an omnidirectional camera," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3944–3949.

[6] M. Liu, C. Pradalier, F. Pomerleau, and R. Siegwart, "The role of homing in visual topological navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2012, pp. 567–572.

[7] T. McNamara, "Mental representations of spatial relations," *Cognit. Psychol.*, vol. 18, no. 1, pp. 87–121, 1986.

[8] G. Franz, "Space, color, and perceived qualities of indoor environments," in *Proc. 19th Int. Assoc. People-Envir. Stud. Conf. Envir. Health Sustainable Devel.*, 2006, pp. 1–8.

[9] K. Bright, G. Cook, and J. Harris, "Colour, contrast and perception: Design guidance for internal built environments," Addkey, 1997, revised 2004. [Online]. Available: http://centaur.reading.ac.uk/11806/

[10] T. King, "Human color perception, cognition, and culture: Why red is always red," in *Proc. SPIE*, 2005, vol. 5667, pp. 234–242.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Lect. Notes Comput. Sci.*, vol. 3951, pp. 404–417, 2006.

[13] M. Liu and R. Siegwart, "DP-FACT: Towards topological mapping and scene recognition with color for omnidirectional camera," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3503–3508.

[14] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[15] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via PLSA," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 517–530.

[16] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 1, pp. 273–280.

[17] S. Vasudevan, S. Gachter, V. Nguyen, and R. Siegwart. (2007). Cognitive maps for mobile robots–An object based approach. *Robot. Autonom. Syst.* [Online]. *55(5)*, pp. 359–371, from Sensors to Human Spatial Concepts. Available: http://www.sciencedirect.com/science/article/B6V16-4MY0MK7-1/2/e379fd59a33b6d0a42355ba120c444e9

[18] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 1406–1413.

[19] O. Booij, B. Terwijn, Z. Zivkovic, and B. Krose, "Navigation using an appearance based topological map," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3927–3932.

[20] L. Zhao, R. Li, T. Zang, L. Sun, and X. Fan, "A method of landmark visual tracking for mobile robot," in *Proc. Ist Int. Conf. Intell. Robot. Appl.*, 2008, pp. 901–910.

[21] C. Valgren and A. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robot. Autonom. Syst.*, vol. 58, no. 2, pp. 149–156, 2010.

[22] M. J. Cummins and P. M. Newman, "FAB-MAP: Appearance-based place recognition and mapping using a learned visual vocabulary model," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 3–10.

[23] C. Wu. (Apr. 10, 2012). "SiftGPU: A GPU implementation of scale invariant feature transform (SIFT)," [Online]. Available: http://www.cs.unc.edu/∼ccwu/siftgpu/

[24] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Patt. Recognit.*, Jun./Jul. 2004, vol. 2, pp. II-506–II-513.

[25] J. Wu and J. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2010.

[26] M. Calonder, V. Lepetit, and P. Fua, "Keypoint signatures for fast learning and recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 58–71.

[27] J. Wu, H. Christensen, and J. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2009, pp. 4763–4770.

[28] J. Wu and J. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 630–637.

[29] A. Ranganathan, "PLISS: Detecting and labeling places using online change-point detection," presented at the Robot. Sci. Syst., Zaragoza, Spain, 2010.

[30] X. Meng, Z. Wang, and L. Wu, "Building global image features for scene recognition," *Pattern Recognit.*, vol. 45, pp. 373–380, 2011.

[31] A. Pretto, E. Menegatti, Y. Jitsukawa, R. Ueda, and T. Arai, "Image similarity based on discrete wavelet transform for robots with low-computational resources," *Robot. Autonom. Syst.*, vol. 58, no. 7, pp. 879–888, 2010.

[32] L. Payá, L. Fernández, A. Gil, and O. Reinoso, "Map building and monte carlo localization using global appearance of omnidirectional images," *Sensors*, vol. 10, no. 12, pp. 11 468–11 497, 2010.

[33] E. Menegatti, M. Zoccarato, E. Pagello, and H. ishiguro, "Image-based monte carlo localisation with omnidirectional images," *Robot. Autonom. Syst.*, vol. 48, no. 1, pp. 17–30, 2004.

[34] A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Wide-baseline visible features for highly dynamic scene recognition," in *Proc. Comput. Anal. Imag. Pattern*, 2009, pp. 723–731.

[35] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[36] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, and Q. Chen, "Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2009, pp. 116–121.

[37] P. Lamon, A. Tapus, E. Glauser, N. Tomatis, and R. Siegwart, "Environmental modeling with fingerprint sequences for topological global localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2003, vol. 4, pp. 3781–3786.

[38] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Aug. 2005, pp. 2429–2434.

[39] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor, "Omni-directional vision for robot navigation," in *Proc. IEEE Workshop Omnidirect. Vis.*, 2000, pp. 21–28.

[40] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2006, pp. 5695–5701.

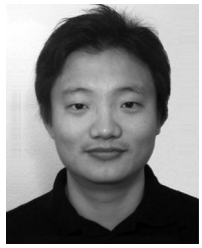
[41] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A robust descriptor for tracking vertical lines in omnidirectional images and its use in mobile robotics," *Int. J. Robot. Res.*, Special Issue Field Service Robots, vol. 28, no. 2, pp. 149–171, 2009.

[42] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[43] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[44] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 2, pp. 524–531.

[45] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 551–556.
[46] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[47] M. Liu, F. Colas, and R. Siegwart, "Regional topological segmentation based on mutual information graphs," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3269–3274.
[48] J. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains," *Neural Comput.*, vol. 23, no. 1, pp. 1–45, 2011.
[49] A. Ranganathan, "PLISS: Labeling places using online changepoint detection," *Autonom. Robot.*, vol. 32, no. 4, pp. 351–368, 2012.
[50] M. Liu, F. Colas, F. Pomerleau, and R. Siegwart, "A markov semi-supervised clustering approach and its application in topological map extraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2012, pp. 4743–4748.
[51] N. Kaempchen and K. Dietmayer, "Data synchronization strategies for multi-sensor fusion," in *Proc. IEEE Conf. Intell. Transport. Syst.*, 2003, pp. 1–9.
[52] M. Liu, L. Wang, and R. Siegwart, "DP-fusion: A generic framework for online multisensor recognition," in *Proc. IEEE Conf. Multisensor Fusion Integr. Intell. Syst.*, Sep. 2012, pp. 7–12.
[53] B. Douillard, D. Fox, and F. Ramos, "A spatio-temporal probabilistic model for multi-sensor multi-class object recognition," *Robot. Res.*, pp. 123–134, 2011.
[54] L. Orváth and P. Kokoszka, "Change-point detection with non-parametric regression," *Stat. J. Theoretical Appl. Stat.*, vol. 36, no. 1, pp. 9–31, 2002.
[55] B. Ray and R. Tsay, "Bayesian methods for change-point detection in long-range dependent processes," *J. Time Series Anal.*, vol. 23, no. 6, pp. 687–705, 2002.
[56] A. Ranganathan and F. Dellaert, "Online probabilistic topological mapping," *Int. J. Robot. Res.*, vol. 30, no. 6, pp. 755–771, 2011.
[57] D. Marinakis and G. Dudek, "Pure topological mapping in mobile robotics," *IEEE Trans. Robot.*, vol. 26, no. 6, pp. 1051–1064, Dec. 2010.
[58] B. Ryu and H. Yang, "Integration of reactive behaviors and enhanced topological map for robust mobile robot navigation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 29, no. 5, pp. 474–485, Sep. 1999.
[59] N. Tomatis, I. Nourbakhsh, and R. Siegwart, "Hybrid simultaneous localization and map building: A natural integration of topological and metric," *Robot. Autonom. Syst.*, vol. 44, no. 1, pp. 3–14, 2003.
[60] H. Choset and K. Nagatani, "Topological simultaneous localization and mapping (slam): Toward exact localization without explicit localization," *IEEE Trans. Robot. Autom.*, vol. 17, no. 2, pp. 125–137, Apr. 2001.
[61] H. Zender, O. Martínez Mozos, P. Jensfelt, G. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robot. Autonom. Syst.*, vol. 56, no. 6, pp. 493–502, 2008.
[62] J. Choi, M. Choi, and W. K. Chung. (2009, Oct.). Incremental topological modeling using sonar gridmap in home environment. in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pp. 3582–3587. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5354247
[63] S. Thrun, "Learning metric-topological maps for indoor mobile robot navigation," *Artif. Intell.*, vol. 99, no. 1, pp. 21–71, 1998.
[64] M. Liu, F. Colas, and R. Siegwart, "Regional topological segmentation based on mutual information graphs," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 3269–3274.
[65] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 219–236, 2007.
[66] M. Liu, C. Pradalier, Q. Chen, and R. Siegwart, "A bearing-only 2D/3D-homing method under a visual servoing framework," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 4062–4067.
[67] A. Argyros, K. Bekris, S. Orphanoudakis, and L. Kavraki, "Robot homing by exploiting panoramic vision," *Autonom. Robot.*, vol. 19, no. 1, pp. 7–25, 2005.
[68] O. Koch, M. Walter, A. Huang, and S. Teller, "Ground robot navigation using uncalibrated cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 2423–2430.
[69] J. Choi, M. Choi, K. Lee, and W. K. Chung, "Topological modeling and classification in home environment using sonar gridmap," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 3892–3898.
[70] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2009.
[71] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[72] A. Romero and M. Cazorla, "Topological slam using omnidirectional images: Merging feature detectors and graph-matching," *Adv. Concepts Intell. Vis. Syst.*, vol. 6474, pp. 464–475, 2010.
[73] E. Einhorn, C. Schröter, and H.-M. Gross, "Building 2d and 3d adaptive-resolution occupancy maps using nd-trees," in *Proc. 55th Int. Sci.* Colloquiium, Ilmenau, Germany, Verlag ISLE, 2010, pp. 306–311
[74] B. Font, F. Jesus *et al.*, "An inverse-perspective-based approach to monocular mobile robot navigation," *Materials*, vol. 18, pp. 2005–2012, 2012.
[75] J. Sethuraman, "A constructive definition of dirichlet priors," *Stat. Sinica*, vol. 4, pp. 639–650, 1994.
[76] R. Neal, "Markov chain sampling methods for dirichlet process mixture models," *J. Comput. Graph. Stat.*, vol. 9, no. 2, pp. 249–265, 2000.
[77] C. Geyer, "Practical Markov chain monte carlo," *Stat. Sci.*, vol. 7, pp. 473–483, 1992.
[78] N. Gagunashvili, "Chi-square tests for comparing weighted histograms," *Nuclear Instrum. Methods Phys. Res. Section A: Accelerat. Spectrometers Detect. Assoc. Equipment*, vol. 614, no. 2, pp. 287–296, 2010.
[79] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2548–2555.
[80] A. Ranganathan and J. Lim, "Visual place categorization in maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2011, pp. 3982–3989.
[81] A. Pronobis and B. Caputo. (2009, May). COLD: COsy localization database. *Int. J. Robot. Res.*, [Online]. *28(5)*, pp. 588–594, Available: http://www.pronobis.pro/publications/pronobis2009ijrr



Ming Liu (S'12) received the B.A. degree in automation from Tongji University, Shanghai, China, in 2005. During his Master's study from Tongji University, he was a Visiting Scholar with Erlangen-Nünberg University and Fraunhofer Institute IISB, Germany, for one year. As a result of his performance and under a special grant, in 2007, he was admitted as a Ph.D. student to Tongji University without attaining a master's degree. Since 2009, he has been working toward the Ph.D. degree in the Department of Mechanical Engineering, ETH Zürich.

His current research interests include autonomous mapping, visual navigation, topological mapping and environment modeling, etc.

Mr. Liu received the Best Student Paper award of IEEE MFI 2012.

Roland Siegwart (M'90–SM'03–F'08) received the Diploma degree in mechanical engineering and the Ph.D. degree in mechatronics from the Eidgenossische Technische Hochschule (ETH) Zurich, Zurich, Switzerland, in 1983 and 1989, respectively.

From 1989 to 1990, he was a Postdoctoral Fellow at Stanford University, Stanford, CA, USA. He was then a part time R&D Director at MECOS Traxler AG, Winterthur, Switzerland. He was also a part-time Lecturer and the Deputy Head of the Institute of Robotics, ETH Zurich, where, since July 2006, he has been a Full Professor of autonomous systems. In 1996, he was an Associate Professor, and later, a Full Professor for autonomous microsystems and robots with the Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, where he was Co-Initiator and the Founding Chairman of the Space Center EPFL and the Vice Dean of the School of Engineering. In 2005, he held visiting positions at the National Aeronautics and Space Administration (NASA) Ames, Moffett Field, CA, USA, and Stanford University, Stanford, CA. He leads a research group of around 35 people working in the field of robotics, mechatronics, and product design. He is a Coordinator of two large European projects, plays an active role in various rover designs of the European Space Agency, and is a Co-Founder of several spin-off companies.

Dr. Siegwart is a member of the Swiss Academy of Engineering Sciences and a Board Member of the European Network of Robotics. From 2004 to 2005, he was the Vice President for Technical Activities and a Distinguished Lecturer from 2006 to 2007. He was an AdCom Member (2007–2009) of the IEEE Robotics and Automation Society.