DP-FACT: Towards topological mapping and scene recognition with color for omnidirectional camera

Ming Liu, Roland Siegwart
Autonomous Systems Lab, ETH Zurich, Switzerland
ming.liu@mavt.ethz.ch, rsiegwart@ethz.ch

***Abstract*—Topological mapping and scene recognition problems are still challenging, especially for online realtime vision-based applications. We develop a hierarchical probabilistic model to tackle them using color information. This work is stimulated by our previous work [1] which defined a lightweight descriptor using color and geometry information from segmented panoramic images. Our novel model uses a Dirichlet Process Mixture Model to combine color and geometry features which are extracted from omnidirectional images. The inference of the model is based on an approximation of conditional probabilities of observations given estimated models. It allows online inference of the mixture model in real-time (at 50Hz), which outperforms other existing approaches. A real experiment is carried out on a mobile robot equipped with an omnidirectional camera. The results show the competence against the state-of-the-art.**

I. Introduction

The topological mapping and scene recognition techniques are efficient ways to model an environment with sparse information. When people describe their surroundings, they normally use unique labels of the places such as “my office”, “the first part of the corridor” etc. It implies that human have a mostly topological representation of their environment [2]. It highly depends on their ability to learn ego positions based on visual hints. This intuitive observation can be extended to similar tasks for mobile robots. For mobile robots, the ability to visually detect scene changes and recognise existing places are essential. Moreover, since robots may have multiple tasks at the same time, these detecting and recognition methods are preferable with an online fashion and with minimum computational and memory cost in real-time.

As far as the state-of-art is concerned, most existing place recognition systems assume a finite set of place labels. And the task is to classify the labels for each image frame. These classifier-based approaches [3] are limited with the applications in predefined or known environments. One of the mainstream techniques for visually scene recognitions are based on object detections [4], [5], [6], [7]. A representative scenario of these methods is to first detect known objects in the scene, then maximize the posterior of the place label given these recognised objects. Methods based on similar concept [8], [9] use keypoint based features for complete scenes. These methods are very robust when the objects are correctly detected. Nevertheless, the state-of-art object detection methods [10], [11] are usually computational expensive. They could be easily unfeasible on computers with limited resources, even with optimizations [12], [13], let alone the robot may have simultaneous tasks other than place recognition.

Several lightweight keypoint descriptors were developed [14], [15] as well and got widely applied in scene recognition problems [16], [17], [18]. Unfortunately, most applications only deal with either categorization of finite number of known places or only available for off-line inferences.

Beside the keypoint based approaches, descriptors using the transformation/inference of whole images [19], [20], [21], [22], [23], [24] are also popular. Amongst the most similar to our previous contribution – FACT [1], is the fingerprint of a place [25], [26]. Both fingerprint and FACT use segments from unwrapped panoramic images. The difference is that both [25] and [26] used laser range finder to help the matching of the descriptors, and FACT used only color information from the segments.

As far as sensors are concerned, omnidirectional vision has shown to be one of the most suited sensors for the scene recognition and visual topological mapping task because its 360° field of view [27], [28]. 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 into radial lines on the images. It means that the vertical lines are well preserved after the transformation [1]. Several other approaches utilize this feature as well, e.g. [25], [29].

Regarding inference approaches, hierarchical probabilistic methods based on statistical techniques won a great success in text mining and biological information processing [30], [31]. In this work, we alternate the classical mixture model to fit them with 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, inferred in real-time online. The related work of similar modeling method will be discussed in next section.

In our previous work [1], we developed a light-weight color based descriptor, which got compared from different perspectives[32], [33], [34], [35], [36]. According to our further study, the major disadvantages of the original approach are as below:

- The matching step was a point estimator, without considering probability and multi-hypotheses.
- The parameter-set was big. There were 5 parameters need to be adjusted.

This work was supported by the EU FP7 project NIFTi (contract # 247870) and EU project Robots@home (IST-6-045350).

- The false positive ratio of scene changing detection was high, therefore it required an off-line refinement.

Considering these shortcomings, we re-factor it into a probability-based framework, using an alternated Dirichlet Process Mixture Model. It enables multi-hypotheses and has the potential to grow to infinite number of clusters. There will be only one primary parameter to be adjusted, which defines the weighting between color features and geometric features. Meanwhile, the descriptors are re-organized by statistical models - histograms are used instead of vector-based representations. For the matching phase of the proposed approach, a nonparametric statistical test is used, replacing the thresholding of numeric distances. It helps to confine the matching result within the open-set of the statistical space.

The objectives that we want to achieve with this paper are as follows:

- Modelling and optimization of a color-based descriptor for unwrapped panoramic images, using hierarchical probability model;
- A concise approach for on-line inference of the proposed Dirichlet Process Mixture Model;
- Implementation of a scene recognition system for topological mapping, while fitting the real-time requirement.

The remainder of this paper is organised as follows. We will start with introducing further related work of hierarchical probability models and features. In section III, we provide a summary of FACT descriptors [1] and its optimizations. As a primary part of this paper, we introduce the novel approach and its inference method in section IV and section V. We introduce our real-time experiment in section VI. The conclusion and future steps of this work are given in the end.

II. Related Work

In most of the related works, change-point detection [37], [18], [38] is the basis to segment a video sequence. In this work, as we are targeting at a lightweight method, the change-point detection is not feasible when using multiple hypothesis methods, such as particle filtering [18]. Instead, we use non-parametric statistic test to evaluate the labeling for each frame separately. This may cause some unstableness in the output label. However, it relief the requirement of saving all the previous data of the sequence. We use an online median filter to smooth the label output, which gives good results shown in section VI.

The theoretical advances in hierarchical probability frameworks, such as LDA [31] and HDP [30], provide a good support for our algorithm. The Dirichlet Process Mixture Model enables countable infinite clusters for the measures, which can be used to define the process of scene recognition well. Fei-fei et al [39] proposed a keypoint based approach using this framework to cluster natural scenes. It can be considered as the most similar work. Nevertheless, the proposed work deals with less featured indoor environments by using much lighter descriptors.

As for color features, beside the fingerprint of place [25], a detailed report on the state-of-art can be found in [40]. Generally speaking, color feature is a weak descriptor, as it can be affected by lighting conditions easily. It is the main reason why we need to use a statistical method in this work to minimize the uncertainty.

III. FACT (Fast Adaptive Color Tags) and DP-FACT

The original *FACT* descriptor [1] at time t is defined as *FACT*$_t$ $:= \{T_{t1}, T_{t2}, \ldots T_{tD}\}$, where T_{tn} is named as a *Tag* which describes a single segment of the omnidirectional image, which is partitioned by dominant vertical lines. Note that $T_{tn} := (U_{tn}, V_{tn}, W_{tn})^T$, where D is the number of *Tag*s; U, V are features represented by UV color space; W is the width of a *Tag*. *DP-FACT* grants *FACT* descriptor with 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. w_t is a distribution over the geometrical space (factored by the width of *Tag*s), while g_t is over the discretized UV color space. Examples of g_t that extracted from two different nodes are shown in figure 1, placed in row order.

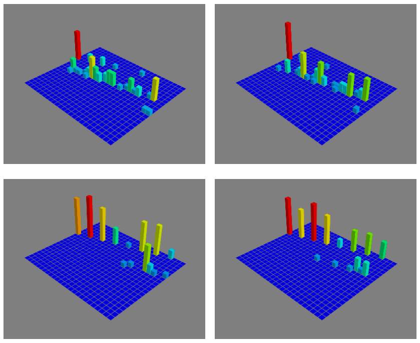


Fig. 1. Example of Histograms over discretized UV space

Intuitively, the distrubution 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 given in section V. For the details of segmentation of omnidirectional images, please refer to [1].

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 re-localizing in an existing topological environment model. The model that we consider in this paper is shown in figure 2. The parameters are depicted in rectangles, and random variables are in circles.

G is a Dirichlet process distributed with base distribution H and concentration parameter α. The base distribution is the mean of the DP and the concentration parameter α is as an inverse variance. The distribution G itself has point masses, and the draw from G will be repeated by sequential draws considering the case of an infinite sequence. Additionally, ϕ_t is an indicator of which cluster does the current image at time t belongs to.

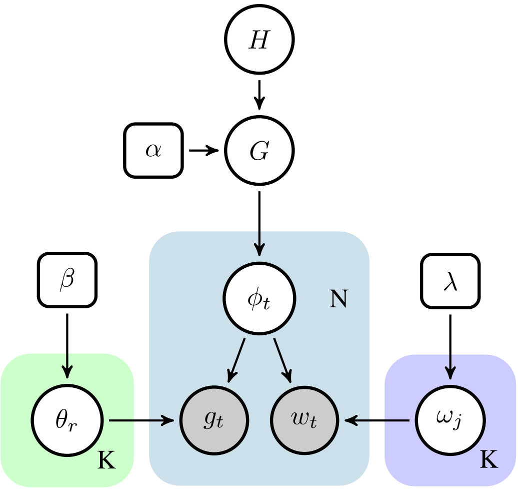


Fig. 2. System Model

The observable random variables from the model are two multinomial distributions g_t and w_t, which reflects two histograms by accumulating the number of features which hit their own discretized space. Taking w_t for an instance, it is a multinomial distribution that represents a single histogram of different width of *Tag*s in one *DP-FACT* feature. The dimensions of a draw 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 draws is represented by N, which is equal to the number of the sequential frames during one experiment over time.

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 $K \times Z$. w_t's of dimension Z are drawn from ω. Z is the number of possible histograms given the maximum number of *Tag*s of 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. Please notice 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,

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

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

Regarding a standard stick-breaking construction process [41], by integrating over G, the drawing of ϕ_t 's follows:

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

where δ_{ϕ_n} is an indicator of a certain frame n is labeled as ϕ_n, i.e. a mass point function locates at ϕ_n. The target problem is then converted to an estimation of $P(\phi_t \mid \phi_{\setminus t}, G, g, w, \omega, \theta; \beta, \lambda)$, where $\phi_{\setminus t}$ is the full set of indicators excluding the current one, namely the historical labeling.

The joint probability can be written directly as,

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

where α, β, λ are parameters. In order to factorize it to independent components, we integrate the joint probability over ω, θ and G,

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

The third part is actually an exception of G, i.e. $E_G\left[p(\phi_1\ \phi_2\ \phi_3\ \phi_4 \cdots\ \phi_N \mid G)\right]$. According to the features of the Dirichlet process, it is proportional to the product $\prod_{t=1}^{N} p(\phi_t \mid \phi_{\setminus t}) \propto p(\phi_N \mid \phi_{\setminus N})$. Therefore,

$$\int_G \int_H \prod_{t=1}^{N} p(\phi_t \mid G) p(G\ ;\ H\alpha)\ dHdG \propto \frac{\sum_{t=1}^{N-1} \delta_{\phi_t} + \alpha \delta_{\phi_{\bar{k}}}}{N-1+\alpha} \tag{2}$$

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

The first two parts can be treated in a similar manner. Take the first part for an instance, using n_v^j representing the number of frames whose width histogram is the v-th element in ω_j within cluster j.

$$\begin{aligned} & \int_\omega \prod_{j=1}^{K} p(\omega_j\ ;\ \lambda) \prod_{t=1}^{N} p(w_t \mid \omega_{\phi_t})\ d\omega \\ =& \prod_{j=1}^{K} \int_{\omega_j} p(\omega_j\ ;\ \lambda) \prod_{t=1}^{N} p(w_t \mid \omega_{\phi_t})\ d\omega_j \\ =& \prod_{j=1}^{K} \int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v - 1} \prod_{v=1}^{Z} \omega_{j,v}^{n_v^j}\ d\omega_j \\ =& \prod_{j=1}^{K} \int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v)}{\prod_{v=1}^{Z} \Gamma(\lambda_v)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v + n_v^j - 1}\ d\omega_j \end{aligned} \tag{3}$$

since

$$\int_{\omega_j} \frac{\Gamma(\sum_{v=1}^{Z} \lambda_v + n_v^j)}{\prod_{v=1}^{Z} \Gamma(\lambda_v + n_v^j)} \prod_{v=1}^{Z} \omega_{j,v}^{\lambda_v + n_v^j - 1}\ d\omega_j = 1 \tag{4}$$

Equation 3 can be continued as,

$$\begin{aligned}&\int_{\omega}\prod_{j=1}^{K}p(\omega_j\ ;\ \lambda)\prod_{t=1}^{N}p(w_t\mid\omega_{\phi_t})\ d\omega\\ &=\prod_{j=1}^{K}\int_{\omega_j}\frac{\Gamma(\sum_{v=1}^{Z}\lambda_v)}{\prod_{v=1}^{Z}\Gamma(\lambda_v)}\frac{\prod_{v=1}^{Z}\Gamma(\lambda_v+n_v^j)}{\Gamma(\sum_{v=1}^{Z}\lambda_v+n_v^j)}\end{aligned}\tag{5}$$

It is similar for the integration over θ. Therefore the joint probability is represented as follows.

$$\begin{aligned}&p(\phi\ \boldsymbol{g}\ \boldsymbol{w}\ ;\ \beta,\lambda)\\ &\propto\prod_{j=1}^{K}\frac{\Gamma(\sum_{v=1}^{Z}\lambda_v)}{\prod_{v=1}^{Z}\Gamma(\lambda_v)}\frac{\prod_{v=1}^{Z}\Gamma(\lambda_v+n_v^j)}{\Gamma(\sum_{v=1}^{Z}\lambda_v+n_v^j)}\\ &\prod_{j=1}^{K}\frac{\Gamma(\sum_{v=1}^{Y}\beta_v)}{\prod_{v=1}^{Y}\Gamma(\beta_v)}\frac{\prod_{v=1}^{Y}\Gamma(\beta_v+n_v^j)}{\Gamma(\sum_{v=1}^{Y}\beta_v+n_v^j)}\\ &\left(\frac{\sum_{t=1}^{N-1}\delta_{\phi_t}+\alpha\delta_{\phi_{\bar{k}}}}{N-1+\alpha}\right)\end{aligned}\tag{6}$$

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

$$p(\phi_t\mid\phi_{\setminus t}\ \boldsymbol{g}\ \boldsymbol{w}\ ;\ \beta,\lambda)\ \propto\ p(\phi_t\ \phi_{\setminus t}\ \boldsymbol{g}\ \boldsymbol{w}\ ;\ \beta,\lambda)\tag{7}$$

we could see that Z is a very big number, which makes the direct inference not possible. Usually sampling methods [42] will be used to estimate the posterior, with considerable time cost. Here we will 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 equation 7 as:

$$\begin{aligned}&p(\phi_t=k\mid\phi\ _{\setminus t}\ \boldsymbol{g}\ \boldsymbol{w})\\ &\propto\frac{\Gamma(\lambda_p+n_p^k)}{\Gamma(\sum_{v=1}^{Z}\lambda_v+n_v^k)}\frac{\Gamma(\beta_q+c_q^k)}{\Gamma(\sum_{u=1}^{Y}\beta_u+c_u^k)}\left(\frac{\sum_{t=1}^{N-1}\delta_k+\alpha\delta_{\phi_{\bar{k}}}}{N-1+\alpha}\right)\\ &=\xi(w_t\mid\omega_{\phi_t})\mu(g_t\mid\beta_{\phi_t})p(\phi_t\mid\phi\ _{\setminus t})\end{aligned}\tag{8}$$

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

V. Matching of DP-FACT

Most existing methods are off-line inference, mainly because the inference is time consuming, for example MCMC (Monte Carlo Markov Chain) sampling method [43] is considered as the standard approach [42]. In order to solve the inference problem in real-time with an on-line manner, the inference of the conditional probabilities may be approximated directly. When it is possible, it reliefs the need to calculate the joint probability. Recall the equation of the posterior of the place labelling depicted in equation 8. It includes three parts. The last part is a representation of a prior CRP (Chinese Restaurant Process) based on the previous observed labels. It can be calculate directly from the history measurements. The first two parts are similar. Usually they are estimated by sampling methods. When we have a closer look at them, we can see that they calculate the gamma function of the count of a certain observed over all the possibilities. In another word, 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, we don't need sampling methods 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. Non-parametric test

Since both observation and existing models are inherently histograms. Therefore the similarity between them can be estimated by non-parametric statistical methods. Here we introduce our approximation of equation 7 using $\chi-2$ test.

$\chi-2$ test is formalized as follows [44].

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

where $\hat{p}_t=\dfrac{n_t+m_t}{N+M}$, $N=\sum_{t=1}^{r}n_t, M=\sum_{t=1}^{r}m_t$, r is the dimension of both histogram; 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 equation 6, the observed distribution is determined by both the history observations and Dirichlet prior. However, the $\chi-2$ test only provides an estimate 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 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:

$$\begin{aligned}&p(\phi_t=k\mid\phi\ _{\setminus t},\ \boldsymbol{g}\ \boldsymbol{w})\equiv p(\phi_t\mid\phi\ _{\setminus t})\cdot\xi(w_t\mid\omega_{\phi_t})\cdot\mu(g_t\mid\beta_{\phi_t})\\ &\propto\left(\frac{\sum_{t=1}^{N-1}\delta_{\phi_t}+\alpha\delta_{\phi_{\bar{k}}}}{N-1+\alpha}\right)e^{-\rho\chi^2(w_t,\omega_k)-(1-\rho)\chi^2(g_t,\theta_k)}\end{aligned}\tag{9}$$

$\rho\in[0,1]$. If $\rho=1$, the estimator 9 will consider the geometry measure, *vice versa*. As a reminder of equation 8, the two targeting conditional probabilities are formalized as follows.

$$\begin{aligned}\xi(w_t\mid\omega_{\phi_t})&\propto e^{-\rho\chi^2(w_t,\omega_{\phi_t})}\\ \mu(g_t\mid\beta_{\phi_t})&\propto e^{-(1-\rho)\chi^2(g_t,\theta_{\phi_t})}\end{aligned}\tag{10}$$

B. Model update

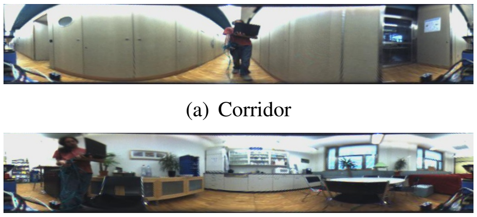
Despite the fast calculation, the non-parametric statistic that we introduced in equation 9 has an inherent disadvantage. We could see that the non-parametric test is a point estimation without considering history informations. In order to remit this disadvantage, a model update algorithm is developed. Comparing with equation 8, where the history information is represented by the counts of occurrences n_p^k and c_q^k, we require a method to take the history data into account. It means that the reference model 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.

$$\begin{aligned}\theta_k^{t+1} &= \frac{n_k^t}{n_k^t+1}\theta_k^t + \frac{1}{n_k^t+1}g_t \\ \omega_k^{t+1} &= \frac{n_k^t}{n_k^t+1}\omega_k^t + \frac{1}{n_k^t+1}w_t\end{aligned} \tag{11}$$

where n_k^t is the number of frames that have been clustered as with label k by time t. Therefore, the update process in equation 11 is a weighted mean over the old knowledge and the new observation at each time step. The advantage of this model update algorithm is obvious: In one hand, it can be calculated on-line with low requirements on computational and space costs; in the other other hand, it reflects the history data in the updated model directly.

VI. Experiments and reasoning

In this section, we introduce our experiment results in an indoor office environment. Our approach is compared with keypoint-based methods in terms of labeling accuracy, performance and inference complexity. Two samples of the unwrapped sample images are shown in figure 3.



(a) Corridor

(b) Coffee Room

Fig. 3. Sample images

A. Comparison in Accuracy

As described in [18], the SIFT feature has the superior accuracy in scene transition detection and recognition accuracy than CENTRIST and Texture based method. In this paper, we compare the proposed *DP-FACT* with SIFT as well as a newly developed lightweight keypoint descriptor BRISK [45].

As for the keypoint based methods(SIFT and BRISK), we use the unwrapped images as inputs. The algorithm is designed as follows. Firstly, keypoint-based feature extraction are then performed on the input images; then we match the current image with reference images which are observed in the past and try to get the most similar; if the matching result is good, we label the current image the same as the best matched reference; otherwise we consider the current image has a new label. The test result is shown in figure 4. In order to ease the comparison, the figures are aligned in time series. The first two figures are the raw output of *DP-FACT* and the result after an on-line median filter over the past 5 frames respectively. Please notice that further off-line smoothing of the labeling can be implemented as well [1], [46], which potentially provides more precise results. The result of keypoint based methods after the same median filtering are given after that. “Compressed Image Sequence” squeeze the whole sequence of observed images, from which the scene change can be intuitively observed. “Experiment Result” shows the output of *DP-FACT* according to the filtered labeling. The “Transition areas” indicate that the robot is at doorway or turning corners, where the scene recognition doesn’t make sense, which is not considered in the statistical results. Comparing with “Ground Truth”, the result of *DP-FACT* detects most transition points correctly and recognizes existing places online. In the contrary, keypoint based methods have high false positive ratio on the transition detection, because the labeling is dominated by massive changes of keypoints even in the save scene. Table I shows the accuracy of scene recognition. As the transition detection for keypoint based method is vague, the scene recognition result is calculated by considering non-repeated labels in the same scene as a group. Since *FACT* requires an off-line filtering, the comparison is not included. We could see that *DP-FACT* has the best recognition accuracy, though color is relative “weaker” features than keypoint descriptors. Two possible reasons can be considered: First is the distortion of the raw omnidirectional images cause non-uniform resolution of the unwrapped images. It makes the keypoint-based features unstable, especially when the keypoints are in different distances; Secondly, *DP-FACT* is structured only in horizontal direction, while keypoints can be possibly detected anywhere in images. This consistency of feature constructions makes the difference maximized between different labels and minimized inside the same one without the influence of unexpected randomness.

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

TABLE I
Accuracy of scene recognition

B. Evaluation in time cost

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

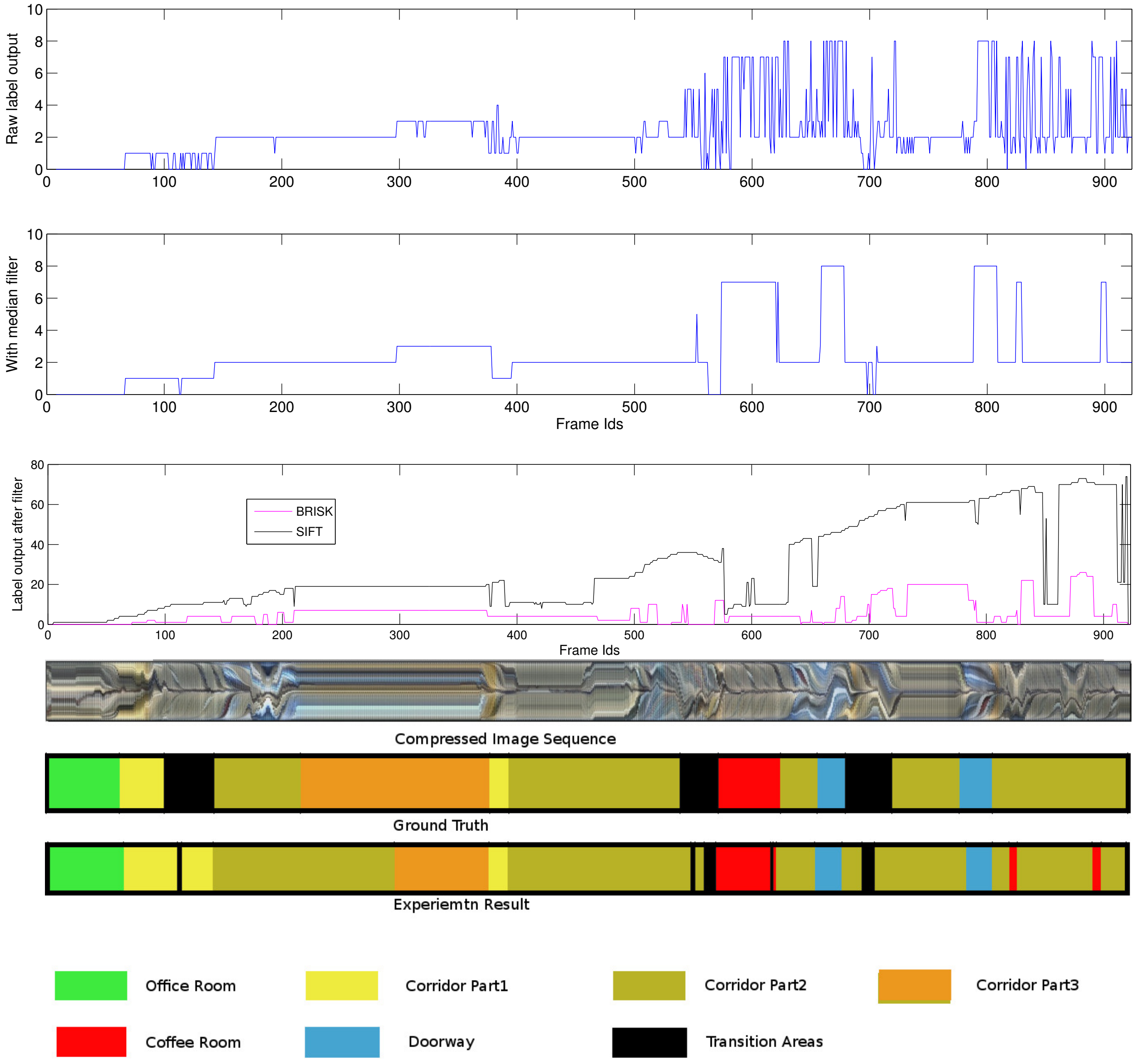


Fig. 4. Experiment results. From top to bottom: raw label output of DP-FACT; result of DP-FACT after median voting filter of 5 frames; result of keypoint based approaches (SIFT, BRISK) after median filter; image sequence in a compressed layout; labeled ground truth; result of DP-FACT; label explanations

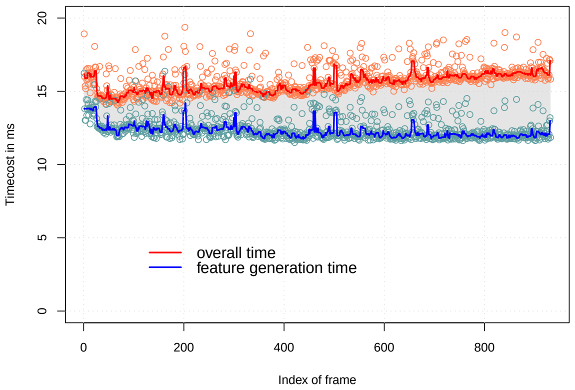


Fig. 5. Time cost of *DP-FACT* over frames. The lines are filtered results out of raw measurements (in circles). The gray area indicates the inference time.

We make a further study of the relation between the inference time and the complexity of the model. Figure 6 depicts a regression result of the inference time over the number of nodes, which is substantially linear. This result implies the potential of the proposed method can be extended to large scale environment without jeopardizing the realtime ability.

Recall the test in figure 4. In addition to the superior recognition accuracy, *DP-FACT* shows faster performance. Figure 7 depicts the comparison in time.

Our aim is to develop an online scene recognition algorithm which can be implemented online with limited computational resources. We launch the algorithm on three different types of CPUs in order to show that our method is feasible for different applications. The result is shown as figure 8. We see that even

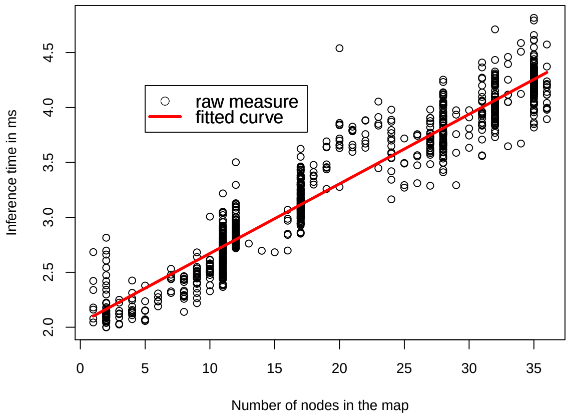


Fig. 6. Inference time vs number of nodes

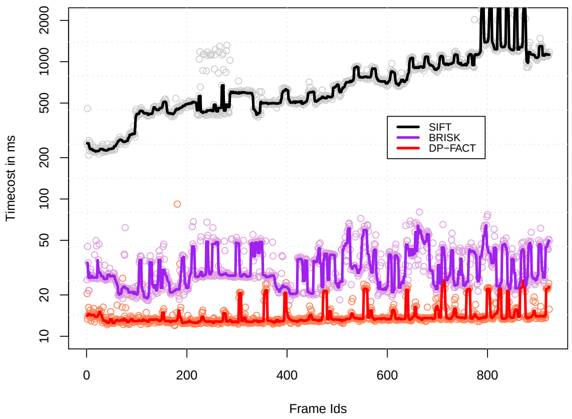


Fig. 7. Time cost comparison

for early CPUs, the algorithm can still reach around 17Hz.

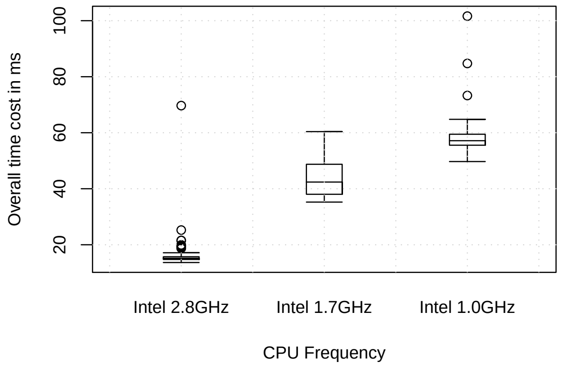


Fig. 8. Performance on difference CPUs, using single core

VII. Conclusion and Future Work

We have presented *DP-FACT*, an on-line scene recognition and topological mapping method using omnidirectional cameras, based on Dirichlet Process Mixture Model. It uses a weak descriptor of the environment - color - to describe features of scenes. The experiment result shows its advantage in the sense of on-line computing and low requirement in computational abilities. Above all, the accuracy and performance of *DP-FACT* is superior of other approaches such as popular keypoint based methods. This study also shows that the inference of a Dirichlet Process Mixture Model can be approximated by reasoning the conditional probability directly. We envision that similar concept can be borrowed to solve other inference problem with large data space as well.

It should be noted that this method only deals with the indoor environments, where vertical lines of the environment are preserved. The results do not imply that the extended applications for semi-structured environment is easily feasible. Not withstanding this limitation, this work does suggest that color based features can be integrated to a real-time online scene recognition and topological mapping robotics system can be envisaged. We can imagine the combination of keypoint and color based methods will help to solve this problem at a hybrid level, without limiting the targeting environment. The results will be carried out in our further study.

References

[1] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, and Q. Chen, "Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 116–121.

[2] T. McNamara, "Mental representations of spatial relations* 1," *Cognitive Psychology*, vol. 18, no. 1, pp. 87–121, 1986.

[3] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.

[4] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via plsa," *Computer Vision–ECCV 2006*, pp. 517–530, 2006.

[5] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003.

[6] S. Vasudevan, S. Gachter, V. Nguyen, and R. Siegwart, "Cognitive maps for mobile robots–an object based approach," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 359 – 371, 2007, from Sensors to Human Spatial Concepts. [Online]. Available: http://www.sciencedirect.com/science/article/B6V16-4MY0MK7-1/2/e379fd59a33b6d0a42355ba120c444e9

[7] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1406–1413.

[8] O. Booij, B. Terwijn, Z. Zivkovic, and B. Krose, "Navigation using an appearance based topological map," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 3927–3932.

[9] L. Zhao, R. Li, T. Zang, L. Sun, and X. Fan, "A method of landmark visual tracking for mobile robot," *Intelligent Robotics and Applications*, pp. 901–910, 2008.

[10] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, p. 404, 2006.

[12] C. Wu, "SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT)." [Online]. Available: http://www.cs.unc.edu/~ccwu/siftgpu/

[13] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

[14] J. Wu and J. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1489–1501, 2010.

[15] M. Calonder, V. Lepetit, and P. Fua, "Keypoint signatures for fast learning and recognition," *Computer Vision–ECCV 2008*, pp. 58–71, 2008.

[16] J. Wu, H. Christensen, and J. Rehg, "Visual place categorization: problem, dataset, and algorithm," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 4763–4770.

[17] J. Wu and J. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 630–637.

[18] A. Ranganathan, "Pliss: Detecting and labeling places using online change-point detection," *Proceedings of Robotics: Science and Systems, Zaragoza, Spain*, 2010.

[19] X. Meng, Z. Wang, and L. Wu, "Building global image features for scene recognition," *Pattern Recognition*, 2011.

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

[21] 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.

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

[23] A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Wide-baseline visible features for highly dynamic scene recognition," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 723–731.

[24] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE transactions on pattern analysis and machine intelligence*, pp. 300–312, 2007.

[25] P. Lamon, A. Tapus, E. Glauser, N. Tomatis, and R. Siegwart, "Environmental modeling with fingerprint sequences for topological global localization," in *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 4. IEEE, 2003, pp. 3781–3786.

[26] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, 2005, pp. 2429–2434.

[27] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor, "Omni-directional vision for robot navigation," in *omnivis*. Published by the IEEE Computer Society, 2000, p. 21.

[28] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. of the IEEE International Conference on Intelligent Systems, IROS06, Beijing, China*, 2006.

[29] ——, "A Robust Descriptor for Tracking Vertical Lines in Omnidirectional Images and its Use in Mobile Robotics," *International Journal of Robotics Research*, 2009, special Issue on Field and Service Robotics.

[30] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[31] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] A. Romero and M. Cazorla, "Topological slam using omnidirectional images: Merging feature detectors and graph-matching," *Advanced Concepts for Intelligent Vision Systems*, vol. 6474, pp. 464–475, 2010.

[33] M. Liu, C. Pradalier, Q. Chen, and R. Siegwart, "A bearing-only 2D / 3D-homing method under a visual servoing framework," in *IEEE International Conference on Robotics and Automation, 2010*, Anchorage Convention District, 2010, pp. 4062–4067.

[34] A. Romero and M. Cazorla, "Topological slam using omnidirectional images: Merging feature detectors and graph-matching," in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2010, pp. 464–475.

[35] E. Einhorn, C. Schröter, and H.-M. Gross, "Building 2d and 3d adaptive-resolution occupancy maps using nd-trees," in *Proceeding 55th Int. Scientic Colloquiium, Ilmenau, Germany*. Verlag ISLE, 2010, pp. 306–311.

[36] J. Blanc-Talon, D. Bone, W. Philips, D. Popescu, and P. Scheunders, *Advanced Concepts for Intelligent Vision Systems: 12th International Conference, ACIVS 2010, Sydney, Australia, December 13-16, 2010: Proceedings*. Springer, 2010.

[37] L. Orváth and P. Kokoszka, "Change-point detection with non-parametric regression," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 36, no. 1, pp. 9–31, 2002.

[38] B. Ray and R. Tsay, "Bayesian methods for change-point detection in long-range dependent processes," *Journal of Time Series Analysis*, vol. 23, no. 6, pp. 687–705, 2002.

[39] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. Ieee, 2005, pp. 524–531.

[40] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1582–1596, 2009.

[41] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[42] R. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, pp. 249–265, 2000.

[43] C. Geyer, "Practical markov chain monte carlo," *Statistical Science*, pp. 473–483, 1992.

[44] N. Gagunashvili, "Chi-square tests for comparing weighted histograms," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 614, no. 2, pp. 287–296, 2010.

[45] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," 2011.

[46] A. Ranganathan and J. Lim, "Visual Place Categorization in Maps," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011.(IROS 2011)*, 2011, p. (page unknown yet).