The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Visual Place Categorization: Problem, Dataset, and Algorithm

Jianxin Wu[†] Henrik I. Christensen[†] James M. Rehg[†]

Abstract— **In this paper we describe the problem of Visual Place Categorization (VPC) for mobile robotics, which involves predicting the semantic category of a place from image measurements acquired from an autonomous platform. For example, a robot in an unfamiliar home environment should be able to recognize the functionality of the rooms it visits, such as kitchen, living room, etc. We describe an approach to VPC based on sequential processing of images acquired with a conventional video camera. We identify two key challenges: Dealing with non-characteristic views and integrating restricted-FOV imagery into a holistic prediction. We present a solution to VPC based upon a recently-developed visual feature known as CENTRIST (CENsus TRansform hISTogram). We describe a new dataset for VPC which we have recently collected and are making publicly available. We believe this is the first significant, realistic dataset for the VPC problem. It contains the interiors of six different homes with ground truth labels. We use this dataset to validate our solution approach, achieving promising results.**

I. INTRODUCTION

Knowing "Where am I" has always being an important research topic in both the robotics and the computer vision communities. Various aspects of this problem have been extensively studied, giving answers in various granularities. For example, place recognition, or global localization, identifies the current position and orientation of a robot [9], [20] and seeks to find the exact parameterization of a robot's pose in a global reference frame. Topological place recognition answers the same question "Where am I", but at a coarser granularity [23]. In topological robot mapping, a robot is not required to determine its 3D location from the landmarks. It is enough to determine a rough location, e.g. corridor or office 113. Places in topological maps do not necessarily coincide with the human concept of rooms or regions [3]. Places in a topological map are usually generated by a discretization of the robot's environment based on certain distinctive features or events in the environment.

An alternative to recognizing specific, unique places is to recognize the semantic category of a place. Scene recognition, or scene categorization, is a term that is usually used to refer to the problem of recognizing the semantic label (e.g. bedroom, mountain, or coast) of a single image [10], [16], [24]. The input images in scene recognition are usually captured by a person, and are ensured to be representative or characteristic of the scene category. It is usually easy for a person to look at an input image in scene recognition, and determine its category label. The learned scene recognizer is generalizable, i.e. it can predict the category of scene images acquired in places that are not present in the training set.

As semantic information is attracting more research efforts in robotics [13], [28], it should be fruitful to combine facets from place recognition, topological mapping, and scene recognition together, and provide an answer to "Where am I" which contains more semantic information. In this paper, we raise a new problem called Visual Place Categorization (VPC). VPC refers to the identification of the semantic category of a place using visual information collected from an autonomous robot platform. For example, consider a robot operating autonomously in an unfamiliar home environment. The robot should be able to identify the type of room that it is in (bedroom, living room, kitchen, etc.), even if it is visiting that particular home for the first time. We utilize a single video camera as the sole sensor in our approach. Cameras have the advantage of being passive and non-obtrusive. In addition, this choice allows us to utilize the existing large body of computer vision research that extracts semantic information from images. A major challenge in categorizing places with a robot is the lack of a powerful attention mechanism that could automatically identify characteristic or distinctive views of a given place. For example, a person taking a picture of a kitchen will naturally frame the image to include representative details such as stove, sink, etc. In contrast, the video obtained from an autonomous robot will include many non-informative images.

This paper makes the following three contributions:

1) We introduce the Visual Place Categorization problem, a novel categorization problem for mobile robot navigation, which is related to scene recognition and place recognition for SLAM.
2) We present the first significant dataset for the VPC problem in home interiors, consisting of image sequences with ground truth labels captured from a variety of different homes.
3) We describe a solution architecture for VPC which supports frame-rate processing. The solution is based on the CENTRIST visual descriptor we recently proposed [25]. We then present some promising experimental results using our new dataset.

II. THE VISUAL PLACE CATEGORIZATION PROBLEM

We define Visual Place Categorization (VPC) as the problem of identifying the semantic category of a place using visual information collected from an autonomous robot platform. Some key aspects of the VPC problem include

- **The use of vision as the main sensing modality**. As cameras are becoming cheaper and on-board computers

[†]The authors are with the Center for Robotics & Intelligent Machines and the School of Interactive Computing, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. wujx2001@gmail.com, {hic, rehg}@cc.gatech.edu.

are becoming faster, vision (EO) sensors are gradually gaining in popularity for robotic systems. Real-time video processing is already employed in a variety situations for robot navigation, including stereo ranging and visual SLAM [4];

- **A focus on recognizing the semantic category of a spatial location**, such as the type of room (e.g. kitchen, living room, etc.) in a home environment, as opposed to its purely geometric or topological characteristics. Metric or topological information is useful in locating the robot on a map, while semantic location information can provide context for task execution. For example, a cleaning robot can adjust its strategy based upon the type of room it is in, and a delivery robot will be more useful if it can distinguish the loading dock of a building from the front reception desk;
- **An emphasis on autonomously-collected image data** (i.e. without human guidance), with the result that many images will not necessarily be characteristic or representative of the underlying category. Many previous works in scene recognition use collections of photographs taken by people. As a consequence, most images tend to capture the key elements of a place category. In contrast, the autonomously-collected frames used in VPC pose a new level of difficulty. Solution approaches must either simulate an attention mechanism or employ new image representations and temporal fusion methods that do not require a characteristic view. We describe a solution based on the latter approach which achieves promising results;
- **The need for good generalization across a wide range of spatial environments**. The ability to accurately predict the semantic category for a place that the robot has never visited before requires the ability to generalize effectively from a training set across a wide range of spatial environments. Given the inherent overlap between room functionalities (e.g. bedrooms that contain a sofa, living rooms that contain a writing desk, family rooms that contain a sink, etc.), it will be necessary to address label ambiguity in naming places and perhaps reason about *sub-locations* such as the smaller functional units inside a room (e.g. a breakfast bar within a kitchen area, or a computer desk within a family room).

We believe there are several application areas in robotics where VPC can provide useful functionality:

- **Human-robot interaction**. A key capability for HRI is the ability to communicate naturally about spatial locations. For example, VPC could enable a wheel-chair robot to understand commands such as *Move into the kitchen* or *Back up into the breakfast nook*.
- **Location-aware robots**. The VPC module could be seamlessly integrated with robot mapping modules such as SLAM (Simultaneous Localization And Mapping) [5]. We expect the synergy between SLAM and VPC to improve both systems. For example, VPC can help a topological mapping method so that it will not combine regions from different semantic categories into one topological location. On the other hand, regions that are both neighbors and belong to the same semantic category could be combined into a larger region that naturally corresponds to a room in an indoor environment (possibly with the help of other constraints such as the convexity constraint in [27]);
- **Object recognition and scene understanding**. The semantic category of an environment exerts strong priors on the objects that may appear within it [21]. Thus successful place categorization can aid object recognition and visual search. We believe that the recognition of place category and the objects contained in a place can be done synergistically. This would enable a robot to quickly identify objects of interest within an unknown environment. Furthermore, the semantic category of the local environment can provide context for a variety of additional sensing tasks.

III. Related work

While we believe our formulation of Visual Place Categorization is novel, it is clearly related to several previous research efforts. In this section, we briefly review the most relevant literature, which can be divided into three broad categories: place recognition and topological SLAM, semantic robot mapping, and scene recognition. As described in Section II, we believe VPC is largely complementary to place recognition and topological SLAM [5], [19], and so we focus our attention on the other two categories of work.

It is difficult to provide an exact definition for either *semantic knowledge* or *semantic mapping*. However, it is important to possess such information in a robot map in order for a robot to interact with a complex and non-static environment. Kuipers [8] defined Spatial Semantic Hierarchy (SSH), a hierarchical structure that encoded spatial knowledge at various abstraction levels. Early attempts tried to detect easy concepts (flat surfaces) such as walls, doors, ceilings, doorways, etc. For example, Liu *et al.* [11] built 3D models that consisted of these simple concepts using both the range sensor and a panoramic camera. Most of these methods used laser scanner data (possibly with the addition of vision sensors).

Problems that have similar formulation to place categorization have been previously presented. Places in the office environment are categorized in the systems of [18], [14], [6]. Topological maps were also built in these systems. The map was classified into place categories including office, laboratory, doorway, corridor, kitchen, and seminar room. These systems used both laser range sensors and cameras. They achieved reasonable recognition accuracies. The categories in these systems, however, are intuitively distinguishable by the geometric shape of objects contained in the scene (e.g. ceiling in the corridor, or door frame in the doorway). We are interested in recognizing more complex semantic categories based on their functionality to humans (e.g. bedroom vs. living room). Images from such complex categories will have

much larger intra-class variations than the categories studied in these earlier systems.

Torralba *et al.* [21] recognized both place instances (e.g. Jason corridor vs. Kevin corridor) and categories (e.g. corridor vs. conference rooms) using data collected from a mobile system. They achieved high accuracy in recognizing place instances. In the category level recognition, they achieved reasonable accuracy in 3 categories (conference rooms, corridor, and office), but failed to recognize other categories (kitchen, elevator, lobby etc.) Our conjecture is that objects in the 3 successful classes have a specific geometric shape, which helps in recognition. As mentioned above, in visual place categorization, images from the same category may have diverse visual patterns. Thus VPC is more complex than the problem reflected in the dataset used in [21]. That dataset was collected by mounting a camera on a person's head. As a consequence, the dataset reflects the attention capabilities of the human observer, which can be readily verified by comparing their dataset to our new VPC dataset.

Pronobis *et al.* [17] also recognized place categories (offices, corridor, printer area, and kitchen). The classifiers were designed to recognize place categories under various changes: weather conditions, moving persons and furniture, etc. However, they do not apply the learned category concepts to new environments. Instead, they tested the learned classifiers in the same part of the building where the training data were collected. The idea in [17] was extended to recognize places in different buildings in [22], which recognized four categories (corridor, bathroom, printer area, and office) in 3 environments. The resolution in [22] is 640x480, which is much smaller than the VPC dataset (1280x720).

IV. The Visual Place Categorization Dataset

In visual place categorization, we strive to categorize more complex and semantically richer places than what have been addressed in previous works. We choose to work with videos from home interiors, which include categories that exhibit larger visual variations, such as bedroom or living room. Since VPC in home interior is a new problem and to our knowledge there is no publicly available dataset that satisfies our problem definition, we collected a VPC Dataset and make it available at `http://categorizingplaces.com/dataset.html`. In this section we will describe the philosophy and procedures we followed in collecting our dataset.

We chose to collect data from home environments. Homes provide many room categories that are naturally defined by their function. It is also a trend to deploy intelligent software and hardware (including robots) in homes, e.g. to help take care of elderly people. We anticipate numerous applications of VPC in home environments. Categories in homes show much larger visual variations than office environments, even for category such as kitchen, which occurs in both environments.

Ideally, high resolution images with well-calibrated focus, appropriate viewing angles, even illumination, and white-balanced colors are desired. However it is difficult to achieve such desired settings simultaneously. We balanced these requirements by choosing a high definition camcorder (JVC GR-HD1) which captures 1280x720 images. We used the automatic settings of the camcorder to let it adjust the camera parameters during recording. The camcorder is able to automatically adapt to illumination changes in different rooms and adjust its focus and white balance. The camcorder is mounted on a rolling tripod to mimic a mobile robot platform. Although it is desirable to use a real robot, mobility and speed issues make this an impractical choice for capturing a large dataset in a wide variety of environments.[1]

To date, we have collected data from 6 homes and manually labeled 11 semantic categories. We asked the volunteers who allowed us to collect data in their homes to keep their homes as natural as possible. We made only two modifications to the home environments that we captured: First, we removed objects that could reveal the identity or the address of the occupants (e.g. family pictures or letters). Second, we closed the blinds in each room and relied upon artificial light. This helped to normalize the illumination environment across homes and times of day. Within each home, we captured two datasets. The first was a continuous run through the entire home, one floor at a time. During the continuous run, the operator mimicked the behavior of a robot following a predefined path through the home environment. He pushed the tripod with the camera facing forward, so that it traveled through all traversable areas in each room. The operator did not look at the captured video during recording, and simply ensured that the tripod followed a smooth path without colliding with any object in the room. Following this continuous capture, we went room-by-room and captured cylindrical panoramic video at two elevation angles. We did not use this second part of the VPC dataset for the experiments in this paper, but it is available for use by interested researchers.

Our protocol for data capture had two consequences for the images that we acquired: First, because the camera viewpoint simply followed the path of the tripod, uninformative views (such as a close-up of a section of wall) are a major portion of the captured video. Second, because the tripod often passed close by major furniture items such as beds and sofas, these objects are typically only partially visible in any specific frame. We believe these are realistic attributes for conventional video data collected by an autonomous platform in a home environment.

The VPC dataset was generated by extracting every third frame from the videos as JPEG (95% quality) images to keep the dataset to a manageable size. Each image is 1280x720 in resolution. Depending upon the size of the home, each home produced images totaling 1 to 2 gigabytes.

We provided manual annotations for this dataset. There are 11 categories (see Table I for category names). We used a special category name *transition* to annotate video segments

[1]We experimented with a PeopleBot platform with an attached Prosilica camera in our initial capture sessions. But we found that the tripod+camera solution made it much easier to quickly capture high-quality images and navigate in small spaces with challenging mobility requirements.

TABLE I: The 11 semantic categories in the VPC dataset, plus a special category named *transition*.

bedroom	bathroom	kitchen	storage closet
living room	dining room	family room	workspace
exercise room	media room	corridor	transition

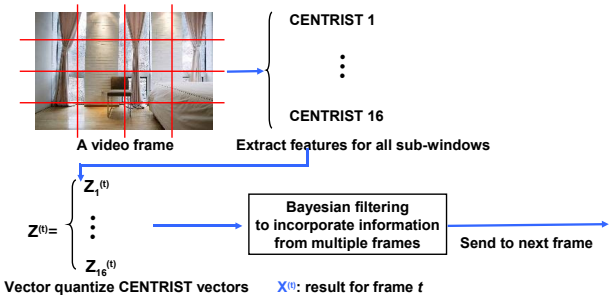


Fig. 1: Diagram of the VPC system.

that are either difficult to categorize or those that contain two or more categories. One category label is attached to a segment of the video (i.e. continuous image frames) instead of a single frame. Because of the autonomous image collection process, frames within a short contiguous time span have a high likelihood to share the same category label. This choice reduces the required manual labeling labor, but still retains enough information for learning place categories.

The homes in the VPC dataset span a wide range of styles and sizes, from modern suburban homes to Craftsman-style urban bungalows. The home owners span a variety of age groups, and include families with and without children. The homes vary in size and age and are designed and decorated in a variety of styles. Both single story and two-story homes are included, and some homes had a finished basement. Note that not all room categories are present in all homes. However, there are five categories that exist in all homes: bedroom, bathroom, kitchen, living room, and dining room.

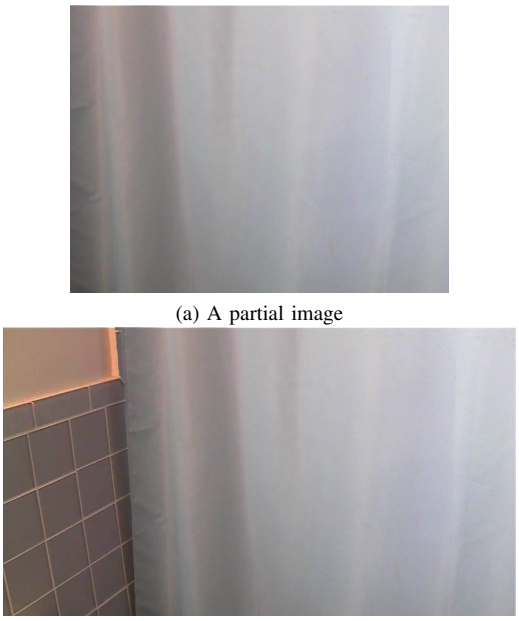
Along with the dataset, we also provided a baseline evaluation package, which uses leave one out cross validation and is based on per-frame accuracy. More evaluation details are described in Section V.

V. The Visual Place Categorization system

The block diagram for our VPC system is shown in Fig. 1. In the following, we will discuss each components in detail. A key aspect of our approach is a visual descriptor, CENTRIST (CENsus TRansform hISTogram) [25] and a Bayesian filtering approach [7], [21].

A. Image Representation

We adopt the "global configurations" approach proposed by Oliva and Torralba [16], which is also followed by quite a few other researchers. They found that perceptual properties such as the degree of naturalness can be reliably captured by their holistic representation, Gist. These perceptual properties were successfully used to recognizing outdoor scene categories in [16]. They showed that scene categories can be estimated without explicitly recognize objects in the scene.



(a) A partial image

(b) The complete image

Fig. 2: A bathroom image is shown in Fig. 2b. The shower curtain "object" (Fig. 2a, cropped from Fig. 2b) is not sufficient on its own to reveal the room category.

These perceptual properties are not as useful in indoor environments. For example, indoor environments all share low degree of naturalness. However, we believe that a holistic approach should still be used. As illustrated in Fig. 2, knowing the object in an image does not automatically tell us the place category. It is the conjunction of the curtain and the tiles on the wall that clearly show that this is a bathroom. Many useful cues such as the tiles in Fig. 2b are often contained in regions that are not objects. Furthermore, recognizing objects in cluttered environments is not necessarily easier than VPC itself. Thus we prefer a holistic representation.

Previously we showed in [25] that the CENTRIST visual descriptor (CENsus TRansform hISTogram) is a representation suitable for encoding place and scene images. In [25], CENTRIST achieved superior performances on both scene recognition (the 15 class scene recognition dataset [10]) and place recognition (the KTH IDOL dataset [17]).

Census Transform (CT) is a non-parametric local transform originally designed for establishing correspondence between local patches [26]. Census transform compares the intensity value of a pixel with its eight neighboring pixels, as illustrated in Eqn. 1. If the center pixel is bigger than (or equal to) one of its neighbors, a bit 1 is set in the corresponding location. Otherwise a bit 0 is set.

$$\begin{array}{|c|c|c|} \hline 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \\ \hline \end{array} \Rightarrow \begin{array}{c} 1\,1\,0 \\ 1\ \ 0 \\ 1\,1\,0 \end{array} \Rightarrow (11010110)_2 \Rightarrow \mathrm{CT} = 214 \qquad (1)$$

The eight bits generated from intensity comparisons can be put together in any order (we collect bits from left to right,

and top to bottom), which is consequently converted to a base-10 number in $[0\ 255]$. Just as other non-parametric local transforms which are based on intensity comparisons (e.g. ordinal measures [2]), Census Transform is robust to illumination changes, gamma variations, etc. Note that the Census Transform is equivalent (modulo a difference in bit ordering) to the *local binary pattern* code $LBP_{8,1}$ [15].

A histogram of CT values for an image or image patch, i.e. CENTRIST, can be easily computed, and we use CENTRIST as our visual descriptor. As shown in previous scene recognition research, incorporating spatial information greatly improves recognition accuracy. Thus we evenly divide an image into $4 \times 4 = 16$ sub-windows, and extract a CENTRIST descriptor from each sub-window. An image is represented by the concatenation of the 16 CENTRIST vectors (spatial CENTRIST).

We index the sub-windows in an image from 1 to 16 by their position. For each $i, 1 \leq i \leq 16$, we collect together all sub-window i CENTRIST descriptors across the entire training set. We then use the k-means clustering algorithm to generate a *visual codebook* with K centers for the i-th sub-window location. In total 16 visual codebooks are created, one for each sub-window location. Any CENTRIST vector from the i-th sub-window position will then be mapped to an integer between 1 and K using the i-th visual codebook. Thus an image is represented by a 16 dimensional vector

$$Z = (z_1, z_2, \ldots, z_{16}), \tag{2}$$

where z_i is the vector quantized index of the CENTRIST descriptor extracted from sub-window position i.

Let X be the category index of a video frame. Then we use a Naive Bayes approach to estimate $P(X|Z)$

$$P(X|Z) \propto P(Z|X)P(X) = \prod_{i=1}^{16} P(z_i|X)P(X), \tag{3}$$

in which $P(z_i|X)$ is easily estimated from the training data (i.e. set to the empirical distribution of the training set).

B. Bayesian filtering

Given that we are using a conventional video camera and we are not specifically identifying representative frames, the probability that a robot will both capture a representative frame and recognize the place category from such a frame is small. Thus it is vital to integrate information from many frames. We maintain a belief (i.e. the probability that the current frame belongs to a certain category) and use a Bayesian filtering approach for updating category beliefs. Specifically, let Z_t be the image observed at time t and $Z_{1:t}$ represent the image history till time t, i.e. the set of images observed from time 1 to t. Correspondingly, let X_t and $X_{1:t}$ be the category label at time t and the history of category labels till time t, respectively. Our purpose is to estimate the distribution $P(X_t|Z_{1:t})$.

The Bayesian filtering process exploits the entire image history to efficiently integrate information from several images. We assume a Markovian property between the category labels X, i.e. $P(X_t|X_{1:t-1}) = P(X_t|X_{t-1})$. Furthermore, we assume that the distribution of the observed image frame Z_t at time t is determined if we know the category label X_t at time t. Thus, the Bayesian filtering process is governed by three distributions [7], [21]:

1) The prior category distribution $P(X_0)$;
2) The category transition distribution $P(X_t|X_{t-1})$; and
3) The observation distribution $P(Z|X)$.

Using the three distributions and our independence assumptions, $P(X_{1:t}, Z_{1:t})$ can be factorized as:

$$P(X_{1:t}, Z_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i|X_{i-1}) \prod_{i=1}^{t} P(Z_i|X_i). \tag{4}$$

In the VPC system, the prior distribution $P(X_0)$ is a discrete uniform distribution since we assume the robot knows nothing about the environment at the beginning. The category transition distribution is specified as $P(X_t|X_{t-1}) = p_e$ if X_t equals X_{t-1}. We set p_e to a large number (e.g. we set $p_e = 0.99$ in our experiments) to reflect the fact that image frames within a consecutive time span have a high likelihood to share the same category label. The rest of the probability mass is shared uniformly among all the other values of X_t that is different from X_{t-1}. The last component, the observation model, is specified by Eq. 3.

After the three distributions are available, the desired quantity can be efficiently updated at each image frame, as shown in [7]:

$$P(X_t|Z_{1:t}) \propto P(Z_t|X_t)P(X_t|Z_{1:t-1}) \tag{5}$$

$$P(X_t|Z_{1:t-1}) = \sum_c P(X_t|X_{t-1} = c)P(X_{t-1} = c|Z_{1:t-1}). \tag{6}$$

A frame t is then classified as the category whose index is $\arg\max P(X_t|Z_{1:t})$ in the Bayesian filtering framework. When Bayesian filtering is not used, we use Eq. 3 to determine the X_t from Z_t alone. Since Bayesian filtering is an inexpensive operation, the running time remains about the same when Bayesian filtering is used on top of Eq. 3.

C. Experimental setup and evaluation methodology

We used $K = 50$ in our experiments, i.e. for each sub-window location, 50 visual codewords are generated. We used the k-means++ variant of k-means [1] to cluster CENTRIST vectors. We are interested in the spatial structure property of an image rather than detailed textural information. Thus, instead of extracting CENTRIST from input video frames, we first compute the Sobel gradients of the input image, and then CENTRIST descriptors are extracted from the Sobel gradient images.

Although there are 11 categories (plus a special *transition* category), only 5 categories are present in all homes. Thus we tested the proposed VPC system on these 5 categories: bedroom, bathroom, kitchen, living room, and dining room. Categorization results on frames whose groundtruth label is not within this set are simply ignored. In each home, the accuracy for a category is computed as the number of correct

TABLE II: Categorization accuracy (in percentages) of all homes and categories when the Bayesian filtering is used.

	bed	bath	kitchen	living	dining	average
home 1	75.76	80.04	12.03	43.90	11.15	44.58
home 2	67.10	32.14	64.37	2.04	13.78	35.89
home 3	80.07	95.32	26.14	3.26	0.00	40.96
home 4	49.77	63.92	69.06	30.50	36.41	49.93
home 5	81.47	86.41	45.05	21.30	0.30	46.91
home 6	35.17	90.81	72.77	22.54	56.00	55.46
average	64.89	74.77	48.24	20.59	19.61	**45.62**

TABLE III: Categorization accuracy (in percentages) of all homes and categories when the Bayesian filtering is not used.

	bed	bath	kitchen	living	dining	average
home 1	55.02	70.32	17.63	62.20	18.69	44.77
home 2	49.05	32.30	53.64	12.24	19.43	33.33
home 3	65.98	88.39	39.12	7.77	2.12	40.68
home 4	36.76	53.07	70.85	28.57	27.17	43.28
home 5	53.77	73.39	41.95	33.08	3.29	41.10
home 6	28.19	76.79	56.17	31.19	48.00	48.07
average	48.13	65.71	46.56	29.18	19.78	**41.87**

categorizations in this category divided by the total number of video frames in this category. The accuracy of a home is computed as the average accuracy of the five categories inside this home.

We used a leave one out cross validation strategy to evaluate the VPC system. The proposed method was applied 6 times. In each run, one home was reserved for testing and all other 5 homes were combined to form a training set. The overall accuracy of our VPC system is the average of the 6 individual homes.

Our visual place categorization system runs at approximately 20 frames per second.

D. Experimental results and comparisons

The categorization accuracies of our visual place categorization system are shown in Tables II and III, which show the results when the Bayesian filtering mechanism is used and when it is not, respectively.

The VPC system achieves a 45.62% overall accuracy when Bayesian filtering is used. Three categories (bedroom, bathroom, and kitchen) have relatively high accuracy (higher or close to 50%). The bathroom and bedroom categories have the highest accuracies, and are close to being useful in practice. However, the living room and dining room categories exhibit poor performance, close to that of random guessing (which is 20%).

The Bayesian filtering method improves both overall system accuracy and the categories with higher accuracy (bedroom, bathroom, and kitchen). However, living room and dining room are sacrificed when Bayesian filtering is applied. For example, the average living room accuracy is reduced from 29.18% to 20.59%. This phenomenon is expected. Bayesian filtering is effectively performing a smoothing operation. The worst categories (living room and dining room) will be treated as noise to some extent in the Bayesian filtering and their performances are hurt. The different effect

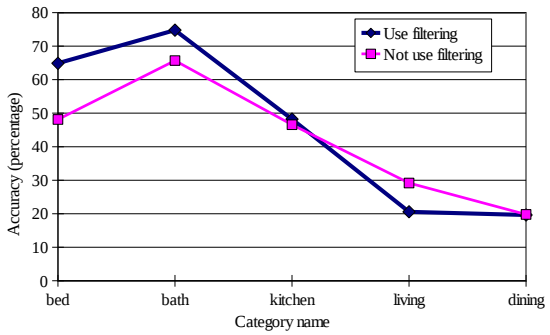


Fig. 3: Effect of using the Bayesian filtering.

of the Bayesian filtering on "good" and "bad" categories are shown in Fig. 3.

It is also interesting to note that although the accuracy of a category varies a lot in different homes (especially the living room and dining room categories), the average accuracy of homes remain relatively stable.

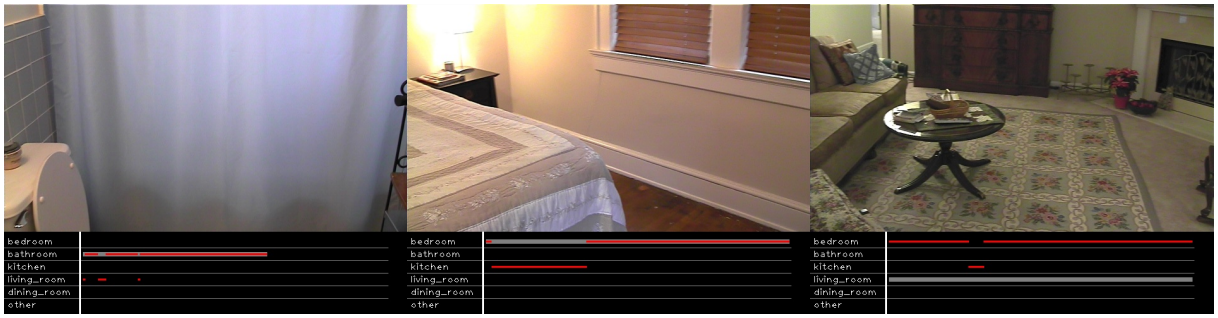
Figures 4a to 4c provides example visualizations for the VPC system results.[2] The gray bar indicates the groundtruth and the red bar is the categorization result. All categories that are not used and the special category *transition* are attributed to the *other* category in Fig. 4. The two bars progressed with time and the end of both bars indicated results for the current frame (e.g. around middle in Fig. 4a and at the end in Fig. 4b and 4c). As shown in Fig. 4, the bathroom and bedroom category are predicated well, with minor fragments in results and small periods of errors. However, the living room in Fig. 4c are poorly recognized.

Our conjecture about the low accuracy in both the living room and dining room categories is as follows. The key objects in these categories are usually very large and our camera can not capture the entire instance of such objects in a single frame. For example, we can only capture half (or even less) of the dining table in one frame due to the limited field of view of our camera. Similarly, our camera encounters the same problem with the big sofa in the living room. This limitation is clearly illustrated by the example frames in Fig. 5. These frames are taken from 4 different homes and in general most of the frames in these two categories suffer from the same limitation.

We used a global image representation and did not specifically detect any characteristic objects. However, we observed that the VPC system usually recovers from errors when such objects (e.g. sink in a kitchen and fireplace in a living room) come into the robot's sight. This observation makes us believe that we could use object recognition to help visual place categorization, and vice versa.

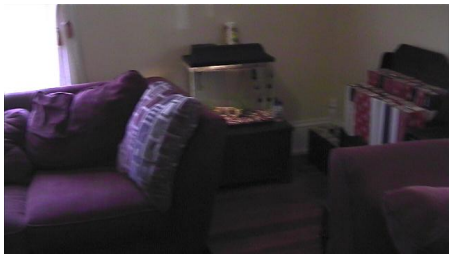
Our results show that Bayesian filtering improves the accuracy of the VPC system. However, the major virtue

[2]A complete video is provided as supplementary material along with this paper, which is the result for the second floor of home 5. Note that this floor only contains a bathroom and a bedroom, which are the best learned categories. Results for other homes and floors are generally inferior to the one shown in this video.

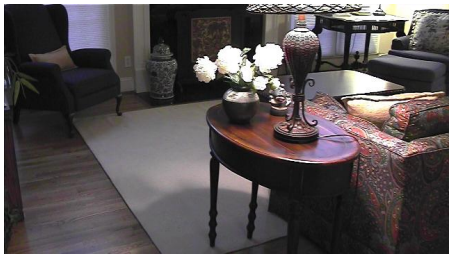


(a) (b) (c)

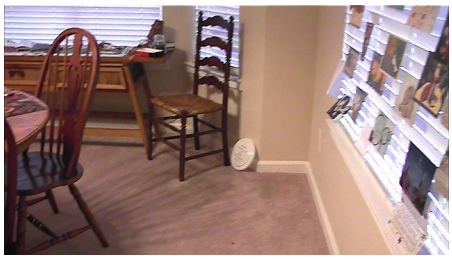
Fig. 4: Example results of the Visual Place Categorization system.



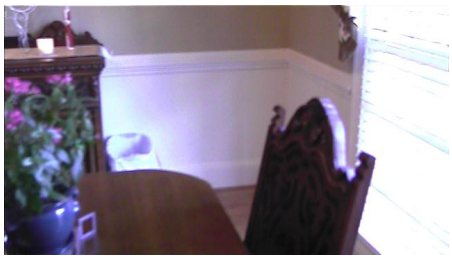
(a) Living room example frame 1



(b) Living room example frame 2



(c) Dining room example frame 1



(d) Dining room example frame 2

Fig. 5: Example of frames in the living room and dining room category.

TABLE IV: Results (overall system accuracy) comparing CENTRIST and SIFT visual descriptors.

	CENTRIST	SIFT
Using filtering	**45.62%**	38.61%
Not using filtering	**41.87%**	35.00%

of Bayesian filtering is to reduce fragmentation in categorization results, i.e. categorized labels now change less frequently. Comparing Fig. 6 with Fig. 4b (which shows results when Bayesian filtering was not used), the predicted labels without applying Bayesian filtering changed so quickly that they would not be useful to a robot system.

The CENTRIST visual descriptor is compared with the SIFT descriptor [12], a very popular visual descriptor. When we extract SIFT descriptors in each sub-window instead of CENTRIST vectors, the overall accuracy is 38.61% when Bayesian filtering is used and 35.00% when it is not. The SIFT-based results are lower than those using CENTRIST, by a large margin. The comparison is summarized in Table IV.

Finally, we want to note that visual place categorization is a challenging problem. For example, in the related work of [22], four place categories in office environments were recognized. While the corridor category was easy (about 55% to 90% accuracy), the other three categories (printer area, bathroom, and office) had very low accuracies (around 10%, which is much lower than the chance probability 25%). In comparison to [22], our CENTRIST visual descriptor and VPC system have shown promising results.

VI. Conclusions and Future Work

We have described the problem of visual place categorization (VPC), introduced the first significant dataset for VPC in home environments, and presented a solution approach based on spatial CENTRIST visual descriptors and Bayesian filtering. We believe that VPC is an important and interesting new problem for robot perception. Through the careful collection and annotation of the VPC dataset which is described in this paper, we hope to promote the study of the VPC problem among the computer vision and robotics research communities. We make this dataset publicly available in conjunction with this paper at `http:`

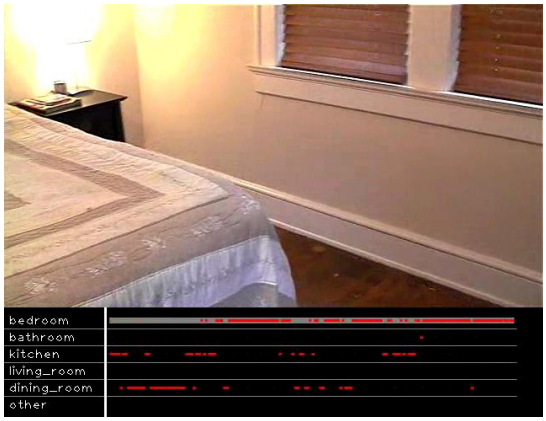


Fig. 6: Example VPC results when Bayesian filtering is not used.

//categorizingplaces.com/dataset.html. Our results demonstrate that it is possible to obtain surprisingly good performance without the need to construct specialized detectors for specific household objects.

We believe our dataset and experiments open up several interesting avenues for future research. Some questions for future exploration include: Can we define a saliency measure or attention mechanism that could be used to identify particularly representative or useful images? What role could geometric reconstruction play in combining image measurements over time and space? Are there more effective discriminative classification methods for making place-level category predictions?

One major limitation that has not yet been discussed is the lack of other sensory inputs, e.g. odometry and laser range sensor readings. Use of these sensors could further reduce the fragmentation of system predictions. For example, the laser range sensor readings should be able detect that the robot has not recently passed through a door and therefore it should have stayed in the same room. Similarly, the room category prediction must not change if odometry data primarily contains rotations. Furthermore, since we only need to provide a single category label for all the frames in the same room, we expect the categorization accuracy to improve by a large percentage if we are able to detect when the robot changes into a different room (e.g. utilizing the work of [6]). We also expect the VPC system to work better if it can exchange information with other modules in a robot, e.g. topological mapping.

REFERENCES

[1] D. Arthur and S. Vassilvitskii. **k-means++**: the advantage of careful seeding. In *18th Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

[2] D. Bhat and S. Nayar. Ordinal measures for image correspondence. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(4):415–423, 1998.

[3] H. Choset and K. Nagatani. Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization. *IEEE Trans. on Robotics and Automation*, 17(2):125–137, 2001.

[4] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *The IEEE Conf. on Computer Vision*, pages 1403–1410, 2003.

[5] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part I. *IEEE Robotics & Automation Magazine*, 13(2):99–108, 2006.

[6] S. Friedman, H. Pasula, and D. Fox. Voronoi random fields: Extracting topological structure of indoor environments via place labeling. In *Int'l Joint Conf. on Artificial Intelligence*, pages 2109–2114, 2007.

[7] M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[8] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1-2):191–233, 2000.

[9] B. Kuipers and P. Beeson. Bootstrap learning for place recognition. In *AAAI Conference on Artificial Intelligence*, pages 174–180, 2002.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 2169–2178, 2006.

[11] Y. Liu, R. Emery, D. Chakrabarti, W. Burgard, and S. Thrun. Using EM to learn 3D models of indoor environments with mobile robots. In *Int'l Conf. on Machine Learning*, pages 329–336, 2007.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. A. Baumann, J. J. Little, and D. G. Lowe. Curious George: An attentive semantic robot. *Robotics and Autonomous Systems*, 56(6):503–511, 2008.

[14] Ó. M. Mozos, C. Stachniss, and W. Burgard. Supervised learning of places from range data using AdaBoost. In *Proc. IEEE Int'l Conf. Robotics and Automation*, pages 1742–1747, 2005.

[15] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

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

[17] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A discriminative approach to robust visual place recognition. In *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, 2006.

[18] A. Rottmann, Ó. M. Mozos, C. Stachniss, and W. Burgard. Semantic place classification of indoor environments with mobile robots using boosting. In *AAAI Conference on Artificial Intelligence*, pages 1306–1311, 2005.

[19] S. Se, D. G. Lowe, and J. J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proc. IEEE Int'l Conf. Robotics and Automation*, pages 2051–2058, 2001.

[20] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust Monte Carlo localization for mobile robots. *Artificial Intelligence*, 128(1-2):99–141, 2001.

[21] A. B. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *The IEEE Conf. on Computer Vision*, pages 273–280, 2003.

[22] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, P. Jensfelt, and H. I. Christensen. Towards robust place recognition for robot localization. In *Proc. IEEE Int'l Conf. Robotics and Automation*, pages 530–537, 2008.

[23] I. Ulrich and I. R. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proc. IEEE Int'l Conf. Robotics and Automation*, pages 1023–1029, 2006.

[24] J. Wu and J. M. Rehg. Where am I: Place instance and category recognition using spatial PACT. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[25] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. Technical Report GIT-GVU-09-05, GVU Center, Georgia Institute of Technology, 2009.

[26] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conf. Computer Vision*, volume 2, pages 151–158, 1994.

[27] Z. Zivkovic, O. Booij, and B. J. A. Kröse. From images to rooms. *Robotics and Autonomous Systems*, 55(5):411–418, 2007.

[28] Z. Zivkovic, O. Booij, B. J. A. Kröse, E. A. Topp, and H. I. Christensen. From Sensors to Human Spatial Concepts: An Annotated Data Set. *IEEE Transactions on Robotics*, 24(2):501–505, 2008.