2016 13th Conference on Computer and Robot Vision

Indoor Place Recognition System for Localization of Mobile Robots

Raghavender Sahdev, John K. Tsotsos
Department of Electrical Engineering and Computer Science and Centre for Vision Research
York University
Toronto, Canada
{sahdev, tsotsos}@cse.yorku.ca

***Abstract*—In this paper we present a method for robots to do visual place recognition and categorization. The robot learns from experience and then recognizes previously observed places in known environments and categorizes previously unseen places in new environments. This system has been practically tested with a novel dataset developed by us to validate the theoretical results of the proposed system. A Histogram of Oriented Uniform Patters (HOUP) descriptor has been used to represent an image and then appropriate classifiers have been used to perform the classification tasks. It is shown that our method not only performs well on our dataset but also on existing datasets. A major contribution of this work is that this is the first real time implementation of a HOUP descriptor on two mobile robot platforms. Finally we built a novel dataset of seventeen indoor places for doing place recognition and validated our method in real time on this dataset.**

Keywords-place recognition; place categorization; HOUP; local binary patterns, support vector machines

I. INTRODUCTION

Autonomous Mobile Robots have been studied by a large number of researchers. One of the most important capabilities is Robot Localization. Robot Localization refers to answering the question for the robot, *"Where am I?"* Localization in general has two aspects qualitative and quantitative. The qualitative aspect of localization refers to knowing the place where the robot is present. For example – In a building, the robot should know that it is on a particular floor in room number 12 (which may be a seminar room, lab, conference room, etc.). The quantitative aspect of Localization allows the robot to have the knowledge about its coordinates in the particular room with a standard reference point. Quantitative localization is addressed by methods like visual odometry and LIDAR based approaches [1], [2].

In this paper our focus is to deal with the qualitative aspect of Localization. We focus on Place Recognition and Place Categorization. Place Recognition gives the robot the ability to know that it has been at a particular place before whereas Place Categorization allows the robot to know that it has been to a similar environment before. We implemented a Histogram of Oriented Uniform Patterns (HOUP) Descriptor as proposed by Fazl-Ersi and Tsotsos [3] and deployed the algorithm on two mobile robots – Virtual Me and Pioneer as shown in Figure 1. The HOUP descriptor is generated to perform the recognition and categorization tasks. For a given image sub block a HOUP descriptor is produced by passing the sub block through a Gabor Filter oriented in different orientations. The output of the Gabor filter is then used to generate Local Binary Patterns similar to the ones proposed by Ojala [4]. These patterns reflect the textural features in the image (curved edges, flat regions, dark spots, bright spots, etc.). We then use Principal Component Analysis to reduce the dimensionality of the descriptor. Finally classifiers like Support Vector Machines [5] and K-nearest neighbors are used for finding the type of place the image is of. We developed a dataset of seventeen different indoor places by moving the robots manually. The robot was made to traverse through all the places initially once and the robot was successfully able to recognize those places with a high accuracy.

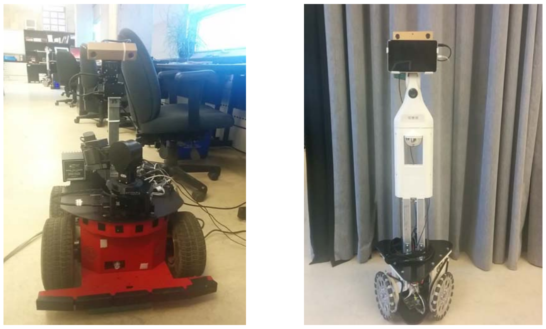


Figure 1. Pioneer (left) and Virtual Me (right) used for place recognition.

II. RELEVANT WORK

Earlier researchers used Laser range finders [6], [7], [8], [9], [10] and sonar based techniques [11]. Although such techniques have been widely used in the past and have displayed high performance, yet it has some drawbacks. It is often restricted to recognizing places with a similar geometric structure. If such a recognition system is asked to distinguish between places with similar geometric structure and a different appearance, it fails to do so. Ulrich and Nourbakhsh [12] developed an appearance based place recognition system for Localization by using color histograms, nearest neighbor matching and a simple voting scheme. They validated their system on 4 different places. Recently Johns and Yang [13] used RANSAC with 2D geometric cliques for learning the expected pair wise

978-1-5090-2491-9/16 $31.00 © 2016 IEEE
DOI 10.1109/CRV.2016.38

geometries for each place. They account for the underlying scene structure and possible view points by doing so.

Place Recognition for outdoor environments has recently been researched [14], [15], [16] and [17]. Chen et al. [15] used 21 layered Convolutional neural networks for doing place recognition. At the expense of accuracy they traded computational cost. Lee et al. [16] used line features to do place recognition in challenging outdoor environments. They leveraged the fact that man-made environments have lines most of the time and these lines are more robust to illumination, viewing direction or occlusions. Indoor place recognition has been done by Zender et al. [6] and Pronobis et al [18]. Another interesting work is that of Paul et al. [19] where they describe a probabilistic framework for appearance based navigation and mapping. Some other probabilistic appearance based methods include [20], [21] and [22].

Several landmark based approaches have been proposed. Such approaches suggest using local image features to represent and classify the scenes. Local Image features characterize limited areas of the image and they often provide more robustness against common image variations. Dudek and Jugessur [23] used visual features in the appearance domain to classify an object or a location. Similar to them we also focus on Qualitative localization of a place. One of the most famous descriptors being used for describing the local features in an image is the Scale Invariant Feature Transform (SIFT) of Lowe (2004) [24]. Lazebnik et al. (2006) [25] describes a method for recognizing scene categories based on approximate global geometric correspondence. Other local features based approaches include the works of Bay et al. [26] and Dallal and Triggs [27]. One of the famous context based approaches is that of computing the gist of a scene proposed by Oliva and Torralba [28], [29] and Oliva [30]. Oliva constructs the global scene representation of an image to build global features from the scene rather than focusing on local features. In this paper we use a similar technique proposed by Fazl-Ersi and Tsotsos [3] and apply it for the task of Place Recognition and validate its performance on our generated dataset.

III. HOUP Descriptor

Histogram of Oriented Uniform Patterns (HOUP) is a distribution based descriptor as suggested by the name itself. The initial image representation that is used to build the histogram describes the frequency content of the image; it can also be viewed as a descriptor based on spatial frequency. Figure 2 gives a general overview of the process of generating a HOUP descriptor for an image.

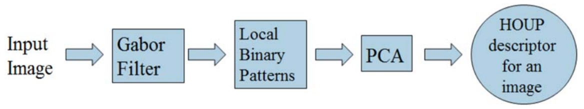


Figure 2. General Pipeline of a HOUP descriptor.

A. The Gabor Filter

Among different oriented filters, Gabor filters have received considerable attention. It has also been shown that these filters possess optimal localization properties in both spatial and frequency domain, and thus are well suited for texture analysis and encoding. Related work has been done by Torralba et al. [31] for encoding images for developing a context based Vision System for place and Object Recognition. In this paper too we use a similar method of initially encoding the image produced by passing it through a Gabor Filter.

Gabor Filters have been widely used for texture analysis, feature extraction, disparity estimation, etc. These filters are special types of filters which allow only a certain band of frequencies to pass through and reject the others. The filter can be mathematically represented as:

$$g(x,y;\lambda,\theta,\varphi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda}+\varphi\right)\right) \quad (1)$$

$$x' = x\cos\theta + y\sin\theta \;\; and \;\; y' = -x\sin\theta + y\cos\theta \quad (2)$$

- θ, theta is the orientation of the normal to the parallel stripes of a Gabor function.
- λ , lambda represents the wavelength of the sinusoidal factors
- φ, phi is the phase offset
- γ, gamma is the spatial aspect ratio
- σ, sigma is the standard deviation of the Gaussian envelope
- x and y are the coordinates of the pixels in the image

After generating the Gabor kernel we convolve the image with the kernel and get the filtered image.

$$v_k(x) = \left|\sum_{x'} i(x')g_k(x-x')\right| \quad (3)$$

Here $v_k(x)$ is the output of the convolved image with the with the Gabor filter $g_k(x-x')$ at a specific frequency and orientation. $i(x')$ is the input image to the Gabor filter. For computing the intermediate stage of the HOUP descriptor for an image sub block, we convolve the image with a Gabor Filter as described. Gabor Filters are generated at 6 different orientations and each orientation's output is then passed to a local binary pattern [4]. Detailed analysis is performed on Gabor coefficients and their joint distribution using local binary patterns. This is to aggregate encoded information at different locations into a low dimension image representation. The suggested aggregation method based on the uniform patterns boosts the discriminative power and generalizability of the representations; it produces scene representations with lower dimensions than most of the existing methods.

The selection of the parameters of the Gabor filters is an important task which needs to be addressed. There does not exist any clever method for selection of the parameters of the Gabor filter. The parameter values depend on the dataset which is being used so there are no generalized set of values for the Gabor filter parameters which produce the best possible performance. For this paper we used a Gabor filter tuned to 6 different orientations giving us $\theta = n\pi/6$ where $n = 0,1,2,3,4$ and 5. φ is set to 0. The remaining parameters γ, λ and σ have been chosen by searching the entire 3D space and set to the ones that give the best performance.

B. Local Binary Patterns

Local Binary Patterns were initially proposed by Ojala [4] for gray-scale texture classification. The method is based on recognizing that certain local binary patterns termed as 'uniform' are fundamental properties of local image texture, and their occurrence histogram proves to be a very powerful texture feature. Ojala derived a generalized gray-scale and rotation invariant operator presentation that allows for detecting the 'uniform' patterns for any quantization of the angular space and for any spatial resolution and presents a method for multi-resolution analysis. The approach of Ojala [4] is very robust in terms of gray-scale variations, since the operator by definition is invariant against any monotonic transformations of the gray scale. The proposed method of local binary patterns is also computationally simple as the operator can be implemented with a few operations in a small neighborhood and a lookup table. The most important property of using the local binary patterns (LBPs) is that certain LBPs termed as 'uniform' represent the fundamental properties (edges, corners and bright/dark spots) of the local image texture and they help in generating a generalized gray scale and rotation invariant operator for detecting these 'uniform' patterns.

Here we use the LBP of a local image region as follows. It's the same as described by Ojala [4].

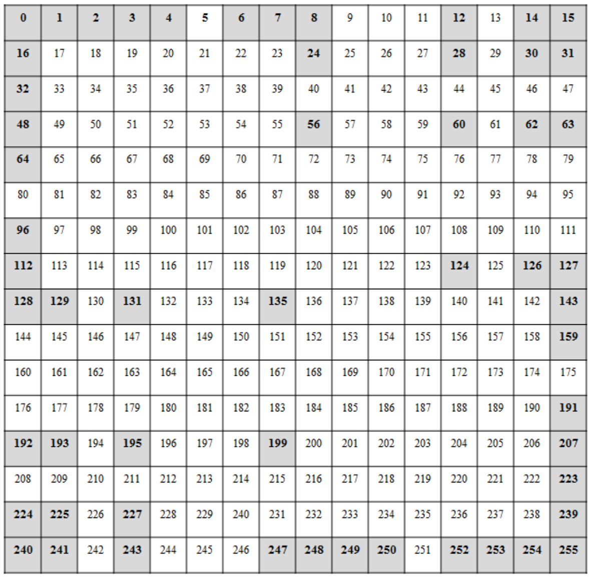
$$T \approx t(s(g_0 - g_c), s(g_1 - g_c), \dots , s(g_{P-1} - g_c)) \qquad (4)$$

where
$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (5)$$

T is the texture in a local neighborhood of a monochrome image as the joint distribution of gray levels of p $(p > 1)$ image pixels. g_c is the central image pixel. In this paper we use $p = 8$. Now for each sign $s(g_p - g_c)$ in equation (4), the terms are multiplied by a binomial factor of 2^p to transform each pixel into a unique number as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_0)2^p \qquad (6)$$

In this paper we set $p = 8$ as we consider a 3x3 neighborhood for a pixel. Equation (6) is computed for each pixel in a neighborhood region of 3 by 3 for an image, each neighborhood is thresholded at the gray value of the center pixel and converted into a binary pattern. The total number of binary pattern that can be generated with $p = 8$ is 256. A Histogram of local binary patterns is generated to count the total number of occurrence of each binary pattern in the image. Out of the 256 patterns only 58 patterns are uniform as shown in Figure 3. A uniform binary pattern is one in which the total number of transitions from 0 to 1 or 1 to 0 is at most 2. These uniform patterns represent the fundamental properties in an image. The pattern #0 (00000000) detects bright spots, patterns #4 (00000100), #8 (00001000), #12 (00001100) detects edges and so on.



0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	9	10	11	**12**	13	**14**	**15**
16	17	18	19	20	21	22	23	**24**	25	26	27	**28**	29	**30**	**31**
32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
48	49	50	51	52	53	54	55	**56**	57	58	59	**60**	61	**62**	**63**
64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
112	113	114	115	116	117	118	119	120	121	122	123	**124**	125	**126**	**127**
128	**129**	130	**131**	132	133	134	**135**	136	137	138	139	140	141	142	**143**
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	**159**
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	**191**
192	**193**	194	**195**	196	197	198	**199**	200	201	202	203	204	205	206	**207**
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	**223**
224	**225**	226	**227**	228	229	240	231	232	233	234	235	236	237	238	**239**
240	**241**	242	**243**	244	245	246	**247**	**248**	**249**	**250**	251	**252**	**253**	**254**	**255**

Figure 3. Highlighted numbers represent the decimal representation of the 58 uniform patterns out of the total 256 binary patterns in an LBP.

We introduce a 59th dimension to represent the non-uniform patterns; this is the sum of all non-uniform patterns. So in total we have 59 dimensional image representations for an image sub block. We then consider computing the Histogram of Oriented Uniform Patterns for each output of the oriented band pass filter. As we have 6 different orientations for the Gabor filter, we get 59 * 6 = 354 dimensional representation for an image sub block. This dimensionality is then reduced by selecting the first N principal components in such a way that the sum of chosen eigenvalues of the principal components accounts for more than 95% of the sum of all components. In our experiments the value of N is set to be 70 as it accounts for more than 95% of the sum of eigenvalues in most cases. So we select the first 70 principal components to act as representations for an image. Hence we have a 70 dimensional representation of an image sub block which we term as the "HOUP" descriptor for the image sub block. This is what we call one candidate feature for the image.

C. Subdivision Scheme

Here we divide an image to generate different features which would provide an informative representation of the image. We divide the given image into 1x1, 2x2, 3x3, 4x4 and 5x5 blocks. So in total we have 55 candidate features. Now a HOUP descriptor for each image is computed. It is observed that highest accuracy is achieved when using the 3x3 sub division scheme; we get 9 features each of 70 dimensionality. So the dimensionality for each image we generate is 70*9 = 630. So we use 630 numbers to represent an image. We do this for all images to generate a training dataset. On the training dataset we use appropriate classifiers to classify the image into its respective place category. For place recognition we use a 1 nearest neighbor classifier and for place categorization we use the support vector machine classifier. We use the LIBSVM tool proposed by [5] in this paper.

D. Feature Selection

Some previous approaches have used feature selection methodology to select the most informative features. Fazl-Ersi and Tsotsos [3] used the feature selection algorithm based on Kernel Alignment proposed by Christianini et al. [32]. It should be noted that we do not use any feature selection algorithm to select informative features because the computational cost of using these methods is not practically realizable for real time application in the field of robotics. We implemented feature selection and it was observed that for place recognition we achieved only 3% improvement in accuracy which was not a substantial increase also it drastically increased the computational time of our algorithm.

IV. Our Dataset

Several datasets exist for indoor visual place recognition such as USC dataset [33] developed by Siagian and Itti. Most of these datasets are limited to the variability they capture in terms of the number of places. The KTH IDOL dataset [18] consists of only 5 indoor places captured by 2 robots under varying illuminations conditions. Another interesting dataset built by Quattoni and Torralba [34] containing 67 scenes was built. However on this dataset current algorithms perform poorly making this dataset impractical for real world robotics applications as it would not be practical to have a robot navigate 67 different places. We in this paper developed a real time dataset of 17 different places. The dataset was built at two different locations under varying illumination conditions during day and night using two different robots. The dataset built used a Point Grey Bumblebee camera. The dataset developed is partly a binocular dataset which has 2 image representations of a scene – the left and right image.

A. Experimental Setup

Here we describe the experimental scenario and the data acquisition devices employed for the evaluation of our visual place recognition system. We tested it on two mobile robot platforms, "Pioneer" and "Virtual ME". The robot platforms used for data acquisition are shown in Figure 1. This dataset has been generated keeping in mind to have a dataset that can be publically used by researchers. It is a challenging novel partly stereo dataset acquired in two different lighting conditions. 11 of 17 places have stereo images captured. Other 6 have monocular images. We only use monocular images in this paper. Figure 4 and 5 show the different places captured in the dataset. In total we have 4 image sequences for the seventeen places captured at day and night. Each of the image sequence has approx. 4000-4200 images with 100-500 images belonging to each place. Each of the image sequences has minor variations in viewpoints for each scene.

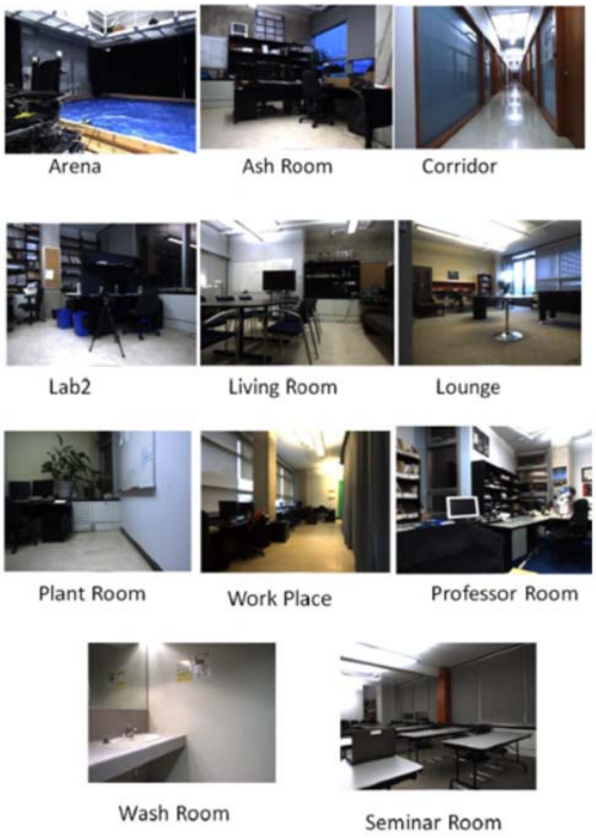


Figure 4. Eleven Places included in the dataset developed at York University, Canada.

Some of the places are treated as a different entity while some are separated by room dividers or curtains to mark off different parts of a big lab. Arena, Workplace, Ash Room and the living Room are different places of one big lab. Lab2, Plant Room and Professor Room are in the second big lab having the three different places as described above. Lounge, Seminar Room and Wash Room are three separate entities used to capture images and generate the dataset. Corridor is a place that essentially links the various places (labs, lounge, washrooms, seminar rooms, etc.) together.

Example picture of the eleven places can be seen in Figure 4.

As already mentioned the visual dataset was developed using the two robot platforms under two different lighting conditions day (when the natural sun light dominates) and night (when the rooms light has a significant effect on the place). The image acquisition was spread over a period of two weeks to generate the dataset. In this way we captured the visual variability that might have occurred.

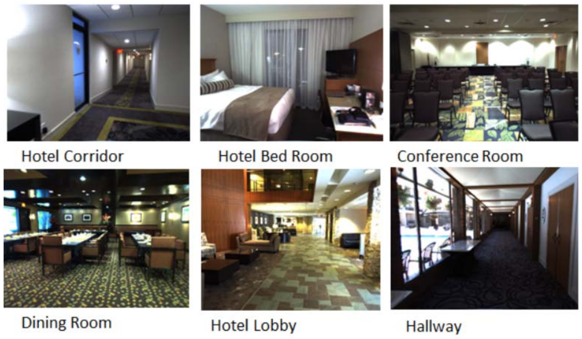


Figure 5. Six places dataset built at the Coast Capri Hotel, Kelowna, British Columbia

B. *Robot Platforms*

Both robots the White robot '*Virtual Me'* and the red colored robot *'Pioneer'* shown in Figure 1 are equipped with a directed perception pan tilt unit and a point grey stereo camera bumble bee. As can be seen in Figure 1, the cameras are mounted at different heights. On *Pioneer* the camera is 88 centimeters above the ground level, whereas on *Virtual me* it is 117 centimeters above the floor. All images were acquired with a resolution of 640 x 480 pixels, with the camera fixed at an upright position. The camera had the freedom to rotate on the spot for Pioneer robot; for virtual me the robot rotated on the spot which gave an indirect effect of having the camera rotate on the spot. The robot (virtual me) and pioneer's camera rotated in order to look around during the acquisition process.

We followed the same procedure during image acquisition with both robot platforms. The robots were manually driven (speed approximately 0.5 meters per second) through all the eleven places while continuously acquiring images at the rate of approximately 3 frames per second. For the different illumination conditions (day and night), the acquisition procedure was performed twice; resulting in two image sequences acquired one after the other giving a total of 4 sequences across a span of two weeks. Example images can be seen in Figure 4 and 5. Due to manual control the path of the robot was slightly different for every sequence. The process of labeling the places was done depending on a key press on the keyboard; a specific key on the keyboard was pressed depending on the place the robot was in at that particular time. Each image was accordingly labeled as belonging to one of the seventeen different places based on the position from where it was taken. For example the robot while standing on the exit of '*Work place'* views the living room is labeled as '*Work place'* because it took the image while it was in the place – *'Work place'.* Similarly for Robot standing in *'Dining Room'* looking at the '*Conference Room'* is labeled as '*Dining Room'* because it is physically in the place – '*Dining Room*'. In such situation we get some miss-classifications when robot is transitioning from one place to other.

C. *Experimental Results*

We conducted four sets of experiments in order to evaluate the performance of our system and test its robustness to different types of variations. We present the results in Table 1 and 2. We started with a set of reference experiments evaluating our method under stable illumination conditions (I). Next we increased the difficulty of the problem and tested the robustness of the system to changing illumination conditions (II) as well as to other variations that may occur in real-world environments. Next we moved on to see whether a model trained on images acquired from one device (robot) can be useful for solving the localization / recognition problems with a different device (robot) in similar illumination condition (III). Finally we modeled a system that would use images trained on one device under a specific lighting condition and test on a different device under different lighting condition (IV). We obtain encouraging results for all the 4 types of experiments conducted as can be seen in the next section. We conducted all 4 experiments for eleven places and conduct experiments (I) and (II) only for the seventeen places.

For the different image sequences different number of images for each place were present in all image sequences of the two robots in two lighting conditions. We built HOUP descriptors for an image sub-block. A sub-division scheme of 3x3 was used giving rise to 9*70 = 630 dimensional representation of each image. The classification algorithm being used is here is the 1 nearest neighbor classifier with the Spearman distance metric. For all the four types of experiments mentioned above same Gabor filter parameters and sub division scheme was used. We did not employ any feature selection method due to its practical infeasibility in our work for mobile robot localization.

We consider 4 different types of experiments conducted. Following types of experiments were conducted:

I. Same Robot Same Lighting Conditions
II. Same Robot Different Lighting Conditions
III. Different Robot Same Lighting Conditions
IV. Different Robot Different Lighting Conditions

Table 1 shows the results of our algorithm on eleven places in our dataset for the two different robots. Table 2 shows the results of our method on seventeen places for a single robot under different illumination conditions.

TABLE I. PERFORMANCE OF OUR METHOD ON OUR DATASET FOR VIRTUAL ME ROBOT FOR SEVENTEEN PLACES

#	Training Set	Testing Set	Lighting Conditions	Accuracy (%)
I	Virtual Me	Virtual Me	Same	98.34
II	Virtual Me	Virtual Me	Different	90.22

TABLE II. PERFORMANCE OF OUR METHOD ON OUR DATASET FOR THE TWO ROBOTS FOR ELEVEN PLACES

#	Training Set	Testing Set	Lighting Conditions	Accuracy (%)
I	Pioneer	Pioneer	Same	98
	VirtualMe	VirtualMe	Same	98
II	Pioneer	Pioneer	Different	93
	VirtualMe	VirtualMe	Different	93
III	Pioneer	VirtualMe	Same	92
	VirtualMe	Pioneer	Same	92
IV	Pioneer	VirtualMe	Different	82
	VirtualMe	Pioneer	Different	85

V. COMPARISON TO EXISTING APPROACHES AND DATASETS

Two sets of experiments were performed one each for place recognition and place categorization to compare our approach with existing methods.

A. The KTH IDOL Dataset – for Place Recognition

This dataset is well known for Topological Place Recognition. It was created by Pronobis et al. [18]. The purpose of this experiment is different from the previous one; this one is a recognition task not a categorization one. However this is also challenging as it provides images of different places under varying lighting conditions. This dataset is built in an office environment and has images belonging to 5 places – "*kitchen, corridor, one person office, two person office* and a *printing area*". The images have been captured by 2 robots Minnie and Dumbo under 3 different lighting conditions – night, sunny and cloudy.

TABLE III. PERFORMACE OF OUR METHOD ON THE KTH IDOL DATASET

#	Train	Test	Lighting	Performance			
				[35] A	[18] B	[3] C	Ours
1	Minnie	Minnie	Same	95.35	95.51	96.61	95.38
	Dumbo	Dumbo	Same	97.62	97.26	98.24	97.22
2	Minnie	Minnie	Diff.	90.17	71.90	92.01	85
	Dumbo	Dumbo	Diff.	94.98	80.55	95.76	88
3	Dumbo	Minnie	Same	77.78	66.63	80.05	72.46
	Minnie	Dumbo	Same	72.44	62.20	75.43	75.48

A) Wu & Rehg, 2011 [35], B)Pronobis et al., 2006 [18], C)Fazl-Ersi & Tsotsos, 2012 [3]

Table 3 lists the accuracy of our proposed methodology in comparison to others. In the paper by Fazl-Ersi and Tsotsos, 2012 [3], for experiment 1, 2 and 3 their feature selection algorithm selected 9, 13 and 23 features. The number of selected features increased with the difficulty of the experiments with maximum being for the third one.

Reasons for the lower accuracies than the best available Fazl-Ersi and Tsotsos [3]: (1) The accuracies reported above are from 9 features by using the 3x3 sub division scheme (2) We have used only 9 features to be used for all the three type of experiments. It can be seen that we get comparable results to the ones reported by Fazl-Ersi and Tsotsos [3]. We are off on average by around 4 %. (3) Feature selection has not been used because our place recognition system has to be used on an actual mobile robot.

Feature selection was implemented for some experiments and it did increase the accuracies by approximately 3-4%. But realizing the practical infeasibility of the feature selection algorithm, it was decided to not use it.

B. The UIUC Dataset – for Place Categorization

The UIUC dataset has been developed by Oliva and Torralba [28], Fei-Fei and Perona [36] and Lazebnik et al. [25]. This is one of the most commonly used databases for scene recognition in the field of Computer Vision. The dataset consist of 15 scene categories – "Suburb, Living Room, Forest, Mountain, Open Country, Street, Store, Bedroom, Industrial, Highway, Coast, Inside City, Office, Tall Building and Kitchen." Each class contains 210 – 410 images.

The standard procedure for experimenting with this dataset is randomly selecting 100 images for training and rest for testing. We here use the same standard protocol used for the dataset; we use 100 images selected from each category for training and use the remaining images in the dataset for testing. The procedure described in Fazl-Ersi [3] uses the feature selection algorithm to select the most informative features; they mention that on an average 43 features are selected from the pool of the 165 candidate features. This leads to 43 * 70 = 3010 dimensional representations to describe a single image. Fazl-Ersi compares the accuracy of his method with other state of the art methods and performs better than them. In his paper [3], he mentions that feature selection selects almost all 1x1 and 2x2 grids whereas 48% and 23% of the 3x3 and 4x4 blocks are selected.

All images have been resized to 256x256 as most of the images are closer to this number. We here use the 3x3 = 9 features to represent an image which leads to 9 * 70 = 630 dimensional representation for an image. The LIBSVM tool [5] is used as the classification algorithm for classifying the images. The LIBSVM tool with a variant of the OSS kernel [37] is used as the underlying kernel measure similar to that in [3]. We here use the LIBSVM tool [5] with the OSS$^+$ kernel to be used as a predefined kernel with the LIBSVM algorithm.

One of the most important tasks involved while using a classifier is to have an appropriate method to normalize the data depending on the classifier used. For example when using the SVM algorithm with a radial basis (rbf) kernel, the performance of the system would be poor unless the data is appropriately normalized. Similarly while using the LIBSVM tool with the OSS$^+$ kernel, the data has been scaled to be between 0 and 1. This is a crucial part and boosts the accuracy by 10 to 15 per cent as opposed to using it without scaling. Additional benefits and the difference that scaling can make for a building a successful system can be found in [38] where in the author mentions about various examples where scaling the data shoots up the accuracy by even 30%.

We achieve an accuracy of 72% by using the 9 features with the linear kernel; After employing the OSS$^+$ kernel, we get an accuracy of 75.33% as opposed to the 86% accuracy achieved by Fazl-Ersi [3]. We are off by 10 % for visual place categorization. Theoretically it is possible to get the accuracy as got by Fazl-Ersi and Tsotsos [3] by implementing the feature selection algorithm, but to obtain a similar accuracy in real time on a robot platform would require faster processors.

Reasons for low accuracy include: (1) we here have not implemented feature selection because of its computational inefficacy (2) we have a very compact representation of a single image with 9 features as opposed to the 43 features being used by Fazl-Ersi [3].

It is expected that after implementing the above points, the accuracy should be comparable to that stated in the paper [3]. However we argue about the infeasibility of having the feature selection algorithm to be implemented as it takes a large amount of time for finding the informative features.

VI. Implementation

We initially used Matlab to implement the algorithm being used in the paper. For deploying the software on the robots the matlab code was converted to C++ for efficient execution of our method. We did not achieve any latency while the robot was doing place recognition. A video of Virtual Me doing Place recognition is available at the link.

https://www.youtube.com/watch?v=k6E12Yp17X8

VII. Conclusion

In this paper we presented a practical real time vision based place recognition system for qualitative localization of a mobile robot. We started out by describing the HOUP descriptor and validated its efficiency through a series of experiments in the subsequent sections. Appropriate classifiers were chosen and used for the place recognition and categorization. We demonstrated for the first time a real time implementation of the HOUP descriptor on two mobile robot platforms. Another contribution of this paper is the development of a novel dataset for indoor place recognition.

Future work would include at aiming to integrate this qualitative localization approach with the quantitative localization using techniques like visual odometry for the robot to exactly know where exactly in a particular place the robot is in. An interesting modification to the presented approach would be using a simple Convolutional Neural Network with a layer of Local Binary Patterns to come up with a potentially better feature representation of an image.

Acknowledgment

The authors acknowledge the support of the NCFRN Network for funding this project. The authors would like to thank Dr. Ehsan-Fazl Ersi and Mr. Asheer Bachoo for the helpful discussions during the course of this project. We would like to thank the Coast Capri hotel for allowing us to build the dataset in the hotel. Finally we acknowledge the support by Birla Institute of Technology and Science, Pilani, India for allowing the first author to pursue this project at York University.

References

[1] J. Zhang and S. Singh, “Visual-lidar odometry and mapping: low-drift, robust, and fast,” *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[2] A. Howard, “Real-time stereo visual odometry for autonomous ground vehicles,” *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[3] E. Fazl-Ersi and J. K. Tsotsos, “Histogram of Oriented Uniform Patterns for robust place recognition and categorization,” *The International Journal of Robotics Research*, vol. 31, no. 4, pp. 468–483, 2012.

[4] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[5] C.-C. Chang and C.-J. Lin, “Libsvm,” *TIST ACM Transactions on Intelligent Systems and Technology ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Jan. 2011.

[6] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. Kruijff, and W. Burgard, “Conceptual spatial representations for indoor mobile robots,” *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, 2008.

[7] A. Rottmann, Ó. M. Mozos, C. Stachniss, W. Burgard, “Semantic place classification of indoor environments with mobile robots using boosting,” In *AAAI,* vol. 5, pp. 1306-1311, July, 2005.

[8] O. Mozos, C. Stachniss, and W. Burgard, “Supervised Learning of Places from Range Data using AdaBoost,” *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*.

[9] J. Asensio, J. Montiel, and L. Montano, “Goal directed reactive robot navigation with relocation using laser and vision,” *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*.

[10] S. Thrun, “Finding landmarks for mobile robot navigation,” *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146)*.

[11] S. Oore, G. E. Hinton, and G. Dudek, “A Mobile Robot That Learns Its Place,” *Neural Computation*, vol. 9, no. 3, pp. 683–699, 1997.

[12] I. Ulrich and I. Nourbakhsh, “Appearance-based place recognition for topological localization,” *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*.

[13] E. Johns and G. Yang, “RANSAC with 2D Geometric Cliques for Image Retrieval and Place Recognition,” *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] M. Milford, W. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, and D. Cox, “Condition-invariant, top-down visual place recognition,” *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[15] Z. Chen, O. Lam, A. Jacobson, M. Milford, "Convolutional neural network-based place recognition," *arXiv preprint* arXiv:1411.1509. 2014 Nov 6.

[16] J. H. Lee, S. Lee, G. Zhang, J. Lim, W. K. Chung, and I. H. Suh, "Outdoor place recognition in urban environments using straight lines," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, "Learning deep features for scene recognition using places database," In *Advances in Neural Information Processing Systems,* 2014.

[18] A. Pronobis, B. Caputo, P. Jensfelt, and H. Christensen, "A Discriminative Approach to Robust Visual Place Recognition," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[19] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," *2010 IEEE International Conference on Robotics and Automation*, 2010.

[20] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, Dec. 2010.

[21] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, Jan. 2008.

[22] W. Maddern, M. Milford, and G. Wyeth, "CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory," *The International Journal of Robotics Research*, vol. 31, no. 4, pp. 429–451, Jan. 2012.

[23] G. Dudek and D. Jugessur, "Robust place recognition using local appearance based methods," *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*.

[24] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*.

[26] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," *Computer Vision – ECCV 2006 Lecture Notes in Computer Science*, pp. 404–417, 2006.

[27] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[28] A. Oliva, A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 3, no. 42, pp. 145-175, 2001.

[29] A. Oliva and A. Torralba, "Chapter 2 Building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research Visual Perception - Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, pp. 23–36, 2006.

[30] A. Oliva, "Gist of the Scene," *Neurobiology of Attention*, pp. 251–256, 2005.

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

[32] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On Kernel Target Alignment," *Innovations in Machine Learning Studies in Fuzziness and Soft Computing*, pp. 205–256.

[33] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 2, pp. 300–312, 2007.

[34] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[35] J. Wu and J. M. Rehg, "CENTRIST: A Visual Descriptor for Scene Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, 2011.

[36] F.-F. Li and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[37] L. Wolf, T. Hassner, and Y. Taigman, "The One-Shot similarity kernel," *2009 IEEE 12th International Conference on Computer Vision*, 2009.

[38] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," 2003, pp. 1-16.