2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Incremental Unsupervised Topological Place Discovery

Liz Murphy[1] and Gabe Sibley[1]

***Abstract*— This paper describes an online place discovery and recognition engine that fuses information over time to create topologically distinct places. A key motivation is the recognition that a single image may be a poor exemplar of what constitutes a place. Images are not 'places' nor are they 'documents'. Instead, by treating image-sequences as a multi-modal distribution over topics – and by discovering topics incrementally and online – it is possible to both reduce the memory footprint of place recognition systems, and to improve precision and recall. Distinctive *key-places* are represented by a cluster topics found from the covisibility graph of a relative simultaneous localization and mapping engine – key-places inherently span many images. A dynamic vocabulary of visual words and density based clustering is used to continually estimate a set of visual topics, changes in which drive the place-recognition process. The system is evaluated using an indoor robot sequence, a standard outdoor robot sequence and a long-term sequence from a static camera. Experiments demonstrate qualitatively distinct themes associated with discovered places – from common place types such as 'hallway', or 'desk-area', to temporal concepts such as 'dusk', 'dawn' or 'mid-day'. Compared to traditional image-based place-recognition, this reduces the information that must be stored without reducing place-recognition performance.**

I. Introduction

Visual appearance-based mapping relies on similarities in sensory input gathered at different times to recognize the same place, and is often coupled with metric SLAM techniques to perform loop-closure. In a typical appearance-based mapping framework, nodes in a topological graph (the map) are associated with images gathered by the robot as it traverses the environment. The decision to create graph nodes or *key-frames* is typically an arbitrary one — based on the distance between frames in space or time — and as such the key-frames do not necessarily represent distinct 'places'.

In this work we approach the selection of *key-places* from a principled semantic viewpoint, modeling image-sequences with a low-dimensional thematic descriptor and monitoring change at that descriptor level in order to obtain 'places' more consistent with a human definition of 'place'. We view the image stream as a story-line and when that story-line starts to change we infer a new place. This move away from representing the world as a set of arbitrary images has two principle benefits: distinct places represent an aggregate view of many key-frames, so fewer representative images need to be kept; and the distance (in image space) between places is further so place recognition performance is improved.

[1]The authors are with the Department of Computer Science, George Washington University, Washington DC, 20052, USA liz_murphy@gwu.edu

This work is made possible by the generous support of Motorola Mobility, Inc.

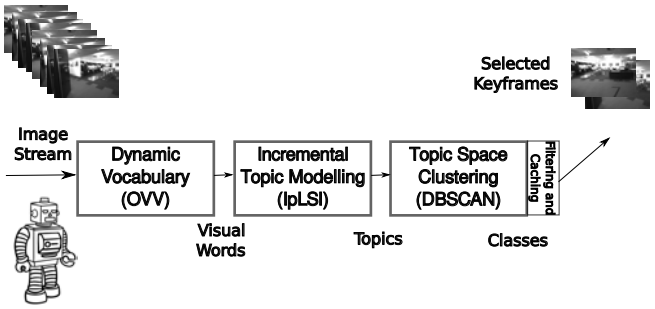


Fig. 1: Block diagram of the topological place discovery approach

The contribution of this work is a technique that performs unsupervised topological place recognition, and which operates on the fly with online visual vocabulary updates and incremental topic modeling. Specifically we integrate an Online Visual Vocabulary (OVV) builder [1] into a relative frame visual SLAM system [2] to describe images. Image descriptions are passed to an incremental online topic modeling engine [3] which produces a low-dimensional thematic descriptor from which distinct topological places are synthesized using density based clustering. Figure 1 outlines the approach. We evaluate the application of our system to the problem of topological place discovery using several robot-gathered image sequence datasets including an indoor sequence, a standard outdoor sequence, and a long term sequence from a static camera.

II. Related Work

Graph based approaches such as appearance-based maps and pose graph SLAM are common environmental representations in contemporary robotics systems, favored because they offer a sparse representation and scale reasonably well as the environment size increases. Recently, attention in the field has turned towards enabling compact graph-based approaches that can operate robustly over the long-term, through information theory based approaches to pruning the graph [4] [5] or by deleting out-of-date views [6].

Others have approached the process of selecting and culling places to represent in the map from a semantic viewpoint, such as by detecting landmarks (distinctive parts of the environment) using Bayesian surprise [7], and identifying topologically distinct places from spherical image sequences using a custom algorithm which monitors changes in the spatial envelope of a spherical GIST global descriptor [8]. In [9], four topological image sequence partitioning techniques are presented which each employ two levels of loop closure to facilitate partitioning the image sequence into topological

places.

In this work we also approach the selection process from a semantic viewpoint, monitoring change in a low-dimensional thematic descriptor. To produce the descriptor we make use of topic modeling, adapting a tool originally designed for text classification to images. Much of the topic modeling literature is devoted to discovering topics in static document corpora with fixed, finite vocabularies, and this has been applied to robotic navigation; to summarize the visual experience of the robot [10] and to detect confusing images in the context of place recognition [11]. Online topic modeling is used in [12] to control the speed of an underwater vehicle for image acquisition using a static vocabulary that produces a representative set of images of the aggregate image stream. While topics are learned online, a batch refinement of the entire dataset is done after updates to the topic model to associate previously seen documents with the updated topics. This approach is needed when the accuracy of topic labeling over the entire dataset is of utmost importance, but this is not the case in our application — we are not interested in retrieving exemplars of *kitchen*, *hallway* or *bathroom* scenes from an indoor dataset, but we want to recognize when the robot moves from the *hallway* into the *kitchen*. In our case the continuity in the model is of primary importance — when we add in new documents we want to ensure that the storyline depicted in the topic assignments of recent documents remains relatively stable. This leads us to draw on a growing body of work that models the changing nature of topics in infinite streaming datasets such as the Web [3], where the topic models are updated online without need for batch refinement over the whole stream, and new 'words' are able to enter the vocabulary and influence topic modeling.

III. Approach

Our approach to online place discovery has three primary components which we discuss in turn here.

A. Dynamic Vocabulary Building

A major downfall of many Bag-of-Words place recognition algorithms is their reliance on static vocabularies that must be learned from existing image corpora. Problems arise when the robot moves to an environment dissimilar to that it was trained on, and doesn't have the words to allow it to describe the scenes it encounters so that they are easily discriminated from others. In this work we synthesize a recent technique, Online Visual Vocabulary (OVV) of [1] and outline its workings briefly here.

OVV builds a forest of visual words using agglomerative clustering. *Elementary* clusters are initialized from feature tracks associated with individual landmarks in the covisibility graph of a relative frame visual SLAM system [2], and each cluster is represented by the mean C_k and covariance R_k of its n constituent feature track descriptors f_k^i.

$$C_k = \frac{\sum_{i=1}^{n} f_k^i}{n} \tag{1}$$

$$R_k = \frac{\sum_{i=1}^{n} (f_k^i - C_k)(f_k^i - C_k)^T}{n-1} \ . \tag{2}$$

OVV has 3 critical operations which build and maintain the forest of words: elementary clusters are *added* to the vocabulary, and existing clusters within the vocabulary are *merged* and *pruned*. Merging and pruning are governed by a metric that seeks to maximize the distance between clusters whilst making individual clusters as compact as possible. The metric is based on Fisher's linear discriminant

$$Q = \frac{tr(S_B)}{tr(S_w)}. \tag{3}$$

S_B is the *between clusters scatter matrix*

$$S_B = \frac{1}{N} \sum_{k=1}^{N} n_k (C - C_k)(C - C_k)^T \tag{4}$$

where C is the global data centroid. S_W is the *within clusters scatter matrix*.

$$S_W = \frac{1}{N} \sum_{k=1}^{N} n_k R_k. \tag{5}$$

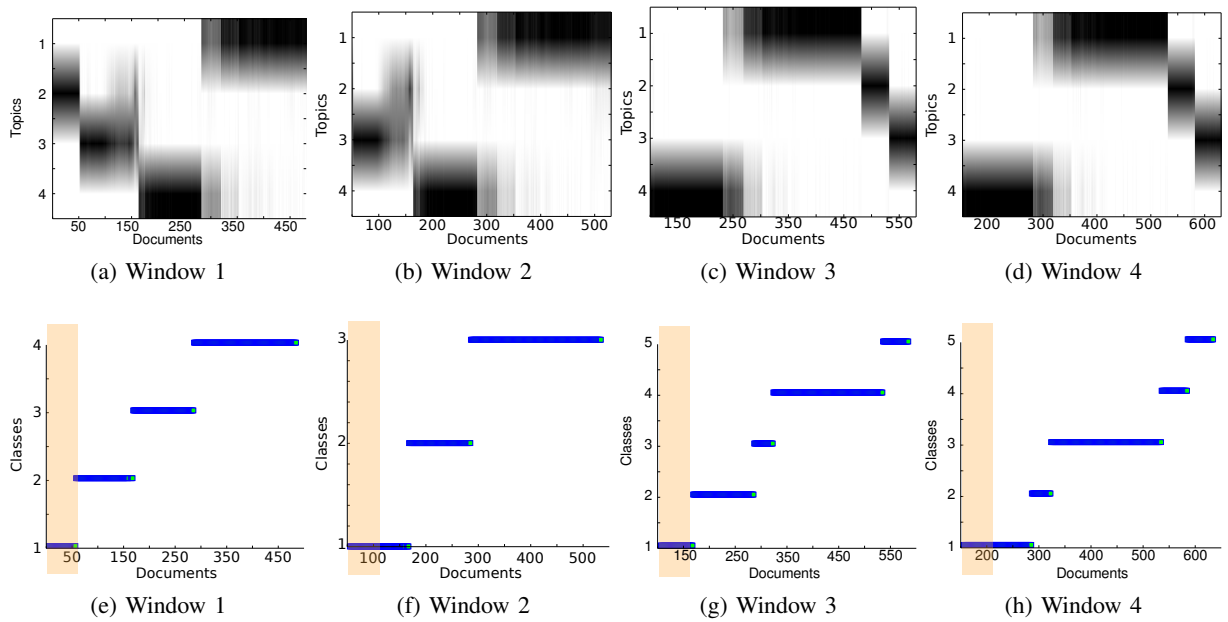
The three critical operations all update the value of S_B and S_W in an incremental manner.

Stability of the vocabulary is ensured as the elementary clusters from the feature tracker are maintained as leaves of the tree of the visual word of which they are a part. Indexing is done in a top-down fashion, where all visual words near the feature are visited according to a threshold criteria. Updates to the vocabulary are triggered automatically by monitoring the feature-cluster association. A feature *fails* to match any of the clusters when it does not fall within three standard deviations of the closest cluster in any dimension. When the success rate of feature matching falls below 0.9 the vocabulary is updated. A transformation matrix ${}^{p}\Gamma_{p-1}$ embodies the change to the vocabulary between update p and $p-1$, and enables the similarity of images indexed at different update stages to be computed.

B. Incremental Topic Modeling

Topic models introduce a latent topic level between words and documents. Much of the topic modeling literature (such as Probabilistic Latent Semantic Indexing (pLSI) [13] and Latent Dirichlet Allocation (LDA)[14] deals with retrospective analysis, that is, all the documents have been seen, the vocabulary is known and a fixed topic model is learned once. By contrast, *incremental* methods operate online, allowing new documents to arrive, the vocabulary to expand, and perform topic updating in a way that preserves the latent semantic indices of documents between updates.

For this reason Incremental probabilistic Latent Semantic Indexing (IpLSI) [3], a variant of pLSI, is used as the topic modeling engine. The latent semantic variables are the topics, z which are placed between documents d and words w. The original pLSI model is generative, and assumes that the distribution of a word and document is conditionally independent. The joint probability of the co-occurrence pair



(a) Window 1 (b) Window 2 (c) Window 3 (d) Window 4

(e) Window 1 (f) Window 2 (g) Window 3 (h) Window 4

Fig. 2: Evolution of topics and classes through 4 sliding window frames. The shaded orange region located between frames 2 and 52 of each sliding window is the 'selection' region. Any class that currently has its first frame in this region is chosen to add an exemplar to the selected key-frames. The midpoint frame of the class is chosen as the representative image.

(w, d) can be written as follows:

$$P(w,d) = \sum_{z \in Z} P(z)P(w|z)P(d|z) \tag{6}$$

$$= P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{7}$$

$$= P(w) \sum_{z \in Z} P(d|z)P(z|w) \tag{8}$$

The parameters $p(w|z)$ and $p(z|d)$ of the pLSI algorithm are estimated via Expectation Maximization (EM) using a training document set D which provides a co-occurrence matrix of words w over documents d with values $f(w, d)$.

Under the vanilla pLSI algorithm, the topic assignments for a new document q can now be estimated using EM while holding $p(w|z)$ fixed, which is usually accomplished in a few iterations. The model is unchanged despite the arrival of the new document q and the additional words w_{new} it may bring. IpLSI, by contrast updates the model using a sliding window approach and relies also on the symmetrical form (8) of (7). The estimation step is:

$$P(z|w,d) = \frac{P(d|z)P(z|w)}{\sum_{z' \in Z} P(d|z')P(z'|w)} \tag{9}$$

and maximization step is:

$$P(d|z) = \frac{\sum_{w \in d} f(w,d)P(z|w,d)}{\sum_{d' \in N} \sum_{w \in d'} f(w,d')P(z|w,d')} \tag{10}$$

$$P(z|w) = \frac{\sum_{d \in N} f(w,d)P(z|w,d)}{\sum_{d \in N} f(w,d)} \tag{11}$$

The model is updated in four steps:

1) **Out-of-date documents and words are removed.** The existing pLSI parameters are updated and normalized, and become $p(w_{old}|z)$ and p$(d_{old}|z)$.
2) **New documents are folded-in** EM is used to estimate $p(z|d_{new})$ using the normal form (not shown for brevity) of (9) and (11) with $p(w_{old}|z)$ held fixed.
3) **New words are added.** Using the symmetrical form of (10), $p(d_{new}|z)$ is calculated using the results of the previous step. It is held fixed, and then (9) and (11) are used to obtain $p(z|w_{new})$.
4) **The pLSI parameters are updated.** Firstly $p(z|w, d)$ is obtained from $p(z|w_{old}, d_{old})$, $p(z|w_{old}, d_{new})$ and $p(z|w_{new}, d_{new})$ which are calculated in Steps (1) - (3). $p(z|w_{new}|d_{old})$ is zero — new words have zero probability of being generated by the old document set. This allows us to use the normal form of (10) to compute $p(w|z)$.

While Steps (2) and (3) require additional EM loops, these are usually quick to converge and take typically 5-15 iterations. The resulting set of parameters serves as incremental training for the next iteration and $p(z|d)$ is used to cluster images in topic space.

C. Clustering and Place Discovery

Figure 2 shows that the topic-document distribution $p(z|d)$ provides a baseline classification of the image stream into classes (topological places). However, it is too fragmented a

classifier for our purposed because it would have indeterminate results in areas where images have non-zero likelihood for more than one topic, as is evident in many frames shown in Figures 2a-2d. Instead, we take the K dimensional topic assignment $p(z|d)$ to be a descriptor of the image and use it to cluster the documents into related classes.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [15] is used for this task. It mandates a minimum density be maintained throughout all members of the class, where density is measured by the number of member points of the same class that fall within a specified Euclidean radius *Eps* of a given point. DBSCAN requires only one additional parameter, *MinPts*, that specifies the minimum number of points required to form a cluster. While simplistic, DBSCAN works well for our purposes. Its runtime complexity is $O(n \log n)$, acceptable when dealing with relatively small sliding windows, and it doesn't require us to know the number of classes ahead of time which is a drawback of alternative techniques such as k-medoids.

DBSCAN clustering identifies temporally smoothed places. The final step in our approach is to take the block of images that DBSCAN has identified as a contiguous place and choose a representative image. Figure 2e-2h illustrates how the classes evolve over a sliding window of 200 frames. While there is some noise (see the appearance of a small, new class not present in Figure 2f in Figure 2g), the majority of classes stay largely intact and slide with the window. In order to obtain a single representative example of each persistent class we define a 'selection candidate' region, shown in orange in Figures 2e - 2h. We calculate the midpoint of classes whose first instance (frame) falls within this region and add the midpoint frame of the class to the collection of selected key-frames. The region is offset by 1 from the window start to avoid selecting duplicate representations of classes that have partially exited the sliding window and may have already contributed an exemplar image to the selective map.

IV. Experimental Results

We test the performance of selective visual mapping against two objectives: (i) choosing *distinct topological places* from the robot's image stream (spatial performance), and (ii) choosing *exemplar representations* of individual places from images gathered each time the robot revisits the same place (temporal performance).

The entire process was run on a MacBook Pro with 16GB of RAM and a 2.2GHz Intel Core i7 processor. OVV is implemented in C++ under the Robot Operating System (ROS) environment. IpLSI and DBSCAN are implemented in MATLAB. For IpLSI we used a sliding window of 500 frames, folding in 50 new frames at a time.

The results are evaluated in terms of three criteria: the *sparsity* of the selective map, *computation time* and *accuracy*. In the spatial case we evaluate accuracy by comparing the performance of the key-place map against the full map using OVV. In the absence of an analogous metric, we can only evaluate the temporal performance qualitatively, judging whether the resulting selection of exemplar views provides an accurate depiction of the different appearance of a place over time.

	Original Frames	Selective Map	Window Frames
Level 7	7030	51	131
New College	3300	35	56

TABLE I: Selective Mapping: Sparsity

A. Spatial Performance

To evaluate the spatial performance of the algorithm we tested it on an indoor and an outdoor environment. The indoor data comprises 7030 frames from 3 loops of a cluttered office environment over which our OVV implementation created 7532 words. The outdoor experiment uses a subset of the New College dataset [16] with 3300 frames and a learned vocabulary of 2268 words.

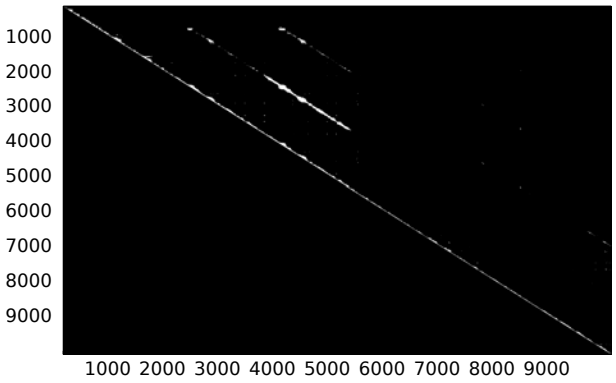
1) Sparsity: Figures 4 and 5 show the images selected from one loop of each of the two test environments. In both cases the spacing of the selected nodes is irregular, it is evident that the algorithm is selecting places based on something other than time or distance traveled. Table I shows that selective mapping chooses to retain only 0.72% of the indoor image stream, and 1.06% of the outdoor image stream. The number of sliding windows of IpLSI is included for comparison, and the ratio between selected images and the number of windows, at 0.39 for indoors and 0.62 for outdoors, illustrates that place identification is independent of the number of window updates.

2) Accuracy: Figure 3a is the image similarity confusion matrix and Figure 3b the precision-recall (PR) curve which details the place recognition performance of OVV on the Indoor dataset. Figure 3b shows that for precision below 0.85, using OVV with selective key-places performs better than OVV over the entire dataset. The shape of the curve at high precision can be explained by the lack of data points in this area for the selective key-places case. Figures 3c and 3d show place recognition performance on the New College park loop. Here selective key-places performs better than using the entire dataset across the board.

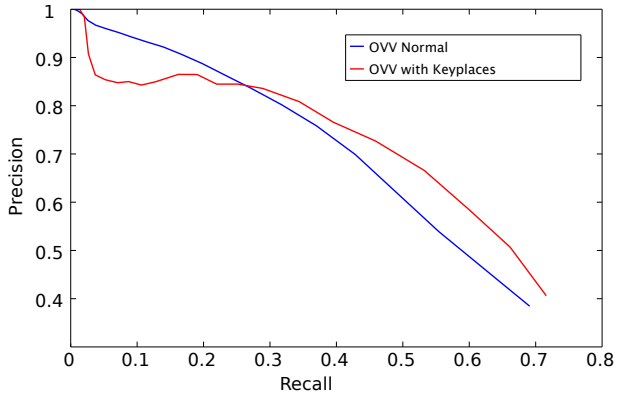
3) Computation Time: Topological place discovery is a multi-stage process and can be viewed as having a time-critical part (dynamic vocabulary creation and indexing) and the topic modeling and clustering stage is a background map maintenance procedure. Frame indexing using OVV averages 0.12 seconds per frame. Updates to the current vocabulary depend on its size, but take at most a few seconds. Table II shows the time required for the IpLSI/DBSCAN MATLAB process to converge over the initial 500 frame window, and to process subsequent 50 frame sliding window updates.

B. Temporal Performance

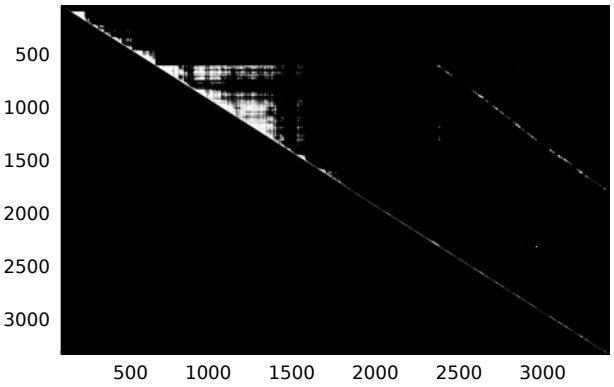
To test the algorithm's ability to select exemplar images of a single place we gathered data from a public camera. We gathered data at 10 minute intervals over a 1 week period and ran OVV. Figure 6 shows the results of applying



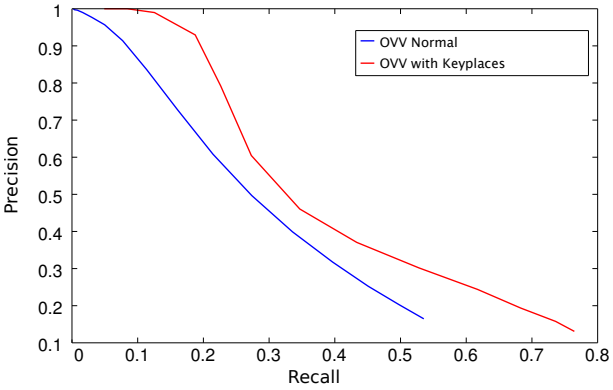
(a) Indoor Similarity Matrix, threshold at 0.35. Off-diagonals show that the robot does 3 loops.



(b) Indoor Precision-Recall Curve



(c) Outdoor Similarity Matrix, threshold at 0.6. The off-diagonal denotes the second loop.



(d) Outdoor Precision-Recall Curve

Fig. 3: Spatial selection results, indoor and outdoor

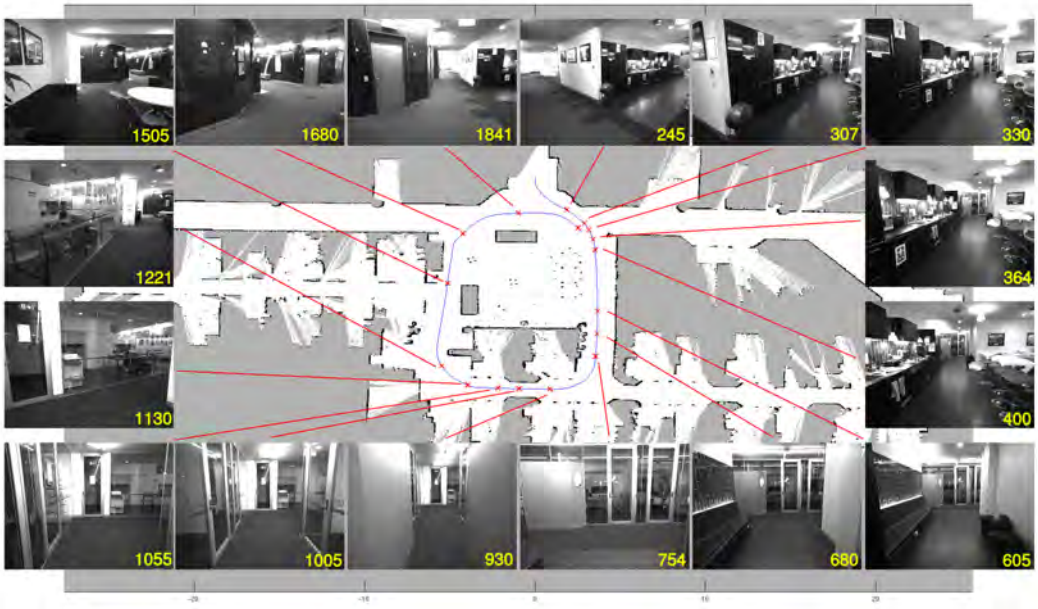


Fig. 4: One loop of level 7 (7030 images) is reduced to these 16 key-places by topics modeling. Places discovered by the topological place discovery engine appear to correspond to the kitchen, hallway, office and meeting room.

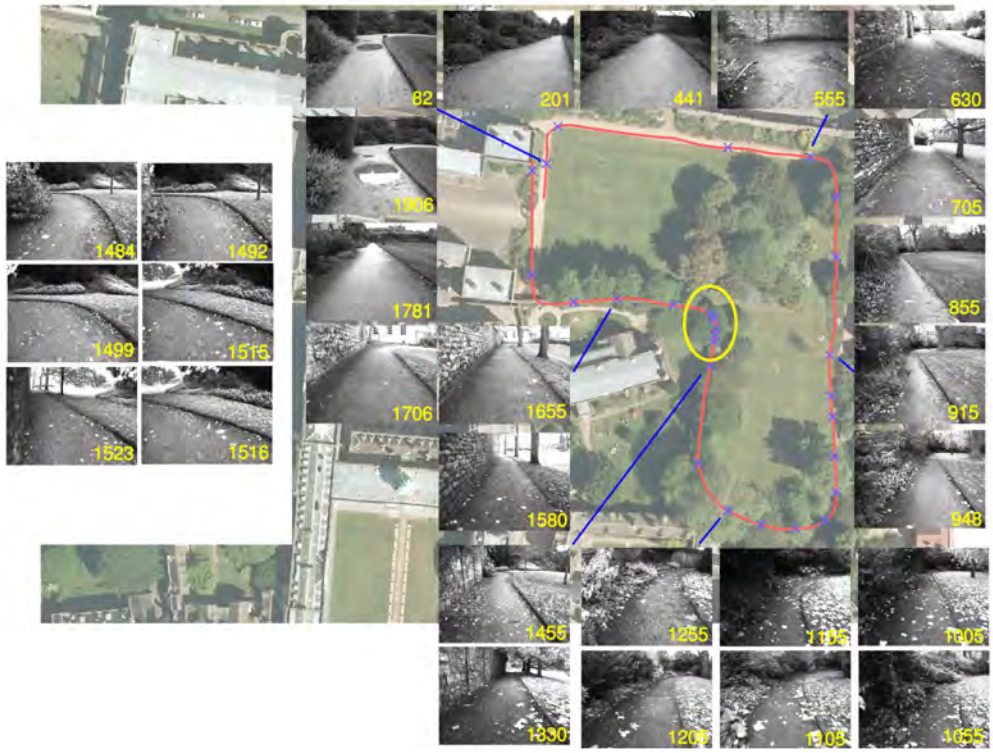


Fig. 5: The entire set of images selected from one loop of the park at New College. The inset shows the 6 images from the circled region. Pitching of the robot at this point causes rapid scene change with minimal distance traveled, hence we see a cluster of discovered places at this point.

	Initial(s)	Window Update(s)
Level 7	6.74	11.65
New College	8.58	11.22

TABLE II: Selective Mapping: Computation Times for Initial 500 frame window and subsequent 50 frame fold-ins.

the sequential algorithm (as above) to one week's worth of images at a single location. Of the 1008 images, selective key-places leaves us with a set of 24 exemplars which show an even mixture of day and night scenes and embody various levels of shadow and lighting conditions.

V. Discussion and Conclusion

This work introduces a novel technique synthesizing dynamic visual vocabularies and incremental topic modeling, to continually discover the presence of topological places in an image stream seen by a robot. Parts of the stream that exhibit continuity in topic space are inferred to have originated from the same topological place. On the other hand, discontinuities in topic space indicate the robot is transitioning to a new topological place. The method requires no training or a-priori knowledge of the environment to which it is being applied. A series of experiments demonstrates that unsupervised topological place discovery to be highly effective at reducing the density of topological maps. Results show that in both indoor and outdoor environments similar precision and accuracy performance is obtained with this skeletal set of key-places as performing place recognition with the same algorithm using the full image stream. Further, the approach is computationally tractable for the purposes of map maintenance. Topological place discovery in the temporal dimension is also demonstrated. This produced succinct exemplar descriptions of appearances of the same place at different times. By modeling image-sequences as a multi-modal distribution of topics – and by discovering topics incrementally and online – it is possible to both reduce the memory footprint of place recognition systems, and to improve precision and recall.

References

[1] T. Nicosevici and R. Garcia, "Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping," *Robotics, IEEE Transactions on*, vol. 28, no. 4, pp. 886–898, aug. 2012.

[2] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "Rslam: A system for large-scale mapping in constant-time using stereo," *International Journal of Computer Vision*, vol. 94, no. 2, pp. 198–214, 2011.

[3] T.-C. Chou and M. Chen, "Using Incremental PLSI for Threshold-Resilient Online Event Analysis," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 3, pp. 289–299, march 2008.

[4] H. Kretzschmar and C. Stachniss, "Information-theoretic compression of pose graphs for laser-based SLAM," *The International Journal of Robotics Research*, vol. 31, no. 11, pp. 1219–1230, 2012.

[5] V. Ila, J. Porta, and J. Andrade-Cetto, "Information-Based Compact Pose SLAM," *Robotics, IEEE Transactions on*, vol. 26, no. 1, pp. 78–93, feb. 2010.

[6] K. Konolige and J. Bowman, "Towards lifelong visual maps," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, oct. 2009, pp. 1156–1163.

Fig. 6: Applying the sequential culling technique to 1 week's worth of camera data taken at 10 minute intervals (1008 images) reduces the set down to these 24 images

.

[7] A. Ranganathan and F. Dellaert, "Bayesian surprise and landmark detection," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, may 2009, pp. 2017–2023.

[8] A. Chapoulie, P. Rives, and D. Filliat, "Topological segmentation of indoors/outdoors sequences of spherical views," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, oct. 2012, pp. 4288–4295.

[9] H. Korrapati, J. Courbon, Y. Mezouar, and P. Martinet, "Image Sequence Partitioning for outdoor mapping," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, may 2012, pp. 1650–1655.

[10] R. Paul, D. Rus, and P. Newman, "How was your day? Online visual workspace summaries using incremental clustering in topic space," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, may 2012, pp. 4058–4065.

[11] R. Paul and P. Newman, "Self help: Seeking out perplexing images for ever improving navigation," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, may 2011, pp. 445–451.

[12] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous Adaptive Underwater Exploration using Online Topic Modelling ," in *13th International Symposium on Experimental Robotics (ISER 2012)*, 2012.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.

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

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Second International Conference on Knowledge Discovery and Data Mining*, E. Simoudis, J. Han, and U. Fayyad, Eds. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[16] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The New College vision and laser data set," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 595–599, May 2009.