Visual Place Recognition using HMM Sequence Matching

Peter Hansen and Brett Browning

Abstract— **Visual place recognition and loop closure is critical for the global accuracy of visual Simultaneous Localization and Mapping (SLAM) systems. We present a place recognition algorithm which operates by matching local query image *sequences* to a database of image *sequences*. To match sequences, we calculate a matrix of low-resolution, contrast-enhanced image similarity probability values. The optimal sequence alignment, which can be viewed as a discontinuous path through the matrix, is found using a Hidden Markov Model (HMM) framework reminiscent of Dynamic Time Warping from speech recognition. The state transitions enforce local velocity constraints and the most likely path sequence is recovered efficiently using the Viterbi algorithm. A rank reduction on the similarity probability matrix is used to provide additional robustness in challenging conditions when scoring sequence matches. We evaluate our approach on seven outdoor vision datasets and show improved precision-recall performance against the recently published seqSLAM algorithm.**

I. INTRODUCTION

Visual place recognition is a core component of many visual Simultaneous Localization and Mapping (vSLAM) systems. Correctly identifying previously visited locations enables incremental pose drift to be corrected using any number of graph-based loop closures techniques (e.g. g^2o [1], COP-SLAM [2]). In this work, we focus on the place recognition component – determining whether a place has been visited before – with an eye towards systems that can provide robust performance under variations in lighting and other atmospheric conditions such as rain, fog and dust.

There has been tremendous work on this topic with most approaches phrasing the problem as one of matching a sensed image against a database of previously viewed images (i.e. images and places are synonymous). The FAB-MAP algorithm [3], [4] remains a popular state-of-the-art algorithm achieving robust image recall performance for outdoor image sequences up to 1000 kilometers. FAB-MAP and many of its competitors, build from the visual Bag-of-Words (BoW) approach popularized in image retrieval [5] and earlier in document retrieval. These approaches rely on extracting scale-invariant image keypoints and descriptors, such as SIFT [6] and SURF [7], from an image. The descriptor vectors are quantized using a dictionary trained on prior data. FAB-MAP uses a probabilistic model of co-occurrences of visual word appearance, while the more traditional BoW approaches use the vector space model approach. The success of BoW approaches however is very dependent on the quality of the visual vocabulary, and in turn the prior data, and on the reliability of extracting the same visual keypoints and descriptors in images with similar viewpoints. The latter is particularly problematic when there are large lighting variations and scene appearance changes such as due to fog.

Recently [8], [9] presented sequence SLAM (seqSLAM) that achieves significant performance improvements over FAB-MAP under extreme lighting and atmospheric variations [9]. Image similarity is evaluated using the sum of absolute differences between contrast enhanced, low-resolution images without the need for image keypoint extraction. For a given query image, the matrix of image similarities between the local query image *sequence* and a database image sequence is constructed. The image recall score is the maximum sum of normalized similarity scores over pre-defined constant velocity paths (i.e. alignments between the query sequence and database sequence images) through the matrix, a process referred to as *continuous* Dynamic Time Warping (DTW). The contrast enhancement and matrix normalization steps are described in section II and are the keys steps for achieving robust performance under extreme lighting or atmospheric changes. The underlying assumption of the continuous DTW is that the vehicle traverses a previously visited path in the environment at a constant multiple of its previous velocity. Recently [10] extended the approach to use odometry to constrain the distance between database and query images to overcome this restriction. The approach taken in this work is to improve the overall flexibility of the sequence alignment procedure, and not rely on odometry which may be unavailable or inaccurate.

DTW is reminiscent of approaches to handle variable speaker speed in speech recognition [11]. There, alignment is solved by efficiently finding the minimum cost path through a similarity matrix using dynamic programming. Here, we propose a similar approach. By phrasing the sequence matching problem as a Hidden Markov Model we can use the Viterbi algorithm [12] to efficiently and optimally align the sequences. We obtain much greater flexibility in state transition models that allow for different velocity variation models, including discontinuous jumps, and produce a meaningful likelihood estimate as an output. The result is, when compared to the continuous DTW used by seqSLAM, that a much larger space of possible paths can be considered without increasing computational load. When coupled with the sequence scoring procedure presented, improved place recognition performance is demonstrated.

In section II, we provide an overview of sequence alignment and the seqSLAM algorithm before presenting our

This publication was made possible by NPRP grant #09-980-2-380 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Hansen is with the QRI8 lab, Carnegie Mellon University, Doha, Qatar phansen@qatar.cmu.edu. Browning is with the Robotics Institute/NREC, Carnegie Mellon University, Pittsburgh PA, USA, brettb@cs.cmu.edu

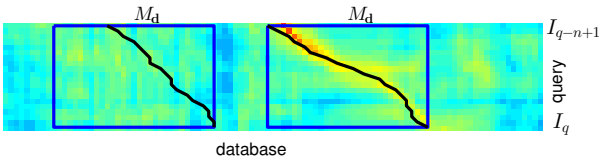


Fig. 1. An image similarity matrix M computed for a query image sequence $I_{\mathbf{q}}$ and all database images. To compute a place recognition score for a database image I_d, the local matrix $M_{\mathbf{d}}$ is selected spanning the previous m database images. A path through the local matrix $M_{\mathbf{d}}$ is found, which aligns the query and database sequences, and the place recognition score computed.

HMM sequence alignment approach (section III) and place recognition scoring procedure (section IV). We compare the performance of the algorithm against seqSLAM on a range of challenging publicly available datasets as detailed in sections V and VI. Finally, we conclude the paper in section VII.

II. BACKGROUND AND RELATED WORK

A. Overview

The goal for a visual place recognition system is to identify a valid database image I_d corresponding to a current query image I_q. The database and query images may belong to different datasets, or may be from the same dataset. For the latter, the database images would be those viewed before the current query image.

We use the same generalized sequence matching fundamentals as seqSLAM as illustrated in Fig. 1. To evaluate a place recognition score between a query image I_q and a database image I_d, a matrix $M_{\mathbf{d}}$ of similarity values between the *sequences* of images $I_{\mathbf{q}} = \{I_{q-n+1}, \cdots, I_q\}$ and $I_{\mathbf{d}} = \{I_{d-m+1}, \cdots, I_d\}$ is computed. A path through the matrix $M_{\mathbf{d}}$ is found maximizing some function of the values along the path (e.g. sum of values). This path defines an alignment/matching between images in the query sequence $I_{\mathbf{q}}$ to images in the database sequence $I_{\mathbf{d}}$. The place recognition score is based on the aligned sequence similarity, and not simply the global one-to-one similarity between the individual images I_q and I_d. Using this sequence matching approach significantly improves the reliability of place recognition.

B. Sequence SLAM

As we compare our algorithm empirically to seqSLAM we first provide a brief overview of seqSLAM.

All images are first converted to low-resolution greyscale images and contrast enhanced. Contrast enhancement is performed by dividing a low-resolution image into multiple cells, each containing $W \times W$ pixels, and normalizing the pixel intensities in each cell to have zero mean and unit standard deviation. The contrast enhanced values in each cell are standard deviations (i.e. z scores) from the mean.

For a query image sequence $I_{\mathbf{q}}$, and all database images, an initial matrix D of image similarity values is constructed. The similarity values are the sum of absolute differences of

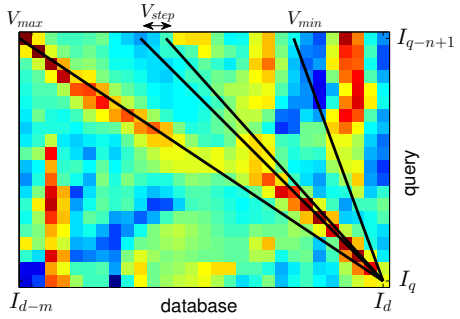


Fig. 2. The continuous DTW of seqSLAM uses a set of pre-defined constant velocity search lines within the limits V_{min} to V_{max} at step sizes V_{step}. These lines are the set of potential paths through the matrix $M_{\mathbf{d}}$.

the contrast enhanced images. The final similarity matrix M is obtained by applying normalization to each value in D:

$$M(q,d) = \frac{D(q,d) - \bar{D}_w}{\sigma_w}, \tag{1}$$

where $\bar{D}_w$ and σ_w are the mean and standard deviations of the values $M(q, d - w/2, \cdots, d + w/2)$ within a fixed sized window width $w = 10$. This normalization provides a 'local best fit' metric within the neighborhood of w database images – see [8] for a more detailed discussion.

For each database image I_d, a *continuous* DTW is used to select the path through the matrix $M_{\mathbf{d}}$, as illustrated in Fig. 2. The continuous DTW uses a set of predefined discretized constant velocity paths (i.e. straight lines) through the matrix between the limits V_{min} and V_{max} at step sizes of V_{step}. The sum of similarity values for each path is computed, and the maximum selected as the place recognition score[1]. Note that this continuous DTW should not be confused with the classical DTW algorithms used extensively for speech recognition [11].

To improve upon seqSLAM, we propose to improve the place recognition score calculation and to exploit a probabilistic framework for more flexible sequence alignment.

III. SEQUENCE ALIGNMENT USING A HIDDEN MARKOV MODEL

The process of finding a path through a similarity matrix $M_{\mathbf{d}}$ is modeled using a Hidden Markov Model (HMM), and the most probable path found using the Viterbi algorithm. Referring to Fig. 3, the HMM is parameterized as follows.

The observations $\mathcal{Z}_{1:n} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_n\}$ are the sequence of n query images $I_{\mathbf{q}}$, and the state space S the set of database images $I_{\mathbf{d}}$. For each observation there is an unobserved hidden variable $\mathcal{X}$ corresponding to one of the database images in the state space. The optimal state sequence $\mathcal{X}^*$ is the one maximizing the conditional probability over all possible state sequences $\mathcal{X}_{1:n} = \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$,

$$\mathcal{X}^* = \underset{\mathcal{X}_{1:n}}{\operatorname{argmax}}\, p(\mathcal{X}_{1:n} | \mathcal{Z}_{1:n}) \tag{2}$$

$$= \underset{\mathcal{X}_{1:n}}{\operatorname{argmax}}\, p(\mathcal{X}_{1:n}, \mathcal{Z}_{1:n}). \tag{3}$$

[1]For seqSLAM, negative matrix values correspond to increased similarity.

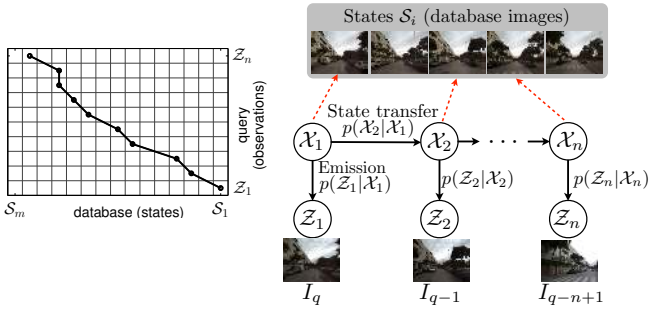


Fig. 3. The Hidden Markov Model. The left shows the observations (query images) and states (database images) represented with respect to the similarity matrix $M_{\mathbf{d}}$ – the path connects the selected state for each observation. The right shows the trellis diagram with emission probabilities and state transfer probabilities labeled. For each observation $\mathcal{Z}$ there is a hidden variable $\mathcal{X}$ corresponding to a database image in the state space.

Here we have made the constraint that the path length n is fixed. In this scenario, the Viterbi algorithm can be used to efficiently find $\mathcal{X}^*$ using dynamic programming. Defining μ_k^i as the largest joint probability of all the path combinations $\mathcal{X}_{1:k}$ ending at state $\mathcal{X}_k = i$, the Viterbi algorithms uses the recursion:

$$\mu_1^i = \underbrace{p(\mathcal{X}_1 = i)}_{\text{initial}}\, p(\mathcal{Z}_1|\mathcal{X}_1 = i) \tag{4}$$

$$\mu_t^i = \underbrace{p(\mathcal{Z}_t|\mathcal{X}_t = i)}_{\text{emission}} \max_{j \in S} \left[\underbrace{p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j)}_{\text{state transfer}}\, \mu_{t-1}^j \right]. \tag{5}$$

The maximum probability for a sequence length n is then

$$\mu(\mathcal{X}^*) = \max_{\mathcal{X}_n}(\mu_n(\mathcal{X}_n)). \tag{6}$$

The above equations find the value of the maximum conditional probability over all state sequences. Storing the argmax at each iteration and backtracking is used to recover the optimal state sequence $\mathcal{X}^*$ (path through matrix).

In the remainder of this section we describe the selection of the emission values, initial state probabilities, and the state transfer probabilities.

1) Emission Matrix: Low-resolution contrast enhanced images are created using the same procedure as seqSLAM described in section II. Recalling that the contrast enhanced values are z scores, we use the below function to compute the similarity matrix $M_{\mathbf{d}}$ with values in the range 0 to 1:

$$M_{\mathbf{d}}(t,i) = \frac{1}{N_r\, N_c} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \text{abs}\left(\Phi(I_{d(i)}(r,c)) - \Phi(I_{q(t)}(r,c))\right), \tag{7}$$

where Φ is the cumulative distribution function of the standard normal distribution with unit standard deviation, and N_r and N_c are the number of low-resolution image rows and columns.

The similarity matrix $M_{\mathbf{d}}$ is converted to the stochastic emission matrix E by normalizing the sum of values in each column to 1,

$$E(t,i) = \frac{M_{\mathbf{d}}(t,i)}{\sum_{t=1}^{n} M_{\mathbf{d}}(t,i)}. \tag{8}$$

The emission matrix stores the conditional probability values $E(t,i) = p(\mathcal{Z}_t|\mathcal{X}_t = i)$.

2) Initial State Probabilities: Referring to Fig. 3, the start point of the path in each local similarity matrix $M_{\mathbf{d}}$ is the lower right corner – the observation $\mathcal{Z}_1$ and state $\mathcal{S}_{i=1}$. The initial state probabilities are therefore

$$p(\mathcal{X}_1 = i) = \begin{cases} 1 & i = 1 \\ 0 & i > 1 \end{cases}. \tag{9}$$

3) State Transition/Transfer Probabilities: The state transition probabilities are the likelihoods of transitioning between states (database images) from one observation (query image) to the next. They are stored in the $m \times m$ state transition matrix A, where m is the number of states, and $A(j,i) = p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j)$ with

$$\sum_{i=1}^{m} A(j,i) = \sum_{i=1}^{m} p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j) = 1 \quad \forall j,t. \tag{10}$$

We use *local* velocity constraints to set the state transition matrix values using the function

$$A(j,i) = \begin{cases} 0 & i < j \\ 1 & 0 \le (i-j) \le (V_{max} + 0.5) \\ exp\left(\frac{-(i-j-V_{max})^2}{2\,V_{max}^2}\right) & \text{otherwise} \end{cases} \tag{11}$$

and then normalize to satisfy (10). This function is a truncated Gaussian distribution with a flattened peak. The state transition matrix A for $m = 15$ states computed using the velocity values $V_{min} = 1/1.5$ and $V_{max} = 1.5$ is shown in Fig. 4 for reference. Note again that the state transitions are defined only between successive observations, and therefore enforce only local velocity constraints.

As a secondary step we use a global velocity mask to limit the number of possible paths through the matrix, as illustrated in Fig. 5. It is constructed using the same values V_{min} and V_{max}, with unshaded regions having a value of 1, and the others 0. Any path through the matrix must lie within the bounds of this mask. To achieve this, any state transfer value in (5) is set to zero if the mask value at cell t,j or t,i is zero. This often requires a temporary re-normalization of the state transfer values A to satisfy (10).

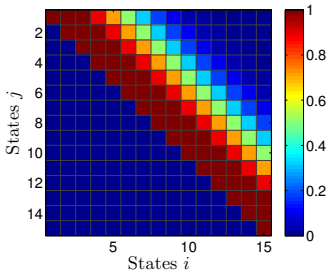
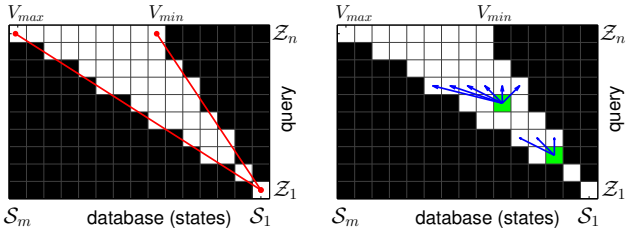


Fig. 4. A sample $m \times m$ state transition probability matrix A computed using (11). For display purposes the values are shown before normalizing the sum of each row to one.



(a) Global velocity mask. (b) Example sets of state transitions.

Fig. 5. The global velocity mask (a) with limits V_{min} and V_{max}. For each iteration of the Viterbi algorithm, the local state transitions are constrained to lie withing the bounds of the global mask – see (b). This restricts the set of possible paths through the matrix to lie within the bounds of the mask.

IV. SEQUENCE SCORE

The HMM sequence alignment procedure in section III finds a path $\mathcal{X}^*$ through the image similarity matrix $M_{\mathbf{d}}$. A place recognition score must be evaluated using this path.

The simplest metric to use is the path probability $\mu(\mathcal{X}^*)$. However, this is particularly unreliable when the query or database images contain visual aliasing or limited appearance changes. In this scenario the probability values in the matrix $M_{\mathbf{d}}$ may all be large, but are ambiguous having no discernible 'best' path. This is illustrated in Fig. 6(a). It shows the path selected in two separate similarity matrices $M_{\mathbf{d}}$. The left matrix is an incorrect place recognition result, but has a higher probability $\mu(\mathcal{X}^*)$ than the correct result on the right.

As discussed in section II, sequence SLAM uses a local normalization of the similarity matrix values M to find a local best match. An alternate approach used in [13] was to compute the eigen-decomposition of a square similarity matrix M, and reconstruct the rank-reduced matrix omitting the first r eigenvalues[2]. They argue that the first r rank one matrices having large eigenvalues are dominant themes in the similarity matrix arising from visual aliasing.

We use a similar spectral-decomposition to [13], but adapted to operate on the matrix $M_{\mathbf{d}}$. The Singular Value Decomposition (SVD) of $M_{\mathbf{d}}$ is found,

$$U\,\Sigma\,V^T = SVD(M_{\mathbf{d}}), \tag{12}$$

and the rank-reduced similarity matrix $\tilde{M}_{\mathbf{d}}$ computed,

$$\tilde{M}_{\mathbf{d}} = U\,\tilde{\Sigma}\,V^T, \tag{13}$$

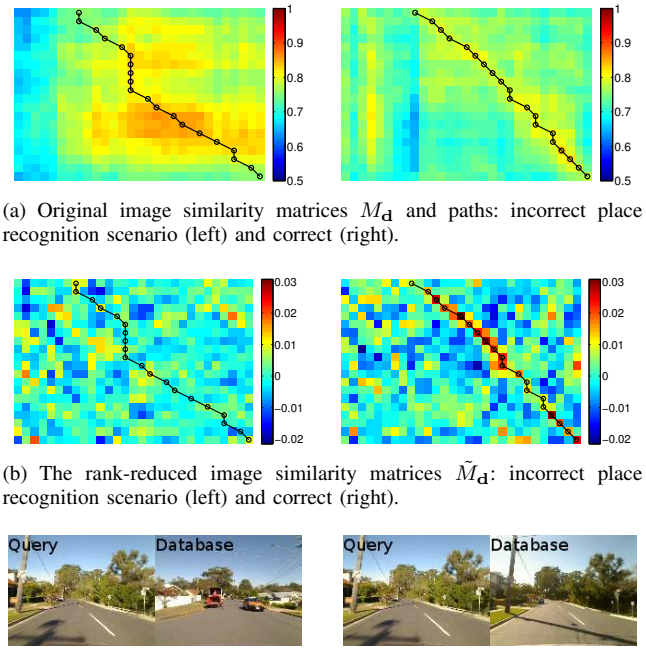
where $\tilde{\Sigma}$ is the $n \times n$ matrix Σ with the first r singular values set to zero. In later experiments we use a heuristic value of $r = 4$. The place recognition score for the query image I_q and database image I_d is selected as the Gaussian weighted sum of rank-reduced similarity values along the path $\mathcal{X}^*$:

$$score = \sum_{i=1}^{n} G(i)\,\tilde{M}_{\mathbf{d}}(\mathcal{X}_i^*), \tag{14}$$

where

$$G(i) = \exp\left(\frac{-(i-1)^2}{2n^2}\right). \tag{15}$$

[2]Applied to a full square similarity matrix M where the query and database images sets are the same.



(a) Original image similarity matrices $M_{\mathbf{d}}$ and paths: incorrect place recognition scenario (left) and correct (right).

(b) The rank-reduced image similarity matrices $\tilde{M}_{\mathbf{d}}$: incorrect place recognition scenario (left) and correct (right).

(c) Query and database images: incorrect place recognition scenario (left) and correct (right).

Fig. 6. The rank-reduction used for improved place recognition scoring. The left columns are an incorrect place recognition scenario, and the right columns a correct scenario. (a) shows the paths computed using the original similarity matrices $M_{\mathbf{b}}$, (b) the rank-reduced matrices $\tilde{M}_{\mathbf{b}}$, and (c) the query and database images. The probability $\mu(\mathcal{X}^*)$ is larger for the incorrect result. The new matching score using the sum of rank-reduced values is larger for the correct result.

Fig. 6(b) shows the rank reduction applied to the original similarity matrices in 6(a). The score in (14) computed using the method described is larger for the true positive result. For a query image sequence $I_{\mathbf{q}}$, the sequence alignment and place recognition score in (14) is computed for all database sequences. The database sequence having the largest score is selected as the match for $I_{\mathbf{q}}$.

V. EXPERIMENTS

Place recognition results using seqSLAM and our HMM-Viterbi algorithm were found for a range of outdoor vision datasets summarized in table I. The datasets include 3 sequences from the KITTI visual odometry training set[3], one sequence from the Málaga urban dataset [14][4], sequences from the St. Lucia multiple times of day dataset [15][5], and sequences collected near our campuses in Pittsburgh and Qatar. Referring to table I, the database and query for St. Lucia and Qatar are different image sets collected at different times. For all other datasets, the database and query are the same image sets. The results for seqSLAM were found using our Matlab implementation of the algorithm.

The same low-resolution image sizes were used by both algorithms with a contrast-enhanced patch size of 8×8 pixels

[3]http://www.cvlibs.net/datasets/kitti/eval_odometry.php
[4]http://www.mrpt.org/MalagaUrbanDataset
[5]https://wiki.qut.edu.au/display/cyphy/St+Lucia+Multiple+Times+of+Day

TABLE I

SUMMARY OF THE DATASETS USED IN THE EXPERIMENTS. THE NOTATIONS 'D' AND 'Q' REFER TO THE DATABASE SEQUENCE AND QUERY SEQUENCE, RESPECTIVELY. FOR THE NUMBER OF FRAMES, THE VALUES IN THE BRACKETS WERE THE ORIGINAL NUMBERS, AND THE VALUES WITHOUT BRACKETS THE NUMBER USED BY SELECTING EVERY k^{th} IMAGE. FOR THE RESOLUTION, THE VALUES IN BRACKETS WERE THE ORIGINAL VALUES, AND THE VALUES WITHOUT BRACKETS THE LOW-RESOLUTION SIZES USED.

Dataset	Database/Query	Length	#frames	Resolution
KITTI 00	D/Q: seq 00	3.72km	2271 [4541]	80×24 [1241×376]
KITTI 02	D/Q: seq 02	5.07km	2331 [4661]	80×24 [1241×376]
KITTI 05	D/Q: seq 05	2.21km	1381 [2761]	80×24 [1226×370]
Málaga	D/Q: extract 10	5.81km	2164 [17310]	32×24 [800×600]
St. Lucia	D: 100909-0845	18.4km	5284 [21135]	32×24 [800×600]
	Q: 110909-1545	18.5km	5032 [20127]	32×24 [800×600]
Qatar	D/Q: seq 1	9.61km	3964 [23783]	40×24 [1200×675]
Pittsburgh	D: seq 1	8.09km	3374 [41112]	32×24 [640×480]
	Q: seq 2	6.66km	2725 [35156]	32×24 [640×480]

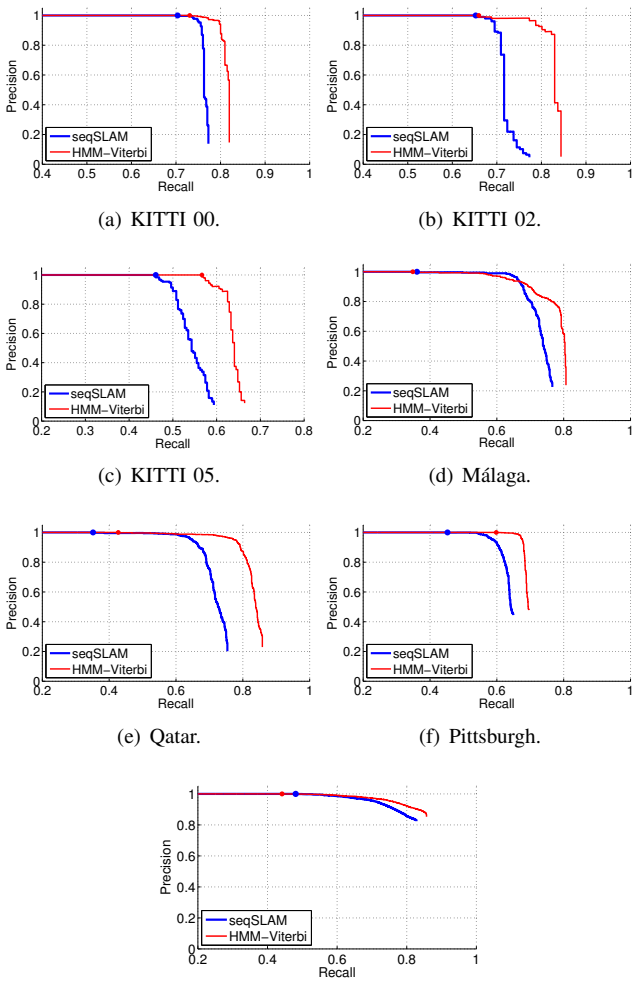
– for stereo datasets, only left images were used. A sequence length of $n = 20$ images was selected for all datasets, and velocity limits of $V_{max} = 1.5$ and $V_{min} = 1/V_{max} = 0.67$. For seqSLAM we use the parameters reported in [8] and use a step size of $V_{step} = 0.02$ pixels, matrix normalization window size $w = 10$, and the maximum path score (sum of similarity values) over all database sequences to select the match for each query sequence. The HMM parameters and sequence scoring parameters given in sections III and IV were used. For the datasets where the query and database images are the same set of images, a query image I_q can only be matched to prior database images.

We evaluate place recognition performance using precision recall. For any dataset, the total number of possible loop closure events and true positives are identified by thresholding the relative position and orientation between candidate query and database image pairs. This is followed with a manual verification. The ground truth pose estimates for each frame in the KITTI datasets are provided. For the remaining datasets, GPS data is used to interpolate an approximate 2 Degree of Freedom (DoF) pose estimate for each frame.

VI. RESULTS AND DISCUSSION

The seqSLAM and HMM-Viterbi precision recall results for all datasets are shown in Fig. 7, and the recall scores for selected precision values provided in table II. The lines connecting the place recognition results for each dataset at 99% precision using the HMM-Viterbi algorithm are displayed in Fig. 8.

For all datasets the HMM-Viterbi place recognition algorithm achieves an increased maximum recall rate, which is



(a) KITTI 00. (b) KITTI 02.

(c) KITTI 05. (d) Málaga.

(e) Qatar. (f) Pittsburgh.

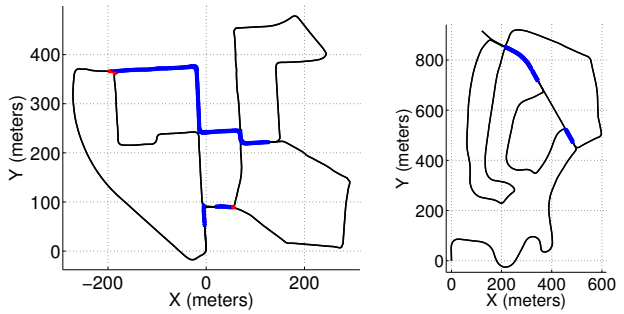
(g) St. Lucia.

Fig. 7. The precision recall results for seqSLAM and HMM-Viterbi. A summary of the datasets is provided in table I. The dots in each figure are located at the largest recall score with a precision value of 1.0.

TABLE II

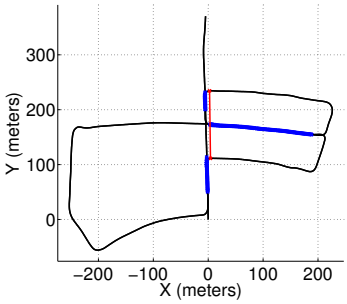
THE RECALL SCORES AT SELECTED PRECISION VALUES. THE FULL RECALL PRECISION CURVES ARE SHOWN IN FIG. 7. FOR EACH DATASET, THE VALUE OF THE BEST PERFORMING ALGORITHM AT THE GIVEN PRECISION LEVEL IS HIGHLIGHTED.

Dataset	Method	Recall		
		1.00 prec	0.99 prec	0.90 prec
KITTI 00	seqSLAM	0.704	0.738	0.758
	HMM-Viterbi	**0.731**	**0.760**	**0.800**
KITTI 02	seqSLAM	0.652	0.652	0.695
	HMM-Viterbi	**0.660**	**0.660**	**0.809**
KITTI 05	seqSLAM	0.461	0.461	0.500
	HMM-Viterbi	**0.566**	**0.566**	**0.613**
Málaga	seqSLAM	**0.361**	**0.568**	0.680
	HMM-Viterbi	0.349	0.548	**0.695**
Qatar	seqSLAM	0.352	0.579	0.670
	HMM-Viterbi	**0.427**	**0.596**	**0.792**
Pittsburgh	seqSLAM	0.452	0.549	0.607
	HMM-Viterbi	**0.598**	**0.660**	**0.681**
St.Lucia	seqSLAM	**0.481**	0.576	0.766
	HMM-Viterbi	0.442	**0.605**	**0.832**

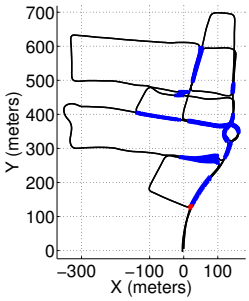


(a) KITTI 00.

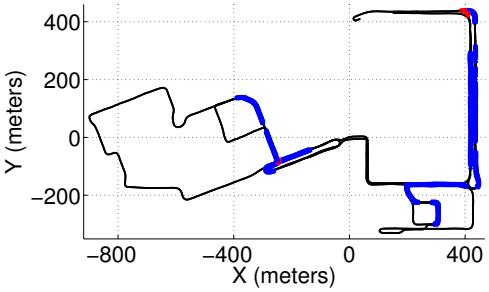
(b) KITTI 02.



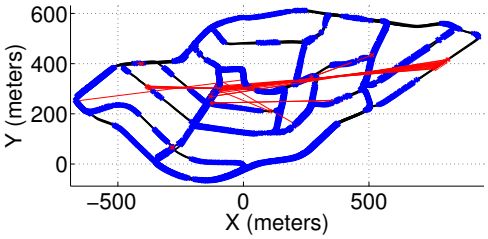
(c) KITTI 05.



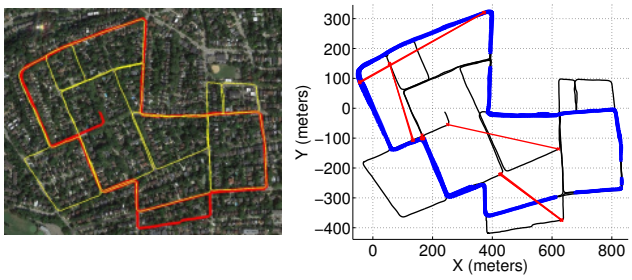
(d) Málaga.



(e) Qatar.



(f) St.Lucia: separate database and query datasets (each traverses same path).



(g) Pittsburgh: left shows dataset (yellow) and query (red) trajectories.

Fig. 8. Place recognition results at 99% precision for each of the datasets using the HMM. The blue lines connect true positive results, and the red lines any false positive results.

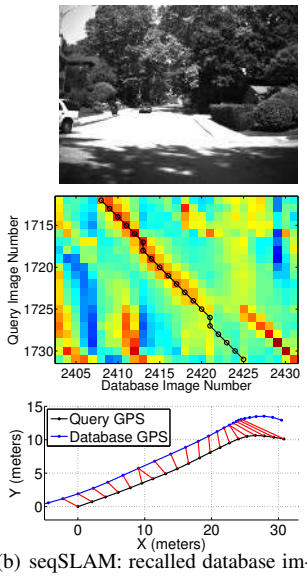
the ratio of the number of true positive place recognition detections to the total number of possible true positives. Moreover, for datasets KITTI 00, KITTI 02, KITTI 03, Qatar and Pittsburgh, the recall rates for HMM-Viterbi is larger than seqSLAM for the high precision values recorded in table II. Although the HMM-Viterbi recall rate at 100% precision is lower than seqSLAM for the St. Lucia dataset, improved recall rates are achieved for precision values of 99% and below. Before discussing the decreased recall rates at high precision values for the Málaga dataset, we first present a typical example where the HMM-Viterbi method outperforms seqSLAM.

Fig. 9 shows a place recognition result from the Pittsburgh dataset using both algorithms. This example highlights the advantage of the HMM path alignment procedure described in section III. For the query image in Fig. 9(a), seqSLAM returned a false positive match (Fig. 9(b)), while HMM-Viterbi correctly identified a true positive database image at 100% precision (Fig. 9(c)). In this scenario there is a non-constant velocity difference between the query sequence and correct database sequence. This is not well approximated by the constant velocity (i.e. linear) search path model employed by seqSLAM which resulted in the inaccurate sequence alignment shown in Fig. 9(b). In contrast, the HMM provides greater flexibility in the state transition models. Referring to Fig. 9(c), this increased flexibility enabled a more optimal alignment between the query sequence and database sequence. This improved sequence alignment and the new scoring metric described in section IV resulted in the true positive place recognition for the query sequence.

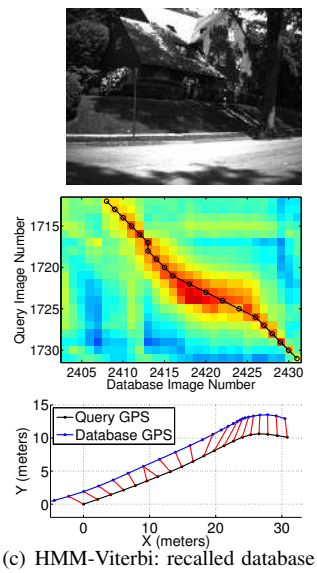
As mentioned, the recall rates for HMM-Viterbi at high precision rates for the Málaga dataset are lower than those for seqSLAM. An example false positive place recognition result using HMM-Viterbi is provided in Fig. 10, showing the query sequence and matched database sequence, the similarity matrix $M_{\mathbf{d}}$, and the rank-reduced similarity matrix $\tilde{M}_{\mathbf{d}}$ used for scoring. The dominant scene appearance between the query and database images are very similar resulting in overall large similarity probabilities in $M_{\mathbf{d}}$. For this example, HMM-Viterbi returned the false positive result at a precision rate of 97%. SeqSLAM also returned a false positive result, but at a much more acceptable precision rate of 86%. We are exploring modifications to the sequence scoring procedure to improve recall rates in scenarios similar to this. This includes automatic and adaptive selection of the number r of singular values set to zero. Additionally, further evaluations of the sequence scoring performance will be tested under more extreme lighting and other climactic variations.

Although performance improvements were demonstrated over the original seqSLAM algorithm, the low-resolution image similarity metric used provides only small viewpoint invariance. For less constrained datasets to those evaluated in which the camera/vehicle returned to the same locations at a similar viewpoint, place recognition performance may degrade. Image descriptors and similarity metrics with improved viewpoint invariance, similar to the image convolution methods in [9], will be explored in future work.

(a) Current query image.

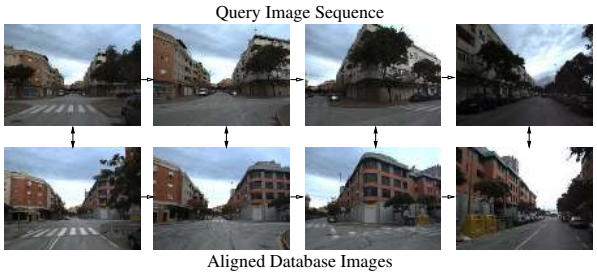


(b) seqSLAM: recalled database image (top), highest scoring path for all database images in similarity matrix M (middle), and aligned sequences (GPS) using the matrix path.

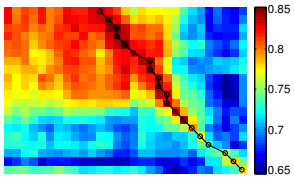


(c) HMM-Viterbi: recalled database image (top), highest scoring path for all database images in similarity matrix M (middle), and aligned sequences (GPS) using the matrix path.

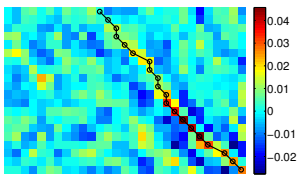
Fig. 9. Database image recall result for the query image in (a) for the Pittsburgh dataset using (b) seqSLAM (false positive) and (c) HMM-Viterbi (true positive at 100% precision). The similarity matrices appear different due to the local normalization of values used by seqSLAM.



(a) Matched sequence alignment (subset shown).



(b) Similarity matrix $M_{\mathbf{d}}$ and path.



(c) Rank-reduced matrix $\tilde{M}_{\mathbf{d}}$.

Fig. 10. A false positive result using HMM-Viterbi for the Málaga dataset: subsets of the images in the matched query/database sequence in (a), original similarity matrix $M_{\mathbf{d}}$ in (b), and rank-reduced matrix $\tilde{M}_{\mathbf{d}}$ used for scoring in (c). The dominant image structures in both sequences appear very similar, resulting in high overall similarity values in $M_{\mathbf{d}}$.

VII. CONCLUSIONS

A visual place recognition system was presented suitable for visual SLAM applications for mobile robots. Local query image sequences are matched to database image sequences by finding a discontinuous path through the matrix of low-resolution contrast enhanced image similarity probabilities. For this, a Hidden Markov Model (HMM) framework is employed with state transition probabilities enforcing local velocity constraints. The Viterbi algorithm is used to compute the most probable path, and a rank reduction of the similarity probability matrix used for matching/place recognition scoring. Experiments using seven outdoor vision datasets were used to compare recall-precision performance against seqSLAM. Overall performance improvements were observed, especially the maximum recall rates. In future work we are exploring alternate sequence scoring techniques and the use of odometry to set the HMM state transitions.

REFERENCES

[1] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g^{2}o: A general framework for graph optimization," in *International Conference on Robotics and Automation*, 2011, pp. 3607–3613.

[2] G. Dubbelman, P. Hansen, B. Browning, and M. B. Dias, "Orientation only loop-closing with closed-form trajectory bending," in *International Conference on Robotics and Automation*, 2012.

[3] M. Cummins and P. Newman, "FAB-MAP: Probabalistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, vol. 27, pp. 647–665, June 2008.

[4] ——, "Appearance-only slam at large scale with FAB-MAP 2.0," *The Int. Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[5] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, October 2003, pp. 1470–1477.

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

[7] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, June 2008.

[8] M. Milford, "Vision-based place recognition: How low can you go?" *International Journal of Robotics Research*, vol. 32, no. 7, pp. 766–789, 2013.

[9] M. Milford and G. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *International Conference on Robotics and Automation (ICRA)*, 2012.

[10] E. Pepperell, P. Corke, and MichaelMilford, "Towards persistent visual navigation using SMART," in *Proceedings of Australasian Conference on Robotics and Automation*, 2013.

[11] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition. 1993," *Prenctice Hall: Englewood Cliffs, NJ*, 2001.

[12] A. Viterbi, "Error bounds for convolution codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260 – 269, 1967.

[13] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 261–286, 2007.

[14] J. Blanco-Claraco, F. Moreno-Dueñas, and J. González-Jiménez, "The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario," *The International Journal of Robotics Research*, vol. (online), pp. 1–8, 2013.

[15] A. Glover, W. Maddern, M. Milford, and G. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day," in *International Conference on Robotics and Automation*, 2010.