Omnidirectional Visual Place Recognition using Rotation Invariant Sequence Matching

[†]Peter Hansen and [‡]Brett Browning

CMU-CS-QTR-126
CMU-RI-TR-15-103

March 22, 2015

[†]Carnegie Mellon University in Qatar
Doha, Qatar

[‡]Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

© Carnegie Mellon University

This publication was made possible by YSREP grant #1-019-2-008 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Abstract

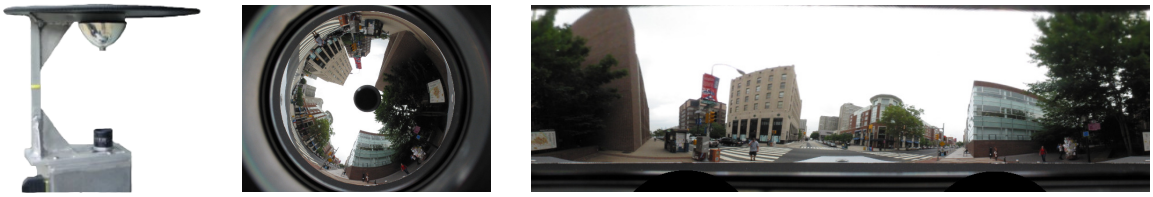
In this work we develop a visual place recognition system for omni-directional panoramic images that makes use of their full 360 degree circumferential angle of view. The system builds on our previous variant of the sequence SLAM algorithm to include panoramic image rotation alignment for improved image similarity scoring, and bi-directional query and database sequence matching within a Hidden Markov Model (HMM) framework for robust place recognition. Three rotation alignment methods are explored including image pixel registration, and both image cross correlation and zero phase alignment in the spherical Fourier domain. All alignment methods operate using low-resolution images for computational efficiency. Experiments using an outdoor panoramic image dataset demonstrate improved precision recall performance using rotation alignment and bi-directional sequence matching. In particular, place recognition is possible in scenarios where a robot traverses a previous path in the opposite direction.

Contents

1 Introduction

Visual place recognition is used within SLAM systems for correctly matching a current query image to a database of previously visited images (i.e. locations). Loop closure algorithms, such as [21, 11], can be implemented to correct for long-range integrated incremental pose estimate errors. In this work, we focus on the development of a place recognition system using using panoramic cameras with the aim of fully exploiting the wide-angle of view to improve visual place recognition performance over more traditional narrow angle of view perspective cameras.

Wide-angle cameras have been used extensively in robots applications, with works on image formation models and calibration [2, 13, 9, 12, 14, 31], relative pose constraints and estimation [28, 24, 27, 26], visual odometry [36], and scale-invariant keypoint detection/descriptors [19, 18, 22] using distortion-invariant image processing techniques [4, 5]. Our interest is visual place recognition for ground vehicle applications using panoramic cameras with a full 360 degree angle of view similar to those in figure 1. Example systems include catadioptric cameras [32, 33] consisting of a camera and reflective surface, or multiple camera arrays whose images are stitched together to produce a single panoramic image.



(a) Omnidirectional catadioptric camera system (left), sample image (middle), and rectified panoramic image (right).

(b) Multiple camera array system (left) and stiched panoramic image (right).

Figure 1: Examples omni-directional camera systems used for outdoor visual SLAM: (a) catadioptric camera and (b) multiple camera array. The panoramic images have a full 360 degree angle of view about the principal axis - the images wrap-around at the left and right edges.

For any camera system selected, place recognition algorithms use image content to derive a place recognition score. Some state of the art systems such as FABMAP [7, 8] use the visual bag-of-words (BoW) framework popularized for image retrieval in [37] and earlier in document retrieval. Detected image feature descriptors (e.g. SIFT [23], SURF [3]) are quantized using a trained corpus of visual words to produce image descriptor that are compared using probabilistic approaches or a vector space model. The success of BoW approaches is highly dependent on the quality of the trained data, and on the ability to repeatably extract the same visual features in images with similar viewpoints under changing lighting conditions. The latter is particularly problematic. An alternate approach is to compute a single 'whole' image descriptor eliminating the

need for training and visual feature extraction. In [1], the authors use an upright SURF descriptor computed over the image and use a discrete Bayesian filtering for robust topometric place recognition. In [29, 30], a low-resolution contrast enhanced image descriptor is used, and local image sequence matching and scoring used to achieve robust place recognition performance even under extreme lighting variations [30].

Regardless of the image descriptor selected, a previously visited place can only be identified if the images being compared contain sufficient scene overlap. This limits the performance of narrow angle of view cameras as small changes in rotational viewpoint between images can significantly decrease image overlap. Panoramic cameras have the distinct advantage that near perfect image overlap can be retained under large viewpoint changes. For the panoramic cameras in figure 1, the same regions of the scene will be imaged under any change in rotation about the axis orthogonal to the ground plane (i.e. typical change in yaw for ground vehicles). The position of the imaged scene points will differ of course.

Any descriptor evaluated for a panoramic image should be invariant to changes in rotation to fully make use of the cameras 360 degree angle of view. Bag of words based descriptors built using extracted visual features are inherently rotation invariant, but as discussed can be unreliable under changes in lighting. We restrict our attention to the more robust low-resolution contrast enhanced descriptor presented as part of sequence SLAM [29, 30]. For this, a one degree of freedom relative change in orientation between image frames is computed, and the descriptors rotated to a common coordinate frame of reference. In section 2 we present a number of computationally efficient methods for computing the one degree of rotation alignment angle between panoramic images that is analogous to a u (horizontal) pixel shift. These include image registration, cross-correlation in the spherical Fourier domain, and a zero-phase canonical image orientation alignment based on [34].

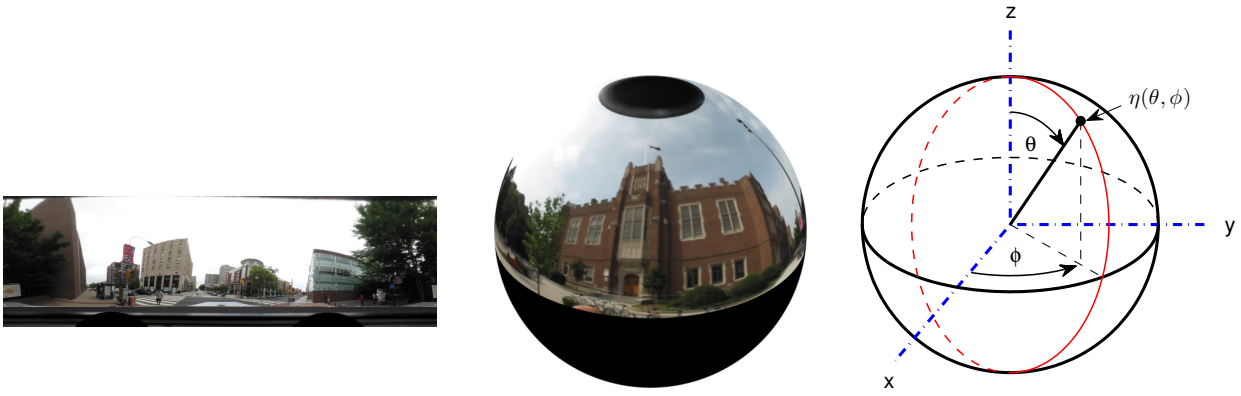
In section 3 we extend our previous visual place recognition system [17], which is a variant of sequence SLAM, to operate using panoramic images with rotation alignment. This includes modification to the Hidden Markov Model (HMM) framework to enable *bi-direction* local sequence matching and scoring. This enable successful place recognition even when a robot traverses a prior path traveling in the opposite direction. Experiments using a hand-labeled outdoor panoramic image database demonstrate improved precision recall performance using rotation alignment and bi-directional sequence matching for place recognition.

2 Rotation Alignment

Any two panoramic images captured at the same position, but different orientations, are related by a change in rotation R. For ground vehicle applications using panoramic camera systems similar to those in figure 1, we assume this rotation is well approximated using a single degree of freedom rotation $R \in SO(1)$ about the cameras principal axis:

$$R = R_z(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}. \quad (1)$$

Referring to figure 2, an image can be rotated by an angle α by applying and horizontal pixel



(a) Original panoramic image. (b) Image mapped to unit view sphere. (c) Spherical image coordinates.

Figure 2: Any central projection panoramic image can be inverse mapped to a function on the unit view sphere centered at the point of projection. The coordinates of on the sphere can be parameterized using an angle of colatitude θ and angle on longitude ϕ. A rotation $R_z(\alpha)$ about the Z axis corresponds to a pixel shift $u(\alpha)$ in the panoramic image.

shift $u(\alpha)$ to the panoramic image,

$$u(\alpha) = \frac{\alpha\, b}{2\pi} N_u, \tag{2}$$

where N_u is the number of image columns. A 'wrap-around' may be required for pixels near the left and right edges of the panoramic images when applying the pixel shift. For convenience in the remaining discussion we denote any image I rotated by an angle α as I^α.

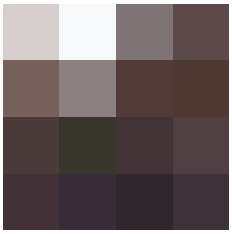
There are two distinct approaches for image alignment allowing a query image I_q and prior database image I_d to be mapped to the same rotational coordinate frame of reference:

- Relative alignment: compute the angle α that aligns the query image to the database image. Images I_q^α and I_d are in the same rotational frame of reference. We explore two methods for relative angle compution including image registration, and image cross-correlation in the spherical Fourier domain.

- Canonical alignment: compute a rotation angle ρ for any image that maps it to a canonical rotational frame of reference. The images I_q^ρ and $I_d^{\rho'}$ are then in the same frame of reference, where $\alpha = \rho - \rho'$. We consider a zero-phase alignment of the spherical harmonic spectrum of the image to compute a canonical angle ρ.

For computational efficiency we use very low resolution $2b \times 2b$ panoramic images for all methods, as illustrated in figure 3. The parameter b is referred to from here on as the image *bandwidth*. In proceeding discussions we use q and d to denote a low resolution $2b \times 2b$ query and database image, respectively. In all examples and experiments, a zero-mean normalization is applied to the low-resolution grayscale images q and b before computing rotation alignment.

(a) Original image: 2048×618 pixels.



(b) $b = 2$: 4×4 pixels.

(c) $b = 4$: 8×8 pixels.

(d) $b = 8$: 16×16 pixels.

(e) $b = 16$: 32×32 pixels.

Figure 3: An original input image (a), and low resolution versions used for rotation alignment with bandwidths (b) $b = 2$, (c) $b = 4$, (d) $b = 8$, and (e) $b = 16$. Color images are show for display purposes only. All alignment techniques presented use grayscale pixel intensity values. Base 2 bandwidths are selected to enable efficient computation of the spherical harmonics.

2.1 Image Registration

Image registration is a simple technique for computing the rotational alignment between a query image q and database image d. The sum of squared error is computed for integer pixel shifts,

$$e(\delta u) = \sum_{u=0}^{2b-1} \sum_{v=0}^{2b-1} d(u, v)\, q(f(u + \delta u), v), \quad \delta u \in \{0, 1, \ldots, 2b - 1\}, \tag{3}$$

where $f(u + \delta u)$ is the function that performs the wrap-around if $u + \delta u$ is larger than $2b - 1$. The integer pixel shift minimizing the error is selected,

$$\delta u^* = \arg\min e(\delta u), \tag{4}$$

and defines the angle α aligning the query image q to the database image d,

$$\alpha = 2\pi \left(\frac{\delta u^*}{2b} \right). \tag{5}$$

2.2 Cross Correlation (Frequency)

Correlation techniques are an alternative to the sum of squared distances registration method described. Here we consider image correlation in the Spherical Fourier Domain (SFT) which has previously been used for image shape matching applications [25, 16], 3D model alignment [6], and specifically for relative orientation and general pose estimation using wide-angle omni-directional images [26, 28, 24, 27]. An overview of image correlation using spherical harmonics is presented in this section, including a summary of the procedure restricted to rotations $R_z(\alpha)$, and practical

implementation details. The reader is referred to [15, 10] for a more comprehensive overview of spherical harmonics and correlation on the sphere.

The correlation of any two functions f and g on the sphere is

$$(f \star g)(R) = \int_{\eta \in \mathbb{S}^2} f(\eta)\, g(R^T \eta)\, d\eta, \tag{6}$$

where $\eta \in \mathbb{S}^2$ is the set of all points on the unit sphere, and $R \in SO(3)$ a rotation that can be parameterized using Euler angles

$$R(\alpha, \beta, \gamma) = R_Z(\gamma)\, R_Y(\beta)\, R_Z(\alpha). \tag{7}$$

Any point $\eta(\theta, \phi)$ on the sphere can be defined using an angle of colatitude $\theta \in [0, \pi)$ and angle of longitude $\phi \in [0, 2\pi)$ – see figure 2(c).

Implementing (6) in a discrete form is challenging as there is not natural/uniform discretization of sample points η on the sphere. The more convenient approached used is to perform the correlation in the spherical Fourier domain. The spherical harmonic expansion (i.e. spectrum) $\hat{f}$ of the function f is

$$\hat{f}_l^m = \int_{\eta \in \mathbb{S}^2} f(\eta) \overline{Y_l^m}(\eta)\, d\eta, \tag{8}$$

$$f(\eta) = \sum_{l \in \mathbb{N}} \sum_{m=-1}^{l} \hat{f}_m^l\, Y_m^l(\eta). \tag{9}$$

where $\overline{Y_l^m}$ is the complex conjugate of the spherical harmonic Y_l^m of degree l and order m. The correlation in (6) using the image spectrums $\hat{f}$ and $\hat{h}$ is

$$h(R) = \sum_{l=0}^{b-1} \sum_{m=-l}^{l} \sum_{p=-l}^{l} \overline{\hat{f}_m^l} \hat{g}_p^l\, U_{pm}^l(R). \tag{10}$$

The matrix elements U^l can be viewed as the spectrum of rotations [24] with values

$$U_{mp}^l\left(R(\alpha, \beta, \gamma)\right) = e^{-im\gamma} P_{mp}^l(\cos(\beta)) e^{-ip\alpha}, \tag{11}$$

where P_{mp}^l are the generalized associated Legendre polynomials. The correlation function in (10) can be evaluated over discrete sets of angles α, β, γ to find the rotational alignment between the functions [26].

Taking the spectrums $\hat{q}$ and $\hat{d}$ of a query image and database image, and restricting rotations to the group $R_z(\alpha)$, the correlation function in (10) reduces to

$$(q \star d)(R(\alpha)) = \sum_{l=0}^{b-1} \sum_{m=-l}^{l} \overline{\hat{q}_m^l} \hat{d}_m^l\, e^{-im\alpha}. \tag{12}$$

The score is computed for the set of $2b$ discrete angles,

$$\alpha = 2\pi \frac{n}{2b}, \quad n \in \{0, 1, \ldots, 2b-1\}, \tag{13}$$

which is the same level of discretization for the image registration method. The angle aligning the query image to the database image is

$$\alpha^* = \arg\max \left(d \star q \right)(R(\alpha))\,. \tag{14}$$

2.2.1 Implementation

We use *s2kit*[1] to compute the spherical harmonic expansion $\hat{f}$ of a $2b \times 2b$ image f. It assumes that the input image f pixels are sample points on the sphere with angles

$$\theta = \frac{\pi(2m+1)}{4b} \qquad m \in \{0, 1, \ldots, 2b-1\}, \tag{15}$$

$$\phi = \frac{\pi m}{b} \qquad n \in \{0, 1, \ldots, 2b-1\}. \tag{16}$$

This angles are selected based on the computationally efficient scheme proposed by Driscol and Heally [10].

The low resolution $2b \times 2b$ images used throughout this work are sampled at the angles ϕ above, but have had their θ sample points 'stretched' to retain more image information in this direction. For convolutions $R(\alpha, \beta, \gamma)$ this would create erroneous results. However, for rotations R_z we can use this sample scheme as rotations are restricted to the angle ϕ.

2.3 Zero-phase

The final alignment technique is a zero-phase alignment based on the method developed for single degree of freedom rotation alignment for panoramic images in [34]. They proposed that any panoramic image can be rotated to a canonical orientation using the zero order horizontal phase shift of the 2D image spectrum. Any two *zero phase* images are in the same rotational frame of reference and can be compared directly without having to compute a relative orientation angle α between them (e.g. using registration or correlation).

Our method is similar, except that we use the spherical harmonic spectrum $\hat{f}$ in place of the $2D$ Fourier transform used in [34]. Figure 4 illustrates the real component of the spherical harmonic functions for increasing degree l. Referring to this figure, the zero-phase shift angle ρ is taken as

$$\rho = \hat{f}_{l=1}^{m=1}, \tag{17}$$

which is the first sectoral harmonic component of the spectrum $\hat{f}$ ($\hat{f}_0^0$ is the DC component). The angle ρ can take any continuous value in the range 0 to 2π. The resulting relative angle α between different zero phase images therefore takes continuous values in the range 0 to 2π. This is different to the registration and correlation methods presented that can only take one of $2b$ discrete angles α. Examples of the zero-phase alignment using images with increasing bandwidth b are illustrated in figure 5.

2.4 Image Alignment Examples and Processing Time

Figure 6 illustrates image alignment for each of the methods discussed for small image sequences. The original (non-aligned) images are also included for reference. The alignment angles α for the registration and spherical correlation methods were computed between successive images and have been applied as a pixel shifts to the original panoramic images. For the zero-phase alignment, the phase shift angles ρ were computed separately for each image and also applied as a pixel shifts to the original panoramic images.

[1] Available `http://www.cs.dartmouth.edu/~geelong/sphere/`

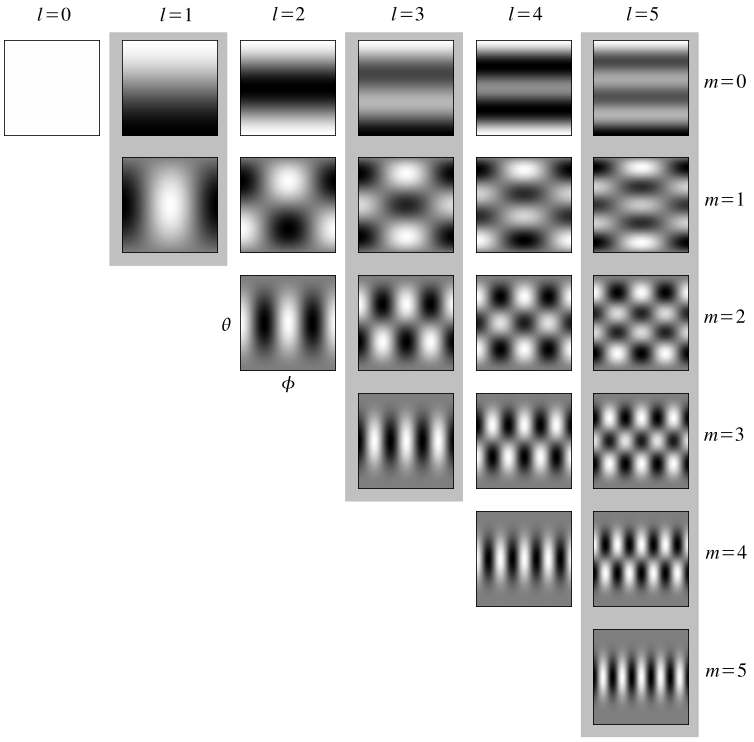


Figure 4: The magnitude of the real components of the spherical harmonic functions Y_l^m shown up to degree $l = 5$ and order $0 \leq m \leq l$. The functions are shown sampled on the θ, ϕ plane with angles $\theta \in [0, \pi)$ and $\phi \in [0, 2\pi)$. Zero phase alignment uses the phase shift of the sectoral image spectrum component $\hat{f}_{l=1}^{m=1}$.

(a) Original input image.

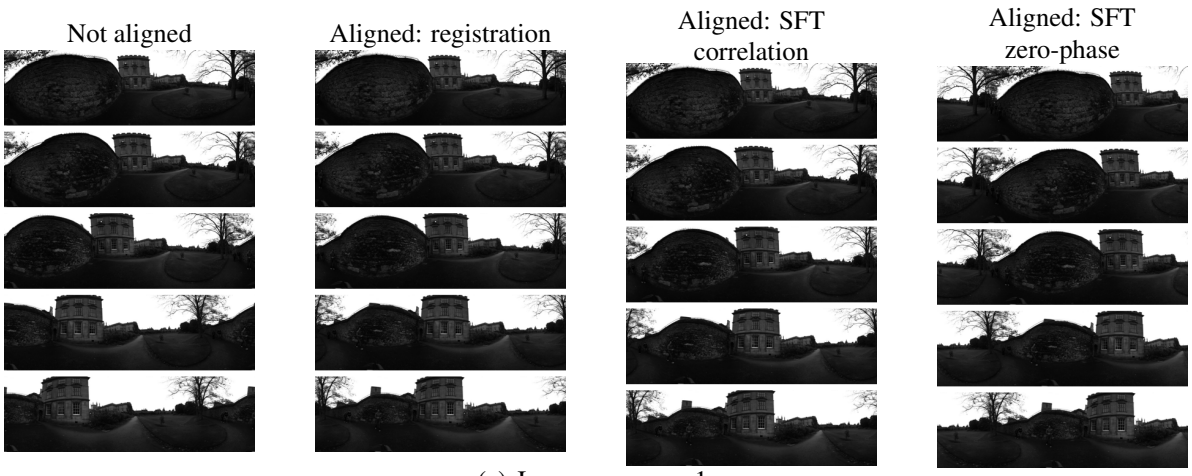
(b) Zero-phase alignment $b = 2$ ($\alpha = 187.30°$).

(c) Zero-phase alignment $b = 4$ ($\alpha = 165.00°$).

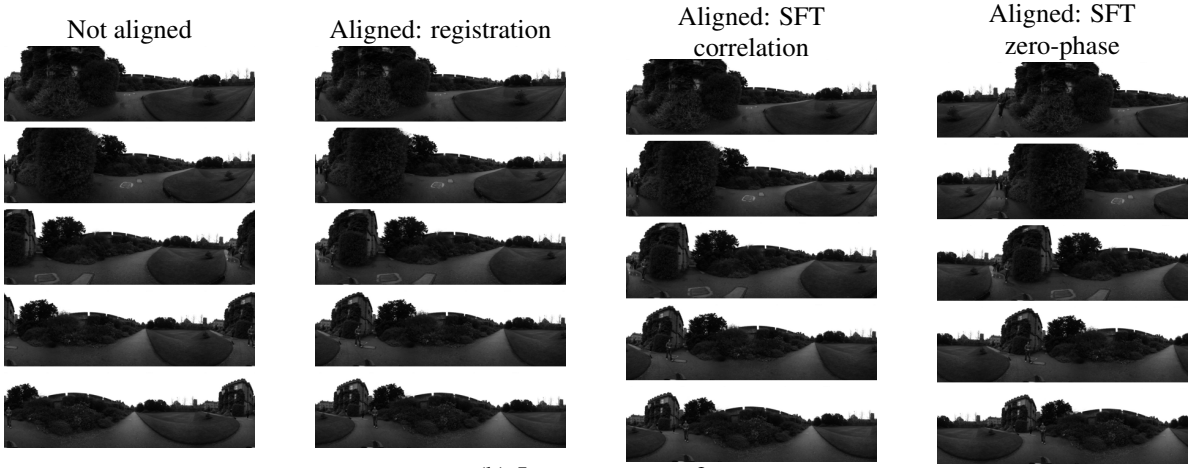
(d) Zero-phase alignment $b = 8$ ($\alpha = 153.52°$).

(e) Zero-phase alignment $b = 16$ ($\alpha = 148.12°$).

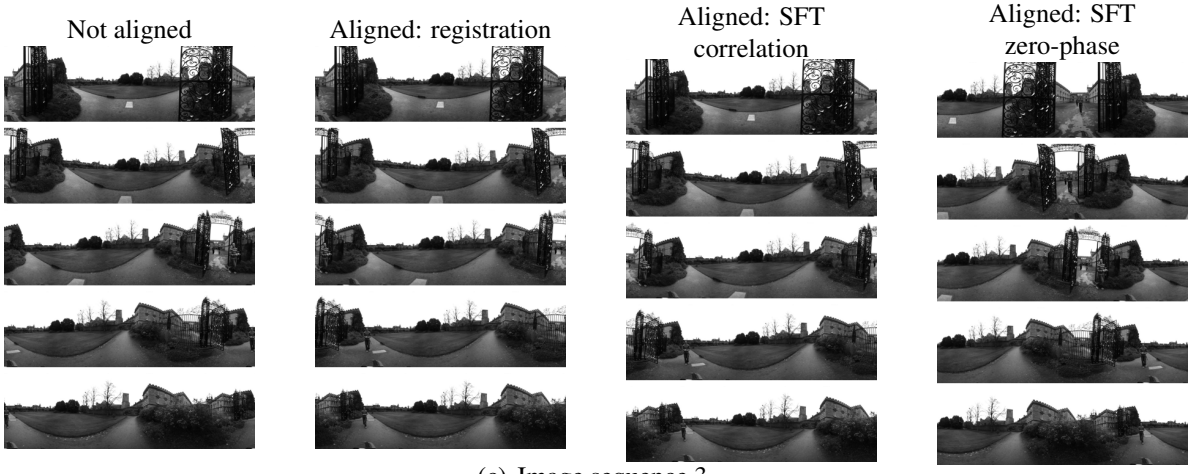
Figure 5: Zero-phase alignment for the original input image in (a) for bandwidths (b) $b = 2$, (c) $b = 4$, (d) $b = 8$ and (e) $b = 16$. In (b) through (e), the left image shows the $2b \times 2b$ image of the inverse transform of the zero-phase component, and the right shows the phase shift ρ applied to the original input images.



(a) Image sequence 1.



(b) Image sequence 2.



(c) Image sequence 3.

Figure 6: Alignment methods applied to three small images sequences. For registration and SFT correlation the angles α have been computed between successive images. For the zero-phase alignment, each individual image has been shifted to its canonical orientation using the zero phase angle ρ.

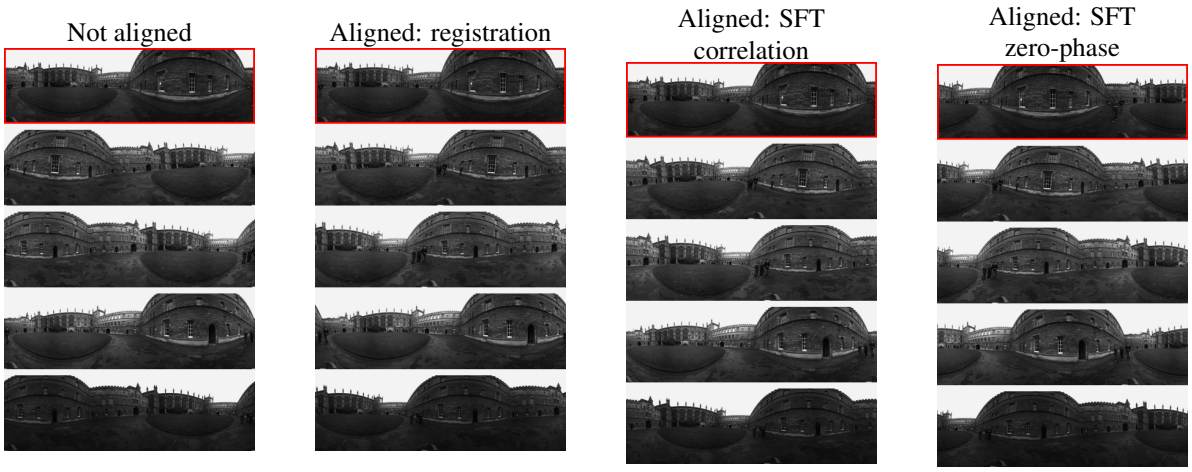


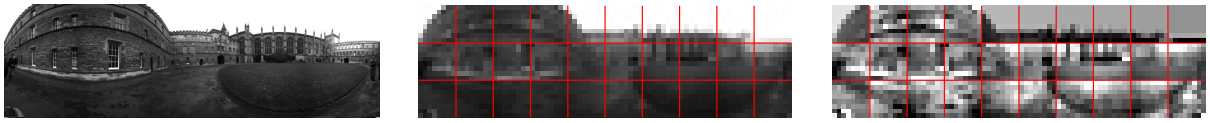
Figure 7: Example image alignment between a query image (top row) and multiple database images taken at a similar location, but not necessarily the same orientation. For registration and SFT alignment, the database images have been rotated to the query image rotation frame of reference.

As a second example, the alignment between a query image, and multiple database images at a similar location, are displayed in figure 7. The angles α for the correlation methods were computed directly between the query image and each database image. Again, the zero-phase alignment is implemented separately for each image without direct comparison between the query and database images.

A summary of processing times for each image alignment method are presented in table 1 for a range of image bandwidths using a Matlab implementation on a desktop computer. The mex functions supplied with s2kit were used to compute the forward spherical Fourier transforms. The zero-phase alignment is consistently the most computationally efficient method across all bandwidths tested. Moreover, assuming the zero phase angles have been precomputed for all database images, the processing time does not depend on the database size.

Table 1: Processing times in seconds to compute the alignment of a query image to 1000 database images. For SFT correlation, the spectra of all database images are assumed to be precomputed. For zero-phase, the phase shift for each database image is assumed to be precomputed. Processing times for SFT correlation and zero-phase include the time taken to compute the spherical Fourier transform.

Bandwidth b	Registration	SFT Correlation	Zero-phase	SFT (s2kit)
4	0.0760	0.0379	0.0003	0.0003
8	0.1140	0.0496	0.0003	0.0003
16	0.2057	0.1117	0.0009	0.0009
32	0.9813	0.3426	0.0025	0.0025
64	11.1154	6.1027	0.0136	0.0136



(a) Original grayscale image.

(b) Low resolution image ($u = 10$, $v = 3$ cells).

(c) Low resolution contrast enhanced descriptor.

Figure 8: Overview of the process for creating low resolution contrast enhanced descriptors. An original image is converted toa low-resolution image with $u \times v$ cells, each 8×8 pixels. A contrast enhancement is applied separately to the values in each cell.

3 Place Recognition

The goal of rotation alignment is to improve visual place recognition performance, and in this section we discuss how it is incorporated into our existing visual place recognition system based on sequence SLAM [29, 30]. This includes how rotation alignment is used to compute and compare image descriptors, and extensions to the place recognition sequence matching and scoring procedure.

3.1 Aligned Image Descriptors and Similarity

Our system uses the same low-resolution contrast-enhanced image descriptors introduced as part of sequence SLAM. Referring to figure 8, a low-resolution grayscale image with $x \times y$ cells is created, each 8×8 pixels in size. Each cell is contrast enhanced separately by removing the mean grayscale pixel intesity and then normalizing to unit standard deviation. The resulting contrast enhanced image descriptor D values are unit standard deviation z-scores. The similarity between a query descriptor D_q and database descriptor D_d is

$$M(q,d) = \frac{1}{N_v N_u} \sum_{u=1}^{N_u} \sum_{v=1}^{N_v} \text{abs}\left(\Phi(D_d(v,u)) - \Phi(D_q(v,u)\right), \tag{18}$$

where $N_v = 8y$ is the number of rows, $N_v = 8x$ is the number of columns, and Φ is the cumulative distribution function of the standard normal distribution with unit standard deviation. Using this function, the similarity values M are probabilities in the range 0 to 1. Note that the size of the descriptors is not the same as the low-resultion $2b \times 2b$ images used for rotation alignment which are typically much smaller.

For alignment using registration and SFT cross correlation, the angle α aligning the current query image to each descriptor image is found. A column shift $u(\alpha)$ using (2) is applied to the low-resolution contrast enhanced query image descriptor D_q before computing the similarity score using D_q^α and D_d in (18). This column shift $u(\alpha)$ could be applied to the low-resolution image before contrast enhancing. This would increase computational expense as contrast enhancement would have to be applied to each shifted low-resolution query image.

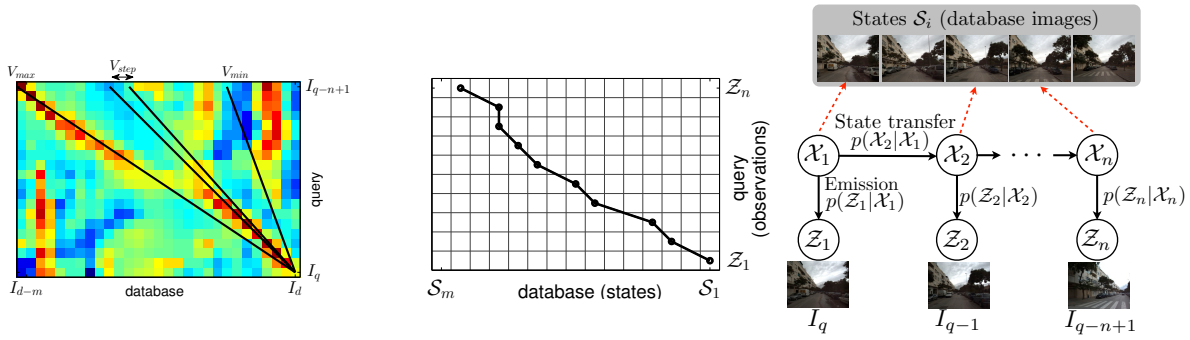
Zero-phase alignment finds a phase shift angle ρ separately for each image. This phase shift angle is converted to a column shift $u(\rho)$ which is applied to the low-resolution images before contrast enhancement is applied to find the final descriptor D^ρ. Applying this shift before contrast enhancement adds no additional computational expense.

3.2 Sequence Matching and Scoring

Using the similarity score $M(q, d)$ between a single query image and single database image alone is not enough for robust place recognition scoring. More robust performance is achieved by matching matching local query image *sequences* to local database image *sequences* using the similarity scores, and then scoring. In section 3.2.1 we provide an overview of our original procedure for sequence matching and scoring for reference. The limitation of this approach is that the query and database sequence being matched must have been traversed while traveling in the same direction. In section 3.2.1 we introduce our revised approach capable of matching sequences irrespective of the direction of travel. This fully exploits the wide-angle of view and rotation alignment.

3.2.1 Original Implementation

To evaluate a place recognition score between a query image I_q and a database image I_d, the matrix $M_\mathbf{d}$ of image similarity scores between the *sequences* of images $I_\mathbf{q} = \{I_{q-n+1}, \ldots, I_q\}$ and $I_\mathbf{d} = \{I_{d-m+1}, \ldots, I_d\}$ is computed. A path through the matrix is found maximizing some function of the similarity values, and a place recognition score computed using the matched path (e.g. integrate similarity values along the path). In sequence SLAM the authors apply a local row-based similarity value normalization (see [29]) and find the path maximizing the sum of normalized similarity values. The paths tested are restricted to straight lines which assumes the robot traverses a previous sequence of locations at a scale multiple of the previous velocity. This was referred to as a *continuous* Dynamic Time Warping (DTW) and should not be confused with the classical DTW algorithms used extensively for speech recognition [35]. In out adapted algorithm we select the best path using a Hidden Markov Model (HMM) framework. As illustrated in figure 9, improved path selection flexibility is achieved that can improve place recognition performance [17].



(a) SeqSLAM constant velocity search lines.

(b) Modified seqSLAM using a HMM.

Figure 9: Sequence SLAM matches local query and database image sequence using a constant velocity assumption - the paths through the matrix are restricted to straight lines. Using a modified HMM framework for sequence matching greatly improves the flexibility of path matching.

Within the HMM framework, the n query images are the observations $\mathcal{Z}_{1:n} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_n\}$, and the database image sequence $I_\mathbf{d}$ the state space $\mathcal{S}$. For each of the n query images 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 state

sequences $\mathcal{X}_{1:n} = \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$,

$$\mathcal{X}^* = \underset{\mathcal{X}_{1:n}}{\operatorname{argmax}}\, p(\mathcal{X}_{1:n}|\mathcal{Z}_{1:n}) \tag{19}$$

$$= \underset{\mathcal{X}_{1:n}}{\operatorname{argmax}}\, p(\mathcal{X}_{1:n}, \mathcal{Z}_{1:n}). \tag{20}$$

The Viterbi algorithm [39] is used to efficiently compute the largest joint probability μ_k^i of all path combinations $\mathcal{X}_{1:k}$ ending at state $\mathcal{X}_k = i$ using dynamic programming:

$$\mu_1^i = \underbrace{p(\mathcal{X}_1 = i)}_{\text{initial}}\, p(\mathcal{Z}_1|\mathcal{X}_1 = i) \tag{21}$$

$$\mu_t^i = \underbrace{p(\mathcal{Z}_t|\mathcal{X}_t = i)}_{\text{emission}}\, \max_{j \in S} \left[\underbrace{p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j)}_{\text{state transfer}}\, \mu_{t-1}^j \right]. \tag{22}$$

The probability of the optimal fixed length n state sequence is

$$\mu(\mathcal{X}^*) = \max_{\mathcal{X}_n}(\mu_n(\mathcal{X}_n)), \tag{23}$$

and the sequence $\mathcal{X}^*$ itself recovered by backtracking through the stored argmax indices at each iteration.

The emission probabilities are the values in the similarity matrix $M_{\mathbf{d}}$ normalized so that the sum of values in each column is equal to one. Referring to figure 10(a), the initial state probability distribution is

$$p(\mathcal{X}_1 = i) = \begin{cases} 1 & i = 1 \\ 0 & i > 1 \end{cases}, \tag{24}$$

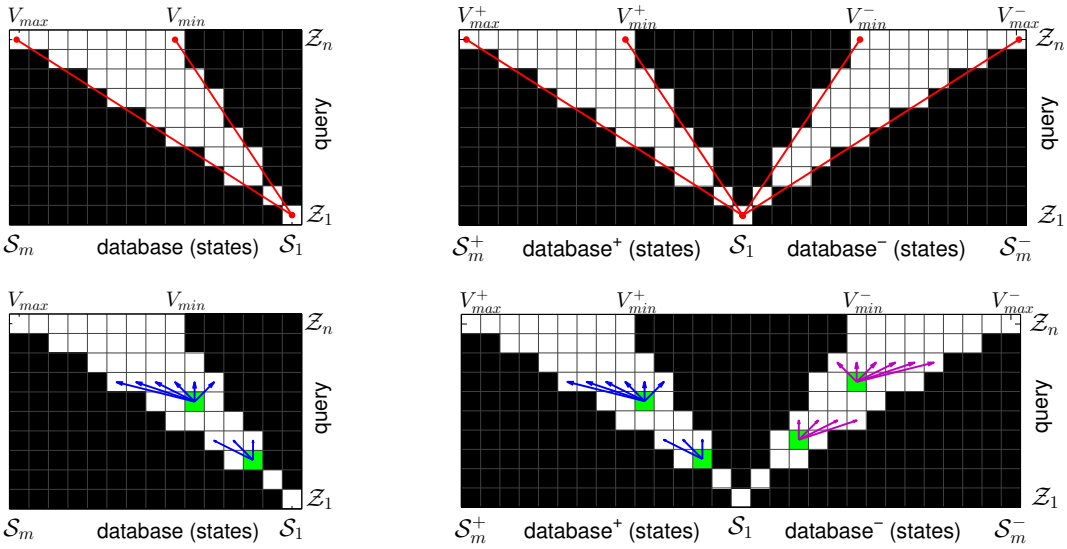
which enforces that the current query image I_q must correspond to the database image I_d. The state transfer/transition probabilities are the likelihoods of transitioning between the m states (database images) from one observation (query image) to the next. We first define *local* velocity constraints to set values of the $m \times m$ state transition matrix $A(j, i) = p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j)$

$$A(j,i) = \begin{cases} 0 & i < j \\ 1 & 0 \leq (i-j) \leq (V_{max} + 0.5) \\ exp\left(\frac{-(i-j-V_{max})^2}{2V_{max}^2}\right) & \text{otherwise} \end{cases} \tag{25}$$

which is a truncated Gaussian distribution with a flattened peak. The matrix is then normalized to satisfy

$$\sum_{i=1}^{m} A(j,i) = \sum_{i=1}^{m} p(\mathcal{X}_t = i|\mathcal{X}_{t-1} = j) = 1 \quad \forall j, t. \tag{26}$$

Next, the *global* velocity mask illustrated in figure 10(a) is used to constrain any path to lie within the bounds of the mask. Any state transfer value in (22) is set to zero if the mask value at cell t, j or t, i is zero. It is clear to see that the original global velocity mask only permits 'positive' velocity paths, that is, the robot must traverse a previous path moving in the same direction.



(a) Original forward velocity state transitions.

(b) Bi-directional state transitions.

Figure 10: Overview of the HMM state transitions for the (a) original single direction sequence matching and (b) the bi-directional sequence matching. The top row shows the 'global' velocity constraints, middle row samples of the state transitions, and bottom row the probability values.

The place recognition score $S(q, d)$ between the query and database image is found by taking the singular value decomposition (SDV) of the matrix $M_{\mathbf{d}}$,

$$U \, \Sigma \, V^T = SVD(M_{\mathbf{d}}), \tag{27}$$

and producing the rank-reduced matrix $\tilde{M}_{\mathbf{d}}$

$$\tilde{M}_{\mathbf{d}} = U \, \tilde{\Sigma} \, V^T, \tag{28}$$

where $\tilde{\Sigma}$ is the $n \times n$ matrix Σ with the first r singular values set to zero. This rank-reduction appraoch is based on [20], and a heuristic value $r = 4$ is selected. The place recognition score is the Gaussian weighted sum of rank-reduced values along the optimal path $\mathcal{X}^*$,

$$S(q, d) = \sum_{i=1}^{n} G(i) \, \tilde{M}_{\mathbf{d}}(\mathcal{X}^*_i), \tag{29}$$

where

$$G(i) = \exp\left(\frac{-(i-1)^2}{2n^2}\right). \tag{30}$$

3.3 Bi-directional

The allowable matrix paths in the original implementation move up and to the left, as evident in figure 10(a). This can be considered a positive/forward velocity search and can only match sequences if the a robot traverses a prior path traveling in the same general direction. To fully

exploit the 360 degree angle of view of panoramic images and rotation alignment, we must enable forward and reverse, or bi-directional, sequence matching.

Referring to figure 10(b), the bi-directional sequence matching approach searches for paths in the similarity matrix $M_{\mathbf{d}}$ spanning the sequences of images $I_{\mathbf{q}} = \{I_{q-n+1}, \ldots, I_q\}$ and $I_{\mathbf{d}} = \{I_{d-m+1}, \ldots, I_d, \ldots, I_{d+m-1}\}$. The same state transitions (local and global) can be mirrored to allow for reverse velocity paths moving up and to the right.

In practice, we first select the matrix $M_{\mathbf{d}}^{+}$ using the set of database images $I_{\mathbf{d}} = \{I_{d-m+1}, \ldots, I_d\}$ and compute the forward velocity place recognition score $S^{+}(q, d)$ using the original implementation. Next, the matrix $M_{\mathbf{d}}^{-}$ using the set of database images $I_{\mathbf{d}} = \{I_d, \ldots, I_{d+m-1}\}$ is selected and mirrored horizontally. The original implementation, operating on the mirrored matrix $M_{\mathbf{d}}^{-}$, is then used to find the reverse place recognition score $S^{-}(q, d)$. The final place recognition score is select as the largest of S^{-} and S^{+}.

4 Experimental Setup

In this section we provide an overview of the experimental setup. This includes a description of the omni-directional dataset used, the ground truth place recognition, and the selected parameter settings.

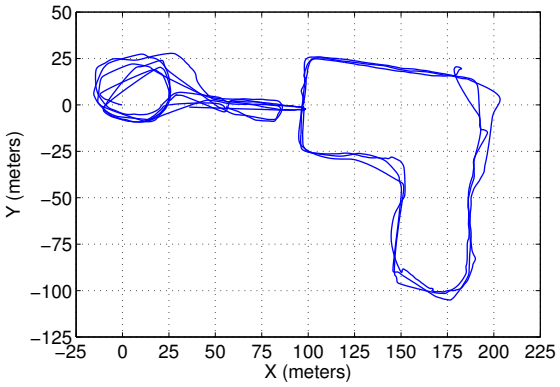
4.1 Omni-directional Dataset

The omni-directional panoramic imagery in the New College vision and laser dataset [38] is used in the experiments in this work. Example images from the dataset were displayed previously in the figures in section 2. The 2048×618 pixel RGB panoramic images in the dataset were produced by stitching together multiple images from a Ladybug camera array (see figure 1(b)) mounted on a Segway robot. There are a total of 8127 images that were logged at 3 frames per second. For our experiments we use every fifth image.

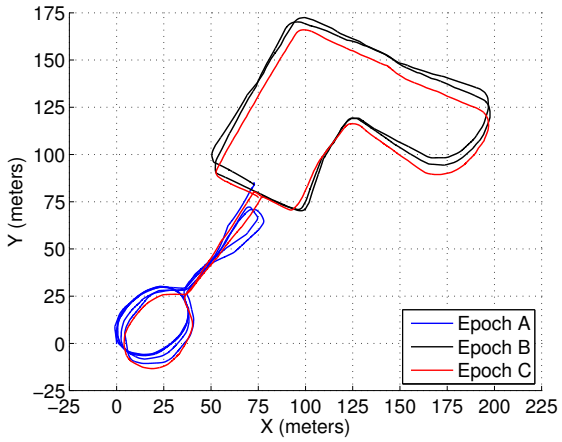
The GPS and visual odometry results from the on-board stereo camera system are displayed in figure 11. The zoomed in views of the visual odometry in 11(c) and 11(d) show that the robot traversed the same paths multiple in both the same and opposite directions.

4.2 Performance Metric and Ground Truth

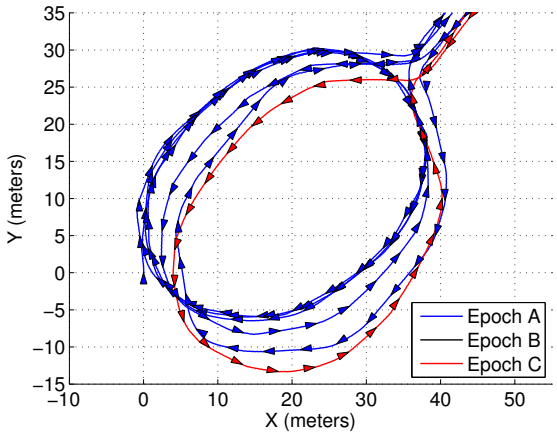
We use precision recall statistics to evaluate the performance of visual place recognition. For each new query image frame I_q, the set of previously logged images become the image database $I_d = \{I_{\mathbf{d}} = \{I_1, \ldots, I_{q-s}\}$. These database image must appear at least s images before the current query image to avoid matching to recent images. The corresponding database image is selected as the one having the largest place recognition score. The matches are ranked in order from highest



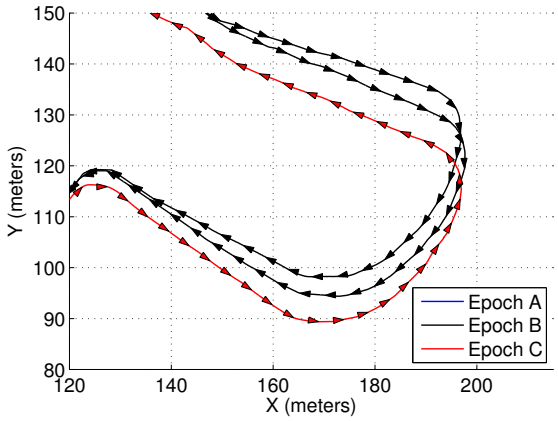
(a) GPS.



(b) Visual Odometry.



(c) Visual Odometry: zoomed view 1. The arrows indicate the direction of travel.



(d) Visual Odometry: zoomed view 2. The arrows indicate the direction of travel.

Figure 11: Camera path: (a) GPS and (b) visual odometry from the stereo camera system. The results are displayed in different coordinate frames of reference. Zoomed in views of the visual odometry are shown in (c) and (d) with the arrows indicating the direction of travel. Observe that the robot traverses the same paths in opposite directions. The visual odometry positions are used to select candidate image sets during the hand-labeling of the ground truth place recognition frames.

to lowest score and the precision recall values computed as

$$\text{precision} = \frac{tp}{tp + fp} \tag{31}$$

$$\text{recall} = \frac{tp}{tp_{total}}, \tag{32}$$

where tp is the number of true positives, fp the number of false positives. The parameter tp_{total} is the total number of possible true positives for the dataset. This is the total number of query images that match to any database image.

Ground truth is required to correctly label true positive and false positive correspondences. The ground truth is an $l \times l$ adjacency matrix G, where l is the number of images in the datasets

Table 2: Summary of the parameter settings used in experiments.

	Parameter	Value	Description
HMM	V^+_{min}	0.67	Minimum global velocity (forward)
	V^+_{max}	1.50	Maximum global velocity (forward)
	V^-_{min}	0.67	Minimum global velocity (reverse)
	V^-_{max}	1.50	Maximum global velocity (reverse)
	n	20	Sequence length
Descriptor	u, v	$10, 3$	Number row and column cells
	w	8	Pixel width per cell
Rotation Alignment	b	16	Bandwidth of image for rotation alignment

used. For each query image, we hand label all true positive databases images matches and populate that row of the adjacency matrix. A database image is considered a match if the same region of the scene is imaged irrespective of the robot orientation. Each query image may be matched to multiple database images. The total number of true positives tp_{total} for the dataset is the number of matrix rows with any non-zero value.

4.3 Parameter Selection

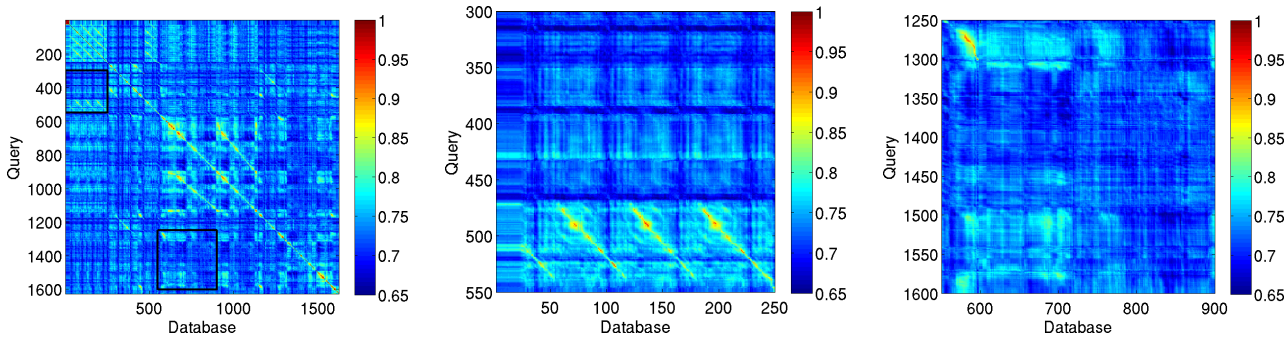
A summary of the selected parameters is provided in table 2. Note that the low-resolution contrast enhanced image descriptor is higher than the $2b \times 2b$ images used for rotation alignment. A low resolution bandwidth b is selected for improved rotation alignment processing efficiency.

5 Results and Discussion

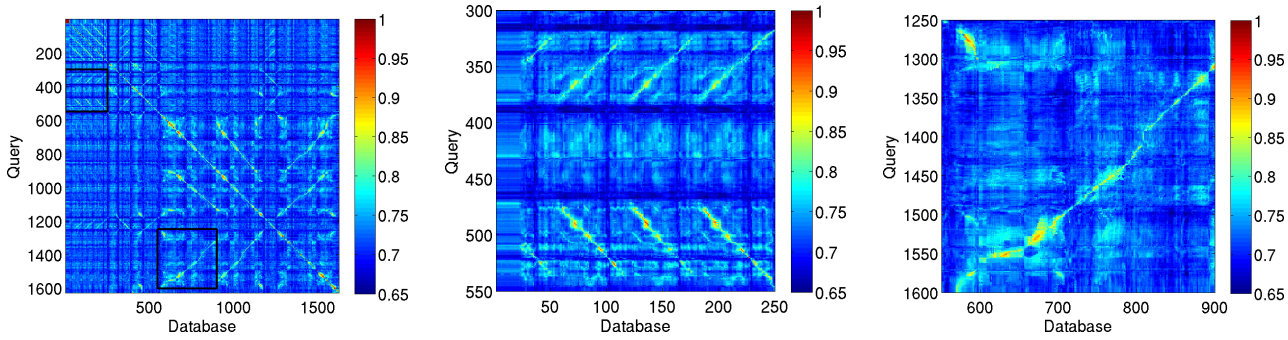
The goal of rotation alignment and the bi-directional sequence matching search is to improve the performance of omni-directional visual place recognition. In this section we compare the place recognition results for each alignment method against benchmark results using no alignment. Moreover, we include results for both the original forward only velocity search (V^+), and the bi-directional search (V^-, V^+). Relevant discussions are presented, as well as directions for future work.

5.1 Image Similarity

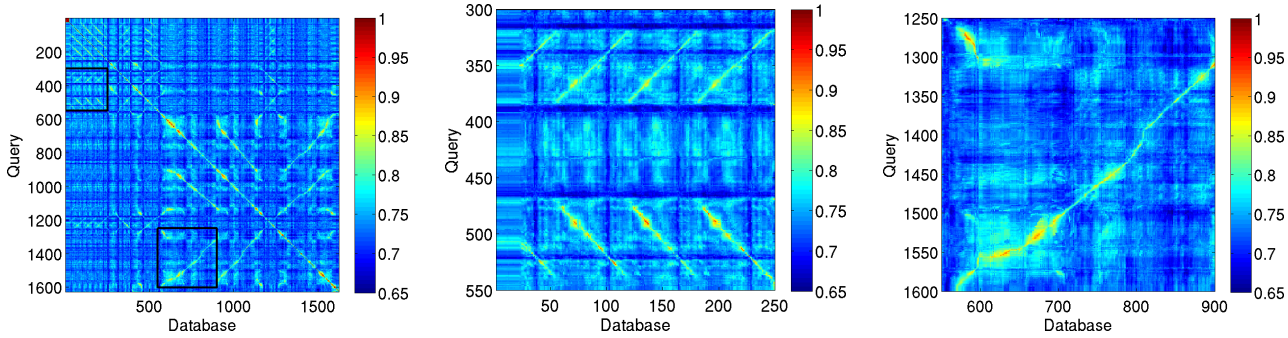
The first step in finding the place recognition results is to compute the image similarity matrix. The image similarity results for each alignment method are presented in figure 12, as well as the benchmark result using no image alignment. Zoomed in sections of the matrix are included to highlight the detail. Although the full similarity matrices are shown, the search for valid sequence matches is restricted to the lower diagonal of each matrix as database images must appear before the current query image for this dataset. Correct query and database sequence matches appear as paths/lines in the matrix with high similarity values.



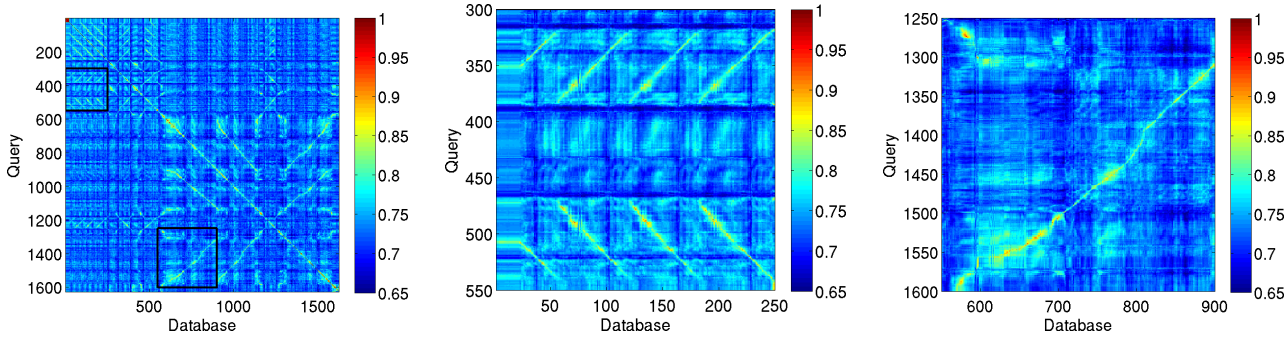
(a) No alignment.



(b) Aligned: Registration.

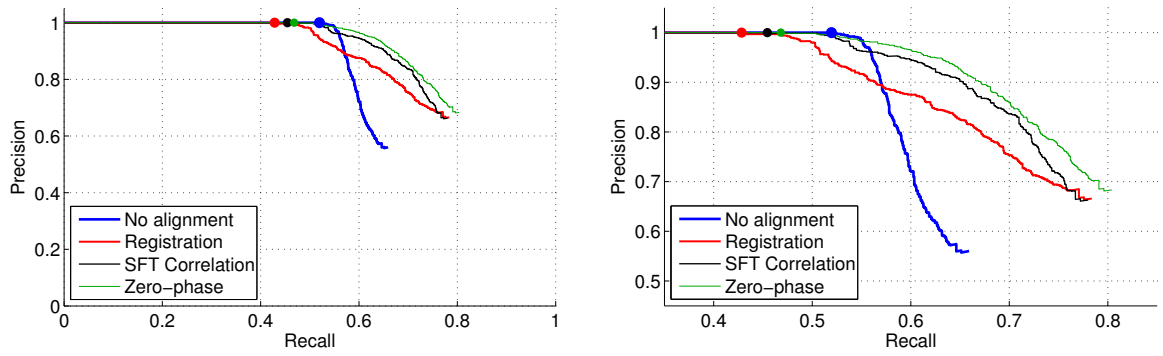


(c) Aligned: SFT cross correlation.

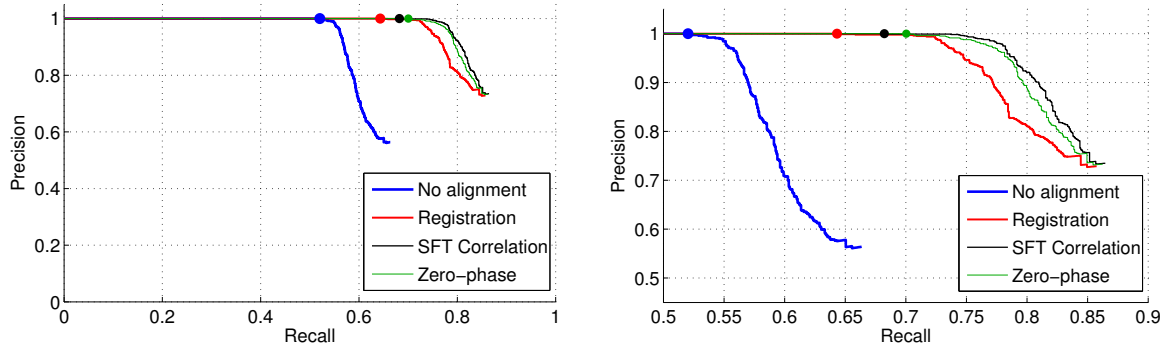


(d) Aligned: Zero-phase.

Figure 12: The image similarity matrices between query and database images using (a) no rotation alignment, (b) registration alignment, (c) SFT correlation and (d) zero-phase alignment. Zoomed in sections where the robot traversed a previously visited path in the opposite direction are displayed. In this scenario paths of high similarity are expected to move up and to the right.



(a) Forward velocity search V^+ only.



(b) Bi-directional velocity search V^-, V^+.

Figure 13: Precision recall results for each alignment method, and the benchmark using no alignment. The results using the original forward velocity V^+ search are in (a), and the bi-directional sequence matching with the addition of the reverse velocity V^- search in (b). Zoomed in versions are provided in the right column to show small variations in the results.

The most noticeable difference when using any rotation alignment is an increase in the number of distinctive local paths in the matrix with high similarity values. These additional paths all move up and to the right and are most clearly seen in the zoomed in views in figure 12. These paths occur when the robot traverses a previous sequence in the environment in the opposite direction. The alignment methods presented in section 2 enable the low-resolution contrast enhanced query and database descriptors to be compared in a common rotational frame of reference. The bi-directional HMM sequence matching presented in section 3 enables these paths of high similarity to be correctly identified and scored.

5.2 Place Recognition Performance

The visual place recognition recall precision curves for each alignment method, and the benchmark without alignment, are presented in figure 13. This includes the results using only the original forward velocity V^+ HMM sequence matching (figure 13(a)), and the bi-directional sequence matching with the addition of the reverse velocity search V^- (figure 13(b)). A summary of recall values at specific precision levels are provided in table 3.

Table 3: Summary of the precision recall results for each alignment method and both forward only and bi-directional sequence matching. The recall values at selected precision levels, and the maximum recall value across all precision levels, are provided. Refer to figure 13 for the full recall precision curves.

Alignment Method	Search	Recall				
		1.00 prec	0.99 prec	0.95	0.90 prec	maximum
No alignment	V^+	**0.520**	**0.546**	0.561	0.569	0.659
Registration	V^+	0.428	0.478	0.517	0.564	0.784
SFT Correlation	V^+	0.454	0.524	0.589	0.653	0.780
Zero phase	V^+	0.468	0.534	**0.624**	**0.671**	**0.804**
No alignment	V^-, V^+	0.520	0.546	0.562	0.570	0.664
Registration	V^-, V^+	0.643	0.725	0.748	0.772	0.857
SFT Correlation	V^-, V^+	0.682	**0.759**	**0.789**	**0.810**	**0.865**
Zero-phase	V^-, V^+	**0.700**	0.745	0.784	0.797	0.863

5.2.1 Forward velocity sequence matching

For the forward only velocity V^+ sequence matching, there is an increase in the maximum recall rates for all rotation alignment methods over the benchmark result using no alignment. Zero phase alignment was found to outperform the registration and zero phase correlation methods having higher recall rates for all levels of precision.

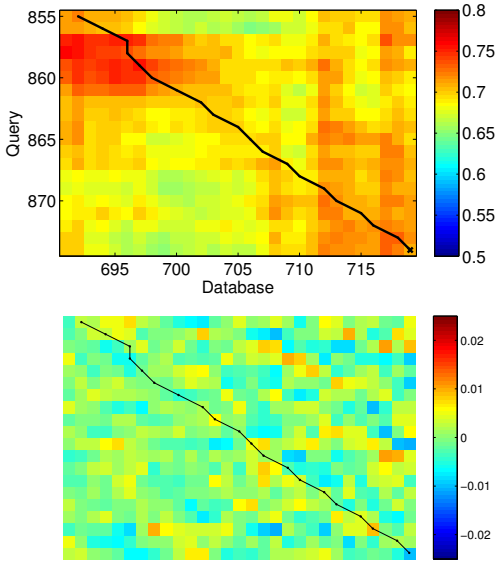
There was a significant increase in the maximum recall rates for all alignment methods. Even though the reverse sequence matching was not used, the rotation alignment still assigns high similarity between query and database images at similar locations but different orientations. This individual similarity score $M(q, d)$ will contribute to the forward velocity place recognition score, and if $M(q, d)$ is sufficiently large it will rank as the best (and correct) match for the query image.

Despite overall performance improvements using rotation alignment, all methods show decreased recall rates at high precision levels. Figure 14 shows an early false positive match using zero phase alignment. The figure includes the similarity matrix and rank-reduced form used for scoring without any rotation alignment in (c), and with zero phase alignment in (c). It is clear that between each method there are variations in both the similarity values and rank-reduced values. The rank-reduced values for zero phase, which are integrated along the matched sequence through the matrix to compute the place recognition score, are much larger leading to the false positive detection.

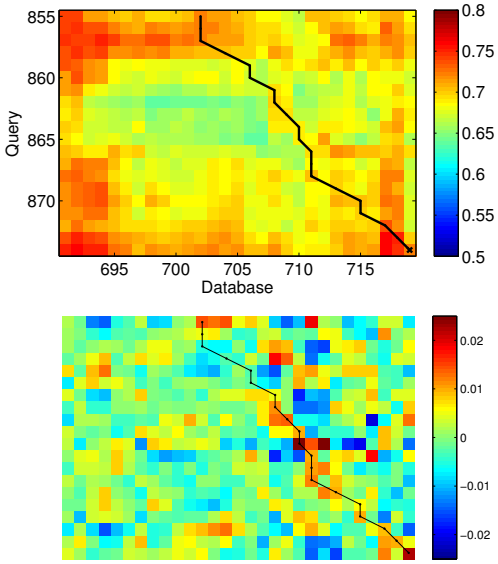
We would expect the similarity values $M(q, d)$ using zero phase alignment, or any rotation alignment in general, to increase over the no alignment result. In effect, image alignment puts the images into a rotational frame of reference maximizing image similarity. However, the $2b \times 2b$ images used for rotation alignment are not the same as the low-resolution contrast enhanced image descriptors. There is therefore no guarantee that alignment will maximize the image descriptor similarity. Improved performance could potentially be achieved by aligning images using the low-resolution contrast enhanced descriptors themselves. This would significantly increase computational expense. An alternate would be to apply an additional normalization similar to contrast enhancement to the $2b \times 2b$ images.

(a) Panoramic query image (top) and false positive panoramic database image (bottom)

(b) The zero phase aligned query and panoramic images in (a).



(c) No Alignment: Original similarity matrix (top) and rank-reduced (bottom) with sequence match shown as black line.



(d) Zero phase alignment: Original similarity matrix (top) and rank-reduced (bottom) with sequence match shown as black line.

Figure 14: An early false positive query/database image match at 0.507 recall using the zero phase alignment. Shown are the original query and database panoramic image pair (a) and their zero phase aligned versions (b). The similarity matrix and rank-reduced form used for scoring without rotation alignment in (d), and the same computed using zero phase alignment in (d).

5.2.2 Bi-directional sequence matching

The results using the bi-directional velocity sequence matching are the most relevant to this work. Significant performance improvements over the benchmark result were observed using all rotation alignment methods. Of these, SFT alignment provided the best overall performance. The results for zero phase alignment are very similar to SFT correlation, and show marginally higher recall rates for precision levels greater than 0.99. Zero phase also has the distinct advantage of being the most computationally efficient of all alignment methods tested, and whose processing time remains fixed irrespective of database size – refer to table 1.

The aerial views of the query and database images matches at 0.99 precision using the bi-directional sequence matching for all rotation alignment methods are provided in figure 15. A

zoomed in section of the map showing the benchmark non-aligned matches and the zero phase matches are shown in figure 15(e). In this section of the map the robot moved along a previously visited sequence of locations in the opposite direction. Using zero phase rotation alignment and the bi-directional sequence matching enabled correct database image matches to be recalled. This improves the flexibility of visual place recognition in real world applications, and demonstrates that the proposed system better makes use of the 360 degree angle of view panoramic imagery.

5.3 Directions for Future Work

All rotation alignment techniques presented operate on low-resolution zero-mean grayscale images. This is only one of many different data normalization techniques used for image registration/correlation. For example, we could perform a contrast enhancement on the $2b \times 2b$ images before alignment. Moreover, for registration there are many alternate descriptor distance metrics that could be used to such as Sum of Absolute Difference (SAD) and Mahalanobis distance. A more extensive comparison of normalization and distance metrics, as well as performance versus the image bandwidth b used to alignment are directions for future work. This would require more extensive experiments using image datasets with increasing variation in atmospheric conditions, especially lighting.

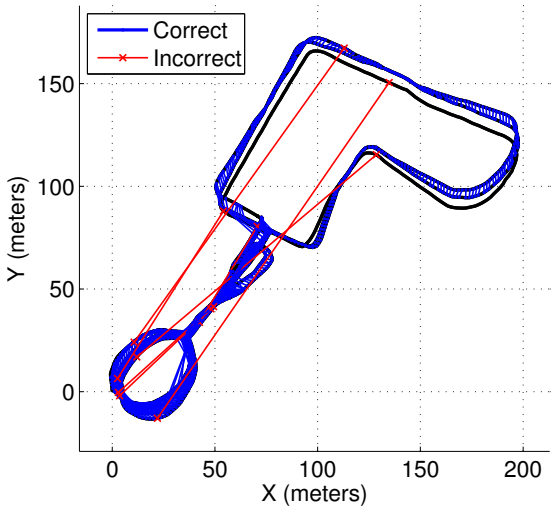
Another direction for future work is improved place recognition scoring. The current implementation integrates rank-reduced image similarity values along the optimal sequence alignment selected using the HMM framework. The decreased recall rates using alignment and the forward only velocity search suggest that improvements to this scoring are required. These improvements may include changes to the heuristic number of singular values set to zero, or more major changes to the overall process itself.

The final direction for future work is to more fully integrate rotation alignment into HMM sequence matching and scoring framework. Currently the rotation alignment angles are used only when finding the initial image similarity values. As a simple example, if the alignment angles between a query sequence stay constant, and the alignment angles between a database sequence stay constant, then one would expect the angles aligning the query images to the database images to remain constant. There is an opportunity to use information/constraints such as this when setting state transition probabilities within the HMM sequence matching framework.

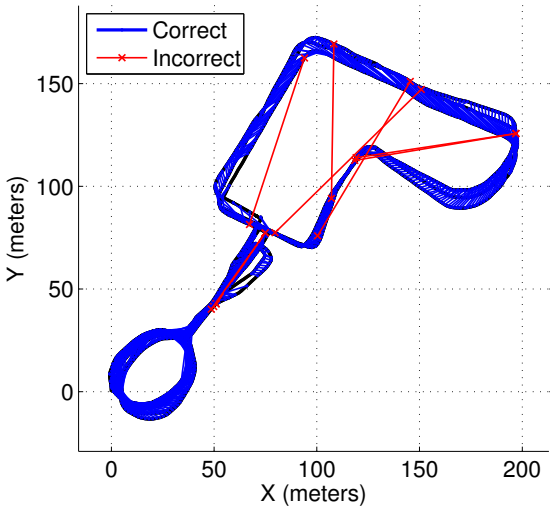
6 Conclusions

Panoramic images with a full 360 degree circumferential angle of view can image the same regions of a scene under extreme rotational viewpoint changes. We presented a visual place recognition system that fully exploits the full wide-angle of view of panoramic images to enable robust place recognition when traversing prior sequences in the environment in any direction. This is particularly advantageous for ground vehicle applications amongst others.

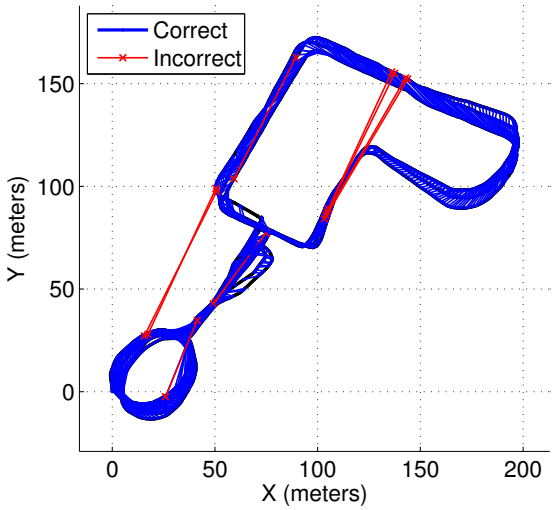
The system builds on our existing variant of the sequence SLAM algorithm. To compare the similarity of image, Low-resolution contrast enhanced image descriptors are rotated to a common rotational coordinate frame of reference. We explored three approaches for efficiently computing the rotational alignment between images, each operating on low resolution bandwidth images. The



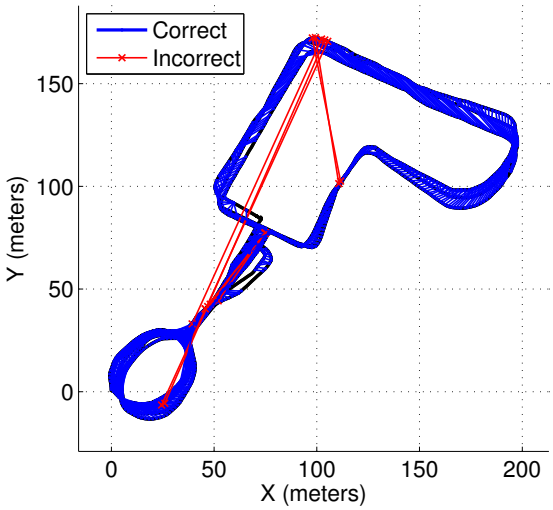
(a) No alignment.



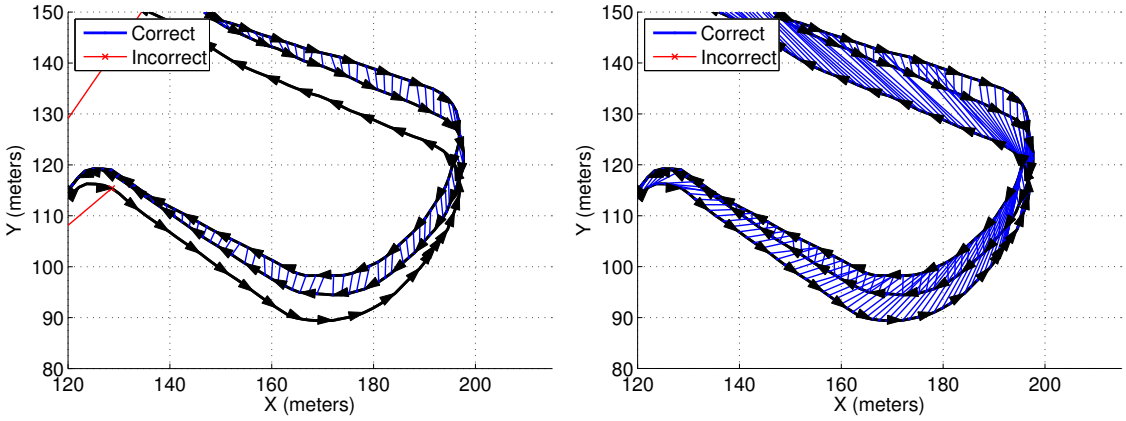
(b) Aligned: Registration.



(c) Aligned: SFT cross correlation.



(d) Aligned: Zero-phase.



(e) Zoomed in view for no-alignment (left) and zero-phase alignment (right).

Figure 15: Place recognition matches at 0.99 precision for all methods using bi-directional sequence matching. The arrows in the zoomed views in (e) show the direction of travel.

included image registration, and both SFT cross-correlation and a zero phase image alignment using the spherical harmonic expansion of the images. Given a matrix of image similarity values, local query image sequences are matched to prior database images sequences to achieve robust place recognition scoring. This matching is analogous to finding a path through the similarity matrix. We extended our existing Hidden Markov Model (HMM) sequence matching framework to enable bi-directional matching.

The combination of rotation alignment and bi-directional matching was shown to significantly improve visual place recognition performance for an outdoor panoramic image dataset. Correct place recognition results were possible even when the robot revisited the same locations traveling in the opposite direction. Of the methods tested, the SFT cross correlation was shown to provide the best recall precision performance. Results for the zero phase alignment were similar, and has the advantage of greater computational efficiency which is independent of the size of the image database.

Finally, a number of directions for future work were presented. These include exploring alternative data normalization techniques for the low bandwidth images used for rotation alignment, further development of the existing rank-reduction scheme used for place recognition scoring, and better integration of the alignment information within the HMM sequence matching framework.

References

[1] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In *Intelligent Vehicles Symposium (IV)*, June 2011.

[2] Simon Baker and Shree Nayar. A theory of single-viewpoint catadioptric image formation. *International Journal of Computer Vision*, 35(2):175–196, 1999.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[4] Thomas Bülow. Multiscale image processing on the sphere. In Luc van Gool, editor, *Proc. 24th Symposium Pattern Recognition of the DAGM*, pages 609–617, 2002.

[5] Thomas Bülow. Spherical diffusion for 3D surface smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1650–1654, December 2004.

[6] Moo K. Chung, Richard Hartley, Kim M. Dalton, and Richard J. Davison. Encoding cortical surface by spherical harmonics. *Statistica Sinica*, 18:1269–1291, 2008.

[7] Mark Cummins and Paul Newman. FAB-MAP: Probabalistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27:647–665, June 2008.

[8] Mark Cummins and Paul Newman. Appearance-only slam at large scale with FAB-MAP 2.0. *The Int. Journal of Robotics Research*, 30(9):1100–1123, 2011.

[9] Kostas Daniilidis and Christopher Geyer. Omnidirectional vision: Theroy and applications. In *Proceedings International Conference on Pattern Recognition*, 2000.

[10] James R. Driscoll and Dennis M. Healy. Computing fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, 15:202–250, 1994.

[11] 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.

[12] Christopher Geyer and Kostas Daniilidis. Equivalence of catadioptric projections and mapping of the sphere. In *Workshop on Omnidirectional Vision*, 2000.

[13] Christopher Geyer and Kostas Daniilidis. A unifying theory for central panoramic systems and practical implications. In *Proceedings European Conference on Computer Vision*, Dublin, Ireland, June 2000.

[14] Christopher Geyer and Kostas Daniilidis. Catadioptric projective geometry. *International Journal of Computer Vision*, 45(3):223–243, 2001.

[15] H. Groemer. *Geometric Applications of Fourier Series and Spherical Harmonics*. Cambridge University Press, 1996.

[16] Boris Gutman, Yalin Wang, Tony Chan, Paul Thompson, and Arthur Toga. Shape registration with spherical cross correlation. In *MICCAI Workshop on Mathematical Foundations in Computational Anatomy*, 2008.

[17] Peter Hansen and Brett Browning. Visual place recognition using HMM sequence matching. In *International Conference on Intelligent Robots and Systems*, pages 4549–4555, 2014.

[18] Peter Hansen, Peter Corke, and Wageeh Boles. Wide-angle visual feature matching for outdoor localization. *The International Journal of Robotics Research*, 29(2-3):267–297, Feb/March 2010.

[19] Peter Hansen, Peter Corke, Wageeh Boles, and Kostas Daniilidis. Scale-invariant features on the sphere. In *International Conference on Computer Vision*, 2007.

[20] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007.

[21] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g^2o: A general framework for graph optimization. In *International Conference on Robotics and Automation*, pages 3607–3613, 2011.

[22] M. Lourencǫ, J.P. Barreto, and F. Vasconcelos. sRD-SIFT: Keypoint detection and matching in images with radial distortion. *IEEE Transactions on Robotics*, 28(3):752–760, June 2012.

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

[24] Ameesh Makadia and Kostas Daniilidis. Rotation recovery from spherical images without correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1170–1175, July 2006.

[25] Ameesh Makadia and Kostas Daniilidis. Spherical correlation of visual representations for 3D model retrieval. *International Journal of Computer Vision*, 89(2-3):193–210, September 2010.

[26] Ameesh Makadia, Christopher Geyer, and Kostas Daniilidis. Correspondence-free structure from motion. *International Journal of Computer Vision*, 75(3), December 2007.

[27] Ameesh Makadia, Christopher Geyer, Shankar Sastry, and Kostas Daniilidis. Radon-based structure from motion without correspondences. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[28] Ameesh Makadia, Lorenzo Sorgi, and Kostas Daniilidis. Rotation estimation from spherical images. In *Proceedings International Conference on Pattern Recogniton*, 2004.

[29] Michael Milford. Vision-based place recognition: How low can you go? *International Journal of Robotics Research*, 32(7):766–789, 2013.

[30] Michael Milford and Gordon Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[31] Branislav Mičušík. *Two-View Geometry of Omnidirectional Cameras*. PhD thesis, Czech Technical University, Centre for Machine Perception, Prague, Czech Republic, June 2004.

[32] Shree Nayar. Catadioptric omnidirectional camera. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 482–488, San Juan, Puerto Rico, June 1997.

[33] Shree Nayar and Simon Baker. Catadioptric image formation. In *Proceedings DARPA Image Understanding Workshop*, New Orleans, 1997.

[34] Tomáš Pajdla and Václav Hlaváč. Zero phase representation of panoramic images for image based localization. In Franc Solina and Ales Leonardis, editors, *Computer Analysis of Images and Patterns*, volume 1689 of *Lecture Notes in Computer Science*, pages 550–557. Springer Berlin Heidelberg, 1999.

[35] Lawrence Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. 1993. *Prenctice Hall: Englewood Cliffs, NJ*, 2001.

[36] Davide Scaramuzza and Roland Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Transactions on Robotics*, 24(5), October 2008.

[37] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision (ICCV)*, pages 1470–1477, October 2003.

[38] Mike Smith, Ian Baldwin, Winston Churchill, Rohan Paul, and Paul Newman. The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599, May 2009.

[39] Andrew Viterbi. Error bounds for convolution codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260 – 269, 1967.