Sequence Searching with Deep-learnt Depth for Condition- and Viewpoint-invariant Route-based Place Recognition

Michael Milford, Stephanie Lowry, Niko Sunderhauf, Sareh Shirazi, Edward Pepperell, Ben Upcroft
Queensland University of Technology Australia
Australian Centre for Robotic Vision
michael.milford@qut.edu.au

Chunhua Shen, Guosheng Lin, Fayao Liu, Cesar Cadena, Ian Reid
The University of Adelaide, Australia
Australian Centre for Robotic Vision

Abstract

Vision-based localization on robots and vehicles remains unsolved when extreme appearance change and viewpoint change are present simultaneously. The current state of the art approaches to this challenge either deal with only one of these two problems; for example FAB-MAP (viewpoint invariance) or SeqSLAM (appearance-invariance), or use extensive training within the test environment, an impractical requirement in many application scenarios. In this paper we significantly improve the viewpoint invariance of the SeqSLAM algorithm by using state-of-the-art deep learning techniques to generate synthetic viewpoints. Our approach is different to other deep learning approaches in that it does not rely on the ability of the CNN network to learn invariant features, but only to produce "good enough" depth images from day-time imagery only. We evaluate the system on a new multi-lane day-night car dataset specifically gathered to simultaneously test both appearance and viewpoint change. Results demonstrate that the use of synthetic viewpoints improves the maximum recall achieved at 100% precision by a factor of 2.2 and maximum recall by a factor of 2.7, enabling correct place recognition across multiple road lanes and significantly reducing the time between correct localizations[1].

1. Introduction

Appearance or condition-invariant place recognition systems must learn and then recognize a route, even when changing time of day, weather conditions or seasons have drastically affected the appearance of the environment. The most successful condition-invariant place recognition systems (Milford and Wyeth 2012, Pepperell, Corke et al. 2013, Sünderhauf, Neubert et al. 2013, Naseer, Spinello et al. 2014) combat extreme changes in appearance by using temporal information in the form of image sequences combined with global descriptors that describe the whole image. As global descriptors are not viewpoint invariant, these approaches fail if the same place is seen from a different viewpoint (Sünderhauf, Neubert et al. 2013).

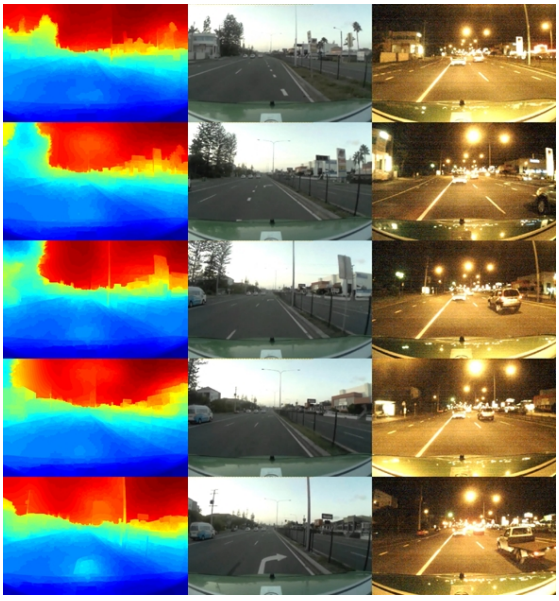

Figure 1: Sample depth images and RGB frames for a correctly matched frame sequence across a day-night cycle and two lane offset using the synthetic viewpoint system presented in this paper.

In road-based applications, viewpoint change along a route tends to come from lateral camera shifts due to lane changes. The relative difference in camera pose is thus generally limited to a few possible discrete lateral shifts in viewpoint. Recent work has proposed a solution using sideways-facing cameras to perform place recognition across lane changes (Pepperell, Corke et al. 2014). Using the assumption that the scene was planar and distant, perspective effects were ignored and image scaling alone was used to determine the likely appearance of the scene. However, the requirement for dual sideways views requires either dual mounted cameras on each car side, or external mounting of a waterproof catadioptric sensor on

[1] This research was partially supported by the ARC Centre of Excellence in Robotic Vision CE140100016 and an ARC Future Fellowship FT140101229 to MM.

the car roof; both relatively expensive solutions involving custom installs.

Our motivation in this paper is the creation of a cheap camera-based GPS module that would easily incorporate into current dashboard-mounted GPS units. To address the challenge of viewpoint change, we use deep convolutional neural fields (Liu, Shen et al. 2014) to estimate the depth of elements in the scene and generate synthetic views of the current scene at different lateral shifts to improve place recognition performance.

The paper proceeds as follows. In Section 2 we review related work, including current viewpoint- and condition-invariant systems, and state of the art depth estimation techniques. Section 3 provides an overview of our approach, comprising the estimation of scene depth from single images using a deep convolutional neural field, the generation of synthetic viewpoint images and the application of a multi-lane version of SeqSLAM to these images to perform place recognition. Section 4 describes the experimental setup and a new dataset that simultaneously tests both condition- and viewpoint invariance, with results presented in Section 5. Finally we conclude in Section 6 with a discussion of several promising areas for future research.

2. Related Work

Here we briefly review the state of the art in depth condition- and viewpoint-invariant place recognition algorithms and single frame depth estimation.

2.1. Condition-invariant place recognition

A visual place recognition system should ideally be both viewpoint-invariant and condition-invariant. Viewpoint invariance – where place recognition occurs regardless of small changes in the camera pose or orientation – can be achieved in unchanging conditions (Cummins and Newman 2008) using keypoint-based feature detectors such as SURF (Bay, Ess et al. 2008). However, when conditions change due to weather, seasonal or illumination variance, keypoint-based feature detectors are unreliable (Valgren and Lilienthal 2007, Furgale and Barfoot 2010, Valgren and Lilienthal 2010, Ranganathan, Matsumoto et al. 2013). Furthermore, visual place recognition systems may need to operate during the night or in snow-covered environments, when images display little texture and contrast, resulting in only a small number of keypoints being detected (Naseer, Spinello et al. 2014).

The most successful prior work on purely condition-invariant place recognition (Milford and Wyeth 2012, Naseer, Spinello et al. 2014) explicitly or implicitly rely on each image being described by a single, dense descriptor. In (Milford and Wyeth 2012, Milford 2013, Milford, Turner et al. 2013, Sünderhauf, Neubert et al. 2013, Hansen and Browning 2014) low-resolution gray-scale images are contrast enhanced and compared using a Sum of Absolute Differences (SAD) calculation. In (Naseer, Spinello et al. 2014), the images are tessellated into a grid of regular cells, and a HOG descriptor (Dalal and Triggs 2005) computed for each cell. Each image is then described by the combined descriptors of all the cells, and images are compared using the cosine distance. While such description techniques allow for condition-invariant localization, the trade-off is the loss of viewpoint invariance (Sünderhauf, Neubert et al. 2013).

Condition-invariant place recognition is achieved by combining these dense image descriptors with temporal information. This paper builds on SeqSLAM (Milford and Wyeth 2012), which matches sequences of images rather than individual images to achieve condition-invariant matching on very long journeys (Sünderhauf, Neubert et al. 2013), with motion blur (Milford, Turner et al. 2013), or with very low-resolution images (Milford 2013). Other approaches formulate the problem of path matching as a minimum cost flow in a data association graph (Naseer, Spinello et al. 2014) or use the Viterbi algorithm (Hansen and Browning 2014).

2.2. Depth estimation

Depth estimation from a monocular image is a challenging problem (Eigen, Puhrsch et al. 2014), particularly in unstructured, outdoor environments. Successful approaches often enforce geometric assumptions, such as box models, to infer the spatial layout of a room (Gupta, Hebert et al. 2010, Hedau, Hoiem et al. 2010) or outdoor scenes (Gupta, Efros et al. 2010), but these models are not applicable for general scene depth estimations, as they only model specific scene structures.

Non-parametric methods (Karsch, Ce et al. 2014) consist of candidate image retrieval, scene alignment and then depth inference using optimizations with smoothness constraints. This approach is based on the assumption that scenes with semantically similar appearance should have similar depth distributions when densely aligned. However, this method is prone to propagate errors through the different decoupled stages and relies heavily on building a reasonable sized image database to perform the candidate retrieval.

In recent years, efforts have been made towards incorporating additional sources of information such as user annotations (Russell and Torralba 2009) and semantic labellings (Liu, Gould et al. 2010, Ladicky, Shi et al. 2014). Depth estimation and semantic labelling can both be improved by jointly performing the two operations (Ladicky, Shi et al. 2014). However, such an approach requires hand-annotation of the semantic labels of the images beforehand as such ground-truth information is

generally not available.

Recently approaches using deep convolutional neural networks (CNNs) have been proposed. Eigen *et al.* (Eigen, Puhrsch et al. 2014) train two CNNs for depth map prediction from a single image. This method tends to learn depths with location preferences, which is prone to fit into specific layouts, and requires a large number of labelled data to cover all possible layouts for training the networks.

In (Tompson, Jain et al. 2014), Tompson *et al.* present a hybrid architecture for jointly training a deep CNN and an MRF for human pose estimation. A unary term and a spatial model are trained separately and then jointly learned as a fine tuning step.

This paper uses the approach from (Liu, Shen et al. 2014), which learns a deep CNN for constructing unary and pairwise potentials of conditional random fields (CRFs). As depth estimation can naturally be formulated as a CRF learning problem (Liu, Shen et al. 2014) this approach has shown state-of-the-art performance on both indoor and outdoor scene depth estimations. Furthermore, as the unary potentials do not contain coordinate information and thus do not learn location preferences this method, in contrast to (Eigen, Puhrsch et al. 2014), can be trained on a standard dataset without requiring additional training data.

3. Approach

Our approach uses a deep convolution neural field model to estimate depth from individual frames in order to generate synthetic viewpoints representing lateral camera shifts. These synthetic viewpoints are processed in parallel by a multi-viewpoint version of SeqSLAM to generate overall place recognition hypotheses.

3.1. Depth Estimation

The system uses a deep convolutional neural field model for depth estimation (Liu, Shen et al. 2014). This approach uses a conditional random field (CRF) which for an image x models the conditional probability of each superpixel depth y by the density function:

$$\Pr(y \mid x) = \frac{1}{Z(x)} \exp(-E(y,x)) \tag{1}$$

The function E is the energy function and is defined as a combination of unary potentials U and pairwise potentials V over the superpixels in N and edges S in x.

$$E(y,x) = \sum_{p \in N} U(y_p, x) + \sum_{(p,q) \in S} V(y_p, y_q, x) \tag{2}$$

Z is the partition function:

$$Z(x) = \int_y \exp(-E(y,x)) \mathrm{d}y \tag{3}$$

A unified deep convolutional neural network (CNN) framework learns the value of U and V. This framework is shown in Figure 2 and is composed of a unary part, a pairwise part and a CRF loss layer, and is trained using back propagation. The unary potential is constructed from the output of a CNN by considering the least square loss, while the pairwise potential enforces smoothness by exploiting consistency information of neighboring superpixels.

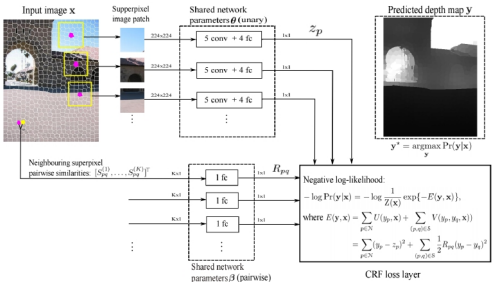

Figure 2: The deep convolutional neural field model for depth estimation (from (Liu, Shen et al. 2014)).

The depth of a new image is predicted using the closed-form maximum a posteriori (MAP) inference:

$$y^* = \arg\max_y \Pr(y \mid x) \tag{4}$$

The network training is based on the VLFeat MatConvNet CNN toolbox (Vedaldi 2013). Training is done on a standard desktop with an NVIDIA GTX 780 GPU with 6GB memory.

During implementation, the first 6 layers of the unary part are initialized using a CNN model trained on the Make3D dataset. These layers are kept fixed and the rest of the network is *pre-trained* with the following settings: momentum is set to 0.9, and weight decay parameters λ_1, λ_2 are set to 0.0005. During pre-training, the learning rate is initialized at 0.0001, and decreased by 40% every 20 epochs. The pre-training takes a total of 60 epochs (with the learning rate decreased twice).

Figure 3: Sample images from the Make3D dataset.

The whole network is then trained with the same momentum and weight decay, using a dropout probability of 0.5 in the first two fully-connected layers. When training the whole network, ~700 superpixel patches need to be processed per image, and each image takes around 10s to process.

Applying the depth prediction model is computationally efficient. The depth prediction involves 3 major steps: super-pixel generation, CRF pairwise input feature generation and the CNN forward step of our model. The network forward step only takes around 0.25 seconds for an input image with 640×480 pixels running on a GPU (NVIDIA Titan), providing the potential for low frame-rate real-time operation on a vehicle.

3.2. Synthetic Viewpoint Generation

Depth masks are used to generate a range of synthetic viewpoints representing lateral shifts (i.e. lane changes) for the vehicle. Image warping is performed using an crude simple inverse depth model – pixel shifts being inversely proportionate to their depth.

Due to the higher accuracy of day-time depth masks, we generate synthetic day-time viewpoints and match them to existing night-time images. Figure 4 shows an example of a day-time image and associated depth mask, accompanied by the warped synthetic view and the night-time image from the same location. Note that the result is far from perfect; amongst other factors, depth compression (the dynamic depth range of the image being smaller than reality) means that the nearby sections of road are not transformed nearly enough. We generated a total of 3 viewpoints approximately covering the range of expected viewpoint variation within the dataset.

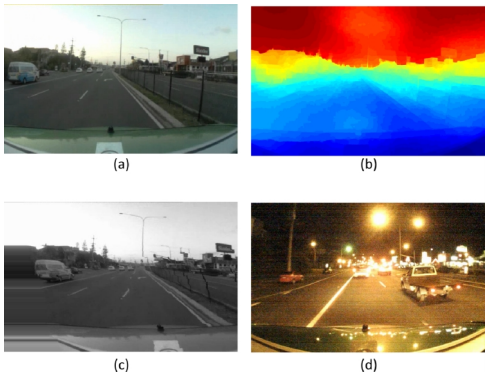

Figure 4: (a) Day-time image, (b) associated depth mask, (c) synthetic sideways shifted viewpoint and (d) corresponding night-time frame from the same location. Note the nearby roadway is not transformed sufficiently due to depth compression issues.

3.3. SeqSLAM

In this section we describe SeqSLAM, which has local best match and sequence recognition components, and the process by which multiple viewpoint streams are processed in parallel.

SeqSLAM frames the image recognition problem not as one of finding the single template that best matches the current image (global best match), but rather as one of finding all the templates within local neighborhoods that are the best match for the current image (local best match). We apply a local contrast enhancement (analogous to a 1D version of patch normalization) process to each element i in the image difference vector $\mathbf{D}$ to produce a contrast enhanced image difference vector $\hat{\mathbf{D}}$:

$$\hat{D}_i = \frac{D_i - \overline{D_l}}{\sigma_l} \tag{5}$$

where $\overline{D_l}$ is the local mean and σ_l is the local standard deviation, in a range of R_{window} templates around template i. Figure 5 shows a schematic of the local contrast enhancement process operating on a number of $\hat{\mathbf{D}}$ vectors calculated at different times.

3.3.1 *Local Best Match*

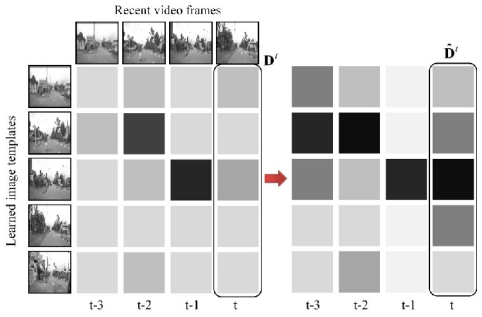

Figure 5: Contrast enhancement of the original image difference vectors increases the number of strongly matching templates. Darker shading = smaller image difference = stronger match.

3.3.2 *Localized Sequence Recognition*

Localized image template matching produces a number of candidate template matches at each time step. To recognize familiar place sequences, a search is performed through the space $\mathbf{M}$ of recent image difference vectors:

$$\mathbf{M} = \left[\hat{\mathbf{D}}^{T-d_s}, \hat{\mathbf{D}}^{T-d_s+1} \ldots \hat{\mathbf{D}}^{T} \right] \tag{6}$$

where d_s determines how far back in time the search goes, and T is the current time. Because car-based scenarios have an accurate source of translational odometry with which frame spacing can be normalized, the search is drastically simplified to a set of 45 degree diagonal trajectories through the difference matrix. A difference score S is calculated for the search trajectory line based on the difference values the line passes through in travelling from time T-d_s to the current time T:

$$S = \sum_{t=T-d_S}^{T} D_k^t \tag{7}$$

where k is the particular difference value the trajectory passes through at time t:

$$k = s + V(d_s - T + t) \quad (8)$$

where s is the template number the trajectory originated in, and V is the unitary trajectory velocity due to the odometry normalization.

After all the trajectory scores have been evaluated, the minimum scoring (i.e. best matching) trajectory for each template is placed in vector **S**. We generate precision-recall performance curves by sweeping a matching threshold over the vector **S** of scores.

3.3.3 Multi-Viewpoint Integration

We generate parallel difference matrices for each synthetic viewpoint and sequence search each matrix independently. The best-matching sequence scores at each point in time are found by taking the minimum sequence matching score across all difference matrices representing each viewpoint.

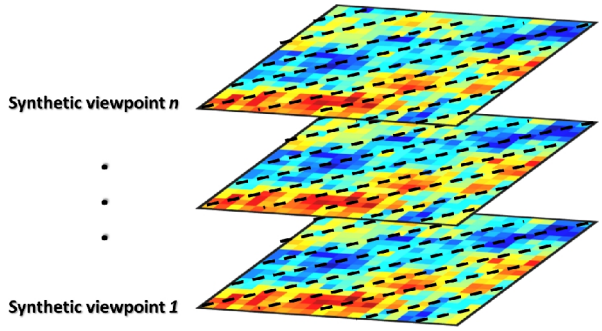

Figure 6: Searching for matching sequences across the difference matrices representing each synthetic viewpoint.

4. Experimental Setup

Experiments were run using a new dataset consisting of two 1.1 km day and night car traverses along a multi-lane road (Figure 7). The night traverse was obtained entirely in one lane, while the day traverse ranged approximately evenly over three lanes, resulting in lateral viewpoint changes of zero, one and two lanes. Raw footage was taken at 30 fps using a forward facing 752×480 pixel Micron camera (model MT9V022I77ATC, approximately 80 degree horizontal FOV). Odometry was collected with an OBDPro USB Scantool and a laptop computer.

4.1. Ground Truth

Ground truth was determined by synchronising the start and end points of the dataset videos and interpolating in-between frame correspondences using odometry, with verification by manual inspection. Ground truth can be considered correct to within approximately 3 metres, and we used an accuracy threshold of 10 metres to generate the precision-recall performance curves.

4.2. Image Pre-Processing and Comparison

SeqSLAM requires an individual image comparison method. In this work, we used a Sum of Absolute Differences (SAD) calculation on resolution reduced, patch-normalized images (Fig. 8) to produce an image difference score d:

$$d = \frac{1}{R_x R_y} \sum_{x=0}^{R_x} \sum_{y=0}^{R_y} \left| A_{x,y} - B_{x,y} \right| \quad (9)$$

where R_x and R_y are the dimensions of the resolution reduced image, and A and B are matrices containing the grayscale pixel intensity values for the two images being compared. Patch-normalization subtracts the local mean intensity and then divides by the local intensity standard deviation. Image comparison was performed over sliding window (x_{offset}, y_{offset}) to provide some additional viewpoint invariance. Cropping was necessary to overcome the poor depth estimation for the road just in front of the car.

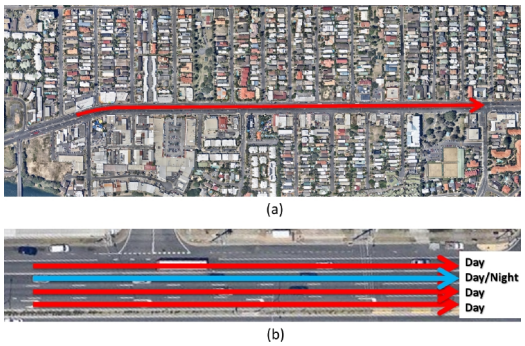

Fig. 7: Experiments were run using a dataset consisting of day and night car traverses along a multi-lane road. The night traverse was obtained entirely in one lane, while the day traverse ranged approximately evenly over three lanes including the same lane as traversed during the day.

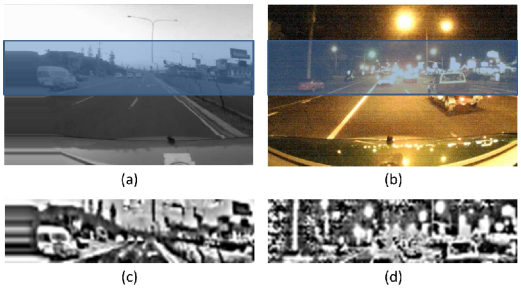

Fig. 8: Original images and low resolution patch-normalized versions used by SeqSLAM

To simplify the difference matrix search, we use odometry to provide a constant spatial frame separation of 1 metre, as in the SMART modification (Pepperell, Corke et al. 2014) of the SeqSLAM algorithm. In contexts where relatively-accurate wheel-based odometry is not available, visual odometry could be used to the same effect (Milford, Vig et al. 2014).

4.3. Parameter Values

TABLE I
PARAMETER LIST

Parameter	Value	Description
R_x, R_y	128,32	Whole image matching resolution
P_{size}	4	Patch-normalization radius
x_{offset}	4	Sliding image comparison window (horizontal)
y_{offset}	2	Sliding image comparison window (vertical)

5. Results

In this section we present results showing the effect on place recognition performance of using deep learning to generate synthetic scene viewpoints. We present precision-recall curves comparing vanilla SeqSLAM performance (using a sliding offset window only) to the depth-based system, and frame correspondence graphs showing the distribution and coverage of correct and incorrect place matches over the entire dataset. Finally we show example frames from correctly and incorrectly matched sequences.

Using vanilla SeqSLAM with image offsetting, a maximum recall of 20% recall at 100% precision is achieved (Figure 9). Allowing precision to drop below 100% results in only a minimal increase up to a maximum recall achieved of 25.6%. Note we are using the more conservative recall calculation where:

$$R = \frac{TP}{\text{total expected matches}} \tag{10}$$

Examining the frame correspondence graph shown in Figure 10, the distribution of matches is primarily centered on a region where the day and night traverses were obtained in the same lane, hence removing the requirement for viewpoint invariance. Correct localization coverage is discontinuous, with the maximum period of incorrect localization being approximately 260 metres in length.

When the synthetic viewpoint system is used, the system achieves a maximum recall of 43.7% at 100% precision, an improvement of 220% over the vanilla SeqSLAM case. Allowing precision to drop below 100% results in an increase up to a maximum recall achieved of 68.9%, with more continuous correct place recognition coverage over the entire dataset. The maximum period of incorrect localization is approximately 118 metres.

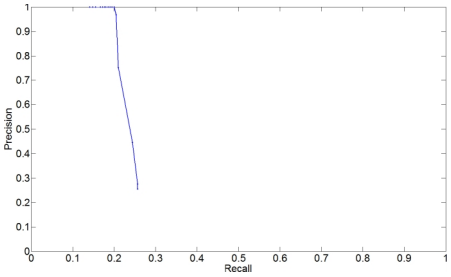

Figure 9: Using vanilla SeqSLAM with image offsetting, a maximum recall of 20% recall at 100% precision is achieved.

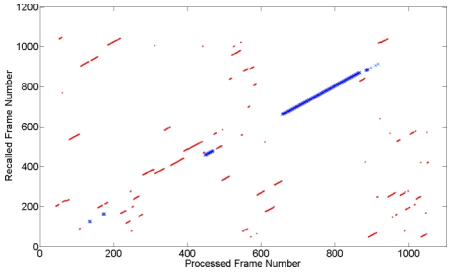

Figure 10: Using vanilla SeqSLAM, the distribution of correct place matches are primarily centered on a region where the day and night traverses were in the same lane, hence removing the requirement for viewpoint invariance (blue cross = true positive place match, red dot = false positive).

Figures 1 and 13-14 show frames from sample sequence matches. Each column shows 5 frames evenly distributed over the 100 frame matching sequence. Figure 13 shows a sequence with one lane viewpoint change that is correctly matched using the synthetic viewpoint system but not vanilla SeqSLAM, while Figure 1 shows a sequence with two lanes viewpoint change that is correctly matched using the synthetic viewpoint system but not vanilla SeqSLAM. Despite the performance improvements, the synthetic viewpoint approach is not perfect; Figure 14 shows an incorrect sequence match; both the synthetic and vanilla methods incorrectly matched this sequence, perhaps due to environmental aliasing.

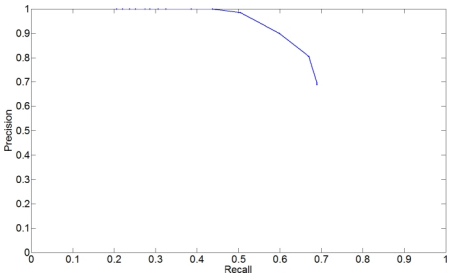

Figure 11: When the synthetic viewpoint system is used, the system achieves a maximum recall of 43.7% at 100% precision, an improvement of 220% over the vanilla SeqSLAM case.

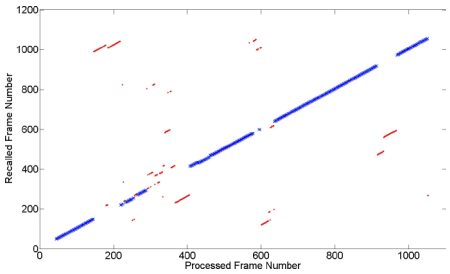

Figure 12: More continuous correct place recognition coverage is achieved over the entire dataset when using synthetic views. The maximum period of incorrect localization is approximately 118 metres.

Figure 15a shows the selected synthetic viewpoints that resulted in the best place matching scores throughout the dataset. Figure 15b shows the ground truth relative lane shift between the night-time traverse and the day-time traverse throughout the dataset. It is interesting to observe that despite periods of "incorrect" viewpoint selection (e.g. between frames 250 and 300 where the system selects the original camera viewpoint whereas there is actually a 2 lane viewpoint change, the system is still able to achieve some correct place recognition matches (Figure 12).

6. Discussion and Future Work

The results presented here demonstrate the potential for using state of the art deep learning-based depth estimation techniques to provide condition-invariant place recognition techniques with more viewpoint invariance. Here we discuss the outcomes and several areas of promising future research.

The depth estimation model performed well on predicting relative depth information instead of predicting an accurate depth map in terms of quantitative measure. However, the estimated depth map sometimes had depth compression issues; the variation in range to the road just in front of the vehicle to near the horizon was too small. There are several possible reasons for this issue. Firstly, ground truth in the training dataset is not accurate especially on distant regions. Secondly, road scene images are not common in the Make3D dataset; it is likely that training our model on more dedicated outdoor road scene images will help improve the performance. As future work, we will collect a larger outdoor RGBD dataset. Finally, the trained model may have had difficulties predicting the vanish direction that goes to infinity, e.g., a road that vanishes at the horizon. A possible solution to this problem is to estimate the surface normals simultaneously.

For the specific scenario of camera-based car localization along a road, one obvious area of future research is to add temporal information into the depth learning model, possibly through some form of model recurrency. Other conventional depth techniques could also be investigated, such as using conventional structure from motion estimates. Ultimately some form of boosting framework incorporating all available depth information would likely result in the best overall performance.

Figure 13: Sample correct place recognition sequence at a one lane offset using the synthetic viewpoint system. This sequence is not correctly matched using vanilla SeqSLAM.

Figure 14: Sample sequence incorrectly matched by both the synthetic viewpoint system and vanilla SeqSLAM.

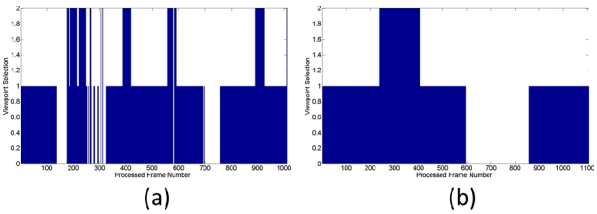

Figure 15: (a) Selected synthetic viewpoints that resulted in the best place matching scores throughout the dataset and (b) ground truth lane viewpoint changes throughout the dataset.

References

[1] Bay, H., A. Ess, T. Tuytelaars and L. Van Gool (2008). "Speeded-Up Robust Features (SURF)." Computer Vision and Image Understanding **110**(3): 346-359.

[2] Cummins, M. and P. Newman (2008). "FAB-MAP: Probabilistic localization and mapping in the space of appearance." International Journal of Robotics Research **27**(6): 647-665.

[3] Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE.

[4] Eigen, D., C. Puhrsch and R. Fergus (2014). Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. Advances in Neural Information Processing Systems (NIPS).

[5] Furgale, P. and T. Barfoot (2010). "Visual Teach and Repeat for Long-Range Rover Autonomy." Journal of Field Robotics **27**(5): 534-560.

[6] Gupta, A., A. A. Efros and M. Hebert (2010). Blocks world revisited: Image understanding using qualitative geometry and mechanics. Computer Vision–ECCV 2010, Springer**:** 482-496.

[7] Gupta, A., M. Hebert, T. Kanade and D. M. Blei (2010). Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. Advances in Neural Information Processing Systems.

[8] Hansen, P. and B. Browning (2014). Visual place recognition using HMM sequence matching. Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on.

[9] Hedau, V., D. Hoiem and D. Forsyth (2010). Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. Computer Vision – ECCV 2010. K. Daniilidis, P. Maragos and N. Paragios, Springer Berlin Heidelberg. **6316:** 224-237.

[10] Karsch, K., L. Ce and K. Sing Bing (2014). "Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling." Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(11): 2144-2158.

[11] Ladicky, L., J. Shi and M. Pollefeys (2014). Pulling things out of perspective. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE.

[12] Liu, B., S. Gould and D. Koller (2010). Single image depth estimation from predicted semantic labels. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE.

[13] Liu, F., C. Shen and G. Lin (2014) "Deep Convolutional Neural Fields for Depth Estimation from a Single Image." ArXiv e-prints **1411**, 6387.

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

[15] Milford, M., I. Turner and P. Corke (2013). Long exposure localization in darkness using consumer cameras. Proceedings of the 2013 IEEE International Conference on Robotics and Automation, IEEE.

[16] Milford, M., E. Vig, W. Scheirer and D. Cox (2014). "Featureless Visual Processing for SLAM in Changing Outdoor Environments." Journal of Field Robotics **31**(5).

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

[18] Naseer, T., L. Spinello, W. Burgard and C. Stachniss (2014). Robust visual robot localization across seasons using network flows. Conference on the Association for the Advancement of Artificial Intelligence.

[19] Pepperell, E., P. Corke and M. Milford (2013). Towards persistent visual navigation using SMART. Proceedings of Australasian Conference on Robotics and Automation. University of New South Wales, Sydney, Australia, ARAA.

[20] Pepperell, E., P. Corke and M. Milford (2014). "Towards Vision-Based Pose-and Condition-Invariant Place Recognition along Routes." Proceedings of the Australasian Conference on Robotics and Automation 2014.

[21] Ranganathan, A., S. Matsumoto and D. Ilstrup (2013). Towards illumination invariance for visual localization. Robotics and Automation (ICRA), 2013 IEEE International Conference on.

[22] Russell, B. C. and A. Torralba (2009). Building a database of 3d scenes from user annotations. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE.

[23] Sünderhauf, N., P. Neubert and P. Protzel (2013). Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA).

[24] Tompson, J. J., A. Jain, Y. LeCun and C. Bregler (2014). Joint training of a convolutional network and a graphical model for human pose estimation. Advances in Neural Information Processing Systems.

[25] Valgren, C. and A. Lilienthal (2007). SIFT, SURF and Seasons: Long-term Outdoor Localization Using Local Features. European Conference on Mobile Robotics (ECMR).

[26] Valgren, C. and A. Lilienthal (2010). "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments." Robotics and Autonomous Systems **58**(2): 157-165.

[27] Vedaldi, A. (2013). "MatConvNet." from http://www.vlfeat.org/matconvnet/.