Enhancing Place Recognition using Joint Intensity - Depth Analysis and Synthetic Data

Elena Sizikova[1], Vivek K. Singh[2], Bogdan Georgescu[2], Maciej Halber[1], Kai Ma[2], and Terrence Chen[2]

[1]Department of Computer Science, Princeton University
[2]Medical Imaging Technologies, Siemens Medical Solutions Inc., Princeton, NJ

Abstract. Visual place recognition is an important tool for robots to localize themselves in their surroundings by matching previously seen images. Recent methods based on Convolutional Neural Networks (CNN) are capable of successfully addressing the place recognition task in RGB-D images. However, these methods require many aligned and annotated intensity and depth images to train joint detectors. We propose a new approach by augmenting the place recognition process with individual separate intensity and depth networks trained on synthetic data. As a result, the new approach requires only a handful of aligned RGB-D frames to achieve a competitive place recognition performance. To our knowledge, this is the first CNN approach that integrates intensity and depth into a joint robust matching framework for place recognition and that evaluates utility of prediction from each modality.

1 Background

Visual place recognition is a task of detecting when two images in an image sequence depict the same location, possibly under camera viewpoint or illumination-related appearance changes [14]. This is a challenging problem in computer vision that is particularly important for intelligent autonomous robot systems. For instance, such systems include (but are not limited to) robots that need to map their positions in space, accurately localize themselves within their environment, and detect when they revisit a previous location. Typically, matches are determined based on similarity of image pairs from widely available and inexpensive RGB-D sensors, such as Kinect. This task is particularly challenging since the RGB appearance of a surface can vary dramatically with viewpoint and lighting. Moreover, depth appearance can have dropouts, noise, and other artifacts that hinder the extraction of repeatable features. Finally, long traversal trajectories contain thousands of scans which require efficient image search strategies.

Convolutional Neural Networks (CNN) have been recently shown to outperform methods relying on handcrafted features in place recognition tasks [1]. Furthermore, [10] and [20] compare performance of convolutional network layers, and conclude that the middle layers of networks trained for related tasks such as semantic place categorization [25], are especially suitable to address this problem. However, such methods are effective only in scenes with viewpoint-independent surface appearances and where training data with intensity image correspondences are available.

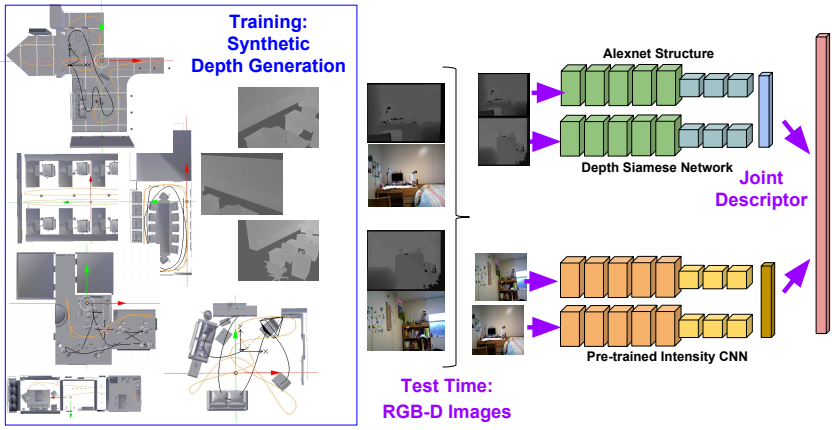

Fig. 1. Overview of the joint depth intensity CNN place recognition system. We use synthetic 3D models of office and living rooms [6] to train a depth Siamese CNN network. We then learn to match the network feature responses on unseen real depth images with intensity feature responses from a pre-trained CNN [25] to obtain a viewpoint-invariant RGB-D descriptor for place recognition.

Due to illumination changes, color information can be ambiguous, which leads to a natural question of how effectively depth information can be used to resolve ambiguities. Previous studies indicate that depth is somewhat inferior to intensity for place recognition [17]. However, such findings are based on a Bag of Words model [13], which produces descriptors that are, in turn, hand-tuned rather than trained on data. Recently, [5, 8, 9] showed that joint RGB-D networks outperform intensity only methods for object detection. Depending on the definition of a 'place match', the problem of place recognition often does not have a sufficient amount of annotated and aligned RGB-D training data to train joint CNN models, and thus it is unclear how much each available modality contributes to performance. At the same time, synthetic depth data was shown to be very useful for dense semantic labelling [6] and for object detection [8], where it is complementary to intensity, but often cannot be directly used to improve joint models due to lack of RGB annotations.

In this paper, we investigate the performance of a depth CNN trained on synthetic depth images for an indoor place recognition task. We obtain training data by synthesizing depth images from computer graphics models of scenes [6], by simulating camera movement along user generated realistic movement trajectories. At training time, the network learns to predict 3D overlap in synthetic depth; while at testing time, it is used to evaluate overlap on real data, in combination with a RGB-D CNN descriptor. We combine the two modalities in a robust matching framework and evaluate the relative contribution of each method using robust statistical analysis. Finally, we release a large dataset of synthetic trajectories with per-frame extrinsics annotations.

2 Approach

We introduce a CNN for matching pairs of RGB-D images that can be used to enhance place recognition under limited data availability. Although CNNs are known to be effective for RGB image matching [21], there are no joint RGB-D indoor place recognition systems. This is due to the fact that existing indoor datasets such as [18, 23] lack the per-frame annotations and RGB-Depth sensor alignments which are necessary to train a larger model.

To address this issue, we synthesize a set of trajectories in computer graphics models of rooms that are then used to create synthetic depth frames. We train a CNN on pairs of the synthetic depth images to learn a match predictor (depth images that overlap significantly). Finally, at testing time, we combine that CNN with one trained on RGB images to recognize matches between real RGB-D images, and evaluate the statistical contribution of each component.

The key idea behind this approach is that synthetic 3D models can be used to produce sufficient amount of depth data to train a CNN, and that it is not necessary to have access to a large collection of aligned RGB and depth images to construct a joint predictor. Although a similar methodology has been used previously for object detection [5, 8, 16, 22], it has never been employed for indoor place recognition.

3 Overview

We evaluate the capacity of joint intensity and depth place recognition method in improving recognition of previously visited places. Our input data consists of a sequence of RGB-D scans, where the correspondences between depth and RGB are known. The scans are acquired as the robot moves along a trajectory and its sensors periodically take snapshots of the current state.

In the next section, we describe the details of network training and descriptor extraction from each RGB-D frame.

3.1 Synthetic Dataset

We employ scenes from the SceneNet dataset [6] to draw custom trajectories and to generate synthetic depth scans to be used as training data for the network. Overall, we generated 134 unique trajectories for the bathroom, living room, and office scenes, where each trajectory circles the room several times to create a variety of realistic loop closure examples. Each trajectory is created by drawing two Bezier curves that circle around a room (see Figure 1). The first curve represents the camera location, and the second curve represents a set of points that the camera observes. Camera movement is animated along these paths, producing a sequence of camera poses and corresponding 640×480-pixel depth frames simulated from a pinhole camera. The trajectories are intended to mimic the way an experienced user would scan the room – i.e., the motion is smooth, every scene object is viewed from multiple viewpoints, with varying time spent in different parts of the scene. The resulting trajectories contain 3,000 camera viewpoints each, and we use every 10-th viewpoint for training.

3.2 Depth Descriptor

Depth descriptors are extracted by calculating depth features from a CNN trained on depth data. Since our goal is to detect same objects under different viewpoints, we use 3D point overlap as a similarity estimator. Note that this is a more challenging metric than translation along the camera trajectory (which is often used for evaluation in place recognition systems), because the same overlap amount allows larger camera pose differences between frames. However, since selected frames are typically passed to a geometric alignment algorithm to generate pairwise transformations, a large amount of overlap is a suitable predictor that the correct transformation will be predicted.

We compute overlap O for each scan pair (i, j) as $O = (P_i \cap P_j)/\max\{P_i, P_j\}$, where P_i and P_j are sets of 3D points in each frame, and the union is the set of points within a threshold ϵ of each other ($\epsilon = 7.5$ cm in all experiments). Normalizing overlap by the maximum number of points ensures that if one of the scans captures a small portion of the other, a case which is visually ambiguous, this example is not selected for training.

Training Setup Each example given to the network for training is a pair of depth images and a label. We generate the set of all pairs of depth images for each trajectory. Each example is assigned a positive label if the 3D overlap between these images is greater than T; otherwise a negative label is assigned. This threshold represents our confidence that such a network yields visually similar pairs, but with sufficient room for viewpoint changes. In our experiments, we find that choosing T of 75% yields the most visually recognizable selections as determined by an observer. We also balance the number of positive and negative examples from each training trajectory. Overall, the training data consists of $1,442,252$ depth image pairs, where depth is encoded using the HHA encoding [4], which is known to be compatible with popular existing CNN architectures. Training process is performed in the Caffe library [11].

Network Architecture and Loss We use a Siamese CNN architecture [3] with two Alexnet branches to learn pose-invariant descriptors. The Siamese architecture equipped with a Contrastive Loss function is known to be especially suitable for distance metric learning from examples [3]. We compare the outputs of several bottom layers and obtain the best performance from the $fc6$ layer which we then use for joint descriptor calculations. To evaluate the quality and amount of our synthetic depth data, we train the network from scratch using stochastic gradient descent with $100,000$ iterations.

3.3 Intensity Descriptor

We employ a pre-trained CNN for descriptor extraction on intensity images, trained for semantic place categorization [25]. In particular, we use an AlexNet CNN trained on 205 scene categories of Places Database (2.5 million images) [25]. We also compare the Caffenet implementation [11]; however, we find PlacesNet layer $fc6$ to be superior (see Supp. Material), echoing conclusion of [20].

3.4 Learning a Joint Descriptor

Given two RGB-D frames, our goal is to estimate their distance. Because the distances between only the depth or only the intensity parts may be unreliable, we combine the distances from both modalities using a robust joint model, which we describe below. Given a pair of frames F_p and F_q, we start by extracting depth and intensity descriptors, (d_p, i_p) and (d_q, i_q), respectively. We then use a small set of aligned RGB-D frames to estimate the joint parameters of the model. In particular, let $D = ||d_p - d_q||$ and $I = ||i_p - i_q||$ be distances in the depth and intensity descriptor spaces, respectively. Our goal is to estimate overlap O as a function of D and I. We consider two models where O is either a linear or polynomial function of D and I, that is, $O \sim \mathrm{D}+\mathrm{I}$ (1) and $O \sim \mathrm{D}+\mathrm{I}+\mathrm{D}\times\mathrm{I}$ (2). Both models are highly sensitive to atypical observations and outliers. To reduce outlier impact, we use a robust MM-type estimator, which is known to deliver highly robust and efficient estimates, to obtain model parameters [24, 15, 12]. Given aligned data from any trajectory (we used ICL Living Room Sequence 3), we estimate coefficients of (2) using 100 bootstrap iterations (i.e. sampling with replacement). In each polynomial model, we balance the number of overlapping and non-overlapping pairs. We describe the statistical properties of these models and their evaluation on a place recognition task in the next section.

4 Evaluation

Model Assessment Our analysis indicates that both linear terms (Depth and Intensity) and the crossterm ($\mathrm{D} \times \mathrm{I}$) are highly statistically significant), and thus all three terms contain meaningful addition in explaining variation of O. The relationship follows our hypothesis that overlap decreases with increasing distances between the intensity or between depth images (we provide values of the coefficients and standard deviations in Suppl. Material). In addition, we compare the mean adjusted R^2 values from models $O \sim D$ (density only), $O \sim I$ (intensity only), $O \sim I + D \times I$ (intensity and depth, non-robust), and $O \sim I + D \times I$ (intensity and depth, robust), which result in the adjusted R^2 values of 0.30, 0.68, 0.69, and 0.70, respectively. These findings indicate importance of each component to explain variability of our system and superiority of the polynomial model $O \sim I + D \times I$. Additionally, the mean coefficients from the robust model perform well across all test datasets, while the mean coefficients from the non-robust model perform substantially worse (see Supp. Material). These findings indicate sensitivity of a conventional estimators to outliers and importance of using a more robust MM-estimator.

Place Recognition Results The goal of our work is to provide a robust view-invariant RGB-D place recognition descriptor, and we evaluate of our method on trajectories from three publicly available benchmark datasets, namely, ICL-NUIM [7], TUM RGB-D [19], and Sun3D [23]. All datasets are pre-processed in the same way as the training dataset to obtain 3D overlap pairs (subsampled to every 10th frame and depth converted to HHA encoding). The ability of descrip-

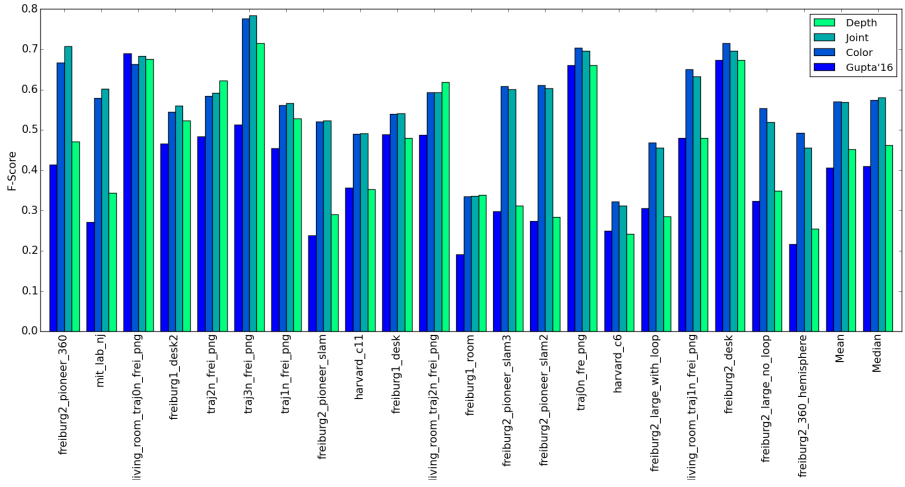

Fig. 2. Best F-scores obtained by each model (under its optimal parameters).

tors to recognize the same location can be evaluated by F-scores. We calculate the precision and recall at equally spaced thresholds between the smallest and the largest descriptor distances of all pairs of scans, for each combination of method and dataset (so that predicted positives are pairs whose descriptor distances are below this threshold, selected among all non-consecutive scan pairs). In each dataset, the true positive pairs are those pairs of frames which have a small overlap in 3D space (less than 30%, the threshold used in geometric alignment algorithms [2]). The F-scores are calculated from precision and recall for all thresholds and for all methods, and the top score is selected for each method.

Figure 2 shows the top F-scores of each method in each dataset. The joint object detection method of [5] consistently ranks lower than both Placesnet [25] and DepthAlexnet (synthetic data only), which is expected without additional fine-tuning. DepthAlexnet outperforms Placenet [25] in 4/20 cases, and the joint model outperforms Placesnet [25] in 12/20 cases, and performs comparably well in others. Interestingly, DepthAlexnet outperforms method of [5] in many cases, suggesting that depth can be a good predictor on its own when sufficient synthetic data is available for training.

5 Conclusion and Future Work

We propose a novel outlier resistant place recognition descriptor in RGB-D. We show that synthetic depth can be employed to train view-invariant CNNs that are useful for place recognition tasks. We also show that combining descriptors from depth and intensity images shows improvement over intensity-only based place recognition, even when only a few aligned RGB-D trajectories are available for training.[1]

[1] Disclaimer: The outlined concepts are not commercially available. Due to regulatory reasons their future availability cannot be guaranteed.

6 Acknowledgements

We thank Thomas Funkhouser and the Princeton Graphics group for research discussions. Finally, we thank the National Science Foundation (NSF-GRFP) for support.

References

1. Chen, Z., Lam, O., Jacobson, A., Milford, M.: Convolutional neural network-based place recognition. In: arXiv (2014)
2. Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: CVPR (2015)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
4. Gupta, S., Girshick, R., Arbelez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV (2014)
5. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016)
6. Handa, A., Pătrăucean, Viorica, Badrinarayanan, V., Stent, S., Cipolla, R.: Understanding real world indoor scenes with synthetic data. In: CVPR (2016)
7. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: ICRA (2014)
8. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: CVPR (2016)
9. Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., Darrell, T.: Cross-modal adaptation for rgb-d detection. In: ICRA (2016)
10. Hou, Y., Zhang, H., Zhou, S.: Convolutional neural network-based image representation for visual loop closure detection. In: arXiv (2015)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACMMM (2014)
12. Koller, M., Stahel, W.A.: Sharpening wald-type inference in robust regression for small samples. Computational Statistics and Data Analysis (2011)
13. Konolige, K., Bowman, J., Chen, J., Mihelich, P., Calonder, M., Lepetit, V., Fua, P.: View-based maps. The International Journal of Robotics Research (2010)
14. Lowry, S., Sunderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: A survey. IEEE Transactions of Robotics (2016)
15. Matas Salibian-Barrera, V.J.Y.: A fast algorithm for s-regression estimates. Journal of Computational and Graphical Statistics (2006)
16. Papon, J., Schoeler, M.: Semantic pose using deep networks trained on synthetic rgb-d. In: ICCV (2015)
17. Scherer, S.A., Kloss, A., Zell, A.: Loop Closure Detection using Depth Images. In: ECMR (2013)
18. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
19. Sturm, J., Magnenat, S., Engelhard, N., Pomerleau, F., Colas, F., Cremers, D., Siegwart, R., Burgard, W.: Towards a benchmark for rgb-d slam evaluation. In: RSS RGB-D Workshop on Advanced Reasoning with Depth Cameras (2011)

20. Sünderhauf, N., Dayoub, F., Shirazi, S., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: arXiv (2015)
21. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR (2014)
22. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: CVPR (2015)
23. Xiao, J., Owens, A., Torralba, A.: SUN3D: A database of big spaces reconstructed using SfM and object labels. In: ICCV (2013)
24. Yohai, V.J.: High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics (1987)
25. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)