IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554047

METRIC LEARNING FOR 2D IMAGE PATCH AND 3D POINT CLOUD VOLUME MATCHING

Baiqi Lai[1], Weiquan Liu[1], Cheng Wang[1], Shuting Chen[2], Xuesheng Bian[1], Xiuhong Lin[1], Chenglu Wen[1], Jonathan Li[3]*

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
[2]Information Engineering School, Chengyi University College, Jimei University, Xiamen, China
[3]Department of Geography and Environmental Management University of Waterloo, Waterloo, Canada
[*]Corresponding author: wqliu@xmu.edu.cn

ABSTRACT

Similarity measure of cross-domain descriptors (2D descriptors and 3D descriptors) between 2D image patches and 3D point cloud volumes provides stable retrieval performance and establishes the spatial relationship between 2D and 3D space, which plays the potential applications in geospatial space, such as 2D and 3D interaction of remote sensing, Augmented Reality (AR) and robot navigation. However, the mature handcrafted descriptors of 2D image patches and 3D point cloud volumes are extremely different, resulting in the huge challenge for 2D image patch and 3D point cloud volume matching. In this paper, we propose a novel network which combines both unified descriptor training and descriptor comparison function training for 2D image patch and 3D point cloud volume matching. First, two feature extraction networks are applied for jointly learning the local descriptors for 2D image patches and 3D point cloud volumes, respectively. Second, a fully connected network is introduced to compute the similarity between 2D descriptors and 3D descriptors. Motivated by the successful indicator system on evaluating 2D patch feature representation, we use the false positive rate at 95% recall (FPR95) and precision based on cross-domain descriptors as the measured metric. The experimental results show that our proposed network achieve state-of-the-art performance in the matching of 2D image patches and 3D point cloud volumes.

Index Terms— cross-domain, 2D and 3D matching, metric learning, image patch, point cloud volume

1. INTRODUCTION

2D image patch matching and 3D point cloud volume matching are widely used in computer vision and robotics. 2D image patch matching can be used for visual depth estimation, image-based multi-view reconstruction, robot pose estimation, visual retrieval, etc. 3D volume matching of point clouds can be used for point cloud map construction, 3D point clouds reconstruction, etc. Especially, the combination of 2D image patch and 3D point cloud volume matching is a solution to establish the spatial relationship between 2D and 3D space, which plays the potential widely applications in geospatial space, such as 2D and 3D interaction of remote sensing and AR [1-2].

Benefit from the development of sensors, multi-sensor systems integrate the strengths of sensors, so that the machine obtains better sensing capabilities. Cross-domain data, such as 2D images and 3D point clouds, means the difference in data structure, data expression ability and data information caused by the data characteristics from different sensors. If the cross-domain descriptors collectively express the relationship between the 2D image patches and 3D point cloud volumes, then the spatial relationship between 2D and 3D space will be established. In this paper, we aim to extract the local cross-domain descriptor of 2D image and 3D point cloud for metric learning.

The traditional manually designed 2D descriptors [3] (e.g. SIFT, SURF, ORB, DAISY, etc.) are extremely different from 3D descriptors [4] (e.g. PFH, FPFH, ROPS, SHOT, etc.). 2D descriptors focus on the relationship between the pixels of the target grid or the pixels of the adjacent grid. While the 3D descriptor is extracted from the point structure information. Thus, using manually designed descriptors cannot complete the cross-domain matching task. Recently, several neural networks are designed for matching 2D image patches and 3D point cloud volumes, e.g. 2D3D-MatchNet [5], Siam2D3D-Net [6] and LCD [7], which simply use the Euclidean distance to measure the descriptor similarities. However, the above learned feature descriptors are not robust enough, making the low accuracy of the 2D image patches and 3D point cloud volumes matching. In comparison, we embed a fully connected network to measure the similarity between cross-domain descriptors, obtaining a similarity model index that works better than the Euclidean distance.

In this paper, for 2D image patch and 3D point cloud volume matching task, we propose a deep learning architecture which combines feature learning and metric

learning. The proposed network guarantees the feature expression ability and robust feature comparison of cross-domain descriptors. Unlike traditional matching network, two branches of the proposed network with different network structure are respectively set for 2D image patch inputs and 3D point cloud volume inputs. Two branches output 2D descriptors and 3D descriptors, which are contacted and fed into a fully connected network to compute the similarity. In detail, the two-tower structure is named 2D-3D feature network, and the fully connected network for metric learning is named metric network, as shown in Fig. 1.

The contributions of this paper include: 1) We introduce a novel network for 2D image patch and 3D point cloud volume matching by an embedding metric network. 2) We first propose to use the false positive rate at 95% recall as the measurement for the learned cross-domain descriptors of 2D image patches and 3D point cloud volumes. 3) Experiment shows cross-domain descriptors achieve state-of-the-art performance in matching of 2D image patches and 3D point cloud volumes.

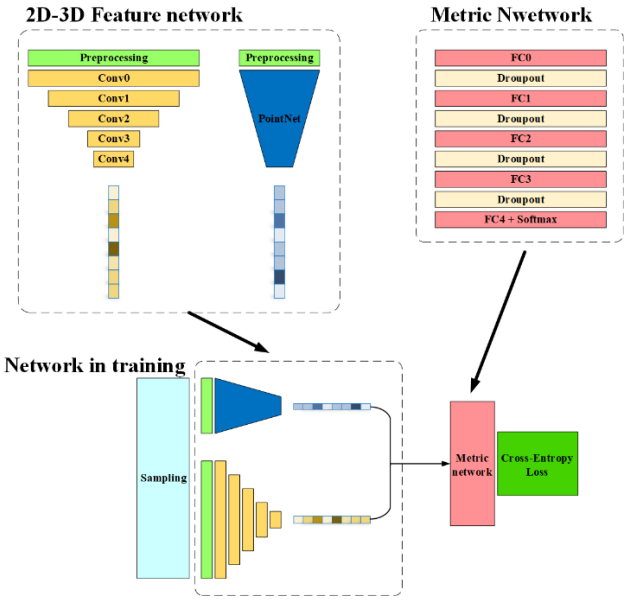

Fig. 1. The proposed network framework, which contains with the 2D-3D feature network and the metric network.

2. NETWORK ARCHITECTURE

In this paper, a deep learning network architecture, which contains with 2D-3D feature network and the metric network, is proposed to jointly learn the cross-domain descriptors to achieve 2D image patch and 3D point cloud volume matching. In detail, the 2D-3D feature network uses Siamese network as framework, but different from traditional Siamese network, 2D-3D feature network obtains two different branches to extract features from the image patch and point cloud volume respectively. The metric network is a fully connected network, receiving both 2D patch descriptors and 3D volume descriptors to compute a similarity. In this section, we introduce the details of the proposed network architecture.

2.1. 2D-3D feature network

Two-tower structure is applied to 2D-3D feature network with two different branches. 2D-3D feature network inputs paired 2D image patches and 3D point cloud volumes and outputs cross-domain descriptors, as shown in Fig. 1. In detail, one branch (image branch) introduces Convolutional Neural Networks (CNNs) architecture to learn descriptors of 2D image patches. The other branch (point cloud branch) uses PointNet [8] as feature extraction, which receives raw point cloud and outputs global 3D descriptors.

For the image branch with CNNs architecture, colored image patches, which are sampled in indoor scene with size 64×64×3, are fed into the network. The outputs of image branch are D-dimensional vectors. In detail, the image branch contains five convolution layers. Except the last layer, Batch Normalization and Tanh activate function are set after each convolution layer. The detailed structure of each layer is set as Table 1. The shorthand notation of PS means the size of convolution kernel, and S means the padding for one operation. D is the embedding size of network, that is the same as the embedding size of 3D volume descriptor.

For the point cloud branch that embed PointNet, raw point clouds are down-sampled to 1024 points which are equipped with coordinate and RGB information. The size of last fully connected network of PointNet are set as D, so that the outputs of point cloud branch are D-dimensional vectors.

Table 1. The details of 2D-3D feature network and metric network.

Name	Type	Output Dim	PS	S
Conv0	C	32	4 x 4	2
Conv1	C	64	4 x 4	2
Conv2	C	128	4 x 4	2
Conv3	C	256	4 x 4	2
Conv4	C	D	4 x 4	4
FC0	FC	256	-	-
FC1	FC	128	-	-
FC2	FC	64	-	-
FC3	FC	32	-	-
FC4	FC	2	-	-

2.2. Metric network

The metric network is consisted with fully connected layers, which is used to compute the similarity between cross-domain descriptors. In detail, metric network receives the concatenation of a pair of descriptors and output two-dimensional vector with value in [0, 1]. The vector expresses the probability that the pair of descriptors is matching or not. The non-linear activate function of metric network is ReLU. In addition, to avoid model failure due to excessive weights in local dimensions, Dropout is applied after each layer. The detailed parameters of metric network are shown in Table 1.

2.3. The preprocessing layers

For the 2D-3D feature network, there are preprocessing layers before the image branch and point cloud branch, as shown in Fig. 1. In detail, the 2D patches are resized as 64×64×3 and the RGB values r, g, b of each pixel are normalized to r/255, g/255, b/255. The 3D volumes are sampled to 1024 points and the RGB values r, g, b of each point is normalized to r/255, g/255, b/255.

3. TRAINING AND PREDICTION

Our proposed network system jointly learns both 2D-3D feature network and metric network. In this section, we introduce the cross-entropy loss function that is used to optimize our proposed network parameters and compute the similarity of cross-domain descriptors. In addition, we constructed as many positive samples and negative samples to ensure the proposed network judge performance.

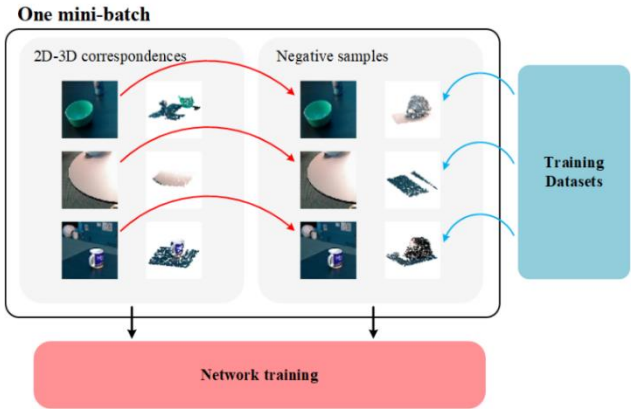

Fig. 2. Negative samples of one mini-batch are generated during training process.

3.1. Data preparation

Paired correspondences of 2D image patches and 3D point cloud volumes are sampled from 3DMatch [9] dataset. The point cloud is reconstructed from the depth maps and contains the corresponding color information. Especially, to construct enough negative samples (non-matching 2D image patches and 3D point cloud volumes) for training, for each mini-batch, 2D image patches randomly sample the same number of non-matching 3D volumes in the training set. Then, these constructed pairs are regard as negative samples. The positive and negative samples are disrupted in this mini-batch, then they are fed into 2D-3D feature network. In the whole training process, the same number of positive samples and negative samples are used for training, which ensures the balance of the training process and avoids over-fitting error, as show in Fig. 2.

3.2. Loss function

The 2D-3D feature network and metric network are trained at the same time, so that we use a unified loss function to optimize both 2D-3D feature network and metric network at the same time. The proposed network tries to minimize the cross-entropy loss function:

$$E = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

In training step, *n* pairs of 2D image patches and 3D point cloud volumes are fed into network in one batch. y_i is the 0/1 labels indicate whether input pair x_i is matching or not, 1 label denotes match and 0 label denotes not match. $\hat{y}_i$ is the output of network and Softmax activate function is set in last network layer, as follow:

$$\hat{y}_i = \frac{e^{f_1(x_i)}}{e^{f_0(x_i)} + e^{f_1(x_i)}}$$

Finally, for input x_i, two-dimensional vector $\begin{pmatrix} f_0(x_i) \\ f_1(x_i) \end{pmatrix}$ is computed as similarity of input pair x_i.

4. EXPERIMENTS

In this section, we first describe the dataset used in this paper. Second, we demonstrate state-of-the-art performance of our proposed network on 2D image patch and 3D point cloud volume matching.

In detail, the matching results of the proposed network is evaluated by the precision and the false positive rate at 95% recall (FPR95). For the training and testing of the proposed network, the dataset is divided into three subsets, Subset1, Subset2, Subset3, which do not intersect each other. Then, the proposed network is trained on one subset and respectively tested on the other two subsets. In addition, the embedding size of the learned descriptors are set as 64, 128, 256 to explore the effect of descriptor dimension, which denoted as Desc64, Desc128, Desc256.

The training stopped after 250 epochs. The training used Nvidia 3090 with 48GB memory. Batch size set as 64. Desc64, Desc 128, Desc 256 took 851 minutes, 997 minutes, and 1221 minutes, respectively.

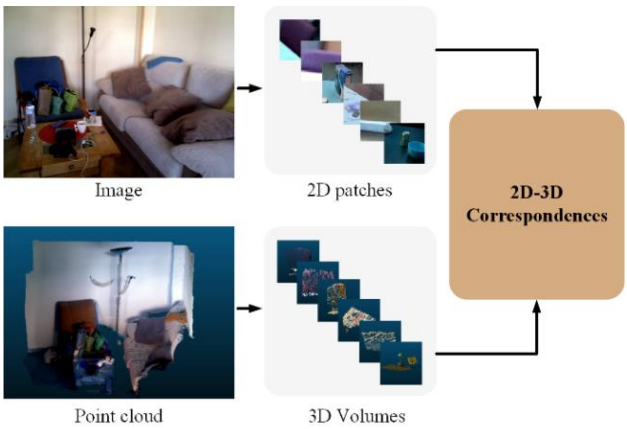

Fig. 3. 2D patches and 3D volumes are collected from Images and point clouds.

Table 2. 2D-3D retrieve performance under precision

		Desc64	Desc128	**Desc256**
Subset1	Subset2	98.502	**98.532**	**98.532**
	Subset3	98.573	98.541	**98.815**
Subset2	Subset1	98.420	98.600	**98.789**
	Subset3	98.430	98.600	**98.754**
Subset3	Subset1	98.487	98.694	**98.801**
	Subset2	98.471	98.677	**98.815**

Table 3. Metric learning performance under FPR95

		Desc64	Desc128	**Desc256**
Subset1	Subset2	0.7506	0.7304	**0.6021**
	Subset3	0.7071	0.7226	**0.6094**
Subset2	Subset1	0.8025	0.7434	**0.6230**
	Subset3	0.7880	0.7029	**0.6281**
Subset3	Subset1	0.7537	0.6691	**0.5969**
	Subset2	0.7672	0.6738	**0.5917**

4.1. Dataset

The matching 2D image patches and 3D point cloud volumes used in the experiments are generated from 3DMatch dataset [10], which is about 600,000 paired 2D-3D correspondences. The point clouds are generated from RGB-D scans, and the correspondences of 2D image patches and 3D point cloud volumes are collected from images and point clouds, as shown in Fig. 3. The Subset1, Subset2, Subset3, contain about 200,000 2D-3D correspondences, are collected from the 600,000 2D-3D correspondences.

4.2. Evaluation under precision

To evaluate the proposed network with precision, one subset contains about 200,000 paired 2D-3D correspondences is used as a training set, and the testing is done on the other two subsets with 2D-3D paired 200,000 correspondences. In training and testing process, we sample 200,000 paired negative samples and mess up with the positive samples. After that, 40,000 paired 2D-3D correspondences are obtained for each subset. The evaluation results are shown in Table 2. The experiments show that the precision achieves up to 98% on other subsets when the proposed network trained on each subset. In addition, the high dimensionality gives improved performance.

4.3. Evaluation under FPR95

The FPR95 is used to evaluate the performance of 2D image patch and 3D point cloud volume matching. In usual image patch matching task, the common methods use several subsets to test the performance of image patch matching. Specifically, the smaller the value of FPR95, the better the performance of the network in image patch matching. In this section, we borrow methods in image patches matching and calculate FPR95 on cross-domain descriptor matching task. After data preparation, each subset has 40,000 pairs of 2D image patches and 3D point cloud volumes. In detail, training on one subset and respectively testing on the other two subset, evaluation results under FPR95 have been shown in Table 3. The FPR95 experimental results are below 1%, which prove the state-of-art performance of the proposed network.

5. CONCLUSION

In this paper, we propose a novel network, combining 2D-3D feature learning and metric learning, for 2D image patch and 3D point cloud volume matching. The proposed network system jointly learns cross-domain descriptors as well as a function to compute similarity of descriptors. In addition, we use the metric learning replaces the artificially designed loss function based on Euclidean distance to achieve non-linear similarity calculation function. Experimental results show that the learned cross-domain descriptors achieve state-of-the-art precision performance and FPR95 performance on 2D image patch and 3D point cloud volume matching. In the future work, we plane to learn robust local cross-domain descriptors of 2D image and 3D point cloud which can be used to retrieve.

6. REFERENCES

[1] Liu W, Lai B, Wang C, et al. “Ground camera image and large-scale 3d image-based point cloud registration based on learning domain invariant feature descriptors”[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 997-1009.

[2] Liu W, Wang C, Bian X, et al. “Learning to match ground camera image and uav 3-d model-rendered image based on siamese network with attention mechanism”[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 17(9): 1608-1612.

[3] Ma J, Jiang X, Fan A, et al. “Image matching from handcrafted to deep features: A survey”[J]. International Journal of Computer Vision, 2021, 129(1): 23-79.

[4] Guo Y, Wang H, Hu Q, et al. “Deep learning for 3d point clouds: A survey”[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[5] Feng M, Hu S, Ang M H, et al. “2d3d-matchnet: learning to match keypoints across 2d image and 3d point cloud”[C]// International Conference on Robotics and Automation (ICRA). 2019: 4790-4796.

[6] Liu W, Lai B, Wang C, et al. “Learning to match 2d images and 3d lidar point clouds for outdoor augmented reality”[C]//IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (IEEE VR). 2020: 655-656.

[7] Pham Q H, Uy M A, Hua B S, et al. “Lcd: learned cross-domain descriptors for 2d-3d matching”[C]//Association for the Advancement of Artificial Intelligence (AAAI). 2020: 11856-11864.

[8] Qi C R, Su H, Mo K, et al. “Pointnet: deep learning on point sets for 3d classification and segmentation”[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 652-660.

[9] Zeng A, Song S, Nießner M, et al. “3Dmatch: learning local geometric descriptors from rgb-d reconstructions”[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1802-1811.