Learning Pairwise Neural Network Encoder for Depth Image-based 3D Model Retrieval

Jing Zhu[1,3] Fan Zhu[1,2] Edward K. Wong[1,3] Yi Fang[1,2*]

[1] NYU Multimedia and Visual Computing Lab
[2] Electrical and Computer Engineering, New York University Abu Dhabi
[3] Computer Science and Engineering, Polytechnic School of Engineering, New York University

ABSTRACT

With the emergence of RGB-D cameras (e.g., Kinect), the sensing capability of artificial intelligence systems has been dramatically increased, and as a consequence, a wide range of depth image-based human-machine interaction applications are proposed. In design industry, a 3D model always contains abundant information, which are required for manufacture. Since depth images can be conveniently acquired, a retrieval system that can return 3D models based on depth image inputs can assist or improve the traditional product design process. In this work, we address the depth image-based 3D model retrieval problem. By extending the neural network to a neural network pair with identical output layers for objects of the same category, unified domain-invariant representations can be learned based on the low-level mismatched depth image features and 3D model features. A unique advantage of the framework is that the correspondence information between depth images and 3D models are not required, so that it can easily be generalized to large-scale databases. In order to evaluate the effectiveness of our approach, depth images (with Kinect-type noise) in the NYU Depth V2 dataset are used as queries to retrieve 3D models of the same categories in the SHREC 2014 dataset. Experimental results suggest that our approach can outperform the state-of-the-arts methods, and the paradigm that directly uses the original representations of depth images and 3D models for retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.5.4 [**Applications**]: Computer vision

Keywords

Neural network; cross-domain; depth image; retrieval

*: Corresponding Author (Email: yfang@nyu.edu)

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806323.

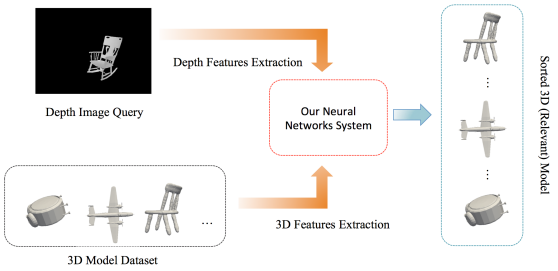

Figure 1: Framework of our proposed cross-domain 3D model retrieval based on depth image query.

1. INTRODUCTION

3D model retrieval has a wide range of applications in the industry, such as architecture and product design. Despite of extensive efforts on 3D model retrieval in recent years, strong requirements on the input queries (normally complete CAD models) have restricted the generalization of its applications. On the other hand, along with the increasing use of handy low-cost depth sensors such as Microsoft Kinect [14], RGBD images are becoming more popular, and as a consequence, large-scale depth images can be more easily accessible. Thus, a novel 3D model retrieval paradigm, depth image-based 3D model retrieval, is motivated and getting popularized in both computer vision and information management communities. A successful depth image-based 3D model retrieval has many potential applications, especially for product designs. While a traditional design process involves a digitization procedure that transforms early stage design drafts into producible CAD models, a depth image-driven retrieval system can avoid such a repetitive and time-consuming procedure by returning a set of relevant existing 3D models based on the depth scan of a query sample, so that designers can directly modify from the retrieved CAD models.

In our work, we exploit the Kinect-typed depth images as queries to retrieve visually similar 3D models from a large collection of 3D model database. In order to alleviate the discrepancy between highly diverged depth images and 3D models, we aim to build an effective cross-domain feature learning framework for depth image-based 3D model retrieval. The pipeline of our framework is shown in Figure 1.

We propose a neural network-based cross-domain feature learning technique by building a neural network pair for depth images and 3D models respectively. In order to reduce

the cross-domain discrepancy, we force unique and identical vectors at the target layers of both neural networks in a supervised fashion for data of the same class. By connecting at the target layers of the neural network pair, data that share the same class label but come from diverged domains can be mapped to the same target vectors. When depth image and 3D model features pass through the neural network pair, values in the hidden layers of both networks are extracted as domain-invariant representations, which are used for depth image-based 3D model retrieval. We evaluate our method using two datasets: the 3D model database comes from SHREC 2014 Large Scale Sketch Track Benchmark [12] and depth images come from NYU Depth V2 dataset [17]. Experimental results demonstrate that our method can achieve outstanding performance. The main contributions of our work are as follows:

- ⋆ We address the challenging depth image-based 3D model retrieval problem with a novel neural network-based approach.
- ⋆ We propose a pairwise neural network technique (PNN) that significantly reduces the cross-domain divergence between depth image features and 3D model features.
- ⋆ The proposed method can achieve outstanding performance on the NYU Depth V2 dataset and the SHREC 2014 benchmark.

2. RELATED WORK

Based on the advancements of RGB-D cameras, depth images have been applied to many computer vision areas, such as object detection [8], gesture recognition, image segmentation, (etc. Saurabh et al.[8] proposed a decision forest approach to deal with the object detection problem in depth images. By cooperating RGB-D contours [2] with the popular convolutional neural networks (CNNs) [7], the proposed approach can achieve outstanding object detection performance. In addition to outputting object detection results at the bounding-box level, pixel-level object inferences are also provided. Wu et al. address the depth image-based gesture recognition problem in a one-shot-learning paradigm, where only one sample can be utilized for training in each gesture category. The NYU Depth v2 dataset [17] is a recently released RGBD dataset that provides diverse indoor scenes with detailed objects' location annotations. Rather than clean and single objects, NYU Depth v2 dataset aims to offer messy indoor scenes with the existence of multiple objects.

As a special machine learning paradigm, transfer learning tackles with the domain mismatch problem. Most existing transfer learning methods [10, 13, 21] operate at the feature learning level and aim to obtain a unified representation for two or multiple mismatched domains (e.g., sketch images vs. 3D shapes, and images vs. texts). Rasiwasia et al. [16] address the image-to-text and text-to-image retrieval problem by investigating the correlations between two modalities and the effectiveness of abstraction, where the canonical correlation analysis (CCA) and the use of abstraction are all proved to be effective. In order to validate the contributions of each separate component, three approaches correlation matching (CM), semantic matching (SM) and semantic correlations matching (SCM) are proposed for the correlation modeling, the abstraction method and the joint working mode of both approaches respectively.

While traditional 3D shape retrieval approaches [3, 11, 15, 19, 4] only consider the same domain data as the input queries, increasing attentions are being paid towards different domain queries. Gao et al. retrieve 3D object based on their 2D views by estimating Hausdorff distances [5] and constructing hypergraphs [6] between objects; Li et al. [12] organize a challenge along with a detailed summary on sketch-based 3D shape retrieval, where human freehand sketches are considered as inputs to the 3D shape retrieval system; Wang et al. [18] also consider using low-cost depth images as queries. In this work, we address the lattest challenge with a neural network-based transfer learning method, which brings two highly variant domains into a new feature space with a low cross-domain discrepancy.

3. APPROACH

We aim to learn pairwise neural networks, which can encode the depth image features and 3D model features respectively, and generate a smooth feature space[1] for both domains. Such pairwise neural networks can be obtained by independently optimizing two discriminative neural networks while connecting both networks at their target layers.

3.1 Discriminative Neural Networks

We consider a 3-layer neural network, which has an input layer, a hidden layer and a target layer. Without loss of generality, we denote $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^M\} \in \mathbb{R}^{M\times D}$ as the input D-dimensional feature, $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^M\} \in \mathbb{R}^{M\times D'}$ as the D'-dimensional hidden layer features and $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{x}}^2, \cdots, \hat{\boldsymbol{x}}^P\} \in \mathbb{R}^{M\times D}$ be the D-dimensional target layer features for either depth images or 3D models. The structure of a single 3-layer neural network is illustrated in Figure 2 (Encoder-1 for example). In our implementation, the input layer features and the target layer features have an identical feature dimension. In the training stage, the weights on all neurons within the network are optimizing towards a minimum discrepancy between the input layer features $\boldsymbol{X}$ and the target layer features $\hat{\boldsymbol{X}}$. Once the weights are optimized, the hidden layer values $\boldsymbol{Y}$ can be extracted as encoded features. In order to enhance the extrapolation capability of the encoded features, we aim to train discriminative neural networks by forcing identical target vectors to instances that come from the same category at the target layer. Thus, the objective function for learning a discriminative neural network can be formulated as:

$$\begin{aligned} \arg\min_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{D}\sum_{i=1}^{D}\|\hat{\boldsymbol{x}}^i - h_{\boldsymbol{W}^l,\boldsymbol{b}^l}(\boldsymbol{x}^i)\|_2^2 + \lambda\sum_{l=1}^{L}\|\boldsymbol{W}^l\|_F^2, \\ \text{s.t.} \quad \hat{\boldsymbol{x}}^i = \hat{\boldsymbol{x}}^j \quad \text{if} \quad q(\boldsymbol{x}^i) = q(\boldsymbol{x}^j) \end{aligned} \tag{1}$$

where $\boldsymbol{W} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, \cdots, \boldsymbol{W}^L\} \in \mathbb{R}^{M\times L}$ is the neuron parameters of the neural network and L is the number of layers, $q(\cdot)$ λ is the balancing parameter and W^l is the weight vector at layer l. The function $h_{W,b}(\boldsymbol{x}^p) = f(\sum_{k=1}^{K} w_k x_k^p + b)$ is a sigmoid function, where

$$f(z) = \frac{1}{1 + exp(-z)}. \tag{2}$$

[1]In a smooth feature space, feature points, which are close to each other, are more likely to share the same class label.

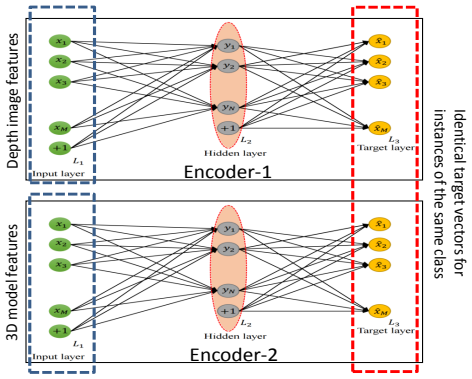

Figure 2: The structure of pairwise neural networks.

3.2 Pairwise Neural Networks

The discriminative neural network can improve the data smoothness in either the depth image domain or the 3D model domain. In order to minimize the data discrepancy between both domains, we use a structure that connects two discriminative neural networks at the target layers. Intuitively, since instances that come from the category are forced to possess identical features at the target ends, the pairwise neural networks are provided with the ability of encoding the same-category cross-domain instances towards similar representations at the hidden layers (see Figure 2). Let $\boldsymbol{X_d} = \{\boldsymbol{x_d^1}, \boldsymbol{x_d^2}, \cdots, \boldsymbol{x_d^{M_d}}\} \in \mathbb{R}^{M_d \times D_d}$ and $\boldsymbol{X_m} = \{\boldsymbol{x_m^1}, \boldsymbol{x_m^2}, \cdots, \boldsymbol{x_m^{M_m}}\} \in \mathbb{R}^{M_m \times D_m}$ be the depth and 3D model domain inputs, and $\hat{\boldsymbol{X}}_{\boldsymbol{d}} = \{\hat{\boldsymbol{x}}_{\boldsymbol{d}}^1, \hat{\boldsymbol{x}}_{\boldsymbol{d}}^2, \cdots, \hat{\boldsymbol{x}}_{\boldsymbol{d}}^{M_d}\} \in \mathbb{R}^{M_d \times D_d}$ and $\hat{\boldsymbol{X}}_{\boldsymbol{m}} = \{\hat{\boldsymbol{x}}_{\boldsymbol{m}}^1, \hat{\boldsymbol{x}}_{\boldsymbol{m}}^2, \cdots, \hat{\boldsymbol{x}}_{\boldsymbol{m}}^{M_m}\} \in \mathbb{R}^{M_m \times D_m}$ be the depth and 3D model domain target vectors, respectively, a unified objective function for learning the pairwise neural networks can be formulated as:

$$
\begin{aligned}
&\arg\min_{\boldsymbol{W_d}, \boldsymbol{b_d}} \frac{1}{M_d} \sum_{i=1}^{M_d} \|\hat{\boldsymbol{x}}_d^i - h_{\boldsymbol{W_d}, \boldsymbol{b_d}}(\boldsymbol{x_d^i})\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{W_d^l}\|_F^2, \\
&\arg\min_{\boldsymbol{W_m}, \boldsymbol{b_m}} \frac{1}{M_m} \sum_{j=1}^{M_m} \|\hat{\boldsymbol{x}}_m^j - h_{\boldsymbol{W_m}, \boldsymbol{b_m}}(\boldsymbol{x_m^j})\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{W_m^l}\|_F^2, \\
&\text{s.t.} \quad \hat{\boldsymbol{x}}_m^i = \hat{\boldsymbol{x}}_m^j = \hat{\boldsymbol{x}}_d^i = \hat{\boldsymbol{x}}_d^j \\
&\text{if} \quad q(\boldsymbol{x}_m^i) = q(\boldsymbol{x}_m^j) = q(\boldsymbol{x}_d^i) = q(\boldsymbol{x}_d^j),
\end{aligned}
\tag{3}
$$

where W_d, b_d, W_m and b_m are parameters of the depth image network and the 3D model network respectively. Computing optimum parameters of the pairwise neural networks is a regression problem. The typical backpropagation algorithm [9], which can efficiently compute the partial derivatives, is applied to obtain the optimum parameters. Once we obtain the optimum $\hat{\boldsymbol{W}}_d$, $\hat{\boldsymbol{b}}_d$, $\hat{\boldsymbol{W}}_m$ and $\hat{\boldsymbol{b}}_m$, neuron values in L_2 layers are extracted as the representations when depth image and 3D model features pass through the networks.

4. EXPERIMENTS

The NYU Depth V2 dataset [17] contains frames of video sequences in a variety of indoor scenes, from where objects in depth images are extracted as queries in our experiments. The database is constructed by selecting 3D models in corresponding categories from the large-scale extended SHREC 2014 benchmark [12]. The numbers of samples in 7 categories (*bathtub*, *bathtub*, *bed*, *chair*, *desk*, *dresser*, *night stand*, *table*) of both datasets are given in Table 1. We follow the Sparse Coding Spatial Pyramid Matching (ScSPM) [20] framework for depth image representations, and extract Local Depth Scale-Invariant Features Transform (LD-SIFT) [1] features from 3D models and obtain 1000-dimensional 3D model histogram features by fitting LD-SIFT features to a Bag-of-Words (BoW) model.

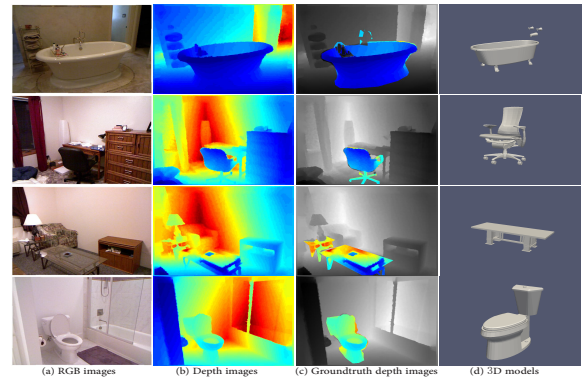

Figure 3: Examples of both the NYU Depth V2 dataset and the SHREC 2014 benchmark. In our experiments, groundtruth depth images (c) are used as queries and 3D models (d) are used as the database.

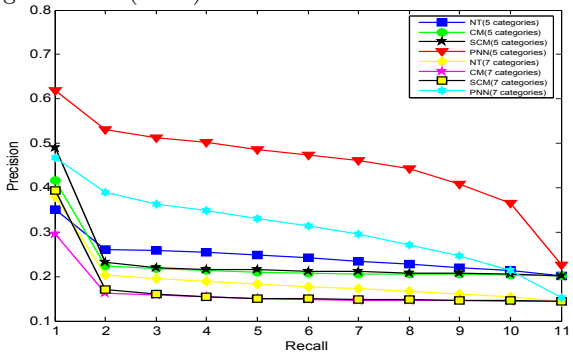

Figure 4: Precision-Recall plot for performance comparison.

We evaluate the proposed approach using 6 common evaluation metrics, Nearest Neighbor (NN), First Tier (FT), Second Tire (ST), E-Measure (E), Discounted Cumulated Gain (DCG) and Average Precision (AP). We conduct experiments using both 5 categories (the first 5 categories as displayed in Table 1) and 7 categories, and compare with the state-of-the-art transfer learning approaches CM and SCM. We also compare with the non-transfer (NT) approach, which directly utilizes the original depth image and 3D model representations for retrieval. Results are reported in Table 2, and the Precision-Recall (PR)-curves for both experiments are given in Figure 4. Experimental results suggest that the proposed PNN method consistently leads the best performance on different settings and evaluation metrics. We can also observe that the performance improvements of PNN over other methods are less significant on 7 categories than 5 categories. We conclude the cause of such performance degradation to the unbalanced numbers of samples across the depth image domain and the 3D model domain. When two domains are supplied with unbalanced numbers of samples for training, the learned neural networks become biased towards the domain with sufficient training samples.

Table 1: Numbers of samples in 7 categories of the NYU Depth V2 dataset and the SHREC 2014 benchmark.

Categories	bathtub	bed	chair	desk
Depth images	52	313	649	192
3D models	76	386	641	147
Categories	dresser	night stand	table	
Depth images	106	302	534	
3D models	140	149	520	

Table 2: Performance metrics comparison of depth image-based 3D model retrieval on the NYU Depth V2 dataset and the SHREC 2014 benchmark.

	NN	FT	ST	DCG	E	AP
7 categories						
NT	0.23	0.14	0.30	0.66	0.02	0.15
CM	0.14	0.14	0.27	0.65	0.02	0.15
SCM	0.14	0.14	0.27	0.65	0.03	0.15
PNN	**0.37**	**0.26**	**0.40**	**0.71**	**0.05**	**0.28**
5 categories						
NT	0.04	0.21	0.39	0.71	0.03	0.21
CM	0.12	0.19	0.39	0.71	0.03	0.20
SCM	0.20	0.18	0.38	0.70	0.02	0.20
PNN	**0.52**	**0.39**	**0.58**	**0.78**	**0.06**	**0.42**

5. CONCLUSIONS

In this work, we address the challenging depth image-based 3D model retrieval problem with a pairwise neural network approach. In order to minimize the discrepancy between highly diverged depth images and 3D models, we build a neural network pair for depth images and 3D models respectively, while enforcing identical target values at output layers of both networks. The proposed PNN method was successfully validated on the NYU Depth Dataset V2 and the extended SHREC 2014 3D shape retrieval benchmark, where the experimental results suggest that PNN can consistently outperform other methods. Moreover, since PNN does not require the correspondence information across different domains, it can be easily generalized.

6. REFERENCES

[1] T. Darom and Y. Keller. Scale-invariant features for 3-d mesh models. *Image Processing, IEEE Transactions on*, 21(5):2758–2769, 2012.

[2] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *International Conference on Computer Vision*, pages 1841–1848. IEEE, 2013.

[3] Y. Fang, M. Sun, and K. Ramani. Temperature distribution descriptor for robust 3d shape retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 9–16. IEEE, 2011.

[4] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 3d deep shape descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2015.

[5] Y. Gao, M. Wang, R. Ji, X. Wu, and Q. Dai. 3-d object retrieval with hausdorff distance learning. *Industrial Electronics, IEEE Transactions on*, 61(4):2088–2098, 2014.

[6] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai. 3-d object retrieval and recognition with hypergraph analysis. *Image Processing, IEEE Transactions on*, 21(9):4290–4303, 2012.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, pages 580–587. IEEE, 2014.

[8] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[10] K. Hu and Y. Fang. 3d laplacian pyramid signature. In *Computer Vision-ACCV 2014 Workshops*, pages 306–321. Springer, 2014.

[11] G. Leifman, R. Meir, and A. Tal. Semantic-oriented 3d shape retrieval using relevance feedback. *The Visual Computer*, 21(8-10):865–875, 2005.

[12] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan, et al. Shrec' 14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics Workshop on 3D Object Retrieval 2014 (3DOR 2014)*, pages 121–130, 2014.

[13] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(6):1134–1148, 2014.

[14] Microsoft. Microsoft kinect description, 2015.

[15] M. Novotni and R. Klein. 3d zernike descriptors for content based shape retrieval. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 216–225. ACM, 2003.

[16] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *International Conference on Multimedia*, pages 251–260. ACM, 2010.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[18] Y. Wang, J. Feng, Z. Wu, J. Wang, and S.-F. Chang. From low-cost depth sensors to cad: Cross-domain 3d shape retrieval via regression tree fields. In *European Conference on Computer Vision*, pages 489–504. Springer, 2014.

[19] J. Xie, Y. Fang, F. Zhu, and E. Wong. Deepshape: Deep learned shape descriptor for 3d shape matching and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1275–1283, 2015.

[20] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition*, pages 1794–1801. IEEE, 2009.

[21] F. Zhu and L. Shao. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2):42–59, 2014.