Pattern Recognition Letters 000 (2017) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Learning domain-invariant feature for robust depth-image-based 3D shape retrieval

Jing Zhu[a,b,d], John-Ross Rizzo[e], Yi Fang[a,c,d,*]

[a] NYU Multimedia and Visual Computing Lab, USA
[b] Department of Computer Science and Engineering, NYU Tandon School of Engineering, New York, USA
[c] Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, New York, USA
[d] Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi, UAE
[e] Depts. of Rehabilitation Medicine and Neurology, NYU Langone Medical Center, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Discriminative neural network
Cross-domain
Depth images
3D shape retrieval

ABSTRACT

In recent years, 3D shape retrieval has been garnering increased attention in a wide range of fields, including graphics, image processing and computer vision. Meanwhile, with the advances in depth sensing techniques, such as those used by the Kinect and 3D LiDAR device, depth images of 3D objects can be acquired conveniently, leading to rapid increases of depth image dataset. In this paper, different from most of the traditional cross-domain 3D shape retrieval approaches that focused on the RGB-D image-based or sketch-based shape retrieval, we aim to retrieve shapes based only on depth image queries. Specifically, we proposed to learn a robust domain-invariant representation between 3D shape and depth image domains by constructing a pair of discriminative neural networks, one for each domain. The two networks are connected by a loss function with constraints on both inter-class and intra-class margins, which minimizes the intra-class variance while maximizing the inter-class margin among data from the two domains (depth image and 3D shape). Our experiments on the NYU Depth V2 dataset (with Kinect-type noise) and two 3D shape (CAD model) datasets (SHREC 2014 and ModelNet) demonstrate that our proposed technique performs superiorly over existing state-of-the-art approaches on depth-image-based 3D shape retrieval task.

1. Introduction

3D shape retrieval has become an important topic in computer vision field with a wide range of applications in engineering, manufacturing, product design, and the medical field. Compared to the within-domain shape retrieval using 3D shapes as queries, cross-domain shape retrieval, such as sketch-based shape retrieval and RGB image-based shape retrieval [6,16,18], is a more attractive yet challenging problem. In recent years, due to the emergence of low-cost depth sensors, e.g. the Kinect and 3D LiDAR systems, RGB-D images of objects can be captured easily. As a consequence, a number of large-scale RGB-D image datasets have become available, and precipitated the problem of cross-domain shape retrieval. Although RGB-D images provide large amounts of information for successful shape retrieval, processing the complex RGB-D images usually requires higher computational consumption on time and space. A shape retrieval system driven only on depth images may be a more efficient and effective system. As shown in Fig. 1, given a depth image query, a depth image-based shape retrieval system can return a set of relevant 3D models from a large 3D model database. As an example, product design users that simply capture depth images of objects could greatly facilitate automated relevant 3D model selection, expediting the step-wise industrial process.

However, due to the high diversity between the raw representation formats of 2D depth images and 3D shapes, it is nearly impossible to build a shape retrieval system by directly matching the depth image queries to corresponding 3D shapes. We can also find the variations from some examples of depth images and their corresponding shapes in Fig. 2. To tackle the variation challenge, intuitively, we can convert the cross-domain data into one single domain to do the (within-domain) retrieval. For example, in the early attempt, instead of directly retrieving 3D shapes from object depth images, [38] transformed the depth image-based shape retrieval problem to a reconstructed model-based shape retrieval problem, which took a noisy 3D model (reconstructed based on depth images from multiple views) as input and outputted a relevant CAD model. However, it is not practical for users to capture depth im-

[*] Corresponding author.
E-mail address: yfang@nyu.edu (Y. Fang).

https://doi.org/10.1016/j.patrec.2017.09.041
0167-8655/© 2017 Published by Elsevier B.V.

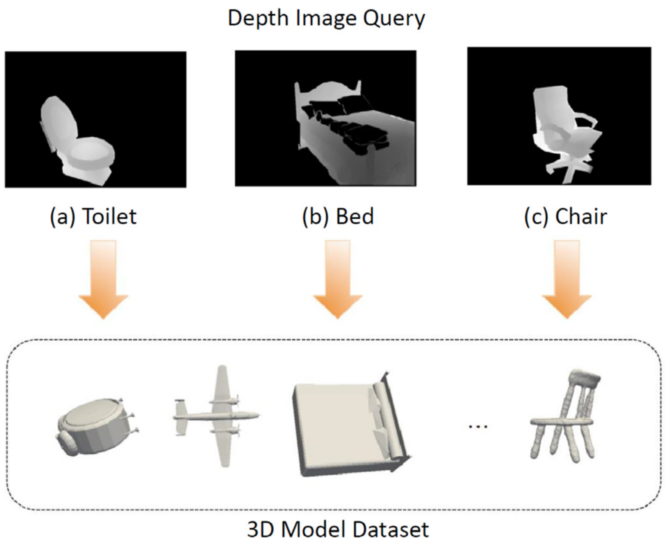

Fig. 1. Illustration of depth image-based 3D model retrieval. Given a depth scan of query sample, a set of relevant 3D models in a large database that are from the same category as query's can be retrieved.

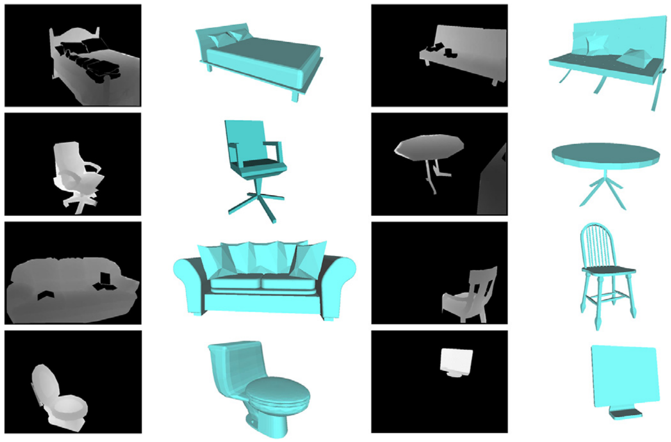

Fig. 2. Examples of depth images (from NYU Depth V2 dataset) and their corresponding 3D shapes (from ModelNet dataset). As we can see from the figure, there is a great variation between 2D depth images and 3D shapes.

ages of a single object from many views. As we can see from the Fig. 2, in the popular RGB-D image datasets, object depth images are captured from only one single angle. Besides, the objects in the depth images are usually incomplete or occluded by other objects, making the retrieval task even more challenging.

Recently, inspired by the successful applications of autoencoders in the computer vision field, Feng et al. [7] proposed to first render some depth images from each 3D model in the dataset, and then trained an autoencoder for each 3D model based on their rendered depth images. Finally, given a depth image query, they got a reconstructed depth image from each autoencoder, and retrieved 3D models by applying a potential model on the reconstructed depth images. Though the performance on a small subset of depth images was promising, it is not easy to apply such approach on large-scale datasets since it required to train one autoencoder for each 3D model. Therefore, we consider to construct a more generalizable deep neural network to learn a cross-domain representation for both 3D shapes and depth images. Due to the distinctive intrinsic properties between depth image domain and 3D shape domain, it is difficult to build a neural network directly on raw 3D shapes and depth images, so we are seeking an indirect way that enables the connection between two domains.

On the other hand, hand-crafted features have shown their excellent performance on many challenging computer vision problems. For example, 3D SIFT have been proven its effectiveness with leading and robust results on 3D shape retrieval [5]. However, most of current existing hand-crafted features are designed for only single domain, either 2D images or 3D shapes, and due to the great discrepancy, it is difficult to find an effective hand-crafted feature working on both 2D depth image and 3D shapes. Although it is unpractical to design a cross-domain hand-crafted feature manually, we can still utilize the advantages of existing hand-crafted features to first reduce the within-domain variation and then train a deep neural network upon the extracted hand-crafted features to handle the cross-domain issues. For better learning, we use two deep neural networks in our proposed model, one for depth images and the other for 3D shapes. To connect these two networks, we define a loss function with constraints on both inter-class and intra-class margins, mapping distinctive input from two domains into the same target space by minimizing the (intraclass) difference between cross-domain data within the same category while maximizing the (interclass) variation among data from different categories. Finally, the final outputs of the trained networks are considered as the domain-invariant representations for given cross-domain data, and relevant 3D shapes can be retrieved by directly comparing the domain-invariant features between 3D shapes and depth image queries.

The experimental results on three popular datasets, where depth images are from NYU Depth V2 dataset and 3D models come from either SHREC 2014 database or ModelNet dataset, suggest that our proposed method significantly outperforms other state-of-the-art approaches. What's more, once the networks are trained, we can perform efficient shape retrieval on given depth image queries since only some matrix computation is required.

In summary, the main contributions of our work include:

- To address the challenging depth-image-based 3D shape retrieval problem, we propose to learn a domain-invariant feature for cross-domain data.
- To better learn the domain-invariant feature, we design a model with two neural networks connected and optimized by a loss function that maximizes the inter-class margin while minimizing the intra-class variance between heterogeneous cross-domain (depth image and 3D model) data.
- The proposed method has been successfully validated on large datasets with superior performance over other state-of-the-arts methods, including those applied on depth image-based shape retrieval, transfer learning methods used on similar task, and the approach that directly uses the original features of depth images and 3D models as representations for retrieval.

2. Related work

Although cross-domain shape retrieval has received many attentions for years, most researchers were working on sketch-based or image-based shape retrieval. Recently, with the increases of the depth image datasets, some researchers have started to look at the depth image-based shape retrieval problem. In this section, we review three key components in depth image-based shape retrieval, including datasets, features and neural network.

2.1. Dataset

Started from decades ago, extending effort has been paid on building 3D shape datasets. Most of current popular 3D shape datasets contain thousands even millions of manually designed CAD models for different kinds of objects. For example, SHREC 2014 Benchmark [17,18] is one of the most popular 3D shape

Please cite this article as: J. Zhu et al., Learning domain-invariant feature for robust depth-image-based 3D shape retrieval, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.09.041

dataset in computer vision field, which is usually used for evaluation on sketch-based 3D shape retrieval. The Princeton ModelNet [41] is another well-known 3D model dataset providing a collection of clean CAD models for more than 600 categories. Other publicly available datasets, e.g. Princeton Shape Benchmark [35] and ShapeGoogle dataset [3], are also widely used for within-domain 3D shape retrieval.

With the advanced development of RGB-D cameras, a large number of RGB-D image datasets have been created. NYU depth V2 dataset [36] is a recent released RGB-D image dataset, which provides a large collection of RGB and depth images for diverse indoor objects, such as cup, desk, etc. The availability of such large-scale RGB-D image dataset enables researchers to solve some traditional challenging computer vision problems using RGB-D images, such as shape reconstruction and object detection [8]. It also leads the popularity of depth images in the graphic and computer vision communities.

2.2. Features

Due to the long history of research on 2D images, a lot of hand-crafted features have been well-defined for different tasks, such as image classification and object recognition, most of which are based on or extended from the classic bag-of-word model, e.g. SIFT features [22], SPM features [14], ScSPM feature [42], etc. Getting inspiration from those 2D image processing approaches, a number of hand-crafted features have been created to address the 3D shape retrieval challenges, such as calculating the probability distribution on geometric properties of an 3D model [25] and finding the symmetry of shapes [12]. Besides the above global descriptors, some local characteristics have also been utilized for more robust shape retrieval. For example, Bronstein et al. [4] bagged the values of multiscale diffusion heat kernel as features to represent 3D models, and Darom and Keller [5] have successfully extended the well-known SIFT features [22] on 3D shapes (known as LD-SIFT) to achieve outstanding performance on shape retrieval task.

In addition to the hand-crafted features, learning-based features are getting more and more popular to address either image or shape problems. As a special machine learning paradigm, transfer learning are mainly used to tackle with the domain mismatch problem. Most current existing transfer learning methods [11,19,45] were operated at the features learning level, aiming to obtain a unified representation for two or more mismatched domains (e.g., sketch images vs. 3D shapes, images vs. 3D shapes and images vs. texts). Rasiwasia et al. [29] addressed the image-to-text and text-to-image retrieval problem by investigating the correlations between two modalities, and measuring the effectiveness of abstraction. In their work, both the canonical correlation analysis (CCA) and the use of abstraction were proven to be effective for retrieval task. To evaluate the contributions of each separate component, three approaches – correlation matching (CM), semantic matching (SM) and semantic correlations matching (SCM) – were proposed for correlation modelling, abstraction utilization and joint working of both approaches, respectively. In another application on cross-domain matching, Zhang et al. [44] proposed to match objects in 2D images with the projected images from multiple generated deformed models.

2.3. Neural network

Inspired by biological neural networks, artificial neural network is a system containing a number of processing elements, providing dynamic outputs according to external inputs. The simplest type of neural network is the perceptron, created by Rosenblatt [32]. Later, Werbos [39] introduced the backpropagation algorithm making neural networks even more popular for machine learning. Nowadays, neural network techniques have achieved great success on various real-world applications, such as biomedicine [20], energy [2], telecommunications [43], geophysics [1], etc. In addition to the traditional neural network structure, there has been increasing interest in deep learning neural networks in recent years [13,15,34], especially in convolutional neural networks (CNN). For example, Malinowski et al. [23] proposed to combine a CNN with a LSTM into an end-to-end architecture for task of answering questions about images, while Pfister et al.[27] utilized CNN to estimate human pose in videos by combining information across the multiple frames using optical flow.

Feng et al.[7] attempted to apply deep neural network techniques on depth image-based shape retrieval. Multiple autoencoders were trained on rendered depth images from corresponding 3D models, one autoencoder for one 3D model. Given a depth image query, the retrieval was performed based on the reconstructed depth images generated from each autoencoder. Generalizing their method on a large 3D model dataset could be very expensive since each 3D model needs one autoencder to be trained for representation. Despite the effort from Feng et al., Zhu et al. [46] recently proposed to build a pair of neural networks for depth image-based 3D shape retrieval. However, random variables were assigned as target vectors to connect the networks in their work, making the retrieval performance greatly depend on the initialization of the random values. Their consideration on within-class variation only also limited the performance on shape retrieval. In this paper, we focus on eliminating these shortcomings by connecting the network pair with a loss function that constraints the inter-class difference as well as the intra-class variance.

3. Approach

We propose to learn a domain-invariant representation for depth image-based shape retrieval using two discriminative neural networks, one for each domain, so that samples from the two domains can be matched without any reconstruction on either domain. Specifically, we first extract hand-crafted features from depth images and 3D shapes respectively, and then learn a network pair upon the extracted hand-crafted features. In this section, we introduce how we extract features in Section 3.1, and followed by a presentation of our network architecture in Section 3.2.

3.1. Feature extraction

Depth image-based shape retrieval is a typical cross-domain matching problem. Usually, raw 3D shapes are represented by points (coordinates), and surfaces (triangles connected by points), while depth images are single-channel images containing distances from sensor to the surfaces of captured objects in each pixel. Since the intrinsic variance between data from these two domains, it is difficult for us to do the retrieval directly on their raw presentations. Therefore, we consider to learn a network pair that could map the highly discriminative data (from two domains) to a common feature space, where the closer feature points are more likely to share the same class label, and the further points are more likely belong to different classes.

As we known, raw 2D images have been taken as the inputs in multiple deep neural networks for end-to-end learning, and achieved outstanding performance. However, it is not very easy to apply neural network on raw presentation of 3D shapes. Inspired by the success of hand-crafted features on within-domain shape retrieval, we extract hand-crafted features from shapes as representations, which could be easily taken as inputs for any kind of neural networks. Given these hand-crafted features, a deep shape representation could be learned via a deep neural network.

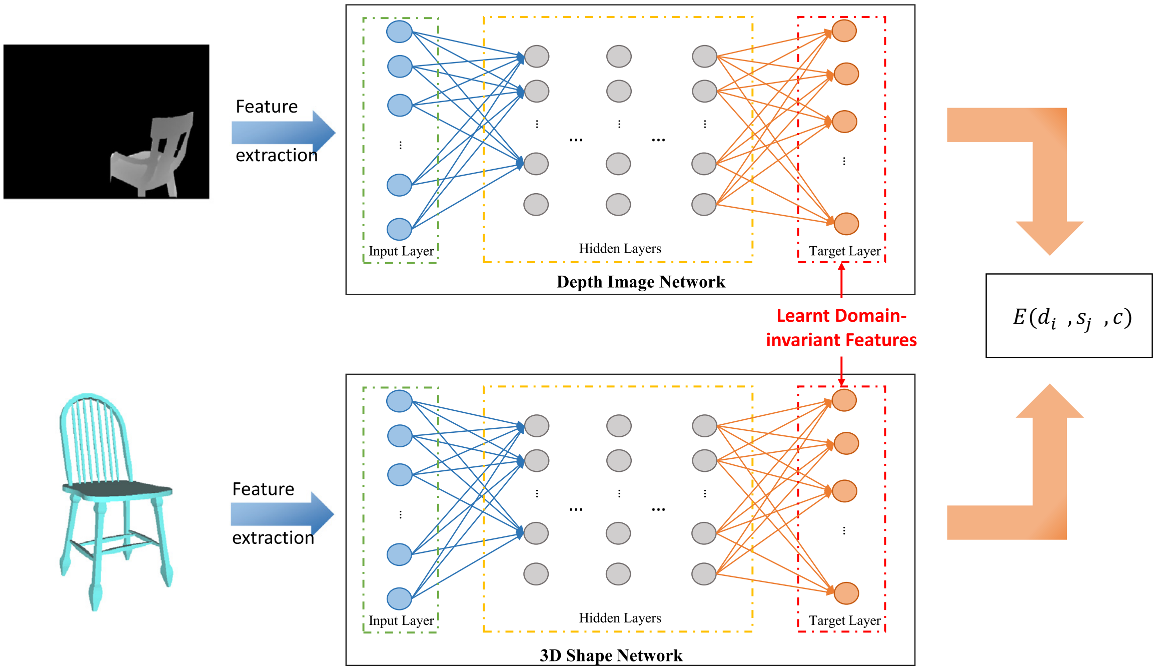

Fig. 3. The pipeline of our proposed method, where two networks are used to handle the cross-domain issues, one for each domain. The two networks share the same architecture, but take different extracted features as inputs. By connecting the two networks with a loss function on the network outputs, a domain-invariant feature can be learned at the target layer.

To keep the same network structure for both shape network and depth image network, we also extract features on depth images as representations. Furthermore, the hand-crafted features have been proven their discriminative ability on either image classification or shape retrieval, so we believe that learning the domain-invariant features from hand-crafted features could eliminate not only the cross-domain invariance but also within-domain difference. Since the feature extraction part is not the contribution for our paper, we just provide brief description about the hand-crafted features we used in our method below.

3.1.1. *Shape features*

For each 3D shape, we adopt the robust Local Depth Scale-Invariant Features Transform (LD-SIFT) features, which is an extension of 2D SIFT [22] features used on 3D meshes. First, some interesting points and local scale are detected by using Difference of Gaussian operator. Then, LD-SIFT features can be computed as the distances from the surrounding vertices to the dominant plane at each interesting point. For more details, please refer to the LD-SIFT paper [5].

3.1.2. *Depth image features*

We extract local features for depth image following the Sparse Coding Spatial Pyramid Matching (ScSPM) [42] framework. After getting SIFT features of each depth image, sparse coding and multi-scale max pooling are applied on the local SIFT features to generate some higher level features. We finally get the ScSPM features by concatenating the SIFT features and the outputs of each sparse coding layer and max pooling layer. For more details, please refer to ScSPM paper [42].

3.2. *Network architecture*

In order to reduce the discrepancy between two domains (depth image domain and 3D shape domain), we adopt two neural networks in our model, one for depth image domain and the other for 3D shape domain. The two networks are “aligned” by a loss function at the target layer. Fig. 3 shows the architecture of our network, where the depth image network and the 3D shape network share the same network architecture.

Let $\boldsymbol{d_i}, \boldsymbol{s_j}$ be the extracted features from depth images and 3D shapes, which also are inputs for the depth image and the 3D shape network, respectively. Then the outputs of network can be computed following:

$$\begin{aligned} &\text{for} \quad l = 1, \quad \boldsymbol{D_i^l} = f\left(\boldsymbol{d}_i * \boldsymbol{W_d^l} + b_d^l\right) \\ &\text{for} \quad l = 2, \dots, L, \quad \boldsymbol{D_i^l} = f\left(\boldsymbol{D}_i^{l-1} * \boldsymbol{W_d^l} + b_d^l\right), \end{aligned} \tag{1}$$

where $\boldsymbol{D_i^l}$ denotes the output of each layer in the depth image network given input $\boldsymbol{d_i}$, and $\boldsymbol{W_d^l}$, b_d^l are the depth image network parameters. Sigmoid function $f(z)$ is adopt as the activation function for the neurons in our network:

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

Similarly, the outputs of 3D shape network can be obtained from:

$$\begin{aligned} &\text{for} \quad l = 1, \quad \boldsymbol{S_j^l} = f\left(\boldsymbol{s}_j * \boldsymbol{W_s^l} + b_s^l\right) \\ &\text{for} \quad l = 2, \dots, L, \quad \boldsymbol{S_j^l} = f\left(\boldsymbol{S}_j^{l-1} * \boldsymbol{W_s^l} + b_s^l\right), \end{aligned} \tag{3}$$

where $\boldsymbol{S_j^l}$ denotes the output of each layer in the 3D shape network for input $\boldsymbol{s_j}$, and $\boldsymbol{W_s^l}$, b_s^l are the shape network parameters. To connect these two networks, a loss function is designed based on the outputs at the target layers (the last layers of the two networks). In the traditional neural network, the loss function is defined by minimizing the variance between the outputs of the network and the target vectors. However, it is pretty hard to define a perfect target vector for cross-domain data by human being. One possible way is to define a soft-max loss upon the outputs of the two networks, and assign one class label for each sample. However, it requires accurate class label for each sample, and training the two networks separately. A weakly-supervised method, that learns a feature space suitable for cross-domain data, would be more desired.

Inspired by metric learning technique, our model takes a pair of samples as inputs at each time during training, one from each domain, and our loss function is composed by two terms: the inter-class margin and the intra-class margin. If the two samples come from the same category, then the variance between outputs of networks is considered as the intra-class margin. Otherwise, the variance can be seen as the inter-class margin. The loss function has the following form:

$$\boldsymbol{E}(\boldsymbol{d_i}, \boldsymbol{s_j}, \boldsymbol{c_{ij}}) = \frac{1}{N_p} \sum_{i=1}^{N_d} \sum_{j=1}^{N_s} (c_{ij} \| \boldsymbol{D_i^L} - \boldsymbol{S_j^L} \|_2^2 - (1 - c_{ij}) \| \boldsymbol{D_i^L} - \boldsymbol{S_j^L} \|_2^2) + \lambda \sum_{l=1}^{L} (\| \boldsymbol{W_d^l} \|_F^2 + \| \boldsymbol{W_s^l} \|_F^2), \tag{4}$$

where N_p is the number of training pairs, N_d is the number of depth image samples, N_s is the number of 3D shape samples, and c_{ij} is the relationship label between the sample $\boldsymbol{d_i}$ and $\boldsymbol{s_j}$. If the two inputs are from the same class, then c_{ij} equals to 1, otherwise, c_{ij} equals to 0. By minimizing the loss calculated from Eq. (4), we can learn a common feature space for cross-domain data with minimum intra-class margin and maximum inter-class margin.

The network training process is an optimization problem, which aims to get optimum parameters of the network so that the loss computed by Eq. (4) is as small as possible. We adopt the classic backpropagation algorithm, which can efficiently compute the partial derivatives and update the parameters with gradients, to obtain the optimum parameters. Once we obtain the optimum $\hat{\boldsymbol{W}}_d$, $\hat{\boldsymbol{b}}_d$, $\hat{\boldsymbol{W}}_s$ and $\hat{\boldsymbol{b}}_s$, given any depth image queries or 3D models, outputs of the corresponding network at the target layer are extracted as the domain-invariant representations. The two networks can be used to generate domain-invariant features independently and simultaneously. Relevant 3D models are then retrieved based on the Euclidean distance calculated between the domain-invariant features of the depth image queries and the 3D models:

$$Dist(\hat{\boldsymbol{D}}_i, \hat{\boldsymbol{S}}_j) = \sqrt{\sum_{k=1}^{m} (\hat{\boldsymbol{d}}_i^k - \hat{\boldsymbol{s}}_s^k)^2}. \tag{5}$$

where $\hat{\boldsymbol{D}}_i, \hat{\boldsymbol{S}}_j$ denote the domain-invariant features for the *ith* depth image query and the j^{th} 3D model, m is the dimension of the output features. The difference between the learned features of the depth image and the 3D model from the same category should be small while the variance is large among those from varied classes. We rank the distances computed by Eq. (5) in ascending order for each depth image query to generate a distance matrix. Then, 3D models with smaller distances in the matrix are retrieved as relevant ones for each depth image query.

4. Experiments

To validate the performance of our proposed method, we comprehensively evaluate our algorithm on one large depth image dataset and two 3D model datasets by conducting experiments with various settings. Retrieval performance is evaluated by common evaluation metrics and precision-recall curves. In all experiments, our method outperforms the-state-of-the-art methods, demonstrating that our proposed method can successfully learn the domain-invariant features for both domains (depth image and 3D shape).

4.1. Datasets

4.1.1. Depth image dataset

The queries of our proposed method are labeled depth images from NYU Depth V2 dataset [36], which is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. The NYU Depth V2 dataset contains 1,449 densely labeled pairs of aligned RGB and depth images with 894 object categories. Since there may be more than one object on a single image, we extract the corresponding depth image regions with different class labels and collect them as multiple object images. In order to guarantee that both the depth image dataset and the 3D model dataset have the same object categories, we use 2,517 depth images for 10 categories in the NYU Depth Dataset V2. The 2,517 depth images are further split into training dataset and test dataset with 1:1 ratio.

4.1.2. 3D model datasets

The 3D models usd in our experiment are from two recent large datasets. One is the large-scale extended SHREC 2014 sketch-based 3D shape retrieval benchmark [17,18] and the other is the ModelNet dataset [41] from Princeton University. The SHREC 2014 benchmark contains 8,987 3D shapes from 171 categories. The number of 3D shapes in each category varies from 1 to 632. In order to match the object categories of depth image dataset, the database is constructed by selecting 3D models in corresponding categories, which contains 2,174 3D shapes from 7 categories. All of the 2,174 3D models will be used for both training and testing. For another test on ModelNet dataset, a subset of 4,315 clean 3D models from 10 categories (ModelNet10) (corresponding to the sample categories from NYU Depth V2 dataset) are used as the 3D shape dataset. The ratio for samples in training set and test set is 1:1.

4.2. Compared approaches

To evaluate the performance, we compare our method with other state-of-the-art methods [46] that address the same depth-image-based 3D shape retrieval problem. Besides that, we select some state-of-the-art transfer learning approaches used for similar cross-modality retrieval task as compared methods. The approach, that retrieves models for given depth image queries using extracted hand-crafted features, is also presented for comparison.

4.2.1. The non-transfer approach (NT)

Without any further learning and processing, the non-transfer approach directly retrieve relevant 3D shapes utilizing the original extracted hand-crafted features as representation for depth image and 3D shapes. The shape retrieval is performed by computing the Euclidean distance between the hand-crafted features of depth image queries and shapes.

4.2.2. Transfer learning methods

Correlation matching (CM) and Semantic Correlations Matching (SCM) are state-of-the-art transfer learning approaches on cross-modality multimedia retrieval [29]. CM learns correlations between two domains with canonical correlation analysis (CCA) [10], and the maximal cross-modality correlations are used for retrieval. SCM is an extension of CM with a higher level of abstraction of the domains, where logistic regression is performed on the maximal correlations that are obtained from CM. Though CM and SCM was not used on depth-image-shape retrieval problem before, they have been validated in similar cross-modality feature learning task. In our experiments, we apply CM and SCM on depth-image-based shape retrieval task with the source code released by the authors.

4.2.3. Pairwise neural network (PNN)

The state-of-the-art approach for cross-domain 3D shape retrieval with depth images as queries. Zhu et al. [46] trained a pair of neural networks based on the features extracted from samples independently. Identical target vectors are assigned for samples from the same category for both networks. The outputs from the hidden layers of the neural network pair were extracted as

Please cite this article as: J. Zhu et al., Learning domain-invariant feature for robust depth-image-based 3D shape retrieval, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.09.041

domain-invariant features to retrieve relevant 3D models for given depth images. The source code is provided by authors.

4.2.4. Proposed method (Ours)

We construct our model using two deep neural networks, one for each domain. Domain-invariant features are learned from the two domains by a loss function that minimizes the intra-class distance while maximizing the inter-class variance between the cross-domain data. Given a depth image query, relevant 3D shapes are retrieved based on the similarity of the domain-invariant features generated by the final outputs of the last layers in the networks.

4.3. Evaluation protocol

In our experiments, the performance of all compared approaches for retrieval are evaluated based on the common widely used six evaluation metrics [35] and precision-recall (PR) curves.

4.3.1. Evaluation metrics

The evaluation metrics include six quantitative statistics (Nearest Neighbor, First Tier, Second Tier, E-Measure, Discounted Cumulated Gain and Average Precision) to evaluate the match retrieval results. Nearest Neighbor (NN) is the average percentage of the closest 3D models that belong to the same category as the depth image queries. Supposed C denotes the total number of 3D models that are in the same category of the query's, First Tier (FT) is the mean percentage of 3D models that are in the same category as the queries' within the top $|C|-1$ matches, and Second Tire (ST) is the mean percentage of relevant 3D models within the top $2*|C|-1$ matches for all queries. E-Measure (E) is the mean of E_q computed by the precision (P_{32}) and recall (R_{32}) of the first 32 retrieved models for every query as $E_q = \frac{2}{\frac{1}{P_{32}}+\frac{1}{R_{32}}}$. With the assumption that matches appearing closer to the top of the ranked list are more relevant, larger weights are assigned for the matches near the top, and Discounted Cumulated Gain (DCG) is the average weighted sum of correct results of a query in the ranked list. Average Precision (AP) computes the average precision of the retrieval for all queries. For all six statistics, higher values indicate better performance.

4.3.2. Precision-Recall (PR) curves

To visualize the performance of retrieval results, precision-recall curves are used to indicate the relation between precision and recall for all depth image queries. They are generated by calculating the standard 11-point interpolated average precision at different recall levels of 0.0, 0.1, $\cdots$, 0.9, 1.0. For a recall level i, interpolated precision is the maximum precision at any recall level that is larger than or equal to i. After obtaining the 11 average precision points, we plot them on a two-dimensional graph with recall on the x-axis and precision on the y-axis.

4.4. Experimental settings

Before the setup of our network model, we first extract features from depth images and 3D models, which are used as inputs for our networks. We follow the Sparse Coding Spatial Pyramid Matching (ScSPM) [42] framework to generate features for depth image network. After obtaining 21504-dimensional ScSPM features, Principal Component Analysis (PCA) [40] is applied to the features to reduce the dimension of the depth ScSPM features to 1000. For 3D shapes, we extract Local Depth Scale-Invariant Features Transform (LD-SIFT) [5] features, and then fit the LD-SIFT feature to a Bag-of-Words (BoW) model to get 1000-dimensional histogram features for each 3D model from the dataset.

In all experiments, our network model is constructed by two 3-layer neural networks with 500 hidden layer size and 1000 target layer size, one for each domain. We report results from two experiments in this paper. For the two experiments, depth images are all from NYU Depth V2 dataset, but 3D models are varied. In the first experiment, 3D models are selected from SHREC 2014 benchmark dataset, while in the other test, 3D models are collected from the ModelNet dataset. The network structure remains the same for the two tests. We can obtain a 1000-dimensional domain-invariant representation for both the depth image queries and the 3D models from our trained network model. When training the neural networks, the learning rate β and regularization term λ are set to different values in different experiments.

4.5. Shape retrieval on SHREC 2014 dataset

Following the experiment settings given by Zhu et al. [46], we test our method using 5-class samples (the first 5 categories as displayed in Table 1) and 7-class samples (as displayed in Table 1) from the datasets. Our model is trained with a depth image training set containing 50% of the depth images randomly selected from each category (670 for 5-category test and 1,014 for 7-category test). The rest of the depth images are used as for testing. All 3D models are used in both training and testing.

We compare our approach with recent PNN [46] and the state-of-the-art transfer learning methods: correlation matching (CM) and semantic correlations matching (SCM) [29]. We also compare our method with the non-transfer approach, which directly utilizes the original depth image extracted ScSPM features and 3D model extracted LD-SIFT features for retrieval. The statistic results are reported in Table 2 with six standard evaluation metrics and in Fig. 4 with precision-recall (PR) curves generated from all compared methods. The retrieval performance of our method is obtaining with the learning rate β and regularization term λ set to 0.02 and 0.0005, respectively.

Experimental results suggest that the proposed method achieves outstanding performance when comparing with other methods (PNN, CM, SCM and NT) on 7-category and 5-category datasets under the 6 metrics. As we can see in Table 2, without any learning process, the extracted hand-crafted features obtain 0.21 and 0.15 average precision on 5-category dataset and 7-category dataset respectively, demonstrating the limited discriminative power of the hand-crafted features. We also observe that applying CM or SCM on extracted hand-crafted features actually does not improve the retrieval performance. The reason might be that CM just learns a linear dependence between data from two domains, which does not represent the correlation between the depth image and shapes well. Though SCM learns a higher abstracted feature space using logistic regression for cross-domain data, the feature space is learned upon maximal correlations obtained by CM, so it is understandable that the performance using SCM on depth-image-based shape retrieval is not good. PNN uses neural networks to learn a mapping between cross-domain data and the defined target values, and gains much higher average precision (0.42) than NT, CM and SCM. PR-curves on Fig. 4 visualize the performance. Our method consistently leads large margin over other state-of-the-art compare methods on both six performance metrics and the PR-curves. This demonstrates the significant improved performance of our model over other methods on cross-domain retrieval and the importance of adding constraint on inter-class term in our loss function.

4.6. Shape retrieval on ModelNet 10 dataset

In this section, we use a subset of ModelNet dataset with 10-category 3D models as our 3D shape database (ModelNet10 dataset). The depth image queries are the same with those in the experiment on SHREC 2014 dataset. The number of samples in 10

Table 1
Number of samples in each category in the constructed dataset, where depth images and 3D models are from NYU Depth V2 dataset and SHREC 2014 benchmark, respectively.

Category	bathtub	bed	chair	desk	dresser	night stand	table
Depth images	57	318	654	197	111	148	539
3D models	109	467	712	204	203	218	402

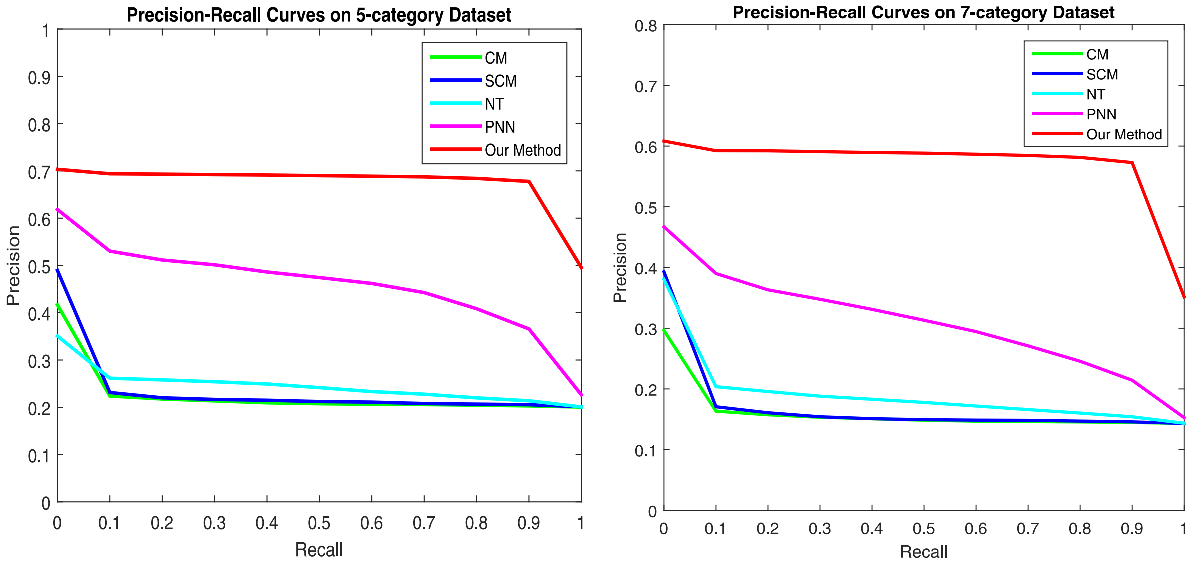

Fig. 4. Precision-Recall plots for performance comparison of state-of-the-art methods on NYU Depth V2 dataset and SHREC 2014 benchmark. The left plot shows the comparisons on 5-category dataset and the right one shows the comparisons on 7-category dataset.

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

	NN	FT	ST	DCG	E	AP
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
Ours	**0.53**	**0.56**	**0.75**	**0.84**	**0.07**	**0.63**
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
Ours	**0.40**	**0.44**	**0.61**	**0.79**	**0.05**	**0.51**

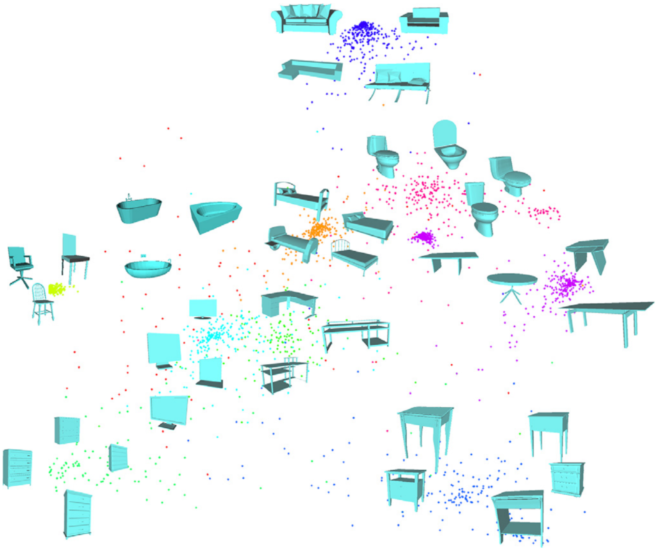

Fig. 5. Visualization of the learned domain-invariant features using our method. Points represent the learned features for both depth image and 3D shape domains. In the figure, we only show some 3D shapes as examples to demonstrate the corresponding class. Cross-domain data from the same category is assigned with the same color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

categories (*bathtub, bed, chair, desk, dresser, monitor, night stand, sofa, table, toilet*) of the two datasets are given in Table 3. Both the depth images and 3D models are split into training set and testing set with 1:1 ratio. Therefore, there are 1,261 depth image and 2,160 3D models in the training set, while 1,256 depth image and 2,155 3D models in the testing set.

When training our model on ModelNet10 dataset, we set the learning rate β and regularization term λ to 0.001 and 0.0001, respectively. We first provide an example to visualize our learned domain-invariant features in Fig. 5 by simply reducing the dimension of the learned features to two using PCA algorithm. Points in the same color represent the cross-domain samples from the same category, and some example shapes are placed next to their corresponding points for better review. As we can see from the visualization figure, most of the cross-domain samples have similar features if they are in the same class. The effective domain-invariant feature learning guarantees the good performance when applying our proposed method for depth image-based shape retrieval.

We also present the statistic results in Table 4 and compare the precision-recall curves against the-state-of-the-art methods in Fig. 6. From the Fig. 6, we can see that our method significantly outperforms other compared methods. More importantly, the whole curve decreases much slower than other methods when

Please cite this article as: J. Zhu et al., Learning domain-invariant feature for robust depth-image-based 3D shape retrieval, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.09.041

Table 3
Number of samples in each category of the constructed dataset, where depth images are from NYU Depth V2 dataset and 3D models come from ModelNet10 dataset.

Category	bathtub	bed	chair	desk	dresser	monitor	night stand	sofa	table	toilet
Depth images	57	318	654	197	111	118	148	307	539	68
3D models	148	514	869	237	244	485	245	690	445	418

Table 4
Performance metrics comparison of depth-image-based 3D model retrieval on the ModelNet10 dataset and NYU Depth V2 dataset.

	NN	FT	ST	DCG	E	AP
NT	0.14	0.13	0.25	0.68	0.01	0.14
CM	0.10	0.12	0.24	0.68	0.01	0.13
SCM	0.18	0.18	0.31	0.71	0.02	0.20
PNN	0.14	0.13	0.26	0.65	0.03	0.15
Ours	**0.33**	**0.33**	**0.46**	**0.76**	**0.04**	**0.41**

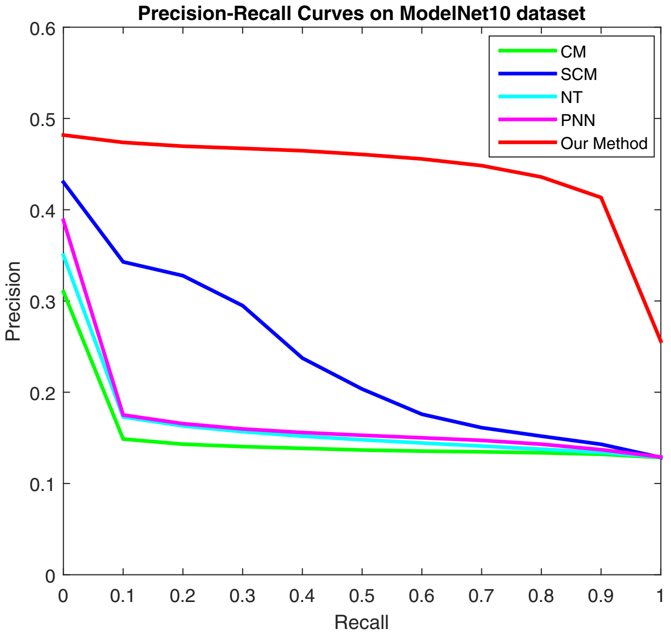

Fig. 6. Precision-Recall plot for performance comparison of state-of-the-art compared methods on NYU Depth V2 dataset and ModelNet 10 dataset.

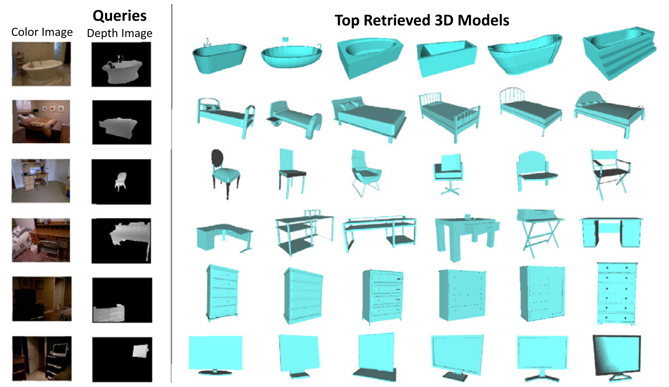

Fig. 7. Examples of successful retrieval using our proposed method (3D models are from the ModelNet10 dataset). Color images are provided for better view, but only the depth images are used as queries. From top to bottom the queries are *bathtub, bed, chair, desk, dresser* and *monitor*. Each row shows the top 6 retrieval results for corresponding query. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

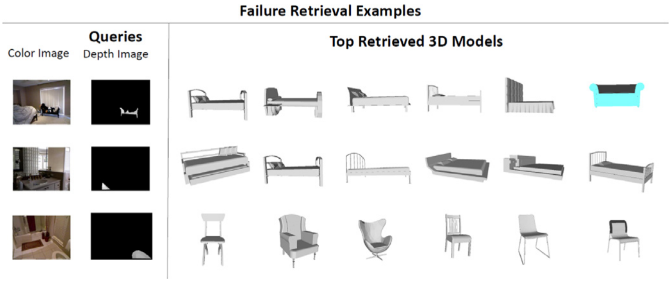

Fig. 8. Examples of failure retrieval. From top to bottom the queries are from *sofa, table* and *toilet*, following with their top 6 retrieval results. 3D Shapes in Cyan denote correct retrieval and shapes in Gray denote the incorrect retrieval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the recall increases, which suggests that our method is more stable. The performance gain of our method is more than 10% when recall reaches 1. As show in Table 4, our method performs best in every metric against other methods, which further demonstrates our method is superior. On the other hand, we observe that PNN surprisingly performed the same as the NT method. The cause might come from the predefined target vector with random values, which could be similar for different classes. The successful retrievals on both SHREC 2014 dataset and ModelNet10 dataset greatly demonstrate the robustness of our proposed method.

Finally, we visualize some successful retrieval examples for queries from different category in Fig. 7. Forbetter view, corresponding color image is provided for each query in the first column, but we did not use any information from color images. The second column shows the depth image queries and each row shows the top 6 retrieval results. The retrieval results demonstrate our method is powerful in learning domain-invariant features. In addition, Fig. 8 presents some failure retrieval examples, in which queries from *sofa* (in first row) and *table* (in second row) retrieve *bed* as their top results. A query from *toilet* might be mismatched to some *chair* models. We conclude the possible reasons as 1) the significant incompleteness of the object in some depth images, for example, the second and third query in Fig. 8 only provide a very small part of object; 2) since the depth images are captured from the real-world environment, there is some occlusion of objects, e.g. the first query of *sofa* is fully covered by a lot of stuffs, such as toys, cloths and cushions, making it have similar view with *bed*. Although there are some failure cases, the statics evaluation strongly demonstrates the effectiveness of our method with superior performance on depth image-based shape retrieval.

5. Translational applications

A method that optimizes 3D-shape identification, as rendered by depth images, has great capacity to facilitate computer vision-focused object identification. More specifically, a system that is able to minimize the computational costs of time and memory will enable the re-allocation of processing power to additional undertakings. This becomes pertinent to translational medical applications that leverage deep learning techniques with real-time object detection/categorization needs. In these approaches, identified 3D

Please cite this article as: J. Zhu et al., Learning domain-invariant feature for robust depth-image-based 3D shape retrieval, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.09.041

shapes may aid in the algorithmic strategies deployed to fuse such information with other inputs, towards the goal of minimum delay semantic labelling with object identification. This operation could take place in parallel with local scene understanding, juxtaposing object identity with concurrent spatial understanding in dynamic environments.

In the visually impaired setting, wearable devices that are configured as assistive technology platforms are one such application where these approaches become attractive. Systems that focus on real-time spatial understanding along with on-board navigation instructions will require expedited obstacle identification methodologies that are performed simultaneously with a host of additional tasks and sub-tasks [26,30,31]. While deep learning and computer vision techniques are still in their nascent stages, as applied to the aforementioned use cases, these systems have the potential to drastically improve the mobility profile of those with low vision or blindness and to reverse the untoward co-morbidities that arise as a result of increased immobility, often a byproduct of trips, falls, and injuries [9,21,24,28,33,37].

6. Conclusions

In this work, we propose to learn a domain-invariant feature using deep learning techniques in an effort to address the challenging problem of depth-image-based 3D shape retrieval. In order to minimize the discrepancy between highly diverged depth images and 3D models, we build a neural network pair for the depth images and 3D models, while connecting the network pair at their target layers. Instead of enforcing identical fixed target values at the output layers of both networks, we add a constraint on the inter-class and intra-class margins in the loss function, enabling the neural network pair to learn a feature space towards minimum intra-class variance and maximum inter-class margin during the training process. Our proposed method has been successfully validated on the NYU Depth Dataset V2, the extended SHREC 2014 3D shape retrieval benchmark, and the Princeton ModelNet dataset. The experimental results have shown that our approach outperforms the state-of-the-art PNN method, other transfer learning methods, and the paradigm that retrieves 3D models by directly using the original extracted hand-crafted features from depth images (ScSPM) and 3D models (3D SIFT). The large improvement margins of our method over other techniques demonstrate its excellent capacity for cross-domain data representation. Moreover, since our model does not require accurate correspondence information across different domains, it can be easily generalized to solve real-world problems, including those that focus on translational medical applications.

References

[1] F. Aminzadeh, W. Sandham, M. Leggett, Geophysical Applications of Artificial Neural Networks and Fuzzy Logic, vol. 21, Springer Science & Business Media, 2013.
[2] T.G. Barbounis, J.B. Theocharis, M.C. Alexiadis, P.S. Dokopoulos, Long-term wind speed and power forecasting using local recurrent neural network models, Energy Conv. IEEE Trans. 21 (1) (2006) 273–284.
[3] A.M. Bronstein, M.M. Bronstein, B. Bustos, U. Castellani, M. Crisani, B. Falcidieno, L.J. Guibas, I. Kokkinos, V. Murino, et al., Shrec 2010: Robust Feature Detection and Description Benchmark, 2010.
[4] A.M. Bronstein, M.M. Bronstein, L.J. Guibas, M. Ovsjanikov, Shape google: geometric words and expressions for invariant shape retrieval, ACM Trans. Graph. 30 (1) (2011) 1.
[5] T. Darom, Y. Keller, Scale-invariant features for 3-d mesh models, Image Process. IEEE Trans. 21 (5) (2012) 2758–2769.
[6] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, M. Alexa, Sketch-based shape retrieval.
[7] J. Feng, Y. Wang, S.-F. Chang, 3d shape retrieval using a single depth image from low-cost sensors, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–9.
[8] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from rgb-d images for object detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 345–360.
[9] R.H. Harwood, Visual problems and falls, Age Ageing 30 (2001) 13–18.
[10] H. Hotelling, Relations between two sets of variates, Biometrika (1936) 321–377.
[11] K. Hu, Y. Fang, 3d laplacian pyramid signature, in: Computer Vision-ACCV 2014 Workshops, Springer, 2014, pp. 306–321.
[12] M. Kazhdan, B. Chazelle, D. Dobkin, A. Finkelstein, T. Funkhouser, A reflective symmetry descriptor, in: European Conference on Computer Vision, Springer, 2002, pp. 642–656.
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[14] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, IEEE, 2006, pp. 2169–2178.
[15] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[16] B. Li, Y. Lu, A. Godil, T. Schreck, B. Bustos, A. Ferreira, T. Furuya, M.J. Fonseca, H. Johan, T. Matsuda, et al., A comparison of methods for sketch-based 3d shape retrieval, Comput. Vis. Image Understand. 119 (2014) 57–80.
[17] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan, et al., Extended large scale sketch-based 3d shape retrieval (2014).
[18] 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), 2014, pp. 121–130.
[19] W. Li, L. Duan, D. Xu, 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) (2014) 1134–1148.
[20] F. Liu, L. Yang, A novel cell detection method using deep convolutional neural network and maximum-weight independent set, in: Medical Image Computing and Computer-Assisted InterventionMICCAI 2015, Springer, 2015, pp. 349–357.
[21] S.R. Lord, S.T. Smith, J.C. Menant, Vision and falls in older people: risk factors and intervention strategies, Clin. Geriatr. Med. 26 (4) (2010) 569–581.
[22] D.G. Lowe, Object recognition from local scale-invariant features, in: International Conference on Computer Vision, vol.2, Ieee, 1999, pp. 1150–1157.
[23] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1–9.
[24] G. McLean, B. Guthrie, S.W. Mercer, D.J. Smith, et al., Visual impairment is associated with physical and mental comorbidities in older adults: a cross-sectional study, BMC Med. 12 (1) (2014) 181.
[25] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, ACM Trans. Graph. 21 (4) (2002) 807–832.
[26] N. Patel, Y. Pan, Y. Li, F. Khorrami, J. Rizzo, T. Hudson, et al., Robust object detection and recognition for the visually impaired, 1st Intl Wksh On Deep Learning for Pattern Recognition (DLPR) (2016).
[27] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: The IEEE International Conference on Computer Vision (ICCV), 2015.
[28] M.L. Popescu, H. Boisjoly, H. Schmaltz, M.-J. Kergoat, J. Rousseau, S. Moghadaszadeh, F. Djafari, E.E. Freeman, Age-related eye disease and mobility limitations in older adults, Invest. Ophthalmol. Vis. Sci. 52 (10) (2011) 7168–7174.
[29] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: International Conference on Multimedia, ACM, 2010, pp. 251–260.
[30] J.-R. Rizzo, T.E. Hudson, R.A. Shoureshi, Smart wearable systems for enhanced monitoring and mobility., in: International Conferences on Modern Materials and Technologies (CIMTEC); 2016; Perugia, Italy, 2016.
[31] J.-R. Rizzo, Y. Pan, T. Hudson, E. Wong, Y. Fang, Sensor fusion for ecologically valid obstacle identification: building a comprehensive assistive technology platform for the visually impaired., 7th Intl Conf on Modeling, Simulation & Applied Optimization (ICMSAO); 2017; Sharjah, U.A.E.
[32] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain., Psychol. Rev. 65 (6) (1958) 386.
[33] S. Sengupta, A.M. Nguyen, S.W. Van Landingham, S.D. Solomon, D.V. Do, L. Ferrucci, D.S. Friedman, P.Y. Ramulu, Evaluation of real-world mobility in age-related macular degeneration, BMC Ophthalmol. 15 (1) (2015) 9.
[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localization and detection using convolutional networks,(2013) arXiv:1312.6229.
[35] P. Shilane, P. Min, M. Kazhdan, T. Funkhouser, The princeton shape benchmark, in: Shape Modeling Applications, IEEE, 2004, pp. 167–178.
[36] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[37] S.W. Van Landingham, J.R. Willis, S. Vitale, P.Y. Ramulu, Visual field loss and accelerometer-measured physical activity in the united states, Ophthalmology 119 (12) (2012) 2486–2492.
[38] Y. Wang, J. Feng, Z. Wu, J. Wang, 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, Springer, 2014, pp. 489–504.
[39] P. Werbos, Beyond regression: New tools for prediction and analysis in the behavioral sciences (1974).

[40] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1) (1987) 37–52.

[41] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.

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

[43] B. Yuhas, N. Ansari, Neural Networks in Telecommunications, Springer Science & Business Media, 2012.

[44] Z. Zhang, T. Tan, K. Huang, Y. Wang, Three-dimensional deformable-model-based localization and recognition of road vehicles, IEEE Trans. Image Process. 21 (1) (2012) 1–13.

[45] F. Zhu, L. Shao, Weakly-supervised cross-domain dictionary learning for visual recognition, Int. J. Comput. Vis. 109 (1–2) (2014) 42–59.

[46] J. Zhu, F. Zhu, E.K. Wong, Y. Fang, Learning pairwise neural network encoder for depth image-based 3d model retrieval, in: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, ACM, 2015, pp. 1227–1230.

Please cite this article as: J. Zhu et al., Learning domain-invariant feature for robust depth-image-based 3D shape retrieval, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.09.041