Deep Correlated Metric Learning for Sketch-Based 3D Shape Retrieval

Guoxian Dai, Jin Xie, Fan Zhu, Yi Fang*
NYU Multimedia and Visual Computing Lab
Department of Electrical and Computer Engineering, NYU Abu Dhabi, UAE
Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, USA
Department of Computer Science and Engineering, NYU Tandon School of Engineering, USA

Abstract

The explosive growth of 3D models has led to the pressing demand for an efficient searching system. Traditional model-based search is usually not convenient, since people don't always have 3D model available by side. The sketch-based 3D shape retrieval is a promising candidate due to its simpleness and efficiency. The main challenge for sketch-based 3D shape retrieval is the discrepancy across different domains. In the paper, we propose a novel deep correlated metric learning (DCML) method to mitigate the discrepancy between sketch and 3D shape domains. The proposed DCML trains two distinct deep neural networks (one for each domain) jointly with one loss, which learns two deep nonlinear transformations to map features from both domains into a nonlinear feature space. The proposed loss, including discriminative loss and correlation loss, aims to increase the discrimination of features within each domain as well as the correlation between different domains. In the transfered space, the discriminative loss minimizes the intra-class distance of the deep transformed features and maximizes the inter-class distance of the deep transformed features at least a predefined margin within each domain, while the correlation loss focuses on minimizing the distribution discrepancy across different domains. Our proposed method is evaluated on SHREC 2013 and 2014 benchmarks, and the experimental results demonstrate the superiority of our proposed method over the state-of-the-art methods.

Introduction

With the advanced development of digitalization techniques, 3D models are widely available in our daily lives across many areas, such as 3D printing, medical imaging and entertainment. The vast amounts of 3D model lead to the pressing demand for effectively searching the desired 3D models. Traditional text-based search could not work well for two main reasons, 1) Only a small number of 3D models are available with text descriptions, which is too limited to retrieve desired 3D models. 2) It is often very hard to describe the very detailed information of complex 3D models simply with texts. Therefore, researchers proposed content-based 3D model retrieval framework, which mainly includes two categories, example-based 3D shape retrieval and sketch-based 3D shape retrieval. Most of the existing works fall into the first group, which is provided with a query 3D model and returns similar models. Example-based 3D shape retrieval is quite straightforward, however, not convenient, since people usually don't always have the desired 3D model example available at hand. Recently, the sketch-based 3D shape retrieval has received more and more attention from computer vision and computer graphics communities. Compared to the example-based framework, the sketches are much more convenient and easier to get, even a young kid could draw simple and comprehensive sketches. Apart from simpleness, sketch is also informative since it is very easy for people to understand the class labels for simple query sketches.

Despite of all the advantages of sketch-based 3D shape retrieval, actually, it is a quite challenging problem. First, sketch and 3D shape come from two different modalities with huge gap. And features extracted from both modalities follow quite different distributions, which makes it very difficult to directly retrieve 3D shapes from query sketches. Secondly, sketches are usually very simple with only several lines. The simpleness, on the contrary, also makes the sketch contain very limited information. The 3D shapes look visually similar as the query sketches only from some certain view angles. Generally, it is very hard to find the "best views" to project 3D shapes, which makes both sketches and 3D shapes visually similar.

The main challenge for sketch-based 3D shape retrieval is the domain discrepancies between these two modalities. In this work, we propose a novel deep correlated metric learning (DCML) method to mitigate the discrepancies between sketch and 3D shape domains. We first extract low-level features for both sketches and 3D shapes. For sketch, we use pre-trained AlexNet (Krizhevsky, Sutskever, and Hinton 2012) to extract features. For 3D shape, we extract 3D-SIFT feature (Darom and Keller 2012), which is further encoded by locality-constrained linear coding (LLC) (Wang et al. 2010) to get a global shape descriptor. Then we learn two deep neural networks to transform the raw features of both domains into a nonlinear feature space, mitigating the domain discrepancy as well as maintaining the discriminations. The loss of the proposed network includes two terms, discriminative term which is constructed with the pairwise distance within each domain and correlation term which is con-

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

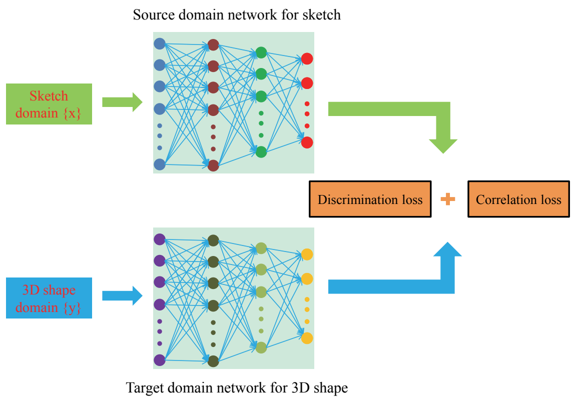

Figure 1: The detailed framework of our proposed deep correlated metric learning network. The whole network structure mainly includes two components, source domain network and target domain network.

structed with the pairwise distance across different domains. The former one minimizes the variations of the deep learned features from the same class and maximizes the variations of the deep learned features from different classes within each domain; the latter one aims to alleviate the domain discrepancy, making the distributions of both domains as consistent as possible. We verify our proposed method on two large scale benchmarks, SHREC 2013 and 2014 datasets, and the experimental results demonstrate the superiority of our proposed method.

The main contribution of our work is that we develop a novel deep correlated metric learning method for sketch-shape cross-domain retrieval, which jointly trains two deep neural networks with one loss to learn two deep nonlinear transformations, one for each domain. The deep learned transformations could map both sketch and shape features from the original space into a nonlinear feature space, maintaining both discrimination within each domain and correlation across different domains. Thus, the distance of the deep learned features from the same class is minimized while the distance of the deep learned features from different classes is maximized by a large margin within each domain; the distribution discrepancy is alleviated by optimizing pairwise across-domain distances. Thus, the deep transformed features are both discriminative within each domain and distribution-consistency across different domains. Therefore, our deep learned features could effective improve the performance of sketch-shape retrieval task.

Related work

Most of the existing works about 3D shape retrieval is the example-based framework, which could be roughly classified into three categories, projection based methods, diffusion based methods and deep learning based methods. For the projection based methods, 3D shapes are projected into a set of 2D images, so that classic image features are adopted to construct shape descriptor, such as LFD (Chen et al. 2003) and ED (Shih, Lee, and Wang 2007). For the diffusion based methods, 3D shape descriptors are derived based on heat diffusion or probability distribution of quantum particles, such as HKS (Sun, Ovsjanikov, and Guibas 2009) and WKS (Aubry, Schlickewei, and Cremers 2011). All the aforementioned methods are just hand-crafted. Inspired by the great success of deep learning in 2D images areas, deep learning is also introduced to 3D areas for shape retrieval. (Xie et al. 2015) use discriminative auto-encoder to extract robust shape descriptor in the hidden layers. (Bai et al. 2015) also proposed a two layer encoding framework for 3D shape matching.

Except for the example-based framework, the sketch-based framework is another promising candidate for retrieving desired 3D shapes. Currently, there are very few works about sketch-based 3D shape retrieval. (Daras and Axenopoulos 2010) proposed a unified 3D shape retrieval system supporting multimedia queries by projecting 3D models into a group of 2D images. The similarities among different models are determined by features extracted from 2D images. (Bronstein et al. 2011) applied bag-of-features (BoF), which was widely used in 2D computer vision, for 3D shape retrieval. In addition, (Eitz et al. 2012) further adopted BoF with Gabor local line based feature (GALIF) for sketch-based 3D shape retrieval. Apart from BoF encoding scheme, (Biasotti et al. 2015) applied the LLC (Wang et al. 2010) scheme for textured 3D shape retrieval. Apart from the aforementioned algorithms, large scale benchmark datasets have also been recently proposed to evaluate the performance of different methods, such as SHREC 2013 (Li et al. 2014) and SHREC 2014 (Li et al. 2015). Sketches of both datasets come from a latest large sketch collection (Eitz et al. 2012). The 3D shapes of SHREC 2013 are mainly collected from Princeton Shape Benchmark (Shilane et al. 2004), while the shapes of SHREC 2014 come from various sources, such as (Tatsuma, Koyanagi, and Aono 2012) and (Li et al. 2012). Different comparison results are reported for both datasets. For SHREC 2013, the best reported result in (Li et al. 2014) is from view clustering and shape context matching (SBR-VC). For SHREC 2014, the best reported result in (Li et al. 2015) is from overlapped pyramid of HOG and similarity constrained manifold ranking, by Tatsuma *et al.*

Recently, deep metric learning has received more and more attention from the computer vision community. Compared to traditional metric learning with a simple linear transformation, deep metric learning inherits advantages from the existing deep learning techniques (Krizhevsky, Sutskever, and Hinton 2012) and could learn much more complex, powerful nonlinear transformation. (Chopra, Hadsell, and LeCun 2005) adopts Siamese network to learn image similarities for face verifications. Generalizing the ideas in both (Chopra, Hadsell, and LeCun 2005) and large margin distance metric learning (Weinberger and Saul 2009), (Hu, Lu, and Tan 2014) proposed a discriminative deep metric learning for face verification, with a marginal distance between positive pair and negative pair. Instead of randomly selecting training pairs in (Hu, Lu, and Tan 2014), (Song et al. 2015) considers all the possible positive pairs and negative pairs in the training set for deep metric learning. Different from deep metric learning with Siamese network in (Chopra, Hadsell, and LeCun 2005; Hu, Lu, and Tan 2014), which adopts pairwise training strategy with two input samples, (Hoffer and Ailon 2015) adopts triplet structure for deep metric learning. Specifically, the triplet structure uses three identical networks with three input examples, one base example, its positive example and negative example. Except for the application of deep metric learning in 2D image areas, it is also introduced to 3D shape areas (Wang, Kang, and Li 2015). (Wang, Kang, and Li 2015) extended the Siamese network for sketch-based 3D shape retrieval by using two based Siamese networks, one for sketch domain and one for 3D shape domain. Their method is based on a strong assumption that all the 3D models are stored upright, which makes it much easier to choose the project view of 3D model. Such assumption can hardly be guaranteed in real application, and without such assumption, it is actually very hard to choose the "best" projection view. The projection results could change greatly, as the view changes.

It is noted that (Wang, Kang, and Li 2015) used similar framework as our proposed method, by extending Siamese network (Chopra, Hadsell, and LeCun 2005) for sketch-based 3D shape retrieval. There are several differences between (Wang, Kang, and Li 2015) and our proposed method: 1) (Wang, Kang, and Li 2015) needs to project 3D model into two different views with a strong assumption that all models are stored upright as default. In fact, the projection results could change dramatically as the projection view changes. However, our proposed method doesn't need projection, neither does the upright assumption. 2) Our proposed method employs a marginal distance for metric learning to increase the discrimination of the deep learned features, while (Wang, Kang, and Li 2015) does not. 3) Our proposed method could outperform (Wang, Kang, and Li 2015) on both SHREC 2013 and 2014 datasets.

Proposed approach

We propose a novel deep correlated metric learning method for sketch-based 3D shape retrieval. Fig. 1 shows the detailed framework of our proposed method. The proposed networks consist of two components, one for sketch domain, referred as source domain network (SDN), and one for 3D shape domain, referred as target domain network (TDN). The proposed method trains both deep neural networks simultaneously with one loss. The loss function includes two terms, discrimination term and correlation term, which minimizes intra-class variations and maximizes inter-class variations within each domain and guarantees the distribution-consistency across different domains.

The proposed method mainly includes two steps: 1) Extracting low-level features for both sketches and 3D shapes. 2) Learning two deep nonlinear transforms to map features of both domains from the original space into a nonlinear feature space, increasing the discrimination of features within each domain as well as mitigating the discrepancy across different domains. The details for each step are introduced as follows.

Feature extraction

Features for both sketches and shapes are extracted separately.

Sketch: Inspired by the outstanding performance of convolutional neural network (CNN) in feature learning (Krizhevsky, Sutskever, and Hinton 2012), we fine-tune AlexNet (Krizhevsky, Sutskever, and Hinton 2012) on sketch dataset and then extract the features in "*fc7*" layer. The feature dimension is 4096.

3D shape: Inspired by Lowe's SIFT (Lowe 2004) in 2D images, (Darom and Keller 2012) extended it into 3D mesh and proposed 3D-SIFT by detecting interest points. We first extract 3D-SIFT for 3D shapes, which are further encoded with the LLC (Wang et al. 2010) scheme to get a global shape descriptor. Readers could refer to (Darom and Keller 2012) and (Wang et al. 2010) for more details about 3D-SIFT and LLC.

Deep correlated metric learning

We denote training examples from source domain (sketch domain) and target domain (3D shape domain) as $S = \{x_1, x_2, x_3, ...\}$ and $T = \{y_1, y_2, y_3, ...\}$, respectively. The transfer functions for SDN and TDN are denoted as $f^s : x \to f^s(x)$ and $f^t : y \to f^t(y)$, respectively. In addition, W_k^s, W_k^t and b_k^s, b_k^t are the weights and bias, connecting layer k and layer $k+1$ of SDN and TDN, respectively; the activations of the i-th example x_i from S and j-th example y_j from T in the k-th layer of SDN and TDN are denoted as $a_k^{i,s}$ and $a_k^{j,t}$, respectively,

$$\begin{aligned} a_{k+1}^{i,s} &= \sigma(W_k^s a_k^{i,s} + b_k^s) = \sigma(r_{k+1}^{i,s}) \\ a_{k+1}^{j,t} &= \sigma(W_k^t a_k^{j,t} + b_k^t) = \sigma(r_{k+1}^{j,t}) \end{aligned} \tag{1}$$

where $\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, $r_{k+1}^{i,s} = W_k^s a_k^{i,s} + b_k^s$, $r_{k+1}^{j,t} = W_k^t a_k^{j,t} + b_k^t$. K_s and K_t denote the total number of layers for SDN and TDN, respectively. Thus, the nonlinear transfer function $f^s(x_i)$ and $f^t(y_j)$ across K_s and K_t layers of SDN and TDN respectively, can be represented as follows,

$$f^s(x_i) = a_{K_s}^{i,s} \quad f^t(y_j) = a_{K_t}^{j,t}. \tag{2}$$

The features extracted from different domains, sketch and 3D shape, suffer the domain discrepancy. Such discrepancy makes it very difficult to directly conduct across-domain retrieval. To effectively perform cross-domain retrieval, the features from both domains should address the following two issues: 1) Within each domain, the features should be as discriminative as possible, 2) Across different domains, the distributions of features from both domains should be as consistent as possible. To this end, we proposed a novel deep correlated metric learning method to mitigate the discrepancy across different domains as well as increase the discrimination within each domain. The proposed method learns two distinct deep neural networks (different weights and different structures) simultaneously to transform features from both domains into a nonlinear feature space. The proposed loss L mainly includes two terms, discriminative term L^d and correlation term L^c,

$$L = \alpha L^d + (1-\alpha)L^c + \lambda(\|W^s\|_F^2 + \|W^t\|_F^2) \tag{3}$$

where L^d aims to minimize intra-class distance of deep transformed features and maximize inter-class distance of deep transformed features with a predefined margin h within each domain. And L^c optimizes the pairwise across-domain distance to mitigate the distribution inconsistency across different domains. Parameter α is the weight to balance between discrimination term and correlation term. $W^s = \{W_1^s, W_2^s, \cdots, W_{K_s}^s\}$ and $W^t = \{W_1^t, W_2^t, \cdots, W_{K_t}^t\}$. $\|W^s\|_F^2$ and $\|W^t\|_F^2$ denote the Frobenius norm of W^s and W^t respectively, which are used to avoid over-fitting.

Discrimination term The discrimination term L^d aims to minimize intra-class distance and maximize inter-class distance within each domain, where L_s^d and L_t^d denote source domain discriminative loss and target domain discriminative loss, respectively,

$$L^d = L_s^d + L_t^d. \tag{4}$$

The source domain discriminative term L_s^d could be rewritten as follows:

$$\begin{aligned} L_s^d &= \sum_{(x_i,x_j)\in P^s} d_+^s(x_i,x_j) + \sum_{(x_i,x_j)\in N^s} d_-^s(x_i,x_j) \\ d_+^s(x_i,x_j) &= \|f^s(x_i) - f^s(x_j)\|_2^2 \\ d_-^s(x_i,x_j) &= \max\{0, h - \|f^s(x_i) - f^s(x_j)\|_2^2\} \end{aligned} \tag{5}$$

where P^s and N^s denote the sets of positive pair and negative pair in source domain S, respectively. The first term minimizes the distance of positive pairs, while the second term is a hinge loss, which pushes away the distance of negative pairs at least a predefined margin h. Similarly, we could formulate L_t^d as follows:

$$\begin{aligned} L_t^d &= \sum_{(y_i,y_j)\in P^t} d_+^t(y_i,y_j) + \sum_{(y_i,y_j)\in N^t} d_-^t(y_i,y_j) \\ d_+^t(y_i,y_j) &= \|f^t(y_i) - f^t(y_j)\|_2^2 \\ d_-^t(y_i,y_j) &= \max\{0, h - \|f^t(y_i) - f^t(y_j)\|_2^2\} \end{aligned} \tag{6}$$

where P^t and N^t denote the sets of positive pair and negative pair in target domain T, respectively.

The overall discriminative loss L^d minimizes intra-class distance and maximizes inter-class by a large margin within both source and target domains.

Correlation term Features from both domains follow different distributions, which makes it hard to directly retrieve objects across different modalities. Thus, a correlation term is further imposed to maintain the distribution consistency across different domains. The correlation term includes two types of pairwise across-domain distances, L_1^c and L_2^c,

$$\begin{aligned} L^c &= L_1^c + L_2^c \\ L_1^c &= \sum_{(x_i,y_j)\in P^c} d_+^c(x_i,y_j) + \sum_{(x_i,y_j)\in N^c} d_-^c(x_i,y_j) \\ L_2^c &= \sum_{c^s,c^t} \sum_{\substack{\forall x_i,x_j\in c^s \\ \forall y_i,y_j\in c^t}} R(x_i,x_j,y_i,y_j) \\ &\quad - \sum_{c^s,d^t}^{c\neq d} \sum_{\substack{\forall x_i,x_j\in c^s \\ \forall y_i,y_j\in d^t}} R(x_i,x_j,y_i,y_j) \end{aligned} \tag{7}$$

where P^c and N^c denote the sets of positive pairs and negative pairs across different domains. L_1^c directly minimizes the distances of positive across-domain pair examples, and maximizes the distances of negative across-domain pair examples at least h, making the distributions of two domains as similar as possible. c^s and c^t denote the set of examples with class label c for source domain and target domain respectively. Except for L_1^c, L_2^c is further imposed to guarantee the distribution-consistency across different domains. $d_+^c(x_i,y_j)$, $d_-^c(x_i,y_j)$ and R are listed as follows,

$$\begin{aligned} d_+^c(x_i,y_j) &= \|f^s(x_i) - f^t(y_j)\|_2^2 \\ d_-^c(x_i,y_j) &= \max\{0, h - \|f^s(x_i) - f^t(y_j)\|_2^2\} \\ R(x_i,x_j,y_i,y_j) &= \Big(\sqrt{\|f^s(x_i) - f^s(x_j)\|_2^2} - \\ &\quad \sqrt{\|f^t(y_i) - f^t(y_j)\|_2^2}\Big)^2. \end{aligned} \tag{8}$$

In L_2^c, x_i and x_j are from the same class, so do y_i and y_j. If (x_i,x_j) and (y_i,y_j) are from the same class, R is minimized; otherwise, R is maximized. Every term in the proposed loss is differentiable, thus the proposed DCML network could be optimized through back-propagation with the stochastic gradient descent method.

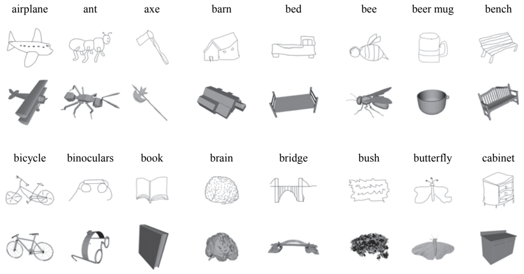

Figure 2: Example of sketches and shapes from SHREC 2013 dataset.

Experiments

Our proposed method is evaluated on two benchmark datasets, SHREC 2013 (Shilane et al. 2004) and SHREC 2014 (Li et al. 2015). We first visualize the distribution of our deep learned sketch and 3D shape features, then we draw precision-recall curve to visualize the retrieval performance of our proposed method and the state-of-the-art methods. Finally, we also calculate a number of standard metrics to evaluate our proposed method, including nearest neighbor (NN), first tier (FT), second tier (ST), discounted cumulative gain (DCG) and mean average precision (mAP). For all the evaluation criterion, our proposed method could outperform the state-of-the-art methods on both datasets, particularly on SHREC 2013 with a huge margin.

Implementation details

In this subsection, we mainly introduce the implementation details for our proposed method. For feature extraction, the sketch feature is extracted from the "*fc7*" layer of AlexNet (Krizhevsky, Sutskever, and Hinton 2012) with the feature size of 4096, while for 3D shape, the feature size of 3D-SIFT is set to 128. In addition, the size of the codebook for LLC is set to 4096, which is generated by regular k-means clustering. The network structures for the sketch and 3D shape domains are set to [4096 2000 1000 100] and [4096 2000 1000 500 100], respectively. In addition, the momentum rate is set to 0.1, learning rate is set to 0.015.

Retrieval on SHREC 2013 dataset

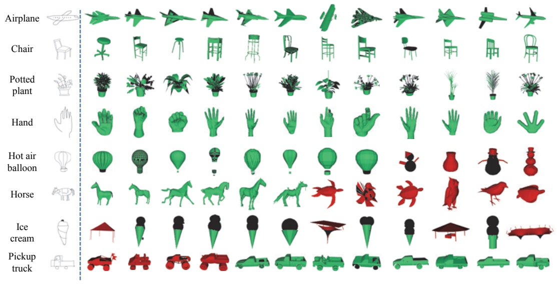

Figure 3: Illustration of retrieved examples on SHREC 2013 dataset.

In this section, we evaluate our proposed method on SHREC 2013 dataset. SHREC 2013 (Li et al. 2014) is large scale benchmark to evaluate algorithms for sketch-based 3D shape retrieval. The benchmark is created by collecting common classes from both the Princeton Shape Benchmark (Shilane et al. 2004) and sketch dataset (Eitz et al. 2012). Fig. 2 shows some examples of sketches and shapes from SHREC 2013 dataset. There are 1258 shapes and 7200 sketches in SHREC 2013, which are grouped into 90 classes in total. The number of shapes in each class is not equal, about 14 in average. While the number of sketches for each class is equal, 80 in total. For each class, there are 50 sketches for training and 30 for testing.

Fig. 3 shows some retrieved examples on SHREC 2013 dataset. The query sketches are listed on the left first column, namely, airplane, chair, potted plant, hand, hot air balloon, horse, ice cream and pickup truck. The top 12 retrieved models are listed on the right side, based on their ranking orders. The correct retrieved models are marked with green color, while the wrong results are marked with red color. As we can see from Fig. 3, for the classes of airplane, chair, potted plant and hand, all the 12 retrieved models are correct ; for the classes of hot air balloon and horse, the proposed method first retrieved correct examples, and then wrong examples, because there are too few examples in these two classes, less than 12. For the last two classes, ice cream and pickup truck, the proposed method retrieved several wrong examples, due to the geometrical similarity among these 3D shapes.

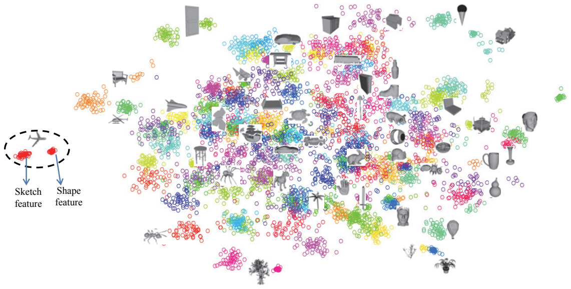

Figure 4: Visualization of the deep learned sketch features and shape features on SHREC 2013 dataset. The features are grouped in different colors by class labels.

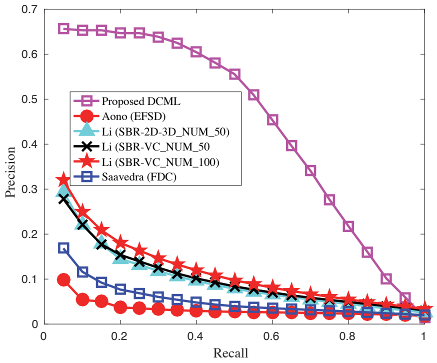

Figure 5: Performance comparisons on SHREC 2013 dataset.

We also visualize the deep learned features by using PCA to reduce the dimension from 100 to 2. Fig. 4 shows the visualization of the distributions of the deep transformed features. All the features are grouped in different colors by their class labels. As we can see in Fig. 4, features with the same label are grouped together, while features with different labels are separated away. Taking the class of airplane as an example, the features of the sketch samples are grouped together, and the features of the 3D shape samples are also grouped together; features from both domains are close to each other, meanwhile away from other classes. This is just a coarse visualization of our proposed method, which could roughly verify the effectiveness of our proposed method.

Precision-recall curve is a common metric to visually evaluate the retrieval performances of different algorithms. Fig. 5 shows the precision-recall curves of our proposed method as well as the state-of-the-art methods reported in (Li et al. 2014) on SHREC 2013 dataset. The magenta curve indicates our proposed method. As we can see in

Table 1: Performance comparisons of different evaluation criteria on SHREC 2013 dataset.

	NN	FT	ST	E	DCG	mAP
COMP	0.164	0.097	0.149	0.085	0.348	0.116
RCDM	0.279	0.203	0.296	0.166	0.458	0.250
KGLR	0.110	0.069	0.107	0.061	0.307	0.086
DSP	0.017	0.016	0.031	0.018	0.240	0.026
WCNN	0.405	0.403	0.548	0.287	0.607	0.469
DCML	**0.650**	**0.634**	**0.719**	**0.348**	**0.766**	**0.674**

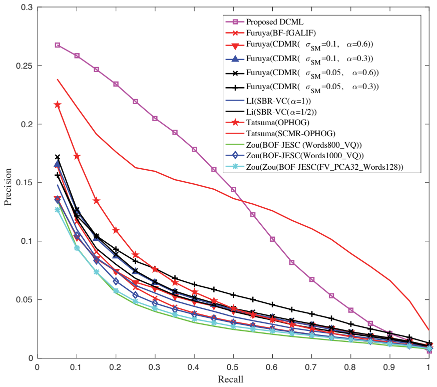

Figure 6: Performance comparisons on SHREC 2014 dataset.

Table 2: Performance comparisons of different evaluation criteria on SHREC 2014 dataset.

	NN	FT	ST	E	DCG	mAP
RCDM	0.109	0.057	0.089	0.041	0.328	0.054
COMP	0.095	0.050	0.081	0.037	0.319	0.050
TSB	0.160	0.115	0.170	0.079	0.376	0.131
WCNN	0.239	0.212	0.316	0.140	0.495	0.228
DCML	**0.272**	**0.275**	**0.345**	**0.171**	**0.498**	**0.286**

Fig. 5, our proposed method could significantly outperform state-of-the-art methods. As the recall value increases, the precision value of our proposed method is at least double times higher than those of the state-of-the-art methods. In addition, the precision value of our method is very stable and decreases very slowly when recall is small. Except for the precision-recall curve, standard metrics, including NN, FT, ST, E, DCG and mAP, are also calculated to evaluate our proposed method and the following methods, namely (Li et al. 2014)(COMP), (Furuya and Ohbuchi 2013)(RCDM), (Saavedra et al. 2012)(KGLR), (Sousa and Fonseca 2010)(DSP) and (Wang, Kang, and Li 2015)(WCNN). Table. 1 shows the comparison results on SHREC 2013 dataset. As we can see in Table. 1, for all the evaluation criteria, our proposed method could outperform the above methods. And the improvement is significant, about more than 30% gain in average compared to the best reported method (Wang, Kang, and Li 2015).

Retrieval on SHREC 2014 dataset

In this subsection, we test our proposed method on SHREC 2014 dataset (Li et al. 2015). SHREC 2014 is a large scale sketch track benchmark for sketch-based 3D shape retrieval, which consists of shapes from various datasets, such as SHREC 2012 (Li et al. 2012) and the Toyohashi Shape Benchmark (TSB) (Tatsuma, Koyanagi, and Aono 2012). The dataset has about 13680 sketches and 8987 3D models in total, grouped into 171 classes. SHREC 2014 dataset is quite challenging due to its diversity of categories and large variations within class. The number of shapes in each class varies from less than 10 to more than 300, while the number sketches for each class is equal to 80. For each group, there are 50 sketches for training and 30 for testing.

Fig. 6 shows the precision-recall curves of our proposed method with the state-of-the-art methods on SHREC 2014 dataset. The magenta curve denotes our proposed method. As we can see in Fig. 6, when the recall value is about less than 0.5, the precision value of our proposed method is higher than that of those methods; while when the recall value is about larger than 0.5, the precision value of our proposed method is inferior to that of those methods. Generally, people are more likely to examine top retrieved objects, instead of latter one, due to time and efforts, *etc.* Based on the above assumption, the precision-recall curve roughly indicates that our proposed method has better retrieval performance compared to the state-of-the-art methods.

In addition, we also compare our proposed method with the following methods, namely, (Furuya and Ohbuchi 2013), (Li et al. 2014), (Tatsuma, Koyanagi, and Aono 2012)(TSB), (Wang, Kang, and Li 2015)(WCNN). The metrics, including NN, FT, ST, E, DCG and mAP, are used to evaluate our proposed method. Table. 2 shows the comparison results of our proposed method and the above methods on SHREC 2014 dataset. In all evaluation criteria, our proposed method could outperform the aforementioned methods, which could verify the effectiveness of our proposed method.

Conclusions

In this work, we developed a novel deep correlated metric learning method for sketch-based 3D shape retrieval. Specifically, we first extracted raw features for sketches and 3D shapes separately. For sketches, we used pre-trained AlexNet to extract features; for 3D shapes, we extracted 3D-SIFT, which was further encoded by LLC to get a global description. Then our proposed method learned two deep nonlinear transformations (one for each domain), by simultaneously training two deep neural networks. The deep learned transformations mapped features from both domains into a nonlinear feature space, guaranteeing the discrimination of features within each domain and distribution-consistency of features across different domains. Our proposed method was evaluated on two benchmarks, SHREC 2013 and SHREC 2014. And the experimental results demonstrated superiority over the state-of-the-art methods.

References

Aubry, M.; Schlickewei, U.; and Cremers, D. 2011. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 1626–1633. IEEE.

Bai, X.; Bai, S.; Zhu, Z.; and Latecki, L. J. 2015. 3D shape matching via two layer coding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37(12):2361–2373.

Biasotti, S.; Cerri, A.; Aono, M.; Hamza, A. B.; Garro, V.; Giachetti, A.; Giorgi, D.; Godil, A.; Li, C.; Sanada, C.; et al. 2015. Retrieval and classification methods for textured 3D models: a comparative study. *The Visual Computer* 1–25.

Bronstein, A. M.; Bronstein, M. M.; Guibas, L. J.; and Ovsjanikov, M. 2011. Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Transactions on Graphics (TOG)* 30(1):1.

Chen, D.-Y.; Tian, X.-P.; Shen, Y.-T.; and Ouhyoung, M. 2003. On visual similarity based 3D model retrieval. In *Computer graphics forum*, volume 22, 223–232. Wiley Online Library.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 539–546. IEEE.

Daras, P., and Axenopoulos, A. 2010. A 3D shape retrieval framework supporting multimodal queries. *International Journal of Computer Vision* 89(2-3):229–247.

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

Eitz, M.; Richter, R.; Boubekeur, T.; Hildebrand, K.; and Alexa, M. 2012. Sketch-based shape retrieval. *ACM Trans. Graph. (Proc. SIGGRAPH)* 31(4):31:1–31:10.

Furuya, T., and Ohbuchi, R. 2013. Ranking on cross-domain manifold for sketch-based 3D model retrieval. In *Cyberworlds (CW), 2013 International Conference on*, 274–281. IEEE.

Hoffer, E., and Ailon, N. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 84–92. Springer.

Hu, J.; Lu, J.; and Tan, Y.-P. 2014. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1875–1882.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, B.; Godil, A.; Aono, M.; Bai, X.; Furuya, T.; Li, L.; López-Sastre, R. J.; Johan, H.; Ohbuchi, R.; Redondo-Cabrera, C.; et al. 2012. Shrec'12 track: Generic 3D shape retrieval. In *3DOR*, 119–126.

Li, B.; Lu, Y.; Godil, A.; Schreck, T.; Bustos, B.; Ferreira, A.; Furuya, T.; Fonseca, M. J.; Johan, H.; Matsuda, T.; et al. 2014. A comparison of methods for sketch-based 3D shape retrieval. *Computer Vision and Image Understanding* 119:57–80.

Li, B.; Lu, Y.; Li, C.; Godil, A.; Schreck, T.; Aono, M.; Burtscher, M.; Chen, Q.; Chowdhury, N. K.; Fang, B.; et al. 2015. A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding* 131:1–27.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Saavedra, J. M.; Bustos, B.; Schreck, T.; Yoon, S. M.; and Scherer, M. 2012. Sketch-based 3D model retrieval using keyshapes for global and local representation. In *3DOR*, 47–50. Citeseer.

Shih, J.-L.; Lee, C.-H.; and Wang, J. T. 2007. A new 3D model retrieval approach based on the elevation descriptor. *Pattern Recognition* 40(1):283–295.

Shilane, P.; Min, P.; Kazhdan, M.; and Funkhouser, T. 2004. The princeton shape benchmark. In *Shape modeling applications, 2004. Proceedings*, 167–178. IEEE.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2015. Deep metric learning via lifted structured feature embedding. *arXiv preprint arXiv:1511.06452*.

Sousa, P., and Fonseca, M. J. 2010. Sketch-based retrieval of drawings using spatial proximity. *Journal of Visual Languages & Computing* 21(2):69–80.

Sun, J.; Ovsjanikov, M.; and Guibas, L. 2009. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, 1383–1392. Wiley Online Library.

Tatsuma, A.; Koyanagi, H.; and Aono, M. 2012. A large-scale shape benchmark for 3D object retrieval: Toyohashi shape benchmark. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 1–10. IEEE.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 3360–3367. IEEE.

Wang, F.; Kang, L.; and Li, Y. 2015. Sketch-based 3D shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1875–1883.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10(Feb):207–244.

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