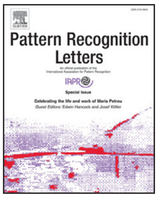
Pattern Recognition Letters 000 (2017) 1–8

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



Training convolutional neural network from multi-domain contour images for 3D shape retrieval

Zongxiao Zhu[a,b,*], Cong Rao[a], Song Bai[c], Longin Jan Latecki[a]

[a] *Temple University, 1805 North Broad Street, Philadelphia, PA 19122, USA*
[b] *South-Central University for Nationalities, 182 Minyuan Road, Wuhan 430074, China*
[c] *Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan 430074, China*

ARTICLE INFO

Article history:
Available online xxx

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D Shape retrieval
Object representation
Convolutional neural network

ABSTRACT

Recent vision-brain physiological experiments [65] have demonstrated that contours, and in particular, contour junctions present in the 2D images are very informative for revealing the 3D structure of the object. Inspired by this observation, we take 2D sketches (or 2D views of 3D sketches) and edge maps of 2D views of 3D models as a unified domain to train the Convolutional Neural Network (CNN). The CNN features are then used for 3D object representation. We show that the CNN can successfully learn the object structure from different types of clues. The performance of the proposed method demonstrates that the semantic gap between the 2D/3D sketches and the 3D models can be bridged without any cross-domain similarity learning. Experiments show that our approach significantly outperforms the state-of-the-art 2D/3D sketch-based 3D retrieval methods.

1. Introduction

The 3D sensors and imaging devices are developing rapidly nowadays. As a result, more and more 3D models have been constructed and are accessible in our daily life. Thus, an efficient way to manage the 3D content is very significant. In this context, retrieving 3D models that are similar to a query 2D/3D sketch or 3D model in a large 3D database has become an important research area. It will lead to many potential applications such as Virtual Reality, 3D Printing/Manufacturing and Robotics. In recent years, several large-scale 3D shape benchmarks, e.g., ModelNet [70], Shape Retrieval Contest (SHREC) [41,51], ShapeNet core55 [55], have been published. Many dedicated algorithms have been developed and tested on three different 3D shape retrieval tasks:

1. 2D sketch-based 3D shape retrieval [35,37,38],
2. 3D sketch-based 3D shape retrieval [34],
3. and 3D model-based 3D shape retrieval [36,55].

In these three retrieval tasks, the first problem to be investigated is an appropriate representation for the 3D model. In these tasks, the 3D objects are often described via multi-view based or volumetric models. And those exploiting the power of Deep Learning techniques have shown outstanding performance in this domain. However, when it comes to 2D/3D sketch-based retrieval, the performance is still not good enough for applications in practice. To compare a hand-drawn 2D/3D sketch directly with a 3D model, most existing methods match 2D sketches or projections of 3D sketches with a set of rendered views of 3D models. As learning based representations often result in much better performance than hand-crafted descriptors, many researchers, including the main organizer [34] of SHREC 2D/3D sketch-based 3D shape retrieval, believe that, “machine learning, especially deep learning, should be utilized instead of selecting and fixing the features beforehand”. More recently, different Convolutional Neural Networks (CNN) [66,75] have been developed to learn cross-domain image representations for 2D sketches and 3D shapes. These methods improve the performance of 2D sketch-based 3D retrieval at the cost of expensive training.

In this paper, we look for an unified representation in all three retrieval tasks. We propose a feasible solution using contours/edges to represent the 2D views. Some of the generated views are demonstrated in Fig. 1, where we examine three different representations: (c, d) the edge maps of 2D views generated from a 3D model (edge maps from 3D models for short in this paper), (f) the 2D views of a line-connected 3D sketch (2D views of 3D sketches for short in this paper), and (g, h) 2D sketches. They look similar by intuition, and it is natural to use all of them as clues to learn contour features for representing a 3D model. As we

[*] Corresponding author at: South-Central University for Nationalities, 182 Minyuan Road, Wuhan 430074, China.
E-mail addresses: zongxiao@mail.scuec.edu.cn, zhuzongxiao@gmail.com (Z. Zhu).

http://dx.doi.org/10.1016/j.patrec.2017.08.028
0167-8655/© 2017 Elsevier B.V. All rights reserved.

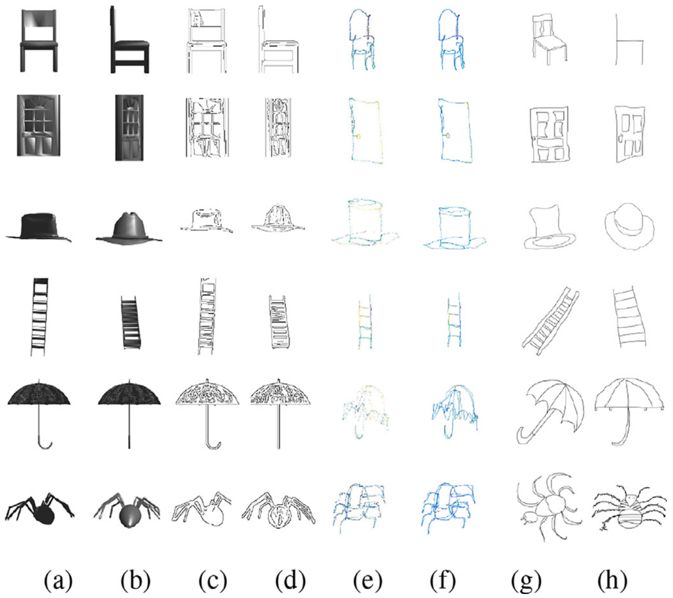


Fig. 1. Different representations for the views of a 3D object: (a) The rendered views of a 3D object at some viewpoint; (b) The rendered views of a 3D object at a different viewpoint; (c) The edge maps of the corresponding views in (a); (d) The edge maps of the corresponding views in (b); (e) The 3D sketches represented as point clouds; (f) The 2D views of corresponding 3D sketches in (e) (generated using [34]); (g) The 2D sketches drawed by human beings; (h) The 2D sketches drawed by different people. It is intuitive to take images in (c), (d), (f), (g) and (h) as the same domain, namely, contours of multiple views for representing a 3D object.

will demonstrate, a CNN is able to capture the visual primitives on the contours and junctions. This allows us to utilize deep learning to bridge the gap between 2D/3D sketches and 3D models, which help improve the performance of 2D/3D sketch-based 3D retrieval.

Besides, there is a solid physiological basis for representing 3D models with contours/edges of 2D views. While the surface features such as color and gradient orientations may help contour detection and surface delineation [47], they do not directly contribute to the high-level representation for visual recognition. In fact, visual recognition and categorization are possible once important contours are present and their relations are determined [10]. Meanwhile, contours and junctions in the 2D images are capable of describing the 3D structural information, since they could imply the arrangements and relations between surfaces in the 3D space [9,24]. For instance, L-junctions indicate points of termination of surfaces, T-junctions signify occlusion in depth, and Y-junctions and arrow-junctions indicate corners facing toward or away from the viewer [52].

The recent studies in [5,16] provide strong evidence that contour junctions indeed play a significant role in human visual recognition and categorization, due to the invariance of junctions to changes in viewpoint. We have also found support for the role of contour junctions in our CNN model built upon the three different contour images: edge maps, 2D views of 3D sketches and 2D sketches. We use the reconstruction method proposed in [46] to find out what information is preserved in certain layer of a CNN. In Fig. 2, we can see that contour junctions are activated clearly in the reconstruction output of the layer 4, which means it is learned by CNN that the junctions are more informative to represent the original image.

From the above analysis, we can see that:

1. Contours and junctions are informative enough for describing all kinds of views and eventually the 3D objects due to their physiological characteristics.

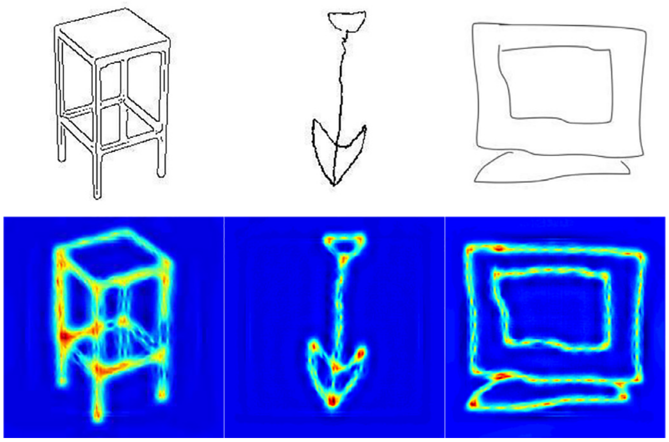


Fig. 2. An evidence that CNN could learn contour pieces and contour junctions for image representation. We can see that most neurons are activated in the locations of contours or junctions.

2. It is intuitive to use edge maps from 3D models, 2D views of 3D sketches and 2D sketches as clues to learn contour features, since they are visually similar thus can be regarded as images in the same domain.

In this paper, we first show that CNNs built upon edge maps from 3D models can work as well as those methods based on multi-view representations in the 3D shape retrieval task. Then, edge maps from 3D models, 2D views of 3D sketches and 2D views are used as clues to learn the CNN model, so as to improve the shape retrieval performance in three different tasks. The rest of this paper is organized as follows. Section 2 reviews related works; Section 3 describes the proposed framework and implementation details; Section 4 discusses the experimental results, and Section 5 concludes the proposed method.

2. Related works

In this section, we introduce related research of 3D shape retrieval over the last three years. Several 3D shape benchmarks play important roles in bringing and comparing new retrieval methods. We summarize these benchmarks and methods in Table 1, where we can find that the related methods are mainly divided into three categories:

1. Non-learning based approach, e.g., Handcraft local/global feature with direct feature matching, Bag-of-Words framework, and Shape Context matching.
2. Distance learning based approach, e.g., Manifold learning and Cross-domain manifold learning.
3. Deep learning based approach, e.g., CNN-based representation, Cross-domain CNN and Indiscriminate-domain CNN.

The research in all three 3D shape retrieval tasks started with non-learning based approaches, such as handcrafted local/global features and direct feature matching, shape context matching or the Bag-of-Words (BoW) framework. Since then, more and more researchers turned to use learning based representations [5,7] including manifold learning and deep learning to develop higher level knowledge-based 3D retrieval algorithms. As shown in rows “MR” and row “2D SC” of Table 1, for tasks such as 3D model-based retrieval and 2D sketch classification, handcrafted features and non-learning based approaches have been gradually replaced by deep learning based approaches in the recent three years.

When it comes to the cross-domain retrieval tasks, it is commonly believed that there exists a semantic gap between the 2D/3D sketches and the 3D models in the database. To bridge this

Table 1
A review of the related works on the three 3D shape retrieval tasks. In the table, "MR" stands for Model-based 3D shape retrieval, "2D SC" stands for 2D Sketch classification and "2D/3D SR" stands for 2D/3D sketch-based 3D shape retrieval.

Tasks	Dataset	Content	Main methods applied on this benchmark
MR	SHREC'14 LSGTB [36]	8987 models (171 classes)	Geometry-based Model [2,49], View-based Model [1,27], Hybrid Model [3,13,15], Direct Feature Matching [74], Bag-of-Words [21], Manifold Learning [48]
	ModelNet40 [70]	12,311 models (40 classes)	Multi-view based Representation [4,28,61], Volumetric Representation [25,58], VRN Ensemble [12], 3D-GAN [69]
	SHREC'16 [55]	51,300 models (55 classes)	Multi-view based Representation [4,61], Volumetric Representation [23]
	SHREC'17 [56]	51,162 models (55 classes)	CNN based on Point Set [23]; 3D Shape Descriptor [33]
2D SC	TU-Berlin [18]	20,000 shapes (250 classes)	HOG with SVM [18], Structured Ensemble Matching [40], Multi-kernel SVM [39], Fisher Vector Spatial pooling [57], Sketch-A-Net [72], Deepsketch [59], GoogLenet [62], Triplet Network [54], Indiscriminate-domain CNN [67]
2D SR	SHREC'13 STB [35]	7200 shapes and 1258 models (90 classes)	DSIFT [21], LD-SIFT [17], HOG [11], Local Features [20], Global Features [8], Direct Feature Matching, Bag-of-Words [19], Shape Context [6]
	SHREC'14 LSSTB [37]	13,680 shapes and 8987 models (171 classes)	Cross-Domain Manifold Ranking [22,73], PCDNN [75], Siamese Network [66]
3D SR	SHREC'16 3DSTB [34]	300 shapes (30 classes) and 1258 models (90 classes)	Cross-Domain Manifold Ranking [34], Siamese Network [66], Multi-view Representation [71]

gap, cross-domain manifold learning schemes and cross-domain neural networks [75] are used to learn domain invariant distance metrics. Recently, cross-domain CNNs such as Siamese Network [66] and Triplet Network [54] offer several solutions to this task. They improve the performance of 2D sketch-based 3D shape retrieval with a more expensive cost of the training process. The related works are summarized in rows "the 2D/3D SR" of Table 1.

Unlike cross-domain CNNs which focus on using different CNN models to process data from different domains, indiscriminate-domain CNNs try to use a single model to handle different types of data. Wang et al. [67] trained a model which can accommodate both images and sketches for sketch-based image retrieval tasks. This model is able to classify both the images and sketches. In this paper, we demonstrate that 2D sketches, 2D views of 3D sketches and edge maps from 3D models can be also used alongside each other. The learned indiscriminate-domain CNN model is able to represent 2D/3D sketches and 3D models with higher retrieval accuracy than the current cross-domain CNNs.

For non-rigid shape retrieval tasks [32,43], point matching [44,76] and point set registration [45] are typically used. So we do not focus on this domain and compare to related methods in this paper. For rigid shape retrieval tasks, previously Su and coworkers [34,61,67] have proposed similar works for each individual task. But in [61], the edge features are used for sketch-based retrieval only, while they are utilized for 3D model-based retrieval in our method; In [67], the same edge views are used for training and testing, while we use different edge views for training and testing in all three tasks; In [34], the CNN is trained upon images of a single domain, namely the projections of 3D sketches, and label matching is applied to achieve the best results. In comparison, our paper is the first one that proposes a single but efficient framework for the three retrieval tasks with outstanding performance on four different benchmarks. The advantage of the proposed framework lies in the clean training and testing strategies and the high versatility. Our framework is much simpler and faster than other proposed frameworks. It provides a practical way of training CNN for cross-domain 3D shape retrieval.

3. Approach

For all the three retrieval tasks described below, we use a multi-view based model to represent the 3D objects and each view is described by the contour features extracted via the introduced CNN architecture. Contours are able to describe several visual properties explicitly, which are not accessible in fully textured color photographs, such as contour orientation, length, curvature, and junctions [65]. Such kind of visual information allows human beings to effectively perceive semantic meanings of scenes [30,31]. An overview of the proposed method can be found in Fig. 3

3.1. Generating views for 2D sketch-based 3D shape retrieval

As 2D sketching seems to be the only mechanism for most people to depict a visual object, many algorithms have been developed in this domain (see Section 2) for 2D sketch-based 3D shape retrieval. The state-of-the-art results come from cross-domain neural networks [75] and the Siamese Network [66]. In contrast to using different CNNs for modeling and 2D views of 3D models separately, we consider these images as a single domain. We use a set of edge maps of 2D views (see Section 3.3) to represent the corresponding 3D model. From Fig. 1 we can see that 2D sketches and edge maps from 3D models look similar, as both of them are collections of main contours and fine details. For 2D sketch-based 3D shape retrieval, we use edge maps from 3D models and 2D sketches for training and testing.

3.2. Generating views for 3D sketch-based 3D shape retrieval

For 3D sketch-based 3D shape retrieval, Li et al. [34] uses handcrafted features, such as localized statistical features or histograms of oriented distances, to compute the distance between 3D sketches or models. Meanwhile, Ye [71] shows better performances via automatically learning the CNN features. We take both 2D views of 3D sketches and edge maps from 3D models for representing 3D sketches or models, and train a CNN with these two types of images without discrimination. As a 3D sketch is stored in the form of a point cloud, we first connect consecutive points with lines to construct a line-connected 3D sketch. Then we project this 3D sketch to generate a set of 2D views. Still, a set of edge maps of 2D views (see Section 3.3) is extracted from each 3D model. From the examples in Fig. 1 (c), (d) and (f), we can see that these images share very similar textures. For 3D sketch-based 3D shape retrieval, we use edge maps from 3D models and 2D views of 3D sketches for training and testing.

3.3. Generating views for 3D model-based 3D shape retrieval

For 3D model-based 3D shape retrieval, most state-of-the-art methods rely on CNNs to attack the problem. Two kinds of CNNs

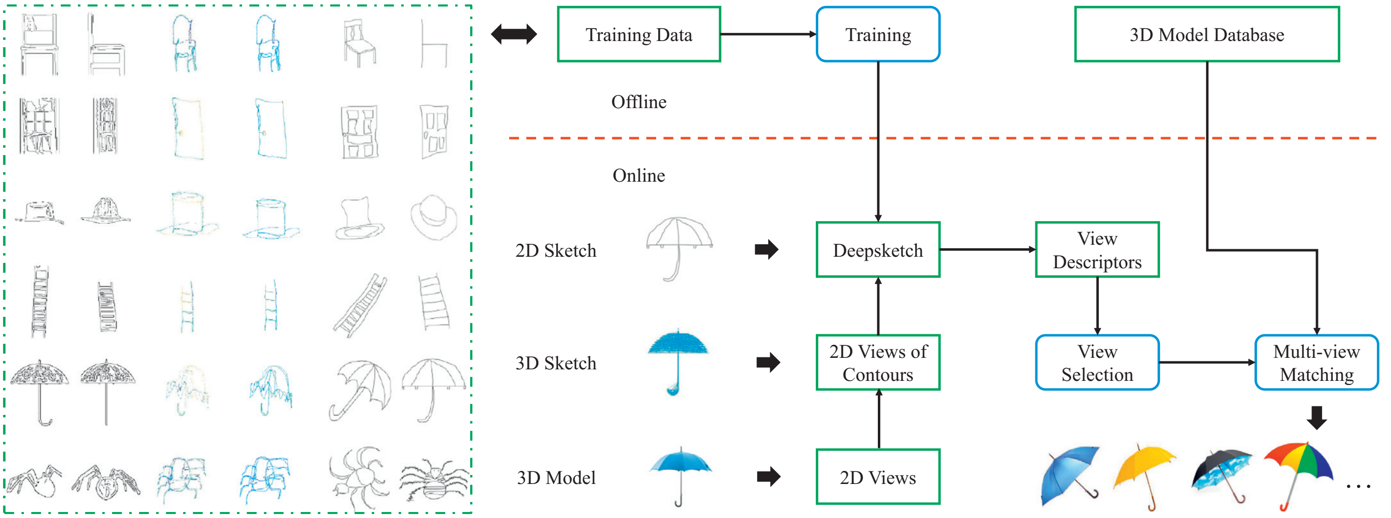


Fig. 3. An overview of the proposed 3D shape retrieval framework.

are mainly considered in this scenario: CNN based upon volumetric representations and CNN based upon multi-view representations. As volumetric representations can not be directly used in 2D/3D sketch based 3D shape retrieval tasks, we leave it as an open question in the future. Multi-view based CNN methods have shown outstanding performance in the 3D shape retrieval tasks. For example, MVCNN [61] aggregates multiple views through a view-pooling layer followed by fully-connected layers. The view pooling layer can combine all streams of views, and the final representation is more informative for retrieval than using the full collection of view-based descriptors of the 3D object. Besides, GIFT [4] samples 64 views uniformly on a view sphere and measures the object similarity by comparing two sets of CNN features of views. These two methods obtain the best results on the normal dataset and the perturbed dataset in SHREC'16 [55] respectively. Unlike these state-of-the-art methods using sophisticated architectures, our 3D model-based retrieval framework is more concise and efficient. Following the first camera setup in MVCNN [61], we assume that the input 3D shapes are upright oriented along a consistent axis, and create 12 rendered views around the axis every 30 degrees. Unlike existing methods using RGB images or depth images as input to the CNN, we take the Canny edge maps of 2D views to train the CNN model. For 3D model-based 3D shape retrieval, we only use edge maps from 3D models for training and testing.

3.4. View representation via deep contour features

For object retrieval, many researchers have found that the retrieval results can be improved if the class distribution predicted by CNN is directly used as the object descriptor. For example, Su et al. [61] use the classification probabilities as the 3D model signature, and obtain the best retrieval results in the normal test of SHREC'16 track [55]. In our experiment, we also find that the retrieval results can be improved significantly when a soft-max layer is appended to the GoogLeNet [62] or Deepsketch [59] model, where the output of the soft-max layer is used as the descriptor of the input 2D view. The comparison of the retrieval results with or without the soft-max layer in the GoogLeNet model can be found in Fig. 4.

3.5. View selection based multi-view matching

All three 3D shape retrieval tasks involve a multi-view matching stage to establish the correspondence between two sets of 2D views of two given 3D objects. The traditional way is to use certain kind of Hausdorff distance. For example, MVCNN [61] averages pairwise distances between two sets of 2D views. GIFT [4] adopts a robust version of Hausdorff distance [26]. In our method, we first determine the dominant class of a set of views by majority voting, and then remove the views belonging to the minority classes. This pre-processing can improve the time efficiency in the multi-view matching process as well as the retrieval accuracy. A pair of well matched 2D views will indicate a good match of two objects, and here we apply the minimum Hausdorff distance to capture the overall dissimilarity of two objects, as defined in the following Eq. (1).

$$D(x_q, x_p) = \min_{q_i \in \nu(x_q)} \min_{p_j \in \nu(x_p)} d(q_i, p_j), \tag{1}$$

where $d(q_i, p_j)$ measures the cosine distance between 2D view/sketch CNN descriptor q_i and p_j, $\nu(x_q)$ is the set of view de-

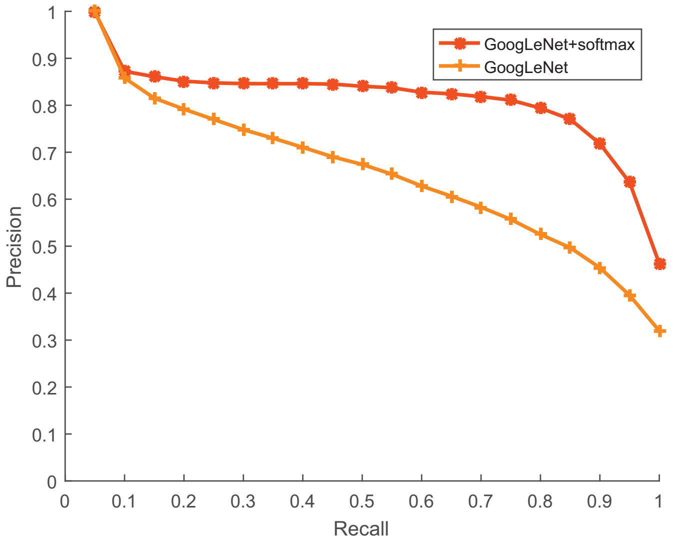
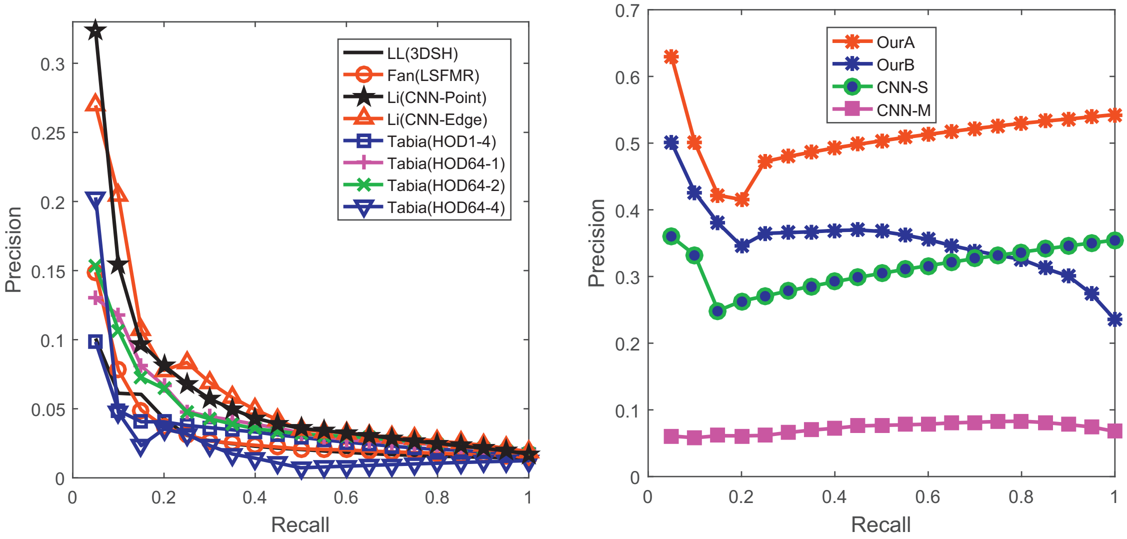


Fig. 4. The comparison of retrieval results on the ModelNet40 dataset with (red) or without (orange) a soft-max layer in the GoogLeNet model. The difference of mAP on the Modelnet40 test dataset is around 8%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

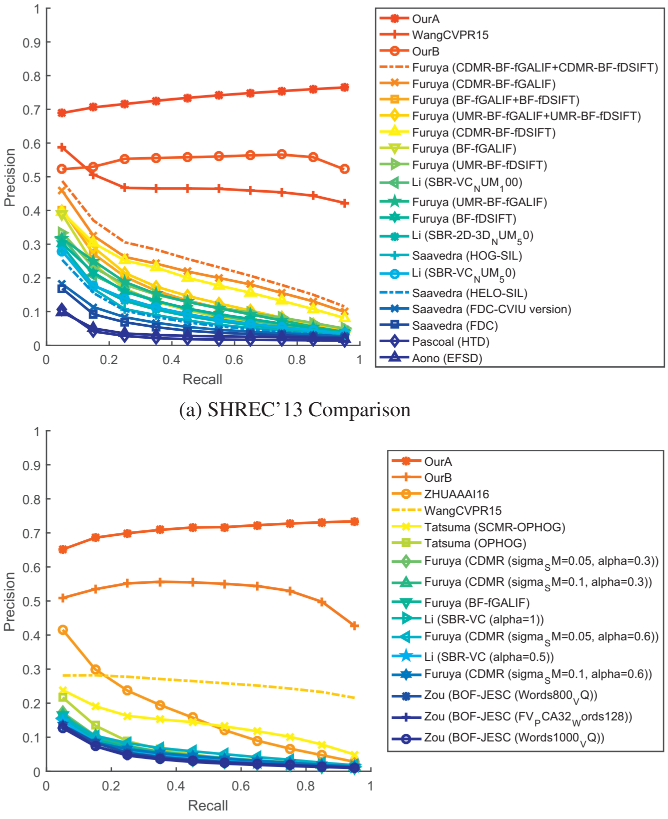
Please cite this article as: Z. Zhu et al., Training convolutional neural network from multi-domain contour images for 3D shape retrieval, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.08.028



(a) Non-learning approaches

(b) Learning-based approaches

Fig. 5. Precision-recall plots for 3D sketch-based 3D shape retrieval.



(a) SHREC’13 Comparison

(b) SHREC’14 Comparison

Fig. 6. Precision-recall plots for 2D sketch-based 3D shape retrieval.

scriptors from the query sketch/model x_q, and $v(x_p)$ is the set of the view descriptors from some 3D object x_p in the database. Note that if x_q represents a 2D sketch, then the set $v(x_q)$ contains only one view descriptor.

Table 2
The detailed CNN architecture for Deepsketch.

Index	Type	Filter	Filter	Stride	Pad
1	Conv	7 × 7	64	2	0
2	ReLU	–	–	–	–
3	Maxpool	3 × 3	–	2	0
4	Conv	5 × 5	128	2	2
5	ReLU	–	–	–	–
6	Maxpool	3 × 3	–	2	0
7	Conv	7 × 7	256	1	1
8	ReLU	–	–	–	–
9	Conv	7 × 7	256	1	1
10	ReLU	–	–	–	–
11	Maxpool	3 × 3	–	2	0
12	Conv	5 × 5	4096	1	1
13	ReLU	–	–	–	–
14	Dropout	–	–	–	–
15	Conv	1 × 1	250	1	1

3.6. The choice of CNN architecture

Our main criterion for selecting the CNN architecture is the performance for sketch classification. Three models, including Sketch-A-Net [72], GoogLeNet [62] and Deepsketch [59], are evaluated on the TU-Berlin benchmark [18]. In this paper, we employ the Deepsketch model since it is much simpler than GoogLeNet and Sketch-A-Net, and the objective function also converges much faster than the other two. The detailed CNN architecture of Deepsketch is described in Table 2.

4. Experiments

The goal of our research is to show that using contour features to represent views can provide sufficient information to infer the characteristics of the whole 3D objects, and succeed in achieving the state-of-the-art performance in different 3D shape retrieval tasks. Besides, these 2D view images from different sources can be used along side each other in training and testing. We first evaluate our method on the 2D sketch-based retrieval task, then extend it to the 3D sketch-based retrieval task and finally the 3D model-based retrieval task (Table 3).

Please cite this article as: Z. Zhu et al., Training convolutional neural network from multi-domain contour images for 3D shape retrieval, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.08.028

Table 3
3D model-based retrieval performance on the normal dataset of SHREC2016. Best results are marked in bold, the second best are in italic.

Method	micro					macro				
	P@N	R@N	F1@N	mAP	NDCG@N	P@N	R@N	F1@N	mAP	NDCG@N
Su [61]	**0.770**	*0.770*	**0.764**	**0.873**	*0.899*	**0.571**	0.625	**0.575**	**0.817**	**0.880**
Bai [4]	0.706	0.695	0.689	0.825	0.896	0.444	0.531	0.454	0.740	0.850
Li [55]	0.508	**0.868**	0.582	0.829	**0.904**	0.147	**0.813**	0.201	0.711	0.846
Wang [68]	0.718	0.350	0.391	0.823	0.886	0.313	0.536	0.286	0.661	0.820
Tastuma [64]	0.427	0.689	0.472	0.728	0.875	0.154	*0.730*	0.203	0.596	0.806
Ours	*0.755*	0.731	*0.726*	*0.865*	*0.899*	*0.492*	0.596	*0.503*	*0.790*	*0.874*

Table 4
3D model-based retrieval performance on perturbed dataset of SHREC2017. Best results are marked in bold, the second best are in italic.

Method	micro					macro				
	P@N	R@N	F1@N	mAP	NDCG@N	P@N	R@N	F1@N	mAP	NDCG@N
Furuya [23]	**0.814**	0.683	*0.706*	*0.656*	*0.754*	**0.607**	0.539	**0.503**	**0.476**	**0.560**
Tastsuma [63]	*0.705*	**0.769**	**0.719**	**0.696**	**0.783**	0.424	*0.563*	0.434	0.418	0.479
Zhou [4]	0.660	0.650	0.643	0.567	0.701	0.443	0.508	0.437	0.406	0.513
Kanezaki [29]	0.655	0.652	0.636	0.606	0.702	0.372	0.393	0.333	0.327	0.407
Deng [42,60]	0.412	*0.706*	0.472	0.524	0.642	0.120	**0.659**	0.164	0.329	0.395
Li [33]	0.496	0.234	0.258	0.172	0.303	0.199	0.373	0.179	0.215	0.336
Mk [50]	0.690	0.012	0.020	0.009	0.043	*0.546*	0.052	0.052	0.047	0.109
Ours	0.681	0.681	0.673	0.638	0.733	0.453	0.503	*0.448*	*0.435*	*0.519*

Table 5
Comparison of 3D sketch-based retrieval performance on the SHREC'16 benchmark (all results except ours are taken from [34]).

SHREC'16 Comparison						
Method	NN	FT	ST	E	DCG	mAP
3DSH	0.029	0.021	0.038	0.021	0.254	0.029
LSFMR	0.033	0.020	0.033	0.018	0.248	0.032
C-Point	0.124	0.044	0.075	0.046	0.294	0.060
C-Edge	0.114	0.056	0.084	0.051	0.302	0.063
Tabia	0.067	0.031	0.057	0.032	0.272	0.044
CNN-S	0.222	0.251	0.320	0.186	0.471	0.314
CNN-M	0.000	0.031	0.108	0.048	0.293	0.072
OursA	**0.413**	**0.471**	**0.527**	**0.302**	**0.617**	**0.508**
OursB	*0.286*	*0.293*	*0.393*	*0.214*	*0.514*	*0.352*

Table 6
Comparison of 2D sketch-based retrieval performance on the SHREC'13 and SHREC'14 datasets.

Method	NN	FT	ST	E	DCG	mAP
SHREC'13 Comparison						
Wang [66]	0.405	0.403	0.548	0.287	0.607	0.469
Furuya [22]	0.279	0.203	0.296	0.166	0.458	0.250
Li [37]	0.164	0.097	0.149	0.085	0.348	0.116
Saavedra [53]	0.110	0.069	0.107	0.061	0.307	0.086
OursA	**0.678**	**0.705**	**0.774**	**0.373**	**0.800**	**0.736**
OursB	*0.469*	*0.521*	*0.653*	*0.301*	*0.681*	*0.548*
SHREC'14 Comparison						
Wang [66]	0.239	0.212	0.316	0.140	0.486	0.228
Tatsuma [63]	0.160	0.115	0.170	0.079	0.376	0.131
Furuya [22]	0.109	0.057	0.089	0.041	0.329	0.055
Li [37]	0.095	0.050	0.081	0.037	0.319	0.050
OursA	**0.642**	**0.681**	**0.736**	**0.351**	**0.794**	**0.712**
OursB	*0.498*	*0.494*	*0.625*	*0.284*	*0.695*	*0.520*

4.1. 2D Sketch-based 3D shape retrieval

We use the SHREC'13 [35] and SHREC'14 [37] benchmarks to test our method for 2D sketch based 3D shape retrieval. Please see Table 1 and [35,37] for more details of these two benchmarks. For the each 3D model, we use 12 views for training. The views are captured at fixed viewpoints with elevation=90, azimuth ∈ {-150, -120, -90, -60, -30, 0, 30, 60, 90, 120, 150, 180}, and the retrieval result is denoted by 'OursA'. The other completely different 12 views are used for testing, and they are captured at fixed viewpoints with azimuth=45, elevation ∈ {-75, -60, -45, -30, -15, 0, 15, 30, 45, 60, 75, 90}. The corresponding retrieval result is denoted 'OursB'. These views are then mixed with 2D sketches to train or test the CNN model. From Fig. 6 and Table 6, it is shown that our contour based representation achieves better performance than all the other approaches on the SHREC'13 and SHREC'14 dataset. Meanwhile, in 2D sketch-based retrieval, the sketches for training and testing are drawn by human. They are more consistent to human perception thus relatively easier to identify. However, in 3D sketches and models, we generate the 2D projections by uniformly sampling the viewpoints on a sphere. Some of the projections may be ambiguous and hard to recognize even for humans. Besides, these 3D object models are not necessarily well aligned, which may also affect the performance of the retrieval results. We will see how this issue affects our method in the next two sections.

4.2. 3D Sketch-based 3D shape retrieval

We evaluate our method for 3D sketch-based 3D shape retrieval on the SHREC'16 benchmark [34] against all participants in this track. Please see Table 1 and [34] for the details of this benchmark. Following the practice in previous methods, all the target 3D models are the same in both training and testing. In our experiment, we use 12 2D views for learning the CNN model, and they are captured at viewpoints with elevation=90, azimuth ∈ {−150, −120, −90, −60, −30, 0, 30, 60, 90, 120, 150, 180}, and the corresponding retrieval result is denoted as 'OursA'. The other completely 12 different views are used for testing, which are captured at viewpoints with azimuth=45, elevation ∈ {−75, −60, −45, −30, −15, 0, 15, 30, 45, 60, 75, 90}, and the retrieval result is denoted as "OursB". The views in the two groups are very different, and if the contour features such as the junction types and angles are invariant to changes in viewpoint, we would expect reasonable retrieval results in both "OursA" and "OursB". From Fig. 5 and Table 5, it is observed that our contour features of views achieve better performances than all other non-learning and learning based approaches. The experiment also demonstrates that the contour features in the

2D projections are capable of describing the structure of a 3D object.

4.3. 3D Model-based 3D shape retrieval

We validate our method for 3D model-based 3D shape retrieval on the ShapeNet Core55 benchmark [14], against all five participants in the SHREC'16 Track [55] and SHREC'17 Track. Please see Table 1 and [14,55] for more details of these benchmarks. The 3D models in the ShapeNet benchmark are converted to the "obj" format with only geometric information, and the model dimensions are normalized within a cube of unit length. In addition, since the ShapeNet benchmark provides consistent upright and front orientation annotations, all models are consistently aligned. This dataset is called the 'normal" dataset. There is also a 'perturbed' version of the dataset, where each model has been randomly rotated. The Precision, Recall, F-score and NDGG are calculated to compare our method with the other methods. These metrics are referred to as P@N, R@N, F1@N and NDGG@N in the table, where N is the total retrieval list length chosen by the method. The organizer also provides two evaluation metrics to combine the retrieval results of different categories: (1) the macro-averaged version is used to give an unweighted average over the entire dataset, and (2) the micro-averaged version assigns equal weights to each query and retrieved shapes. In 3D object retrieval, the F measure, mAP and NDCG indicate the overall performance thus we compare mainly based on these metrics. On the normal dataset of SHREC2016, it is observed that our retrieval mAP is only 0.8% behind the best result of Su [61], while our CNN architecture is much simpler and more efficient than theirs. In Table 4, our mAP is 5.8% lower than the best result of Tastsuma [63] and 1.8% behind the second result from Furuya [23]. Tastsuma et al. translate the center of the 3D model to the origin and then normalize the size and the rotation of the 3D model. They use an improved version of Neighbor Set Similarity [63] for ranking the retrieved shapes. Furuya et al. convert each 3D model into a 3D oriented point set by sampling the surfaces of the 3D model, which is more expressive than the edge maps from 3D models. All these techniques could be used to improve our retrieval results.

5. Conclusion

Among various methods for 3D shape retrieval, the deep learning based representations tend to gradually replace traditional learning or non-learning based approaches. In this paper, we address the problem by developing a uniform representation for different 3D shape retrieval tasks, namely 2D/3D sketch-based and 3D model-based 3D shape retrieval. We propose to represent 3D objects with multi-view based description, where each view is described by the contour features extracted by the learned CNN. We demonstrate that the proposed contour-based representation is successful in 3D model-based 3D shape retrieval task and achieves the state-of-the-art performance. As it is an uniform representation for edge maps from 3D models, 2D views of 3D sketches and 2D sketches, we train the CNN model with these images as the same domain. It is showed in our experiments that the performance of 2D/3D sketch-based 3D shape retrieval tasks are significantly improved compared to existing methods.

Acknowledgments

This work was supported by the 2016 China Scholarship Council (CSC) scholarship (No. 201507780014). It was also supported by Natural Science Foundation of Hubei Province of China (No. 2015CFB597). This work was supported in part by NSF grant IIS-1302164.

References

[1] P.F. Alcantarilla, A. Bartoli, A.J. Davison, Kaze features, in: European Conference on Computer Vision, 2012, pp. 214–227.
[2] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, T. Seidl, 3d shape histograms for similarity search and classification in spatial databases, in: International Symposium on Spatial Databases, 1999, pp. 207–226.
[3] M. Aono, H. Koyanagi, A. Tatsuma, 3d shape retrieval focused on holes and surface roughness, in: Signal and Information Processing Association Annual Summit and Conference, 2013, pp. 1–8.
[4] S. Bai, X. Bai, Z. Zhou, Z. Zhang, L.J. Latecki, Gift: a real-time and scalable 3d shape search engine, Comput. Vision Pattern Recognit. (2016) 5023–5032.
[5] X. Bai, C. Rao, X. Wang, Shape vocabulary: a robust and efficient shape representation for shape matching, Trans. Image Process. 23 (9) (2014) 3935–3949.
[6] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[7] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[8] S. Biasotti, I. Pratikakis, U. Castellani, T. Schreck, A. Godil, R. Veltkamp, Sketch-based 3d model retrieval by viewpoint entropy-based adaptive view clustering (2013).
[9] I. Biederman, Recognition-by-components: a theory of human image understanding., Psychol. Rev. 94 (2) (1987) 115.
[10] I. Biederman, G. Ju, Surface versus edge-based determinants of visual recognition, Cognit. Psychol. 20 (1) (1988) 38–64.
[11] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: International Conference on Image and Video Retrieval, 2007, pp. 401–408.
[12] A. Brock, T. Lim, J. Ritchie, N. Weston, Generative and discriminative voxel modeling with convolutional neural networks, arXiv:1608.04236 (2016).
[13] B. Bustos, D. Keim, D. Saupe, T. Schreck, D. Vranic, Automatic selection and combination of descriptors for effective 3d similarity search, in: Symposium on Multimedia Software Engineering, 2004, pp. 514–521.
[14] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: an information-rich 3d model repository, arXiv:1512.03012 (2015).
[15] J.-y. Chen, B. He, X.-h. Wang, Hpal information entropy based combination methods for 3d model retrieval, J. Syst. Simul. 1779 (1777).
[16] H. Choo, D.B. Walther, Contour junctions underlie neural representations of scene categories in high-level human visual cortex: contour junctions underlie neural representations of scenes, Neuroimage 135 (2016) 32–44.
[17] T. Darom, Y. Keller, Scale-invariant features for 3-d mesh models, Trans. Image Process. 21 (5) (2012) 2758–2769.
[18] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects? Transactions on Graphics 31 (4) (2012). 44–1.
[19] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa, Sketch-based image retrieval: benchmark and bag-of-features descriptors, Visual. Comput. Graphics 17 (11) (2011) 1624–1636.
[20] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, M. Alexa, Sketch-based shape retrieval., Trans. Graphics 31 (4) (2012). 31–1.
[21] T. Furuya, R. Ohbuchi, Dense sampling and fast encoding for 3d model retrieval using bag-of-visual features, in: International Conference on Image and Video Retrieval, 2009, p. 26.
[22] T. Furuya, R. Ohbuchi, Ranking on cross-domain manifold for sketch-based 3d model retrieval, in: International Conference on Cyberworlds, 2013, pp. 274–281.
[23] T. Furuya, R. Ohbuchi, Deep aggregation of local 3d geometric features for 3d model retrieval., in: British Machine Vision Conference, 2016.
[24] A. Guzmán, Decomposition of a visual scene into three-dimensional bodies, in: Joint Computer Conference, Part I, 1968, pp. 291–304.
[25] V. Hegde, R. Zadeh, Fusionnet: 3d object classification using multiple data representations, arXiv preprint: 1607.05695(2016).
[26] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the hausdorff distance, Pattern Analysis and Machine Intelligence 15 (9) (1993) 850–863.
[27] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.
[28] E. Johns, S. Leutenegger, A.J. Davison, Pairwise decomposition of image sequences for active multi-view recognition, arXiv preprint: 1605.08359(2016).
[29] A. Kanezaki, Y. Matsushita, Y. Nishida, Rotationnet: Joint learning of object classification and viewpoint estimation using unaligned 3d object dataset, arXiv:1603.06208 (2016).
[30] J. Kim, I. Biederman, Where do objects become scenes? J. Vision 9 (8) (2009). 779–779.
[31] J.G. Kim, I. Biederman, C.-H. Juan, The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: a transcranial magnetic stimulation study, J. Neurosci. 31 (22) (2011) 8320–8324.
[32] Z. Lahner, E. Rodola, F.R. Schmidt, M.M. Bronstein, D. Cremers, Efficient globally optimal 2d-to-3d deformable shape matching, in: Computer Vision and Pattern Recognition, 2016, pp. 2185–2193.
[33] B. Li, H. Johan, 3D model retrieval using hybrid features and class information, Multimed. Tools Appl. 62 (3) (2013) 821–846.
[34] B. Li, Y. Lu, F. Duan, S. Dong, Y. Fan, L. Qian, H. Laga, H. Li, Y. Li, P. Liu, et al., Shrec'16 track: 3d sketch-based 3d shape retrieval (2016).

Please cite this article as: Z. Zhu et al., Training convolutional neural network from multi-domain contour images for 3D shape retrieval, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.08.028

[35] B. Li, Y. Lu, A. Godil, T. Schreck, M. Aono, H. Johan, J.M. Saavedra, S. Tashiro, SHREC'13 Track: large scale sketch-based 3D shape retrieval, 2013.

[36] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, Q. Chen, N.K. Chowdhury, B. Fang, et al., A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries, Comput. Vision Image Understanding 131 (2015) 1–27.

[37] 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, pp. 121–130.

[38] B. Li, T. Schreck, A. Godil, M. Alexa, T. Boubekeur, B. Bustos, J. Chen, M. Eitz, T. Furuya, K. Hildebrand, et al., Shrec'12 track: Sketch-based 3d shape retrieval., in: 3DOR, 2012, pp. 109–118.

[39] Y. Li, T.M. Hospedales, Y.-Z. Song, S. Gong, Free-hand sketch recognition by multi-kernel feature learning, Comput. Vision Image Understanding 137 (2015) 1–11.

[40] Y. Li, Y.-Z. Song, S. Gong, Sketch recognition by ensemble matching of structured features., in: British Machine Vision Conference, 2013.

[41] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H. Van Nguyen, R. Ohbuchi, et al., A comparison of methods for non-rigid 3d shape retrieval, Pattern Recognit. 46 (1) (2013) 449–461.

[42] Z. Lian, A. Godil, X. Sun, Visual similarity based 3d shape retrieval using bag-of-features, in: Shape Modeling International Conference, 2010, pp. 25–36.

[43] J. Ma, W. Qiu, J. Zhao, Y. Ma, A.L. Yuille, Z. Tu, Robust l2e estimation of transformation for non-rigid registration., Trans. Signal Process. 63 (5) (2015) 1115–1129.

[44] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, Trans. Image Process. 23 (4) (2014) 1706–1721.

[45] J. Ma, J. Zhao, A.L. Yuille, Non-rigid point set registration by preserving global and local structures, Trans. Image Process. 25 (1) (2016) 53–64.

[46] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Computer Vision and Pattern Recognition, 2015, pp. 5188–5196.

[47] D. Marr, Vision: A computational approach, 1982.

[48] R. Ohbuchi, T. Furuya, Distance metric learning and feature combination for shape-based 3d model retrieval, in: ACM Workshop on 3D Object Retrieval, 2010, pp. 63–68.

[49] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, Trans. Graphics 21 (4) (2002) 807–832.

[50] M. Patrick, Binvox Code-3d Mesh Voxelizer, Princeton University, 2011.

[51] I. Pratikakis, M. Savelonas, F. Arnaoutoglou, G. Ioannakis, A. Koutsoudis, T. Theoharis, M. Tran, V. Nguyen, V. Pham, H. Nguyen, et al., Shrec'16 track: Partial shape queries for 3d object retrieval.

[52] N. Rubin, The role of junctions in surface completion and contour matching, Perception 30 (3) (2001) 339–366.

[53] J.M. Saavedra, B. Bustos, T. Schreck, S.M. Yoon, M. Scherer, Sketch-based 3D Model Retrieval Using Keyshapes for Global and Local Representation., in: 3DOR, 2012, pp. 47–50.

[54] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, ACM Trans. Graphics 35 (4) (2016) 119.

[55] M. Savva, F. Yu, H. Su, M. Aono, B. Chen, D. Cohen-Or, W. Deng, H. Su, S. Bai, X. Bai, Shrec'16 track: Large-scale 3d shape retrieval from shapenet core55, (2016).

[56] M. Savva, F. Yu, H. Su, et al., Shrec'17 track: Large-scale 3d shape retrieval from shapenet core55 (2017).

[57] R.G. Schneider, T. Tuytelaars, Sketch classification and classification-driven analysis using fisher vectors, Trans. Graphics 33 (6) (2014) 174.

[58] N. Sedaghat, M. Zolfaghari, T. Brox, Orientation-boosted voxel nets for 3d object recognition, arXiv:1604.03351 (2016).

[59] O. Seddati, S. Dupont, S. Mahmoudi, Deepsketch: deep convolutional neural networks for sketch recognition and similarity search, in: International Workshop on Content-Based Multimedia Indexing, 2015, pp. 1–6.

[60] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).

[61] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: International Conference on Computer Vision, 2015, pp. 945–953.

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[63] A. Tatsuma, H. Koyanagi, M. Aono, A large-scale shape benchmark for 3d object retrieval: Toyohashi shape benchmark, in: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012, pp. 1–10.

[64] A. Tatsuma, A. Masaki, Food image recognition using covariance of convolutional layer feature maps, Trans. Inf. Syst. 99 (6) (2016) 1711–1715.

[65] D.B. Walther, D. Shen, Nonaccidental properties underlie human categorization of complex natural scenes, Psychol. Sci. (2014).

[66] F. Wang, L. Kang, Y. Li, Sketch-based 3d shape retrieval using convolutional neural networks, in: Computer Vision and Pattern Recognition, 2015, pp. 1875–1883.

[67] X. Wang, X. Duan, X. Bai, Deep sketch feature for cross-domain image retrieval, Neurocomputing (2016).

[68] Y. Wang, W. Deng, Self-restraint object recognition by model based cnn learning, in: International Conference on Image Processing, 2016, pp. 654–658.

[69] J. Wu, C. Zhang, T. Xue, W.T. Freeman, J.B. Tenenbaum, Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling, arXiv:1610.07584 (2016).

[70] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: a deep representation for volumetric shapes (2015).

[71] Y. Ye, Applying deep learning to scene sketch recognition and 3D sketch-based 3D model retrieval. Ph.D. thesis, Texas State University, 2016.

[72] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, T. Hospedales, Sketch-a-net that beats humans, arXiv:1501.07873 (2015).

[73] D. Zhou, J. Weston, A. Gretton, O. Bousquet, B. Schölkopf, Ranking on data manifolds, Neural Inf. Process. Syst. 16 (2004) 169–176.

[74] X. Zhou, K. Yu, T. Zhang, T.S. Huang, Image classification using super-vector coding of local image descriptors, in: European Conference on Computer Vision, 2010, pp. 141–154.

[75] F. Zhu, J. Xie, Y. Fang, Learning cross-domain neural networks for sketch-based 3d shape retrieval, in: AAAI Conference on Artificial Intelligence, 2016.

[76] Z. Zhu, G. Wang, J. Liu, Z. Chen, Fast and robust 2d-shape extraction using discrete-point sampling and centerline grouping in complex images, IEEE Trans. Image Process. 22 (12) (2013) 4762–4774.

Please cite this article as: Z. Zhu et al., Training convolutional neural network from multi-domain contour images for 3D shape retrieval, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.08.028