HIGH-LEVEL SEMANTIC FEATURE FOR 3D SHAPE BASED ON DEEP BELIEF NETWORKS

Zhenbao Liu[1], *Shaoguang Cheng*[1], *Shuhui Bu*[1,*], *Ke Li*[2]

[1]Northwestern Polytechnical University, Youyi west road #127, Xi'an 710072, China
[2]Information Engineering University, Longhai Road #66, Zhengzhou 450052, China
[*]bushuhui@nwpu.edu.cn

ABSTRACT

Deep learning has emerged as a powerful technique to extract high-level features from low-level information, which shows that hierarchical representation can be easily achieved. However, applying deep learning into 3D shape is still a challenge. In this paper, we propose a novel high-level feature learning method for 3D shape retrieval based on deep learning. In this framework, the low-level 3D shape descriptors are first encoded into visual bag-of-words, and then high-level shape features are generated via deep belief network, which facilitates a good semantic preserving ability for the tasks of shape classification and retrieval. Experiments on 3D shape recognition and retrieval demonstrate the superior performance of the proposed method in comparison to the state-of-the-art methods.

Index Terms— 3D Shape classification, 3D shape retrieval, Bag-of-words, Deep learning, Deep belief networks

1. INTRODUCTION

In recent decades, with the rapidly developed machine learning techniques, multimedia content analysis and modeling has achieved promising progresses. However, almost all methods still can not generate high-level perceptions similar to that generated by human brain, due to the shallow representation level. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG) have achieved remarkable advances, and consequently theory and architecture of brain can be understood through these devices. From latest research on structural and functional brain organization, we get that the cognition results from the dynamic interactions of distributed brain areas operating in large-scale networks, although it has long been assumed that cognitive functions are attributable to the isolated operations of single brain areas [1]. The brain also appears to process information through multiple stages of transformation and representation.

Inspired by the architectural depth of the brain, neural network has been extensively studied for decades to train deep multi-layer neural networks to imitate human brain, however, training deeper networks consistently yielded poor results. Recently, deep belief networks (DBNs) [2] has introduced to overcome the problem. The input data are trained layer by layer in a restricted Boltzmann machine (RBM) [3] by means of contrastive divergence (CD) [3]. In addition, deep learning extends the generalization ability against large intra-class variations, which is benefited by feature learning from low-level features without pre-defined models. Lee et al. [4] proposed a convolutional DBN which is a scalable generative model for learning hierarchical representations from unlabeled images. They showed that the method can achieve excellent performance on several visual recognition datasets. Ciresan et al. [5] proposed a multi-column deep neural networks for image classification. On the very competitive MNIST handwriting benchmark, the method achieves near-human performance.

Although deep learning has been successfully applied to natural language processing, computer vision, and other fields, the technique is difficult to be applied to 3D shape analysis due to the intrinsic differences between the structure of 3D graph data and traditional speech data or 2D images. 3D shapes are one of important multimedia formation, which has been extensively applied in the domains of multimedia, graphics, virtual reality, amusement, design, and manufacturing [6] due to the rich information preserving the surface, color, and texture of real objects. In recent years, with the rapidly developed 3D techniques, the increasing of 3D model data requires effective retrieval and classification techniques for their management and reusing.

There have been many solutions to 3D shape recognition, matching, classification, and retrieval in the last decade. These solutions are directly related with shape description [6]. Geodesic of 3D shape has an immanent property that its value is robust against articulated deformation, thereby it is a good choice for generating descriptor. Jain et al. [7] propose to

This work is supported partly by grants from NSFC (61202185, 61003137, 41201390), NWPU Basic Research Fund (JC201202, JC201220), Shaanxi Natural Science Fund (2012JQ8037), and Open Project Program of the State Key Lab of CAD&CG (A1306), Zhejiang University.

use first eigenvalues of the geodesic distance matrix of a 3D object to generate 3D shape descriptor which are isometry-invariant. Different from above method which uses shape's geometric information, view-based methods [8, 9] are based on an assumption that if two 3D models are geometrically similar they also look similar from all different viewing angles. These methods do not require utilization of geometric attributes or topological relationship, and hence are capable of handling 3D models with degeneration, holes, and missing patches. An important problem existent in view based 3D model retrieval is how to effectively organize and build the relationship of many views of 3D objects, for example, constructing hypergraphs of views [10].

The above discussed methods have the problem that low-level features are not sufficient to characterize the high-level semantics of 3D shapes. Inspired by the success of deep learning in computer vision, in this paper we seek for the possibility to apply high-level feature learning into 3D shape field. We propose a novel framework which builds a bridge between 3D shape and deep learning. The core idea is to extract bag-of-visual-features (BoVF) from multi-views of the shapes, and then generate high-level features for 3D shape retrieval via deep learning. Specifically, we first generate 200 depth images from a given shapes, and then scale invariant feature transform (SIFT) algorithm [11] is adopted to extract visual features from the depth images. In the second step, we generate a visual vocabulary from visual features of training shapes, and BoVF of each shape are computed for high-level feature learning. Next, deep belief network is utilized to learn a higher representative model, and high-level features of newly inputed shapes are computed by the model. Finally, shape classification and retrieval are performed based on the high-level features. The major advantages of our method are that the high-level features are learned from a set of shapes through supervised or un-supervised methods, therefore, better discriminability and generalization can be achieved, and moreover, the performance can be further improved with the increasing of training data number. Experiments are conducted in both 3D shape retrieval and recognition tasks. Results and comparison with state-of-the-art methods indicate that the proposed method can achieve better performance.

2. HIGH-LEVEL VISUAL FEATURE LEARNING FOR 3D SHAPES

The proposed high-level feature learning method for 3D shapes is carried out in the following two stages, and the flowchart of the proposed method is depicted in Fig. 1.

1. Low-level visual feature: We use visual features of depth images act as low-level features. Depth images are rendered at position of vertices of dodecahedron, in addition, for overcoming the rotation sensitive of depth images, the object is rotated in 10 angles, and consequently, 200 depth images are obtained. Next, we adopt SIFT features [11] to represent shape's visual features which are extracted from the depth images of shapes. The bag-of-words paradigm is used to generate order-irrelevant features for high-level feature learning.

2. High-level feature learning: Because deep learning is able to generate more discriminative and robust high-level features from shallow features via multi-layer network, we introduce it to learn high-level features more suitable for 3D shape classification and retrieval.

2.1. Low-level visual Feature Extraction

The low-level visual feature extraction for 3D models in our algorithm can be summarized as follows.

Pose normalization. For a given 3D mesh, translate the center of its mass to the origin and then scale the maximum polar distance of the points on its surface to one. Rotation normalization is not performed, but this will be compensated to some extent because of our way to extract views and local features.

Image rendering from multi-views. In this research, depth view images are rendered from 20 vertices of a regular dodecahedron whose mass center is also located in the origin. To make the feature robust for rotation, we rotate the dodecahedron 10 times and extract views at vertices of dodecahedron. The rotation angle must be set carefully to ensure that all the cameras are distributed uniformly and able to cover different viewing angles for a 3D model. This has been elaborated in Light Field Descriptor (LFD) [8], however different from the LFD, we discard the binary images and only use the depth images for shape retrieval task. Hence a 3D object is represented by 200 depth images from different views. The rendering image we use in our experiments has a size of 256×256.

SIFT feature extraction. After depth-images are rendered, we extract the scale and rotation invariant visual features using SIFT algorithm on each depth-image. The SIFT algorithm is carried out with two steps. First, calculate the multi-scale, multi-orientation and the position information of the interesting points detected by Difference of Gaussian method. Second, compute the SIFT descriptors in these interesting points. We use the default parameters for visual feature extraction, and the output feature vector has 128 dimension. The SIFT descriptor is designed to be robust to noise and illumination changes of images, moreover, it is stable to various changes of 3D viewpoints [12], which is a expectant property to do some compensation for lack of rotation normalization. According to our experiments, SIFT features' number vary from 20 to 40 due to the different content of every depth image. Finally, a 3D model is approximately represented with up to 5000-7000 SIFT features.

Feature encoding and histogram generation. Each SIFT feature extracted from 3D models is coded as a visual

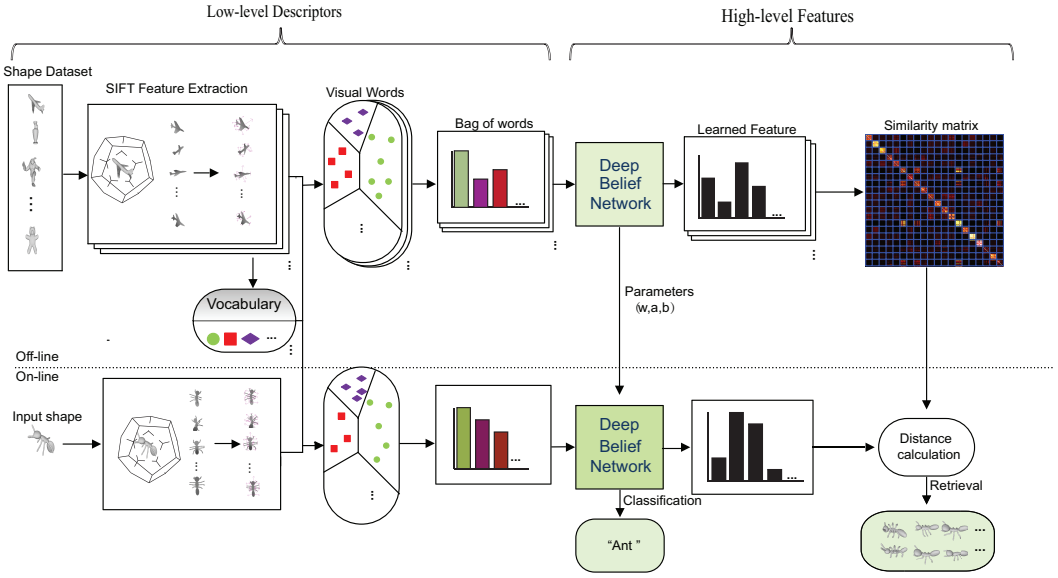

Fig. 1. The flowchart of the proposed method.

word through searching for the nearest neighbor in the vocabulary. Prior to encoding features, the vocabulary is learned with K-means algorithm via clustering the features collected from every view of every model into a specified vocabulary size. In order to obtain satisfactory discrimination capability, we set the number of words from 1000 to 3000 in experiments. In the experiment, we randomly choose 50% models per category to generate the visual vocabulary due to the tremendous computing workload of the clustering process.

2.2. Feature Learning via Deep Learning

In this work, for achieving an optimal balance between inter-class variance and intra-class affinity, it is necessary to further learn high-level features from low-level features. Deep learning aims to enlarge the inter-class distance and reduce the intra-class distance, hence suitable for the task. Due to the DBN has shown good performance and is a probabilistic approach, we adopt the DBN as the feature learning method to extract high-level features for the 3D shapes.

2.2.1. Restricted Boltzmann Machines

Restricted Boltzmann machine (RBM) is a two-layer undirected graphical model where the first layer consists of observed data variables, and the second layer consists of latent variables. The visible layer is fully connected to the hidden layer via pair-wise potentials, while both the visible and hidden layers are restricted to have no within-layer connections.

The joint distribution $p(\mathbf{v}, \mathbf{h}; \theta)$ over the visible units $\mathbf{v}$ and hidden units $\mathbf{h}$, given the model parameters $\theta = \{\mathbf{w}, \mathbf{a}, \mathbf{b}\}$, is defined in terms of an energy function $E(\mathbf{v}, \mathbf{h}; \theta)$ of

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}, \quad (1)$$

where $Z = \sum_v \sum_h exp(-E(\mathbf{v}, \mathbf{h}; \theta))$ is a normalization factor or partition function. For a Bernoulli (visible)-Bernoulli (hidden) RBM, the energy is defined as

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j, \quad (2)$$

where w_{ij} represents the symmetric interaction between the visible unit v_i and the hidden unit h_j, b_i and a_j are biases, and V and H are the number of visible and hidden units.

Because there are no direct connections between hidden units in a RBM, the conditional probabilities can be efficiently calculated as

$$p(h_j = 1|\mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^{V} w_{ij} v_i + a_j\right), \quad (3)$$

$$p(v_i = 1|\mathbf{h}; \theta) = \sigma\left(\sum_{j=1}^{H} w_{ij} h_j + b_i\right), \quad (4)$$

where $\sigma(x) = 1/(1 + exp(-x))$ is a sigmoid activation function.

The marginal probability that the model assigns to a visible vector $\mathbf{v}$ is

$$p(\mathbf{v}; \theta) = \frac{\sum_h exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}. \quad (5)$$

The derivative of the log probability of training vector with respect to a weight is

$$\frac{\partial log(p(\mathbf{v}))}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}, \quad (6)$$

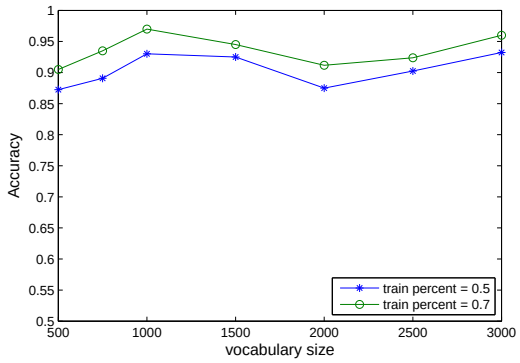

Fig. 2. Averaged classification accuracy on different visual vocabulary size BoW_n (SHREC 2007).

where the angle brackets are used to denote expectations under the distribution specified by the subscript that follows. This leads to a very simple learning rule for performing stochastic steepest ascent in the log probability of the training data

$$\Delta w_{ij} = \epsilon \left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \right), \tag{7}$$

where ϵ is a learning rate. Unfortunately, $\langle v_i h_j \rangle_{model}$ is extremely expensive to compute exactly, so that the CD approximation [3] to the gradient is used for updating the weightings.

2.2.2. Deep Belief Networks

Stacking a number of the RBMs and learning layer by layer from bottom to top gives rise to a DBN. It has been shown that the layer-by-layer greedy learning strategy [2] is effective, and the greedy procedure achieves approximate maximum likelihood learning.

In our work, the bottom layer RBM is trained with the input data of BoVF, and the activation probabilities of hidden units are treated as the input data for training the upper-layer RBM. The activation probabilities of the second-layer RBM are then used as the visible data input for the third-layer RBM, and so on. After obtain the optimal parameters $\theta = \{\mathbf{w}, \mathbf{a}, \mathbf{b}\}$, the inputed BoVF are processed layer by layer with Eq. (3) till the final layer. And the last layer's output $o(\mathbf{X})$ is used as the high-level features. In the retrieval, L_2 distance of the features is used to measure the similarity of two shapes $\mathbf{X}$ and $\mathbf{Y}$ as

$$d_s(\mathbf{X}, \mathbf{Y}) = ||o(\mathbf{X}) - o(\mathbf{Y})||_2. \tag{8}$$

3. EXPERIMENTS

In order to assess the proposed method, we use two standard 3D shape benchmarks to evaluate the performances on classification and retrieval. Experiments by using different parameters are performed to obtain optimal parameters, and then we use the optimal parameters to evaluate the retrieval performance.

In this study, SHREC 2007 watertight models [13] and McGill 3D shape benchmark [14] are used as the test data. The SHREC 2007 watertight dataset [13] is made up of 400 watertight mesh models, subdivided into 20 classes, each of which contains 20 objects with different geometrical variations and also articulated deformations. The complete McGill shape database contains 457 models including shapes with articulating parts and without articulation. The set of articulated shapes consists of 255 models in 10 categories, and there are 20-30 models per categories.

The low level visual feature extraction including image rendering, SIFT feature extraction, and vocabulary generation totally takes about 23400 seconds for SHREC 2007 watertight models and 26800 seconds for McGill 3D shape benchmark. The high level feature learning phase costs about 600 seconds. All the experiments are conducted on the platform with 8G memory and Intel i3 core processor.

3.1. Experiments on Classification

In order to assess the classification performance of the proposed method, we first study the performance while setting different parameters on SHREC 2007 dataset. We use average classification accuracy as the evaluation measure for the following experiments. The training data are randomly selected from the SHREC 2007 dataset, and the remaining data are treated as test data. First, let us see how the different word numbers affect the classification accuracy. We set the number to 500, 750, 1000, 1500, 2000, 2500 and 3000 respectively, and obtain different results, which is shown in Fig. 2. In addition, we use different size of training data to evaluate the performance. As can be seen, generally, small dictionary size leads to lower classification accuracy. Although larger dictionary size achieves better performance, the calculation time of the visual features increases rapidly, which causes low computation performance. Therefore, an optimal words number of 1000 is selected for the following experiments.

From the results of above experiments, we selected $BoW_n = 1000$ to apply into the proposed method for the following experiments. We construct four layers for DBN including input and output layers. The number of nodes in each hidden layer is empirically set to be 1000, 800. In the classification experiments, we randomly select 50% models in each category as training samples, and left models as test data. The average classification accuracies of the proposed method on SHREC 2007 and McGill dataset are 93% and 89%, respectively. Only using the BoVF, the average classification accuracies on SHREC 2007 and McGill dataset are 83% and 78%, respectively. From the comparison results, we can conclude that the high-level feature which extracted by the deep network can achieve much higher recognition accuracy, and therefore the high-level feature is more discriminative.

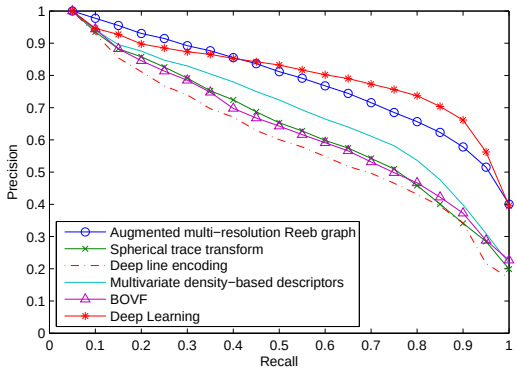

Fig. 3. Recall-precision curve of some state-of-the-art methods and proposed method on SHREC 2007 dataset.

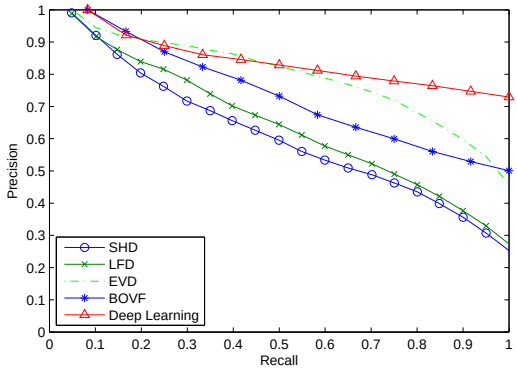

Fig. 4. Recall-precision curve of some previous methods and the proposed method on McGill dataset.

3.2. Experiments on Retrieval

Evaluation metrics. We use six standard evaluation metrics to assess the performance of the proposed method. They are precision-recall curve, nearest neighbor (NN), first tier (FT), second tier (ST), E-measure (E), and discounted cumulative gain (DCG), where the detailed definitions can be found in [15].

Experiments on SHREC 2007. First, we use the SHREC 2007 dataset to evaluate the retrieval performance of the proposed method. In the experiments, we use the shapes in training dataset to train a DBN model, and then use the meshes in SHREC 2007 to generate high-level features for similarity calculation.

The recall-precision curves of some state-of-the-art methods and proposed method are plotted in Fig. 3. From the figure, we can see that the proposed method achieves better overall retrieval performance.

The numerical evaluation measures are listed in Table 1. From the table, we can conclude all the measures have remarkable improvements from using BoVF to high-level features. The improvement of DCG index is 6.84%, which demonstrates that the proposed method has the capability to improve retrieval performance by using high-level features. We can also find that nearest neighbor (NN) has the minimal improvement in comparison with other evaluation metrics. This is mainly because the NN only checks the validity of the most nearest of the retrieval results, while our proposed method can ameliorate the whole retrieval performance.

Table 1. Retrieval performance of proposed method using standard measures on SHREC 2007 dataset.

Method	NN(%)	FT(%)	ST(%)	E(%)	DCG(%)
BoVF	88.35	54.36	33.38	47.00	85.44
Proposed Method	91.25	69.89	42.70	59.90	92.28

Table 2. Retrieval performance of proposed method using standard measures on McGill dataset .

Method	NN(%)	FT(%)	ST(%)	E(%)	DCG(%)
BoVF	87.31	44.59	28.69	41.69	81.37
Proposed Method	90.15	61.81	39.79	58.29	89.23

Experiments on McGill dataset. In order to assess the robustness of the proposed method, we also perform an experiment on McGill dataset of articulated 3D shapes [14]. The pre-calculated BoVF and DBN model are used to generate high-level features for shape retrieval. The recall-precision curves of the proposed method and some other methods are shown in Fig. 4, which includes shape harmonic descriptor (SHD) [16], light-field descriptor (LFD) [8], eigenvalue descriptor (EVD) of affinity matrix [7]. From the figure, we can see that the average retrieval performance of the proposed high-level features is superior than other methods. Same as the previous experiment, through the deep learning, the retrieval accuracy is also noticeably improved from that using BoVF. The six standard evaluation metrics on McGill dataset are listed in Table 2.

4. CONCLUSION

In this paper, we present a novel high-level feature learning approach for 3D shapes. In the first stage, low-level 3D shape visual features and BoVF features are extracted to represent order-irrelevant features. And then DBN is adopted to learn semantic high-level features. This high-level feature is employed to perform 3D shape recognition and retrieval. Experimental results demonstrate that the learned high-level features can achieve better performance compared to other methods. The experiments results also show that the learned high-level features are more discriminative which can sup-

press intra-class variation and enhance the inter-class similarity separation. From the retrieval results shown in Fig. 3 and Fig. 4, the overall retrieval performances are boosted by using the learned high-level feature compared to that using the low-level features.

Different from traditional deep learning methods used in image processing, we adopt visual BoVF as the input for deep learning in this work, because the order of images taken from views are not directly comparable. Furthermore, directly using low-level features as input for deep learning may require lots of training data. But usually 3D shape dataset only contains a small amount of shapes, thereby, it is impossible to achieve commendable accuracy. Through the proposed method, only a small amount training samples are required to train a DBN model, and adequate performance can be accomplished. Nevertheless, the retrieval performance can be further improved by utilizing more training data.

At present we only investigate SIFT as the low-level descriptors, although promising results are obtained, better retrieval and recognition performance may be achieved by using other visual descriptors or sampling scheme. In addition, in practical applications especially for Kinect, only singe-view image and depth image are available for shape recognition or retrieval, thereby, it is necessary to develop a method which can take full advantage of RGB-D information.

5. REFERENCES

[1] Menon V Bressler S L, "Large-scale brain networks in cognition: emerging methods and principles," *Trends in cognitive sciences*, vol. 14, no. 6, pp. 277–290, 2010.

[2] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[4] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[5] Dan Ciresan, Ueli Meier, and Jürgen Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[6] Zhenbao Liu, Shuhui Bu, Kun Zhou, Shuming Gao, Junwei Han, and Jun Wu, "A survey on partial retrieval of 3d shapes," *Journal of Computer Science and Technology*, vol. 28, no. 5, pp. 836–851, 2013.

[7] Varun Jain and Hao Zhang, "A spectral approach to shape-based retrieval of articulated 3d models," *Computer-Aided Design*, vol. 39, no. 5, pp. 398–407, 2007.

[8] Yu-Te Shen, Ding-Yun Chen, Xiao-Pei Tian, and Ming Ouhyoung, "3d model search engine based on light-field descriptors," *EUROGRAPHICS Interactive Demos, Granada, Spain*, pp. 1–6, 2003.

[9] Takahiko Furuya and Ryutarou Ohbuchi, "Dense sampling and fast encoding for 3d model retrieval using bag-of-visual features," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, p. 26.

[10] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai, "3-d object retrieval and recognition with hypergraph analysis," *Image Processing, IEEE Transactions on*, vol. 21, no. 9, pp. 4290–4303, 2012.

[11] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] Zhouhui Lian, Afzal Godil, and Xianfang Sun, "Visual similarity based 3d shape retrieval using bag-of-features," in *Shape Modeling International Conference (SMI), 2010*. IEEE, 2010, pp. 25–36.

[13] RC Veltkamp and FB ter Harr, "Shrec 2007 3d shape retrieval contest," Tech. Rep., Utrecht University, Technical Report UU-CS-2007-015, 2007.

[14] Kaleem Siddiqi, Juan Zhang, Diego Macrini, Ali Shokoufandeh, Sylvain Bouix, and Sven Dickinson, "Retrieving articulated 3-d models using medial surfaces," *Machine Vision and Applications*, vol. 19, no. 4, pp. 261–275, 2008.

[15] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser, "The princeton shape benchmark," in *Shape Modeling Applications, 2004. Proceedings*. IEEE, 2004, pp. 167–178.

[16] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3d shape descriptors," in *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. Eurographics Association, 2003, pp. 156–164.