Learning High-Level Feature by Deep Belief Networks for 3-D Model Retrieval and Recognition

Shuhui Bu, *Member, IEEE*, Zhenbao Liu, *Member, IEEE*, Junwei Han, *Member, IEEE*, Jun Wu, *Member, IEEE*, and Rongrong Ji, *Senior Member, IEEE*

***Abstract*—3-D shape analysis has attracted extensive research efforts in recent years, where the major challenge lies in designing an effective high-level 3-D shape feature. In this paper, we propose a multi-level 3-D shape feature extraction framework by using deep learning. The low-level 3-D shape descriptors are first encoded into geometric bag-of-words, from which middle-level patterns are discovered to explore geometric relationships among words. After that, high-level shape features are learned via deep belief networks, which are more discriminative for the tasks of shape classification and retrieval. Experiments on 3-D shape recognition and retrieval demonstrate the superior performance of the proposed method in comparison to the state-of-the-art methods.**

***Index Terms*—3-D model recognition, 3-D model retrieval, bag-of-words, deep belief networks, deep learning.**

I. Introduction

THREE-DIMENSIONAL models have been extensively applied in the domains of multimedia, graphics, virtual reality, amusement, design, and manufacturing [1] due to the rich information preserving the surface, color, and texture of real objects. A huge number of publicly available models such as Google 3D Warehouse have been quickly spread online. With the development of RGB-D devices, e.g., Microsoft Kinect, users can obtain 3D models in a convenient and efficient way, which further leads to the explosion of 3D data. This rapidly increasing of 3D model data requires effective retrieval and classification techniques [2] for their management and reusing.

Manuscript received December 07, 2013; revised June 05, 2014; accepted August 18, 2014. Date of publication August 28, 2014; date of current version November 13, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61202185, Grant 61003137, Grant 91120005, Grant 61473231, Grant 61373076, Grant 61422210, and Grant 61103062, the Fundamental Research Funds for the Central Universities under Grant 310201401-(JCQ01009, JCQ01012) and Grant 2013121026, the Shaanxi Natural Science Fund under Grant 2012JQ8037, the Open Project Program of the State Key Lab of CAD&CG of Zhejiang University under Grant A1306, and Xiamen University under the 985 Project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan. *(Corresponding author: Junwei Han.)*

S. Bu, Z. Liu, J. Han, and J. Wu are with Northwestern Polytechnical University, Xi'an 710072, China (e-mail: bushuhui@nwpu.edu.cn; liuzhenbao@nwpu.edu.cn; jhan@nwpu.edu.cn; junwu@nwpu.edu.cn).

R. Ji is with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian 361005, China (e-mail: rrji@xmu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2351788

Up to present, a large amount of methods for 3D shape recognition and retrieval have been proposed. Among them, a commonly adopted pipeline is to first extract low-level descriptors, and then train classifiers for shape recognition or compare descriptor similarity for shape retrieval. Many local descriptors [1], [3] for 3D shape have been proposed to describe 3D shapes from different facets. These descriptors are used to characterize important local geometric statistics of 3D surfaces, which are distinctively discriminative with other local regions.

Despite the progressive improvement on recognition, matching, and retrieval, the performance of the existing 3D shape descriptors are still far from satisfactory. The main issue lies in the insufficiency in describing complex 3D shapes using *local* statistic, i.e., each type of local descriptors only catches a piece of geometric characteristics, moreover 3D shape is composed of complex topological structure and visibly variational geometry. In order to further enhance their discriminability, it is natural to consider how to combine multiple descriptors together to provide complementary information. One commonly adopted scheme is to regard a 3D shape as a collection of primitive elements, each of which is separately described by various low-level descriptors. Then middle-level features are extracted from low-level descriptors to enable a stronger description and generalization ability. One approach is the bag-of-features (BoF), in which similar geometric information repeatably appearing on 3D shapes are packed into a same bag, and the shape similarity is then measured via the difference between the bag occurrence histograms. Shape Google [4], [5] is one of the representative works. In this method, heat kernel signature [6] and its scale invariant version [7] are used as local descriptors considering global heat diffusion, and a vocabulary of geometric words are generated by means of unsupervised clustering. After that, spatially sensitive bag-of-features (SS-BoF) are introduced to encapsulate the relationship between spatially close words. Lavoue [8] considers spatial information of bag-of-features by introducing a histogram of pairs of visual words, which are built by describing dense point samples with local Fourier descriptor on local Laplace-Beltrami space. Toldo *et al.* [9] adopt sparse segmented region descriptor instead of local point descriptor for middle-level feature extraction, based on which assigned all part descriptors of objects into words.

Recently, with the rapid progress of machine learning, feature learning which can improve the discriminability of low-level feature is becoming a hot research topic. Castellani *et al.* [10] propose a middle-level feature extraction scheme through

learning hidden states from local descriptors. In the method, local patches are modeled as a stochastic process through a set of circular geodesic pathways and learned via hidden Markov model. Bu *et al.* [11] propose shift-invariant ring feature based on iso-geodesic rings and shift-invariant sparse coding for 3D shape analysis. It represents the local region of one feature point efficiently and has great performance on correspondence and retrieval tasks. In the research of Shape Google [4], [5], despite the introduction of SS-BoF, the authors also present a similarity-sensitive hashing method to achieve the best discriminativity and compact representation. The Laplacian-based descriptors achieve state-of-the-art performance, however, they usually focus on different aspect properties of shapes and suitable for specified task. In order to provide a generic feature descriptor for 3D shape, Litman *et al.* [12] propose a learning scheme for the construction of optimized spectral descriptors. In order to select most significant features for shape retrieval and classification, Barra *et al.* [13] propose a method utilizing multiple kernel learning to find optimal linear combination of kernels in classification and retrieval. To address the problem of automatic recognition of functional parts of man-made 3D shapes, Laga *et al.* [14] use graph to represent 3D shape, and then model the context of a shape part as walks in the graph. In the method, the similarity computation can be efficiently performed using graph kernels. Leng *et al.* [15] present an interactive learning mechanism, which establishes a mapping from feature points in low-level feature space to points in high-level semantic space. The mechanism receives long-term relevance feedback from users via recorded retrieval history, which is adopted to capture users' semantic information to refine retrieval results. Best view selection is an important procedure in view-based shape retrieval, in order to achieve better performance, Laga [16] propose a framework to automatically select best views of 3D models by learning sets of 2D views that maximize the similarity between shapes of the same class and also the views that discriminate shape of different classes. To cope with the problem of low compactness and discrimination power of view-based descriptors, Tabia *et al.* [17] adopt vectors of locally aggregated tensors to generate descriptor, and then use Principal Component Analysis to reduce the dimension of the descriptor. Gao *et al.* [18] propose a 3D object retrieval method with Hausdorff distance learning. In their method, relevance feedback information is employed to select positive and negative view pairs with a probabilistic strategy.

The advantage of the above discussed methods is that they use higher-order features or learning approaches to achieve higher accuracy in retrieval or recognition. In this paper, we provide an alternative high-level feature learning framework based on deep learning [19] for 3D shape recognition and retrieval, which differs from the existing approaches by focusing on automatically building a deep hierarchical structure that pools low-level 3D shape descriptors and come out with a high-level, better descriptive shape representation by learning from a large collection of shapes through deep belief networks.

As the key innovation leveraged in this paper, deep learning [19] has been studied in natural language processing, speech recognition, image processing, and computer vision, demonstrating significant improvement by introducing more discriminative and robust high-level features from shallow features via multi-layer network. In addition, it 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.* [20] adopted a convolutional deep belief network 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.* [21] proposed a multi-column deep neural networks for image classification. On the very competitive MNIST handwriting benchmark, the method achieves near-human performance.

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. However, borrowing deep learning into 3D shape is not straightforward. As we know, image data is represented by a rectangular grid structure (can be regarded as a graph with fixed connections), which has the merit that multi-level or hierarchical representation can be easily achieved. In addition, the pixels of image can be represented in a fixed sequence, such as row-major order which converts a 2D array into a linear array by storing one row after the other. It results in that raw image pixel data or feature can be directly input to deep learning framework. However, this framework does not work for 3D models because 3D meshes are with different tessellations and without any order, which is intrinsically different from image data.

In order to overcome the problem, we propose a novel framework which builds a bridge between 3D shape and deep learning. The core idea is to extract a middle-level position-independent feature from any low-level 3D descriptors, and then generate high-level features for 3D shape retrieval via deep learning. The positional independence denotes that the middle-level feature is irrelevant with the order of low-level feature, as to handle the afore-mentioned paradox. The benefit of these middle-level features lies in that they encode not only the geometric information but also their spatial relationship. In addition, this feature has a form of 2D as same structure of normal images. Therefore it can be input into deep learning framework for learning high-level features, which provides structural information and have more powerful generalization. The main contributions of this work are two-fold.

- We propose a deep learning framework for shape classification and retrieval. To the best of our knowledge, it is the first time to apply deep learning into 3D shape retrieval.
- Our work can extract higher level features from local descriptors, and these features have better discriminability and generalization against intra-class large geometrical variations.

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 the promising performance.

II. Related Work

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, and next we will give a brief discussion on different description means. Comprehensive and excellent reviews on descriptor and shape

retrieval can be found in an early work [1] and latest work [3]. According to different levels in shape representation, for example, shape-level and part-level (element-level), state-of-the-art methods can be divided into two gross categories: global description and local description.

Global description only considers the whole shape characteristics of a 3D model, and early representative works include Euclidean distance distribution [22] and spherical harmonics descriptor [23]. 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.* [24] propose to use first eigenvalues of the geodesic distance matrix of a 3D object to generate 3D shape descriptor which are isometry-invariant. Smeets *et al.* [25] propose a similar approach which is based on spectral decomposition of the geodesic distance matrix (SD-GDM). Another different type of methods take into account global topological information of 3D shape, and convert 3D mesh into low-dimensional topology structure equivalent to the connection of a 3D model. For example, skeletal representations of 3D volumetric objects are used to assist in retrieving shapes with similar skeletons [26]. Mademlis *et al.* [27] extract medial surfaces of a 3D shape, and decompose the whole shape into several parts for constructing a graph. Tierny *et al.* [28] construct the Reeb graph of a closed 2-manifold 3D object surrounding vertices located on the extremity of prominent components, and segment it into a set of Reeb charts for 3D model retrieval. Barra and Biasotti [29] similarly defines shape parts and their connection by aggregating the 10 kernels obtained from 10 different Reeb graphs. The merit of Reeb graph is that it consists of only one dimensional graph structure excluding any degenerate surfaces, which can simplify the problem of shape retrieval.

View-based methods are special type of global descriptor which based on an assumption that if two 3D models are geometrically similar they also look similar. Different from above mentioned methods, these methods do not require the utilization of geometric attributes or topological relationship, and hence are capable of handling 3D models with degeneration, holes, and missing patches. They commonly rely on generating panoramic view [30] or many projections of a 3D object at full viewpoints, for example, captured from cameras located on an unit sphere [31] or cube, and measure the similarity among 3D models by visual similarity which is considered as a direct way consistent with human perception. Most of view-based methods depend on camera array settings for capturing views of 3D object, in order to relax the restriction, a camera constraint-free view-based approach [32] is proposed to achieve better performance. 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 hypergraph of views [33]. In order to avoid the inefficiency from a large number of view comparison, Gao *et al.* [34] adopt only a small set of query views to obtain less computational cost during the comparison with target shapes.

Different from global representation of a 3D shape, local description captures important geometric changes on local regions of 3D surface, which are distinctively discriminative with local regions from another shape. An earlier and representative work is spin images [35], and it has been introduced into 3D shape retrieval [36]. Sipiran *et al.* [37] adopt 3D Harris detector to detect interest points for 3D shape retrieval, which can be seen as extension from 2D Harris detector measuring the variation in the gradient of a given function (e.g., the intensity function of a image) due to shifts in a local window. 3D SURF descriptor [38] is recently proposed for classifying and retrieving similar shapes.

The above local descriptions depend on extrinsic properties constrained by location and orientation of 3D mesh, or a local coordinate system defined on a mesh vertex. In order to provide intrinsic properties of shapes, local descriptors without specifying the descriptor position relative to an arbitrarily defined coordinate system have been studied. Laplace Beltrami operator, which is a generalization of the Laplacian from flat space to manifold, is considered appealing for 3D shape retrieval benefiting from sparse, symmetric, and intrinsic properties and its robustness to rigid transformation and deformation. Retrieval methods [39]–[41] extract main eigenvalues and eigenvectors of Laplace matrix generated on local regions to match different regions of 3D shapes. In a recent work [42], 3D shape is also partitioned into several connected iso-surfaces (annuluses) of conformal factors, and expressed with a graph which node substitutes each annulus. Heat kernel signature [6], a recently proposed local descriptor, absorbs much attention from researchers. The rich local geometric information, invariance to isometric deformations, and multi-scale characteristic make the signature capable of working well in 3D shape retrieval and matching [5], [7], [43]. In order to overcome the influence of diffusion time changes under different shape scales [7], Fourier transform is imposed on heat kernel signature at each given vertex to obtain time invariants.

In this work, we aim at building high-level features of 3D model by virtue of deep learning framework, and forming a highly discriminative description to recognize intra-class models with large geometric variations and separate irrelevant objects. The difference between global description mentioned above and proposed method lies in that extracted shape feature are with higher descriptive power and not easy to be influenced by whole geometrical and topological changes. Although high-level features from deep learning are based on local geometrical description, low-level local descriptors are difficult to discriminate different parts with approximate descriptor values, and middle-level local descriptors such as bag-of-words only considers clusters of these local descriptors. We build the connection relationship between high-level feature and local descriptor via deep networks. Moreover, our work do not require the generation of many views compared with view based methods, and hence reduce the storage cost and comparison cost for 3D model retrieval systems.

III. High-Level Feature Learning for 3-D Shapes

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

A. Low-Level 3-D Shape Descriptors

In this research, we adopt scale-invariant heat kernel signature and average geodesic distance as the low-level 3D shape descriptors, since these two local descriptors are robust against

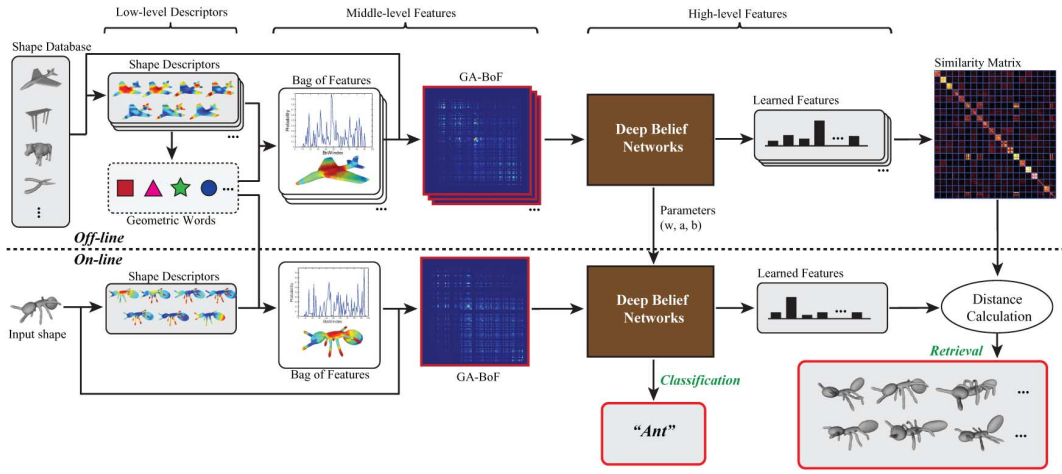

Fig. 1. Flowchart of the proposed method.

non-rigid and complex shape deformations. More importantly, they both consider the global shape information, which is necessary for global shape retrieval.

Heat Kernel Signature (HKS): HKS has caught much attention in shape retrieval [5], [7], [43], because it can capture rich local geometric information, is invariant to isometric deformations, and has multi-scale characteristic. The heat diffusion over a manifold mesh X is described by the heat equation

$$(\Delta + \frac{\partial}{\partial t})u(x,t) = 0 \tag{1}$$

where Δ represents the negative semi-definite Laplace-Beltrami operator, $u(x,t)$ is the distribution of heat on the manifold at point x at time t. The initial condition is some initial heat distribution $u(x,0)$ which typically is a Dirac delta function on the analyzed vertex x, and boundary conditions are required if the manifold has a boundary. The equation describes the conduction of heat u on the surface X with respect to time t. The fundamental solution $K_t(x,y)$ of the heat equation is called the heat kernel, which can be understood as the amount of heat transfered from a source vertex x to a target vertex y in time t.

In order to obtain a compact point signature, Sun *et al.* [6] proposed to use the diagonal of the heat kernel as a local descriptor, referred to as heat kernel signature (HKS). For each point x on the shape, its HKS is an n-dimensional feature vector as

$$HKS(x) = c(x)\left(K_{t1}(x,x), \ldots, K_{tn}(x,x)\right) \tag{2}$$

where $c(x)$ is chosen in such a way that $\|HKS(x)\|_2 = 1$. The HKS expresses the physical meaning that when a quantity of heat is placed on a given vertex how much the heat is remained after time t elapses. This descriptor captures information about the neighborhood of a point x on the shape at a time scale defined by t, the information varies from small neighbor with small t to global shape with larger t. In addition, the computation of HKS relies on the first eigenfunctions and eigenvalues of the Laplace-Beltrami operator, which can be done efficiently in the eigen decomposition.

Scale-Invariant Heat Kernel Signature (SI-HKS): However, a limitation of the HKS is that it is sensitive to the scale of shape, for example, when the shape becomes large the region described by the HKS becomes small in the same time range. In order to resolve the problem, Bronstein *et al.* [7] proposed a scale-invariant heat kernel signature (SI-HKS) by Fourier transform of the difference of the HKS, as follows:

$$\begin{aligned} h_{diff}(x) = (&logK_{\alpha^{\tau_2}}(x,x) - logK_{\alpha^{\tau_1}}(x,x), \ldots, \\ &logK_{\alpha^{\tau_m}}(x,x) - logK_{\alpha^{\tau_{m-1}}}(x,x)), \end{aligned} \tag{3}$$

$$SIHKS(x) = \left|(\mathcal{F}_{h_{diff}}(x))(\omega_1, \ldots, \omega_n)\right| \tag{4}$$

where $\mathcal{F}$ is the discrete Fourier transform, and $\omega_1, \ldots, \omega_n$ denote a set of frequencies at which the transformed vector is sampled. The scale-invariant values are obtained by first calculating the differences of logarithmic heat kernels between time step τ_m and τ_{m-1}, and then selecting the absolute values of Fourier transform of the differences as features. Usually, a large m is used to make the representation being insensitive to large scaling factors.

Average Geodesic Distance: The average geodesic distance (AGD) is introduced by Hilaga *et al.* [44] for the purpose of shape matching. Let $g(x_i, x_j)$ be the geodesic distance between two vertices x_i and x_j on the mesh X, the average distance of x_i to all other vertices is defined as follows:

$$A_n(x_i) = \sqrt[n]{\frac{\sum_{x_j \in X} g(x_i, x_j)}{Area(X)}} \tag{5}$$

where $Area(X)$ denotes the total area of the mesh. However, it still suffers from the influence of scale changes as it is only used to normalize the AGD value. Therefore, in its original implementation [44], a minimum value of $A_n(x)$ is used as normalization factor to make the descriptor scale-invariant.

$$AGD_n(x_i) = \frac{A_n(x_i)}{min_{x_j \in X} A_n(x_j)}. \tag{6}$$

For any $n \geq 1$, the value of AGD_n measures how "isolated" a point is from the rest points of the surface. In addition, its local maxima coincide with the tips of the surface.

However, the original AGD is not robust when using extremum value as a normalization factor, e.g. the use of the intra-class geometric variations make the local descriptor change easily. It is therefore difficult to be applied to generate

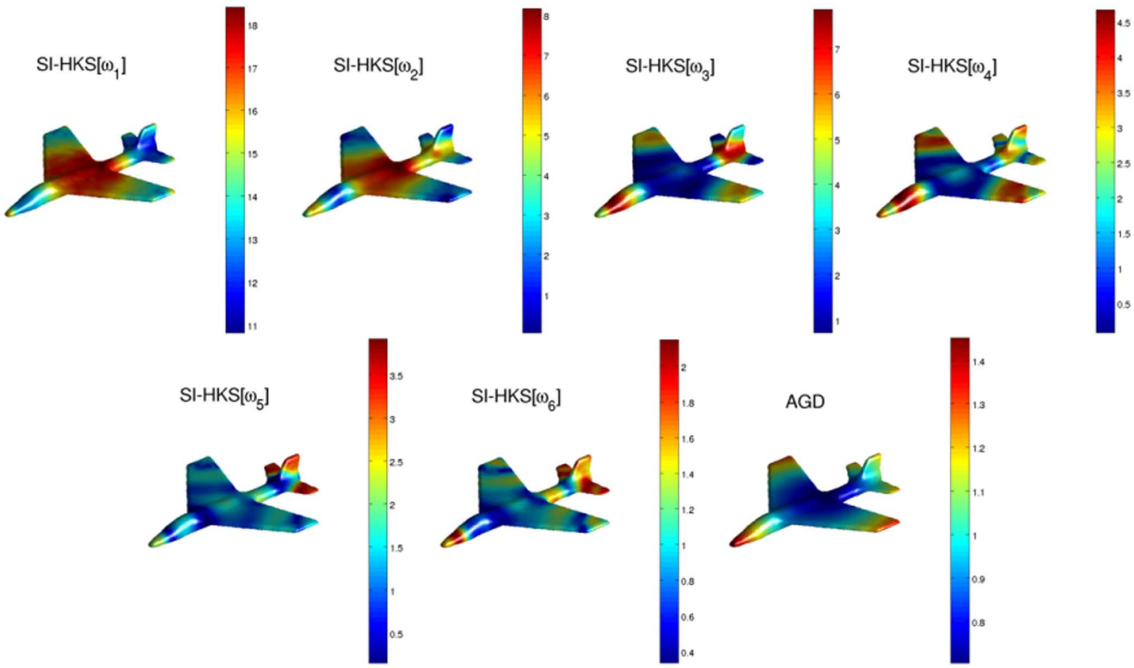

Fig. 2. The visualization of geometric descriptors of a model.

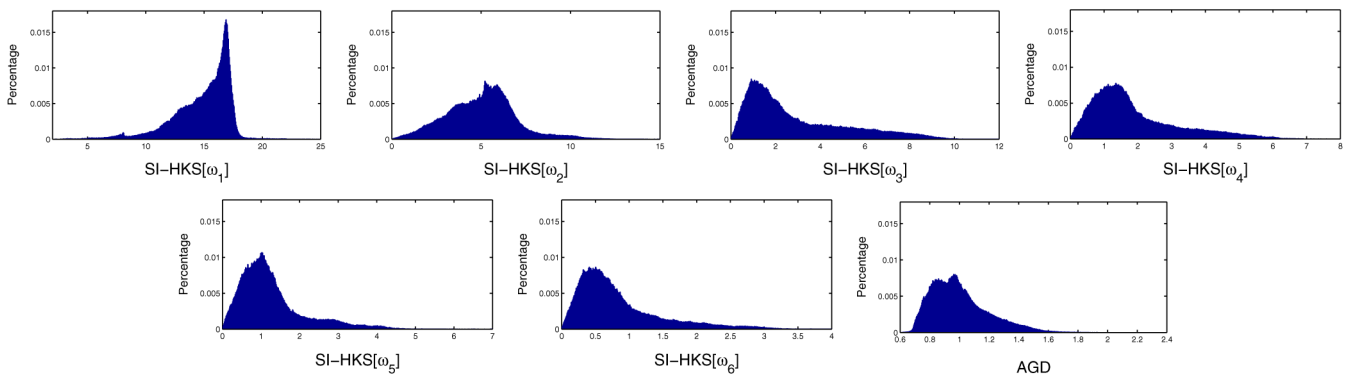

Fig. 3. Histogram of each descriptor component over all models in SHREC 2007.

bag-of-words from a set of models. We then modify the normalization factor in (6) to the mean of $A_n(x)$ as follows:

$$AGD(x_i) = \frac{A_n(x_i)}{\sum_{x_j \in X} A_n(x_j)/N} \tag{7}$$

where N is the vertex number of the mesh. For any model, the modified AGD descriptor has a fixed mean value 1.

Low-Level Descriptors: Finally, we concatenate the first six frequency components of SI-HKS and AGD descriptor to form a low-level shape descriptor as

$$F(x_i) = (SIHKS(x_i)[\omega_1, \ldots, \omega_6], AGD(x_i)) \tag{8}$$

where the dimension of the feature is $M = 7$. Although these two features are just simply combined without weighting, in the following step the feature weighting will be considered. Fig. 2 shows an example of extracted descriptors. For the SI-HKS, the time-scale is set to be $[1, 20]$ with an interval of 0.2, the number of eigenfunction is set to 100, and the log time base $\alpha = 2$. The parameter of n in Eq. (5) is set to 1.

B. Middle-Level Features

In this step, bag-of-features are computed to represent the occurrence probability of geometric words. To this end, we do not adopt the widely adopted clustering method to generate geometric vocabulary, due to the value ranges of each dimension in the descriptor are different as illustrated in Fig. 3, and consequently the contributions to the feature discrimination are also different. Thereby, without any weighting scheme to the descriptors, the clustering lacks defensing against noisy feature and can not handle feature variances. In this work, we adopt Minkowski metric and feature weighting [45] for k-means to generate geometric words more precisely.

After the geometric words $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ of size K is obtained, the next step is to quantize the low-level descriptor space in order to obtain a compact representation. For each point $x \in X$ with the descriptor $F(x)$, we define the feature distribution $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))^T$ as a $K \times 1$ vector whose entries are

$$\phi_i(x) = c(x) exp\left(-\frac{||F(x) - c_i||_2^2}{k_{BoF}\sigma_{min}^2}\right) \tag{9}$$

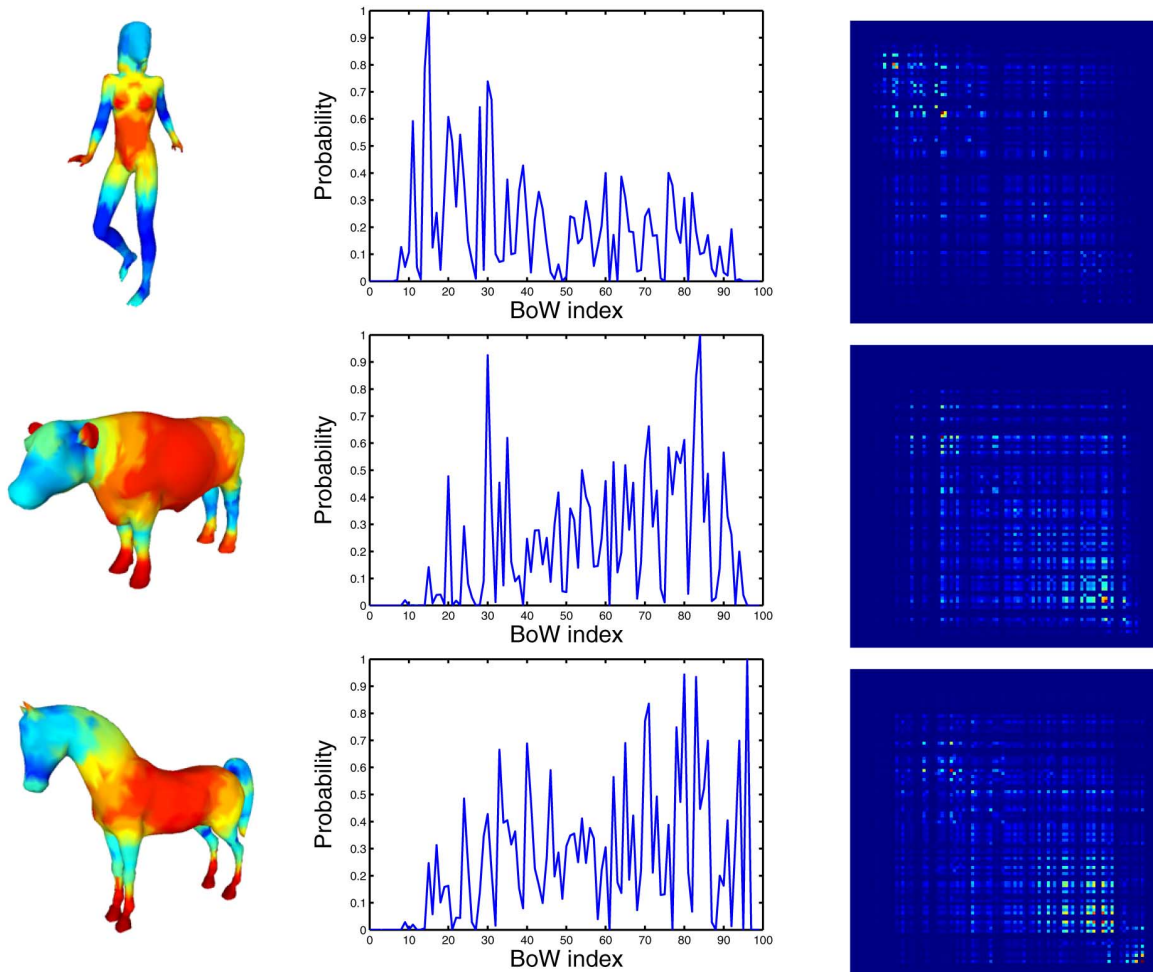

Fig. 4. Some examples of bag-of-words distributions and GA-BoF features. Left: Bag-of-words indices are visualized at mesh's vertex. Middle: Bag-of-words distribution over the mesh. Right: GA-BoF features.

where the constant $c(x)$ is selected to satisfy $\|\phi(x)\|_1 = 1$. The above equation is a soft vector quantization. The advantage is that it can generate more representative probability feature values, and $\phi_i(x)$ can be interpreted as the probability of the point x to be associated with the geometric word c_i. Two parameters are utilized to control effects of vector quantization, where σ_{min} is the minimum distance between any two geometric words, and k_{BoF} is a parameter controlling the decay coefficient along the distance to geometric word. Left column figures in Fig. 4 visualize the index of most probabilistic word on the mesh, and middle figures show the overall BoF distribution by summing all BoF distributions of vertices on the mesh. From the figures in left column, we can intuitively see that for the same type of shapes, the overall distributions of most probabilistic word index have similar patterns, e.g. four-leg animals, a cow and a horse. In contrast, the distributions for models in different category are distinctly different.

The disadvantage of bag-of-features is the fact that they only consider the occurrence distribution of the words and ignore the structural relationship between them, which decreases their discrimination. For geometric shapes, only using features of vertex leads to limited descriptive capability, and the structural information of vertices is more critical for representing 3D topological connections. This issue also exists in natural language processing, in which an effective approach is to extend the vocabulary to include not only words but also contextual phrases. The situation is analogous to the structural information of vertices, which can be represented by means of spatially related geometric words. Therefore, inspired by Shape Google [4], [5], we use the geodesics on the mesh to measure the spatial relationship between each pair of BoFs on vertices. Different from Shape Google, we consider geodesic distance instead of heat kernel to avoid the possible influence from time scale and shape size, and introduce the geodesics-aware bag-of-features (GA-BoF)

$$\mathbf{v}(X) = N(X) \sum_{x_i \in X} \sum_{x_j \in X} \phi(x_i)\phi(x_j)^T exp\left(-k_{gd}\frac{g_d(x_i, x_j)}{\sigma_{gd}}\right) \tag{10}$$

where $N(X)$ is a normalization factor which makes features have a fixed maximum value of 1, σ_{gd} is the maximal geodesic distance of any pair of vertices on the mesh, and k_{gd} denotes the decay rate of distances, which is selected empirically.

The resulting $\mathbf{v}$ is a $K \times K$ matrix, which represents the frequency of geometric words i and j appearing within a specified geodesic distance. This expression provides occurrence probability of geometric words and relationship between them. Moreover, it provides a position-independent representation of shape, where the positional independence denotes the middle-level fea-

ture is irrelevant with the order of low-level features or vertices. Right column in Fig. 4 shows GA-BoF of three shapes. As can be seen, the patterns of GA-BoF are similar between shapes in same class, while they are obviously different between irrelevant shapes.

C. Feature Learning via Deep Learning

In order to further deeply mine the relationship of features from intra-class shapes and inter-class shapes in a large dataset, it is necessary to further learn high-level features from middle-level features for 3D shapes. However, due to the intrinsic differences between structures of 3D shapes and images or speech data, it is difficult to input the raw mesh into the deep learning directly. The GA-BoF can be regarded as a relationship matrix each entry of which represents the occurrence probability of two geometric words within a specified geodesic distance. Furthermore, all the shapes have the same size of GA-BoFs, and the middle-level feature is invariant to the order and the number of vertices on the mesh. Therefore, it is appropriate to construct a deep learning network.

Recent works on deep belief networks (DBNs) [46], [47] have shown that it is feasible to learn multiple layers of non-linear features that are useful for object classification. The input data are trained layer by layer in a restricted Boltzmann machine (RBM) [48] by means of contrastive divergence (CD) [48]. Due to the fact that DBN has shown good performance and is a probabilistic approach, we adopt DBN as the feature learning method to extract high-level features for the 3D shapes.

Restricted Boltzmann Machines: In order to make the paper more self-contained, we succinctly discuss the concept of restricted Boltzmann machine (RBM). 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 pairwise potentials, while both the visible and the 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

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z} \tag{11}$$

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

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j \tag{12}$$

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

$$p(h_j = 1 | \mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^{V} w_{ij} v_i + a_j\right) \tag{13}$$

$$p(v_i = 1 | \mathbf{h}; \theta) = \sigma\left(\sum_{j=1}^{H} w_{ij} h_j + b_i\right) \tag{14}$$

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

$$p(\mathbf{v}; \theta) = \frac{\sum_h exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}. \tag{15}$$

The probability that the network assigns to a training data can be raised by adjusting the weights and biases to lower the energy of that data and raise the energy of other data, especially those that have low energies and therefore make a big contribution to the partition function. The derivative of the log probability of training vector with respect to a weight is

$$\frac{\partial log(p(\mathbf{v}))}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{16}$$

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 to perform stochastic steepest ascent in the log probability of the training data

$$\Delta w_{ij} = \varepsilon \left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \right) \tag{17}$$

where ε is a learning rate. It is noted that $\langle v_i h_j \rangle_{model}$ is extremely expensive to compute exactly, so that the CD approximation [48] to the gradient is used for updating the weightings.

D. 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 [46] 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 GA-BoF, and the activation probabilities of hidden units are treated as the input data for training the upper-layer RBM, and so on.

Unsupervised Feature Learning: In the shape retrieval task, it is difficult to obtain all class labels of the training data, thereby, the un-supervised training is suitable for the mission. We use the un-labeled 3D shape data to train the DBN layer-by-layer. After obtain the optimal parameters $\theta = \{\mathbf{w}, \mathbf{a}, \mathbf{b}\}$, the inputed GA-BoFs are processed layer-by-layer with Eq. (13) 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

$$d_s(\mathbf{X}, \mathbf{Y}) = \|o(\mathbf{X}) - o(\mathbf{Y})\|_2. \tag{18}$$

Supervised Feature Learning: Usually, DBNs are trained in an un-supervised fashion. If labels are provided to the networks, recognition performance can be further improved, benefiting from joint optimization of hidden layer and class labels. In the shape recognition task, we fully utilize the information of training set, including their features and corresponding labels, so as to obtain a higher discriminative DBN. A discriminative RBM [49] is adopted as the top layer of the DBN. After the model is trained, by inputing the GA-BoF of the test data into the model, the class label output is used to recognize the shape.

IV. Experiments

In order to assess the proposed method, we use three standard 3D shape benchmarks and a mixed dataset to evaluate the performances on classification and retrieval. Several experiments by using different parameters are performed to discover optimal parameters, and then we use the optimal parameters to evaluate the classification and retrieval performance.

Dataset: In this study, SHREC 2007 watertight models [50], Princeton shape benchmark [51], SHREC 2011 Non-rigid 3D watertight dataset [52], and a mixed dataset are used as the test data.

The SHREC 2007 watertight dataset [50] 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 data set contains not only natural objects but also man-made objects. SHREC 2011 non-rigid dataset [52] consists of 600 watertight triangle meshes that are transformed from 30 original models. The Princeton shape benchmark (PSB) [51] contains a database of 3D polygonal models collected from the web. There are 1,814 models which are divided into three levels of classifications from geometric level to functional level, and then to high level which just consists of 'manmade' and 'natural' objects.

Implementation Details: The major part of the code is written in MATLAB, and some parts of the codes are written in C++. The experiments are run on a computer with a 3.2 GHz Intel Xeon CPU and 8 GB of RAM. The required time for computing low-level descriptors per shape is about 3-12 seconds, depending mostly on the number of vertices. Computational time for generating the BoWs is 10-20 minutes depending on the number of clusters. The calculation of GA-BoF for a model will cost 5-9 seconds which related to the number of BoWs and vertices. The DBN training is performed off-line, and the required time is about 20-30 minutes depending on the number of hidden layer nodes and iteration number, while the computation of high-level features by DBN is negligible.

A. Experiments on Classification

In order to assess the performance of the proposed method, we first study the performance while setting different parameters. 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 64, 80, 100, 128, and 160 respectively, and obtain different performances, which is shown in Fig. 5. In addition, we use different number of training data to evaluate the performance. As can be seen, generally small dictionary size of BoW leads to lower classification accuracy. Although larger dictionary size achieves better performance, the calculation time of GA-BoF increases rapidly, which causes low computation performance. Therefore, an optimal BoW number of 100 is selected for the following experiments. From the Fig. 5, we can also get that the averaged accuracy when using only 10% training examples (2 shapes) can achieve relatively acceptable results. Therefore, the proposed method are not heavily rely on the number of training examples. Different from image applications, usually there do not exist so many 3D shapes for training, thereby it is a special advantage for 3D shape applications.

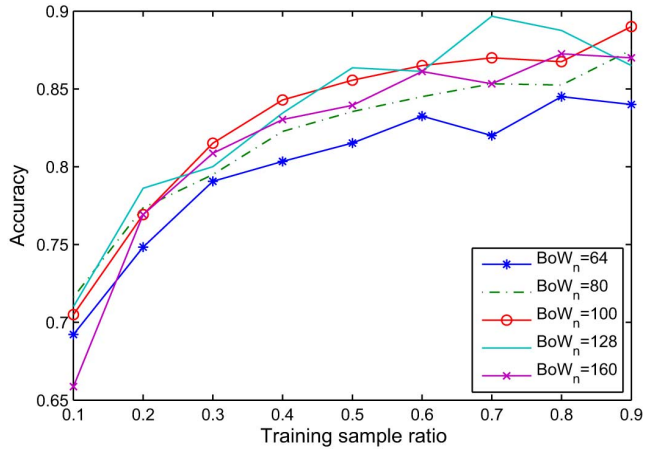

Fig. 5. Averaged classification accuracy on different bag-of-words number BoW_n (SHREC 2007).

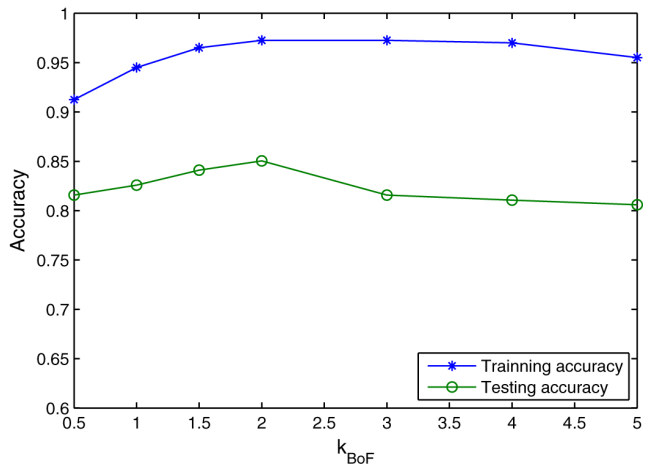

Fig. 6. Averaged classification accuracy on different k_{BoF} (SHREC 2007).

In the second experiment, we use different k_{BoF} to evaluate the classification accuracy of training and testing data by using the BoW number of 100. The results are plotted in Fig. 6. From this figure, we can conclude that when k_{BoF} is 2, the best accuracy is achieved. When k_{BoF} is larger than 2 it drops slowly. Because, this parameter controls how closely BoW are selected as the BoF, larger value will make fewer similar BoW are selected.

Next, we study the effects of different k_{gd}. This parameter controls the decay rate for calculating the GA-BoF. If small value is set, a pair of vertices with large geodesic distance still contribute to the GA-BoF, on the contrary small contribution will be made to GA-BoF. When its value is turned larger, the 2D GA-BoF will degrade to BoF because of losing the neighborhood relationship. Fig. 7 shows the classification accuracy under different k_{gd}. From the figure, we can see that when $k_{gd} = 10$, the proposed method can achieve best performance.

In addition, we also check the influence from different generation methods for bag-of-words. While using traditional k-means to generate geometric vocabulary, the average accuracy is 83.0%. By using weighted k-means, the classification accuracies are 83.5%, 84.5%, and 85.0% when β are 2, 3, and 4, respectively. From the results, we can conclude that by using weighted k-means, the performance can slightly improved.

From the results of above experiments, we selected $BoW_n = 100$, $k_{BoF} = 2$, $k_{gd} = 10$ as optimal parameters, and weighted

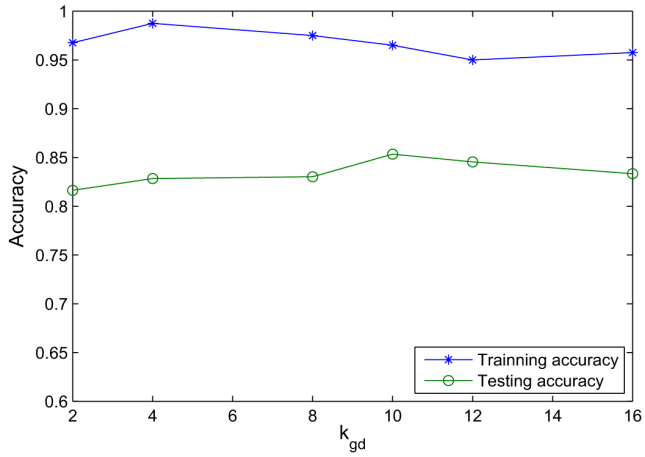

Fig. 7. Averaged classification accuracy on different k_{gd} (SHREC 2007).

k-means with $\beta = 4$ to apply into the proposed method for the following experiments. The DBN is consisted of three hidden layers, and node numbers of hidden layers from bottom to up are empirically set to be 1000, 800, and 400, respectively.

In the following classification experiments, we randomly select 50% shapes in each category as training samples, and left shapes as test data. We first use SHREC 2007 and SHREC 2011 to test the proposed method (3D-DL), where the confusion matrices of classifications on both datasets are visualized in the Fig. 8. In addition, we use support vector machine (SVM) to train and test the shape classification by using the middle-level feature: GA-BoF. The DBNs can be regarded as one type of feature dimension reduction method, in order to demonstrate the performance of DBNs compared with other similar techniques, we also do experiments with Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Linear Discriminant Analysis (LDA), and Locally Linear Embedding (LLE) to perform similar feature dimension reduction of GA-BoF before SVM classification. Furthermore we adopt Latent Semantic Analysis (LSA) to analysis the classification accuracy, different from above mentioned methods the first step is to construct BoW-shape matrix which is similar to term-document matrix in natural language analysis, and then use singular value decomposition (SVD) to reduce the dimension of the feature of shape prior to the SVM training and classification. The comparison results are listed in Table I. From the numerical results, we can draw the conclusion that through the process of DBNs, the characterization capability of high-level feature is much better than other linear or nonlinear feature enhancement methods. Due to the fact that shapes in SHREC 2011 only contain articulated deformation and the shape variance is small, the shape classification accuracy is much higher than that in SHREC 2007.

In the last experiment on classification, we choose the Princeton shape benchmark [51] as the dataset to assess the robustness and multi-level classification of the proposed method. Since some shapes are non-watertight, the geodesic calculation of pairs of vertices is impossible. Although approximate mesh can be made through connecting each vertex's k-nearest neighbor vertices, this approximation may degrade the representation accuracy of original shape. As a consequence, in the low-level feature extraction, only the SI-HKS is adopted, and the Euclidean distance is used as the spatial measure

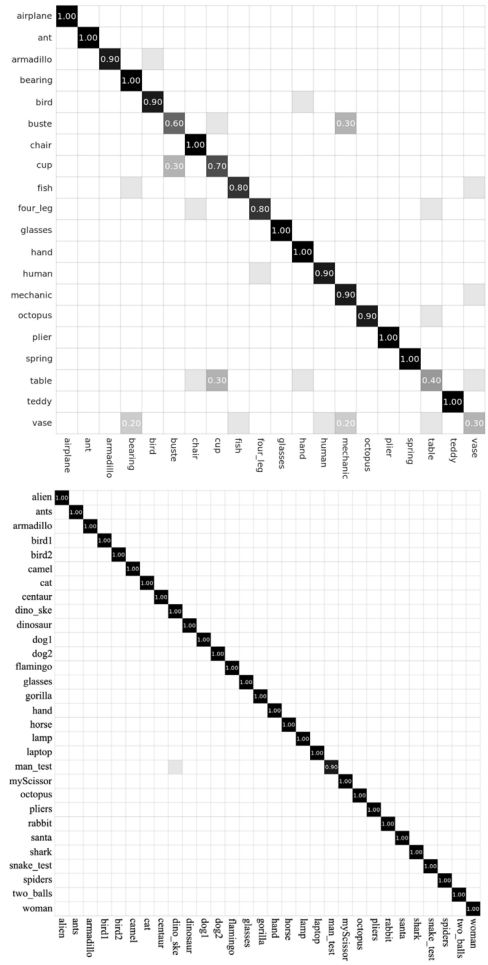

Fig. 8. Confusion matrix calculated by using the proposed method on SHREC 2007 and SHREC 2011 dataset. Top: SHREC 2007. Bottom: SHREC 2011.

TABLE I
RECOGNITION PERFORMANCE (AVERAGE CLASSIFICATION ACCURACY %) BY USING DIFFERENT FEATURES AND FEATURE ENHANCEMENT METHODS ON SHREC 2007, SHREC 2011, AND PSB DATASETS

Dataset	GA-BoF	PCA	MDS	LDA	LLE	LSA	3D-DL
SHREC 2007	68.0	78.5	79.5	71.5	72.5	76.5	**85.0**
SHREC 2011	93.2	98.3	98.3	90.7	97.3	98.3	**99.7**
PSB - base	38.3	43.8	44.5	40.2	42.1	43.5	**51.3**
PSB - L1	42.7	47.1	48.4	43.9	46.5	47.2	**56.6**
PSB - L2	50.7	54.6	55.9	51.7	53.7	56.4	**64.6**
PSB - L3	71.4	76.5	77.1	72.5	73.7	75.3	**85.1**

for GA-BoF computation. Four levels of classifications are conducted, the levels are 'base' (fine grained), 'L1', 'L2', and 'L3' (coarsest). In the 'base' level, same functional objects are divided into several classes based on the geometric differences. In the 'L1' level, shapes are divided into 42 categories including animals, plants, airplanes, furniture, etc. In the 'L2' level, the classification has seven categories: animal, building, furniture, household, plant, vehicle, and a miscellaneous category. For the last level, the classification partitions models into manmade and natural. The classification results are also listed in Table I for better comparison. From the above results, we can conclude that the high-level feature extracted by the deep network is more discriminative, and it can achieve much higher recognition accuracy.

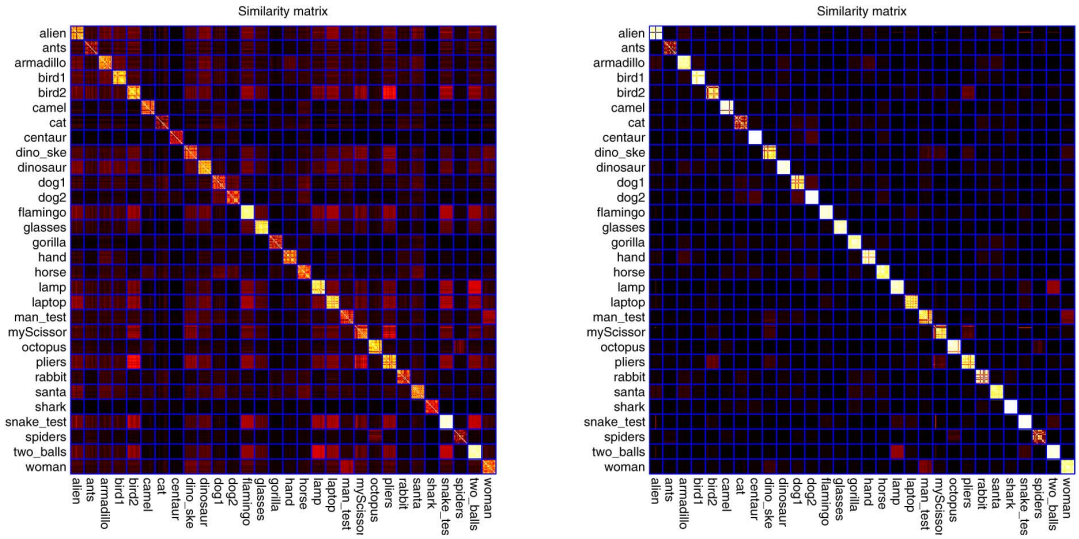

Fig. 9. Similarity matrices computed based on GA-BoF and the proposed method, for all the models from the SHREC 2011 dataset. Brighter color implies greater degree of similarity. Left: Calculated by using GA-BoF features. Right: Calculated by using learned features.

B. 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 [51].

Training Dataset: In order to train a higher discriminative and domain independent DBN, we use the training data not only from a single dataset, but from a mixed dataset which is aggregated from multiple 3D shape datasets including SHREC 2007 [50], McGill 3D Shape Benchmark [56], SHREC 2011 [52], Princeton shape benchmark [51], and TOSCA shapes [57]. The number of shapes is 1400, which are divided into 58 categories. The shapes in our training dataset contain articulated deformation, in addition shapes have sufficient and diverse variations, and consequently it is helpful to boost the generalization of the DBN.

Experiments on SHREC 2011: First, we use the SHREC 2011 dataset to evaluate the retrieval performance of the proposed method. In the experiment, we use the shapes in training dataset to train a DBN, and then use the meshes in SHREC 2011 to generate high-level features for similarity calculation.

The similarity matrices which are computed based on GA-BoF and the proposed method are shown in the Fig. 9. The similarity matrix calculated by using GA-BoF is shown in the left of Fig. 9, and similarity matrix calculated by using deep learning is shown in the right. From the figure, we can see that after the feature learning, the intra-class similarity is increased, and the inter-class similarity is decreased, and consequently, the proposed high-level feature can improve the retrieval performance.

The recall-precision curves of some state-of-the-art methods and the proposed method are plotted in Fig. 10. From the figure, we can see that the proposed method achieves better overall retrieval performance, although the precisions of the proposed method are not good than spectral decomposition of geodesic distance matrix (SD-GDM) [25] when the recall is under 0.4. The numerical evaluation measures are listed in Table II. From the table, we can conclude that although NN, FT, and E measures obtained by the proposed method are not as good as that generated by SD-GDM, ST and DCG are shown to be superior than SD-GDM. The numerical result are consistent with the recall-precision curve in Fig. 10, and we can conclude that the proposed method has better overall retrieval performance. It is need to be note that the ST values in SHREC 2011 are two times

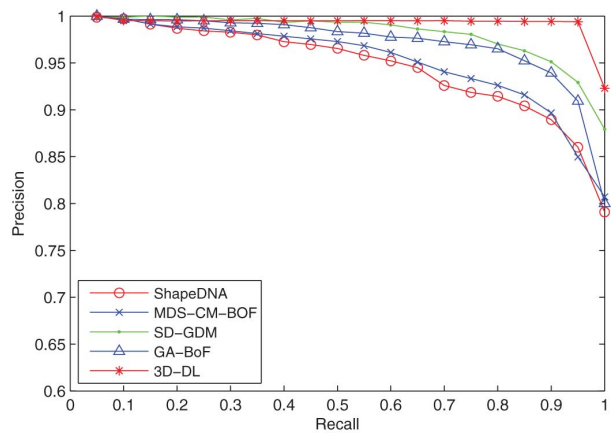

Fig. 10. Recall-precision curve of some state-of-the-art methods and the proposed method on SHREC 2011 dataset.

TABLE II
RETRIEVAL PERFORMANCE ON THE SHREC 2011 DATASET

Methods	NN(%)	FT(%)	ST(%)	E(%)	DCG(%)
FOG [52]	96.8	81.7	90.3	66.0	94.4
BOW-LSD [41]	95.5	67.2	80.3	57.9	89.7
MDS-CM-BOF [53]	99.5	91.3	96.9	71.7	98.2
BOGH [52]	99.3	81.1	88.4	64.7	94.9
LSF [52]	99.5	79.9	86.3	63.3	94.3
ShapeDNA [54]	99.2	91.5	95.7	70.5	97.8
Harris3DGeoMap [52]	56.2	32.5	46.6	32.2	65.4
HKS [52]	83.7	40.6	49.7	35.3	73.0
MeshSIFT [55]	99.5	88.4	96.2	70.8	98.0
SD-GDM [25]	**100.0**	**96.2**	98.4	**73.1**	99.4
GA-BoF	98.6	91.0	97.4 (48.7)	68.3	97.2
3D-DL	99.7	94.1	**99.2 (49.6)**	72.3	**99.6**

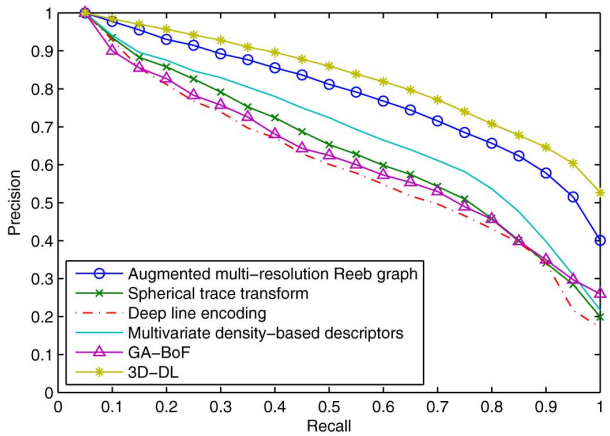

Fig. 11. Recall-precision curve of some previous methods and the proposed method on SHREC 2007 watertight dataset.

TABLE III
THE PRECISION VALUES OF 20, 40, 60, AND 80 RETURN ITEMS ON SHREC 2007 WATERTIGHT DATASET

Methods	20 (%)	40 (%)	60 (%)	80 (%)
DLE [50]	54.6	32.9	24.1	19.0
MDD [50]	62.6	36.6	26.2	20.5
STT [50]	56.4	34.6	25.2	19.9
SI-MSC [50]	60.4	36.6	26.2	20.5
aMRG [50]	71.4	41.4	29.0	22.5
ERG [29]	62.4	41.5	30.5	24.4
GA-BoF	56.7	33.3	24.1	19.0
3D-DL	**72.9**	**42.1**	**31.7**	**25.8**

TABLE IV
THE RECALL VALUES OF 20, 40, 60, AND 80 RETURN ITEMS ON SHREC 2007 WATERTIGHT DATASET

Methods	20 (%)	40 (%)	60 (%)	80 (%)
DLE [50]	54.6	65.8	72.4	76.3
MDD [50]	62.6	73.2	78.6	82.1
STT [50]	56.4	69.2	75.6	79.8
SI-MSC [50]	60.4	73.2	78.8	82.2
aMRG [50]	71.4	82.8	87.2	90.2
ERG [29]	62.4	82.9	91.6	**97.5**
GA-BoF	56.7	66.7	72.2	76.0
3D-DL	**72.9**	**84.2**	**92.6**	94.7

their correct values. In oder to give a clear comparison we list both values, where the correct values are listed in brackets.

Experiments on SHREC 2007 Watertight Dataset: In order to assess the robustness of the proposed method, we also perform an experiment on the SHREC 2007 watertight dataset [50]. The pre-calculated geometric vocabulary and the DBN 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. 11, which include depth line encoding (DLE) [50], multivariate density-based descriptor (MDD) [50], spherical trace transform (STT) [50], silhouette intersection and multi-scale contour (SI-MSC) [50], augmented multi-resolution Reeb graph (aMRG) [50], and extended Reeb graphs (ERG) [29]. Numerical values for the averaged precision and recall on all the models in the dataset are listed in Tables III and IV, respectively. We list these values of returned 20, 40, 60, 80 items, which are 1, 2, 3, and 4 times the size of each class, and these results show that our method is comparable to the representative methods evaluated on the dataset.

TABLE V
THE DCG VALUES OF VARIOUS TRANSFORMATIONS ON SHREC 2007 WATERTIGHT DATASET

	Strength			
Transform	1	2	3	4
None	91.67			
Holes	91.24	91.08	90.85	90.17
Micro holes	91.48	91.32	91.05	90.92
Scale	91.62	91.49	91.27	91.11
Sampling	91.01	90.84	90.39	90.13
Noise	91.64	91.59	91.45	91.30

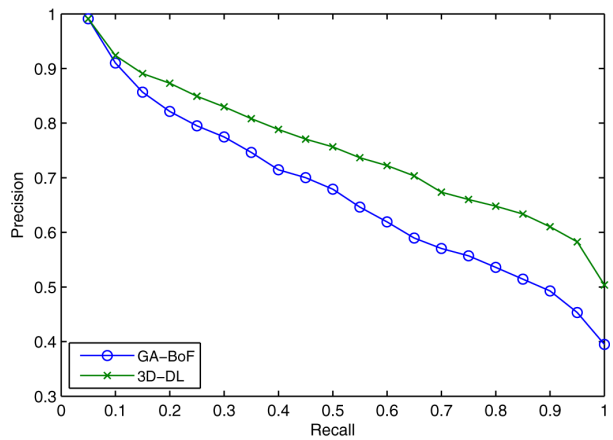

Fig. 12. Recall-precision curve of the proposed method on a mixed dataset.

We also apply various transformations on the dataset to evaluate the robustness of the proposed method. For each transformation, different transformation strengths are adopted to the models from slight (level 1) to strong (level 4). The DCG values computed based on different transformations are listed in Table V. Because we use SI-HKS and AGD which are robust against noise and scale, furthermore, geodesic distance is adopted as distance measure to calculate middle-level feature, therefore, the retrieval performances do not decline significantly.

Experiments on Mixed Dataset: Finally, we use a mixed 3D shape dataset as testing data to evaluate the domain irrelevant property and robustness of the proposed method. The construction of the test dataset is similar to that of training dataset. The shapes are randomly selected from multiple datasets as mentioned above, and the selection ensures that they are not included in the training dataset. There are 620 meshes in the test dataset which are divided into 31 categories, and there are 20 meshes in each category. In this experiment, we also use the shapes in the training dataset to train a DBN, and then use the DBN to generate high-level features for shapes in the mixed dataset.

The recall-precision curves of the proposed method are shown in Fig. 12. By using the DBN to compute high-level features, the precision is greatly improved from the GA-BoF. The numerical evaluation measures are listed in Table VI. From the table, we can conclude all of the measures are improved from using GA-BoF to high-level features. The DCG improvement is 9.4%, which demonstrates that the proposed method has the capability to improve retrieval performance by using learned features.

In addition, we perform an experiment by using variant number of training data and fixed test data. In this experiment,

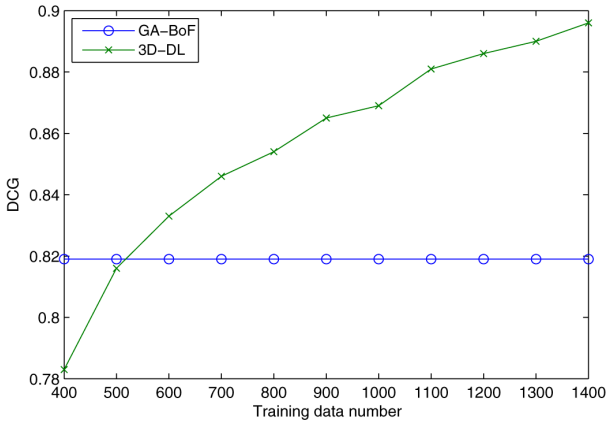

Fig. 13. DCG values of the proposed method under different training data numbers.

TABLE VI
RETRIEVAL PERFORMANCE OF THE PROPOSED METHODS USING FIVE STANDARD MEASURES ON THE MIXED DATASET

Methods	NN(%)	FT(%)	ST(%)	E(%)	DCG(%)
GA-BoF	84.1	59.4	35.6	50.7	81.9
3D-DL	**85.8**	**67.7**	**39.8**	**56.8**	**89.6**
Improvements (%)	2.0	14.0	11.8	12.0	9.4

in order to avoid the influence from varying vocabulary, we use fixed geometric vocabulary to generate GA-BoFs. The results are plotted in Fig. 13 and listed in Table VI. Due to the GA-BoF is independent to the training data, the DCG value by using GA-BoF is a const value of 81.9%. However, through the feature learning, the retrieval performance is improved with the increasing of training data size, and the retrieval performance could be further improved by using more training data.

V. CONCLUSION

In this paper, we present a novel three-level feature extraction framework for recognition and retrieval of 3D shapes. In the first stage, low-level 3D shape descriptors are extracted to represent the intrinsic geometric properties. And then a geometric vocabulary is built to compute local structure insensitive bag-of-features. Finally, DBN is adopted to learn structural high-level features. These high-level features are employed to perform 3D shape recognition and retrieval. The experiments results show that the learned high-level features are more discriminative which can suppress intra-class variation and enhance the inter-class similarity separation. From the retrieval results shown in Fig. 10 and Fig. 12, the overall retrieval performances are boosted by using the learned high-level feature compared to that using the middle-level features.

Different from traditional deep learning methods used in computer vision, we adopt middle-level features as the input for deep learning in this work. The reason lies in that 3D shapes have complex structures, and also lies in that it is hard to seek a beginning position and make a comparable sequence. In addition, 3D shapes are usually known to be poor in features, and thus descriptors are less informative compared to image data. Therefore, directly applying deep learning for high-level feature learning causes low performance. Furthermore, 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 difficult to achieve commendable accuracy. Through the proposed method, only a small amount training samples are required to train the DBN, and adequate performance can be accomplished. As shown in Fig. 5, even 2 training samples for each category are used in training, the classification accuracy of test data is 70% for 20 types of shapes. Therefore, the proposed method has the merit of efficient training. Nevertheless, the retrieval performance can be further improved by utilizing more training data as demonstrated in Fig. 13.

Limitations: However, based on the theory of deep learning, using the middle-level features may cause the lower generalization, because during the middle-level features processing some information is lost, especially the local geometric connections. In addition, in the proposed method, the features for the deep learning input are global features. Although we adopt geodesic-sensitive BoFs to preserve the geometrical relationship of BoFs, the local connection information is still missing. Hence, the proposed method is difficult to apply to more sophisticated tasks such as segmentation, partial retrieval, and symmetric detection.

Future Works: First, at present we only investigate SI-HKS and AGD as the low-level descriptors. Other local descriptors will be studied in the following research. Second, the GA-BoF is adopted as the middle-level features in current work. It is necessary to study other methods which can preserve more structural information for feature learning. Third, it is necessary to research novel deep learning method that can directly processing graph-based data, including 3D mesh data, communication network, and traffic network, which makes it have wider application fields with better performance.

REFERENCES

[1] J. W. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools Appl.*, vol. 39, no. 3, pp. 441–471, 2008.
[2] A. D. Bimbo and P. Pala, "Content-based retrieval of 3d models," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 20–43, 2006.
[3] Z. Liu, S. Bu, K. Zhou, S. Gao, J. Han, and J. Wu, "A survey on partial retrieval of 3D shapes," *J. Comput. Sci. Technol.*, vol. 28, no. 5, pp. 836–851, 2013.
[4] M. Ovsjanikov, A. M. Bronstein, M. M. Bronstein, and L. J. Guibas, "Shape Google: A computer vision approach to isometry invariant shape retrieval," in *Proc. 2009 IEEE 12th Int. Conf. Comput. Vis. Workshops*, 2009, pp. 320–327.
[5] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov, "Shape Google: Geometric words and expressions for invariant shape retrieval," *ACM Trans. Graph.*, vol. 30, no. 1, pp. 1–22, 2011.
[6] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.
[7] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Proc. 2010 IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2010, pp. 1704–1711.
[8] G. Lavoué, "Combination of bag-of-words descriptors for robust partial shape retrieval," *Vis. Comput.*, vol. 28, no. 9, pp. 931–942, 2012.
[9] R. Toldo, U. Castellani, and A. Fusiello, "Visual vocabulary signature for 3D object retrieval and partial matching," in *Proc. 2nd Eurograph. Conf. 3D Object Retrieval*, 2009, pp. 21–28.
[10] U. Castellani, M. Cristani, S. Fantoni, and V. Murino, "Sparse points matching by combining 3D mesh saliency with statistical descriptors," *Comput. Graph. Forum*, vol. 27, no. 2, pp. 643–652, 2008.
[11] S. Bu, P. Han, Z. Liu, K. Li, and J. Han, "Shift-invariant ring feature for 3D shape," *Vis. Comput.*, vol. 30, no. 6, pp. 867–876, 2014.

[12] R. Litman and A. M. Bronstein, "Learning spectral descriptors for deformable shape correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 171–180, Jan. 2014.
[13] V. Barra and S. Biasotti, "Learning kernels on extended Reeb graphs for 3D shape classification and retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, 2013, pp. 25–32.
[14] H. Laga, M. Mortara, and M. Spagnuolo, "Geometry and context for semantic correspondences and functionality recognition in man-made 3D shapes," *ACM Trans. Graph.*, vol. 32, no. 5, p. 150, 2013.
[15] B. Leng and Z. Xiong, "Modelseek: An effective 3D model retrieval system," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 935–962, 2011.
[16] H. Laga, "Semantics-driven approach for automatic selection of best views of 3D shapes," in *Proc. 3rd Eurograph. Conf. 3D Object Retrieval*, 2010, pp. 15–22.
[17] H. Tabia, D. Picard, H. Laga, and P.-H. Gosselin, "Compact vectors of locally aggregated tensors for 3d shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, 2013, pp. 17–24.
[18] Y. Gao, M. Wang, R. Ji, X. Wu, and Q. Dai, "3D object retrieval with Hausdorff distance learning," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2088–2098, Apr. 2014.
[19] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[20] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. ACM Conf. Mach. Learn.* , 2009, pp. 609–616.
[21] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2012, pp. 3642–3649.
[22] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Matching 3D models with shape distributions," in *Proc. SMI 2001 Int. Conf. Shape Model. Appl.*, 2001, pp. 154–166.
[23] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Proc. 2003 Eurograph./ACM SIGGRAPH Symp. Geometry*, 2003, pp. 156–164, Eurographics Association.
[24] V. Jain and H. Zhang, "A spectral approach to shape-based retrieval of articulated 3D models," *Comput.-Aided Des.*, vol. 39, no. 5, pp. 398–407, 2007.
[25] D. Smeets, T. Fabry, J. Hermans, D. Vandermeulen, and P. Suetens, "Isometric deformation modelling for object recognition," in *Computer Analysis of Images and Patterns*. New York, NY, USA: Springer, 2009, pp. 757–765.
[26] N. D. Cornea, M. F. Demirci, D. Silver, S. Dickinson, and P. Kantor, "3d object retrieval using many-to-many matching of curve skeletons," in *Proc. 2005 Int. Conf. Shape Model. Appl.*, 2005, pp. 366–371.
[27] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Combining topological and geometrical features for global and partial 3D shape retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 819–831, Aug. 2008.
[28] J. Tierny, J.-P. Vandeborre, and M. Daoudi, "Partial 3d shape retrieval by reeb pattern unfolding," *Comput. Graph. Forum*, vol. 28, no. 1, pp. 41–55, 2009.
[29] V. Barra and S. Biasotti, "3D shape retrieval using kernels on extended Reeb graphs," *Pattern Recog.*, vol. 46, no. 11, pp. 2985–2999, Nov. 2013.
[30] P. Papadakis, I. Pratikakis, T. Theoharis, and S. Perantonis, "Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval," *Int. J. Comput. Vis.*, vol. 89, no. 2, pp. 177–192, 2010.
[31] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3d model retrieval," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 223–232, 2003.
[32] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, and T.-S. Chua, "Camera constraint-free view-based 3D object retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2269–2281, Apr. 2012.
[33] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Feb. 2012.
[34] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: Efficient 3D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, Oct. 2011.
[35] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, May 1999.
[36] Y. Liu, H. Zha, and H. Qin, "The generalized shape distributions for shape matching and analysis," in *Proc. IEEE Int. Conf. Shape Model. Appl.*, 2006, pp. 16–16.
[37] I. Sipiran, B. Bustos, and T. Schreck, "Data-aware 3D partitioning for generic shape retrieval," *Comput. Graph.*, vol. 37, no. 5, pp. 460–470, 2013.
[38] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool, "Hough transform and 3D surf for robust three dimensional classification," in *Proc. Comput. Vis.–ECCV 2010*, 2010, pp. 589–602.
[39] H.-Y. Wu, H. Zha, T. Luo, X.-L. Wang, and S. Ma, "Global and local isometry-invariant descriptor for 3d shape comparison and partial matching," in *Proc. 2010 IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2010, pp. 438–445.
[40] A. Dubrovina and R. Kimmel, "Matching shapes by eigendecomposition of the Laplace-Beltrami operator," in *Proc. Symp. 3D Data Proc.*, 2010, vol. 2, no. 3.
[41] G. Lavoue, "Bag of words and local spectral descriptor for 3d partial shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, 2011, pp. 41–48.
[42] K. Sfikas, T. Theoharis, and I. Pratikakis, "Non-rigid 3d object retrieval using topological information guided by conformal factors," *Vis. Comput.*, vol. 28, no. 9, pp. 943–955, 2012.
[43] A. Kovnatsky, M. M. Bronstein, A. M. Bronstein, D. Raviv, and R. Kimmel, "Affine-invariant photometric heat kernel signatures," in *Proc. Eurograph. Conf. 3D Object Retrieval*, 2012, pp. 39–46.
[44] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, "Topology matching for fully automatic similarity estimation of 3d shapes," in *Proc. 28th Annu. Conf. Comput. Graph. Interactive Tech.*, 2001, pp. 203–212.
[45] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering," *Pattern Recog.*, vol. 45, no. 3, pp. 1061–1075, 2012.
[46] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[47] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[48] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[49] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 536–543.
[50] D. Giorgi, S. Biasotti, and L. Paraboschi, CNR - IMATI, Genoa, Italy, "Watertight models track," Tech. Rep. 9, 2007.
[51] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton shape benchmark," in *Proc. Shape Model. Appl.*, 2004, pp. 167–178, IEEE.
[52] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H. Van Nguyen, and R. Ohbuchi, "SHREC'11 track: Shape retrieval on non-rigid 3D watertight meshes," *3DOR*, vol. 11, pp. 79–88, 2011.
[53] Z. Lian, A. Godil, X. Sun, and H. Zhang, "Non-rigid 3D shape retrieval using multidimensional scaling and bag-of-features," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, 2010, pp. 3181–3184.
[54] M. Reuter, F.-E. Wolter, and N. Peinecke, "Laplace-spectra as fingerprints for shape matching," in *Proc. 2005 ACM Symp. Solid Physical Model.*, 2005, pp. 101–106.
[55] C. Maes, T. Fabry, J. Keustermans, D. Smeets, P. Suetens, and D. Vandermeulen, "Feature detection on 3D face surfaces for pose normalisation and recognition," in *Proc. 2010 4th IEEE Int. Conf. Biometrics: Theory Appl. Syst. (BTAS)*, 2010, pp. 1–6.
[56] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson, "Retrieving articulated 3D models using medial surfaces," *Mach. Vis. Appl.*, vol. 19, no. 4, pp. 261–275, 2008.
[57] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, *Numerical Geometry of Non-Rigid Shapes*. New York, NY, USA: Springer, 2008.

Shuhui Bu (M'09) received the M.Sc. and Ph.D. degrees from the College of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan, in 2006 and 2009, respectively.

He was an Assistant Professor with Kyoto University, Kyoto, Japan, from 2009 through 2011. He is currently an Associate Professor with Northwestern Polytechnical University, Xi'an, China. He has published approximately 30 papers in major international journals and conferences, including IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, TVC, C&G, ACM MM, ICPR, CGI, SMI, etc. His research interests are concentrated on computer vision and robotics, including 3D shape analysis, image processing, pattern recognition, 3D reconstruction, and related fields.

Zhenbao Liu (M'11) received the Ph.D. degree in computer science from the College of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan, in 2009.

He was a visiting scholar with the GrUVi Lab, Simon Fraser University, Burnaby, Canada, in 2012. He is currently an Associate Professor with Northwestern Polytechnical University, Xi'an, China. His research interests include 3D shape analysis, matching, retrieval, and segmentation.

Junwei Han (M'10) received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2003.

From 2003 to 2010, he was a Research Fellow with Nanyang Technological University, Singapore, The Chinese University of Hong Kong, Shatin, Hong Kong, Dublin City University, Dublin, Ireland, and the University of Dundee, Dundee, U.K. He also previously worked as a Visiting Researcher with Microsoft Research Asia, Beijing, China, and University of Surrey, Surrey, U.K. Since 2010, he has been a Full Professor with Northwestern Polytechnical University, Xi'an, China. He has published over 100 papers on the topics of computer vision, multimedia processing, brain imaging analysis, and remote sensing image processing.

Dr. Han is currently an Associate Editor for *Multidimensional Systems and Signal Processing*.

Jun Wu (M'12) received the BS degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2001, and the M.Sc. and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2004 and in 2008, respectively.

From 2003 to 2004, he was a visiting student at Microsoft Research Asia, Beijing, China. From August 2005 to October 2005, he was a Visiting Scholar with the Department of Computer Science, University of Hamburg, Hamburg, Germany. From 2008 to 2010, he was Research Staff with the Intelligent Systems Lab Amsterdam, University of Amsterdam, Amsterdam, the Netherlands. He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests include machine learning, multimedia analysis, and multimedia information retrieval.

Rongrong Ji (M'11–SM'14) received the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China.

He was a Research Intern at Microsoft Research Asia, Beijing, China, from 2007 to 2008. In 2010, he was a Visiting Student with Peking University, Beijing, China. He was a Postdoc Research Fellow with Columbia University, New York, NY, USA, from 2010 to 2013. He is currently a Professor with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Amoy, China. He is the author of over 50 tier 1 journal and conference papers. His research interests include image and video search, content understanding, mobile visual search and recognition, as well as interactive human-computer interface.

Dr. Ji is the recipient of the Best Paper Award at ACM Multimedia 2011 and the Microsoft Fellowship 2007. He is the guest editor of IEEE MULTIMEDIA MAGAZINE, *Neurocomputing*, *Signal Processing*, *ACM Multimedia Systems*, *Multimedia Tools and Applications*, etc., Special Session Chair of ICMR 2014, VCIP 2013, MMM 2013, PCM 2012, etc., Program Committee Member of over 30 flagship international conferences including CVPR 2013, ICCV 2013, ACM Multimedia 2014 to 2010, etc., and reviewer for over 20 IEEE/ACM Transactions, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.