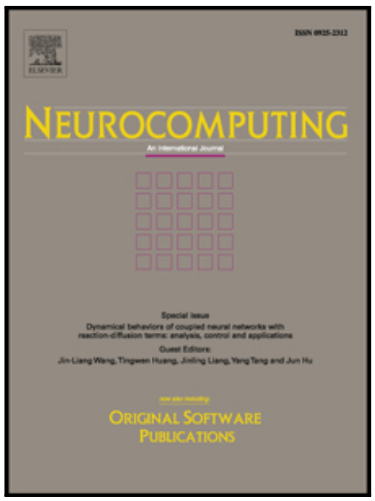

3D Shape Recognition and Retrieval based on Multi-modality Deep Learning

Shuhui Bu, Lei Wang, Pengcheng Han, Zhenbao Liu, Ke Li

PII: S0925-2312(17)30257-6
DOI: 10.1016/j.neucom.2016.06.088
Reference: NEUCOM 18053

To appear in: *Neurocomputing*

Received date: 21 February 2016
Revised date: 21 June 2016
Accepted date: 22 June 2016

Please cite this article as: Shuhui Bu, Lei Wang, Pengcheng Han, Zhenbao Liu, Ke Li, 3D Shape Recognition and Retrieval based on Multi-modality Deep Learning, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.06.088

3D Shape Recognition and Retrieval based on Multi-modality Deep Learning

Shuhui Bu[a], Lei Wang[a], Pengcheng Han[a], Zhenbao Liu[a,*], Ke Li[b]

[a]*Northwestern Polytechnical University, China*
[b]*Information Engineering University, China*

Abstract

For 3D shape analysis, an effective and efficient feature is the key to popularize its applications in 3D domain where the major challenge lies in designing an effective high-level feature. The three-dimensional shape contains various useful information including visual information, geometric relationships, and other type properties. Thus the strategy of exploring these characteristics is the core of extracting effective 3D shape features. In this paper, we propose a novel 3D feature learning framework which combines different modality data effectively to promote the discriminability of uni-modal feature by using deep learning. The geometric information and visual information are extracted by Convolutional Neural Networks (CNNs) and Convolutional Deep Belief Networks (CDBNs), respectively, and then two independent Deep Belief Networks (DBNs) are employed to learn high-level features from geometric and visual features. Finally, a Restricted Boltzmann Machine (RBM) is trained for mining the deep correlations between different modalities. Extensive experiments demonstrate that the proposed framework achieves better performance.

Keywords: 3D Shape, Recognition, Retrieval, Deep Learning, Multi Modality

1. Introduction

With the advent of information epoch, 3D shapes as one type of multimedia data, have been extensively used in the fields of both computer graphics and computer vision applications such as multimedia games, medical diagnosis, industry design, information retrieval, and so forth. All these applications require effective and automatic storage, recognition, and retrieval for 3D models. Thus, it is critical to establish an efficient shape search engine, by which users can obtain 3D models in a convenient way and further explore them. And the core of search engine needs effective retrieval and classification techniques for the management and reusing of 3D shapes. In the last decades, a lot of efforts have been conducted on the analysis and retrieval of texts and images, and consequently great performances have been achieved. However, as the characteristics of 3D shapes are much different with texts and images, these successful recognition and retrieval methods can not be applied to 3D models directly, hence the analysis and understanding of 3D shapes is still a long-standing research topic.

There have been many solutions to 3D shape recognition, matching, and retrieval problems in recent years. Reviewing the implementations of these solutions, we find they are directly related with shape descriptor, which is used to characterize important characteristics to discriminate with other shapes or local regions. The comprehensive reviews can be found in an early work [1] and the latest works [2, 3, 4, 5, 6, 7]. Many early predominant shape features dependent on human designed or hand-crafted, capture some specific information like geometry, topology, and part-level structure from 3D models.

3D shape is composed of complex topological structure and visibly variational geometry; consequently, only limited information can be extracted with hand-crafted feature methods. To further improve the performance of 3D shape descriptor, an alternative approach is to learn hidden states from complex 3D data. The recent great success of automatic feature learning methods has aroused intensive interests in computer vision and machine learning fields. These approaches can learn features automatically from training data, which not only reduces workload but also extracts more efficient de-

*Corresponding author
Email addresses: bushuhui@nwpu.edu.cn (Shuhui Bu), wanglei_nwpu@mail.nwpu.edu.cn (Lei Wang), 18709221546@163.com (Pengcheng Han), liuzhenbao@nwpu.edu.cn (Zhenbao Liu), like19771223@163.com (Ke Li)

scriptors compared with the way of designing features according to human prior knowledge. Especially, the fast development of deep learning techniques [8, 9, 10] improves the ability of feature representation, which has boosted performances in recognition tasks.

It seems difficult to directly adopt deep learning techniques for 3D shape descriptors extraction, since 3D models are usually represented as 2D manifolds that are different from the representation of 2D images. And there is no a standard procedure for encoding 3D geometry models. To address the issue, the most common idea is converting 3D shapes into image representations and then using deep learning techniques to deal with these images, which can be found in recent works. For instance, Xie et al. [11] adopt the multi-view depth image representation and propose multi-view deep extreme learning machine (MVD-ELM) to achieve fast and quality projective feature learning for 3D shapes. Zhu et al. [12] also project 3D shapes into 2D space and use autoencoder for feature learning on 2D images. High accuracy 3D shape recognition performance is obtained by both these methods.

However, only analyzing 3D data from view-based aspect is still not enough for 3D shape understanding, since when converting 3D shapes into 2D images, the 3D spatial geometry information is inevitably lost. In the real world, humans comprehend objects through all kinds of information which is various but impossibly independent from each other. For instance, videos often include visual and audio signals, images related to the title and labels. So for 3D shape, it includes multi-view images captured from various angles and shape intrinsic properties. Due to these characteristics describing the same object, they have some highly non-linear relationships. However, these characteristics from different modalities have varied kinds of representations and structures. The features of 3D objects often include the information of geometric structure and topological relationship. Because of this, it is a challenge to mine the hidden non-linear relationships between different modalities features.

In this paper, we provide a solution having comprehensive consideration about both extrinsic properties and intrinsic features of 3D shapes. We propose a novel scheme to fuse different modality data of 3D shapes into a deep learning framework. The core idea is operating deep learning techniques to combine advantages of geometry-based algorithm making use of the complex topological relationships and geometric properties of 3D model itself, and that of visual-based feature method extracting visual characteristics of 3D model from different viewing images. In brief, Convolutional Deep Belief Networks (CDBNs) and Convolutional Neural Networks (CNNs) are adopted to learn 3D shapes from geometry-based modality and view-based modality, respectively. Next these two modalities are fused with a Restricted Boltzmann Machine (RBM) to obtain more discriminative features. The scheme consists of following three major parts:

1. **View-based feature learning**: First, each 3D shape is represented by a set of 2D images from different views. Next, since CNNs have outstanding ability to extract visual-features in computer vision community, all these projections are used to train the CNNs for acquiring the visual representations of 3D shapes.
2. **Geometry-based feature learning**: Because the convolution operation has advantages of invariant to rotation and translation, in addition, integrating it into neural networks achieves weight sharing which boosts the training due to the reduction of the parameter number, CDBNs are adopted to learn the geometric features. 3D shapes are first transformed to volumetric representation which is easily input into CDBN model, and then with it to learn the geometric representations of 3D shapes.
3. **Modality feature fusion**: Above mentioned two types of features represent different aspect information from 3D shapes. We use DBNs to further explore their high-level representations, which are called high-level visual descriptor (HVD) and high-level geometric descriptor (HGD). Then a RBM is employed to associate the two modality high-level features, which mines their non-linear information and generates stronger representative feature which called as 3D multi-modality feature (3D MMF).

This framework has three advantages as mentioned below. First of all, different modalities are fused to comprehensively understand 3D shape. Moreover, using different deep learning techniques in different feature extraction procedures makes full advantages of various deep learning methods extracting distinct properties from 3D models. Thirdly, unlike other machine learning methods which need to tune parameter manually for obtaining the best performance, there are no parameters to be tuned in the whole learning procedure. The proposed scheme is learning itself automatically.

Several experiments are conducted in 3D shape recognition and retrieval tasks. Results and comparisons with related descriptors indicate that the proposed framework reaches promising performance.

2. Related Work

View-based descriptors. These type features rely on a collection of 2D projections from different views to describe the shape of 3D objects, and they are efficiently robust against 3D shape representation artifacts like holes and noise.

An early work researched by Murase et al. [13] recognizes 3D objects with compact representations obtained by automatically varying pose and illumination. Another particular work is light field descriptor (LFD) [14], which extracts a set of geometric and Fourier descriptors from object silhouettes rendered from several different viewpoints. Experiments indicate that LFD is invariant to translation, scale, and rotation, in addition it is robust against noise or degeneracy. Gao et al. [15] 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. Laga [16] proposes a framework to automatically select the best views of 3D models by learning sets of 2D views that not only maximize the similarity between shapes of the same class, but also make the views discriminate shapes in different classes. Chen et al. [14] utilize a visual similarity-based 3D model retrieval system with the faith that if two 3D models are similar, they also look similar from all viewing angles to complete retrieval task. Bonaventura et al. [17] present an information-theoretic framework to compute the shape similarity between 3D polygonal models. To deal with the problem of low compactness and discrimination power of view-based descriptors, Tabia et al. [18] adopt vectors of locally aggregated tensors to generate descriptors, and then use principal component analysis to reduce the dimension of the descriptors. 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 [19]. In order to avoid the inefficiency from a large number of view comparisons, Gao et al. [20] adopt only a small set of query views to obtain less computational cost during the comparison with target shapes.

We can find that view-based descriptors not only benefit from existing image processing technologies that have achieved great performances, but also require no explicit virtual model information, which contributes to the convenience of extraction and robustness of features.

Geometry-based descriptors. These type of features directly work on the native 3D representations, such as complex polygon meshes, voxel-based discretizations, point clouds, or implicit surfaces.

An earlier and representative work is spin images [21]. Darom et al. [22] extend the spin images to possess the capability of scale-invariant and interest point detection. Sipiran et al. [23] adopt 3D Harris detector to locate interesting points for 3D shape retrieval, which can be seen as an extension from 2D Harris detector measuring the variation in the gradient of a given function (e.g., the intensity function of a image). 3D SURF and SIFT descriptors extracted from 3D voxel grids [24, 25] are proposed for classifying and retrieving similar shapes. Laplace Beltrami operator, which is a generalization of the Laplacian from flat space to manifold, is appealing for 3D shape retrieval because of sparse, symmetric, and intrinsic properties of its robustness to rigid transformation and deformation. Retrieval methods [26, 27, 28, 29] extract main eigenvalues and eigenvectors of Laplace matrix generated on local regions to match different regions of 3D shapes. Laplace-Beltrami operator also provides an efficient way of computing a conformal map from a manifold mesh to a homeomorphous surface with constant Gaussian curvature. The histogram of conformal factors [30] serves as a robust pose-invariant signature of 3D shape, which is regarded as an attribute of a graph node to identify segmented parts in bipartite graph matching for 3D shape retrieval [31]. In a recent work [32], 3D shape is also partitioned into several connected iso-surfaces (annuluses) of conformal factors, and expressed with a graph where node substitutes each annulus.

Heat kernel signature [33], a local descriptor designed on polygon meshes, provides rich local geometric information which makes the signature invariant to isometric deformation and has multi-scale characteristics, thereby achieving better performance in 3D shape retrieval and matching [34, 35, 36, 37]. In order to overcome the influence of diffusion time change under different shape scales [34], Fourier transform is imposed on heat kernel signature at each given vertex to obtain scale invariant. Another work uses intrinsic shape context (ISC) [38] to characterize the local shape property. In the method, the shape context is processed in an intrinsic local polar coordinate system, therefore it is intrinsic and invariant to isometric deformation. Furthermore, Fourier transform is applied to the original shape content data to deal with orientation ambiguity.

Deep Learning descriptors. Above mentioned descriptors are largely "hand-designed" according to the prior knowledge about the geometric property of the shape surfaces or volumes, and some do not generalize well across various domains. Therefore, to achieve adaptive 3D feature generation feature learning based

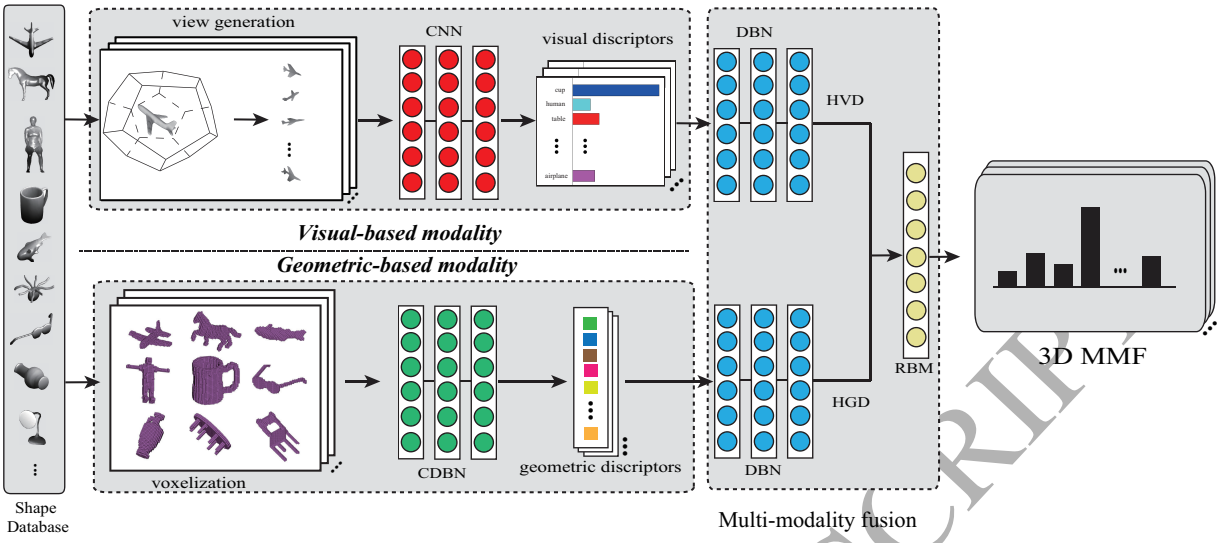

Figure 1: The flowchart of the proposed method. (Only off-line training process is illustrated.)

methods attract attention of many researchers in recent years. Deep learning techniques as powerful feature extraction tools even more become hot spots.

Wu et al. [39] propose the expectation work that learns 3D model descriptors from the voxel-based formation of an object using 3D CDBNs, which obtain good results of shape classification on Princeton ModelNet. Bu et al. [40] propose shift-invariant ring feature (SI-RF) based on iso-geodesic rings and shift-invariant sparse coding for 3D shape analysis. It represents the local region of a feature point efficiently and has great performance on correspondence and retrieval tasks. Xie et al. [11] adopt the multi-view depth image representation and propose multi-view deep extreme learning machine (MVD-ELM) to achieve fast and quality projective feature learning for 3D shapes. Su et al. [41] propose multi-view CNN for 3D shape recognition where the multi-view features are integrated with an extra CNN. Zhu et al. [12] project 3D shapes into 2D space and use autoencoder for feature learning on 2D images. High accuracy 3D shape retrieval performance is obtained by aggregating the features learned on 2D images. Zhao et al. [42] propose Retinex-based Importance Feature (RIF) and Relative Normal Distance (RND) for 3D free form shapes based on the human visual perception characteristics and surface geometry respectively. Chen et al. [43] propose the multi-modal support vector machine to combine three modalities of image feature,i.e., Sift descriptor, Outline Fourier transform descriptor, and Zernike Moments descriptor to discriminate the multiple classes of object. Leng et al. [44] propose a 3D model based on Deep Boltzman Machines (DBM) and semi-supervised learning method to recognize 3D shape.

Though the above mentioned methods have achieved tremendous advancements on classification, matching, and retrieval, it is still far from satisfactory in order to apply 3D objects in more realms. The main issue lies in the fact that geometry-based methods and view-based methods only use partial information of 3D object. In detail, geometry-based methods utilize the complex topological structure and geometric properties of 3D model itself but ignore the visual similarities between 3D objects. Conversely, view-based methods only consider the visual characteristics of model from different viewing images. In order to overcome the shortages, we try to use deep learning techniques to learn and fuse distinct modalities from geometry and view based aspects. The main contributions of this work can be concluded as two aspects:

- **Multi-modal Fusion**: To further improve the performance, multi-modal fusion is adopted to learn 3D shape intrinsic non-linear relationships. Through fusing multi-modal descriptors, complementary visual and geometric information can be encapsulated to increase the accuracies of classification and retrieval.
- **Deep Features**: CNN and CDBN are used to extract visual and geometric feature of 3D shape. CNN has a strong capability to extract the visual feature, while CDBN has the ability to generate high representative feature from 3D object.

Our framework takes full advantage of CNN and CDBN, therefore, more comprehensive descriptions can be extracted.

3. 3D Multimodality Feature

Geometric and visual information are two significant aspects of 3D shape researches. In our framework, we extract these two type descriptors separately, and then fuse them to generate the 3D MMF, which is high discriminative and effective. The flowchart of the proposed method is depicted in Fig. 1 which indicates that the architecture of suggested multi-modal feature fusion contains two modality inputs: geometric descriptors and visual descriptors. Traditional geometric feature are designed using complex 3D shape structure coping with great abundance points. Taking down sampling in preprocess is an effective way to decrease the computational time of generating various features. In our framework, the pretreatments of CNN and CDBN model are voxelization and depth images generating without down sampling method. In the geometric feature extraction, 3D shapes are converted from mesh form into the voxel representation which is close to the original 3D object, therefore we do not need down sampling. In the visual feature extraction, we take the depth image as the input which also does not require down sampling due to the conversion of 3D shape to multiple images. The details of every extracted step are found as follows.

3.1. Geometric descriptors Extraction by CDBN

Traditional geometric descriptors are designed using complex 3D shape structure with human prior knowledge, which increases the workload and decreases the efficiency for designers coping with large amounts of 3D shapes used in various applications. CDBN is a powerful tool to automatically learn highly discriminative features because of its unsupervised and deep learning networks. 3D shape is composed of complex topological structure and variational geometry, CDBN seems difficult to be used in 3D shape analysis directly. Therefore, we first discretize the 3D shape into regularized grid and regard the voxelization as input of 3D CDBN to extract the geometric descriptor.

Voxelization. Voxelization is that we transform the 3D shape mesh form into the voxel representation which is close to the original 3D object. It not only contains information about the surface of the model, but also describes the internal properties of the model. This kind of representation is one type spatial relation reserving certain significant geometrical information, which discretizes the 3D model and reduces the original complex 3D structure for easily applying 3D CDBN to extract intrinsic 3D geometric features. We use 3D matrix to represent the geometric aspect of 3D shape with probability distribution of binary variables. In the 3D matrix, if one voxel is inside the 3D mesh, the corresponding matrix item, which represent the probability of shape distribution, is set to 1; otherwise the probability value is set to 0. Then we take the 3D matrix as the input of CDBN to extract geometric descriptors. In this work, we extend the CDBN implementation to support 3D data.

Geometric descriptors. For 2D images, DBN [9] is a powerful probabilistic models used to model the joint probabilistic distribution over pixels and labels. However it is a challenge to adapt the model from 2D pixel data to 3D voxel data. A 3D voxel volume with reasonable resolution would have the bigger data than an image with ordinary size and there are a huge number of parameters in a fully connected DBN, which make the model hard to be trained effectively. So we use convolution to reduce model parameters by weight sharing. Compared to the traditional convolutional deep learning, we ignore the pooling layers which may bring about greater uncertainty for feature generation.

The energy of a convolutional layer in our model is defined as

$$E(\mathbf{v},\mathbf{h}) = -\sum_{f}\sum_{j}\left(\mathbf{h}_j^f\left(\mathbf{W}^f * \mathbf{v}\right)_j + \mathbf{c}^f\mathbf{h}_j^f\right) - \sum_{l}\mathbf{b}_l\mathbf{v}_l, \tag{1}$$

where f denotes feature channel, j denotes the index of hidden units, and l indicates the index of visible units. The sign "*" represents the convolution operation. In the function, $\mathbf{h}_j^f$ denotes each hidden unit in feature channel f, $\mathbf{v}_l$ represents the visible unit which is the 3D voxel input, and $\mathbf{W}^f$ is the convolution filter. $\mathbf{c}^f$ and $\mathbf{b}_l$ are bias terms of hidden unit $\mathbf{h}_j^f$ and visible unit $\mathbf{v}_l$, respectively. Similar to [45], we also allow for a convolution stride.

We set a 3D shape as a 30×30×30 voxel grid with 3 extra cells of padding in both directions to reduce the convolution border artifacts. We put forward to see the labels as standard one of K softmax variables. The final architecture of our model is illustrated in Fig. 2. The first layer has 32 filters of size 8 and stride 2; the second layer has 160 filters of size 5 and stride 2; the third layer has 512 filters of size 4; each convolution filter is connected to all the feature channels in the previous layer; the fourth layer is a standard fully connected RBM with 2000 hidden units; and the fifth and final layer with 1000 hidden units takes as input a combination of multinomial label variables and Bernoulli feature variables.

The 3D CDBN model is trained in two steps includ-

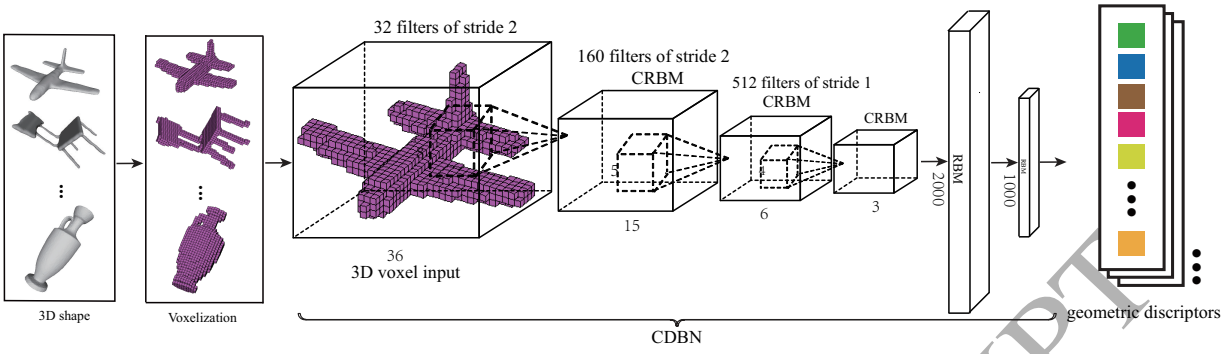

Figure 2: Architecture of our 3D shape CDBN model. For illustration purpose, we only draw one filter for each convolutional layer.

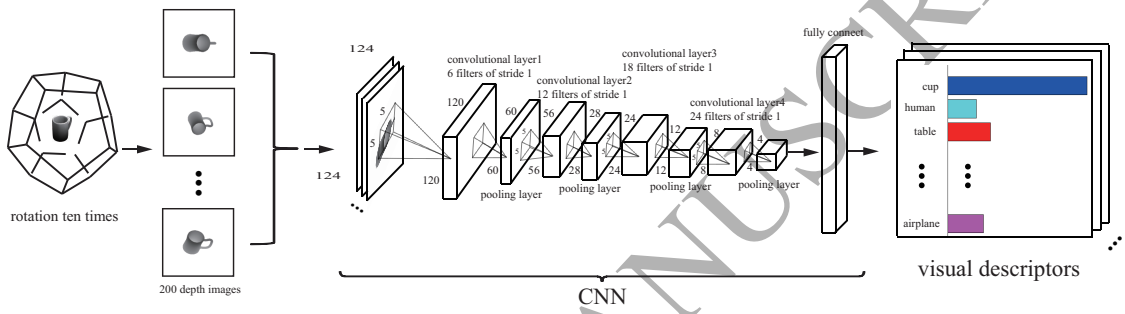

Figure 3: Architecture of our 3D shape CNN model. For illustration purpose, we only draw one filter for each convolutional layer.

ing layer-wise pre-training and generative fine-tuning procedures. During pre-training, the first four layers are trained separately with standard Contrastive Divergence [46] algorithm, and the top layer is trained using Fast Persistent Contrastive Divergence (FPCD) [47]. Once the lower layer is learned, the weights are fixed and the hidden activations are fed into the next layer as input. In our fine-tuning procedure, we adopt a method similar to wake sleep algorithm [9]. In the wake phase, we propagate the data bottom-up and use the activations to collect the positive learning signal. In the sleep phase, we maintain a persistent chain on the topmost layer and propagate the data top-down to collect the negative learning signal. This fine-tuning procedure mimics the recognition and generation behavior of the model and works well in practice. Once the weights of whole networks have been learned, we use forward computation to generate the geometric descriptor $o(\mathbf{X}_{shape})$ using input data of voxelization.

3.2. *Visual descriptor Extraction by CNN*

The popular way of analyzing 3D shapes from view perspective is to convert the 3D model into 2D images from various angles. In theory, these 2D images should contain information from 3D model as much as possible. In our visual descriptor generation procedure, we first project 3D shape into 2D images from 20 directions and adopt CNN to further extract visual features. The details of our algorithm are summarized as follows.

3D Model Pretreatment. In this part, we set the origin point on the center of 3D model mass and then measure the maximum polar distance of the points to one on its surface. Rotation normalization is not performed, but this will be compensated to some extent as described in the following.

Depth Images Collection. Depth images, one type of 2D images, are rendered from 20 vertices of a regular dodecahedron whose mass center is also located in the origin. In the proposed method, we rotate the regular dodecahedron 10 times to make the feature robust against rotation. The rotation angle should be set carefully to ensure that all the cameras are distributed uniformly and able to cover different viewing angles for a 3D model. We consider that dodecahedron has 20 vertices which can generate a moderate data size leading to high computational performance and significant information. The strategy is similar with LFD in view extraction but slightly different with it, we discard the binary images and only use the 2D depth images. Fi-

nally a 3D object is represented by 200 images, each of which has the size of 256×256.

In the depth image rendering, effective information concentrates on the center of the image. Therefore, we remove the borders of depth images and crop the images to the sizes of 124×124 from 256×256 intending to filter out interfering and redundant information, which makes the data compact. In addition, this processing can boost the following CNN feature learning due to the image size is smaller than the original depth images. Because the effective input range of CNN model is from 0 to 1, depth map is not suitable as the input for CNN model. Therefore we normalize the range of each dimension to [0, 1].

Visual descriptors. CNN, a powerful deep learning technique, has achieved great performance of extracting image features in computer vision community. From above procedure, we obtain 2D images containing rich visual information about 3D model. Therefore, CNN is used to extract visual features for each image of 3D shape. As shown in the Fig. 3, the CNN consisting of 4 convolutional layers followed by one fully connected layer and a softmax classification layer, is used to extract features of 2D images. For each layer l, we have:

$$\mathbf{F}_l = pool(sigmoid(\mathbf{W}_l * \mathbf{F}_{l-1} + \mathbf{b}_l)), \tag{2}$$

where $l \in \{1, .., 4\}$, $\mathbf{b}_l$ is the bias parameter of the l-th layer, $\mathbf{W}_l$ is the convolutional kernel. The initial feature map is the 2D images $\mathbf{F}_0$. The sigmoid function is threshold function which is the non-linear symmetric squashing units. The pool operation is a function considering a neighborhood of activations and generating one activation in every neighborhood. Max-pooling operator is regarded as the pool function, which gets the maximum activation in the neighborhood and brings the built-in invariance to translations. The network consists of four convolutional layers. The numbers of filter are set to 6, 12, 18, 24 from the *1st* to the last convolutional layer, and filter size and pooling size of all layers are set to the same values 5 and 2, respectively. In this framework, we use back-propagation method [48] to learn the weights of whole network with input depth images from 3D shape and corresponding label. After CNN model trained completely, for each input depth image, we generate corresponding CNN feature $o(\mathbf{X}_{2D})$ using forward formula of CNN.

Due to a 3D shape surrounded by a dodecahedron which is rotated ten times, 200 depth images are generated to represent individual 3D shape. In other words, visual descriptors $o(\mathbf{X}_{view})$ which are seen as view-based features consist of 200 CNN features $o(\mathbf{X}_{2D})$. If there are K categories in the database, the CNN feature $o(\mathbf{X}_{2D})$ is $1 \times K$ array. So we concatenate 200 $o(\mathbf{X}_{2D})$ into one vector called as visual descriptors $o(\mathbf{X}_{view})$. The visual descriptors can be described as

$$o(\mathbf{X}_{view}) = [o(\mathbf{X}^1_{2D}), o(\mathbf{X}^2_{2D}), ..., o(\mathbf{X}^j_{2D}), ..., o(\mathbf{X}^{200}_{2D})], \tag{3}$$

where $o(\mathbf{X}_{view})$ represents each 3D shape visual descriptor, $o(\mathbf{X}^j_{2D})$ denotes each CNN feature in the shapes, and $j \in [1, 200]$. The visual descriptor for one 3D model is a vector with the size of $200 \times K$. Because visual descriptors contain visual information of 3D shape from all necessary angles, they are better than $o(\mathbf{X}_{2D})$ to represent 3D shape.

3.3. Multi-modal Feature Fusion

Geometric descriptors and visual descriptors stand for spatial characteristics and visual properties of 3D shape, respectively. Therefore, the 3D shape information of two descriptors are complementary. The direct way is to build a RBM over the concatenated geometry-based and view-based feature. While the joint model trained in this way is limited as a shallow model, as a consequence, it is too hard to represent the highly non-linear correlations and extremely different statistical properties between both modalities. In our work, to associate geometry-based and view-based data comprehensively, we first extract high-level descriptors from both geometric descriptors and visual descriptors. By this means, information from specific modality is weakened and more information in high-level features reflects the attributes of 3D models. In another word, high-level features remove the modality-specific information and only reserve the attributes of 3D models.

High-level Descriptors. It is well known that DBNs can extract the deep structural information from features or raw data, which boosts the discrimination ability of generated high-level features. We use DBNs to further explore the intrinsic visual feature distribution of view images for view-based modality features and geometric non-linear relations between voxels for geometry-based modality features, respectively. In another word, High-level Geometric Descriptors (HGD) extracted by DBNs from geometric descriptors and High-level Visual Descriptors (DVD) extracted by DBNs from visual descriptors remove the modality-specific information and only reserve the attributes of 3D models.

The architecture of using DBNs is illustrated in the right part of Fig. 1. Stacking a number of the RBMs and learning layer by layer from bottom to top gives rise to a single DBN. It has been shown that the layer-by-layer greedy learning strategy [9] is effective, and the

greedy procedure achieves approximate maximum likelihood learning. In our work, for each DBNs the bottom layer RBM is trained with the input data $o(\mathbf{X}_{shape})$ or $o(\mathbf{X}_{view})$, 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. The newly inputted geometric descriptors or visual descriptors are processed layer by layer till the final layer after obtaining the optimal parameters for each DBN. And the last layers output $h(\mathbf{X}_{shape})$ and $h(\mathbf{X}_{view})$ are seen as the high-level geometric descriptors and high-level visual descriptors. In order to make the paper more self-contained, we succinctly discuss the concept of restricted Boltzmann machines. The RBM is a two layer, bipartite, undirected graphical model with a set of binary hidden unit $\mathbf{h}$, a set of (binary or real-valued) visible units $\mathbf{v}$, and symmetric connections between these two layers represented by a weighted matrix W. 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

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}, \quad (4)$$

where $Z = \sum_v \sum_h exp(-E(\mathbf{v}, \mathbf{h}; \theta))$ is a normalization factor or partition function and the marginal probability that the model assigns to a visible vector $\mathbf{v}$ is

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

For a Bernoulli (visible)-Bernoulli (hidden) RBM, the energy is

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j, \quad (6)$$

where w_{ij} represents the symmetric interaction between visible unit v_i and hidden unit h_j, b_i and a_j the biases, and V and H are the numbers of visible and hidden units. The conditional probabilities can be efficiently calculated as

$$p(h_j = 1|\mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^{V} w_{ij} v_i + a_j\right), \quad (7)$$

$$p(v_i = 1|\mathbf{h}; \theta) = \sigma\left(\sum_{j=1}^{H} w_{ij} h_j + b_i\right), \quad (8)$$

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

Table 1: Time usage statistics of our framework. The dataset is SHREC 2007, and the time unit is minute.

Different modality	Procedures	Time (minute)
Visual	Depth images	~ 3.2
	CNN	~ 61.3
Geometry	Voxelization	~ 2.6
	CDBN	~ 140.5
Multi-modality	HVD	~ 13.4
	HGD	~ 15.4
	3D MMF	~ 3.1

In principle, the RBM parameters can be optimized by performing stochastic gradient ascent on the log-likelihood of training data. Unfortunately, computing the extract gradient of the log-likelihood is intractable. Instead, the CD approximation [46] is typically used, which has been shown to work well in practice.

3D Multi-modality Feature. A RBM after the DBNs, is employed to associate both modalities to learn the 3D MMF $h(\mathbf{X}_{joint})$ for 3D model. As the Fig. 1 described, the 3D MMF combine high-level geometric descriptors and high-level visual descriptors. Because the 3D MMF $h(\mathbf{X}_{joint})$ come from both $h(\mathbf{X}_{shape})$ and $h(\mathbf{X}_{view})$ by using RBM, they contain spatial properties of 3D model itself and visual similarities of 3D shape. So $h(\mathbf{X}_{joint})$ are more discriminative and robust.

For the recognition tasks, softmax regression is used on the learned 3D MMFs to perform one-vs-all classification. For the retrieval tasks, L_2 distance of the 3D MMF is utilized to measure the similarity of two shapes $\mathbf{X}$ and $\mathbf{Y}$ as

$$d_s(\mathbf{X}, \mathbf{Y}) = \|h(\mathbf{X}_{joint}) - h(\mathbf{Y}_{joint})\|_2. \quad (9)$$

4. Experiments

We use three standard 3D shape benchmarks including SHREC 2007 watertight models [49], SHREC 2011 non-rigid 3D watertight dataset [50], McGill shape benchmark [51] to assess the proposed methods performances on classification and retrieval tasks.

The SHREC 2007 watertight dataset is made up of 400 watertight mesh models, divided into 20 classes, each of which contains 20 objects with different geometrical variations and also articulated deformations. The dataset contains not only natural objects but also man-made objects. SHREC 2011 non-rigid dataset consists of 600 watertight triangle meshes that are transformed from 30 original models. McGill shape benchmark contains 457 models including shapes with articulating parts and without articulation. The set of artic-

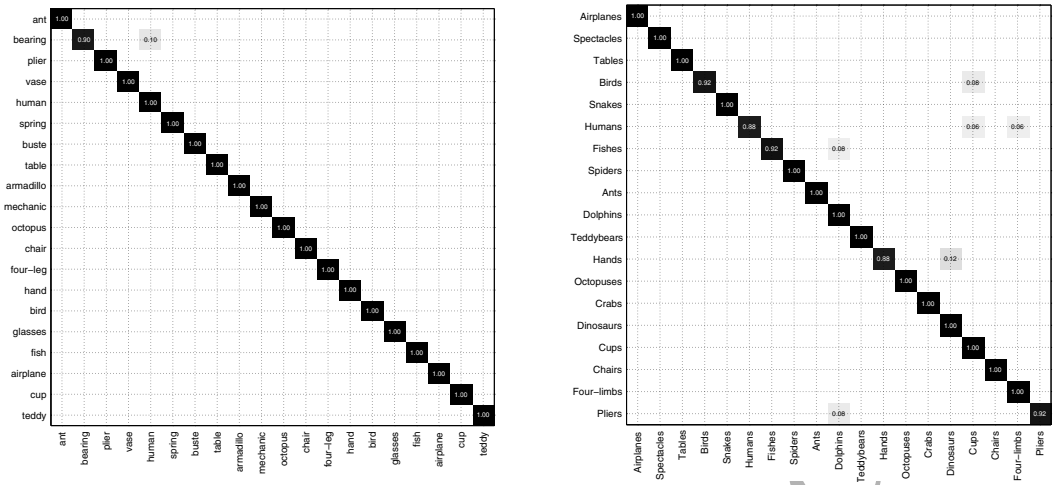

Figure 4: Confusion matrices calculated by using the proposed method on SHREC 2007 (left) and McGill (right).

ulated shapes consists of 255 models in 10 categories, and there are 20∼30 models per category.

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. At the same time, we use GPU to speed up the part deep model training in the whole framework. To speed up the multi-modal fusion, we implemented a deep learning toolbox [1], in which all matrix operations were carried out on the GPU using the Cudamat library.

In the proposed method, various deep learning methods are adopted to learn high-level features. For the CNN, each layer has different kernel and stride sizes, hence the analytical complexity is difficult to be summarized. In addition, the CNN is trained with stochastic gradient descent method, therefore, the required time is related to the epoch number. The complexity analyses of CDBN and DBN have the same problems. In order to show the computational effectiveness, Table 1 lists required computation time for each step. Generally, the off-line training requires hours to train the deep model. When classifying a given 3D model, the required computation time is less than 0.5 second.

4.1. Network Designing

Review the whole framework, the network architecture is significant to achieve good performance.

[1]The source code of our deep learning toolbox is available at https://github.com/shaoguangcheng/DeepNet

Table 2: The average classification results of proposed method. The unit in this table is percentage.

Feature	SHREC 2007	SHREC 2011	McGill
Geometric descriptors	82.00	70.00	81.69
Visual descriptors	89.22	73.75	85.22
3D MMF	**99.50**	**95.40**	**97.47**

First, in the step of learning visual descriptors, convolutional layer number in CNNs affects the recognition accuracy and computation speed. Higher classification accuracy can be obtained with more number of layers, but fast speed with less ones. In our work, with number of layers increasing the computing speed will significantly decline causing low computation performance and the accuracy of the classification is no longer obviously increasing. In order to achieve good performance on both the computing speed and the classification precision, we choose 4 layers as an appropriate layer number for CNNs.

Second, during geometric descriptors learning, the grid size is also critical to performance. Generally, if the grid size is larger, the classification accuracy is higher, nevertheless the computing speed is lower. For achieving balanced performance, we choose 36×36×36 as a moderate grid size.

Third, in the step of learning the 3D MMF, we construct four layers for each DBNs including input and output layers. Since geometric descriptors and visual descriptors are two modal features, different network configurations are set for each DBNs. For high-level

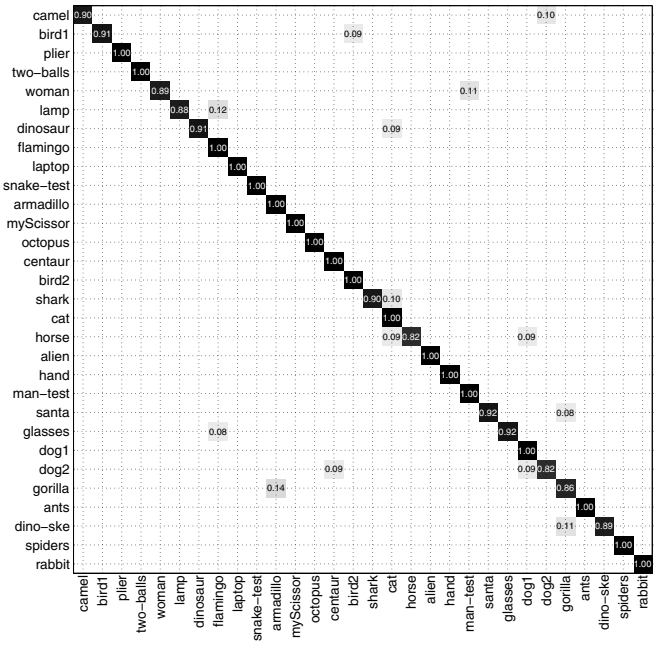

Figure 5: Confusion matrix calculated by using the proposed method on SHREC 2011.

visual descriptors, the number of nodes in each hidden layer is set to 3000 and 1000, the node number of output layer is set to 500. For high-level geometric descriptors, the corresponding node numbers are set to 5000, 2000, 500, respectively. To extract 3D MMF with the last RBM model after two modal DBNs, the number of hidden notes is set as 4800.

4.2. Experiments on Classification

Shape classification experiment is tested for evaluating whether the feature is qualified to correctly classify set of shapes. The average classification accuracy is taken as the evaluation metric for the following experiments. For each dataset of the three shape benchmarks, we randomly select 50% models in each category as training samples, and rest models as test data.

We conduct classification experiments on SHREC 2007, SHREC 2011, and McGill datasets with three type features including visual descriptors, geometric descriptors, and the 3D MMF, respectively. The average classification accuracies of each type feature are observed in the Table 2. From the Table 2, we can clearly conclude that the 3D MMF achieves much better classification performance in comparison with the results obtained from only using single modality feature. This can be explained by the fact that the geometry and view-based modalities only reflect partial properties of 3D

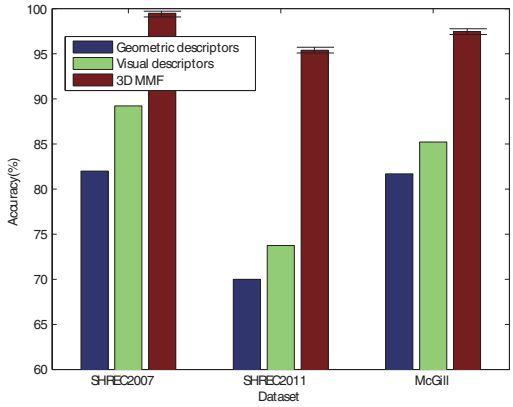

Figure 6: The average accuracies on different dataset. The upper and lower bounds are drawn with black line on the 3D MMF bar.

model, hence, we can obtain more discriminative power when both different modalities are considered. Among three datasets, the results on SHREC 2011 have the lowest accuracy because of the small shape variance leading to insensitive description. From the table, we notice that the classification performance of the geometric descriptors is worse compared with visual descriptor. The main reason is that operation of voxelization loses certain in-

formation like topological relationships, which leads to inadequate performance, though voxelization is easily used for 3D shape analysis with CDBN model. Actually, the RBM is a probabilistic and unsupervised model leading to slightly different result in each experiment. Generally, we take the average accuracy as the result of each dataset. From the Fig. 6, we can clearly see that the accuracy of 3D MMF is slightly fluctuant with repeated experiment. In the figure, the upper and lower bounds are drawn with black line on the 3D MMF bar. In order to prove the importance of fusing feature procedure, we concatenate the geometry-based and view-based feature to a vector directly and get a new feature. With the 94.5% accuracy on dataset SHREC 2007, we know that the direct way is not better than 3D MMF which can reach 99.5%.

In the field of machine learning, confusion matrix [52] is a specific table layout that allows visualization of the performance of an algorithm. A confusion matrix contains information about actual and predicted classifications done by a classification system. To further analyze the recognition results in detail, the confusion matrices of classifications on three datasets with the 3D MMF are visualized in the Figs. 4 and 5. From the results, we can draw the conclusion that the proposed method has a promising prospect for recognition. The abscissa denotes the actual classification and the ordinate indicates the predicted classification. As we can see left in Fig. 4, it is SHREC 2007 confusion matrix. Obviously, there is only an error classification that 10% of 'bearings' are classified as 'human' falsely as 'breaing' is similar to 'human'. Different colors donate distinct probabilities. In this confusion matrix, the contrasting light and dark panes represent various probabilities. The light pane means the low probability indicating that the abscissa class does not likely appear on the ordinate class. On the contrast, the dark pane shows that probability of the abscissa class appearing on the ordinate class is much greater.

4.3. Experiments on Retrieval

For the retrieval tasks, there are 6 standard evaluation metrics used to assess the performance of the recommended 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 [53]. We use the models trained in the classification experiments to calculate the 3D MMF for every 3D shape. The Eq. (9) is utilized to describe the similarity between two models.

Table 3: The precision values of 20, 40, 60, and 80 return items on SHREC 2007. The unit in this table is percentage.

Methods	20	40	60	80
DLE [49]	54.6	32.9	24.1	19.0
MDD [49]	62.6	36.6	26.2	20.5
STT [49]	56.4	34.6	25.2	19.9
SI-MSC [49]	60.4	36.6	26.2	20.5
aMRG [49]	71.4	41.4	29.0	22.5
ERG [54]	62.4	41.5	30.5	24.4
3D MMF	**97.2**	**49.4**	**33.0**	**24.8**

Table 4: The recall values of 20, 40, 60, and 80 return items on SHREC 2007. The unit in this table is percentage.

Methods	20	40	60	80
DLE [49]	54.6	65.8	72.4	76.3
MDD [49]	62.6	73.2	78.6	82.1
STT [49]	56.4	69.2	75.6	79.8
SI-MSC [49]	60.4	73.2	78.8	82.2
aMRG [49]	71.4	82.8	87.2	90.2
ERG [54]	62.4	82.9	91.6	97.5
3D MMF	**97.2**	**98.8**	**99.1**	**99.3**

Retrieval Experiments on SHREC 2007. We take retrieval experiments on SHREC 2007 dataset to evaluate the retrieval performance. The recall-precision curves of our method and some state-of-the-art approaches are plotted in Fig. 7, which includes depth line encoding (DLE) [49], multivariate density-based descriptor (MDD) [49], spherical trace transform (STT) [49] and augmented multi-resolution Reeb graph (aMRG) [49]. Numberical values for the averaged precision and recall on all the models in the dataset are listed in Tables 3 and 4, respectively. We list these values of returned 20, 40, 60, and 80 items, which are 1, 2, 3, and 4 times the size of each class. From the figure and tables, the fact is clear that the recommended approach achieves the best retrieval result overall. Geometric descriptors from voxels and visual descriptors from views are employed for the retrieval experiments. Although the performance of only visual feature is obviously superior compared with its competitors, the 3D MMF is still better than it. The reason is that uni-modal feature can merely deliver specific aspects information of 3D shape, but the suggested method fuses shape and view based modality information. Thus the 3D MMF contain both intrinsic attributes and extrinsic properties of 3D models. Meanwhile the 3D MMF increases the intra-class similarity and reduces the inter-class similarity. As a consequence, the retrieval performance is improved.

Table 5 lists the numerical evaluation results. From

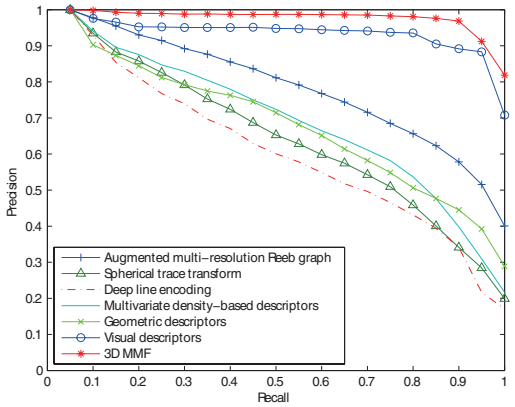

Figure 7: Recall-precision curve of some state-of-the-art methods and proposed method on SHREC 2007.

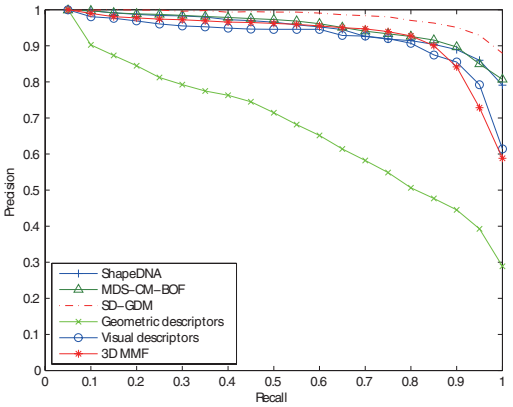

Figure 8: Recall-precision curve of some state-of-the-art methods and proposed method on SHREC 2011.

Table 5: Retrieval performance of proposed method using standard measures on SHREC 2007. The unit in this table is percentage.

Feature	NN	FT	ST	E	DCG
Geometric descriptors	85.50	57.82	36.11	50.70	86.78
Visual descriptors	96.50	86.98	47.99	68.61	97.26
3D MMF	**99.75**	**91.77**	**49.07**	**71.43**	**99.16**

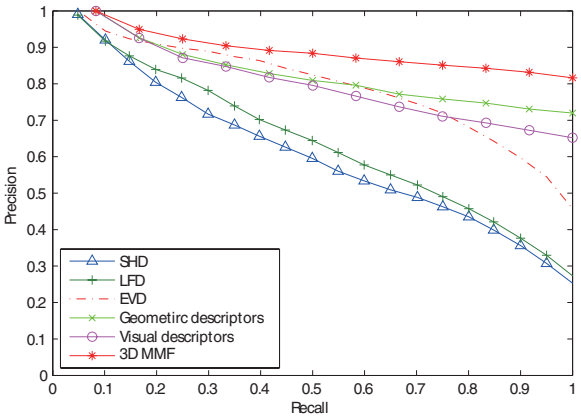

Figure 9: Recall-precision curve of some state-of-the-art methods and proposed method on McGill.

the table, we can clearly see that all of the measures are highly improved from using uni-modal features to multi-modal features. The average improvement of NN, FT, ST, E, DCG index are respectively 8.75%, 19.37%, 7.02%, 11.78%, 7.14%, which demonstrates that the 3D MMF generated by multi-modal fused method has the outstanding capability to improve retrieval performance.

Retrieval Experiments on SHREC 2011 and McGill. We also conduct retrieval experiments on SHREC 2011 and McGill datasets to evaluate the retrieval performance. The recall-precision curves of our method and some state-of-the-art approaches including ShapeDNA [55], Multidimensional Scaling, Clock Matching, and Bag-of-Features (MDS-CM-BOF) [56], Spectral Decomposition of the Geodesic Distance Matrix (SD-GDM) [50], The Spherical Harmonics Descriptor (SHD) [57], Light Field Distribution (LFD) [14] and Eigenvalue descriptor (EVD) [58] are plotted in Figs. 8 and 9. Tables 6 and 7 list the numerical evaluations. The results show that the retrieve performance of the proposed method has promising prospect.

5. Conclusion

In this paper, we put forward a novel multi-modal feature extraction and fusion method for recognition and retrieval of 3D shapes. First, geometric descriptors and visual descriptors are extracted as geometry-based feature and view-based feature through CDBNs and CNNs, respectively. Then two DBNs are adopted to learn structural high-level descriptors. Furthermore, to discover the deep interrelation across modalities, we utilize a RBM to fuse these high-level features. Experiments conducted on standard benchmarks for classification and retrieval tasks have demonstrated that recommended methods achieve much better performance in comparison with state-of-the-art approaches. The experiments results show that the joint representations are more discriminative which can suppress intra-class vari-

Table 6: Retrieval performance of proposed method using standard measures on SHREC 2011. The unit in this table is percentage.

Feature	NN	FT	ST	E	DCG
Geometric descriptors	85.50	57.81	36.11	50.70	86.78
Visual descriptors	96.83	86.02	46.51	**89.39**	96.86
3D MMF	**98.00**	**86.85**	**46.80**	67.76	**97.35**

Table 7: Retrieval performance of proposed method using standard measures on McGill. The unit in this table is percentage.

Feature	NN	FT	ST	E	DCG
Geometric descriptors	88.84	62.45	37.68	59.66	88.37
Visual descriptors	87.30	54.12	36.23	51.91	86.51
3D MMF	**92.12**	**73.19**	**42.12**	**68.55**	**92.38**

ation and enhance the inter-class similarity separation.

Different from the traditional shape analysis methods in computer vision, we sufficiently consider both intrinsic properties and extrinsic visual similarities. In addition, we do not simply fuse the geometric and visual features to train the model, instead, we take the strategy that first learns high-level features for each modality through DBN to remove the modality-specific information, and then high-level features are concatenated to learn multi-modal feature for shape analysis. By using this strategy, the highly non-linear correlations between geometry-based and view-based modalities can be modeled comprehensively.

Limitations. In the procedure of generating geometric descriptors, accuracy will be higher with bigger size of grid, but computation is exponent increasing. Therefore, an appropriate method should be designed to balance these two aspects. In our framework, to learn the joint representation for 3D shape, we concatenate the high-level features from different modalities. However, the information carried by each modality feature is not identical, as can be seen from the experiments on SHREC 2007 dataset, the visual descriptors contains more information than geometric descriptors. Therefore a solution should be sought to represent the importance of different modalities. Moreover, in the proposed method, features for deep learning are global so that local information of 3D shapes is missing more or less. Hence, the proposed method is difficult to be applied for more sophisticated tasks such as segmentation, partial retrieval, and symmetric detection.

Future work. First, at present we only investigate geometric and visual descriptors in our framework. In order to describe 3D shapes much better, we will explore the possibility to combine global features and local features from each modality. Second, 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 methods that can directly process graph-based data, including 3D mesh data, communication network, and traffic network, which lead to its wider applications with better performance.

Acknowledgements

This work is partly supported by grants from National Natural Science Foundation of China (61573284, 61672430), the Fundamental Research Funds for the Central Universities (310201401-(JCQ01009,JCQ01012)), Open Projects Program of National Laboratory of Pattern Recognition (NLPR).

References

[1] J. W. Tangelder, R. C. Veltkamp, A survey of content based 3d shape retrieval methods, Multimedia tools and applications 39 (3) (2008) 441–471.

[2] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H. Van Nguyen, R. Ohbuchi, et al., A comparison of methods for non-rigid 3d shape retrieval, Pattern Recognition 46 (1) (2012) 449–461.

[3] Z. Liu, S. Bu, K. Zhou, S. Gao, J. Han, J. Wu, A survey on partial retrieval of 3d shapes, Journal of Computer Science and Technology 28 (5) (2013) 836–851.

[4] S. Bu, Z. Liu, J. Han, J. Wu, R. Ji, Learning high-level feature by deep belief networks for 3-d model retrieval and recognition, Multimedia, IEEE Transactions on 16 (8) (2014) 2154–2167.

[5] S. Bu, S. Cheng, Z. Liu, J. Han, Multimodal feature fusion for 3d shape recognition and retrieval, MultiMedia, IEEE 21 (4) (2014) 38–46.

[6] S. Bu, P. Han, Z. Liu, J. Han, H. Lin, Local deep feature learning framework for 3d shape, Computers & Graphics 46 (2015) 117–129.

[7] Z. Liu, X. Wang, S. Bu, Human-centered saliency detection, IEEE Transactions on Neural Networks and Learning Systems 27 (6) (2016) 1150–1162.

[8] Y. Bengio, Learning deep architectures for ai, Foundations and trends in Machine Learning 2 (1) (2009) 1–127.

[9] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[10] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[11] Z. Xie, K. Xu, W. Shan, L. Liu, Y. Xiong, H. Huang, Projective feature learning for 3d shapes with multi-view depth images, in: Computer Graphics Forum, Vol. 34, Wiley Online Library, 2015, pp. 1–11.

[12] Z. Zhu, X. Wang, S. Bai, C. Yao, X. Bai, Deep learning representation using autoencoder for 3d shape retrieval, in: Security, Pattern Analysis, and Cybernetics (SPAC), 2014 International Conference on, IEEE, 2014, pp. 279–284.

[13] H. Murase, S. K. Nayar, Visual learning and recognition of 3-d objects from appearance, International journal of computer vision 14 (1) (1995) 5–24.

[14] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, M. Ouhyoung, On visual similarity based 3d model retrieval, in: Computer graphics forum, Vol. 22, Wiley Online Library, 2003, pp. 223–232.

[15] Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai, 3d object retrieval with hausdorff distance learning, Industrial Electronics, IEEE Transactions on 61 (4) (2014) 2088–2098.

[16] H. Laga, Semantics-driven approach for automatic selection of best views of 3d shapes, in: Proceedings of the 3rd Eurographics conference on 3D Object Retrieval, Eurographics Association, 2010, pp. 15–22.

[17] X. Bonaventura, J. Guo, W. Meng, M. Feixas, X. Zhang, M. Sbert, 3d shape retrieval using viewpoint information-theoretic measures, Computer Animation and Virtual Worlds 26 (2) (2015) 147–156.

[18] H. Tabia, D. Picard, H. Laga, P.-H. Gosselin, Compact vectors of locally aggregated tensors for 3d shape retrieval, in: Eurographics Workshop on 3D Object Retrieval, 2013.

[19] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3-d object retrieval and recognition with hypergraph analysis, Image Processing, IEEE Transactions on 21 (9) (2012) 4290–4303.

[20] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, N. Zhang, Less is more: efficient 3-d object retrieval with query view selection, Multimedia, IEEE Transactions on 13 (5) (2011) 1007–1018.

[21] A. E. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3d scenes, Pattern Analysis and Machine Intelligence, IEEE Transactions on 21 (5) (1999) 433–449.

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

[23] I. Sipiran, B. Bustos, T. Schreck, Data-aware 3d partitioning for generic shape retrieval, Computers & Graphics.

[24] J. Knopp, M. Prasad, G. Willems, R. Timofte, L. Van Gool, Hough transform and 3d surf for robust three dimensional classification, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 589–602.

[25] R. López-Sastre, A. García-Fuertes, C. Redondo-Cabrera, F. Acevedo-Rodríguez, S. Maldonado-Bascón, Evaluating 3d spatial pyramids for classifying 3d shapes, Computers & Graphics.

[26] J. Hu, J. Hua, Salient spectral geometric features for shape matching and retrieval, The visual computer 25 (5-7) (2009) 667–675.

[27] H.-Y. Wu, H. Zha, T. Luo, X.-L. Wang, S. Ma, Global and local isometry-invariant descriptor for 3d shape comparison and partial matching, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 438–445.

[28] A. Dubrovina, R. Kimmel, Matching shapes by eigendecomposition of the laplace-beltrami operator, in: Proceedings of Symposium on 3D Data Processing, Vol. 2, 2010.

[29] G. Lavoue, Bag of words and local spectral descriptor for 3d partial shape retrieval, in: Eurographics Workshop on 3D Object Retrieval, 2011.

[30] M. Ben-Chen, C. Gotsman, Characterizing shape using conformal factors, in: Eurographics Workshop on 3D Object Retrieval, 2008, pp. 1–8.

[31] L. Shapira, S. Shalom, A. Shamir, D. Cohen-Or, H. Zhang, Contextual part analogies in 3d objects, International Journal of Computer Vision 89 (2-3) (2010) 309–326.

[32] K. Sfikas, T. Theoharis, I. Pratikakis, Non-rigid 3d object retrieval using topological information guided by conformal factors, The Visual Computer 28 (9) (2012) 943–955.

[33] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative multi-scale signature based on heat diffusion, in: Computer Graphics Forum, Vol. 28, Wiley Online Library, 2009, pp. 1383–1392.

[34] M. M. Bronstein, I. Kokkinos, Scale-invariant heat kernel signatures for non-rigid shape recognition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 1704–1711.

[35] M. Spagnuolo, M. Bronstein, A. Bronstein, A. Ferreira, et al., Affine-invariant photometric heat kernel signatures (2012) 39–46.

[36] M. Ovsjanikov, A. M. Bronstein, M. M. Bronstein, L. J. Guibas, Shape google: a computer vision approach to isometry invariant shape retrieval, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 320–327.

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

[38] I. Kokkinos, M. M. Bronstein, R. Litman, A. M. Bronstein, Intrinsic shape context descriptors for deformable shapes, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 159–166.

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

[40] S. Bu, P. Han, Z. Liu, K. Li, J. Han, Shift-invariant ring feature for 3d shape, The Visual Computer 30 (6-8) (2014) 867–876.

[41] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, arXiv preprint arXiv:1505.00880.

[42] Y. Zhao, Y. Liu, Y. Wang, B. Wei, J. Yang, Y. Zhao, Y. Wang, Region-based saliency estimation for 3d shape analysis and understanding, Neurocomputing 197 (2016) 1–13.

[43] F. Chen, R. Ji, L. Cao, Multimodal learning for view-based 3d object classification, Neurocomputing 195 (2016) 23–29.

[44] B. Leng, X. Zhang, M. Yao, Z. Xiong, A 3d model recognition mechanism based on deep boltzmann machines, Neurocomputing 151 (2015) 593–602.

[45] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[46] G. E. Hinton, Training products of experts by minimizing contrastive divergence, Neural computation 14 (8) (2002) 1771–1800.

[47] T. Tieleman, G. Hinton, Using fast weights to improve persistent contrastive divergence, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1033–1040.

[48] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, Citeseer, 1990.

[49] D. Giorgi, S. Biasotti, L. Paraboschi, Watertight models track, Tech. rep., Technical Report (2007).

[50] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H. Van Nguyen, R. Ohbuchi, et al., Shrec'11 track: Shape retrieval on non-rigid 3d watertight meshes., 3DOR 11 (2011) 79–88.

[51] J. Zhang, K. Siddiqi, D. Macrini, A. Shokoufandeh, S. Dickinson, Retrieving articulated 3-d models using medial surfaces and their graph spectra, in: Energy minimization methods in computer vision and pattern recognition, Springer, 2005, pp. 285–300.

[52] J. T. Townsend, Theoretical analysis of an alphabetic confusion matrix, Perception & Psychophysics 9 (1) (1971) 40–50.

[53] P. Shilane, P. Min, M. Kazhdan, T. Funkhouser, The princeton shape benchmark, in: Shape Modeling Applications, 2004. Proceedings, IEEE, 2004, pp. 167–178.

[54] V. Barra, S. Biasotti, 3d shape retrieval using kernels on extended reeb graphs, Pattern Recognition 46 (11) (2013) 2985–2999.

[55] M. Reuter, F.-E. Wolter, N. Peinecke, Laplace-beltrami spectra as 'shape-dna' of surfaces and solids, Computer-Aided Design 38 (4) (2006) 342–366.

[56] Z. Lian, A. Godil, X. Sun, H. Zhang, Non-rigid 3d shape retrieval using multidimensional scaling and bag-of-features, in: Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE, 2010, pp. 3181–3184.

[57] M. Kazhdan, T. Funkhouser, S. 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.

[58] R. Gal, D. Cohen-Or, Salient geometric features for partial shape matching and similarity, ACM Transactions on Graphics (TOG) 25 (1) (2006) 130–150.

Biography

Shuhui Bu received the PhD degree in College of Systems and Information Engineering from University of Tsukuba, Japan in 2009. Currently, he is an associate professor at Northwestern Polytechnical University, China. Prior to joining Northwestern Polytechnical University, he was an assistant professor at Kyoto University, Japan. His research includes 3D shape analysis, image processing, and computer vision.

Lei Wang received the BS degree from Northwestern Polytechnical University of China in 2013. Currently, he is working toward the MS degree at Northwestern Polytechnical University, China. His research interests include artificial intelligence, 3D shape analysis, image processing, and computer vision.

Pengcheng Han received the BS degree from Northwestern Polytechnical University of China in 2013. Currently, he is working toward the MS degree at Northwestern Polytechnical University, China. His research interests include artificial intelligence, 3D shape analysis, image processing, and computer vision.

Zhenbao Liu received the Ph.D. degree in computer science from the College of Systems and Information Engineering, University of Tsukuba, Japan, in 2009. He is an associate professor at Northwestern Polytechnical University, Xi'an. He was a visiting scholar in the GrUVi Lab of Simon Fraser University in 2012. His research interests include 3D shape analysis, matching, retrieval and segmentation.