Discriminative Auto-encoder for Robust 3D Shape Retrieval

Anonymous CVPR submission

Paper ID ****

Abstract

With the recent development in 3D acquisition and printing technology, there has been a rapid growth of available three dimensional (3D) models in diverse areas such as engineering, medicine and biology, etc. It is therefore of great interest to develop the effective and efficient methods for 3D shape retrieval. In this paper, we propose a high-level shape feature learning scheme for retrieval with a deep auto-encoder. First, the distributions of heat kernel signature (HKS) at different scales are developed to represent shape as the input of the auto-encoder. Then, by imposing the Fisher discrimination criterion on the neurons in the hidden layer of the neural network, a discriminative auto-encoder is proposed to extract the intrinsic shape features so that they have small within-class scatter but big between-class scatter. Finally, the neurons in hidden layers are concatenated to form a shape descriptor for retrieval from multiple discriminative auto-encoders, which can be formed by using multiple distributions of HKS at different scales as the input of the auto-encoder. The proposed method is evaluated on three datasets with large geometry variations, i.e., Mcgill, SHREC 10 ShapeGoogle and Protein datasets. Experimental results on the benchmark datasets demonstrate the effectiveness of the proposed method within applications of 3D shape matching and searching.

1. Introduction

With the recent advancement of 3D data acquisition and printing technology, we have observed an explosive growth of the 3D meshed surface models in a variety of fields, such as engineering, entertainment and medical imaging [16, 14, 12, 5, 4, 1]. Due to the data-richness of 3D models, shape retrieval for 3D model searching, understanding and analyzing has been receiving more and more attention. Using a shape as a query, the shape retrieval algorithm aims to find the similar shapes to the query. The performance of a shape retrieval algorithm mainly relies on shape descriptor which can effectively capture the distinctive properties of shape, while it is invariant to different classes of transformations. Moreover, the shape descriptor should be insensitive to both topological and numerical noise such as consistent behavior even with topological short-circuits and numerical noises. Also, it is a intrinsic and compact representation to describe the hidden geometric structures of the mesh for fast and effective shape matching. Once the shape descriptor is formed, the similarity between two shapes is determined by the similarity between the shape descriptors and used for retrieval.

The shape descriptor for shape matching and retrieval has been extensively studied in geometry community [20, 9, 6, 21, 18]. In the past decades, plenty of shape descriptors have been proposed, such as the $D2$ shape distribution [6], statistical moments of the model [21, 17], Fourier descriptor [3], Light Field Descriptor [10], Eigenvalue Descriptor (EVD) Although these shape descriptors can represent the shape effectively, they are either sensitive to non-rigid transformation or topological changes, or computationally inefficient. To be invariant to the isometric transformation, the local geometric features are extracted to represent the shape, such as spin images, histograms [11, 8, 7, 13]. Although these shape descriptors can describe deformable shapes well, they are sensitive to local geometric noise and are not able to characterize the global structure of the shape. In addition, some of these descriptors require large amounts of space to store the local descriptors in order for the global shape matching.

Apart from the earlier shape descriptors, another popular approaches to shape retrieval are diffusion based methods [19, 2, 16, 15]. Based on the Laplace-Beltrami operator, global point signature (GPS) [] was proposed to represent shape. Since the eigenfunctions of the Laplace-Beltrami operator are able to robustly characterize the point on meshed surface, each vertex is represented by a high dimensional vector of scaled eigenfunctions of the Laplace-Beltrami operator evaluated at the vertex. The high dimensional vector is called GPS. Another widely used shape signature is heat kernel signature (HKS) [], where Sun et al. proposed to use the diagonal of the heat kernel as a local descriptor to represent shape. HKS is invariant to isometric deformations, insensitive to the small perturbations of the surface. Both

GPS and HKS are point based signatures, which characterizes each vertex on the meshed surface using a vector.

In the aforementioned methods, the shape descriptors are hand-crafted rather than learned from a set of training shapes. In [15], the authors applied the bag-of-features (BOF) paradigm to learn the shape descriptor. The dictionary of words is learned by the K-means clustering method from a set of HKSs of shapes. Then a histogram of pairs of spatially-close words over the learned dictionary is formed as the shape descriptor for retrieval. Based on K-means clustering, Lavouè et al. combined the standard and spatial BOF descriptors for shape retrieval. Since K-means clustering can be viewed as a special case of sparse coding, Litman et al. employed sparse coding to learn the dictionary of words instead of K-means clustering. The histogram of encoded representation coefficient over the learned dictionary is used to represent shape for retrieval. Moreover, in order to obtain the discriminative representation coefficients, a class-specific dictionary is constructed in a supervised way.

Recently, the deep auto-encoder has been widely used in many challenging tasks such as image denoising, image classification and face recognition due to the favorable ability of modeling the nonlinearity by mapping the high dimensional feature to the low dimensional discriminative feature in the hidden layer of the network. Inspired by great success of the deep auto-encoder in computer vision and pattern recognition, in this paper, we develop a novel auto-encoder based shape descriptor for retrieval, which imposes the Fisher discrimination criterion on the hidden layer to make the hidden layer features discriminative. It is expected that the neurons in the hidden layer have small within-class scatter but big between-class scatter. Moreover, in order to much more effectively represent shape, by using the HKS histograms at different scales as the inputs of the auto-encoder, we train a stacked discriminative auto-encoder and concatenate all neurons in the hidden layers as the high-level learning shape descriptor for retrieval. The proposed shape descriptor is verified on the representative and benchmark shape datasets, showing very promising performance.

The rest of the paper is organized as follows. Section 2 briefly introduces HKS and auto-encoder. Section 3 presents the proposed shape descriptor with the discriminative auto-encoder. Section 4 performs extensive experiments and Section 5 concludes the paper.

2. Background

2.1. Heat Kernel Signature

The 3D model is represented as a graph $G = (V, E, W)$, where V is the set of vertices, E is the set of edges and W is the weigh value for each edge. Given a graph constructed by connecting pairs of data points with weighted edges, the heat kernel $H_t(x, y)$ measures the heat flow across a graph, which is defined to the amount of the heat passing from the vertex x to the vertex y within a certain amount of time. The heat flow across the surface is governed by the heat equation $u(x, t)$, where x denotes one vertex on the meshed surface and t denotes the diffusion time. Provided that there is an initial heat distribution on meshed surface at $t = 0$, the heat kernel provides the fundamental solution of the heat equation, which is associated with the Laplace-Beltrami operator L by:

$$\frac{\partial H_t}{\partial t} = -LH_t \tag{1}$$

where H_t denotes the heat kernel and t is the diffusion time. The solution of Eq. (1) can be obtained by the eigenfunction expansion by the Laplace-Beltrami operator described below.

$$H_t = exp(-tL) \tag{2}$$

By the spectral theorem, the heat kernel can be further expressed as follows:

$$H_t(x, y) = \sum_i e^{-\lambda_i t}\phi_i(x)\phi_i(y) \tag{3}$$

where λ_i is the i^{th} eigenvalue of the Laplacian, ϕ_i is the i^{th} eigenfunction, and x and y denotes the vertex x and y, respectively. Heat kernel signature (HKS) of the vertex x at time t, S_x^t, is defined as the diagonal of the heat kernel of the vertex x taken at time t:

$$\begin{aligned} S_x^t &= H_t(x, x) \\ &= \sum_{i=0} e^{-\lambda_i t}\phi_i(x)\phi_i(x) \end{aligned} \tag{4}$$

HKS, as a point signature, can capture information of the neighborhood of a point x on the shape at a scale defined by t. In the following section, without the specific instruction, we use t to represent the scale of HKS, where $t = 1, 2, \cdots, T$.

2.2. Auto-encoder

An auto-encoder neural network usually consists of two parts, i.e., encoder and decoder. The encoder, denoted by F, maps the input $\boldsymbol{x} \in \mathcal{R}^{d\times 1}$ to the hidden layer representation, denoted by $\boldsymbol{z} \in \mathcal{R}^{r\times 1}$, where d is the dimension of the input and r is the number of neurons in the hidden layer. In the auto-encoder neural network, one neuron in the layer l is connected to all the neurons in the layer $l + 1$. We denote the weight and bias connecting the layer l and the layer $l + 1$ by $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$, respectively. The output of the layer is called the activation function. Usually, the activation function is non-linear, such as sigmoid function $\sigma(\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{x}}}$

or tanh function $\sigma(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}}-e^{-\boldsymbol{x}}}{e^{\boldsymbol{x}}+e^{-\boldsymbol{x}}}$. Therefore, the output $f(\boldsymbol{a}^l)$ of the layer $l+1$ is :

$$f_{l+1}(\boldsymbol{a}^l) = \sigma(\boldsymbol{W}^l\boldsymbol{a}^l + \boldsymbol{b}^l) \tag{5}$$

where $f_{l+1}(\boldsymbol{a}^l)$ is the activation function in the layer $l+1$ and $\boldsymbol{a}^l$ is the neurons in the layer l. Thus, the encoder $F(\boldsymbol{x})$ of k hidden layers can be represented as follows:

$$F(\boldsymbol{x}) = f_k(f_{k-1}(\cdots, f_1(\boldsymbol{x}))) \tag{6}$$

The decoder, denoted by G, maps the hidden layer representation $\boldsymbol{z}$ back to the input $\boldsymbol{x}$. It is defined:

$$\boldsymbol{x} = f_{2k}(f_{2k-1}(\cdots, f_{k+1}(\boldsymbol{z}))) \tag{7}$$

Denote by $\boldsymbol{W}$ and $\boldsymbol{b}$ the weights and biases of all layers in the auto-encoder, respectively, where $\boldsymbol{W} = [\boldsymbol{W}^1, \boldsymbol{W}^2, \cdots, \boldsymbol{W}^{2k-1}]$ and $\boldsymbol{b} = [\boldsymbol{b}^1, \boldsymbol{b}^2, \cdots, \boldsymbol{b}^{2k-1}]$. To optimize the parameters $\boldsymbol{W}$ and $\boldsymbol{b}$, the standard auto-encoder minimizes the following cost function:

$$\begin{aligned} <\hat{\boldsymbol{W}}, \hat{\boldsymbol{b}}> = & argmin_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{2}\sum_{i=1}^{N} \|\boldsymbol{x}_i - G(F(\boldsymbol{x}_i))\|_2^2 \\ & + \frac{1}{2}\lambda\|\boldsymbol{W}\|_2^2 \end{aligned} \tag{8}$$

where x_i represents the i^{th} one of the N training samples, parameter λ is the positive scalar. In Eq. (8), the first term is the reconstruction error and the second term is the regularization term that prevents overfitting. An efficient optimization method can be implemented by the restricted Boltzman machine and back-propagation framework. The reader can see [] for more details.

3. Shape descriptor based on discriminative auto-encoder

We detail the proposed framework of the discriminative auto-encoder based shape descriptor, which comprises three components, namely, HKS histogram, discriminative auto-encoder and 3D shape descriptor. In the HKS histogram component, the distribution of heat kernel signatures of shape is extracted as the low-level feature to input the discriminative auto-encoder. Then we train a discriminative auto-encoder to learn a high level feature in the hidden layer in the discriminative auto-encoder component. In the 3D shape descriptor component, we form a descriptor from all hidden layer representations of a stacked discriminative auto-encoder for shape retrieval.

3.1. HKS histogram

Suppose there are C shapes, each of which has J samples. We use $\boldsymbol{y}_{i,j}$ to index the j^{th} sample of the i^{th} shape. For each shape $\boldsymbol{y}_{i,j}$, we extract HKS feature $\boldsymbol{S}_{i,j} \in \mathcal{R}^{T\times N}$, where $\boldsymbol{S}_{i,j} = [\boldsymbol{S}_{i,j}^1, \boldsymbol{S}_{i,j}^2, \cdots, \boldsymbol{S}_{i,j}^T]$, $\boldsymbol{S}_{i,j}^t$ denotes the heat kernel signature of the shape $\boldsymbol{y}_{i,j}$ at the t^{th} scale, $t = 1, 2, \cdots, T$, N is the number of vertices of shape $\boldsymbol{y}_{i,j}$ and T is the number of scales. For the scale t, we calculate the distribution of heat kernel signatures of N vertices of the shape $\boldsymbol{y}_{i,j}$ to form the HKS histogram $\boldsymbol{h}_{i,j}^t$. The HKS histograms can convert different dimensional HKSs of different shapes to the same dimensional distributions, which can be suitable to the input of the auto-encoder. Moreover, the HKS histogram, as a low-level feature, has the potential to be robust to different geometry transformations. Thus, it can facilitate to extract invariant high-level feature in the hidden layer for retrieval.

In addition, we normalize the HKS histogram, which is centralized by the mean and variance of the HKS histograms over all training samples from C classes, namely,

$$\boldsymbol{h}_{i,j}^t = \frac{\boldsymbol{h}_{i,j}^t - \boldsymbol{h}^t}{\boldsymbol{v}^t} \tag{9}$$

where $\boldsymbol{h}^t$ and $\boldsymbol{v}^t$ are the mean and variance of all training HKS histograms $\boldsymbol{h}_{i,j}^t$.

3.2. Discriminative auto-encoder

In this subsection, we propose a discriminative auto-encoder to extract discriminative high-level feature for shape retrieval. In order to boost the discriminative power of the hidden layer features, we impose a Fisher discrimination criterion on them. Given the input $\boldsymbol{x}_i^t$ of the shape class i at the scale t, $\boldsymbol{x}_i^t = [\boldsymbol{h}_{i,1}^t, \boldsymbol{h}_{i,2}^t, \cdots, \boldsymbol{h}_{i,J}^t]$, we denote by $\boldsymbol{z}^t$ the hidden layer features of the auto-encoder from all classes. We can write $\boldsymbol{z}^t$ as $\boldsymbol{z}^t = [\boldsymbol{z}_1^t, \boldsymbol{z}_2^t, \cdots, \boldsymbol{z}_C^t]$, where $\boldsymbol{z}_i^t = [\boldsymbol{z}_{i,1}^t, \boldsymbol{z}_{i,2}^t, \cdots, \boldsymbol{z}_{i,J}^t]$, $\boldsymbol{z}_{i,j}^t$ is the hidden layer feature of the j^{th} sample from the class i, $i = 1, 2, \cdots, C$, $j = 1, 2, \cdots, J$. Based on the Fisher discriminative criterion, the discrimination can be achieved by minimizing the within-class scatter of $\boldsymbol{z}^t$, denoted by $S_w(\boldsymbol{z}^t)$, and maximizing the between-class scatter of $\boldsymbol{z}^t$, denoted by $S_b(\boldsymbol{z}^t)$. $S_w(\boldsymbol{z}^t)$ and $S_b(\boldsymbol{z}^t)$ are defined as:

$$\begin{aligned} S_w(\boldsymbol{z}^t) &= \sum_{i=1}^{C} \sum_{\boldsymbol{z}_{i,j}^t \in i} (\boldsymbol{z}_{i,j}^t - \boldsymbol{m}_i^t)(\boldsymbol{z}_{i,j}^t - \boldsymbol{m}_i^t)_i)^T \\ S_b(\boldsymbol{z}^t) &= \sum_{i=1}^{C} n_i(\boldsymbol{m}_i^t - \boldsymbol{m}^t)(\boldsymbol{m}_i^t - \boldsymbol{m}^t)^T \end{aligned} \tag{10}$$

where $\boldsymbol{m}_i^t$ and $\boldsymbol{m}^t$ are the mean vector of $\boldsymbol{z}_i^t$ and $\boldsymbol{z}^t$, respectively, and n_i is the number of samples of class i. Intuitively, we can define the discriminative regularization term $L(\boldsymbol{z}^t)$ as $tr(S_w(\boldsymbol{z}^t))) - tr(S_b(\boldsymbol{z}^t)))$. Thus, by incorporating the discriminative regularization term into the standard auto-encoder model, we can form the following objective

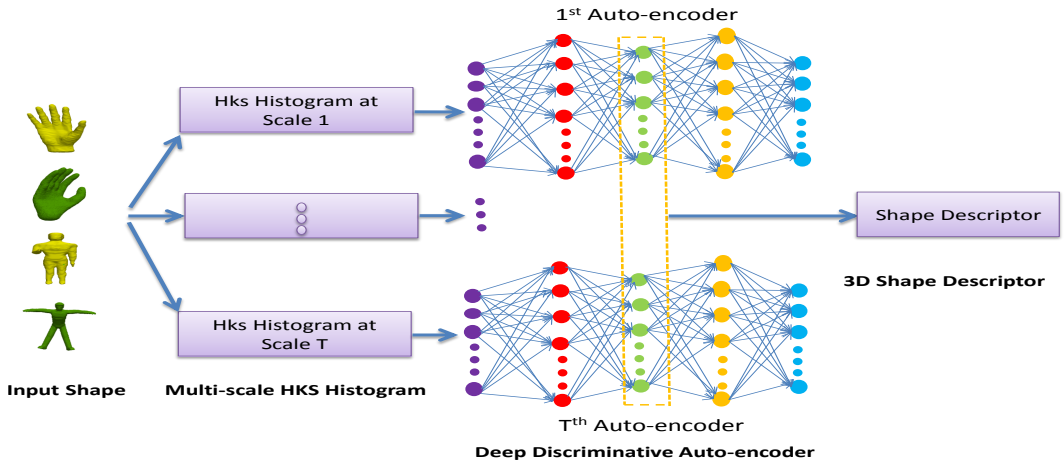


Figure 1. The framework of the proposed discriminative auto-encoder based shape descriptor.

function of the discriminative auto-encoder:

$$J(\boldsymbol{W}^t, \boldsymbol{b}^t) = argmin_{\boldsymbol{W}^t, \boldsymbol{b}^t} \sum_{i=1}^{C} \|\boldsymbol{x}_i^t - G(F(\boldsymbol{x}_i^t))\|_2^2 + \frac{1}{2}\lambda\|\boldsymbol{W}^t\|_2^2 + \frac{1}{2}\gamma(tr(S_w(\boldsymbol{z}^t)) - tr(S_b(\boldsymbol{z}^t))). \tag{11}$$

For the sample $\boldsymbol{h}_{i,j}^t$, we define the following functions:

$$J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t) = \|\boldsymbol{h}_{i,j}^t - G(F(\boldsymbol{h}_{i,j}^t))\|_2^2 \tag{12}$$

$$L_0(\boldsymbol{z}_{i,j}^t) = (\boldsymbol{z}_{i,j}^t - \boldsymbol{m}_i^t)(\boldsymbol{z}_{i,j}^t - \boldsymbol{m}_i^t)_i)^T + (\boldsymbol{m}_i^t - \boldsymbol{m}^t)(\boldsymbol{m}_i^t - \boldsymbol{m}^t)^T \tag{13}$$

To optimize the objective function of the discriminative auto-encoder, we adopt the back-propagation method of the error. We denote by $W_{m,n}^{l,t}$ by the weight associated with the connection between the unit p in the layer l and the unit q in the layer l. Also, $b_m^{l,t}$ is the bias associated with the connection with the unit p in the layer l. The partial derivatives of the overall cost function $J(\boldsymbol{W}^t, \boldsymbol{b}^t)$ can be computed as:

$$\frac{\partial J(\boldsymbol{W}^t, \boldsymbol{b}^t)}{\partial \boldsymbol{W}^{l,t}} = \sum_{i=1}^{C} \sum_{\boldsymbol{h}_{i,j}^t \in i} \frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}} + \lambda \boldsymbol{W}^{l,t} + \gamma \sum_{i=1}^{C} \sum_{\boldsymbol{z}_{i,j}^t \in i} \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}} \tag{14}$$

$$\frac{\partial J(\boldsymbol{W}^t, \boldsymbol{b}^t)}{\partial \boldsymbol{b}^{l,t}} = \sum_{i=1}^{C} \sum_{\boldsymbol{h}_{i,j}^t \in i} \frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}} + \gamma \sum_{i=1}^{C} \sum_{\boldsymbol{z}_{i,j}^t \in i} \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}} \tag{15}$$

Denote by $\boldsymbol{\delta}^{l,t}$ the error of the output layer L in the auto-encoder. For the output layer (the layer L), we have:

$$\boldsymbol{\delta}^{L,t} = -(\boldsymbol{h}_{i,j}^t - \boldsymbol{a}^{L,t})\sigma'(\boldsymbol{u}^{L,t}) \tag{16}$$

where $\boldsymbol{a}^{L,t}$ is the activation of the output layer and $\sigma'(\boldsymbol{u}^{L,t})$ is the derivative of the activation function in the output layer. For other layers $l = L-1, L-2, \cdots, 2$, with the back-propagation method in [], the error $\boldsymbol{\delta}^{l,t}$ can be recursively obtained by the following equation:

$$\boldsymbol{\delta}^{l,t} = ((\boldsymbol{W}^{l,t})^T \boldsymbol{\delta}^{l+1,t})\sigma'(\boldsymbol{u}^{L,t}) \tag{17}$$

Therefore, the partial derivatives of the function $J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)$, $\frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}}$ and $\frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}}$ can be computed :

$$\frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}} = \boldsymbol{\delta}^{l+1,t}(\boldsymbol{a}^{l,t})^T \qquad \frac{\partial J_0(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}} = \boldsymbol{\delta}^{l+1,t} \tag{18}$$

Since $\boldsymbol{z}^t = \boldsymbol{W}^{k-1}\boldsymbol{a}^{k-1} + \boldsymbol{b}^{k-1}$, for $l \neq k$, $\frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}} = 0$ and $\frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}} = 0$. For the hidden layer, i.e., $l = k$, $\frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial W_{p,q}^{l,t}}$ and $\frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial b_p^{l,t}}$ can be computed as follows:

$$\frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial W_{p,q}^{k-1,t}} = \frac{\partial z_{i,j,p}^{k,t}}{\partial W_{p,q}^{k-1,t}} \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial z_{i,j,p}^{k,t}} = a_n^{k-1,t} \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial z_{i,j,p}^{k,t}} \qquad \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial b_p^{k-1,t}} = \frac{\partial L_0(\boldsymbol{z}_{i,j}^t)}{\partial z_{i,j,p}^{k,t}} \tag{19}$$

The partial derivative of $L_0(\boldsymbol{z}_{i,j}^t)$ with respect to $z_{i,j,p}^{k,t}$ can

be obtained:

$$\frac{\partial L(\boldsymbol{z}^t)}{\partial \boldsymbol{z}_p^{k,t}} = 2(1-\frac{1}{n_i})(z_{i,j,p}^{k,t} - m_{i,p}^t) + 2(\frac{1}{n_i} - \frac{1}{\sum n_i})(m_{i,p}^t - m_p^t) \tag{20}$$

Therefore, based on Eqs. (18), (19) and (20), for $l \neq k$, the partial derivatives of the objective function of the discriminative auto-encoder with respect to $\boldsymbol{W}^{l,t}$ and $\boldsymbol{b}^{l,t}$ can be obtained by Eq. (18). For the hidden layer, $\frac{\partial J(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{x}_i^t)}{\partial \boldsymbol{W}^{l,t}}$ and $\frac{\partial J(\boldsymbol{W}^t, \boldsymbol{b}^t, \boldsymbol{x}_i^t)}{\partial \boldsymbol{b}^{l,t}}$ can be computed:

$$\frac{\partial J(\boldsymbol{W}^{l,t}, \boldsymbol{b}^{l,t}, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{W}^{l,t}} = (\boldsymbol{\delta}^{l+1,t} + 2(1-\frac{1}{n_i})(\boldsymbol{z}_{i,j}^{k,t} - \boldsymbol{m}_i^t) + 2(\frac{1}{n_i} - \frac{1}{\sum n_i})(\boldsymbol{m}_i^t - \boldsymbol{m}^t))(\boldsymbol{a}^{l,t})^T \tag{21}$$

$$\frac{\partial J(\boldsymbol{W}^{l,t}, \boldsymbol{b}^{l,t}, \boldsymbol{h}_{i,j}^t)}{\partial \boldsymbol{b}^{l,t}} = \boldsymbol{\delta}^{l+1,t} + 2(1-\frac{1}{n_i})(\boldsymbol{z}_{i,j}^{k,t} - \boldsymbol{m}_i^t) + 2(\frac{1}{n_i} - \frac{1}{\sum n_i})(\boldsymbol{m}_i^t - \boldsymbol{m}^t) \tag{22}$$

With the discriminative auto-encoder, we can learn discriminative high level feature in the hidden layer to form the shape descriptor for retrieval.

3.3. 3D Shape Descriptor

In this subsection, we use the activations of the hidden layer of the discriminative auto-encoder to form the shape descriptor. In order to characterize the intrinsic structure of the shape more effectively, we train multiple discriminative auto-encoders by setting the HKS histograms at different scales to the inputs of the discriminative auto-encoder. That is, for each scale t, we can learn $\boldsymbol{W}^t$ and $\boldsymbol{b}^t$ from a set of training HKS histograms, i.e.,$\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_C^t$, $t = 1, 2, \cdots, T$. Thus, T discriminative auto-encoders can be formed by T groups of HKS histograms. Once the multiple discriminative auto-encoders are trained, we can concatenate the activations of all hidden layers to form a shape descriptor.

Denote the t^{th} encoder of the multiple discriminative auto-encoders by G^t, which corresponds to the input of the HKS histogram at the scale t The shape descriptor of the j^{th} shape from the class i, i.e., activations in the hidden layers of the multiple discriminative auto-encoders, can be represented :

$$\boldsymbol{\alpha}_{i,j} = [G^1(\boldsymbol{h}_{i,j}^1), G^2(\boldsymbol{h}_{i,j}^2), \cdots, G^T(\boldsymbol{h}_{i,j}^T)] \tag{23}$$

Fig. shows the descriptors of two different shapes with pose changes. From this figure, one can see that.

4. Experimental Results

We conducted the experiments for shape matching and retrieval to evaluate performance of the proposed 3D shape descriptor. We define a universal time unit $\tau = 0.01$ and take 101 sampled time values for the computation of the HKS descriptor. And 128 bins are used to form the HKS histogram, which results in the 128-dimensional input of the discriminative auto-encoder. We train a 4 layered auto-encoder, whose layer sizes are 128,1000,500,30, respectively. Moreover, in Eq. (11), λ and γ are set to , respectively.

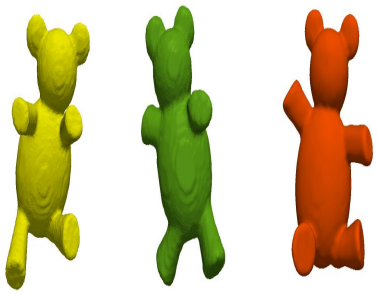
4.1. Shape Matching Performance

The shape matching is a key step in 3D model retrieval. A good shape descriptor should be robust to represent the 3D model with pose changes, topological changes and noise corruption. The models used in the experiment were chosen from the McGill dataset []. We evaluate performance of the proposed shape descriptor from the two aspects.

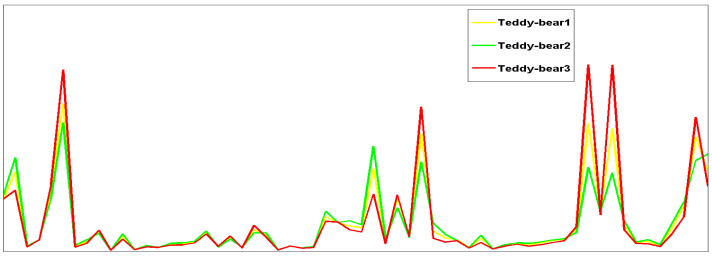
Consistency over articulated shapes In this experiment, we test the performance of the proposed shape descriptor on the deformed shape models. We choose the Teddy-bear and Human models with different poses. The shape descriptors of the deformed shapes are illustrated in Fig. . From the figure one can see that the descriptors of the model with different pose changes are very similar, which demonstrates that the proposed shape descriptor has the potential to consistently represent the shapes with pose changes. On the other hand, the shape descriptors of different models are distinctive. This verifies that the hidden layer features in the proposed discriminative auto-encoder have small within-class variations but large between-class variations.

Resistance to noise By perturbing the vertices of the mesh with various levels of the numerical noise, we will demonstrate that the proposed shape descriptor is robust to noise. The noise, a 3-dimensional vector, is randomly generated from a multivariate normal distribution, $Noise \sim N_3(\mu, R * \Sigma)$, where $\mu = [E[X_1], E[X_2], \cdots, E[X_k]]$ is the 3-dimensional mean vector of the coordinates of all vertices, $\Sigma = [Cov[X_i, X_j]]$ is the 3×3 covariance matrix of all vertices, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, k$, and R denotes the ratio between the variance of noise and variance of the coordinates of the vertices.

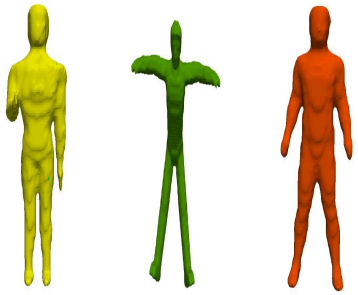
Fig. show the clean Crab and Hand models, and their noisy models (i.e., noisy model 1 and noisy model 2), respectively. The noisy model 1 and noisy model 2 are corrupted by the noise of $R = 0.001$ and $R = 0.04$, respectively. Particularly, in the noisy model 3, geometric structures of the mesh have been moderately deteriorated. As indicated in Fig. , the variations of the proposed shape descriptors of the clean and noisy models (plotted with the yellow, green and red curves, respectively) is small. The test



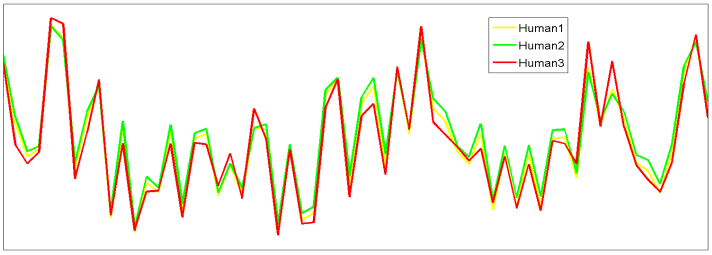
(a) Teddy-bear models:Teddy-bear1, Teddy-bear2, Teddy-bear3.



(b) Descriptors of the Teddy-bear models



(c) Human models:Human1, Human2, Human3



(d) Descriptors of the Human models

Figure 2. Descriptors of the Teddy-bear model and the Human model. The last column shows the descriptors of the shapes, which are plotted by the yellow, green and red curves, respectively.

demonstrates that the proposed shape descriptor formed by the deep discriminative auto-encoder is robust to noise.

4.2. 3D Shape Retrieval Performance

In order to demonstrate effectiveness of our method, we test the proposed shape descriptor on three benchmark datasets of 3D models, i.e., McGill[], SHREC'10 Shape-Google [] and protein [] datasets. Before searching, we pre-calculate shape descriptors of all queries in the database. Therefore, every shape is transformed to a compact 1D vector (e.g. 30-dimension). And we use L_2 norm to compute the distance between the two shape descriptors.

4.2.1 McGill Shape Dataset

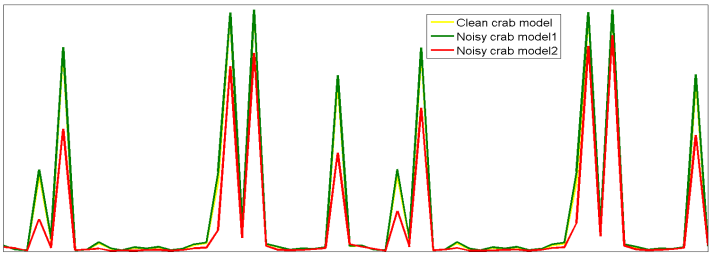
The McGill 3D shape dataset is a challenging dataset, which contains 255 objects with significant part articulations. They are from 10 classes: ant, crab, spectacle, hand, human, octopus, plier, snake, spider and teddy-bear. Each class contains one 3D shape with a variety of pose changes. Fig. shows some examples in the McGill shape dataset.

We compare our proposed method to the state-of-the-art methods: the Hybrid BOW [], the PCA based VLAT method [], the graph-based method [], the hybrid 2D/3D approach [] and covariance descriptor []. We evaluated the proposed method with different performance measures, namely, Nearest Neighbor (NN), the First Tier (1-Tier), the Second Tier (2-Tier) and the Discounted Cumulative Gain (DGC). The retrieval performance of these methods is illustrated in Table 2. From this table, compared to the state-of-the-art methods [], we can see that the proposed method can achieve the best performance on the 4 performance measures. Although there are large nonrigid deformations with the objects in the McGill shape dataset, due to the discriminative feature representation in the hidden layer of the discriminative auto-encoder, our proposed method is robust to nonrigid deformations.

Table 1. Retrieval results on the McGill dataset.

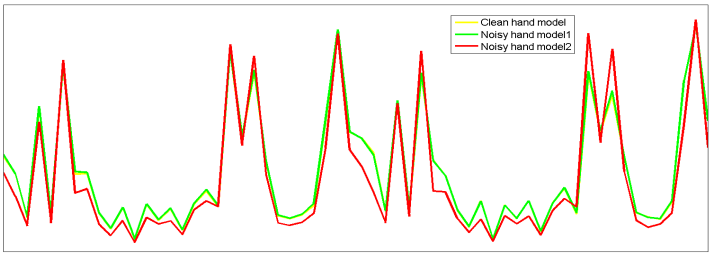
Methods	NN	1-Tier	2-Tier	DCG
Covariance method []	0.977	0.732	0.818	0.937
Graph-based method []				
PCA-based VLAT []				
Hybrid BOW []	0.957	0.635	0.790	0.886
Hybrid 2D/3D []				
Deep shape descriptor				

(a) Clean and noisy models of the shape Crab.



(b) Descriptors of the clean and noisy crab models.

(c) Clean and noisy models of the shape Human.



(d) Descriptors of the clean and noisy human models.

Figure 3. Descriptors of the noisy crab and hand models. The left three columns show the clean model, the noisy model with noise of $R = 0.001$ and the noisy model with noise of $R = 0.04$, respectively. The last column shows the descriptors of the noisy models, which are plotted by the yellow, green and red curves, respectively.

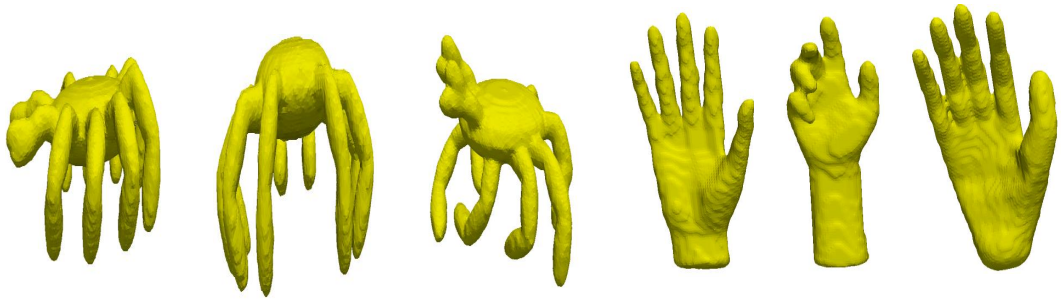


Figure 4. Example shapes in the McGill dataset.

Figure 5. Example shapes in the SHREC'10 ShapeGoogle dataset.

4.2.2 SHREC'10 ShapeGoogle Dataset

SHREC'10 ShapeGoogle dataset [] contains 1184 synthetic shapes. In this dataset, there are 715 shapes from 13 classes

are generated with the five simulated transformations, i.e., isometry, topology, isometry+topology, partiality and triangulation, and there are 456 unrelated distractor shapes. Following the setting in [], in order to make the dataset more challenging, all shapes are re-scaled to have the same size and the samples in the dataset which have the same attribute are considered to be of the same class. For example, male and female shapes are considered to be from the same class. Fig. shows some examples of the ShapeGoogle dataset.

We compared our deep shape descriptor to the bag of feature (BOF) descriptor based on standard vector quantization (VQ) [], sparse coding with unsupervised dictionary learning (DL) [] and sparse coding with supervised DL []. We used the mean average precision criterion to evaluate our proposed method. For each query, the retrieval was performed on the other 54 shapes of the same class and 1105 negative samples. Evaluation results are summarized in Table 3. From this table, one can see that our proposed deep shape descriptor is superior to the BOF descriptors based on standard VQ [], sparse coding with unsupervised DL [] and sparse coding with supervised DL [] in the case of different transformations.

Table 2. Retrieval results on the SHREC'10 ShapeGoogle dataset.

Transformation	VQ []	UDL []	SDL []	DSD
Isometry	0.977	0.732	0.818	0.937
Topology				
Isometry+Topology				
Partiality	0.957	0.635	0.790	0.886
Triangulation				

5. Conclusions

The data-richness of 3D models urges us to focus on robust and intelligent model analysis and understanding. In this paper, We propose a deep shape descriptor with the discriminative auto-encoder for shape matching and retrieval. By imposing the Fisher discrimination criterion on the feature representation in the hidden layer of the auto-encoder, we develop a discriminative auto-encoder so that the feature representation in the hidden layer have small within-class scatter but large between-class scatter. Then, with the multiscale HKS histogram, we train a stacked discriminative auto-encoder to extract all features in the hidden layers to form the deep shape descriptor. The deep shape descriptor demonstrates its performance in various tests for matching and retrieving 3D shapes with deformations, topological short-circuits and numerical noises.

References

[1] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Efficient computation of isometry-invariant distances between surfaces. *SIAM J. Sci. Comput.*, 28:1812–1836, September 2006. 1

[2] A. M. Bronstein, M. M. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro. A gromov-hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Int. J. Comput. Vision*, 89:266–286, 2010. 1

[3] D.-Y. Chen, X.-P. Tian, Y. te Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. *Computer Graphics Forum*, 22:223–232, 2003. 1

[4] X. Chen, A. Golovinskiy, and T. Funkhouser. A benchmark for 3D mesh segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2009. 1

[5] F. De Goes, S. Goldenstein, and L. Velho. A hierarchical segmentation of articulated bodies. *Computer Graphics Forum*, 27:1349–1356, 2008. 1

[6] M. Elad, A. Tal, and S. Ar. Content based retrieval of vrml objects - an iterative and interactive approach. *Proc. Sixth Eurographics Workshop Multimedia*, pages 97–108, 2001. 1

[7] R. Gal, A. Shamir, and D. Cohen-Or. Pose-oblivious shape signature. *IEEE Transactions on Visualization and Computer Graphics*, 13:261–271, 2007. 1

[8] D. Huber, A. Kapuria, R. Donamukkala, and M. Hebert. Parts-based 3d object classification. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:82–89, 2004. 1

[9] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design*, 37(5):509 – 530, 2005. Geometric Modeling and Processing 2004. 1

[10] V. Jain and H. Zhang. A spectral approach to shape-based retrieval of articulated 3d models. *Computer-Aided Design*, 39(5):398 – 407, 2007. Geometric Modeling and Processing 2006. 1

[11] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999. 1

[12] S. Katz, G. Leifman, and A. Tal. Mesh segmentation using feature point and core extraction. *The Visual Computer*, 21:649–658, 2005. 1

[13] R. Ohbuchi, K. Osada, T. Furuya, and T. Banno. Salient local visual features for shape-based 3d model retrieval. *Shape Modeling and Applications, 2008. SMI 2008. IEEE International Conference on*, pages 93–102, 2008. 1

[14] R. Osada, T. Funkhouser, B. Chazelle, and D. Dokin. Shape distributions. *ACM Transactions on Grphics*, 33:133–154, 2002. 1

[15] M. Ovsjanikov, A. Bronstein, and M. Bronstein. Shape google: a computer vision approach to invariant shape retrieval. *Proc. NORDIA*, 2009. 1, 2

[16] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. *Proceedings of the fifth Eurographics symposium on Geometry processing*, pages 225–233, 2007. 1

[17] D. Saupe and D. V. Vranic. 3d model retrieval with spherical harmonics and moments. *DAGM*, pages 392–397, 2001. 1

[18] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. *In Shape Modeling International*, pages 167–178, 2004. 1

[19] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *SGP '09: Proceedings of the Symposium on Geometry Processing*, pages 1383–1392, 2009. 1

[20] J. W. H. Tangelder and R. C. Veltkamp. A survey of content based 3d shape retrieval methods. *In Shape Modeling International*, pages 145–156, 2004. 1

[21] D. V. Vranic, D. Saupe, and J. Richter. Tools for 3d-object retrieval: Karhunen-loeve transform and spherical harmonics. *IEEE MMSP 2001*, pages 293–298, 2001. 1