Heat Diffusion Long-Short Term Memory Learning for 3D Shape Analysis

Fan Zhu, Jin Xie, and Yi Fang(✉)

NYU Multimedia and Visual Computing Lab,
Department of Electrical and Computer Engineering,
New York University Abu Dhabi, Abu Dhabi, UAE
{fan.zhu,jin.xie,yfang}@nyu.edu

Abstract. The heat kernel is a fundamental solution in mathematical physics to distribution measurement of heat energy within a fixed region over time, and due to its unique property of being invariant to isometric transformations, the heat kernel has been an effective feature descriptor for spectral shape analysis. The majority of prior heat kernel-based strategies of building 3D shape representations fail to investigate the temporal dynamics of heat flows on 3D shape surfaces over time. In this work, we address the temporal dynamics of heat flows on 3D shapes using the long-short term memory (LSTM). We guide 3D shape descriptors toward discriminative representations by feeding heat distributions throughout time as inputs to units of heat diffusion LSTM (HD-LSTM) blocks with a supervised learning structure. We further extend HD-LSTM to a cross-domain structure (CDHD-LSTM) for learning domain-invariant representations of multi-view data. We evaluate the effectiveness of both HD-LSTM and CDHD-LSTM on 3D shape retrieval and sketch-based 3D shape retrieval tasks respectively. Experimental results on McGill dataset and SHREC 2014 dataset suggest that both methods can achieve state-of-the-art performance.

Keywords: 3D shape retrieval · Recurrent neural network · Long-short term memory · Heat kernel signature

1 Introduction

Researches on 3D-meshed surface models have been receiving exponentially increasing attentions with the sustainability growing expectations on virtual reality, which is believed to be the revolutionary technology that can completely reshape our lives. In fact, virtual reality isn't exclusive for gaming anymore, it has already sprawled into many areas. For example, virtual reality movies are becoming the mainstream with Hollywood directors. Since the virtual world is established in a 3D space, researchers have been paying efforts to the development of multiple areas of 3D computer vision, which covers 3D correspondence, 3D shape retrieval, 3D segmentation, *etc.* The performance of these 3D analysis systems heavily rely on the quality of 3D shape representations, thus how to

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 305–321, 2016.
DOI: 10.1007/978-3-319-46478-7_19

effectively describe a 3D shape in machine language is of premier importance for 3D shape analysis.

Popular strategies of building 3D shape representations mainly include the projection-based approaches and the heat kernel-based approach. Intuitively, the projection-based approaches aim to transform the 3D shape representation problem into a well developed-image representation problem by projecting a 3D shape from multiple viewpoints and consequently obtaining multiple projection images, where either handcrafted features (*e.g.,* scale invariant feature transform (SIFT) [25]) or deep learning features (*e.g.,* convolutional neural networks (CNN) [22]) are used to represent these projection images. On the other hand, the heat kernel-based approach estimates geometrical relationships between 3D mesh points throughout sequential diffusion time. A typical example of the heat kernel-based 3D shape representation is the heat kernel signature (HKS) [34]. Due to the unique property of the heat kernel, HKS is invariant to geometrical transformations, however, the temporal information along the heat diffusion time has not been utilized by HKS.

In this work, we aim to develop a new 3D shape representation by utilizing the heat flows on 3D shape surfaces and the corresponding temporal dynamics of the heat flows within the diffusion period. Inspired by the advancements of deep learning techniques, *e.g.,* CNN [22] and recurrent neural networks (RNN), we learn the temporal dynamics of heat flows using the long-short term memory (LSTM) [15]. While RNN can in principle learn sequential data by storing information of a recent time point with the internal memory, LSTM, as a special type of RNN, is equipped with architecture that is capable of storing "long-term" memories in addition to storing "short-term" memories. Thus, by learning the heat flows with LSTM, we are able to extract joint information between diffusion time-steps that are either consecutive or with a large interval. Figure 1 illustrates the pipeline of the HD-LSTM learning framework. We start by computing the heat kernel features (*i.e.,* HKS) from 3D shapes, and learn the heat diffusion kernel distributions (shown in Fig. 1) overall all sampling time-steps through HD-LSTM, where the heat diffusion kernel distribution is the histogram of heat diffusion values given a fixed time-step. We then guide the input features towards discriminative 3D shape representations through a supervised LSTM learning structure, where the category information of training samples are supplied to the output end of LSTM in the form of discriminative vectors. When the heat flows sequentially pass through the HD-LSTM, its "forget gate layer" can selectively throw away the previous heat flow from the cell state, and determine how much we decide to update the current state value using the past data. Benefiting from the easy generalization property of HD-LSTM, we extend HD-LSTM to a cross-domain learning structure CDHD-LSTM, which minimizes the cross-domain discrepancy by connecting HD-LSTM to a 3-layer neural network and guiding same-category cross-domain data toward identical targets. Our contributions are threefolds:

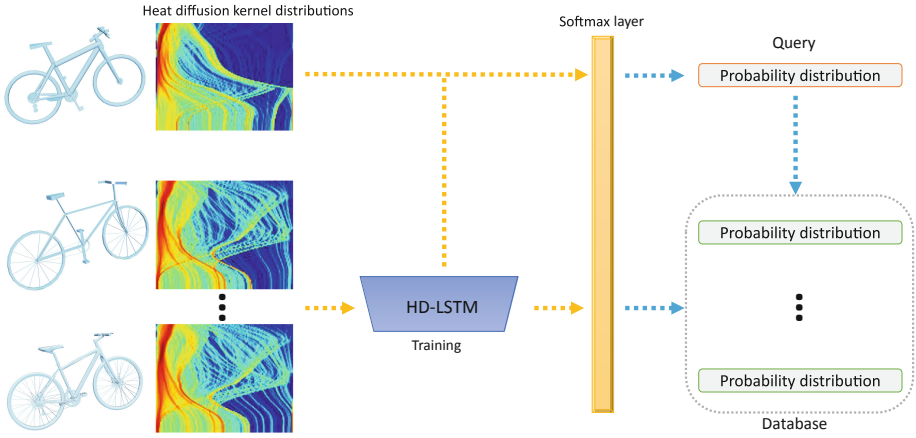

Fig. 1. We discover the temporal dynamics of heat diffusions and correspondingly propose HD-LSTM to learn discriminative 3D shape representations based on heat diffusions.

- We explore the temporal dynamics of heat flows over multiple diffusion time-steps, and we propose a novel deep learning 3D shape representation by learning sequential heat kernel features using HD-LSTM with a supervised structure.
- We extend the supervised HD-LSTM structure to a cross-domain setting, and propose a cross-domain deep learning strategy CDHD-LSTM for learning domain-invariant representations to address the sketch-based 3D shape retrieval problem.
- We conduct experiments on both 3D shape retrieval and sketch-based 3D shape retrieval tasks to evaluate the effectiveness of HD-LSTM and CDHD-LSTM. Experimental results demonstrate that both methods can achieve state-of-the-art performance on popular benchmarks.

2 Related Works

The challenges for developing 3D shape representations include the complexity of 3D models [35], structural variations of 3D models [5], noise, *etc.* There are extensive investigations in the literature on the topic of building effective 3D shape representations to address these challenges. Early approaches mainly rely on "handcrafted" features. One classical strategy is characterizing neighboring point signatures through shape distributions, including spin images [18] and shape context [2], which are both invariant under rigid shape transformations. Another approach is the point-based signature, which characterizes vertexes on 3D shape surfaces with vectors. Popular point-based signatures include the global point signature (GPS) [28], whose vector components are obtained from scaled

eigenfunctions of the Laplace-Beltrami operator, and HKS [34], which is obtained by computing histograms of heat diffusions on shape surfaces. Some other shape descriptors are designed based on geodesic distances [11,14]. Intuitively, geodesic distance-based methods are invariant under isometric deformations, however, they are sensitive to topological noise. While the aforementioned methods operate directly on native 3D shapes (*e.g.,* polygon meshes and point clouds), some 3D shape representations are extended from well-established image representation techniques, including the extension of SURF feature to 3D voxel grids [21] and the extension of SIFT feature to represent 2D projection images of 3D shapes [7] and some recent work that employ CNN to perform deep learning on 2D projection images [33,36]. In general, existing 3D shape representation learning approaches are based on the following taxonomy: (1) volumetric methods, *e.g.,* Wu *et al.* [38] and Sedaghat *et al.* [29]; (2) 2D projection methods, *e.g.,* Maturana *et al.* [26], Su *et al.* [33] and Shi *et al.* [31]; and (3) shape distribution methods *e.g.,* DeepShape [39], where our approach in this work belongs to the third category. Beyond the regular 3D shape retrieval task, some studies, including Su *et al.* [33], Wang *et al.* [36] and Zhu *et al.* [41], attempt to build domain-invariant 3D shape representations to directly compare 3D shapes with hand-drawn sketches.

With the inspiring victory of AphaGo [32] over the world champion Go player Lee Sedol in recent days, the deep learning technology is stepping in public's eyes in a real sense. As one of the important deep learning techniques behind the victory of AlphaGo, CNN has already achieved many revolutionary successes in a wide range of applications, *e.g.,* image classification [22], recommender systems [27] and human action recognition [16]. Also, some recent work employ CNN to perform deep learning on 2D projection images of 3D shapes, so as to develop 3D shape representations [33,36]. Different from feed-forward neural networks, RNN builds an internal state that allows cycled signal flows within the neural network. Benefiting from such a property, RNN is applicable of dealing with sequential data, *e.g.,* speech classification [13] and caption generation [20]. However, RNN is practically hard to train due to the vanishing gradient problem [3]. In addition, RNN is incapable of dealing with long-term dependencies with a standard structure. The LSTM architecture, as a special type of RNN, can avoid the vanishing gradient problem by performing gradient descent with back-propagation through time [15], and it is also capable of learning long-term dependencies. LSTM has demonstrated its capability for learning sequential data in tasks such as image caption generation [17] and action recognition [9].

Our approach is to utilize the favorable sequential data learning capability of the LSTM architecture, aiming to explore the temporal dynamics of diffusion flows on 3D shape surfaces. We guide LSTM with a supervised learning structure, so that the learned 3D shape representations are discriminative. To our knowledge, this is the first work that attempts to learn 3D shape representations with a LSTM architecture.

3 Heat Diffusion Long-Short Term Memory

3.1 Heat Diffusion on 3D Shape Surfaces

We start by revisiting some preliminary knowledge on the Laplace-Beltrami operator, the heat operator and heat kernel [34]. Let $\mathcal{M}$ denote a Riemannian manifold, the heat diffusion process is governed by the Laplace-Beltrami operator of Δ_M:

$$\Delta_{\mathcal{M}}\mu(u,t) = -\frac{\partial\mu(u,t)}{\partial t}. \tag{1}$$

It is verified that the Laplace-Beltrami operator of $\Delta_{\mathcal{M}}$ and the heat operator H_t satisfy the following relation:

$$H_t = e^{-t\Delta_{\mathcal{M}}}. \tag{2}$$

Since both operators share the same eigenfunctions, if we denote λ as an eigenvalue of the Laplace-Beltrami $\Delta_{\mathcal{M}}$ corresponding to a eigenfunction, $e^{\lambda t}$ is an eigenvalue of the heat operator H_t corresponding to the same eigenfunction. The heat kernel $k_t(u,v)$ is introduced to measure the amount of heat that has been transformed from point u to point v on the 3D shape surface at time t. Given an initial heat distribution $f : \mathcal{M} \to \mathbb{R}$, for any $\mathcal{M}$, there exists the following relation between the heat kernel $k_t(u,v)$ and the heat operator H_t:

$$H_t f(u) = \int_M k_t(u,v)f(v)dv, \tag{3}$$

where dv is the volume form at $v \in \mathcal{M}$. Assuming the Riemannian manifold $\mathcal{M}$ is compact, the heat kernel can then be expressed in the form of its eigen-decomposition:

$$k_t(u,v) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(u)\phi_i(v), \tag{4}$$

which can then be used to compute the HKS of each vectex u on the 3D shape surface at time t:

$$\begin{aligned} S_t(u) &= k_t(u,u) \\ &= \sum_{i=0} e^{-\lambda_i t}\phi_i(u)^2, \end{aligned} \tag{5}$$

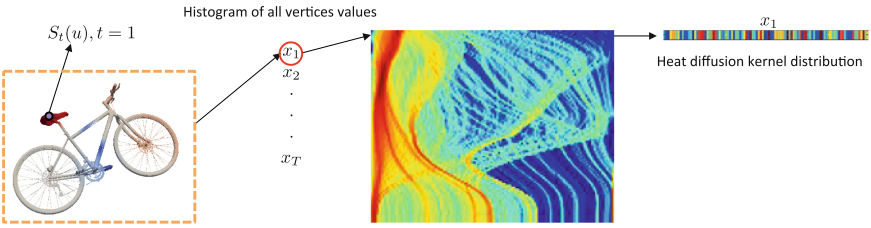

Fig. 2. Illustration of the heat diffusion kernel distribution of a 3D shape. (Color figure online)

where $S_t(u)$ is defined as the diagonal of the heat kernel $k_t(u, v)$. We then use heat diffusion kernel distribution x_t of HKS values $S_t(u)$ for all vectices at the diffusion time t. Figure 2 illustrates how the heat diffusion kernel signature values on the surface of a *bicycle* model can be computed, and how the heat diffusion kernel distribution can be correspondingly obtained from HKS values. Given the time-step $t = 1$, a red point on the 3D shape surface denotes a high HKS value, where a high HKS value is equivalent to the "corner" point, which contains the most valuable information within the neighboring vertices. The heat diffusion kernel distribution x_1 at the time step $t = 1$ can then be obtained by projecting all HKS values $S_1(u), \forall u$ onto a histogram.

3.2 Learning Heat Diffusion with Long-Short Term Memory

Extended from the original LSTM architecture, LSTM has some variants, including the "peephole" architecture [12] and the gated recurrent unit architecture [6]. In this work, we propose to learn heat kernel probability distributions over multiple diffusion time steps using HD-LSTM, which is designed based on the basic LSTM architecture. A memory cell of LSTM contains four main components, including an input gate, a self-recurrent neuron, a forget gate and an output gate. When we feed the heat diffusion kernel distribution x_t as the input to HD-LSTM, the activation at the input gate, the candidate value $\hat{C}_t$ and the activation at the memory cell can be computed as:

$$I_t = \sigma(W_I x_t + U_I h_{t-1} + b_I), \tag{6}$$

$$\hat{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{7}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{8}$$

where $\sigma(\cdot)$ is a sigmoid layer that determines how much information are going through this layer and outputs values $O_\sigma \in (0, 1]$, and the $tanh(\cdot)$ layer outputs values $O_{tanh} \in (-1, 1)$. The forget gate determines the new cell state C_t by deciding how much information of the earlier heat diffusion kernel distributions should be forgotten. Given the values of the input gate activation i_t, the forget gate activation f_t and the candidate value $\hat{C}_t$, the new cell state C_t can be obtained using:

$$C_t = f_t * C_{t-1} + I_t * \hat{C}_t. \tag{9}$$

The output gate value o_t can then be obtained based on the input heat diffusion kernel histogram x_t, the hidden layer value at the previous time step h_{t-1} and the updated cell state value C_t through:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o), \tag{10}$$

and the new hidden layer value h_t can be computed using:

$$h_t = o_t * tanh(C_t). \tag{11}$$

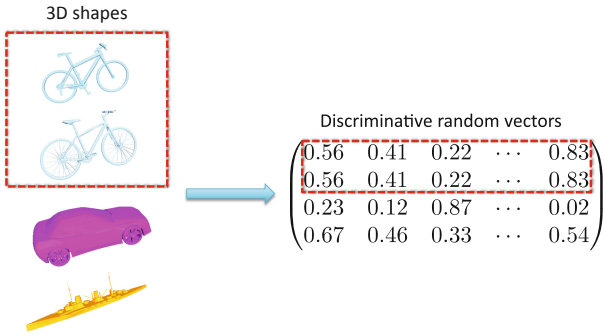

Fig. 3. Generating discriminative random vectors from groundtruth labels of 3D shapes. Red dashed rectangular denote 3D shapes that belong to the same object category and their corresponding identical discriminative random vectors. (Color figure online)

In above formulations, W_I, W_c, W_f, W_o, U_I, U_c, U_f, U_o and V_o are weight parameters of the model, and b_I, b_f, b_c and b_o are bias vectors.

In order to guide HD-LSTM toward learning discriminative 3D shape representations, we transform groundtruth labels of training 3D shapes into the form of discriminative vectors Y, and assign these vectors to the hidden layer unit h_t at each time step t. 3D shapes that belong to the same category will be assigned identical discriminative vectors at the outputs of hidden units, so that HD-LSTM can encourage the intra-class distance of learned 3D shape representations to be low. Figure 3 shows how discriminative vectors are generated based on the groundtruth category information of training 3D shapes. The top two 3D shapes on the left side are both *bicycles*, thus they are mapped to identical vectors within the red dashed rectangular. Experimental results suggest that using random values for entries of the discriminative vectors can lead to good performance. Previous investigations [40] also demonstrated the effectiveness of using random vectors. In the training phase, HD-LSTM minimizes the reconstruction error between discriminative vectors Y and the hidden unit outputs h_t through time:

$$\arg\min_{W,U,V,b} \sum_{i=1}^{N} \sum_{t=1}^{T} \|Y^i - h_t^i\|_2^2, \tag{12}$$

where N is the total number of training samples, T is the total number of sampling time steps of the heat diffusion kernel on 3D shape surfaces and W, U, V, b are abbreviated forms of above defined weight parameters and bias vectors of the HD-LSTM model. Once we obtain the optimal values of $\hat{W}$, $\hat{U}$, $\hat{V}$ and $\hat{b}$, the output of hidden unit $\hat{h_t^i}$ at each time step can be considered as a discriminative representation of the 3D shape i. We then train a softmax layer [4] using outputs of hidden units of all training samples ($i \in [1, N]$) through all heat diffusion kernel time steps ($t \in [1, T]$), so that the predicted probability P_t^i for the j-th class of the output unit $\hat{h}_t^i$ (while $\hat{h}_t^i$ corresponds to heat diffusion

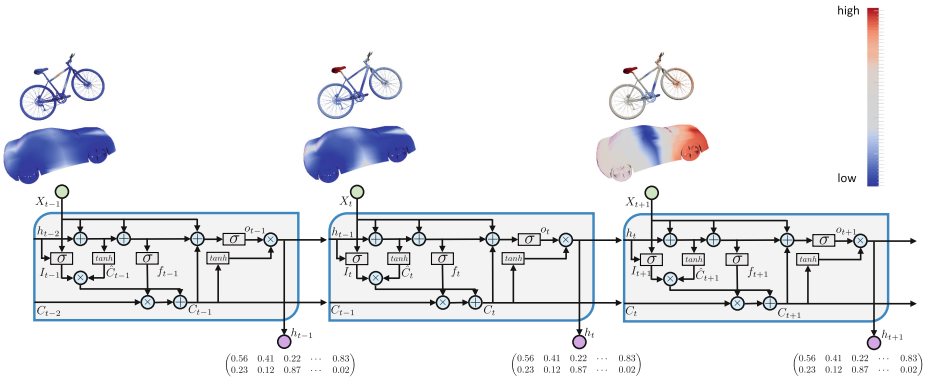

Fig. 4. Learning HD-LSTM from temporal dynamics of heat diffusion on 3D shape surfaces. We show heat diffusion kernel values on the 3D shape surfaces in consecutive 3 time steps, where the red points denote high values and blue points denote low values. The sequential inputs to HD-LSTM are histograms of heat diffusion kernel values at each time step. (Color figure online)

kernel distribution x_t^i at the input gate of HD-LSTM) can be computed through:

$$P_t^i(y = j|h_t^i) = \frac{e^{\hat{h_t^i}^T w_j}}{\sum_{k=1}^{K} e^{\hat{h_t^i}^T w_k}}. \tag{13}$$

P_t^i is a J-dimensional vector, where J equals to the number of classes of the dataset. Finally, in order to obtain a global representation of a 3D shape i, we compute the average of P_t^i through all time steps:

$$P^i = \frac{1}{T} \sum_{t=1}^{T} P_t^i. \tag{14}$$

3.3 3D Shape Retrieval

We evaluate the performance of HD-LSTM on the 3D shape retrieval task, where retrieval is conducted by computing the dissimilarity matrix D' between the query 3D shapes P_q and the database 3D shapes P_d based on L_2 norm using Euclidean distance:

$$D'_{ij} = \sqrt{(P_q - P_d)^2}. \tag{15}$$

4 Cross-Domain Heat Diffusion Long-Short Term Memory

By guiding LSTM with discriminative vectors at the outputs of hidden units, a favorable property of HD-LSTM is that it can be easily generalized to learning

multi-view data by connecting HD-LSTM to another neural network through discriminative vectors. Thus, we further propose a CDHD-LSTM architecture to address the sketch-based 3D shape retrieval problem [24] based on HD-LSTM. Inspired by some recent work that represent sketches using CNN [33], in this work, we consider each sketch as an image and compute the CNN feature for each sketch using a per-trained CNN [30]. We denote $X_s = \left\{x_s^1, x_s^2, \cdots, x_s^M\right\}$ as CNN features of M training sketches. In order to map both CNN features of sketches and heat diffusion kernel distributions of 3D shapes into a unified feature space, we establish a bridge between both domains by connecting a 3-layer neural network to the output units of HD-LSTM, where the 3-layer neural network contains an input layer, a hidden layer and a target layer. The input layer takes CNN sketch features as inputs to the neural network, and we follow the same strategy as in HD-LSTM to assign discriminative vectors to the target layer. The 3-layer neural network and HD-LSTM can be connected by assigning identical discriminative vectors Y^i for data that come from the same category. More specifically, learning the 3-layer neural network can be achieved by minimizing the reconstruction error between at the target layer:

$$\arg\min_{W,b} \frac{1}{M} \sum_{i=1}^{M} \|Y^i - \sigma_{W^l,b^l}(x_s^i)\| + \varphi \sum_{l=1}^{L} \|W\|_F^2, \tag{16}$$

where $W = \left\{W^1, W^2, \cdots, W^L\right\} \in \mathbb{R}^{P \times L}$ is the neuron parameters of the 3-layer neural network, b is the bias neuron value, P is the number of neurons and L is the number of layers. Once the optimal values of W and b are obtained, we extract neuron values r^i at the target layer when a query sketch x_s^i from the testing set passes through the 3-layer neural network. Consider a sketch sample x_s^i and a 3D shape sample x_t^i that belong to the same object class c', since both of the cross-domain samples are mapped towards the same discriminative vector $Y^{c'}$, the data smoothness can be preserved between the learned sketch representation r^i and the learned 3D shape representation h_t^i. Similar as the strategy adopted in HD-LSTM, we jointly train a softmax layer using both learned sketch representations $r^i, \forall i \in [1, M]$ and learned heat diffusion $h_t^i, \forall i \in [1, N], t \in [1, T]$. For a sketch-based 3D shape retrieval system, the predicted probability histograms P_s are representations for query sketches, while the mean of predicted probability histograms P_d are representations for 3D shapes in the database. The architecture of CDHD-LSTM is illustrated in Fig. 5. Note that Y is a generalized interpretation of the discriminative vectors. In fact, while the discriminative vectors for 3D models Y_m and the discriminative vectors for sketches Y_s have an identical dimension K', the vector numbers are very likely to be different (*i.e.,* $M \neq N$). The optimization of the 3-layer neural network is separate from HD-LSTM, and can be implemented using the commonly used backpropagation algorithm [37].

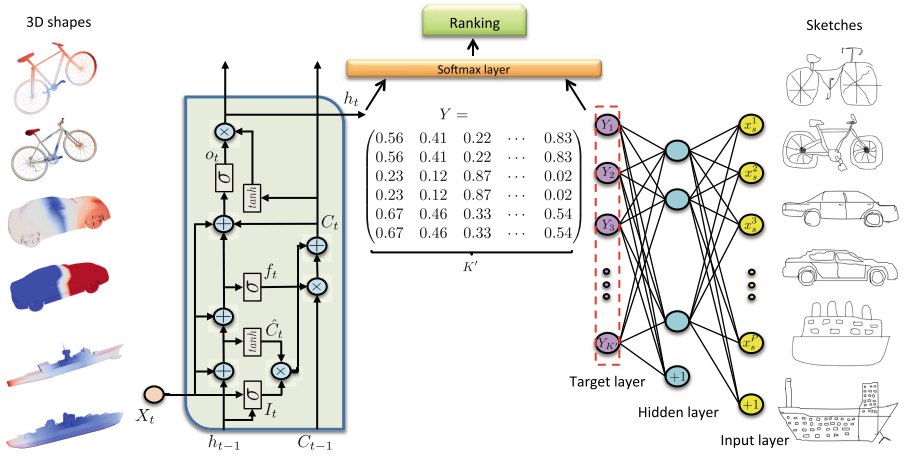

Fig. 5. Learning domain-invariant representations for sketch-based 3D shape retrieval using the CDHD-LSTM architecture. CDHD-LSTM is constructed by connecting a 3-layer neural network to HD-LSTM at the output ends, where the connection is established by sharing identical discriminative random vectors for sketches and 3D shapes that come from the same category.

5 Experiments

5.1 3D Shape Retrieval

In order to demonstrate the effectiveness of the proposed HD-LSTM method, we conduct experiments on 3D shape retrieval tasks using the McGill dataset. The 5 commonly used evaluation metrics, nearest neighbor (NN), first tier (1-Tier), second tier (2-Tier), discounted cumulated gain (DCG) and average precision (AP) are used for evaluating the performance of the proposed methods and comparison methods.

McGill shape dataset: The McGill dataset contains 255 objects with significant partial deformations. These objects come from 10 object categories, including *ant*, *crab*, *spectacle*, *hand*, *human*, *octopus*, *plier*, *snake*, *spider* and *teddy bear*, where each object category contains 3D shapes with a wide range of pose variations. We conduct the retrieval experiment by randomly choosing 10 shapes per class to train HD-LSTM while using the remaining shapes as query data.

We empirically set the dimension of the discriminative random vectors K' as 120 and the learning rate as 0.1. When computing the heat diffusion kernel descriptors (*i.e.,* HKS) of 3D shapes, the universal time unit τ and the total heat diffusion sampling time step value T are defined as 0.01 and 101 respectively. HKS values on 3D shape surfaces are projected to 128-dimensional histograms. We set the maximum iteration of LSTM to 50. As illustrated in Fig. 6, HD-LSTM normally can converge to a low reconstruction error between 20 $\sim$ 30 iterations on McGill dataset.

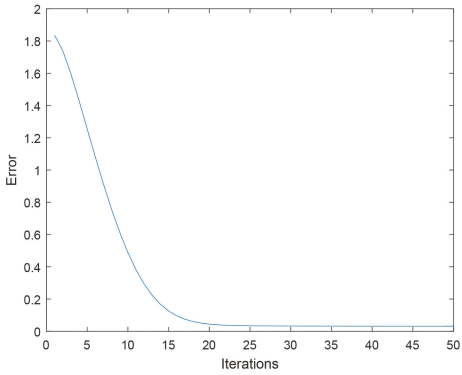

Fig. 6. Convergence of the reconstruction error when training HD-LSTM on the McGill shape dataset.

We use the 128-dimensional heat diffusion kernel distributions over 101 time steps as local features for each 3D shape, and construct Bag-of-Words (BoW) as a baseline by projecting the local features onto a dictionary, which contains 120 dictionary atoms. When evaluating the performance of the proposed HD-LSTM approach, we show experimental results of the cases when we use the softmax layer to obtain category probabilities for retrieval and when we directly use outputs of hidden units for retrieval. We also show comparisons with state-of-the-art methods, including the Hybrid BoW [23], the covariance method [35], the graph-based method [1] and the DeepShape method [39]. The retrieval performance of the proposed HD-LSTM method and state-of-the-art methods are illustrated in Table 1. Among the methods that HD-LSTM compares to, DeepShape [39] models HKS features with heat diffusion kernel distribution and learns discriminative shape representations with Autoencoder, while [35], [23] and [1] are based on the covariance, the bag-of-words model and graph matching of local shape descriptors respectively. It can be observed that the performance of the baseline BoW method is relatively poor when comparing with other approaches. By learning the dynamics of the heat diffusion on 3D shapes, the temporal information are utilized through the learning process. Also, the supervised HD-LSTM approach can enhance the discriminative power of shape representations. Consistent with our assumption, experimental results suggest that the proposed HD-LSTM method achieves a leading performance, and when incorporating with the softmax layer the performance of HD-LSTM can be further improved.

5.2 Sketch-Based 3D Shape Retrieval

We evaluate the performance of CDHD-LSTM on the sketch-based 3D shape retrieval task using the extended large scale SHREC 2014 dataset [24], which contains 13,680 2D sketch image queries of 171 object classes (with an identical number of 80 sketches for each class) from the human sketch recognition dataset

Table 1. Performance comparison between the proposed HD-LSTM method and the state-of-the-art methods on the McGill dataset.

Methods	NN	1-Tier	2-Tier	DCG	AP
Hybrid BoW [23]	0.95	0.63	0.79	0.88	–
Covariance method [35]	0.97	0.73	0.81	0.93	–
Graph-based method [1]	0.97	0.74	0.91	0.93	–
DeepShape [39]	**0.98**	0.78	0.83	–	–
BoW	0.80	0.40	0.54	0.70	0.46
HD-LSTM (without softmax)	0.97	0.88	0.83	0.88	0.90
HD-LSTM (with softmax)	**0.98**	**0.92**	**0.95**	**0.95**	**0.94**

[10] and 8,987 3D shapes of corresponding 171 object classes from a combination of multiple 3D datasets. The sketch data contains a training split of 8,550 sketches and a testing split of 5,130 sketches. When training the CDHD-LSTM, we use the training split of sketch data and all 3D shapes in the database. In the sketch-based 3D shape retrieval phase, the testing split of sketch data are used as queries and all 3D shapes are considered as the database.

A unique property of the SHREC 2014 3D shape dataset is the numbers of 3D shapes are highly unbalanced across different categories that the number in each class can vary from 1 to 632. Thus, we follow [24] to evaluate the performance of the sketch-based 3D shape retrieval system based on the reciprocally weighted evaluation metric. Specifically, a reciprocal weight is assigned to each query instance based on the number of available 3D shapes that belong to the same category as the query. Assuming a sketch query z belong to class $l_q(z)$, the weight $\hat{w_r}(z)$ assigned to the retrieval result in response to query z can be defined as:

$$\hat{w}_r(z) = \frac{1}{p_z}, \tag{17}$$

where p_z indicates the number of available 3D shapes that belong to class $l_q(z)$. The 5 evaluation metric scores, NN, FT, ST, DCG and AP are obtained by further dividing another global weight $\hat{w_g}$, which can be computed using:

$$\hat{w}_g = \sum_{z=1}^{Z} \frac{1}{p_z}, \tag{18}$$

given that Z is the total number of sketch queries.

We use the pre-trained CNN on ImageNet [8] to extract sketch features at the sketch input end of CDHD-LSTM. Each sketch image is first resized to 231×231 pixels as the input to CNN [30], which contains 5 convolutional layers and 2 fully connected layers, where values of 4096 neurons in the 7-th layer are extracted as the 4096-dimensional sketch image representation. In order to reduce the computational cost, we perform dimensionality reduction on 4096-dimensional sketch CNN features using PCA [19] and reduce the dimension to 100. We use

the same diffusion time steps and the number of heat diffusion kernel distribution bins as in HD-LSTM, and we set the discriminative dimension and learning rate as 100 and 0.01 respectively.

Table 2. Reciprocally weighted performance metrics comparison on different datasets of the extended large-scale SHREC'14 benchmark for the Query-by-Sketch retrieval.

Method	NN	FT	ST	DCG	AP
		$1.0e-05*$			
BF-fGALIF	0.43	0.27	0.41	2.03	0.34
CDMR ($\sigma_{\mathrm{SM}} = 0.1$, $\alpha = 0.6$)	0.18	0.14	0.22	0.12	0.15
CDMR ($\sigma_{\mathrm{SM}} = 0.1$, $\alpha = 0.3$)	0.38	0.25	0.38	0.18	0.30
CDMR ($\sigma_{\mathrm{SM}} = 0.05$, $\alpha = 0.6$)	0.33	0.27	0.40	0.18	0.31
CDMR ($\sigma_{\mathrm{SM}} = 0.05$, $\alpha = 0.3$)	0.44	0.30	0.45	0.20	0.36
SBR-VC ($\alpha = 1$)	0.25	0.14	0.26	1.86	0.19
SBR-VC ($\alpha = 0.5$)	0.25	0.15	0.27	1.87	0.19
OPHOG	0.52	0.29	0.45	2.08	0.34
SCMR-OPHOG	0.52	0.39	0.61	2.17	0.49
BOF-JESC (VQ = 800)	0.33	0.14	0.26	1.88	0.22
BOF-JESC (VQ = 1000)	0.31	0.13	0.20	1.82	0.18
BOF-JESC (FV)	0.32	0.14	0.19	1.74	0.15
HD-LSTM	0.28	0.14	0.22	0.33	0.29
CDHD-LSTM (without softmax)	0.86	0.44	0.93	3.33	0.68
CDHD-LSTM (with softmax)	**0.91**	**0.54**	**1.03**	**3.37**	**0.75**

In order to demonstrate the cross-domain learning capability of CDHD-LSTM, we learn HD-LSTM on 3D shapes of the SHREC 2014 dataset and perform 3D shape retrieval without learning sketch features using the 3-layer neural network (since the dimension of sketch features is 300, the distances between sketch queries and 3D shapes in the database can be directly computed). We denote the brutal cross-domain retrieval method as HD-LSTM. Similar as the regular 3D shape retrieval, both experimental results of CDHD-LSTM when using and not using the softmax layer are shown in Table 2. We compare with the-state-of-the-arts methods, bag-of-features of dense SIFT (BF-DSIFT), cross-domain manifold ranking (CDMR), shape context matching (SBR-VC), overlapped pyramid of histograms of oriented gradients (OPHOG), similarity constrained manifold ranking-overlapped pyramid of histograms of oriented gradients (SCMR-OPHOG) and bag-of-features junction-based extended shape context (BOF-JESC) [24]. The Precision-Recall curve of the proposed method and the state-of-the-art methods are shown in Fig. 7.

It can be observed that the performance of the original HD-LSTM is weak due to the high cross-domain discrepancy between the 3D shape representations

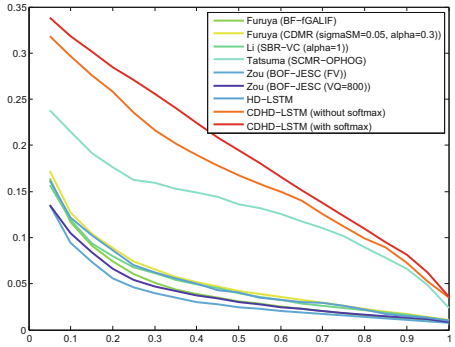

Fig. 7. Precision-Recall plot of performance comparisons on the extended large-scale SHREC 14 sketch-based 3D shape retrieval dataset.

and the sketch representations. After performing CDHD-LSTM learning, same-category instances of both 3D shapes and sketches are encouraged to map to identical target, so that the data smoothness can be preserved within the new feature space. As shown in Table 2, the CDHD-LSTM method can achieve significant improvements over HD-LSTM, and it also outperforms the state-of-the-art methods. Similar as in the regular 3D shape retrieval experiment, improved performance can be observed when incorporating CDHD-LSTM with a softmax layer for retrieval.

6 Conclusions

In this work, we explored temporal dynamics of heat diffusion kernel distributions, and thus proposed to learn novel 3D shape representations by utilizing relationships between different heat diffusion sampling time steps. Based on the sequential data learning method LSTM, we propose a supervised learning structure HD-LSTM that learns discriminative 3D shape representations by guiding the heat diffusion kernel distributions toward discriminative random vectors at the outputs of hidden units. Employing the generalization capability of HD-LSTM, we further propose a CDHD-LSTM structure for learning domain-invariant representations by connecting the output end of HD-LSTM to a 3-layer neural network. Since cross-domain data that belong to the same category are guided to approach an identical discriminative vector, the data smoothness within learned representations can be preserved. We evaluated the effectiveness of HD-LSTM and CDHD-LSTM structures on the regular 3D shape retrieval task and the sketch-based 3D shape retrieval task respectively. Experimental results on the MacGill shape dataset and the extended SHREC 2014 dataset suggest both HD-LSTM and CDHD-LSTM can achieve state-of-the-art performance.

References

1. Agathos, A., Pratikakis, I., Papadakis, P., Perantonis, S.J., Azariadis, P.N., Sapidis, N.S.: Retrieval of 3D articulated objects using a graph-based representation. In: 3DOR 2009, pp. 29–36 (2009)
2. Belongie, S., Malik, J., Puzicha, J.: Shape context: a new descriptor for shape matching and object recognition. In: Advances in Neural Information Processing Systems, vol. 2, p. 3 (2000)
3. Bengio, Y., Boulanger-Lewandowski, N., Pascanu, R.: Advances in optimizing recurrent networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8624–8628 (2013)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2006)
5. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM J. Sci. Comput. **28**(5), 1812–1836 (2006)
6. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint (2014). arXiv:1406.1078
7. Darom, T., Keller, Y.: Scale-invariant features for 3-D mesh models. IEEE Trans. Image Process. **21**(5), 2758–2769 (2012)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
9. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
10. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. Graph. **31**(4), 44 (2012)
11. Gal, R., Shamir, A., Cohen-Or, D.: Pose-oblivious shape signature. IEEE Trans. Vis. Comput. Graph. **13**(2), 261–271 (2007)
12. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**(10), 2451–2471 (2000)
13. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649 (2013)
14. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3D shapes. In: Annual Conference on Computer Graphics and Interactive Techniques, pp. 203–212. ACM (2001)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
17. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415 (2015)
18. Johnson, A.E.: Spin-images: a representation for 3-D surface matching. Ph.D. thesis, Citeseer (1997)
19. Jolliffe, I.: Principal Component Analysis. Wiley Online Library (2002)

20. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
21. Knopp, J., Prasad, M., Willems, G., Timofte, R., Gool, L.: Hough transform and 3D SURF for robust three dimensional classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 589–602. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15567-3_43
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
23. Lavoué, G.: Combination of bag-of-words descriptors for robust partial shape retrieval. Vis. Comput. **28**(9), 931–942 (2012)
24. Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T., et al.: A comparison of methods for sketch-based 3D shape retrieval. Comput. Vis. Image Underst. **119**, 57–80 (2014)
25. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
26. Maturana, D., Scherer, S.: Voxnet: a 3D convolutional neural network for real-time object recognition. In: IEEE International Conference on Intelligent Robots and Systems, pp. 922–928. IEEE (2015)
27. Van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Advances in Neural Information Processing Systems, pp. 2643–2651 (2013)
28. Rustamov, R.M.: Laplace-Beltrami eigenfunctions for deformation invariant shape representation. In: Eurographics Symposium on Geometry processing, pp. 225–233. Eurographics Association (2007)
29. Sedaghat, N., Zolfaghari, M., Brox, T.: Orientation-boosted voxel nets for 3D object recognition. arXiv preprint (2016). arXiv:1604.03351
30. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint (2013). arXiv:1312.6229
31. Shi, B., Bai, S., Zhou, Z., Bai, X.: DeepPano: deep panoramic representation for 3-D shape recognition. IEEE Sig. Process. Lett. **22**(12), 2339–2343 (2015)
32. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
33. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3D shape recognition. In: IEEE International Conference on Computer Vision, pp. 945–953 (2015)
34. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: Computer Graphics Forum, vol. 28, pp. 1383–1392. Wiley Online Library (2009)
35. Tabia, H., Laga, H., Picard, D., Gosselin, P.H.: Covariance descriptors for 3D shape matching and retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4185–4192 (2014)
36. Wang, F., Kang, L., Li, Y.: Sketch-based 3D shape retrieval using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875–1883 (2015)
37. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proc. IEEE **78**(10), 1550–1560 (1990)

38. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: a deep representation for volumetric shapes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920 (2015)
39. Xie, J., Fang, Y., Zhu, F., Wong, E.: Deepshape: deep learned shape descriptor for 3D shape matching and retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1275–1283 (2015)
40. Zhang, Y., Shao, M., Wong, E., Fu, Y.: Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In: IEEE International Conference on Computer Vision, pp. 2416–2423 (2013)
41. Zhu, F., Xie, J., Fang, Y.: learning cross-domain neural networks for sketch-based 3D shape retrieval. In: AAAI (2016)