

3D Keypoint Detection Based on Deep Neural Network with Sparse Autoencoder

Xinyu Lin, Ce Zhu, Qian Zhang, Yipeng Liu

Abstract

Researchers have proposed various methods to extract 3D keypoints from the surface of 3D mesh models over the last decades, but most of them are based on geometric methods, which lack enough flexibility to meet the requirements for various applications. In this paper, we propose a new method on the basis of deep learning by formulating the 3D keypoint detection as a regression problem using deep neural network (DNN) with sparse autoencoder (SAE) as our regression model. Both local information and global information of a 3D mesh model in multi-scale space are fully utilized to detect whether a vertex is a keypoint or not. SAE can effectively extract the internal structure of these two kinds of information and formulate high-level features for them, which is beneficial to the regression model. Three SAEs are used to formulate the hidden layers of the DNN and then a logistic regression layer is trained to process the high-level features produced in the third SAE. Numerical experiments show that the proposed DNN based 3D keypoint detection algorithm outperforms current five state-of-the-art methods for various 3D mesh models.

Keywords: 3D computer vision, 3D keypoint detection, deep neural network, sparse autoencoder

1 Introduction

Detection of 3D keypoints has been a very popular approach within 3D computer vision for various applications, such as object registration [1], 3D shape retrieval [2], object matching [3], mesh segmentation [4] and simplification [5]. Researchers have proposed various methods to extract 3D keypoints from the surface of 3D mesh models over the last decades. Most of 3D keypoint detection algorithms are based on geometric methods [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Godila and Wagan [6] proposed a method for detecting the 3D salient local features on the basis of voxel grid inspired by the Scale Invariant Feature Transform (SIFT) algorithm [20]. Sipiran and Bustos [7] proposed an effective and efficient extension of the Harris operator [21] for 3D objects. Lee et al. [5] defined mesh saliency in a scale-dependent manner utilizing a center-surround operator on Gaussian-weighted mean curvatures and used it as a measure of regional importance for 3D mesh models. Holte utilized Difference-of-Normals operator to address the problem of detecting 3D keypoints [10]. Castellani et al. [11] proposed a salient point detection algorithm where sparse 3D keypoints are selected robustly by exploiting visual saliency principles on 3D mesh models. Besides of the methods mentioned above, there are other methods based on Laplacian spectrum [3, 22, 23], which extract salient geometric feature points in Laplace-Beltrami spectral domain instead of spatial domain.

As is described in [24], using geometric methods to detect 3D keypoints lacks enough flexibility to meet the requirements for various applications because of the following three reasons: 1) Different tasks have different requirements for 3D keypoint detection algorithm: high recall is necessary in some tasks while others may require high precision [25]. 2) Geometric methods usually assume that the vertices with sharp changes in the 3D models are 3D keypoints, but in fact, these vertices may be noise or local variation. 3) Using geometric methods encounters various difficulties when

*All the authors are with School of Electronic Engineering / Center for Robotics, University of Electronic Science and Technology of China (UESTC), Chengdu, China. (email: xinyulin@std.uestc.edu.cn; eczhu@uestc.edu.cn; zhangqian@std.uestc.edu.cn; yipengliu@uestc.edu.cn)

semantic ambiguity is considered. All of these reasons drive researchers to find a new framework to detect 3D keypoints.

In recent years, some researchers proposed 3D keypoint detection algorithms based on machine learning [24, 26, 27], which could solve these problems mentioned above to some extent. But most of them only utilized local information to detect 3D keypoints, lacking corresponding global information such as Laplacian spectrum. Teran and Mordohai [24] proposed a 3D keypoint detection algorithm using a random forest [28] as the classifier, where several geometric detectors are used to produce attributes. Creusot et al. [26] utilized a linear method, namely Linear Discriminant Analysis and a non-linear method, namely AdaBoost [29] to detect 3D keypoints from 3D face scans. Salti and Tombari [27] cast 3D keypoint detection as a binary classification between points whose support can be correctly matched by a pre-defined 3D descriptor (SHOT descriptor [30]) or not, and the same with [24], random forest [28] is used as the classifier.

In this paper, we propose a new 3D keypoint detection algorithm on the basis of deep learning. Here we formulate the 3D keypoint detection as a regression problem using deep neural network (DNN) with sparse autoencoder (SAE) [31] as our regression model. Both local information and global information of a 3D mesh model in multi-scale space are fully utilized to detect whether a vertex is a keypoint or not. SAE can effectively extract the internal structure of these two kinds of information and formulate high-level features for them, which is beneficial to the regression model. Three SAEs are used to formulate the hidden layers of the DNN and then a logistic regression [32] layer is trained to process the high-level features produced in the third SAE. These four layers are stacked together to formulate a DNN as the regression model of the proposed 3D keypoint detection algorithm. Numerical experiments demonstrate that the DNN approach proposed by us outperforms the existing five state-of-the-art methods for various 3D mesh models.

The main contributions of this paper are summarized as follows: (a) It's the first time that DNN with SAE has been used to detect 3D keypoints. (b) Both local information and global information of a 3D mesh model in multi-scale space are fully utilized to detect whether a vertex is a keypoint or not. (c) Numerical experiments on the datasets [33] are presented to verify the performance of the proposed DNN based 3D keypoint detection algorithm.

The rest of this paper is organized as follows. Introduction of DNN with SAE is described in Section 2. Section 3 presents the attributes and the training process. The proposed DNN based 3D keypoints detection algorithm will be displayed in Section 4 and performance study and result analysis are shown in Section 5. Finally, conclusions are presented in Section 6.

2 Deep Neural Network with Sparse Autoencoder

2.1 Sparse Autoencoder

Autoencoder [34] is an unsupervised learning algorithm, in which target values are set to be equal to the input values. It tries to learn the function $\mathbf{h}_{\mathbf{W},\mathbf{b}} \approx \mathbf{x}$, where \mathbf{x} represents the unlabeled dataset $\{x_1, x_2, \dots, x_m\}, x_i \in \mathbb{R}^m$. The left part of Fig. 1 displays the framework of an autoencoder, in which the hidden layer (layer 2 in Fig. 1) contains the internal structure of input data. Similar to Principal Component Analysis [35] (PCA), autoencoder can learn a low-dimensional representation of the input data by limiting the number of hidden units.

Sparse autoencoder [31, 36] is a variant of autoencoder, which can be formed by imposing a sparsity constraint on the hidden units. It can also learn interesting structure of the input data, even if the dimension of hidden layer is larger than the dimension of input layer. The overall cost function of sparse autoencoder is

$$J_{sparse}(\mathbf{W}, \mathbf{b}) = \frac{1}{2} \|\mathbf{h}_{\mathbf{W},\mathbf{b}} - \mathbf{x}\|^2 + \beta \sum_{j=1}^S \text{KL}(\rho || \hat{\rho}_j) \quad (1)$$

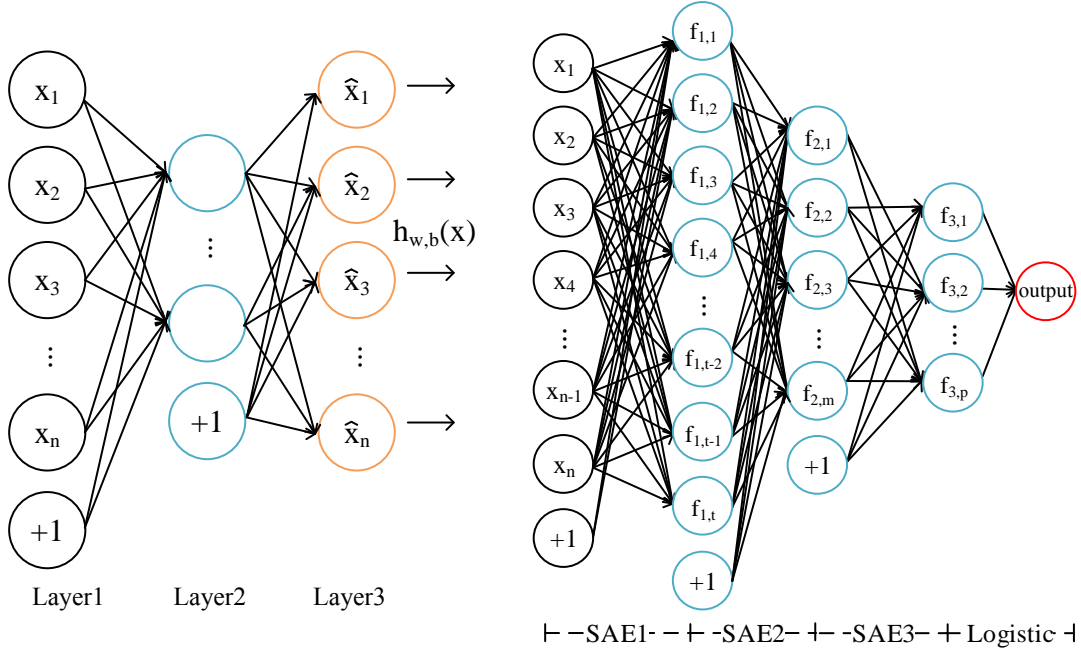


Figure 1: Left: the framework of an autoencoder. Right: DNN with three SAEs as the hidden layers and a logistic regression layer

where

$$\text{KL}(\rho || \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j} \quad (2)$$

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^m [a_j^2(x_i)] \quad (3)$$

where $a_j^2(x)$ denotes the activation of hidden unit j when the network is given a specific input x and $\hat{\rho}_j$ represents its average activation. The second term of (1) is a penalty term which penalizes $\hat{\rho}_j$ deviating significantly from ρ and (approximately) enforce the constraint $\hat{\rho}_j = \rho$, where ρ is a sparsity parameter, typically a small value close to zero. $\text{KL}(\rho || \hat{\rho}_j)$ is the Kullback-Leibler divergence [37] between a Bernoulli random variable with mean ρ and a Bernoulli random variable with mean $\hat{\rho}_j$, and it is a standard function for measuring the level of difference between two different distributions. β controls the weight of the sparsity penalty term. S is the number of neurons in the hidden layer.

2.2 Deep Neural Network with Sparse Autoencoder

Neural networks with multiple hidden layers can be useful for solving classification and regression problems with complex data. Each layer can learn features at different level of abstraction. However, there exist some problems when training neural networks with multiple hidden layers in practice, e.g. the local minimum problem of weights.

In this paper, we select SAE to formulate the hidden layers of DNN for the following three reasons. The first reason is its excellent performance according to [31, 36]. SAE can learn interesting structure of the input data effectively if there exists any correlation in the input data. Besides, SAE has great ability for feature processing according to [38]. Using deep sparse autoencoder (DSAE, stacked by a few successive SAEs) can learn high-level features of the input data effectively. Each SAE in DSAE can learn features at different levels (from low level to high level). In addition, the process of training SAEs can be viewed as a stage of pre-training of DNN, which makes the initial weights of DNN close to the position of global optimization. So, it can effectively avoid the local minimum problem of weights mentioned in the last paragraph.

To formulate the DNN regression model, we firstly train three SAEs and select the encoder part to formulate the hidden layers of the DNN. As is shown in right part of Fig. 1, the input of the first SAE is the original input data and its hidden layer is regarded as input of the second SAE. Likewise, the hidden layer of second SAE is regarded as the input of the third SAE. This process can be viewed as a stage of high-level features formulation according to [38]. Then we train a logistic regression layer to process the high-level features produced in the third SAE. Finally, we stack the four layers together to formulate a DNN as our regression model.

3 Attributes and Training Process

For 3D mesh models, there is no additional information other than the position of vertices and the connectivity information among these vertices. If we have enough 3D mesh models with ground truth, we can directly utilize these information to train the DNN regression model. However, the training data in 3D mesh models with ground truth, are limited. So, to improve the performance of DNN regression model, what we first need to do is preprocessing the original data to formulate the attributes as the inputs to our DNN regression model.

3.1 Attributes

In this section, we utilize both local and global information of 3D mesh models in multi-scale space to formulate the attributes as the inputs to our DNN regression model. For local information, we utilize three types of geometric properties of surface of a 3D mesh model: (1) the Euclidean distance between neighborhood rings to the tangent plane; (2) the angles of normal vectors between the vertex and its neighborhood rings; (3) various curvatures. For global information, we consider the properties of log-Laplacian spectrum of a 3D mesh model [23]. For any vertex v in a 3D mesh model $\mathbf{M}(x, y, z)$, let \mathbf{f} be its attribute. It can be formulated by

$$\mathbf{f} = [\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_\Omega]^T \quad (4)$$

$$\mathbf{f}_i = [\mathbf{f}_d, \mathbf{f}_\theta, \mathbf{f}_c, \mathbf{f}_{ls}], i = 0, 1, 2, \dots, \Omega \quad (5)$$

where $\mathbf{f}_i, i = 0, 1, 2, \dots, \Omega$ represents the information in scale i of a 3D mesh model which can be denoted as $\mathbf{M}_\delta(x, y, z)$. $\mathbf{M}_\delta(x, y, z)$ can be calculated by

$$\mathbf{M}_\delta(x, y, z) = \mathbf{M}(x, y, z) * \mathbf{G}(x, y, z, \delta) \quad (6)$$

$$\mathbf{G}(x, y, z, \delta) = \frac{1}{(\sqrt{2\pi}\delta)^3} e^{-\frac{(x^2+y^2+z^2)}{2\delta^2}} \quad (7)$$

where $\delta \in \{0, \varepsilon, 2\varepsilon, \dots, \Omega\varepsilon\}$ is the standard deviation of 3D Gaussian filter and ε amounts to 0.3% of the length of the main diagonal located in the bounding box of the model. $\mathbf{M}_0(x, y, z)$ indicates that it is the original mesh model $\mathbf{M}(x, y, z)$. $*$ is the convolution operator.

\mathbf{f}_i is made up of four parts: three types of local information ($\mathbf{f}_d, \mathbf{f}_\theta, \mathbf{f}_c$) and one type of global information (\mathbf{f}_{ls}).

3.1.1 Local Information

According to the second fundamental form of a surface [39], the most direct measurement which indicates the curvature degree of a vertex v on the surface of a 3D mesh model is the Euclidean distance between neighborhood vertices to the tangent plane of the vertex v . The angles of normal vectors between the vertex and its neighborhood rings is another important geometric property for the surface of a 3D mesh model [10].

Here, we calculate the first two types of local information \mathbf{f}_d and \mathbf{f}_θ . In this paper, a 3D mesh model (triangular mesh models of arbitrary topology) is represented as a set of vertices \mathbf{V} and faces \mathbf{F} with adjacency information between these entities. For different vertices in a 3D mesh model, they may have different numbers of neighborhood vertices. We utilize an adaptive technique in [7] to find the neighborhood vertices of a vertex. Let v be the analyzed vertex and $\mathbf{V}_k(v), k = 1, 2, 3, 4, 5$ be the k -ring neighborhood vertices around v . As is shown in Fig. 2, blue dots, magenta dots, green

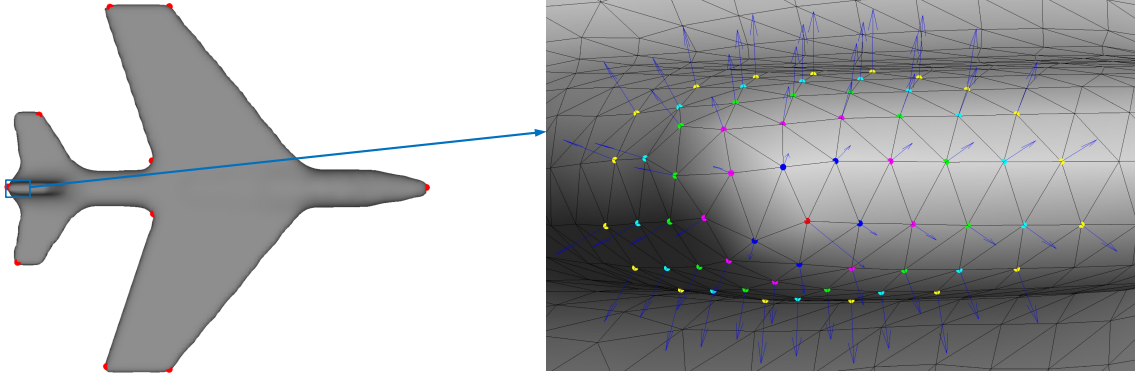


Figure 2: Left: airplane mesh model with ground truth; Right: part of airplane mesh model. (Vertex v (red dot) and its neighborhood rings: $\mathbf{V}_1(v)$, blue dots; $\mathbf{V}_2(v)$, magenta dots; $\mathbf{V}_3(v)$, green dots; $\mathbf{V}_4(v)$, cyan dots; $\mathbf{V}_5(v)$, yellow dots.)

dots, cyan dots and yellow dots represent the first, second, third, fourth and fifth ring around vertex v respectively. For any vertex v in a 3D mesh model, we utilize \mathbf{n} to represent its normal vector. Let v_{kj} be the j -th point in $\mathbf{V}_k(v)$ and \mathbf{n}_{kj} be its normal vector. Let d_{kj} be the Euclidean distance of v_{kj} to the tangent plane and θ be the angle of normal vector between v and v_{kj} , both of which can be calculated by

$$d_{kj} = \frac{|\mathbf{n}^T [x_{kj}, y_{kj}, z_{kj}]^T - \mathbf{n}^T [x_v, y_v, z_v]^T|}{\|\mathbf{n}\|_2} \quad (8)$$

$$\theta_{kj} = \min(\arccos(\frac{\mathbf{n}^T \mathbf{n}_{kj}}{\|\mathbf{n}\|_2 \|\mathbf{n}_{kj}\|_2})) \quad (9)$$

where (x_v, y_v, z_v) is the coordinate of v and (x_{kj}, y_{kj}, z_{kj}) is the coordinate of v_{kj} . Let \mathbf{d}_k be $[d_{k1}, d_{k2}, \dots, d_{kN_k}]^T$ and $\boldsymbol{\theta}_k$ be $[\theta_{k1}, \theta_{k2}, \dots, \theta_{kN_k}]^T$, where N_k is the number of vertices in $\mathbf{V}_k(v)$. Because there are different numbers of neighborhood vertices for different vertices in a 3D mesh model, we utilize six types of statistical properties of \mathbf{d}_k and $\boldsymbol{\theta}_k$ to formulate \mathbf{f}_d and \mathbf{f}_θ respectively in order to form the fixed dimension of attributes. \mathbf{f}_d and \mathbf{f}_θ can be calculated by

$$\mathbf{f}_d = [\max(\mathbf{d}_k), \min(\mathbf{d}_k), \max(\mathbf{d}_k) - \min(\mathbf{d}_k), \text{mean}(\mathbf{d}_k), \text{var}(\mathbf{d}_k), \text{harmmean}(\mathbf{d}_k)] \quad (10)$$

$$\mathbf{f}_\theta = [\max(\boldsymbol{\theta}_k), \min(\boldsymbol{\theta}_k), \max(\boldsymbol{\theta}_k) - \min(\boldsymbol{\theta}_k), \text{mean}(\boldsymbol{\theta}_k), \text{var}(\boldsymbol{\theta}_k), \text{harmmean}(\boldsymbol{\theta}_k)] \quad (11)$$

where $\text{mean}(\cdot)$, $\text{var}(\cdot)$ and $\text{harmmean}(\cdot)$ are the arithmetic average, variance and harmonic average respectively.

As is shown in Table 1, for any vertex v located in flat area of a 3D mesh model, all six types of statistical properties of \mathbf{d}_k and $\boldsymbol{\theta}_k$ should be small. For any vertex v located in edges of a 3D mesh model, $\max(\cdot)$, $\max(\cdot) - \min(\cdot)$ and $\text{var}(\cdot)$ should be large. However, $\min(\cdot)$ and $\text{var}(\cdot)$ should be small. For any vertex v belonging to 3D keypoints of a 3D mesh model, $\max(\cdot)$, $\min(\cdot)$, $\text{mean}(\cdot)$ and $\text{harmmean}(\cdot)$ should be large. However, $\max(\cdot) - \min(\cdot)$ and $\text{var}(\cdot)$ should be small.

Besides of the two geometric properties mentioned above, curvatures [40] are frequently used to detect saliency of 3D mesh models according to [5, 13, 15, 24, 26]. In this paper, four types of curvatures are used to formulate the \mathbf{f}_c and it can be formulated as

$$\mathbf{f}_c = [c_1, c_2, \frac{c_1 + c_2}{2}, c_1 c_2] \quad (12)$$

where c_1 and c_2 are principal curvatures. $(c_1 + c_2)/2$ is mean curvature, and $c_1 c_2$ is gaussian curvature.

Table 1: The effect of six types of statistical properties presented on different regions of a 3D mesh model

	$\max(\cdot)$	$\min(\cdot)$	$\max(\cdot) - \min(\cdot)$	$\text{mean}(\cdot)$	$\text{var}(\cdot)$	$\text{harmmean}(\cdot)$
Flat areas	small	small	small	small	small	small
Edges	large	small	large	relatively large	large	small
3D keypoints	large	large	small	large	small	large

3.1.2 Global Information

As a powerful tool, Laplacian spectrum is widely used to analyze global information of 3D mesh models according to [3, 22, 23, 41, 42, 43]. The Laplacian matrix \mathbf{L} of a 3D mesh model is a symmetric matrix and it can be decomposed as:

$$\mathbf{L} = \mathbf{B} \mathbf{\Lambda} \mathbf{B}^T \quad (13)$$

where $\mathbf{\Lambda} = \text{Diag}\{\lambda_f, 1 \leq f \leq \Psi\}$ is a diagonal matrix arranged in ascending order, in which λ_f is the eigenvalue of \mathbf{L} . The corresponding eigenvectors are utilized to form the orthogonal matrix \mathbf{B} . Ψ is the number of vertices of a 3D mesh model. $\mathbf{H}(f) = \{\lambda_f, 1 \leq f \leq \Psi\}$ is the Laplacian spectrum of a 3D mesh model.

In this paper, we get global information via log-Laplacian spectrum used in [23], which is defined as

$$\mathbf{L}(f) = \log(|\mathbf{H}(f)|) \quad (14)$$

Spectral irregularity \mathbf{R} is utilized to calculate the mesh saliency and it can be formulated as

$$\mathbf{R}(f) = |\mathbf{L}(f) - \mathbf{J}_\Gamma(f) * \mathbf{L}(f)| \quad (15)$$

where $\mathbf{J}_\Gamma(f) = \frac{1}{\Gamma}[1, 1, \dots, 1]$ is $1 \times \Gamma$ vector. To bring this representation back to the spatial domain, the composition should be formulated as:

$$\mathbf{S} = \mathbf{B} \mathbf{R}_1 \mathbf{B}^T \odot \mathbf{W} \quad (16)$$

where $\mathbf{R}_1 = \text{Diag}\{\exp(\mathbf{R}(f)) : 1 \leq f \leq \Psi\}$ is a diagonal matrix. \mathbf{W} is the distance-weighted adjacency matrix and \odot is Hadamard product.

Let \mathbf{s}_k be $[s_{k1}, s_{k2}, \dots, s_{kN_k}]^T$ and s be the element of \mathbf{S} . The same with last section, we also utilize the six types of statistical properties of \mathbf{s}_k to formulate \mathbf{f}_{ls} . It can be calculated by

$$\mathbf{f}_{ls} = [\max(\mathbf{s}_k), \min(\mathbf{s}_k), \max(\mathbf{s}_k) - \min(\mathbf{s}_k), \text{mean}(\mathbf{s}_k), \text{var}(\mathbf{s}_k), \text{harmmean}(\mathbf{s}_k)] \quad (17)$$

3.2 Training Process

We use the same datasets as in [22, 24, 33] where a web-based application is developed and utilized to collect ground truth of 3D keypoints on 43 mesh models. These mesh models are organized in two datasets. The first one (Dataset A) is constituted by 24 triangular mesh models and annotated by 23 human subjects. Another one (Dataset B) is constituted by 43 triangular mesh models and annotated by 16 human subjects. Similar to [24], for all the experiments, we select two-thirds of Dataset A and two-thirds of Dataset B to train our DNN regression model. The remained mesh models are used as test datasets and they are never used in training the DNN regression model. As is shown in Table 2, for Dataset A, the representative of clusters $\sigma \in \{0.01, 0.02, \dots, 0.1\}$ and $n \in \{11, 12, \dots, 22\}$ are placed in the positive class. For Dataset B, the representative of clusters $\sigma \in \{0.01, 0.02, \dots, 0.1\}$ and $n \in \{8, 9, \dots, 15\}$ are placed in the positive class. All other vertices are placed in the negative class.

To train the DNN regression model, we need to train three SAEs firstly. All the parameters related to the three SAEs are displayed in Table 3. Then a logistic regression layer is trained. Finally, we stack the four layers together to formulate a DNN as our regression model and fine tuning is done by performing backpropagation on the DNN to improve the performance of DNN regression model.

Table 2: Training datasets

Datasets	Mesh models	n	σ	Positive samples	Negative samples
A	16	$\{11, \dots, 22\}$	all	17115	148565
B	28	$\{8, \dots, 15\}$	all	18427	222034

Table 3: Parameters for DNN regression model

	Dimension of input layer	Dimension of hidden layer	ρ	β
SAE1	665	800	0.15	4
SAE2	800	200	0.15	4
SAE3	200	50	0.1	4
Logisitc	50	1		

4 Proposed DNN Based 3D Keypoint Detection Algorithm

The outline of the proposed DNN based 3D keypoint detection algorithm can be divided into the following four steps:

- Utilize 3D Gaussian filter to construct scale space for a 3D mesh model $\mathbf{M}(x, y, z)$ and get a series of evolved 3D mesh models $\mathbf{M}_\delta(x, y, z)$.
- Use multi-scale information to calculate the attributes for every vertex of a 3D mesh model in the same way as described in Section 3.
- For every vertex of a 3D mesh model, put its attributes to the well-trained DNN regression model and get a regression value, and then get the saliency map of this 3D mesh model.
- Select the local maxima of saliency map of a 3D mesh model as the 3D keypoints. Compare the value of DNN regression ρ for every vertex with those points in its neighborhood rings $V_k(v), k = 1, 2, 3, 4, 5$ and select the maximal one as the 3D keypoints.

5 Numerical Experiments

In this section, we evaluate the performance of DNN based 3D keypoint detection algorithm and compare it with five state-of-the-art methods - namely 3D Harris [7], HKS [8], Salient Points [11], Mesh Saliency [5] and Scale Dependent Corners [9]. All of the five methods are referenced algorithms for performance evaluation in [22, 24, 33].

5.1 Datasets and Evaluation Metrics

As is described in section 3.2, we utilize the remained one-third of Dataset A and one-third of Dataset B as the test datasets. Some evaluation methods usually measured the repeatability rate according to varying factors, such as model deformation, scale change, different modalities, noise, and topological change [44]. Different from them, Dutagaci et al. utilized three evaluation metrics - namely False Positive Error (FPE), False Negative Error (FNE) and Weighted Miss Error (WME) to evaluate the performance of 3D keypoints detection algorithms [33].

Besides of the evaluation metrics mentioned above, Teran [24] also adopted the Intersection Over Union (IOU) [29] as their main metric to evaluate the performance of the 3D keypoint detection algorithms. It can be calculated by

$$IOU(r) = \frac{TP}{FN + FP + TP} \quad (18)$$

where $FP = N_A - N_C$ is the number of false positives and $FN = N_G - N_C$ represents the number of false negatives. $TP = N_C$ is the number of true positives. N_G is the number of ground truth

points, N_C is the number of correctly detected points and N_A denotes the number of detected 3D keypoints by the algorithm. r is localization error tolerance [33].

Similar to Dutagaci et al. [33], Song et al. [22] and Teran et al. [24], we utilize the same two datasets to evaluate the performance of six algorithms in terms of four evaluation metrics - namely IOU, FNE, FPE and WME. The IOU is chosen as the main metric because FNE and FPE can be misleading in isolation according to the discussion in [24]. Besides, a 3D video sequence produced in [45] is used to verify the stability of the proposed DNN based 3D keypoints detection algorithm.

5.2 Experimental Results

In this section, we present the performance results of the proposed DNN based 3D keypoint detection algorithm and other five state-of-the-art 3D keypoint detection algorithms. In all the experiments, Ω is 6 and Γ is 9. It is important to note that the 3D keypoints detected by the six 3D keypoint detection algorithms are constant when n/σ varies, but the ground truth is variable when n/σ varies according to [33].

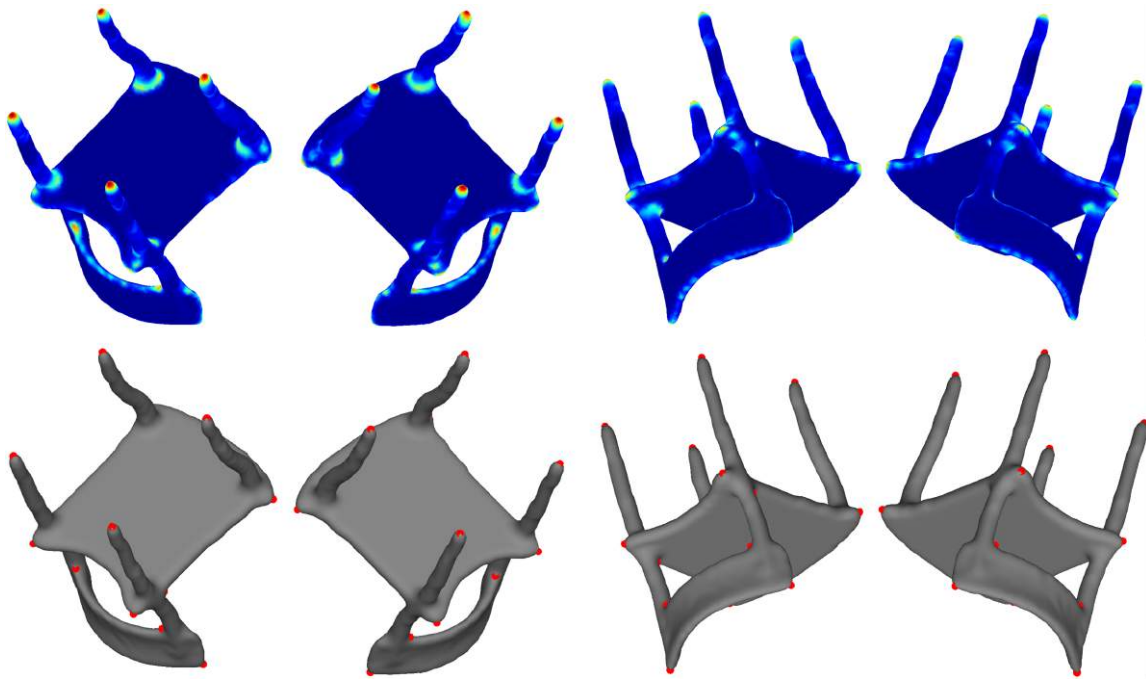
Fig. 3 displays two types of visualizations of the results. Fig. 3(a) displays the 3D keypoints of chair model detected by our proposed DNN based 3D keypoint detection algorithm (second row) and corresponding saliency maps (first row) from four different viewpoints. Fig. 3(b) displays 3D keypoints and corresponding saliency maps of five frames (the 1st, 25th, 50th, 75th and 100th frame respectively) which come from a 3D video sequence [45] from the same viewpoint. Visualizations of comparative results can be found from Fig. 4, where 3D keypoint of chair model in Dataset A and armadillo model in Dataset B are detected by six methods. More visualizations of the results can be found in supplementary material. From the visualizations of the results, we can see that the 3D keypoints detected by our approach are more in accord with human visual characteristics. Besides, the 3D keypoints detected by our approach are stable according to Fig. 3(b), where the distribution and location of the detected 3D keypoints are almost the same, except for a keypoint lied on the shoulder of the first frame of the 3D video sequence.

Fig. 5 gives IOU graphs with respect to localization error tolerance r for six 3D keypoint detection algorithms. We compare the performance of six methods at $n = 8/\sigma = 0.03$, $n = 8/\sigma = 0.05$, $n = 2/\sigma = 0.03$ and $n = 2/\sigma = 0.05$ for Dataset A in terms of IOU evaluation metrics. In the same way, we compare the performance of six methods at $n = 11/\sigma = 0.03$, $n = 11/\sigma = 0.05$, $n = 2/\sigma = 0.03$ and $n = 2/\sigma = 0.05$ for Dataset B in terms of IOU evaluation metrics. Our proposed approach performs best in terms of IOU evaluation metric, especially when r and n are relatively large.

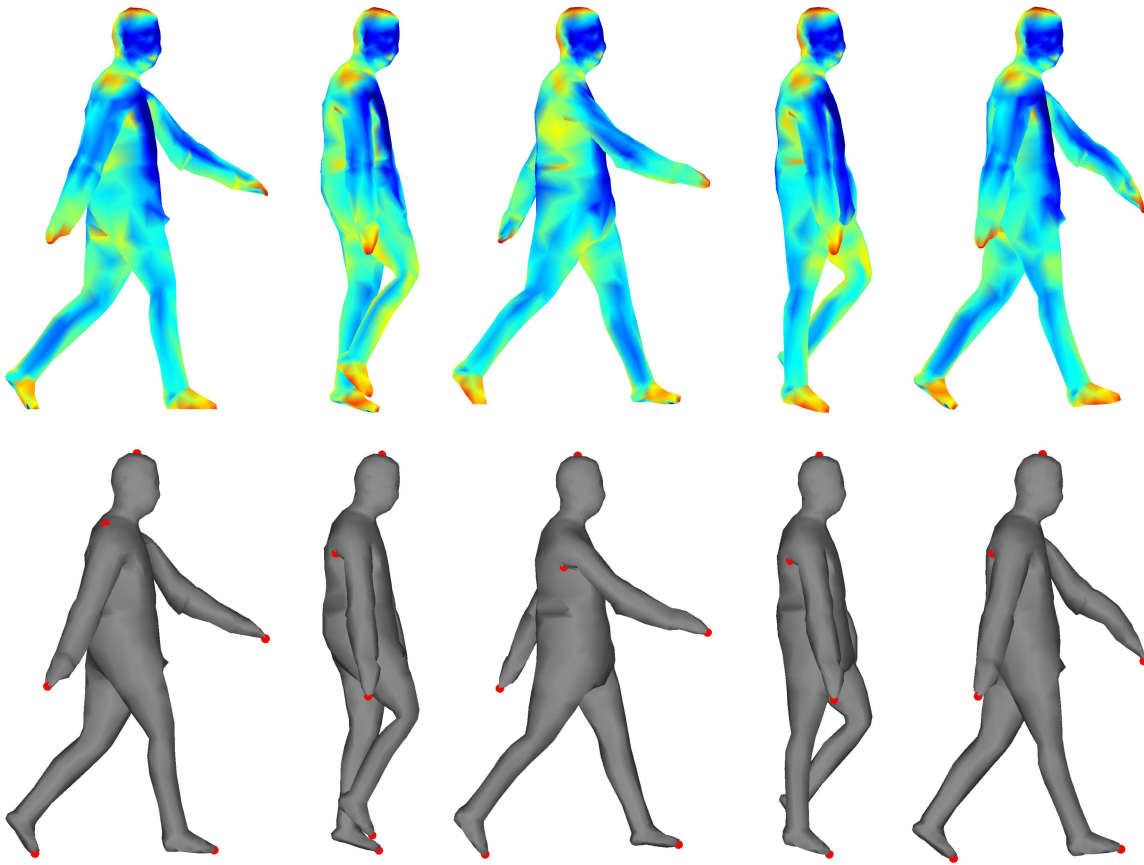
To reach an overall comparison, we average the 4 types of evaluation metric scores over all settings. Fig. 6 gives average IOU, FNE, FPE and WME graphs with respect to localization error tolerance r for six 3D keypoint detection algorithms, where $n \in \{2, 3, \dots, 23\}/\sigma \in \{0.01, 0.02, \dots, 0.1\}$ is for Dataset A and $n \in \{2, 3, \dots, 16\}/\sigma \in \{0.01, 0.02, \dots, 0.1\}$ is for Dataset B. Digital results over all settings are summarized in Table 4. From Fig. 6 and Table 4, we can see that our proposed approach performs best, especially when localization error tolerance r is relatively large.

Table 4: Average IOU, FNE, FPE, WME on Test Dataset A ($n \in \{2, 3, \dots, 23\}, \sigma \in \{0.01, 0.02, \dots, 0.1\}$) and B ($n \in \{2, 3, \dots, 16\}, \sigma \in \{0.01, 0.02, \dots, 0.1\}$)

	IOU-A	FNE-A	FPE-A	WME-A	IOU-B	FNE-B	FPE-B	WME-B
Mesh saliency	0.078	0.248	0.919	0.232	0.063	0.225	0.935	0.210
SD-corners	0.061	0.328	0.936	0.326	0.072	0.379	0.924	0.380
3D-Harris	0.102	0.343	0.879	0.329	0.084	0.330	0.910	0.307
Salient points	0.111	0.376	0.875	0.361	0.122	0.295	0.868	0.274
HKS	0.216	0.671	0.530	0.656	0.218	0.686	0.519	0.662
DNN	0.275	0.509	0.561	0.484	0.272	0.516	0.578	0.490

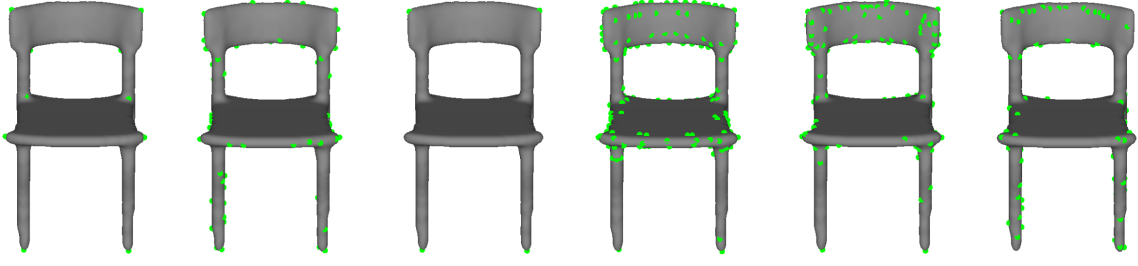


(a)

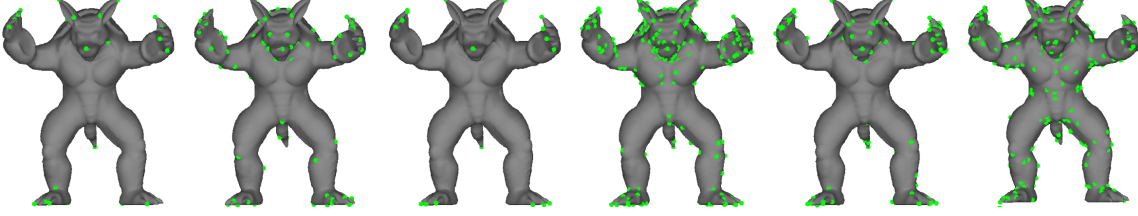


(b)

Figure 3: Saliency maps (first row) and 3D keypoints (second row) detected by the proposed DNN based 3D keypoint detection algorithm.



(a) 3D keypoints of chair model in Dataset A detected by six algorithms



(b) 3D keypoints of armadillo model in Dataset B detected by six algorithms

Figure 4: 3D keypoints detected by six algorithms. (DNN: first column; Salient points: second column; HKS: third column; Mesh saliency: fourth column; 3D-Harris: fifth column; SD corners: sixth column)

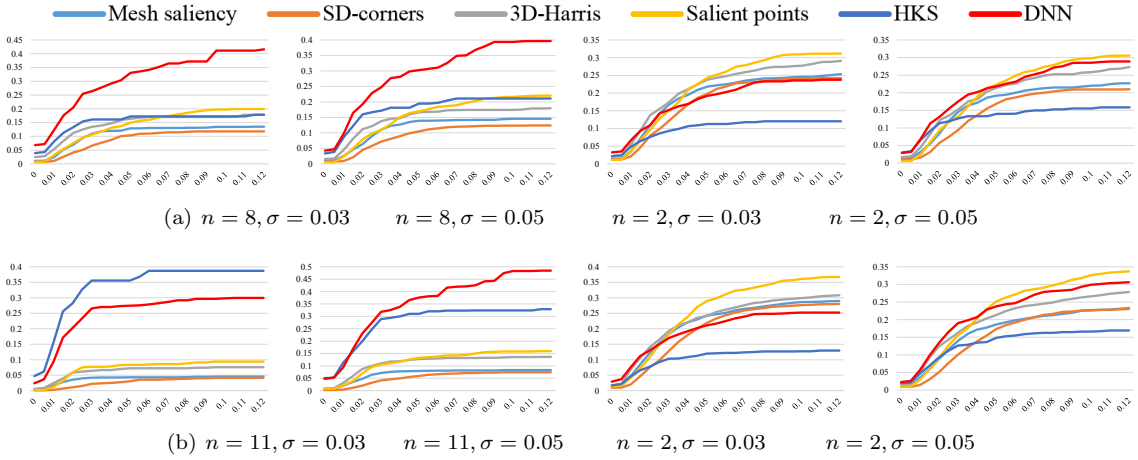


Figure 5: IOU curves for Dataset A (first row) and Dataset B (second row) at various n/σ pairs

6 Conclusion

In this paper, we propose a new 3D keypoint detection algorithm on the basis of deep learning by formulating the 3D keypoint detection as a regression problem using DNN with SAE as our regression model. It's the first time that DNN with SAE has been used to detect 3D keypoints. Both local information and global information of a 3D mesh model in multi-scale space are fully utilized to detect whether a vertex is a keypoint or not. Three types of geometric properties of surface of a 3D mesh model are used to formulate the local information: 1) the Euclidean distance between neighborhood rings to the tangent plane; 2) the angle of normal vector between the vertex and its neighborhood rings; 3) various curvatures. For global information, we consider the properties of log-Laplacian spectrum of a 3D mesh model used in [23]. SAE can effectively extract the internal structure of these two kinds of information and formulate high-level features for them, which is beneficial to the regression model. Three SAEs are used to formulate the hidden layers of the DNN and then a logistic regression layer is trained to process the high-level features produced in the

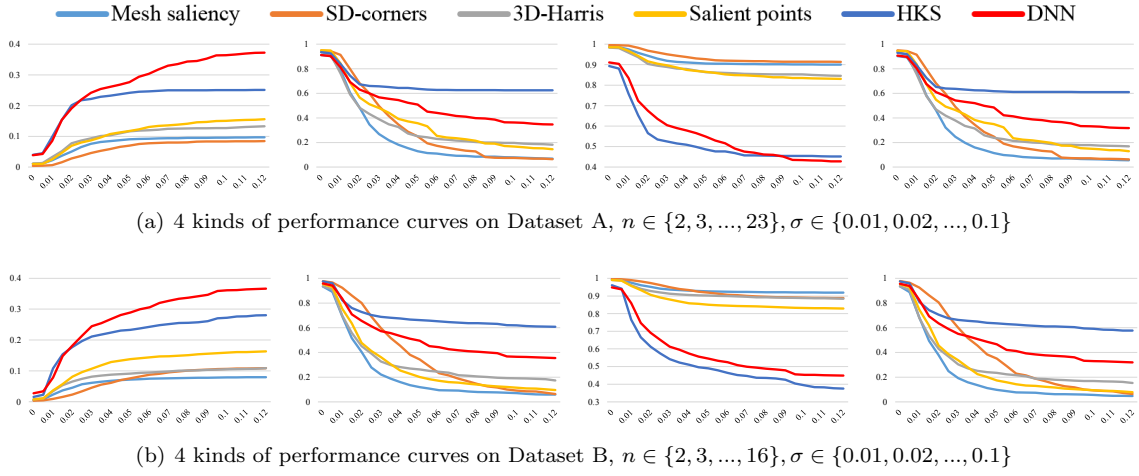


Figure 6: Four kinds of performance curves on Dataset A and B. IOU (first column), FNE (second column), FPE (third column), WME (fourth column)

third SAE. These four layers are stacked together to formulate a DNN as the regression model of our 3D keypoint detection algorithm.

Experimental results indicate that the proposed DNN based 3D keypoint detection algorithm outperforms other five state-of-the-art methods in terms of IOU metric, especially when localization tolerance error r is relatively large. Besides, the 3D keypoints detected by our approach are stable and more in accord with human visual characteristics.

References

- [1] Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration. In: Proceedings of the Third Eurographics Symposium on Geometry Processing. SGP '05, Aire-la-Ville, Switzerland, Switzerland, Eurographics Association (2005)
- [2] Funkhouser, T., Kazhdan, M.: Shape-based retrieval and analysis of 3d models. In: ACM SIGGRAPH 2004 Course Notes. SIGGRAPH '04, New York, NY, USA, ACM (2004)
- [3] Hu, J.X., Hua, J.: Salient spectral geometric features for shape matching and retrieval. *The Visual Computer* **25**(5) (2009) 667–675
- [4] Katz, S., Leifman, G., Tal, A.: Mesh segmentation using feature point and core extraction. *The Visual Computer* **21**(8) (2005) 649–658
- [5] Lee, C.H., Varshney, A., Jacobs, D.W.: Mesh saliency. *ACM Transactions on Graphics* **24**(3) (2005) 659–666
- [6] Godil, A., Wagan, A.I.: Salient local 3d features for 3d shape retrieval. *Proceedings of SPIE* **7864** (2011) 78640S–78640S–8
- [7] Sipiran, I., Bustos, B.: Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer* **27**(11) (2011) 963–976
- [8] Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. *Computer Graphics Forum* **28**(5) (2009) 1383–1392
- [9] Novatnack, J., Nishino, K.: Scale-dependent 3d geometric features. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (Oct 2007) 1–8

- [10] Holte, M.B.: 3d interest point detection using local surface characteristics with application in action recognition. In: Image Processing (ICIP), 2014 IEEE International Conference on. (Oct 2014) 5736–5740
- [11] Castellani, U., Cristani, M., Fantoni, S., Murino, V.: Sparse points matching by combining 3d mesh saliency with statistical descriptors. *Computer Graphics Forum* **27**(2) (2008) 643–652
- [12] Wang, S., Gong, L., Zhang, H., Zhang, Y., Ren, H., Rhee, S.M., Lee, H.E.: Sdtp: a robust method for interest point detection on 3d range images. *Proceedings of SPIE* **9025** (2014) 90250O–90250O–9
- [13] Akagndz, E., Ulusoy, .: Scale and orientation invariant 3d interest point extraction using hk curvatures. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. (Sept 2009) 697–702
- [14] Zaharescu, A., Boyer, E., Varanasi, K., Horaud, R.: Surface feature detection and description with applications to mesh matching. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (June 2009) 373–380
- [15] Chen, H., Bhanu, B.: 3d free-form object recognition in range images using local surface patches. *Pattern Recognition Letters* **28**(10) (2007) 1252–1262
- [16] Zhong, Y.: Intrinsic shape signatures: A shape descriptor for 3d object recognition. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. (Sept 2009) 689–696
- [17] Mian, A., Bennamoun, M., Owens, R.: On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision* **89**(2) (2009) 348–361
- [18] Unnikrishnan, R., Hebert, M.: Multi-scale interest regions from unorganized point clouds. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on. (June 2008) 1–8
- [19] Darom, T., Keller, Y.: Scale-invariant features for 3-d mesh models. *Image Processing, IEEE Transactions on* **21**(5) (2012) 2758–2769
- [20] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60**(2) (2004) 91–110
- [21] Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, Britain (Sept 1988)
- [22] Song, R., Liu, Y., Martin, R.R., Rosin, P.L.: 3d point of interest detection via spectral irregularity diffusion. *The Visual Computer* **29**(6) (2013) 695–705
- [23] Song, R., Liu, Y.H., Martin, R.R., Rosin, P.L.: Mesh saliency via spectral processing. *ACM Transactions on Graphics* **33**(1) (2014) 6:1–6:17
- [24] Teran, L., Mordohai, P.: 3d interest point detection via discriminative learning. In: Proceedings of the 13th European Conference on Computer Vision Conference on Computer Vision, Zurich, Switzerland (Sept 2014)
- [25] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(10) (2005) 1615–1630
- [26] Creusot, C., Pears, N., Austin, J.: A machine-learning approach to keypoint detection and landmarking on 3d meshes. *International Journal of Computer Vision* **102**(1) (2013) 146–179
- [27] Salti, S., Tombari, F., Spezialetti, R., Stefano, L.D.: Learning a descriptor-specific 3d keypoint detector. In: Computer Vision (ICCV), 2015 IEEE International Conference on. (Dec 2015) 2318–2326

- [28] Breiman, L.: Random forests. *Machine learning* **45**(1) (2001) 5–32
- [29] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1) (1997) 119–139
- [30] Tombari, F., Salti, S., Di Stefano, L.: Unique signatures of histograms for local surface description. In: *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III. ECCV’10*. Springer-Verlag, Berlin, Heidelberg (2010) 356–369
- [31] Ng, A.: Sparse autoencoder. *CS294A Lecture notes* **72** (2011)
- [32] Cox, D.R.: The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* **20**(2) (1958) 215–242
- [33] Dutagaci, H., Cheung, C.P., Godil, A.: Evaluation of 3d interest point detection techniques via human-generated ground truth. *The Visual Computer* **28**(9) (2012) 901–917
- [34] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786) (2006) 504–507
- [35] Jolliffe, I.: *Principal component analysis*. Wiley Online Library (2002)
- [36] Goodfellow, I., Lee, H., Le, Q.V., Saxe, A., Ng, A.Y.: Measuring invariances in deep networks. In: *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc. (2009) 646–654
- [37] Kullback, S., Leibler, R.A.: On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1) (1951) 79–86
- [38] Le, Q.V.: Building high-level features using large scale unsupervised learning. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. (May 2013) 8595–8598
- [39] Harris, W.: The second fundamental form of a surface and its relation to the dioptric power matrix, sagitta and lens thickness. *Ophthalmic and Physiological Optics* **9**(4) (1989) 415–419
- [40] Flynn, P.J., Jain, A.K.: On reliable curvature estimation. In: *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR ’89., IEEE Computer Society Conference on*. (June 1989) 110–116
- [41] Pauly, M., Kobbelt, L.P., Gross, M.: Point-based multiscale surface representation. *ACM Transactions on Graphics* **25**(2) (2006) 177–193
- [42] Lévy, B., Zhang, H.R.: Spectral mesh processing. In: *ACM SIGGRAPH 2010 Courses. SIGGRAPH ’10*, New York, NY, USA, ACM (2010) 8:1–8:312
- [43] Zhang, J., Zheng, J., Wu, C., Cai, J.: Variational mesh decomposition. *ACM Transactions on Graphics* **31**(3) (2012) 21:1–21:14
- [44] Tombari, F., Salti, S., Di Stefano, L.: Performance evaluation of 3d keypoint detectors. *International Journal of Computer Vision* **102**(1) (2012) 198–220
- [45] Huang, P., Hilton, A., Starck, J.: Shape similarity for 3d video sequences of people. *International Journal of Computer Vision* **89**(2) (2010) 362–381