Yuya Ohishi, Ryutarou Ohbuchi, Densely sampled Local Visual Features on 3D mesh for retrieval, *short paper*, *Proc. WIAMIS 2013*, pp.1-4, July 3-5, 2013, Paris, France.

DENSELY SAMPLED LOCAL VISUAL FEATURES ON 3D MESH FOR RETRIEVAL

Yuya Ohishi, Ryutarou Ohbuchi

sst623.ohishiAT_gmail.com, ohbuchiAT_yamanashi.ac.jp

University of Yamanashi, Kofu, Yamanashi-ken, Japan

ABSTRACT

The Local Depth-SIFT (LD-SIFT) algorithm by Darom, et al. [2] captures 3D geometrical features locally at interest points detected on a densely-sampled, manifold mesh representation of the 3D shape. The LD-SIFT has shown good retrieval accuracy for 3D models defined as densely sampled manifold mesh. However, it has two shortcomings. The LD-SIFT requires the input mesh to be densely and evenly sampled. Furthermore, the LD-SIFT can't handle 3D models defined as a set of multiple connected components or a polygon soup. This paper proposes two extensions to the LD-SIFT to alleviate these weaknesses. First extension shuns interest point detection, and employs dense sampling on the mesh. Second extension employs remeshing by dense sample points followed by interest point detection a la LD-SIFT. Experiments using three different benchmark databases showed that the proposed algorithms significantly outperform the LD-SIFT in terms of retrieval accuracy.

1. INTRODUCTION

Shape-based comparison, clustering, recognition, retrieval, or mining of three-dimensional (3D) shape models has become an important area of study as 3D shape model has become a mainstream multi-media data type. These operations require features or descriptors for 3D shape. In this paper, we aim at features that accept diverse surface-based shape representations including manifold mesh and polygon soup, are invariant to non-rigid deformations and 7-DOF similarity transformations.

Our proposed algorithms are based on *Local Depth SIFT* (LD-SIFT) proposed by Darom and Keller [2]. The LD-SIFT computes interest points on a (densely sampled) 2D manifold mesh embedded in 3D space and extracts local visual feature at these interest points. These local features are integrated into a feature vector per 3D model by using *Bag-of-Features* (*BoF*) approach (e.g., [10]) for efficient model-to-model comparison. The LD-SIFT performs well for certain class of 3D models, namely, those defined as a densely sampled manifold mesh. For other classes of 3D models, e.g., a sparsely sampled 3D model, a 3D model consisting of multiply connected-components, or a polygon soup model, the LD-SIFT performs poorly.

We designed our algorithms with the aim of alleviating these weaknesses of the LD-SIFT. To improve invariance to mesh sampling density, our first algorithm computes LD-SIFT local features at densely and randomly generated points on model's surfaces. To improve invariance against shape representation, our second algorithm remeshes the input with dense set of points, followed by the interest point detector and local features of the LD-SIFT. Experimental evaluation showed that the two proposed algorithms have higher retrieval accuracy than the LD-SIFT for broader class of shape representations and/or sampling density. The proposed algorithms also perform competitively with some other algorithms in the literature.

We briefly review the LD-SIFT algorithm in the next section, followed in Section 3 by the description of proposed algorithms. Experiments and results are presented in Section 4, and conclusions will be stated in Section 5.

2. RELATED WORK

The LD-SIFT [2] is in concept similar to Lowe's *Scale Invariant Feature Transform* (SIFT) [5] for 2D images. The LD-SIFT tries to detect scale-invariant interest points on a 3D mesh model. It then computes local geometrical features at the interest points. While 2D image pixels have 2D rectangular grid connectivity, 3D meshes have irregular connectivity. Consequently, the algorithm to compute multi-scale pyramid necessary for interest point detection on a 3D mesh is different from that for 2D images. Also, local feature computed at each interest point is different.

The LD-SIFT accepts as its input a densely and evenly sampled 2D manifold mesh embedded in 3D space, and recursively applies mesh density-invariant (to a certain extent) smoothing operation to the mesh. The LD-SIFT then computes *Difference of Gaussian* (*DoG*) meshes from the sequence of repeatedly Gaussian-smoothed mesh. Figure 1 illustrates the DoG mesh generation, in which the most detailed (smallest scale) DoG mesh is indicated as level 0. The LD-SIFT then finds interest points as local maxima or minima, in both scale and location, of the DoG meshes.

To extract view-based geometrical feature, the LD-SIFT finds local coordinate frame at each interest point by applying Principal Component Analysis to the distribution of the vertices in the locality. The LD-SIFT then renders a small depth image (e.g., of size 21×21 pixels) that covers the area around the interest point. (Figure 2.) For an interest point i, scale $S_i = E\sqrt{g}D_i$ is the size of the area to be

rendered and the area from which a local tangential plane is computed. Parameter E controls the scale, D_i is the average distance from the interest point i to the vertices adjacent to it, and g is the number of Gaussian smoothing applied to the mesh. The larger the scale S_i, the wider the area covered in a local depth image. One or more SIFT features are then extracted from each depth image. A set of dozens to thousands of local features per 3D model is integrated into a feature vector per 3D model by using BoF approach for efficient model-to-model comparison.

The LD-SIFT performed well for the 3D models it is designed for. However, there are several drawbacks. Interest points for the LD-SIFT can't be computed properly if the mesh is sampled very sparsely, the mesh contains high-aspect ratio polygons, or the mesh is not manifold. Figure 3 shows two 3D models of horse having different sampling densities; 11,105 vertices and 346 vertices. The former produced 812 interest points on the most detailed of the DoG meshes (DoG level 0). However, the latter, sparsely sampled mesh produced only 36 interest points. Retrieval accuracy using such a small number of interest points (and thus local features) is not very good.

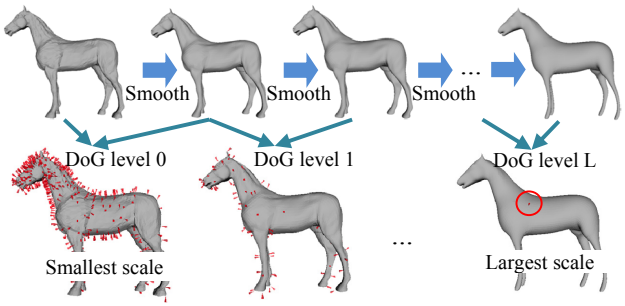


Fig. 1. Computing interest points for the LD-SIFT [2].

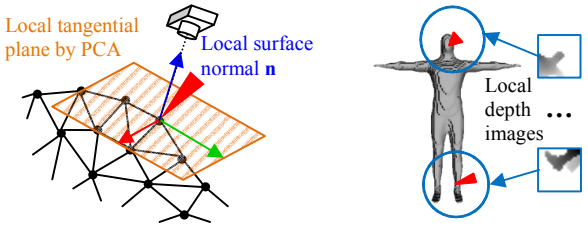


Fig. 2. At each interest point, local coordinate frame is computed by PCA, and a local depth image is rendered from the direction of normal vector to be used for extracting SIFT [5] features.

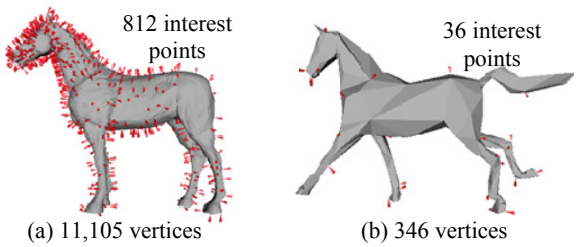


Fig. 3. Examples of LD-SIFT [2] interest points.

3. ALGORITHMS

3.1. Resampled LD-SIFT (RLD-SIFT)

Our first algorithm, *Resampled LD-SIFT*, or *RLD-SIFT*, remeshes the input 3D model by N_{RLD} points generated densely and uniformly over the 3D model surfaces. By doing so, sparsely sampled mesh (i.e., low polygon count mesh) and high-aspect ratio polygons are taken care of.

Sample points on the surfaces are generated so that the number of points per area is uniform over the 3D model. Assume that the total area of the 3D model is A. Then, a point should be placed per area $s = A / N_{RLD}$. For the ith triangular face f_i having an area(f_i), its number of points q_i is determined by $q_i = \text{floor}(\text{area}(f_i)/s + \Delta q_{i-1})$ where $\Delta q_i = \text{area}(f_i)/s + \Delta q_{i-1} - q_i$. Given the number of points q_i, each one of the points on a triangle is placed at position $\mathbf{P} = (1-\sqrt{u_1})\,\mathbf{t}_1 + \sqrt{u_1}(1-u_2)\mathbf{t}_2 + \sqrt{u_1}(u_2 \cdot \mathbf{t}_3)$. Here, $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$ are vertices of the triangle, and u_1 and u_2 are *Sobol's quasi-random number sequences* (QRNS)[8]. Using QRNS produces sampling having lower variance in density than pseudo-random sequence. Then each sample point is connected to its k nearest points by Euclidian distance to generate the resampled 3D mesh model. Note that the resampled mesh is in general not manifold.

After the remeshing, interest points are detected by using steps identical to that of the LD-SIFT, *i.e.,* recursive Gaussian smoothing for DoG meshes followed by detection of local maxima and minima on the DoG meshes.

Feature computation at the interest point is also nearly identical to the LD-SIFT; a local tangential plane is found via PCA, a small depth image is rendered of the area around the interest point from the direction of the plane's normal vector, and SIFT features are extracted from the depth image. We made a change in the image generation. The LD-SIFT generated an image from a camera placed "outside" of the surface that "looks into" the mesh. For many 3D models that are not solid (e.g., those defined by closed manifold mesh), distinction of inside/outside is meaningless. Often used convention of vertex traverse order is often not coherent from a mesh modeling software to the other. Thus, the RLD-SIFT and the DLD-SIFT renders two images, one "looking into" and the other "looking out of", at the interest point for SIFT feature extraction.

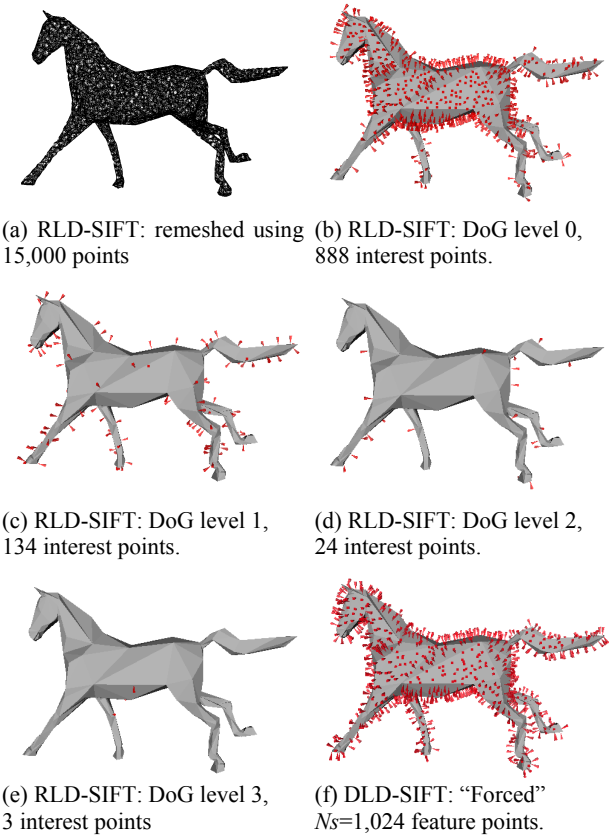
The RLD-SIFT has scale parameter C and image size I, in addition to number of points N_{RLD} and neighbor size k for remeshing. We evaluated effects of these parameters in the experiments.

Figure 4a shows the resampled mesh, which used N_{RLD}=15,000 and k=10, of the sparsely sampled horse (346 vertices) model. RLD-SIFT interest points computed from the resampled mesh are shown in Figures 4b~4e. Using the LD-SIFT, the original model produced only 36 interest points. However, the RLD-SIFT produced 888 interest points as shown in Figure 4b from the same sparsely sampled model. Number of interest points drops quickly at coarser (i.e., higher DoG level) DoG meshes as shown in the Figure 4b~4e.

3.2. Dense LD-SIFT (DLD-SIFT)

The *Dense LD-SIFT*, or *DLD-SIFT*, densely and uniformly generates "feature points" for local feature computation. We used the same point generation algorithm as for remeshing of the RLD-SIFT for this purpose. The number of local features per 3D model, N_{DLD} is a parameter. For the DLD-SIFT, local feature scale parameter S_i is indicated relative to the diameter of a minimum enclosing sphere of the 3D model to be compared. For example, $S_i = 1.0$ indicates a scale identical to the enclosing sphere of the model. Another parameter is the size I in pixels of square depth image for SIFT feature extraction.

Figure 4f shows an example of 1,024 DLD-SIFT "forced" feature points generated on the sparsely sampled 364 vertex model of horse shown in Figure 3b.



(a) RLD-SIFT: remeshed using 15,000 points

(b) RLD-SIFT: DoG level 0, 888 interest points.

(c) RLD-SIFT: DoG level 1, 134 interest points.

(d) RLD-SIFT: DoG level 2, 24 interest points.

(e) RLD-SIFT: DoG level 3, 3 interest points

(f) DLD-SIFT: "Forced" *Ns*=1,024 feature points.

Fig. 4. Remeshing for the RLD-SIFT (a) and RLD-SIFT interest points detected (b~e). The DLD-SIFT forces feature points (f).

3.3. Integrating local features for comparison

Comparing a 3D model described by using a set of local features can be costly if each of many local features is matched against each other among sets. To reduce cost of the comparison, proposed algorithms use BoF approach [10] to integrate the set of local features into a feature vector per 3D model. BoF approach converts, by vector quantization, each local feature into a *visual word* using a *visual codebook* having vocabulary size V. Frequencies of visual words are counted to generate a histogram; the histogram becomes the feature vector for the 3D model. The visual codebook is learned by clustering from local features extracted from 3D models in the database. In our implementation, we used k-means clustering. An optimal number of visual words vary depending on the dataset to be queried.

Distance D between a pair of feature vectors $\mathbf{x}_p$ and $\mathbf{x}_q$ is computed by using the symmetric version of *Kullback-Leibler Divergence* (*KLD*);

$$D(\mathbf{x}_p, \mathbf{x}_q) = \sum_{k=1}^{V} (x_{p,k} - x_{q,k}) \log\left(\frac{x_{p,k}}{x_{q,k}}\right) \quad (1)$$

4. EXPERIMENTS AND RESULTS

We conducted experimental evaluation of the proposed algorithms by using three standard 3D model retrieval benchmarks. In experiments below, we varied N_{RLD} and C for the RLD-SIFT and scale S_i for the DLD-SIFT to quantify their effects. There are other parameters. We compared the depth images sizes of I=21, 41, and 61, and chose I=21 for all the experiments shown below. Based on preliminary experiments, we chose N_{DLD}=1,024 for the DLD-SIFT as retrieval accuracy saturated at the point. Similarly, we chose N_{RLD}=15,000 and k=10 for the RLD-SIFT. Optimal size V of visual vocabulary depends on the database to be retrieved. We varied V in the range 500~5000 and chose best performing values of V for each of the LD-SIFT, RLD-SIFT and DLD-SIFT.

Experiments are performed by using three benchmark databases. The *McGill Shape Benchmark* (*MSB*) [6] is a set of highly articulated (non-rigid), watertight, densely-sampled, yet less geometrically varied/detailed models. MSB consists of 255 models divided into 10 classes. The *Princeton Shape Benchmark* (*PSB*) [9] contains models using varied shape representations, i.e., polygon soup, open manifold mesh, closed manifold mesh, etc., with high variance in sampling density. The PSB contains two subsets of 907 models each, the "training" set and the "test" set. We used the PSB test set partitioned into 92 classes for evaluation. The *Engineering Shape Benchmark* (*ESB*) [4] includes 867 models of mechanical parts divided into 45 classes. Models in the ESB are difficult target for interest point detection as they contain flat surfaces and sharp corners. While ESB models are watertight manifolds, sampling density vary significantly. Figure 5 shows examples of models from the three benchmark databases.

As index of retrieval accuracy, we use *R-precision*, which is a ratio, in percentile, of the models retrieved from the desired class C_k (i.e., the same class as the query) in the top R retrievals, in which R is the size of the class $|C_k|$.

Figure 6 shows effect of local feature scale parameter E on retrieval accuracy of the DLD-SIFT. The MSB, with its articulated models, prefer smaller scale, while the PSB with its mostly rigid objects having complex geometries prefers larger scale. Similar tendency is observed for the RLD-SIFT.

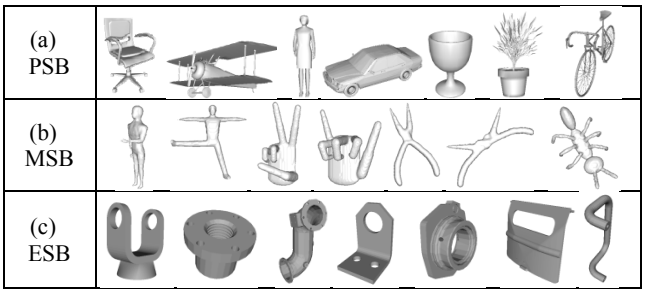


Fig. 5. Examples of benchmark database models.

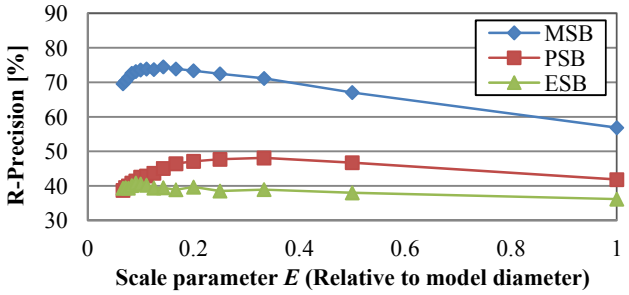


Fig. 6. DLD-SIFT scale parameter and retrieval accuracy.

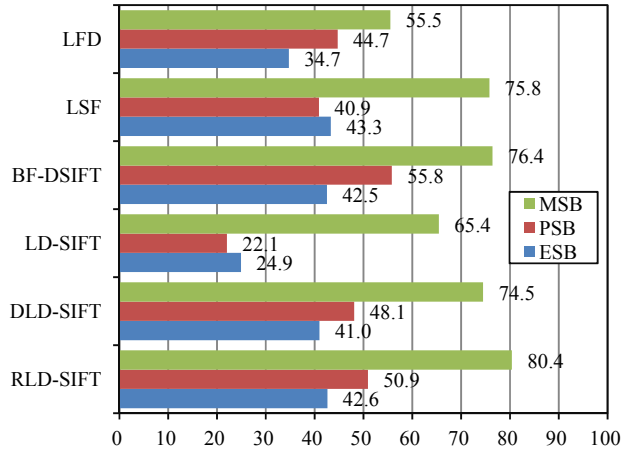


Fig. 7. Comparison of retrieval accuracy. (R-Precision [%])

We compared retrieval accuracies of the proposed algorithms, the LD-SIFT, DLD-SIFT and RLD-SIFT with three other algorithms, the *Light Field Descriptor* (*LFD*) [1], the *Bag-of-Features Dense-SIFT* (*BF-DSIFT*) [3], and the *Local Statistical Feature* (*LSF*) [7]. We used our own implementations for the LSF and the BF-DSIFT. Executable for the LFD was downloaded from the author's web site.

Figure 7 shows the R-Precision [%] for the 6 algorithms. For the MSB, RLD-SIFT did the best, with more than 4% margin in R-Precision over the quite capable BF-DSIFT and LSF. Both of the DLD-SIFT and RLD-SIFT did much better (with 10~15% margin) than the LD-SIFT, even though the LD-SIFT should handle densely sampled meshes of the MSB well. The PSB contains polygons soup and other "unruly" shape models, and the ESB contains high-aspect ratio polygons and low-sampling density meshes. For these benchmarks, the LD-SIFT didn't do well; it came in the last among the six algorithms. For the PSB, the BF-DSIFT did best, followed by RLD-SIFT and then DLD-SIFT. For the ESB, accuracies of the LSF, BF-DSIFT, and RLD-SIFT are about equal.

5. CONCLUSION

In this paper, we proposed two improvements to the Local Depth SIFT (LD-SIFT) [2], a local geometrical features for 3D model retrieval. The LD-SIFT can only handle a 3D shape represented as a densely sampled, singly connected mesh. To alleviate this weakness, one of the proposed features, Dense LD-SIFT (DLD-SIFT) used (forced) dense sampling on the input mesh, while the other Resampled LD-SIFT (RLD-SIFT) used dense remeshing followed by interest point detection. Our experimental evaluation using three benchmark databases showed that both DLD-SIFT and RLD-SIFT significantly outperform the original.

ACKNOWLEDGEMENTS

This research has been funded in part by the Ministry of Education, Culture, Sports, Sciences, and Technology of Japan (No. 18300068)

REFERENCES

[1] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, M. Ouh-young, "On Visual Similarity Based 3D Model Retrieval", *Computer Graphics Forum*, **22**(3), 223-232, 2003.

[2] Tal Darom, Yosi Keller, "Scale-Invariant Features for 3-D Mesh Models", *IEEE Trans. on Image Proc.*, **21**(5), (2012).

[3] T. Furuya, R. Ohbuchi, Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features, *Proc. ACM CIVR 2009*, (2009).

[4] S. Jayanti, Y. Kalyanaraman, N. Iyer, K. Ramani, "Developing an engineering shape benchmark for CAD models", *CAD*, **38**(9), 939-953, 2006.

[5] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *IJCV*, **60**(2), 91-110, 2004.

[6] McGill 3D Shape Benchmark, http://www.cim.mcgill.ca/~shape/benchMark/.

[7] Y. Ohkita, Y. Ohishi, T. Furuya, R. Ohbuchi, "Non-rigid 3D Model Retriecal Using Set of Local Statistical Features", *Proc. IEEE ICME 2012 Workshop on Hot3D,* 2012, 593-598, (2012).

[8] Press et al., Numerical Recipes in C –The art of Scientific Computing, *Cambridge Universe Press*, 1992, pp.309-315.

[9] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton Shape Benchmark", *Proc. SMI'04*, 167-178, 2004, http://shape.cs.princeton.edu/benchmark/.

[10] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images". *Proc. ICCV 2005*.