Identifying 3D Models by Matching Rendered Image and Depth Image with Using a New Combined Descriptor

Dayou Jiang¹*

¹ Department of Computer Science and Technology, Anhui University of Finance and Economics, Bengbu, Anhui, 233000, China
*Corresponding author's e-mail: 120200032@aufe.edu.cn

Abstract. This paper presented a 3D model identification method by matching pairs of rendered images and depth images. Firstly, the technique uses the HARRIS detector and PIIFD detector (partial intensity invariant feature descriptor) to detect feature points of rendered-depth image pair of shape and then uses LGHD (Log-Gabor Histogram Descriptor) to descript these feature points. Secondly, the shape identification processing is conducted, including normalizing shape pose, capturing multi-view pairs of rendered images and depth images, and image matching. Finally, 100 pairs of 3D models are used for the experiment. The experiment results show that the proposed method has an efficient performance on rendered-depth image matching and can be used for 3D model identification.

1. Introduction
The current content-based 3D model retrieval techniques can be mainly classified into four categories based on feature representation: feature-vector-based methods, statistic-based methods, graph-based methods, and view-based methods [1]. The view-based approach concept comes from two 3D models represented by a set of 2D images projected from multiple viewpoints, and these 2D images are evenly distributed on the view sphere [2].

Some global image feature descriptors are commonly used to describe the images. 2.5D spherical harmonics transformation and 2D shape histogram are proposed to retrieve 2D shape drawings [3]. Compared with global image features, local features have many advantages. For example, they are robust to occlusion and clutter and can distinguish large object databases. In recent years, many feature descriptors and detection techniques have been proposed, such as SIFT (scale-invariant feature transform) [4], SURF (Speeded-Up Robust Features) [5], Harris Corner Detector [6], MSER (Maximally Stable Extremal Regions) [7], BRISK (Binary Robust Invariant Scalable Keypoints) [8], and FAST (Features from Accelerated Segment Test) feature detector [9]. Many techniques have been proposed to adapt descriptors of SIFT/SURF to multispectral images from the perspective of descriptors. The EOH (Edge-Oriented Histogram) [10] exploits only edge points in local windows rather than all pixels and utilizes five bins for computing descriptors. It has a better matching performance on multispectral images than SIFT but does not assign a primary orientation to key points. The PIIFD is invariant to image rotation, partially invariant to image intensity, affine transformation, and viewpoint/perspective change [11]. The LGHD [12] describes the neighborhood of feature points combining frequency and spatial information using multi-scale and multi-oriented Log-Gabor filters, which is proposed to suit the task of matching feature points between images with non-linear intensity variations. The existing view-based 3D object identification algorithm mainly using depth information of 3D models [13-14]. The rendered images and depth images of 3D models from multiple viewpoints...
are combined to identify the 3D models in the paper. The rest of the article is organized as follows: a brief review of related work is given in section 2. The overview of the proposed method is given in section 3. Section 4 discusses the experimental results. Section 5 concludes the paper.

2. Related works

2.1. EOH descriptor
A feature descriptor captures the distinct intensity patterns around an interest point. Many feature descriptors have been proposed that are superior to patches in both dimensionality and tolerance against variations. The EOH represents the distribution of 5 types of edges in each local area called sub-regions. The key point region is split up into 16 sub-regions. These edges are categorized into five types, as shown in Table 1. Each local histogram contains five bins. The 80 elements are normalized and used as the vector of characteristics.

| types | horizontal | vertical | 45° | 135° | non-directional |
|-------|------------|----------|-----|------|-----------------|
| filters | 1 2 1      | -1 0 1   | 2 2 -1 | -1 2 2 | -1 0 1          |
|        | 0 0 0      | -2 0 2   | 2 -1 -1 | -1 -1 2 | 0 0 0          |
|        | 1 2 1      | 1 0 1    | 1 1 1 | 1 1 1 | 1 0 1          |

2.2. Keypoints detected using different feature detectors
The feature descriptors' performance is studied when used in pair of rendered and depth images' matching. Fig.1 shows the keypoints detected using different feature detectors for depth images. The blob detectors like MSER and BRISK detect a few keypoints as SIFT. The corner detector FAST, SURF, and LOG (Laplacian of the Gaussian) also find many interest points. The HARRIS sees more point features than FAST at the edge and corner. The PIIFD detects the maximum feature points result from using the Harris detector.

![Keypoints detected using different feature detectors](image)

3. Proposed work
In the experiments, the framework proposed by Mikolajczyk [15] was used to evaluate the performance of the different descriptors when they were used in the Rendered-Depth image matching. The performance measure is computed as follows:

\[
\text{Performance} = \frac{\text{Recall}_(\text{Alg.1})}{\text{Recall}_(\text{SIFT})}
\]
\[ \text{Recall} = \frac{N_{\text{corr\_match}}}{N_{\text{corr\_spond}}} \] (2)

The Recall is the ratio of correctly matched points to total correspondences. The \( N_{\text{corr\_spond}} \) represents the number of features detected in the given image, \( N_{\text{corr\_match}} \) represents the correctly matched features. The Recall_{(Alg.i)} describes the different algorithms, Recall_{(SIFT)} is the Recall value using SIFT algorithm. The performance value higher than one means that the evaluated algorithm obtains more interest points than those computed by SIFT in the given pair images. Fig.2 shows the keypoints matched in a pair of images by using these methods.

![Figure 2. Keypoints matched for rendered-depth image pairs using the existing algorithms.](image)

Under no rotation, the SIFT-EOH method has more matching pairs than SIFT. The SURF, PIIFD also offer better performance than SIFT and are similar to SIFT-EOH. Among there, the FAST-LGHD has the best performance. Considering the poor performance of HARRIS, BRISK and FAST in a couple of images, they were excluded from the performance comparison experiment of 100 pairs of images. The performances of algorithms to SIFT are shown in Fig.3.

![Figure 3. The performances of algorithms to SIFT.](image)
Inspired by the sound characteristics of the EOH descriptor for the multispectral image, EOH based algorithms are considered. These algorithms all use the corresponding feature detector to detect the feature points and then use edge-oriented histogram descriptors for image matching. Fig.4 shows the keypoints matched in a pair of images by using these methods.

![Figure 4. Keypoints matching for a rendered-depth image using the EOH based algorithms.](image)

The performances of these EOH based algorithms to SIFT are shown in Fig.5. The figure presents that the HARRIS-EOH and SURF-EOH have better performance than others.

![Figure 5. The performances of algorithms to SIFT.](image)

The original LGHD algorithm employs the FAST feature detector to detect the feature points at first. Here, HARRIS and PPIFD detectors first use LGHD as the feature descriptor for points feature detection. As shown in Fig.6, combining the HARRIS detector and PPIFD detector can detect more feature points than FAST.
4. Experimental Result

The proposed approach has been evaluated with a dataset containing 100 pairs of Rendered-Depth images. These images have been obtained using the 3D graphics toolkit, namely OpenSceneGraph [16]. The dataset of 3D models is from NTU 3D Model Benchmark ver.1 [17]. The toolkit reads standard 3D CG file formats (.obj, .3ds, .stl, etc.) into MATLAB plat. The camera is controlled using sphere orientation (elevation, azimuth, and yaw) or camera matrix. These images' image size is 400 by 400, and all models' camera site faces the models for obviously compared experimental efficiency. Some pairs of Rendered and Depth images from the model benchmark are shown in Table 2.

Table 2. Sample image pairs from models benchmark.

| Rendered Images | Depth Images | Rendered Images | Depth Images |
|-----------------|--------------|-----------------|--------------|
| ![Sample Images](image1) | ![Sample Images](image2) | ![Sample Images](image3) | ![Sample Images](image4) |

The performance of the algorithm for model identification is measured using the equation as below:

\[
\text{Accuracy} = \frac{N_{Identified}}{N_{Tested}} \times 100
\]  

(3)

\(N_{Identified}\) is the number of models to be correctly identified, \(N_{Tested}\) is the total number of models to be identified. In the experiment, whether the model has been correctly identified is measured by the maximum number of correctly matched points. For each model's Rendered image, if the model's corresponding depth image has the maximum number than other depth images of any other models, this model is correctly identified. Otherwise, it is false being recognized.

The identification processing is implemented subsequently in the following steps:
- Pose Normalization: Normalize 3D objects concerning the canonical coordinate frame to ensure that their mass centers coincide with the origin, here, using the PCA approach to normalize the 3D models [18].
- Capture Rendered-Depth image: Capture rendered-buffer and corresponding depth-buffer views on a given unit geodesic sphere's vertices whose mass center is also located in origin. Here, 24 different poses still exist for a normalized model, as shown in Fig.7.

![Figure 7. Sample of a pair of Rendered-Depth images of 24 poses.](image)

- Image matching: Using the proposed HARRIS-PIIFD-LGHD method for Rendered-Depth image pairs matching. Measure the dissimilarity between two 3D models by calculating the number of their matching pairs. Fig.8 shows the sample of matching with the different images.

![Figure 8. Sample of multi-view shape matching.](image)

Although Fig.8 (b) and Fig.8 (d) show many matching pairs, the number of matching pairs is still obviously less than Fig.8 (a) and Fig.8 (c). Table 3 shows the matching results using 100 pairs of images.

| Accuracy (%) | Under no rotation | Under rotation |
|--------------|------------------|----------------|
|              | 100              | 84             |

5. Conclusions
The paper presented a 3D model identification method using a new combined descriptor: HARRIS-PIIFD-LGHD for models' rendered-depth images. The experimental results demonstrated that the EOH-based descriptors have better performance than those without using EOH. However, this method is more complicated to calculate and is not suitable for model recognition in complex situations.
References
[1] Iyer, N., Jayanti, S., Lou, K., et al. (2005) Three-dimensional shape searching: state-of-the-art review and future trends. Computer-Aided Design., 37(5): 509-530.
[2] Li, L., Wang, H. Z., Chin, T. J., Suter, D., Zhang, S. S., (2010) Retrieving 3D CAD Models Using 2D Images with Optimized Weights. In:2010 3rd International Congress on Image and Signal Processing, Yantai. pp. 1586-1588.
[3] Pu, J. T., and Ramani, K., (2006) On visual similarity based 2D drawing retrieval. J. Computer-Aided Design., 38(3):249-259
[4] Lowe, D. G., (1999) Object recognition from local scale-invariant features. In: International Conference on Computer Vision, Corfu, Greece. pp. 1150–1157
[5] Bay, H., Ess, A., Tuytelaars, T., Gool, L.V., (2008) Speeded up robust features (surf). Comput Vision Image Underst., 110(3): 346-359
[6] Harris, C., Stephens, M., (1988) A combined corner and edge detector. In: Alvey Vision Conference, 15(50): 10-5244.
[7] Matas, J., Chum, O., Urba, M., Pajdla, T., (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British Machine Vision Conference, Cardiff, UK. pp. 384-396
[8] Leutenegger, S., Chli, M., Siegwart, R., (2011) BRISK: Binary Robust Invariant Scalable Keypoints. In Proceedings of the IEEE ICCV, Barcelona. pp. 2548-2555
[9] Rosten, E., Drummond, T., (2005) Fusing Points and Lines for High-Performance Tracking. In: Proceedings of the IEEE ICCV, Beijing. pp. 1508–1511
[10] Jain, A.K., Vailaya, A., (1966) Image Retrieval Using Color and Shape. Pattern Recognition. 29(8):1233-1244
[11] Chen, J., Tian, J., Lee, N., et al. (2010) A partial intensity invariant feature descriptor for multimodal retinal image registration. IEEE Trans. Biomed. Eng. 57:1707–1718.
[12] Aguilera, C. A., Sappa, A. D., Toledo, R., (2015) LGHD: A feature descriptor for matching across non-linear intensity variations. In:2015 IEEE International Conference on Image Processing (ICIP), Quebec City. pp. 178-181.
[13] Jain HP., Subramanian, A., (2010) Real-time upper-body human pose estimation using a depth camera. In: International Conference on Computer Vision. Berlin, Heidelberg. pp. 227-238.
[14] Hong, Y.Y., Kim, J.W., (2017) A 2d-view depth image-and cnn-based 3d model identification method. Applied Sciences. 7(10): 988.
[15] Mikolajczyk, K., Schmid. C., (2005) A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers. 27(10):1615–1630.
[16] OpenSceneGraph Public License (OSGPL). http://www.openscenegraph.org/
[17] Shilane, P., Min, P., Kazhdan, M., Funkhouser, T., (2004) The Princeton Shape Benchmark. In: Proceedings Shape Modeling Applications. Genova, Italy. pp. 167-178.
[18] Lian, Z. H., Godil, A., Sun, X. F., (2010) Visual Similarity based on 3D Shape Retrieval Using Bag-of-Feature. In:2010 Shape Modeling International Conference. Aix-en-Provence. pp.25-36.