3D Shape Retrieval from a Photo Using Intrinsic Image

Shoki Tashiro*, and Masaki Aono†
* E-mail: tashiro@kde.cs.tut.ac.jp
†E-mail: aono@tut.jp Tel: +81-532-44-6764
Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan

***Abstract*—In recent years, 3D shape objects have spread on the Internet. Using a 2D photo as a query for 3D shape retrieval is usually much easier than preparing a 3D shape object or drawing a 2D sketch. We propose a new method for photo-based 3D shape retrieval using a so-called "Intrinsic Image." Intrinsic Image enables us to separate a given 2D photo into "Reflectance" and "Shading" images. We have observed that during the separation, texture information is primarily captured by "Reflectance," while shape information is left within "Shading." After the separation, we employ Histogram of Oriented Gradients (HOG) to extract the feature vector from "Shading" images, and apply principal component analysis (PCA) to obtain robustness against rotation, which has been the biggest problem of HOG. We conducted experiments with a commonly available 3D shape benchmark, compared our proposed method with the previous methods, and demonstrated that our method outperformed them in terms of 1st-Tier, 2nd-Tier, and P@1.**

I. Introduction

Thanks to the popularity of 3D printers and 3D scanning devices such as Kinect, 3D shape objects have increasingly appeared on the Internet, and have been applied in many fields including computer-aided design and manufacturing (CAD/CAM), computer-aided architecture, computer-aided medical operation, and entertainment. Re-using existing 3D shape objects allows us to reduce the cost of creating them from scratch. Hence, efficient management and retrieval methods have been in great demand.

On the other hand, the multi-modality of the query input for retrieving 3D shape objects has become more and more important and diversified. In the past, preparing a 3D shape object as a query was the only means to search similar objects. By now, a variety of different ways for queries have been introduced, including 2D sketches and 3D point clouds [1]. In particular, sketch-based 3D shape retrieval methods have become popular because of the wide availability of a vast amount of sketch data through SHREC (Shape Retrieval) [1, 2]. By the same token, 2D photo-based query has another potential of adding multi-modality to 3D shape retrieval. An overview of content-based 3D shape retrieval is shown in Fig. 1.

In this paper, we focus on a 3D shape retrieval method from a photo by taking advantage of "Intrinsic Image" decomposition to extract 3D shape features. Preparing a photo is as easy as preparing a sketch because we can take pictures using the smartphones or the tablets equipping the camera. A photo has more information than a sketch because it has colors, textures, and optical shading information. Although devices capturing 3D information such as the Kinect are available, they are usually an expensive way to obtain 3D information if one attempts to capture data from large objects. This is why we focus on a photo as the query of 3D shape retrieval. For simplicity, we assume a photo has a clear background and a clear-looking object shape with colors and textures.

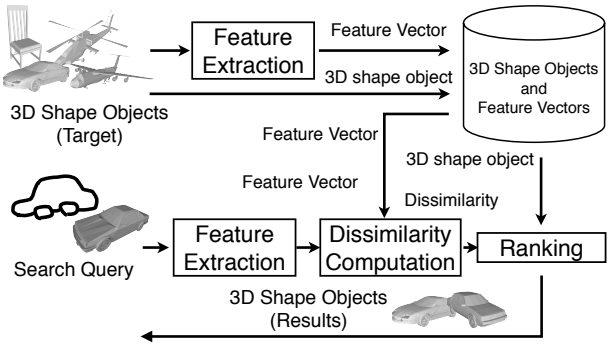

Fig. 1. Overview of content-based 3D shape retrieval

The difficulty of using a photo as the query for 3D shape retrieval lies in the fact that most objects have colors and textures, as well as different surface optical reflections under different lighting conditions. Recently, so-called "Intrinsic Images" [3, 4] are drawing attention, which is a technique to split a photo image into "Reflectance" and "Shading" components. Ideal "Reflectance" contains colors and textures with no shading information. In contrast, ideal "Shading" is supposed to have shape information estimated by gray-scaled values. Although 3D shape models do not always have apparent colors or textures, a photo naturally comes with them.

Our main idea in this paper is to extract "Shading" information by means of an "Intrinsic Image" from a given 2D photo, and to match "Shading" with computer-generated depth buffer images, hoping to achieve higher accuracy to search 3D shapes, similar to the object in a photo. We hypothesize that "Shading" has less texture information than the original image, and is easier to match with non-textured 3D shape objects. For the feature vector, we adopt the method proposed by Aono et al. [5], which employs HOG (Histogram of Oriented Gradients) [6] as features to match depth-buffer images. Their method suffers from rotational invariance of HOG, and attempted to alleviate the artifacts by adding 2D

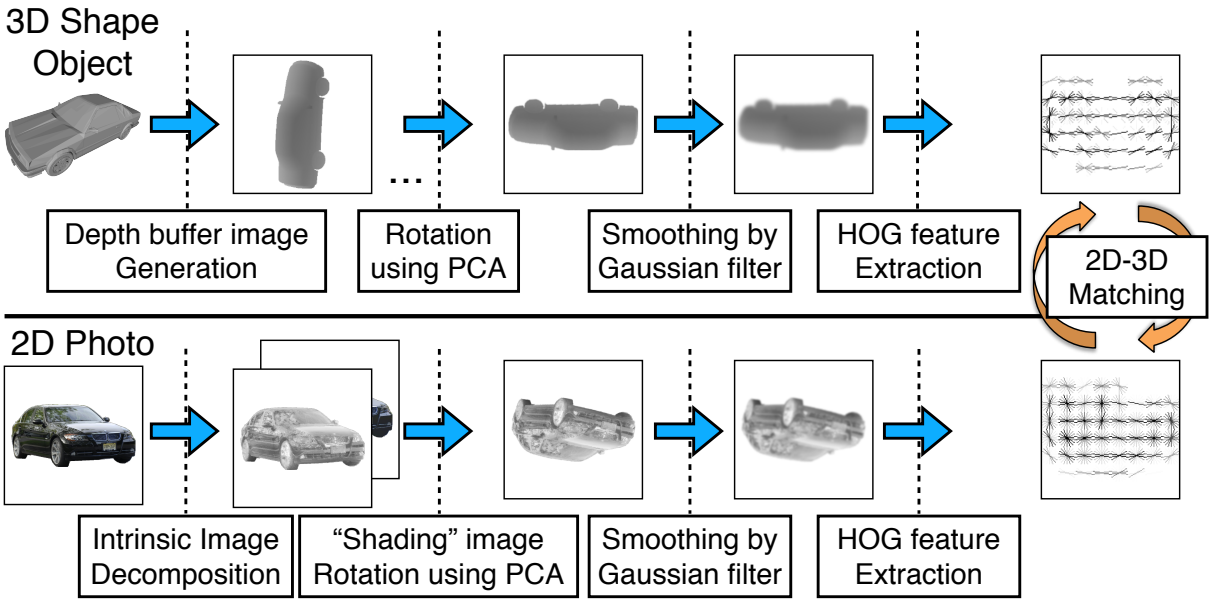

Fig. 2. Overview of our proposed system for 3D shape retrieval from a photo. Intrinsic Image decomposition is the main characteristic of the system.

rotated images in advance. In contrast, before extracting HOG features, we apply PCA to the input photo to solve the problem of estimating the orientation of the object. We conducted the experiment to demonstrate the effectiveness of an "Intrinsic Image," using the Princeton Shape Benchmark (PSB) [7].

II. Related Work

To our knowledge, Ansary et al. [8] were the first to propose a photo-based 3D shape retrieval method which employed Zernike moment features from the silhouette image generated from 320 views, applied adaptive view clustering, and performed probaility-based dissimilarity computation. Aono et al. [5] proposed another method for 3D shape retrieval from a 2D photo, based on depth-buffer images for extracting HOG features and silhouette images for extracting Zernike moment features. By utilizing depth-buffer images, their method could capture depth information in a 2D image and achieved higher search accuracy than Ansary et al. Daras et al. [9] proposed a 3D shape retrieval method which accepted 3D shape objects, sketch images, and a photo as the search query. Tatsuma et al. [10] proposed a benchmark data for photo-based 3D shape retrieval, consisting of 1875 gray-scale photos, one hundred 3D shape objects with 5 different labels, and 100 non-labeled 3D shape objects.

III. Intrinsic Image Based 3D Shape Retrieval from a 2D Photo

We propose a new method for 3D shape retrieval from a 2D photo, taking advantage of Intrinsic Image decomposition. Intrinsic Image decomposition aims to produce "Reflectance" and "Shading" which are the "intrinsic" properties of an image [3, 4]. An ideal "Reflectance" has color and texture information, but no optical shading information, while ideal "Shading" has optical shading information with no colors and textures. Our idea here is to focus on the "Shading" information, discarding "Reflectance," and attempting to extract shape information from the optical shading information. It should be noted that previous methods [5, 8] did not consider removing colors and textures from a photo, often suffering from the artifacts of colors and textures when matching the 2D image features with the features extracted from rendered images of a 3D shape object. The overall flow of our proposed method is illustrated in Fig. 2. The first step for the conversion from a 3D shape object to a 2D projected image is performed by depth buffer rendering. In the following, we describe the other steps in Fig. 2.

A. Intrinsic Image Decomposition

Initially, the 2D photo image as an input query is assumed to have not only optical shading, but color and texture information. The Intrinsic Image technique [4] makes it possible to isolate this information, and decompose into two separate images, "Reflectance" and "Shading." Let $\mathbf{R}$ denote "Reflectance" and $\mathbf{S}$ denote "Shading." The Intrinsic Image technique attempts to find the optimal $\mathbf{R}^*$ and $\mathbf{S}^*$, satisfying the following equation:

$$\mathbf{R}^*, \mathbf{S}^* = \arg\min_{\mathbf{R},\mathbf{S}} p(\mathbf{R}, \mathbf{S}|\mathbf{I})$$

where p is the probability distribution function and $\mathbf{I}$ is the 2D photo image, which satisfies $I_i = R_i \cdot S_i$ for pixel i. To find the optimal $\mathbf{R}^*$ and $\mathbf{S}^*$, the probability distribution function p proposed by Krähenbühl [11, 12], with a fully connected conditional random field, has been employed. The example of a given 2D photo and the "Shading" is shown in Fig. 3.

Fig. 3. Example of a 2D photo (left) and the "Shading"(right)

B. Image Rotation using PCA

After decomposing a given 2D photo into "Reflectance" and "Shading," we wish to generate HOG features from the "Shading" component. However, since HOG is not robust against rotation, previous approaches have suffered from the rotation invariance. For instance, Aono et al. [5] intentionally generated multiple rotated images to cope with this problem. We estimate the direction of the image orientation by means of PCA [13] instead of generating rotated images to achieve robustness against rotation. It should be noted that PCA is for estimating the image orientation without a priori learning, not for reducing dimensionality.

C. Smoothing Image by Gaussian Filter

We also consider removing unintentionally incurred noise in "Shading" during Intrinsic Image decomposition. To reduce such noise, we apply smoothing with a Gaussian filter of size 17×17 before extracting The HOG features.

D. HOG Extraction

After smoothing the image, we extract HOG features. A HOG feature vector is computed by the following steps:

1) Compute gradient magnitude m and orientation θ
2) Compute an orientation histogram cell by cell
3) Normalize the histogram block by block
4) Concatenate all blocks of histograms into one feature vector

Gradient magnitude m and orientation θ are defined by following formula:

$$\begin{aligned} m(x,y) &= \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \\ \theta(x,y) &= \tan^{-1}\frac{f_y(x,y)}{f_x(x,y)} \end{aligned}$$

where

$$\begin{aligned} f_x(x,y) &= L(x+1,y) - L(x-1,y) \\ f_y(x,y) &= L(x,y+1) - L(x,y-1) \end{aligned}$$

and $L(x,y)$ denotes the value at the pixel (x,y). The HOG feature vector tends to be very high dimensional. For instance, given a 256×256 image, assuming one cell consists of 16×16 pixels, one block has 3×3 cells, the number of the orientations of gradients is 9, and the total dimension of this HOG feature vector is $15876 (= 3^2 \times 14^2 \times 9)$. In our proposed method, we adopt 32×32 pixels for one cell, 2×2 cells for one block, and 9 bins for the orientation of gradients, resulting in 1764 dimensions in total.

E. 2D-3D Matching

After computing the feature vectors from each 3D shape object and the input photo, we compute their dissimilarity. Suppose $\mathbf{f}^{\mathcal{I}}$ is a feature vector of an input photo $\mathcal{I}$ and $\mathbf{f}_i^{\mathcal{M}}$ (where $i = 1, \cdots, n$) are feature vectors of a 3D shape object $\mathcal{M}$, where n is the number of viewpoint for rendering. We define dissimilarity D between $\mathcal{I}$ and $\mathcal{M}$ as follows:

$$D(\mathcal{M}, \mathcal{I}) = \min_{i=1,\cdots,n} d(\mathbf{f}^{\mathcal{I}}, \mathbf{f}_i^{\mathcal{M}})$$

where d is the dissimilarity measure. We choose Manhattan distance for computing d after trial and error.

IV. Experiments and Results

To demonstrate the effectiveness of our proposed method, we conducted experiments. We employed the Princeton Shape Benchmark (PSB) [7] for the 3D shape benchmark data, which includes 907 models with 92 labels. We selected 20 classes out of 92 in our experiments. For the input photo data, we collected from the Internet ten images for each class at random. For simplicity, a noisy background was eliminated manually when necessary. We evaluated the search performance of our proposed method using the evaluation measures including 1st-Tier, 2nd-Tier, P@1, Recall, and Precision. We compared our proposed method (using "Shading" with Gaussian Smoothing) with Zernike moment features [8], Fourier spectral features [9], and composite features of HOG and Zernike moment without PCA orientation estimation [5]. We set the dimension of the Zernike moment feature vector to 49 and the Fourier spectral feature to 1024 by applying a low-pass filter.

To generate the 2D representation of a 3D shape object, we perform multi-view rendering from 92 views, and produced a collection of depth-buffer images. Table I summarizes the comparison of 1st-Tier, 2nd-Tier, and P@1 with PCA. The corresponding Recall-Precision graphs are shown in Fig. 4. In Table I and Fig. 4, "ZM" denotes the Zernike moment feature, "FSF" denotes the Fourier spectral feature, and "HOG(N)+ZM" denotes the composite feature of HOG and Zernike moment without PCA orientation estimation nor smoothing by Gaussian filter.

TABLE I
Comparison in 1st-Tier, 2nd Tier, and P@1

Method	1st-Tier	2nd-Tier	P@1
Proposed	**0.3155**	**0.4177**	**0.6400**
ZM	0.1951	0.2758	0.3750
FSF	0.1969	0.2888	0.4000
HOG(N)+ZM	0.2359	0.3385	0.4500

From Table I and Fig. 4, we demonstrated that our method (HOG feature with PCA orientation estimation, Gaussian Smoothing, and "Shading" by Intrinsic Image decomposition) outperformed other methods.

V. Conclusion

In this paper, we proposed a method for 3D shape retrieval method from a photo as a query, focusing on the

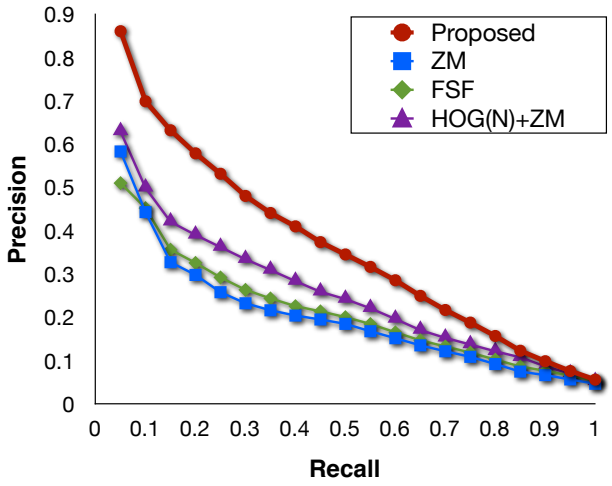

Fig. 4. Comparison in Recall-Precision graph

Intrinsic Image technique to decompose the photo image into "Reflectance" and "Shading." After extracting HOG feature from "Shading" image, we applied PCA to achieve rotation invariance. From the experiments with PSB, we demonstrated that our method outperformed previous methods for retrieving 3D shapes from a photo.

In the future, we will investigate further the feature extraction methods to obtain more search accuracy as well as the fully automatic background elimination methods.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-In-Aid (B) 26280038.

References

[1] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, Hongbo Fu, Takahiko Furuya, Haisheng Li, Jianzhuang Liu, Henry Johan, Ryuichi Kosaka, Hitoshi Koyanagi, Ryutarou Ohbuchi, Atsushi Tatsuma, Yajuan Wan, Chaoli Zhang, and Changqing Zou. A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, 131(0):1 – 27, 2015.

[2] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2012*, 31(4):31:1–31:10, Jul. 2012.

[3] Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1459–1472, Sep. 2005.

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159:1–159:12, Jul. 2014.

[5] Masaki Aono and Hiroki Iwabuchi. 3D shape retrieval from a 2D image as query. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–10, Dec. 2021.

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference*, volume 1, pages 886–893 vol.1, June 2005.

[7] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Proceedings of the Shape Modeling International 2004*, SMI '04, pages 167–178, 2004.

[8] Tarik Filali Ansary, Jean-Phillipe Vandeborre, and Mohamed Daoudi. On 3D retrieval from photos. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, 3DPVT '06, pages 687–694. IEEE Computer Society, 2006.

[9] Petros Daras and Apostolos Axenopoulos. A 3D shape retrieval framework supporting multimodal queries. *International Journal of Computer Vision*, 89(2-3):229–247, Sep. 2010.

[10] A. Tatsuma, S. Tashiro, and M. Aono. Benchmark for photo-based 3d shape retrieval. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–4, Dec 2014.

[11] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011.

[12] Philipp Kräehenbüehl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 513–521. JMLR Workshop and Conference Proceedings, May 2013.

[13] Jose M. Saavedra and Benjamin Bustos. An improved histogram of edge local orientations for sketch-based image retrieval. In *Proceedings of the 32Nd DAGM Conference on Pattern Recognition*, pages 432–441, 2010.