View-Based 3D Object Retrieval: Challenges and Approaches

Yue Gao and Qionghai Dai
Tsinghua University, China

View-based 3D object retrieval is an emerging research topic that has numerous geographic-related applications in many fields, such as CAD and virtual city navigation.

The rapid development of computer graphics hardware and 3D technologies has increasingly lead to the use of 3D objects in various applications,[1,2] especially in the entertainment, medical, and architectural design industries. As a result, the need for effective and efficient 3D object retrieval methods has increased significantly as well. For instance, 3D object retrieval can help reduce the costs of model design by nearly 80 percent in the CAD field.

In general, 3D object retrieval methods can be divided into one of two categories based on either 3D models or multiple views. In 3D model-based methods, each 3D object is represented by a virtual 3D model, which can be created using statistics-, extension-, volume-, or surface-geometry-based methods, all of which use the 3D model data. Many practical applications cannot obtain a 3D model, however, so a virtual 3D model must be reconstructed. This approach is computationally expensive, and the poor performance of reconstruction methods often results in low-quality 3D models.

View-based 3D object retrieval methods, on the other hand, use a single view or multiple views for 3D object representation. These views can be obtained with either a group of cameras or a virtual camera array. Figure 1 shows several example views used to describe 3D objects. Such view-based methods do not require a 3D model, and the ubiquity of mobile devices with cameras makes it easy to obtain images of real objects. Online multiview data of 3D objects have become increasingly available as many e-business websites, such as Amazon and eBay, provide multiple views for most of their products. Under these circumstances, it is possible to conduct a view-based object search. For location-based mobile applications, view-based methods also provide new search opportunities with the help of cameras. Compared with model-based methods, view-based methods is more discriminative for 3D objects,[3,4] which can lead to better object retrieval performance.[2,5,6]

A general view-based 3D object retrieval process consists of four stages: view capture, view selection, feature extraction, and object matching. Here, we focus on the recent progress in view-based 3D object retrieval, which has been widely used in CAD applications, for example. We first survey the key technologies and challenges in view-based 3D object retrieval and then discuss the state-of-the-art methods and future research directions in the field.

Retrieving 3D Objects with Multiple Views

We can define the view-based 3D object retrieval task as follows: Each object consists of one or more views, and given one query object, the objective is to find all relevant and/or similar objects from the 3D object database under the view-based representation.

View-based 3D object retrieval has several main challenges.

- *View capture*. Views are the fundamental elements for view-based 3D object analysis. Most existing methods use a camera array that consists of a group of cameras capturing views from different directions.
- *Representative view selection*. Although a large number of views can provide rich information, they also introduce redundant and noisy data and result in high computational costs.
- *Feature extraction*. It is still difficult to extract features for multiple views because of the special characteristics of 3D data.

1070-986X/14/$31.00 © 2014 IEEE Published by the IEEE Computer Society

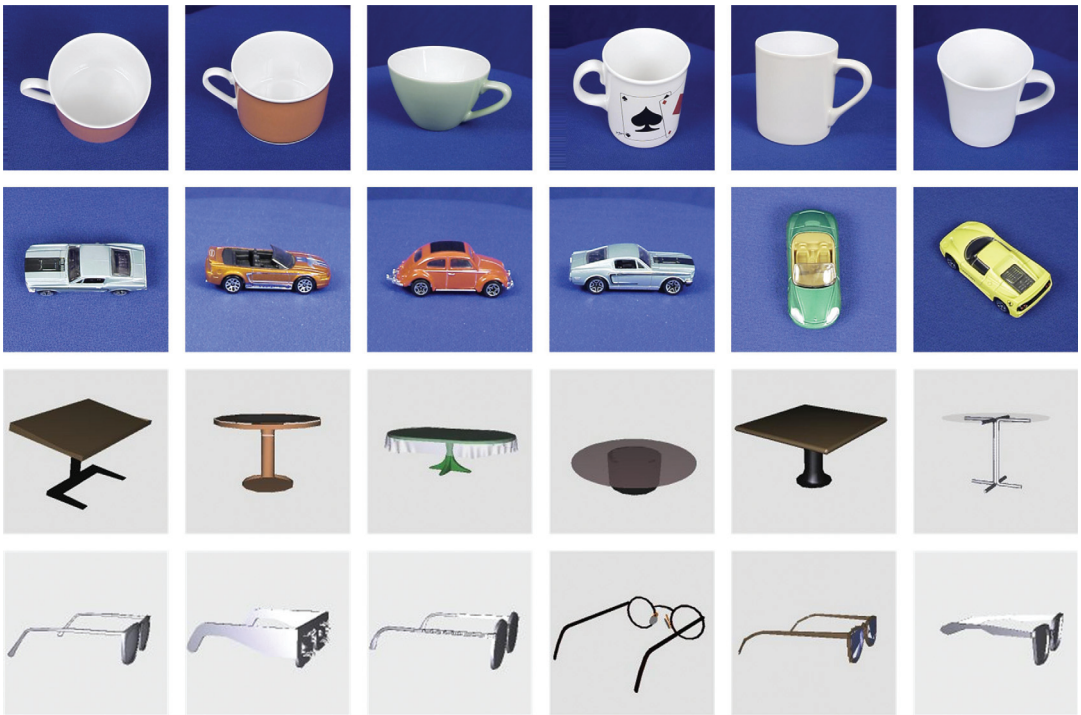

Figure 1. Example multiple views of 3D objects. View-based 3D object retrieval methods use a single view or multiple views for 3D object representation and do not require the 3D model.

The spatial correlation among different views should be taken into consideration, which still requires further investigation.

- *Object matching using multiple views*. Most of the existing image retrieval tasks are based on one-to-one image matching. View-based 3D object retrieval, however, focuses on multiple view matching. Thus, it is challenging to determine how best to conduct many-to-many view matching and estimate the relevance among different 3D objects.

View Capture and Representative View Selection

A compact group of views can provide adequate and concise information for 3D object description. View capture and representative view selection are two fundamental steps in view-based 3D object retrieval. Existing methods can be divided into four paradigms: generating representative views with a predesigned camera array,[7] generating representative views from a large view pool,[8–10] conducting synthesized view generation,[11] and performing incremental view selection.[12]

Predesigned Camera Array

In these methods, a camera array, consisting of tens to hundreds of cameras, is predesigned to generate multiple representative views of each 3D object. Common camera array configurations include regular dodecahedrons, cubes, and 32-hedrons. The number of representative views is determined by a predesigned camera array.

Lighting field descriptor (LFD)[7] is a typical predesigned camera array method that generates representative views by using 10 silhouettes obtained from the vertices of a dodecahedron over a hemisphere. This approach can describe the spatial structure from different directions. LFD uses Zernike moments and Fourier descriptors as the features of representative views. To determine the similarity between two 3D objects, this method finds the best match between two groups of LFD views, using the spatial structure in the matching procedure.

Petro Daras and Apostolos Axenopoulos proposed a compact multiview descriptor (CMVD) method that generates 18 characteristic views using 18 vertices of the corresponding bounding 32-hedron for each 3D object.[4] This method uses both binary and depth images. The two 3D models are compared by feature matching selected views using 2D features, such as 2D Polar-Fourier transform, 2D Zernike moments, and 2D Krawtchouk moments. The testing object is rotated to obtain the best matched direction for the query object. The minimal sum of the distance from the selected rotation

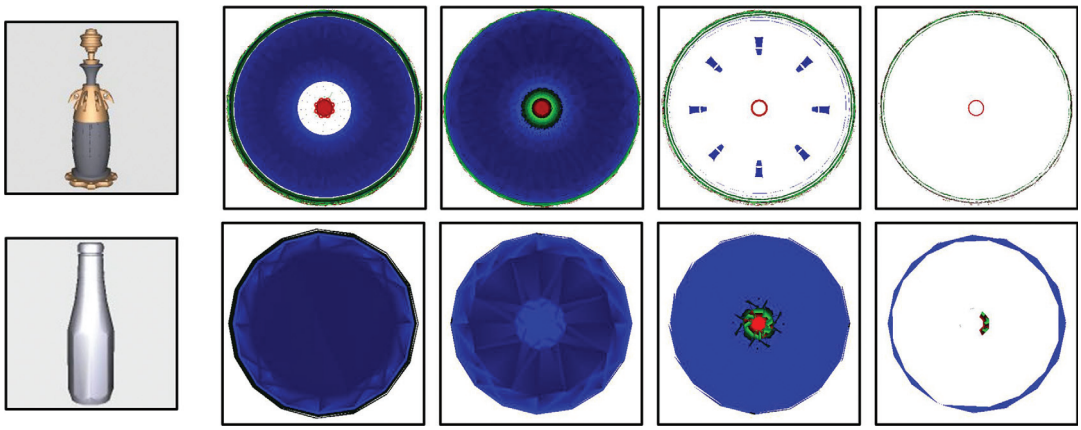

Figure 2. Example spatial structure circular descriptor (SSCD) images for 3D objects.[5]

direction is measured as the distance between the two compared objects.

Jau-Ling Shih and his colleagues proposed an elevation descriptor (ED) that extracts six range views for each 3D model from a bounding box.[13] The 3D model's altitude information from the six directions is employed as a description, and the 3D models are compared by matching the EDs. To match two groups of ED views, the minimal distance between each pair of views is calculated to measure the distance. Although the ED view information is compact (only six views), too much spatial information is lost, which limits the performance of ED for 3D object retrieval.

Representative View Selection from a Pool

In this type of method, a large group of initial views are first captured, and representative views are selected from this view pool.

A typical method of view selection from a pool is Adaptive Views Clustering (AVC).[9] The view pool contains 320 views. The view selection process involves using adaptive view clustering with Bayesian information criteria, and generally about 20 to 40 views are selected as the representative views for each 3D object. In this method, the object matching is achieved by a probabilistic method. Another representative method was presented by S. Mahmoudi and M. Daoudi.[14] In this method, seven views from three principal and four secondary directions are obtained to index objects, and contour-based features are extracted from each view for multiview matching.

Synthesized View Generation

In synthesized view generation methods, a view generation process is conducted to project the 3D model information to a synthesized view.

The panoramic object representation for accurate model attributing (Panorama) method uses panoramic views to represent surface information.[11] The panoramic view of a 3D model is obtained by projecting the model onto the lateral surface of a cylinder that is aligned with one principal axe and centered at the object's centroid.

The spatial structure circular descriptor (SSCD) synthesized view method is invariant to rotation and scaling.[15] The spatial information is represented by an SSCD that includes several SSCD images. In this method, a minimal bounding sphere of the 3D model is generated first, and all points from the model surface are projected to the bounding sphere. Attribute values represent the surface spatial data. The bounding sphere is further projected onto a plane's circular region, which is used to describe the surface information. Figure 2 shows example SSCD images.[15]

Incremental View Selection

This type of method incrementally selects the representative views from a large pool rather than selecting views in one round. Typically, this category involves query view suggestion (QVS),[12] which incrementally selects query views by using relevance feedback from users. In QVS, the view clustering is conducted on the pool of initial views, and one view is selected from each view cluster as the candidate query view. In the first round, one query view is selected to conduct 3D object retrieval. Then the user's relevance feedback is employed to select a new query view in each instance. A distance metric is learned for the newly selected

view, and all query views are combined with learned weights. In this way, incrementally selected query views can help guarantee the discrimination of selected query views.

Feature Extraction

Feature extraction is an important task for multiple view representation. Widely used features include 2D Polar-Fourier transform, 2D Zernike moments, 2D Krawtchouk moments, and Fourier descriptors. In recent years, bag-of-words (BoW) feature approaches have been used for image descriptions as well as for view-based 3D object retrieval. A general framework includes extracting local features, quantizing local features into a set of visual words with a pretrained dictionary, and generating BoW histograms. This method is robust to variances in occlusions, viewpoints, illumination, scale, and backgrounds. Generally, scale-invariant feature transform (SIFT) features are extracted from representative views of one 3D object, and a BoW feature is generated for a 3D model representation. To perform 3D model matching, the distance between two BoW features can be measured using either KL divergence or other distance metrics.

The BoW method was first used in view-based 3D model retrieval by Takahiko Furuya and Ryutarou Ohbuchi.[16] In this method, each 3D model is rendered into a set of depth images, and the SIFT features are extracted to generate a BoW feature for each 3D model. Two 3D models are matched using the distance between the two BoW features.

Ohbuchi and his colleagues further proposed using Kullback-Leibler divergence (KLD) to measure the distance between two BoW features in 3D object retrieval.[17] In this method, a semisupervised manifold learning method is used for model class recognition, which projects the original feature space onto a lower-dimension manifold. The relevance feedback is then employed to capture the semantic class information by using the manifold ranking. The BoW method is computationally expensive. Furuya and Ohbuchi later introduced an accelerated method by employing a graphics processing unit, and they used dense sampling to extract feature points in 3D model views.[18]

Object Matching with Multiple Views

For object matching, the general distance measures, such as Euclidean, Minkowski, and Mahalanobis distances, can be used in view-based 3D object retrieval. However, these conventional approaches usually adopt simple methods to integrate the distances from view pairs across two objects, while ignoring the higher-order information among these multiple views. We first summarize some typical distance measures here.

Let O_1 and O_2 denote two sets of views from two objects; v' and v'' denote the views from these two sets, respectively; and $d(.,.)$ indicate the distance between two views.

The *average distance* is the average of all distances between any view pairs across the two compared objects, which can be written as

$$D_{\text{avg}}(O_1, O_2) = \frac{1}{|O_1||O_2|} \sum_{v' \in O_1} \sum_{v'' \in O_2} d(v', v'') \tag{1}$$

The *minimal distance* measures the minimal distance from all view pairs. It is defined as

$$D_{\min}(O_1, O_2) = \min_{v' \in O_1 v'' \in O_2} d(v', v'') \tag{2}$$

The *Hausdorff distance* is the longest of all distances from a point in one set to the closest point in the other set. It is defined as

$$\begin{aligned} D_{\text{Haus}}(O_1, O_2) \\ = \max\{&\max_{v' \in O_1}\{\min_{v'' \in O_2} d(v', v'')\} \\ &\max_{v' \in O_2}\{\min_{v'' \in O_1} d(v', v'')\}\} \end{aligned} \tag{3}$$

The *sum-min distance* is the sum of all distances from a point in one set to the closest point in the other set. It is defined as

$$D_{\text{sum\_min}}(O_1, O_2) = \frac{1}{|O_1|} \sum_{v' \in O_1} \min_{v'' \in O_2} d(v', v'') \tag{4}$$

The matching schemes between two groups of views have been thoroughly investigated in recent years using bipartite graph formulation, probabilistic matching, and learning-based methods. In one approach, the two compared groups of views are formulated in a bipartite graph structure, and the weighted bipartite graph matching is conducted to measure the distance between 3D objects.[6] This method first selects representative views from the pool of initial views with corresponding weights. These selected views are used to construct a weighted bipartite graph, and the proportional max-weighted bipartite matching is based on this graph.

To better estimate the relevance among multiple views, probabilistic matching frameworks

have been introduced in view-based 3D object retrieval.[8,9] One approach clusters all views for the query object first,[8] and these view clusters are used to train the query Gaussian models. This method trains both positive and negative matching models using positive and negative matched samples, respectively, which are further combined to conduct object relevance estimation.

Researchers have also investigated learning-based methods for view-based 3D object retrieval, which can learn the underneath structure to determine the global relevance of the whole dataset. The hypergraph formulation is a typical learning-based method in which the relationship among 3D objects is formulated in a hypergraph structure, where each vertex denotes one view from one object.[10] The edges in the hypergraph are generated via view clustering, where each view cluster generates one edge, and all objects with views in this cluster are connected by the edge. Given the query object, a semisupervised learning process is conducted in this structure to learn the relevance score for each object to the query. With training samples, this method can be further expanded to view-based 3D object classification.

Although typical distance measures can be applied in view-based 3D object retrieval, they are limited in their ability to model the relationship among multiple views. Probabilistic methods[8,9] can generate better results than typical distance measures and benefit from probabilistic matching between two groups of multiple views.[10] Learning-based method have the advantage of using the underneath structure among all compared 3D objects. However, this type of method is limited by computational costs. Also, for a large-scale 3D object database, it is hard to learn the global information using only one big graph structure.

Future Directions

Although significant progress has been achieved in view-based 3D object retrieval, many challenges still require further investigation, including large-scale data management, feature extraction, multiview matching, multimodal data. and geolocation-based applications.

Large-Scale Data

The development of 3D technologies has resulted in a rapidly increasing volume of 3D object data. This large-scale 3D object data presents added challenges for view-based 3D object retrieval. Determining how to handle large-scale data and building efficient indices will attract much research attention in the future.

Feature Extraction

Feature extraction is a common but hard issue for multimedia information retrieval. With the recent progress in deep learning methods, a possible way to further enhance existing features is to optimally learn the features from the original 3D data with deep learning.

Multiview Matching

Although many multiview matching algorithms exist for 3D object retrieval, it is still challenging to learn the relationship among 3D objects at the semantic level using multiple views. Thoroughly investigating the correlation among multiple views from different directions is important for 3D object retrieval as well as other applications, such as 3D reconstruction, classification, and calibration.

Multimodal Data

The sheer number of data capture applications and devices has lead to diverse data-generation methods. Mobile devices with cameras, such as tablets and smartphones, can capture images any time and anywhere, and their power continues to grow. With the rapid development of social media, sites such as Flickr, Facebook, Foursqure, and Twitter are enabling increasing amounts of user-generated content. Different sensors are also providing data from diverse implementations, such as surveillance and remote imaging systems. This multimodal information brings with it rich data for 3D object analysis, which in turns creates challenges for cross-media and cross-platform data processing.

Geolocation-Based Applications

There is a great deal of potential for geolocation-based applications that can take advantage of view-based methods, such as mobile- and location-based search. Individual users and surveillance systems generate enormous amounts of geolocation-based imagery. Together with geolocation-based Web content, such as Twitter tweets and Foursquare check-ins, a massive amount of regional, multimodal data is being produced. All these data provide new opportunities for the next stage in mobile search. For example, intelligent cities have become a hot

topic in recent years. In this context, multiple-view computing includes tracking, model reconstruction, recognition, and virtual city navigation. Other applications include location-based recommendations and online social media analysis. All these applications can benefit from view-based object analysis, which can contribute rich information from multimodal data resources. The main challenges in these applications for view-based object retrieval relate to multimodal and multiresource data management.

Conclusion

View-based 3D object retrieval is an essential topic with many emerging applications. The next stage of research in this field will need to not only focus on the key technologies for view-based object retrieval but also extend it to general domains, which can certainly benefit from the achievements of view-based object analysis.

MM

References

1. A. Bimbo and P. Pala., "Content-Based Retrieval of 3D Models," *ACM Trans. Multimedia Computing, Comm., and Applications*, vol. 2, no. 1, 2006, pp. 20–43.
2. B. Bustos et al., "Feature-Based Similarity Search in 3D Object Databases," *ACM Computing Surveys*, vol. 37, no. 4, 2005, pp. 345–387.
3. Y. Gao et al., "3D Object Retrieval with Hausdorff Distance Learning," *IEEE Trans. Industrial Electronics*, vol. 61, no. 4, 2014, pp. 2088–2098.
4. P. Daras and A. Axenopoulos, "A 3D Shape Retrieval Framework Supporting Multimodal Queries," *Int'l J. Computer Vision*, vol. 89, no. 2, 2010, pp. 229–247.
5. P. Shilane et al., "The Princeton Shape Benchmark," *Proc. Shape Modeling Int'l*, 2004, pp. 1–12.
6. Y. Gao et al., "3D Model Retrieval Using Weighted Bipartite Graph Matching," *Signal Processing: Image Comm.*, vol. 26, no. 1, 2011, pp. 39–47.
7. D.Y. Chen et al., "On Visual Similarity Based 3D Model Retrieval," *Computer Graphics Forum*, vol. 22, no. 3, 2003, pp. 223–232.
8. Y. Gao et al., "Camera Constraint-Free View-Based 3D Object Retrieval," *IEEE Trans. Image Processing*, vol. 21, no. 4, 2012, pp. 2269–2281.
9. T.F. Ansary, M. Daoudi, and J.P. Vandeborre, "A Bayesian 3D Search Engine Using Adaptive Views Clustering," *IEEE Trans. Multimedia*, vol. 9, no. 1, 2007, pp. 78–88.
10. Y. Gao et al., "3D Object Retrieval and Recognition with Hypergraph Analysis," *IEEE Trans. Image Processing*, vol. 21, no. 9, 2012, pp. 4290–4303.
11. P. Papadakis et al., "Panorama: A 3D Shape Descriptor Based on Panoramic Views for Unsupervised 3D Object Retrieval," *Int'l J. Computer Vision*, vol. 89, nos. 2–3, 2010, pp. 177–192.
12. Y. Gao et al., "Less Is More: Efficient 3D Object Retrieval with Query View Selection," *IEEE Trans. Multimedia*, vol. 11, no. 5, 2011, pp. 1007–1018.
13. J. Shih, C. Lee, and J. Wang, "A New 3D Model Retrieval Approach Based on the Elevation Descriptor," *Pattern Recognition*, vol. 40, no. 1, 2007, pp. 283–295.
14. S. Mahmoudi and M. Daoudi, "3D Models Retrieval by Using Characteristic Views," *Proc. Int'l Conf. Pattern Recognition*, vol. 2, 2002, pp. 457–460.
15. Y. Gao, Q. Dai, and N. Zhang, "3D Model Comparison Using Spatial Structure Circular Descriptor," *Pattern Recognition*, vol. 43, no. 3, 2010, pp. 1142–1151.
16. T. Furuya and R. Ohbuchi, "Dense Sampling and Fast Encoding for 3D Model Retrieval Using Bag-of-Visual Features," *Proc. ACM Int'l Conf. Image Video Retrieval*, 2008, article no. 26.
17. R. Ohbuchi et al., "Salient Local Visual Features for Shape-Based 3D Model Retrieval," *Proc. IEEE Shape Modeling Int'l Conf.*, 2008, pp. 93–102.
18. R. Ohbuchi and T. Furuya, "Scale-Weighted Dense Bag of Visual Features for 3D Model Retrieval from a Partial View 3D Model," *Proc. IEEE ICCV 2009 Workshop on Search in 3D and Video*, 2009, pp. 63–70.

Yue Gao is with the Department of Automation at the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing. His research interests include large-scale multimedia retrieval and live social media analysis. Gao has a PhD in control theory and engineering from Tsinghua University. He is a senior member of IEEE and a member of ACM. Contact him at kevin.gaoy@gmail.com.

Qionghai Dai is a professor in the Department of Automation at the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing. His research interests include signal processing, broadband networks, video processing, and communication. Dai has a PhD in Computer science and automation from Northeastern University, China. He is a senior member of IEEE. Contact him at qionghaidai@tsinghua.edu.cn.