SHREC’15 Track: 3D Object Retrieval with Multimodal Views

Yue Gao†, Anan Liu†, Weizhi Nie†, Yuting Su†, Qionghai Dai†, Fuhai Chen†, Yingying Chen‡, Yanhua Cheng§, Shuilong Dong‖, Xingyue Duan‖, Jianlong Fu‖, Zan Gao‖, Haiyun Guo‡, Xin Guo‡, Kauqi Huang§, Rongrong Ji§, Yingfeng Jiang§, Haisheng Li§, Hanqing Lu§, Jianming Song§, Jing Sun§, Tieniu Tan§, Jinqiao Wang§, Huanpu Yin‖, Chaoli Zhang§, Guotai Zhang§, Yan Zhang§, Yan Zhang§, Chaoyang Zhao§, Xin Zhao† and Guibo Zhu‡.

†School of Medical, University of North Carolina at Chapel Hill, USA.
‡School of Electronic Information Engineering, Tianjin University, China.
§Department of Automation, Tsinghua University.
‖School of Information Science and Engineering, Xiamen University, China.
§Department of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China.
‡National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China.
§College of Computing & Digital Media, DePaul University, United States.
‖School of Computing, Tianjin University of Technology, China.
§School of Computer Science and Information Engineering, Beijing Technology and Business University, China.

Abstract

This paper reports the results of the SHREC’15 track: 3D Object Retrieval with Multimodal Views, which goal is to evaluate the performance of retrieval algorithms when multimodal views are employed for 3D object representation. In this task, a collection of 505 objects is generated and both the color images and the depth images are provided for each object. 311 objects are selected as the queries and average retrieval performance is measured. The track attracted six participants and the submission of 26 runs, to two tasks. The evaluation results show a promising scenario about multimodal view-based 3D retrieval methods, and reveal interesting insights in dealing with multimodal data.

1. Introduction

View-based 3D object retrieval aims to retrieve 3D objects which are represented by a group of multiple views. Most of existing methods start from 3D model information, while it is hard to obtain the model information in real world applications. In the case where no 3D model is available, a 3D model construction procedure is required to generate the virtual model via a collection of images for model-based methods. We notice that 3D model reconstruction is computationally expensive and that its performance is highly restricted to sampled images, which severely limits practical applications of model-based methods.

With the widely applied color and/or depth visual information acquisition devices, such as Kinect and mobile devices with cameras, it becomes feasible to record color and/or depth visual information for real objects. In this way, the application of 3D object retrieval can be further extended to real objects in the world. Starting from the Lighting Field Descriptor [CTSO03a] at 2003, much research attention has focused on view-based methods in recent years. Ankerst et al. [AKKS99] proposed an optimal selection of 2D views from a 3D model, which focuses on numerical characteristics obtained from the 3D model representative features. Shih et al. [SLW07] proposed Elevation Descriptor (ED) feature, which is invariant to translation and scaling of 3D models. However, it is not suitable for 3D model which consists of a set of 2D images. Tarik et al. [ADV07] pro-

© The Eurographics Association 2015.

DOI: 10.2312/3dor.20151066
posed a Bayesian 3D object search method, which utilizes 
X-means [CTSO03b] to select characteristics views and ap-
plies Bayesian model to compute the similarity between dif-
ferent models. Gao et al. [GWT∗12] proposed a hypergraph 
learning method for 3D object retrieval, in which the re-
vance among 3D objects is formulated in a hypergraph struc-
ture. Although extensive research efforts have been dedi-
cated to view-based 3D object retrieval, it is still a challenging 
task from the following aspects: view acquisition and selec-
tion, feature extraction and object distance measure.

In the track of 3D Object Retrieval with Multimodal Views,
we aim to concentrate focused research efforts on this in-
teresting topic. The objective of this track is to retrieve 3D 
objects by using multimodal views, which are color images 
and depth images for each 3D object. Our collection is com-
piled of 505 objects, in which 311 objects are selected as 
the queries. Six groups were participated in this track and 
26 runs were submitted for two tasks. The evaluation results 
show a promising scenario about multimodal view-based 3D 
retrieval methods, and reveal interesting insights in dealing 
with multimodal data.

2. Dataset and Queries
A real world 3D object dataset with multimodal views, 
Multi-view RGB-D Object Dataset (MV-RED)†, is collect-
ed for this contest. The MV-RED dataset consists of 505 ob-
jects, which can be divided into 60 categories, such as apple,
cap, scarf, cup, mushroom, and toy. For each object, both 
RGB and depth information were recorded simultaneously 
by 3 Microsoft Kinect sensors from 3 directions. That is,
there are two types of imaging data, i.e., RGB and depth,
for each object.

This dataset was recorded using with three Kinect sensors 
(the 1st generation) but under two different camera settings,
as shown in Fig.1(a) and Fig.1(b), respectively. 202 objects 
were recorded using the first camera array and 303 object-
s were recorded using the other one. For data acquisition,
Camera 1 and Camera 2 captured 360 RGB and depth im-
gages respectively by uniformly rotating the table controlled 
by a step motor. Camera 3 captured only one RGB image 
and one depth image in the top-down view. Using this set-
ing, 721 RGB images and 721 depth images can be cap-
tured for each object. For each RGB and depth image, the 
image resolution is 640 × 480. We then uniformly sampled 
the images from Camera 1 and 2 with the step of 10 degrees 
and a compact dataset with 73 RGB and 73 depth images for 
each object is generated. Foreground segmentation results 
for RGB images are provided.

All these 505 objects belong to 60 categories. Here the 
categories containing no less than 10 objects are selected as 
the queries, leading to 311 queries in total. In our track, t-
wo 3D object retrieval tasks are launched, which employ the 
complete version and the concise version of data respective-
ly. In each task, these 311 objects are used as the query ob-
ject once. The contest consists of two versions, i.e., retrieval 
on the whole dataset (721 views) and the compact dataset 
(73 views).

Figure 1: The recorded scene for each object.

3. Evaluation
To evaluate the performance of all participated methods,
the following evaluation criteria, which have been widely 
employed in existing 3D object retrieval works [CTSO03a,
GWJ∗14, SLW07], are employed.

1. Precision-Recall Curve comprehensively demonstrates 
retrieval performance; it is assessed in terms of average 
recall and average precision, and has been widely used in 
multimedia applications.
2. NN evaluates the retrieval accuracy of the first returned 
result.
3. FT is defined as the recall of the top τ results, where τ is 
the number of relevant objects for the query.
4. ST is defined as the recall of the top 2τ results.
5. F-Measure (F) jointly evaluates the precision and the re-
call of top returned results. In our experiments, top 20 
retrieved results are used for F1 calculation.
6. Normalized discounted cumulative gain (NDCG) is a s-
statistic that assigns relevant results at the top ranking po-
sitions with higher weights under the assumption that a 
user is less likely to consider lower results.
7. Average normalized modified retrieval rank (ANMRR) is 
a rank-based measure, and it considers the ranking infor-
mation of relevant objects among the retrieved objects. A 
lower ANMRR value indicates a better performance, i.e., 
relevant objects rank at top positions.

In this paper, all of evaluation results are based on distance 
matrices submitted by all of participators.

4. Participants
Six groups participated in this track and 26 runs were sub-
mitted. The participant details and the corresponding con-
tributors are shows as follows.

† http://media.tju.edu.cn/mvred/
1. GMM-Zernike and GMM-HoG submitted by Zan Gao, Guotai Zhang, Yan Zhang, Yingfeng Jiang and Jianming Song from Tianjin University of Technology, China.
2. IVA-Deep4 and IVA-DeepColor submitted by Haiyun Guo, Jinqiao Wang, Chaoyang Zhao, Yingying Chen, Jianlong Fu, Guibo Zhu and Haoming Lu from National Laboratory of Pattern Recognition, China.
3. BGM-Color and BGM-HoG submitted by Xin Guo, Jing Sun and Xingyue Duan from the College of Computing & Digital Media, DePaul University, USA.
4. CAS-ECR, CAS-ECKM and CAS-ECSR submitted by Xin Zhao, Yanhua Cheng, Kaiqi Huang and Tieniu Tan from Center for Research on Intelligent Perception and Computing, China.
5. XMU-GS and XMU-GS-FB submitted by Rongrong Ji, Yan Zhang and Fuhai Chen from Xiamen University, China.
6. ZFCE-BoF and ZFCE-MVM submitted by Haisheng Li, Shuilong Dong, Huanpu Yin, Chaoli Zhang from Beijing Technology and Business University, China.

The brief summarization is provided in Table 1.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Method Name</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tianjin University of Technology</td>
<td>GMM-Zernike</td>
<td>Graph Matching</td>
</tr>
<tr>
<td>GMM-HoG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Laboratory of Pattern Recognition</td>
<td>IVA-Deep4</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>Institute of Automation</td>
<td>IVA-DeepC</td>
<td></td>
</tr>
<tr>
<td>Chinese Academy of Sciences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The College of Computing &amp; Digital Media</td>
<td>BGM-Color</td>
<td>Gaussian Model</td>
</tr>
<tr>
<td>DePaul University</td>
<td>BGM-HoG</td>
<td></td>
</tr>
<tr>
<td>XMU-GS</td>
<td></td>
<td>Greedy Search</td>
</tr>
<tr>
<td>XMU-GS-FB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xiamen University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center for Research on Intelligent Perception</td>
<td>CAS-CSR</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>and Computing Institute of Automation</td>
<td>CAS-ECKM</td>
<td></td>
</tr>
<tr>
<td>Chinese Academy of Sciences</td>
<td>CAS-ECSR</td>
<td></td>
</tr>
<tr>
<td>Beijing Technology and Business University</td>
<td>ZFCE-BoF</td>
<td>Spatial Distance</td>
</tr>
<tr>
<td>ZFCE-MVM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Methods

5.1. 3D Model Retrieval based on GMM by Tianjin University of Technology (GMM-Zernike/GMM-HoG)

Each 3D object is represented by a view set to convey the 3D structure information through the relationships among such views. Give the query object Q, the retrieval task is to find the matched objects from all of dataset. Let VQ = {v1Q, ..., vmQ} denote the view set of the query object Q with m views, and let VC = {v1C, ..., vmC} denote the view set of object C in the MV-RED dataset with m views. Here, let Δ denote the binary variable related to two hypotheses: Δ = 1 indicates that C is relevant to Q and Δ = 0 otherwise. Until now, the similarity between Q and M is defined as the following likelihood ratio:

\[ S(Q, C) = p(C|Q, Δ = 1) - p(C|Q, Δ = 0), \]  

where \( p(C|Q, Δ = 1) \) denotes the probability of M given Q when C is relevant to Q and \( p(C|Q, Δ = 0) \) denotes the probability of C given Q when C is not relevant to Q. The next task is to train \( p(C|Q, Δ = 1) \) and \( p(C|Q, Δ = 0) \) by using the testing dataset. Finally, Eq. 1 is used to handle the model retrieval problem.

In this track, each object provides RGB image and depth images. Thus, Zernike moment feature is extracted from each RGB image and Hog feature is extracted from each depth image, leading to a 49-D Zernike moment feature vector and a 81-D HoG feature vector, respectively. Here, the hierarchical agglomerative clustering method is employed to group all query views into clusters. One representative view is then selected from each cluster, and only the representative views are used for retrieval. It is noted that this procedure is also conducted for each object in the testing database.

A Gaussian model is learned to model the feature distribution in each cluster. Let x be the feature of the training view; the model can be defined as:

\[ p(q|x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(d(x, \mu_i))^2}{2\sigma_i^2}\right), \]  

where \( g_i(a|\mu_i, \sigma_i^2) \) denotes the i-th Gaussian component, \( w_i \) indicates the weight of the i-th Gaussian component, and n is the number of Gaussian models. The probability of one view belonging to the i-th Gaussian component is calculated by:

\[ g_i(a|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(d(x, \mu_i))^2}{2\sigma_i^2}\right), \]  

where \( d(x, \mu_i) \) is the Euclidean distance between x and \( \mu_i \), \( \mu_i \) and \( \sigma_i \) are the parameters for the Gaussian model. It is noted that, generally, there are quit a few training samples. Therefore, each gaussian component is generated as follows. For the i-th Gaussian component \( p(q) = \sum_{i=1}^n w_i g_i(a|\mu_i, \sigma_i^2) \), the parameters are leaned by:

\[ w_i = \frac{n_i}{n_{all}}, \]  

\[ \mu_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \psi_k^Q, \]  

\[ \sigma_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (d(\psi_k^Q - \mu_i))^2. \]  

where \( n_{all} \) is the total number of views of the query object, \( n_i \) is the number of views in the i-th cluster, and \( \psi_k^Q \) is the feature vector of views in the cluster. According to these learning processes, the parameters of \( p(C|Q, Δ = 1) \) and \( p(C|Q, Δ = 0) \) can be learned. The best retrieval result
should satisfy the following objective function:

\[ r = \arg \max_i p(C|Q, \Delta = 1) - p(C|Q, \Delta = 0). \quad (7) \]

In our results, two groups of experimental results using Zernike moment feature and HoG feature, i.e., GMM-Color and GMM-HoG, were submitted.

5.2. Learning Multiview Deep Feature by National Laboratory of Pattern Recognition (IVA-Deep4/IVA-DeepColor)

CNN was first introduced by LeCun [LS88] in the early 1990s and has shown record-breaking performance in many visual recognition tasks. The general pipeline of CNN feature extraction has two steps: the first step is to train CNN model in a supervised way; then deep features can be extracted from the last several layers of CNN. For this contest, three kinds of CNN features are extracted with three different CNN models respectively. Figure 2 shows the overview of the multiview deep CNN features.

![Figure 2: Overview of multiview deep CNN features for 3D object retrieval.](image)

Specifically, a 19-layer deep CNN model is used, which is pre-trained on ILSVRC’12 to classify each image into 1000 classes to extract the first kind of CNN structure features. On the one hand, this kind of CNN features can deliver rich semantic and structure information and intuitively suppress background noise. On the other hand, it is not quite sensitive to color information, which is rather crucial to object retrieval. In addition, color information is an effective supplement for structure and shape features. Therefore, the crawled color image dataset from Google is further utilized to learn a deep color CNN model with 10 dominant colors to extract color features which not only deliver rich color information but also are robust to light change.

The above two CNN features are both extracted from RGB images. However, apart from RGB image, depth image is another important view to describe 3D object, especially the information of shape and distance. To transform raw depth maps into efficient CNN features before encoding, the depth image is represented with an image with three channels at each pixel. Afterwards, the CNN pre-trained on RGB images can be adapted to generate powerful CNN features for depth images. This kind of deep depth CNN features involves rich shape and structure information.

Since rich semantic, color, shape and depth features have been extracted from each view of one 3D object, each feature is projected into similarity metric space and the similarity score for each view can be obtained. Then these complementary multi-view deep CNN scores can be combined by a weighted fusion scheme to obtain more comprehensive and accurate retrieval results. The experiments show that the deep depth features obtain a low F-measure scores than deep color features and deep structure features. The reason is that the depth images are very small for each object due to the object is small, and the depth information is not obvious for different objects such as “Apple” and “Orange”. While the deep color and deep structure features achieve better results with the fc7 output, they could effectively capture the semantic, color and shape information.

5.3. 3D Model Retrieval based on Bipartite Graph Matching by DePaul University (BGM-Color/BGM-HoG)

As there are too much redundant information in multiple views, especially in 721 views for each object, the original 2D images of each object need to be clustered by taking advantage of both visual and spatial information to remove the redundancy. The rule for image clustering is to maximize the inner-class correlation while minimizing the inter-class correlation. Consequently, the view-constrained clustering method can be formulated as an energy minimization problem. The objective function consists of two parts, data terms and smooth terms and can be defined as:

\[ C' = \arg \max C \sum_{i=1}^{m} E(v_i) + \sum_{i,j=1}^{m} E(v_i, v_j) \neq j, v_i, v_j \in C, \quad (8) \]

where \( E(v_i) \) represents energy of view \( i \), which term represents the contribution of this view for this cluster \( C; E(v_i, v_j) \) represents the correlation between different views. If two different views \( v_i \) and \( v_j \) belong to \( C; E(v_i, v_j) \) should have a higher value. The sum of \( E(v_i, v_j) \) and \( E(v_i) \) represents the entire energy of one specific clustering strategy.

Thus, \( E(v_i) \) measures the agreement between cluster \( C \) and the observed data \( v_i \). It can be computed by:

\[ E(v_i) = D_1(f_i, f_{center}), \quad (9) \]

where \( f_{center} \) represents the feature of center point in \( C; f_i \) represents feature of \( v_i \); \( D_1(f_i, f_{center}) \) represents similarity between \( v_i \) and \( v_{center} \), which is computed by Euclidean distance. \( E(f_i, f_j) \) affects the correlation among \( v_i \) and \( v_j \) in \( v_{center} \). It can be formulated by:

\[ E(v_i, v_j) = E(v_i) \cdot E(v_j) \cdot D_2(v_i, v_j), \quad i \neq j \quad (10) \]

where \( E(v_i) \) and \( E(v_j) \) are computed according to Eq.9; \( D_2(f_i, f_j) \) represents similarity between \( v_i \) and \( v_j \), which is computed by:

\[ D_2(v_i, v_j) = D_1(f_i, f_j) \cdot D_2(v_i, v_j), \quad (11) \]

where \( D_1(f_i, f_j) \) is the computed by Euclidean distance.
$D_i(v_i, v_j)$ represents the spatial similarity between different two views, which is computed by spherical distance between $v_i$ and $v_j$. The centre of the sphere is the center of this 3D model.

Finally, Eq. 8 can be converted to:

$$C' = \underset{C}{\text{arg max}} \sum_{i=1}^{m} D_1(f_i, f_{center}) + \sum_{i,j=1}^{m} E(v_i) \cdot E(v_j) \cdot D_2(v_i, v_j)$$

subject to $i \neq j$, $v_i, v_j \in C$  

(12)

After the above processes, the original clustering problem has been successfully converted into one Energy Maximization problem. Graph cut is applied to get a set of sub-clusters.

Here the Kuhn Munkres method [Kuh55] is employed to solve the problem. As the Kuhn Munkres method aims to solve the maximal matching problem, the object function should be modified. First an $n \times n$ edge costs matrix $C$ is created, where $c_{ij} = W - w_{ij}$, and $W > w_{ij}$. The missing edges (similarity value is zero) are given a large cost($W$). Using the above definitions, the objective function of the max-weighted bipartite matching is changed to the following equation:

$$\lambda_M = \underset{\Lambda_M \in \Lambda_1^m \times \Lambda_2^n}{\text{arg max}} \sum_{i,j=1}^{\Lambda_1^m \times \Lambda_2^n} (W - w_{a(i), b(i)})$$

(13)

Given a bipartite graph $G = \{U, V, E\}$ and an $n \times n$ edge cost matrix $C$, the Hungarian algorithm will output a complete max-weighted bipartite matching $M_{\text{Match}}$ [CCGR10]. The bipartite matching results are used to compare two 3D objects.

5.4. 3D Model Retrieval Based on Greedy Search by Xiamen University (XMU-GS/XMU-GS-FB)

In this method, three types of features are extracted for each image, including 49-D Zernike moment [Hu62], 120-D Fourier descriptor [Bra65], and BoWs. The main idea is to formulate the relationship between two 3D objects using three bipartite graphs, which are constructed using the three features respectively. The detailed algorithm is introduced as follows.

Each object is described by a set of views $\{V_1, V_2, ..., V_n\}$, and the SIFT feature is extract on the dense sampling points. The size of employed vocabulary is $N_v = 512$. Then each view can be represented by an $N_v$ dimension vector. To capture the shape information, Fourier descriptor and Zernike moment, are extracted from each image respectively, leading to one $n \times 120$ matrix $M_{FD}$ and one $n \times 49$ matrix $M_{Zernike}$.

To compare two 3D objects $O_1$ and $O_2$, the corresponding feature matrices, $M_1 = \{f_1^1, f_2^1, ..., f_n^1\}$ and $M_2 = \{f_1^2, f_2^2, ..., f_n^2\}$, can be generated first, where $f_j^1$ represents BoW feature for each view. The Euclidean distance is used to measure the distance between $f_j^1$ and $f_j^2$. Then a $n^3 \times n^2$ matrix $M^T$ can be achieved to represent the relationship between $O_1$ and $O_2$. Eq. 14 is utilized to compute the view matching results in different feature space between $O_1$ and $O_2$.

$$X^* = \underset{X}{\text{arg max}} \sum_{x=1}^{X} X \cdot M^T$$

subject to $X \in \{0, 1\}^{n \times n^2}$

(14)

where greedy algorithm is leveraged to handle this optimization problem to get the best matching results $X$. According to different matching results in different feature space, Eq. 15 is used to generate the final matching score.

$$S = \sum_{i,j=1}^{(\lambda_1 M_{BoW} + \lambda_2 M_{FD} + \lambda_3 M_{Zernike})}$$

$$M_{BoW} = X_{BoW} \odot M_{BoW}$$

$$M_{FD} = X_{FD} \odot M_{FD}$$

$$M_{Zernike} = X_{Zernike} \odot M_{Zernike}$$

(15)

where $\lambda_1 = 0.014$, $\lambda_2 = 0.98$ and $\lambda_3 = 0.006$ is the weight for different feature matrix, $S$ is the final matching score, which is used to represent similarity between $O_1$ and $O_2$. 3D object retrieval is based on the matching score $S$ between the query object and the objects in the database.

In XMU-GS-RF, the user relevance feedback information is introduced in the retrieval process, where top 10 returned results are manually labeled as relevant or irrelevant to the query. Then the top 100 returned results are re-ranked by using the minimal distance to the labeled positive samples and the query.

5.5. Enhanced CKM by Center for Research on Intelligent Perception and Computing (CAS-ECKM)

CKM [BSWR12] adapts a single-layer feature learning networks based on K-means clustering for 2D images [CNL11]. To keep the feature learning process as effective as [CNL11], CKM takes the depth channel as the fourth channel of the RGB channels and directly learns features from the four channels. By using the state-of-the-art image pre-processing and feature encoding of [CNL11], CKM can obtain useful translational invariance of low-level features from raw data such as edges, and can be robust to small deformations of objects. However, it is experimentally shown find that extracting features from RGB modality and depth modality individually and fusing their SVM classifiers can make CKM more powerful. Furthermore, the two derived data modalities, gray-scale and surface normals, can provide additional advantages for object recognition. In the end, RGB and gray-scale were combined to capture visual appearance of the RGB view, while depth and surface normals were leverage to capture shape cues of the depth view. The framework of the enhanced CKM is shown in Fig. 3.

The enhanced CNN-RNN method is proposed based on the original CNN-RNN model [SLNM11] [CZHT14]. CNN-RNN mainly consists of three steps: resizing all the images to the same scale, extracting low level feature for each image by a single convolutional layer, and finally applying multiple fixed-tree RNNs to learn high order feature representation based on the low level feature responses. Although CNN-RNN can learn powerful features from the raw data, such artificial processing of the first step, i.e., resizing all the images to the same scale by simply cropping or warping the images, may degrade the performance of the learned features. In order to adopt CNN-RNN for images of arbitrary sizes, the first step of CNN-RNN is replaced by a spatial pyramid matching layer together with a re-organization step, as shown in Fig.5. SPM can split each feature map into multiple subregions, and aggregate the responses in each subregion by max-pooling in the algorithm. The number of subregions determine the output size regardless of the variable input sizes of feature maps, then the fixed-tree RNNs can compose the fixed-size re-organization feature maps to high order features as [SLNM11, CZHT14]. CNN-SPM-RNN is employed to extract features for each modality of RGB, gray-scale, depth and surface normals, respectively. For each object, the RGB feature and gray-scale feature are concatenated to represent the appearance information, while depth feature and surface normal feature are combined to capture shape cues.

5.7. CNN-SPM-RNN by Center for Research on Intelligent Perception and Computing (CAS-CSR)

CNN-SPM-RNN [CNL11] is building on the unsupervised feature learning structure of CNN-RNN [SLNM11] [CZHT14]. CNN-RNN mainly consists of three steps: resizing all the images to the same scale, extracting low level feature for each image by a single convolutional layer, and finally applying multiple fixed-tree RNNs to learn high order feature representation based on the low level feature responses. Although CNN-RNN can learn powerful features from the raw data, such artificial processing of the first step, i.e., resizing all the images to the same scale by simply cropping or warping the images, may degrade the performance of the learned features. In order to adopt CNN-RNN for images of arbitrary sizes, the first step of CNN-RNN is replaced by a spatial pyramid matching layer together with a re-organization step, as shown in Fig.5. SPM can split each feature map into multiple subregions, and aggregate the responses in each subregion by max-pooling in the algorithm. The number of subregions determine the output size regardless of the variable input sizes of feature maps, then the fixed-tree RNNs can compose the fixed-size re-organization feature maps to high order features as [SLNM11, CZHT14]. CNN-SPM-RNN is employed to extract features for each modality of RGB, gray-scale, depth and surface normals, respectively. For each object, the RGB feature and gray-scale feature are concatenated to represent the appearance information, while depth feature and surface normal feature are combined to capture shape cues.

5.8. BoF and MVM Method by Beijing Technology and Business University (ZFCE-BoF/ZFCE-MVM)

This method extracts four features from each binary image: Zernike moments feature, Fourier feature, Circularity feature, Eccentricity feature, and the four features compose the hybrid shape descriptor ZFCE. Noted that binary image is expressed as view in the following subsections. This method uses two strategies to achieve the similarity computation for a query, which is Bag-of-Feature (BoF) approach and multiple view matching (MVM) in each angle.

BoF: 3D model can obtain global feature by BoF approach about the view feature of Zernike moments and
Fourier. To calculate global feature, method generates a codebook of visual words in advance. The visual word is thus defined as the center of a cluster obtained by applying K-means clustering to the view features, which are extracted from 3D models’ view sets in the MV-RED dataset (505 models). K-means clustering is performed with K=512. Then, the frequency histogram of vector quantized view features into visual words becomes a global feature vector for the Target dataset model. Finally, k-nearest-neighborhood algorithm is adopted to gain the global feature of the Query dataset (311 models) model by counting the number of view feature, which falls into the corresponding visual word.

This method combines the 4 features by linear weight, and the weights of Zernike moments feature, Fourier feature, Circularity and Eccentricity can be set as 0.2, 0.3, 0.2, 0.3 and 0.3, 0.4, 0.1, 0.2 for concise version and complete version respectively.

MV/M: For each angle, 4 features are used to calculate similarity distance between query model and test model. In addition, three typical distance measures (Minimal distance, Average distance, Hausdorff distance) are used to calculate similarity distance between two different models.

Average distance:

\[
D_{\text{ave}}(O_1, O_2) = \frac{1}{|O_1||O_2|} \sum_{v' \in O_1} \sum_{v'' \in O_2} d(v', v''),
\]

Hausdorff Distance:

\[
D_{\text{haus}}(O_1, O_2) = \max \left\{ \max_{v' \in O_1} \min_{v'' \in O_2} d(v', v''), \max_{v'' \in O_2} \min_{v' \in O_1} d(v', v'') \right\},
\]

where \(O_1\) and \(O_2\) denote the view sets of two objects, \(v'\) and \(v''\) denote the views in these two sets, and \(d(v', v'')\) indicates the distance between two views. Hausdorff distance [DJ94] is used in Zernike moments feature, while Average distance is used in rest features. As for \(d(v', v'')\), Manhattan distance is employed in Zernike moments feature and Fourier feature, and Euclidean distance is employed in Circularity and Eccentricity feature.

The matching algorithm can be described specifically as follows: first, for each feature in each angle, the proposed method calculates similarity distance of the view set respectively and the similarity distance is 0 when the view set of a angle does not exist. Then this approach gains similarity distance of two models by summing the 4 angles’ similarity distances based on a feature. Noted that here the summed similarity distance will be multiplied by 73/37 for concise version or 721/371 for complete version if the compared two models are under different recording settings. Finally, this approach combines the 4 features by linear weight, and the weights of Zernike moments feature, Fourier feature, Circularity and Eccentricity can be set as 0.5, 0.3, 0.1, 0.1 and 0.5, 0.4, 0.1, 0.0 for concise version and complete version respectively.

6. Results

In this section, we present the results of the six groups that submitted 26 runs for two tasks on the compact dataset and the complete dataset, respectively. Fig. 6 and Fig. 7 demonstrate the quantitative evaluation results from MV-RED-73 and MV-RED-721 respectively. Fig. 8 and Fig.9 show the Precision-Recall curves from MV-RED-73 and MV-RED-721 respectively.
4. The results using 721 images do not have significant improvement than the results using 73 views for almost all the methods. For some methods, the performance is even degraded when more views are employed. This observation demonstrates that more data not only provide more information, but also introduce noise data, which may have negative impact on 3D object representation.

7. Conclusion

In conclusion, this track has attracted research attention on 3D object retrieval using multimodal views. It is a challenging task and all the data in the testing dataset come from real objects. We have six groups who have successfully participated in the track and contributed 26 runs for 2 tasks. This track serves as a platform to solicit the existing view-based 3D object retrieval methods. Also all the participated methods have achieved improved performance, the task is still challenging and the results are far from satisfactory and practical applications. There is still a long way for view-based 3D object retrieval.

8. Acknowledgements

We would like to express our deepest gratitude to Zhongyang Wang, Xixi Li, Qun Cao, Xiaorong Zhu, Ning Xu, Fan yu, Yang Li, Xiaoxue Li, Yaoyao Liu, Fuwu Li, Yang Shi, Yahui Wang, Xixi Li, Qun Cao, Xiaorong Zhu, Ning Xu, Fan yu, based 3D object retrieval. The authors from Tianjin University was supported in part by the National Natural Science Foundation of China (61472275, 61170239, 61202168, and 61303208), the Tianjin Research Program of Application Foundation and advanced Technology, the grant of Elite Scholar Program of Tianjin University.

References


