View-Based 3-D Model Retrieval: A Benchmark

An-An Liu, *Member, IEEE*, Wei-Zhi Nie, Yue Gao, *Senior Member, IEEE*, and Yu-Ting Su

***Abstract*—View-based 3-D model retrieval is one of the most important techniques in numerous applications of computer vision. While many methods have been proposed in recent years, to the best of our knowledge, there is no benchmark to evaluate the state-of-the-art methods. To tackle this problem, we systematically investigate and evaluate the related methods by: 1) proposing a clique graph-based method and 2) reimplementing six representative methods. Moreover, we concurrently evaluate both hand-crafted visual features and deep features on four popular datasets (NTU60, NTU216, PSB, and ETH) and one challenging real-world multiview model dataset (MV-RED) prepared by our group with various evaluation criteria to understand how these algorithms perform. By quantitatively analyzing the performances, we discover the graph matching-based method with deep features, especially the clique graph matching algorithm with convolutional neural networks features, can usually outperform the others. We further discuss the future research directions in this field.**

***Index Terms*—3-D model retrieval, benchmark, deep learning, graph matching.**

I. Introduction

RECENTLY, view-based 3-D model retrieval is becoming important in diverse domains [9], [24], [28], [36], [40], [47], [52], such as computer-aided design, digital entertainment, medical diagnosis, e-business, and location-based mobile applications since the rapid development of computer graphics hardware and 3-D technologies for modeling, reconstruction, printing and so on have produced increasing number of 3-D models [2], [31], [63]. Given a query 3-D model represented by a group of multiview 2-D images [10], [22], [53] or sketch [11], [38], [62] or range image [54], 3-D model retrieval aims to find the relevant models from the 3-D model database [13]. Especially, when only partial shape information is available, this task will be much more difficult [9]. Although this task has been studied for several years [35], [45], [69], it is still quite challenging due to the variation of viewpoints, illuminations, model sizes, model styles, the number of multiview 2-D images, etc. Therefore, it is critical to concurrently evaluate the state-of-the-art methods on the popular datasets to discover their strength and weakness and help to identify future research directions in this field.

Manuscript received April 6, 2016; revised November 29, 2016; accepted January 23, 2017. Date of publication February 15, 2017; date of current version February 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61472275, Grant 61502337, and Grant 61100124, in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 15JCYBJC16200, in part by the China Scholarship Council under Grant 201506255073, and in part by the Elite Scholar Program of Tianjin University under Grant 2014XRG-0046. This paper was recommended by Associate Editor M. Last. *(Corresponding authors: An-An Liu; Wei-Zhi Nie.)*

A.-A. Liu, W.-Z. Nie, and Y.-T. Su are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: liuanan@tju.edu.cn; weizhinie@tju.edu.cn).

Y. Gao is with the School of Software and TNList, Tsinghua University, Beijing, China

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2664503

A. Motivation

Although the current methods have demonstrated superior performance on this task, there still exist three key problems.

1) Although many novel methods [21], [39], [52], [63] have been proposed for 3-D model retrieval in recent years, to the best of our knowledge, few work has been done for systematic evaluation of the state of the arts.
2) Most of the current methods leverage the hand-crafted visual features for model retrieval. To the best of our knowledge, there are few methods which explore the feasibility of deep feature learning on this task.
3) Currently, there are two kinds of datasets for the evaluation in this area. ETH is the most popular real world 3-D model dataset for this task. However, it is only a small scale dataset with 80 models and is not challenge enough for this task. NTU60, NTU216, and PSB are the most popular virtual 3-D model dataset for this task. However, there is lack of toolbox for multiview image capturing with the preset virtual camera array.

Therefore, the current situation has serious negative influence on the large-scale and systematic performance evaluation of multiple features and algorithms.

B. Contributions

The main contributions are summarized as follows.

1) *Systematic Evaluation:* We systematically investigate and evaluate this task by: a) proposing a clique graph-based method and b) reimplementing six representative method for 3-D model retrieval, including two distance-based methods, three statistical model-based methods, and one graph matching-based methods.
2) *Deep Feature Learning:* We systematically evaluate the deep features for 3-D model retrieval since it is essential to discover whether the deep learning technique, which has achieve significant success in more and more areas in computer vision, can facilitate the development of this new field.
3) *Dataset:* We build a challenging real world 3-D model dataset, multiview 3-D model dataset (MV-RED). For fair comparison on the dataset with virtual models (such as NTU60, NTU216, and PSB), we
    a) We build a challenging real world 3-D model dataset, multiview 3-D model dataset (MV-RED).
    b) For fair comparison on the dataset with virtual models (such as NTU60, NTU216, and PSB), we

develop a toolbox, which can capture the multiview 2-D images of one virtual 3-D model for the view-based retrieval.

The rest of this paper is structured as follows. The related work is introduced in Section II. The methods, datasets, and evaluation criteria are presented in Sections III–V, respectively. The experimental results will be detailed in Section VI. Section VII concludes this paper.

II. Related Work

View-based 3-D model retrieval mainly contains two key techniques, feature representation and similarity measure of pairwise models [5], [12], [27], [51], [59], [68]. We will briefly review the related work from both sides.

A. Feature Representation

For view-based 3-D model retrieval, each 3-D model usually contains a large group of initial captured multiview images. This view set not only provide rich information but also produce redundancy. To overcome this limitation, it is important to select characteristic views from a pool of views [26]. The objective of view selection is to determine a small set of compact but discriminative views for 3-D object representation. Existing view selection methods can generally be divided into two categories: unsupervised methods and interactive methods. In the unsupervised methods, a common procedure is to conduct view clustering to select characteristic views from each view cluster. Ansary *et al*. [1] applied X-means to adaptively extract characteristic views. Since the unsupervised methods usually neglect the discriminative objective of view selection, the interactive view selection method is developed to employ the user relevance feedback information to discriminatively select views [67]. Gao *et al.* [20] first extracted the candidate views in the unsupervised manner and showed them to the users for confirmation. The relevance feedback information is employed to update the characteristic views based on the samples labeled as being relevant and irrelevant. In this framework, the characteristic views can be incrementally selected in an interactive manner. With the characteristic views, it is necessary to represent them in specific feature spaces. Several methods [18], [42] often utilize Zernike moments [33] as the visual representation since Zernike moments are a class of orthogonal comments and their rotation invariance is important for shape representation. Daras and Axenopoulos [8] proposed a novel compact multiview descriptor, which is invariant to articulation and global deformation of 3-D models. Gao *et al.* [15] proposed the spatial structure circular descriptor, that can preserve the global spatial structure of 3-D models and is invariant to rotation and scaling. Li and Johan [37] took advantage of Zernike moments, Fourier descriptors, depth information, and Ray-based features (ZFDR) [61] to design a hybrid 3-D shape descriptor ZFDR. Shih *et al*. [56] proposed the elevation descriptor, which is invariant to translation, rotation and scaling. Most recently, Nie *et al.* [48] implemented the convolutional neural networks (CNN) features for view representation and significantly augmented the performance of 3-D model retrieval.

B. Similarity Measure

The similarity measure methods can be roughly classified into three categories.

1) The distance-based method directly measures the similarity between pairwise models in terms of specific distance metrics [66]. The Euclidean distance [7] and the Hausdorff distance (HAUS) [29] are two popular metrics for 3-D model retrieval. Recently, Gao *et al.* [19] proposed to learn a view-level Mahalanobis distance to estimate the HAUS between pairwise models.
2) The statistical model-based method takes advantages of the feature distribution of multiview images to generate the statistical model [60], [65], which can be utilized to measure the comparability between a query model and a specific category. Ansary *et al*. [1] applied Bayesian model to compute the similarity between different models. Gao *et al*. [18] proposed a camera constraint-free view-based method (CCFV) for 3-D mode retrieval. The proposed CCFV model can be generated on the basis of the query Gaussian models by combining the positive matching model and the negative matching model. This method can remove the constraint of static camera array settings for view capturing and can be applied to any view-based 3-D model database.
3) The graph matching-based method is widely applied to leverage the multiview information and the latent context for 3-D model retrieval [44], [71]. In [21], the weighted bipartite graph was built with the characteristic views and the matching results were used to measure the similarity between two 3-D models. Liu *et al.* [42] designed a second-order graph for 3-D model representation with both node-wise and pair-wise attributes and realized graph matching by mathematically formulating it as a Rayleigh quotient maximization with affinity constraints. To explore the higher order relationship among models, Gao *et al.* [20] proposed a hyper-graph analysis approach for this task by avoiding the estimation of the pairwise distance between models. Although these graph matching-based methods can achieve improvement to some extent, the current image-wise similarity measure is not robust enough and can easily have negative influence on graph matching due to the existence of redundant and noisy data. Furthermore, it is necessary to discover and preserve the local and global structural attributes conveyed in the graph for effective 3-D model retrieval.

III. Methods

As introduced above, the state-of-the-art 3-D model retrieval methods mainly contain two key steps, feature representation and similarity measure. For the basic evaluation, Zernike moments are selected for visual representation as [25], [30], [55], and [58]. The histogram of oriented gradients (HoGs) is also evaluated for feature extraction since it can capture the characteristics of local shapes [6]. Although deep learning has achieved breakthrough results on multiple applications [64], few work has been done on 3-D model retrieval

with deep learning methods. Therefore, it is important to systematically evaluate deep feature learning in this new research area. In this paper, we applied the 16-layer deep CNN model (VGG-16) [32] to generate the deep feature for visual representation. This model is pretrained on imagenet large scale visual recognition competition (ILSVRC)12 to extract the CNN structure features. This kind of CNN features can deliver rich semantic and structure information and intuitively suppress background noise. During training, the image is passed through a stack of convolutional (conv.) layers, where the filters are utilized with a very small receptive field 3×3. The convolution stride is fixed to 1 pixel and the spatial padding of conv. layer input is applied to preserve the spatial resolution after convolution. Spatial pooling is carried out by five max-pooling layers, which follows some of the conv. layers. Max-pooling is performed over a 2×2 pixel window, with stride 2. A stack of convolutional layers is followed by three fully connected (FC) layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels. The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks. All hidden layers are equipped with the rectification (ReLU) nonlinearity. The network architecture is shown in Fig. 1. Here we extracted the features from layer $\mathrm{FC}-7$.

We selected and reimplemented seven popular 3-D model retrieval methods based on shape retrieval contest 2015 and 2016 track on "multiview and multimodal 3-D object retrieval" [16], [17]. These popular methods are listed in Table I and briefly introduced as follows.

A. Distance-Based Method

This kind of methods directly leverages specific distance estimation of pairwise 3-D models for retrieval. Theoretically, it aims to select the most informative view of each model for model representation and set-to-set distance measure. Two popular methods are listed below.

1) *HAUS [13]:* Hierarchical agglomerative clustering is implemented for view clustering and the one with the smallest distance to the rest views in each cluster is selected as the characteristic view. In the HAUS, the distance between each point from one set and the closest point from the other set is determined. Then, the HAUS is calculated as the maximal point-wise distance. It can be calculated as follows:

$$\mathrm{HAUS}(\mathcal{O}_1, \mathcal{O}_2) = \max\left\{\begin{matrix}\max_{v\in\mathcal{O}_1}\left\{\min_{u\in\mathcal{O}_2} d(v,u)\right\} \\ \max_{v\in\mathcal{O}_2}\left\{\min_{u\in\mathcal{O}_1} d(v,u)\right\}\end{matrix}\right\} \tag{1}$$

where $\mathcal{O}_1$ and $\mathcal{O}_2$ are two compared sets, v and u are the points from $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, and $d(v,u)$ is the distance between two views.

2) *Nearest Neighbor (NN) [13]:* The NN-based method is similar to HAUS. The only difference is that NN leverages the minimal distance of all view pairs across two models for similarity measure. It can be calculated as follows:

$$\mathrm{NN}(\mathcal{O}_1, \mathcal{O}_2) = \min_{v\in\mathcal{O}_1, u\in\mathcal{O}_2} d(v,u). \tag{2}$$

TABLE I
LIST OF THE EVALUATED METHODS (M: MATLAB AND MC: MIXTURE OF MATLAB AND C/C++)

Method	Feature	Code
HAUS [13]	Zernike/HoG/CNN	M
NN [13]	Zernike/HoG/CNN	M
AVC [1]	Zernike/HoG/CNN	MC
SCCV [43]	Zernike/HoG/CNN	MC
CCFV [18]	Zernike/HoG/CNN	MC
WBGM [14]	Zernike/HoG/CNN	MC
CGM [41]	Zernike/HoG/CNN	MC

B. Statistical Model-Based Method

This kind of method learns the statistical model for each category of 3-D models to infer the comparability between the query model and individual category. Basically, it aims to adaptively learn the view-invariant characteristics of each kind of models with multiview cues.

1) *Adaptive Views Clustering (AVC) [1]:* AVC selected the optimal 2-D characteristic views $V = \{V_k\}_{k=1}^{K}$ of a 3-D model based on the adaptive clustering algorithm. With the characteristic views of 3-D models, AVC learns individual probabilistic Bayesian model of each category and utilizes the inference score to classify the query model. Considering a query model $\mathcal{Q}$, we wish to find the model $\mathcal{O}_i \in \widetilde{\mathcal{O}}$, which has the highest probability $P(\mathcal{O}_i|Q)$. Since each model is represented by its characteristic views, $P(\mathcal{O}_i|Q)$ can be computed by

$$P\left(\mathcal{O}_i\middle|\mathcal{Q}\right) = \sum_{k=1}^{K} P\left(\mathcal{O}_i\middle|V_k^{\mathcal{Q}}\right) P\left(V_k^{\mathcal{Q}}\middle|\mathcal{Q}\right). \tag{3}$$

2) *CCFV-Based Method [18]:* CCFV is proposed to remove the constraint of the setting of static camera array for view capturing and 3-D model retrieval.
For a query model $\mathcal{Q}$ with the characteristic view set $\{v_k^{\mathcal{Q}}\}_{k=1}^{K}$ and a candidate model $\mathcal{O}$ with the characteristic view set $\{v_c^{O}\}_{c=1}^{C}$, the relationship between both models can be defined by a binary variable δ. $\delta = 1$ if $\mathcal{O}$ is relevant to $\mathcal{Q}$ and $\delta = 0$, otherwise. Therefore, we aim to find the model with $\delta = 1$. The similarity between both can be measured by

$$S(\mathcal{Q}, \mathcal{O}) = P(\mathcal{O}|\mathcal{Q}, \delta = 1) - P(\mathcal{O}|\mathcal{Q}, \delta = 0). \tag{4}$$

Here both models, $P(\mathcal{O}|\mathcal{Q}, \delta = 1)$ & $P(\mathcal{O}|\mathcal{Q}, \delta = 0)$, are formulated with the Gaussian model and are individually trained with positive and negative matched samples, respectively. Then, $S(\mathcal{Q}, \mathcal{O})$ can be utilized to rank all candidates in the descending order as the retrieval list.

3) *Spatial Context-Constrained View-Based Method (SCCV) [43]:* Different from AVC and CCFV, SCCV aims to leverage the spatial context for characteristic view extraction and model retrieval. First, SCCV encodes both visual feature and spatial context to construct a graph for view clustering ($\mathbb{C}$).

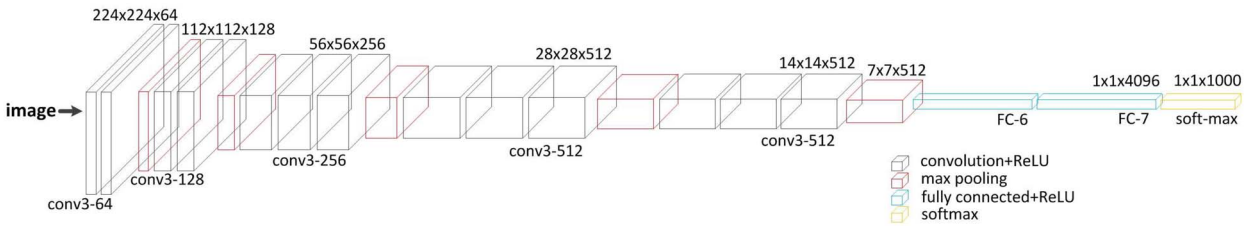


Fig. 1. Network architecture for VGG-16.

The objective function consisting of two parts, data terms and smooth terms, can be defined as follows:

$$\mathbb{C}^* = \arg\max_{\mathbb{C}} \sum_{i=1}^{m} D_1(f_i, f_c) + \sum_{i,j=1}^{m} E(v_i)E(v_j)D_2(v_i, v_j) \quad \text{s.t.} \quad v_i, v_j \in \mathbb{C}. \tag{5}$$

$D_1(f_i, f_c)$ denotes the contribution of the ith view for this cluster $\mathbb{C}$ by computing the similarity between the ith view (f_i) and the center (f_c). $E(v_i)E(v_j)D_2(v_i, v_j)$ denotes the correlation between pairwise views by computing the product of the contributions of individual views ($E(v_i)$ & $E(v_j)$) and the similarity between both views ($D_2(v_i, v_j)$). If both views belong to the same cluster, this term should have a higher value. After the above process, the problem of view clustering can be converted into the energy maximization problem. The graph-cut algorithm can be applied iteratively to get a set of subclusters. Then, the random walk algorithm can be implemented to update the weight for each view and we select the view with the highest weight in each subcluster as the characteristic view. Finally, the Bayesian model as utilized in AVC is learned to infer the score of the query model given a specific category for 3-D model retrieval.

C. Graph Matching-Based Method

Different from the aforementioned methods, graph matching-based methods explicitly leverage multiview information to solve the many-to-many distance measure. Essentially, it aims to construct the graph structure by discovering the salient features of individual view images and/or the visual/spatial relationship between pairwise/high-order view images. Generally speaking, any graph matching method can be utilized to measure the similarity between pairwise 3-D models. Several popular methods are introduced below.

1) Weighted Bipartite Graph Matching [14]: Weighted bipartite graph matching (WBGM) first extracts the characteristic views from the query model and the candidate model. The initial weights of characteristic views are initialized and further updated based on the correlations among them. The two groups of views are formulated as the two subsets of the weighted bipartite graph and the proportional max-weighted bipartite matching is employed to calculate the matching between 3-D objects. Given two 3-D models ($X = \{v_i^X\}_{i=1}^n$ & $Y = \{v_i^Y\}_{i=1}^n$) with n characteristic views, and the corresponding weight vectors ($\{w_i^X\}_{i=1}^n$ & $\{w_i^Y\}_{i=1}^n$), the weighted bipartite graph, $\mathcal{G} = \{\mathcal{X}, \mathcal{Y}, \mathcal{E}\}$ can be constructed for similarity measure between pair models. Each node of $\mathcal{G}$ represents one view in X or Y. Each edge of the edge set, $\mathcal{E} = \{w_{ij}\}$, denotes a correlation between pairwise nodes. Let $d(v_i^X, v_j^Y)$ denote the distance between both views. Then, w_{ij} can be computed by $w_{ij} = (w_i^X + w_j^Y)d(v_i^X, v_j^Y)/2$. Let $\Lambda = \{\Lambda_k\}$ be the set of matching solutions for the constructed weighted bipartite graph. In Λ_k, $v_i^k \in \mathcal{V}$ and $u_i^k \in \mathcal{U}$ are two matching nodes. The Kuhn–Munkres method is a popular solution to achieve the optimal matching, Λ^*. First, the edge matrix, W, is converted to the cost matrix C, where $c_{ij} = \mu - w_{ij}$ and $\mu \geq \max(w_{ij})$. Then the objective function of max-weighted bipartite matching can be formulated as

$$\Lambda^* = \arg\max_{\Lambda_k \in \Lambda} \sum_{1 \leq i \leq n} c_{v_i^k, u_i^k}. \tag{6}$$

The Hungarian algorithm can be implemented to achieve the optimal matching and similarity measure between pairwise 3-D models. The ranking list can be generated with the pairwise similarities in the descending order.

2) Clique Graph Matching [41]: Clique graph matching (CGM) replaces individual node (the basic unit in one graph) of the classic graph by one clique, which consists of K nearest neighbors in the specific feature subspace and can convey the local structural attributes in the star model. A clique-graph $\tilde{G} = \{\tilde{V}, \tilde{A}\}$ consists of the clique set $\tilde{V}$ and the attribute set $\tilde{A}$ associated with individual cliques. Each clique $\tilde{V}_i \in \tilde{V}$ can be represented by the star model, $\tilde{V}_i = \{\tilde{c}_i, \{\tilde{l}_{ij}\}_{j=1}^k, \{\tilde{e}_{ij}\}_{j=1}^k\}$, where $\tilde{c}_i$ denotes the center of the clique, $\{\tilde{l}_{ij}\}_{j=1}^k$ denotes k leave nodes of the clique, $\{\tilde{e}_{ij}\}_{j=1}^k$ denotes k edges linking the center node and each leave node. The order of one clique, $\delta(\tilde{V}_i)$, equals to the number of the nodes in it. $\tilde{A}_i \in \tilde{A}$ is the attribute for the ith clique, which represents the importance of this clique in the entire clique-graph and can be calculated based on the term frequency by dividing the total frequency of the nodes in $\tilde{V}_i$ by the produce of the clique number and the node number in each clique.

Given two clique-graphs, $\tilde{G}^p = \{\tilde{V}^p, \tilde{A}^p\}$ and $\tilde{G}^q = \{\tilde{V}^q, \tilde{A}^q\}$, the similarity between both can be represented by $J(X, \tilde{G}^p, \tilde{G}^q)$, which means the similarity of both clique-graph can be computed by considering their structure characteristics ($\tilde{G}^p$ and $\tilde{G}^q$) and the clique-to-clique correspondence X. By denoting the clique-to-clique similarity as $S_{i,m}$, where $\tilde{V}_i \in \tilde{G}^p$ and $\tilde{V}_m \in \tilde{G}^q$, the vector $S = \{S_{i,m}\}_{i=1,\ldots,N^p, m=1,\ldots,N^q} \in R^{N^p N^q \times 1}$, where N^p and N^q, respectively, denotes the clique numbers in $\tilde{G}^p$ and $\tilde{G}^q$, can be used to represent the similarities

between all pairwise cliques. We define the solution of clique-graph matching as a binary indicator matrix $\bar{X} \in \{0,1\}^{N^p \times N^q}$. If $\tilde{V}_i \in \tilde{G}^p$ is matched to $\tilde{V}_m \in \tilde{G}^q$, $X_{i,m} = 1$. Otherwise, $X_{i,m} = 0$. Moreover, it is natural to impose one-to-one constrains that make $\bar{X}$ as a permutation matrix

$$\bar{X} \cdot 1_{N^q \times 1} \leq 1_{N^p \times 1}, \quad \bar{X}^\top \cdot 1_{N^p \times 1} \leq 1_{N^q \times 1} \tag{7}$$

where $1_{N^p \times 1}/1_{N^q \times 1}$ denotes an all-ones vector with size N^p/N^q and the inequalities hold for every element. For a convenient representation, the vectorized version of $\bar{X}$, X, is used.

Obviously, CGM can be formulated as an integer linear programme to discover the indicator vector X^* that maximizes the following objective function:

$$\begin{aligned} \left(X, \tilde{G}^p, \tilde{G}^q\right)^* &= \arg \max_{X, \tilde{G}^p, \tilde{G}^q} \left(S^\top \cdot X\right) \\ \text{s.t. } \bar{X} \cdot 1_{N^q \times 1} &\leq 1_{N^p \times 1}, \quad \bar{X}^\top \cdot 1_{N^p \times 1} \leq 1_{N^q \times 1}. \end{aligned} \tag{8}$$

However, it is nontrivial to optimize (8) directly since both clique structure information in individual clique-graph and similarity measure of pairwise cliques from different clique-graphs are latent variables. The solution for optimization contains two key steps.

1) The sparse-subspace clustering method can be employed to discover the clique structure information.
2) The local graph matching with unary and pairwise correspondences is implemented to align both cliques from two clique-graphs and the clique-wise similarity can be further computed. After achieving the clique-to-clique similarity, the objective function in (8) can be easily solved by integral linear programming.

CGM is originally proposed as a general graph matching algorithm [49]. In this paper, we find that it can benefit measuring the similarity between pairwise 3-D models due to two main reasons.

1) CGM can represent the cluster of multiview images with similar visual pattern by one clique. It can effectively overcome the difficulty in characteristic view extraction, which is highly required in most of current methods [1], [14], [43], and consequently avoid the negative influence by the existence of outliers (imperfect characteristic views) for graph matching.
2) CGM can simplify the complicated edge correspondence into the formulation only with clique-wise and edge-wise correspondences. Therefore, it can significantly increase the complexity of the local structure with more nodes in one clique to convey more high-order attributes of one cluster of multiview images with similar visual pattern, while only sacrificing little computational complexity. Consequently, we implement it for this task and the comparative experiments in Section VI will demonstrated its superiority.

IV. Datasets

Three virtual 3-D model datasets (NTU60, NTU216, and PSB) and two real-world 3-D model datasets (ETH and MV-RED) are utilized for quantitative evaluation. They are,

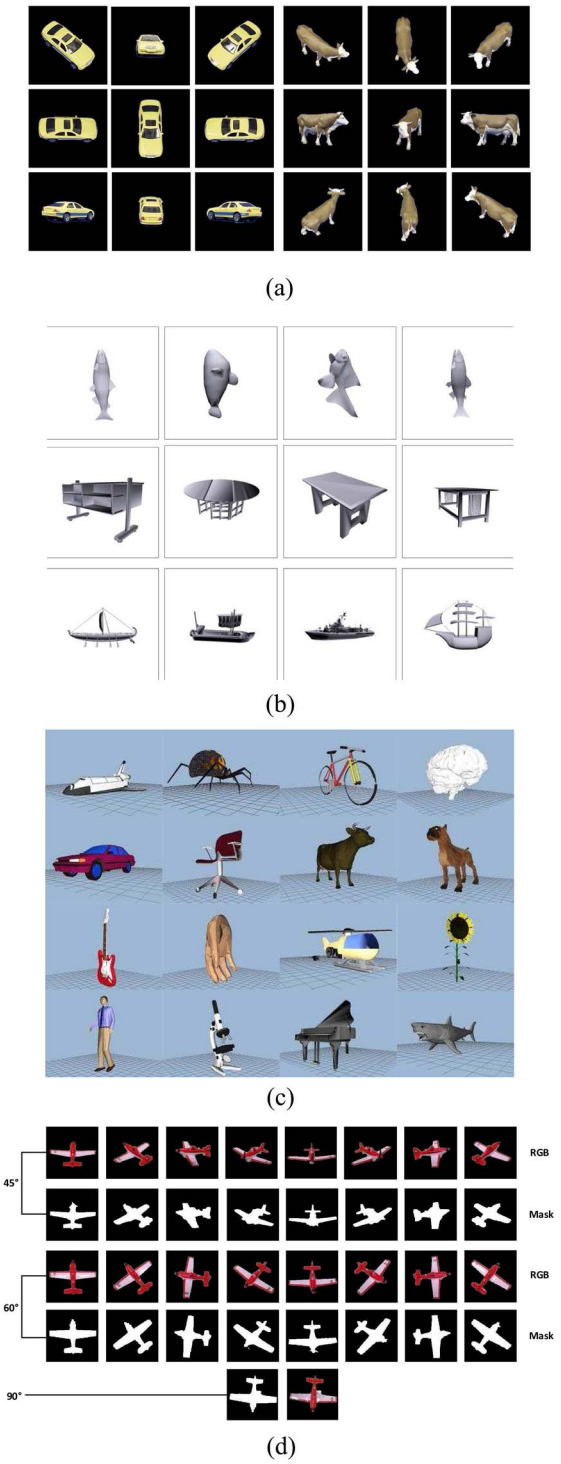


Fig. 2. 3-D model samples from different datasets. (a) ETH. (b) NTU60/NTU216. (c) PSB. (d) MV-RED.

respectively, introduced as follows. Some samples are shown in Fig. 2.

1) *NTU60/NTU216 [3]:* The NTU dataset contains 500 models from 50 categories. Since this dataset only provides 3-D models, which cannot be directly employed for view-based methods, we capture views of individual model by setting a virtual camera array consisting of 60 cameras. The cameras are set on the vertices of a polyhedron with the same structure of Buckminsterfullerene (C60).

Consequently, individual 3-D models can be represented by a set of images from 60 views. We term NTU in this setting as NTU60. Furthermore, we capture NTU views from 216 angles. All of them form 6 faces of each 3-D model and each face contains 36 views. Therefore, individual model contains 216 views. We term NTU in this setting as NTU216.

2) *PSB [57]:* PSB contains 1814 virtual 3-D models from 161 classes. Similar to NTU, we generate the view set for each model by setting the virtual camera array with 60 cameras.
3) *ETH [34]:* ETH contains 80 models from eight categories and provides each model with 41 views, which are captured using the camera array spaced evenly over the upper viewing hemisphere and the position for each camera is set by subdividing the faces of an octahedron to the third recursion level.
4) *MV-RED:* We contribute a real-world 3-D model dataset, multiview model dataset (MV-RED). MV-RED consists of 505 objects from 60 categories. Each object was recorded simultaneously by three cameras from three directions. For data acquisition, Camera-45° and Camera-60° captured 36 RGB images every 10° by uniformly rotating the table controlled by a step motor. Camera-90° captured only 1 RGB image in the top-down view. Therefore, each object contains 73 images. The image resolution is 640 × 480. Foreground segmentation was implemented for the dataset and masks were provided. To the best of our knowledge, MV-RED is the most challenging real-world 3-D model dataset since it contains much more samples with the change of illumination and significant intraclass difference.

V. Evaluation Criteria

For the evaluation on each dataset, each 3-D model is selected as the query once for retrieval. Two kinds of visual features and seven methods are evaluated on five datasets, respectively. The optimal parameters of individual method are selected by cross validation. For quantitative comparison, the following popular criteria are employed as [19], [21], and [50].

1) Precision-recall curve (PR-curve) is able to comprehensively demonstrate the retrieval performance, which illustrates the precision and recall measures by varying the threshold to distinguish relevance and irrelevance in model retrieval. The area under curve (AUC) of PR-curve can be calculated for quantitative evaluation.
2) NN evaluates the retrieval accuracy of the first returned result.
3) First tier (FT) is defined as the recall of the top κ results, where κ is the number of relevant objects for the query.
4) Second tier (ST) is defined as the recall of the top 2κ results.
5) F-measure (F) jointly evaluates the precision and recall of top returned results. In our experiments, top 20 retrieved results are used for calculation.
6) Discounted cumulative gain (DCG) [1] is a statistic that assigns relevant results at the top ranking positions with higher weights under the assumption that a user is less likely to consider lower results.
7) Average normalized modified retrieval rank (ANMRR) [46] is a rank-based measure, and it considers the ranking information of relevant objects among the retrieved objects. A lower ANMRR value indicates a better performance, i.e., relevant objects rank at top positions.

TABLE II
Gain by CGM With Different Features Against Competing Methods (%)

CGM+Zernike	AUC↑	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓
NTU60	12-47	3-41	21-53	20-38	11-40	12-32	6-14
NTU216	16-51	23-93	9-23	3-13	2-20	3-32	11-26
PSB	23-187	17-103	23-41	15-74	4-132	12-146	10-27
MV-RED	3-70	7-116	3-33	4-21	1-13	11-83	11-27
CGM+CNN	AUC↑	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓
NTU60	10-70	2-40	14-82	14-62	17-54	16-61	10-25
NTU216	12-66	10-56	4-19	7-13	4-13	8-54	6-48
PSB	18-113	18-125	11-82	9-118	7-103	8-163	12-36
MV-RED	2-53	11-18	7-31	2-28	3-24	9-27	7-21
CGM+HoG	AUC↑	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓
NTU60	2-19	10-14	6-106	3-73	2-98	7-121	4-23
NTU216	0.5-102	19-132	14-89	7-60	7-84	17-103	7-23
PSB	7-246	195-229	74-184	43-134	15-136	27-194	4-162
MV-RED	5-284	0.5-293	0.4-223	1-154	1-139	1-251	0.4-294

VI. Experimental Results

In this section, we first evaluate individual methods on five datasets. Then we statistically analyze the performances and speed with respect to different methods, features and datasets.

A. Comparison Among Individual Methods

The PR-curve with AUC on five datasets are shown in Fig. 5. The quantitative evaluation on five datasets are shown in Fig. 6. With the comparison, we have the following observations.

1) CGM can consistently outperform all the others, regardless of what feature is utilized, since it can discover the cliques to preserve the local structures and further leverage clique-wise and edge-wise similarity for CGM to preserve the global correspondence. The achieved gain by CGM with Zernike/CNN/HoG against competing methods are shown in Table II.
2) WBGM only keeps the characteristic views of one model, which can be regarded as the members of the corresponding cliques in CGM. However, it is usually difficult to select characteristic views from multiple similar views and moreover the image-wise similarity measure is not robust enough when outliers exist. Therefore, WBGM only with the node-wise attribute works worst. CGM cannot only discover the high-order attributes conveyed in the multiview images of one 3-D model but also replaces the image-wise node in WBGM with the clique, consisting of a set images with similar visual pattern. Although there might exist outliers, the clique-wise similarity measure can significantly enhance inliers while suppressing outliers. Therefore, CGM can theoretically and experimentally outperform WBGM.
3) SCCV, CCFV, and AVC consider the feature distribution of multiview images and utilize the statistical models for 3-D model learning and inference.

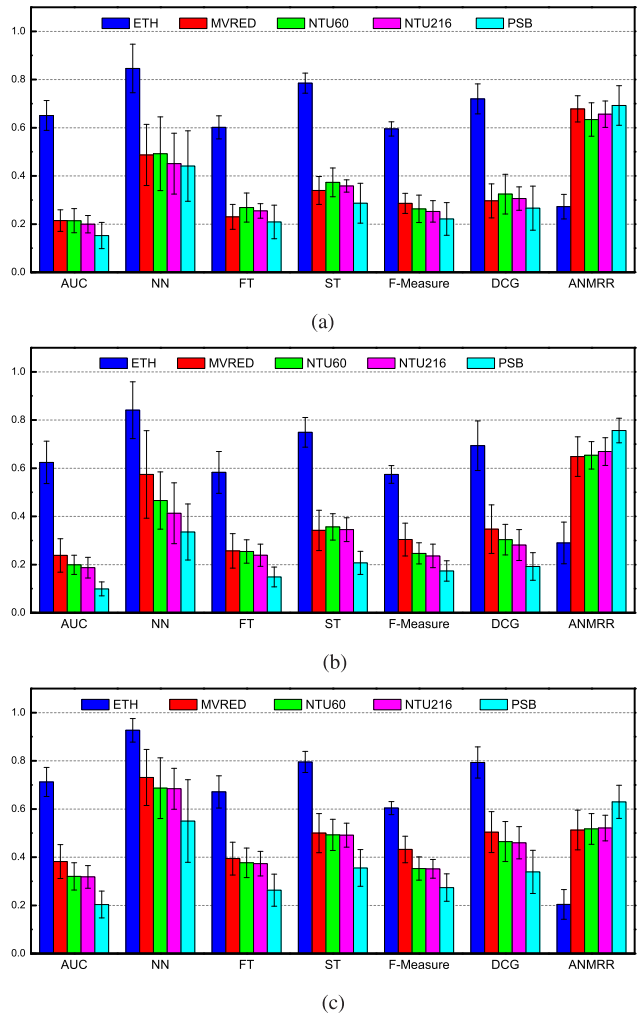


Fig. 3. Statistical analysis of the performances on different datasets. (a) Comparison@Zernike. (b) Comparison@HoG. (c) Comparison@CNN.

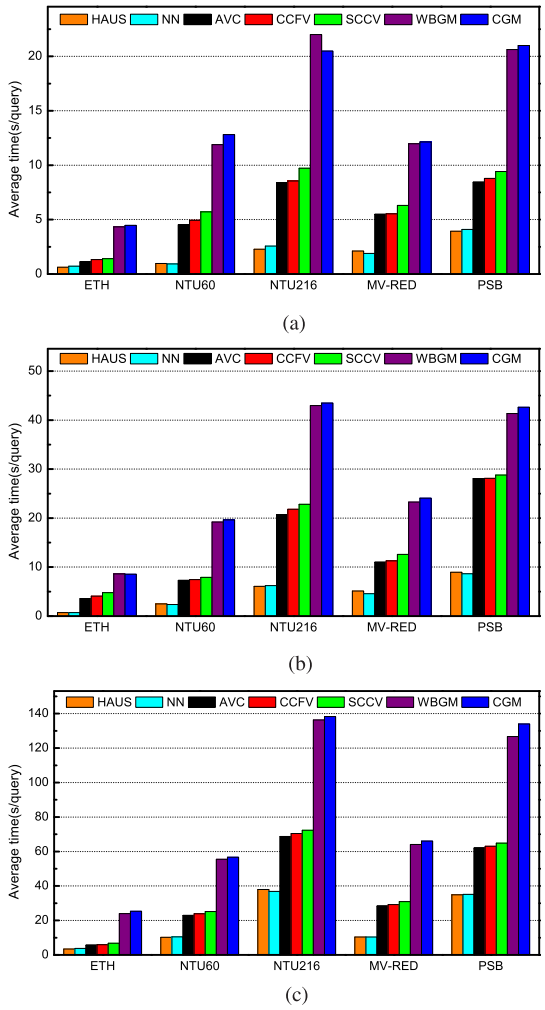


Fig. 4. Speed comparison on different datasets with respect to different methods and features. Sample number in individual dataset: ETH:80; NTU60/NTU216:500; MV-RED:505; and PSB:1814. View number of individual sample: ETH:41; NTU60:60; NTU216:216; MV-RED:73; and PSB:60. (a) Comparison@Zernike. (b) Comparison@HoG. (c) Comparison@CNN.

Since AVC only leverages the visual information of multiview 2-D images, it is usually difficult to select the most informative views for model learning and consequently AVC works worst competing against the others. Comparatively, SCCV leverages both visual and spatial information for characteristic view extraction. The comparison experiments in Fig. 5 demonstrate that the extracted characteristic views are more informative and consequently improve the performances. Different from AVC and SCCV, CCFV takes advantages of all 2-D images of one 3-D model to learn both positive and negative Gaussian mixtrure model to augment the discrimination. Therefore, it can also improve the performance comparing against AVC and achieve competing performance to SCCV.

4) HAUS and NN are distance-based methods. NN measures the Euclidean distance between pairwise views and is very sensitive to the feature representation. Comparatively, HAUS can directly measure the set-to-set distance and is more robust to NN. Therefore, HAUS can consistently outperform NN as shown in Fig. 5. To our surprise, HAUS with Zernike moments can usually achieve competing methods to SCCV, CCFV, and WBGM without statistical model learning and complex graph matching as shown in Fig. 5. This indicates that HAUS is more practical for real application with less computational complexity. However, when CNN features are used, HAUS cannot significantly improve the performance as other methods do and obviously works worse than the competing methods (SCCV, CCFV, and WBGM) especially on NTU60, NTU216, and MV-RED.

B. Statistical Analysis

The statistical analysis of the performances by different methods and different features on five datasets are, respectively, shown in Figs. 3 and 7. By comparison, we have the following observations.

1) Generally speaking, the graph matching-based methods can consistently outperform the distance-based methods and the statistical model-based methods on all datasets. The graph matching-based methods usually select the

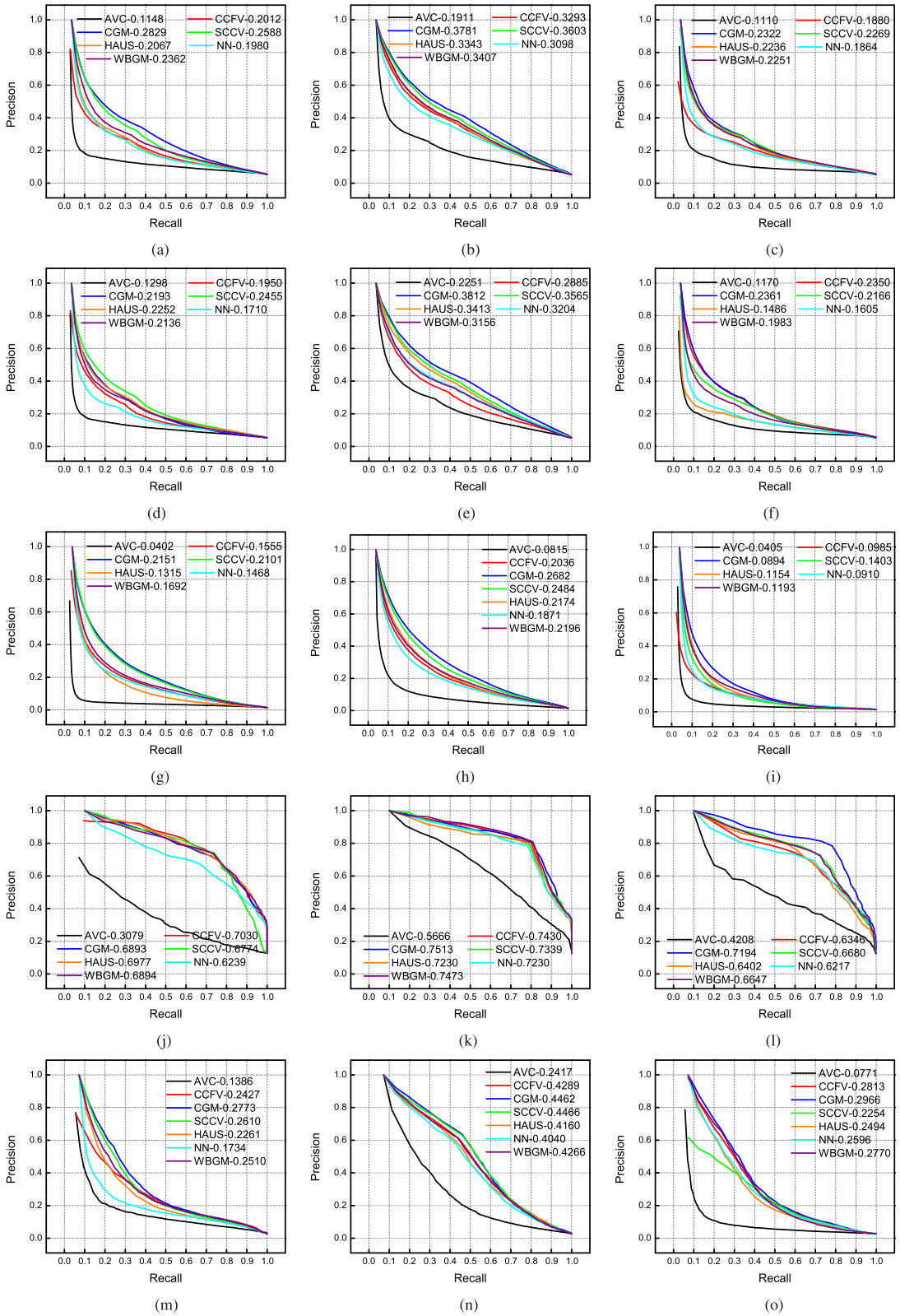


Fig. 5. Performance comparison on five datasets (Z: Zernike moments; C: CNN features; H: HoG features; rows 1–5 correspond to the performances on NTU60, NTU216, PSB, ETH, and MV-RED, respectively). (a) PR-Z@NTU60. (b) PR-C@NTU60. (c) PR-H@NTU60. (d) PR-Z@NTU216. (e) PR-C@NTU216. (f) PR-H@NTU216. (g) PR-Z@PSB. (h) PR-C@PSB. (i) PR-H@PSB. (j) PR-Z@ETH. (k) PR-C@ETH. (l) PR-H@ETH. (m) PR-Z@MV-RED. (n) PR-C@MV-RED. (o) PR-H@MV-RED.

characteristic views or cliques to keep the local characteristics from different views and then leverage graph matching for similarity measure. The strict constraints by graph matching can usually benefit achieving a relatively reasonable similarity measure. Furthermore, the graph matching-based methods can usually achieve the

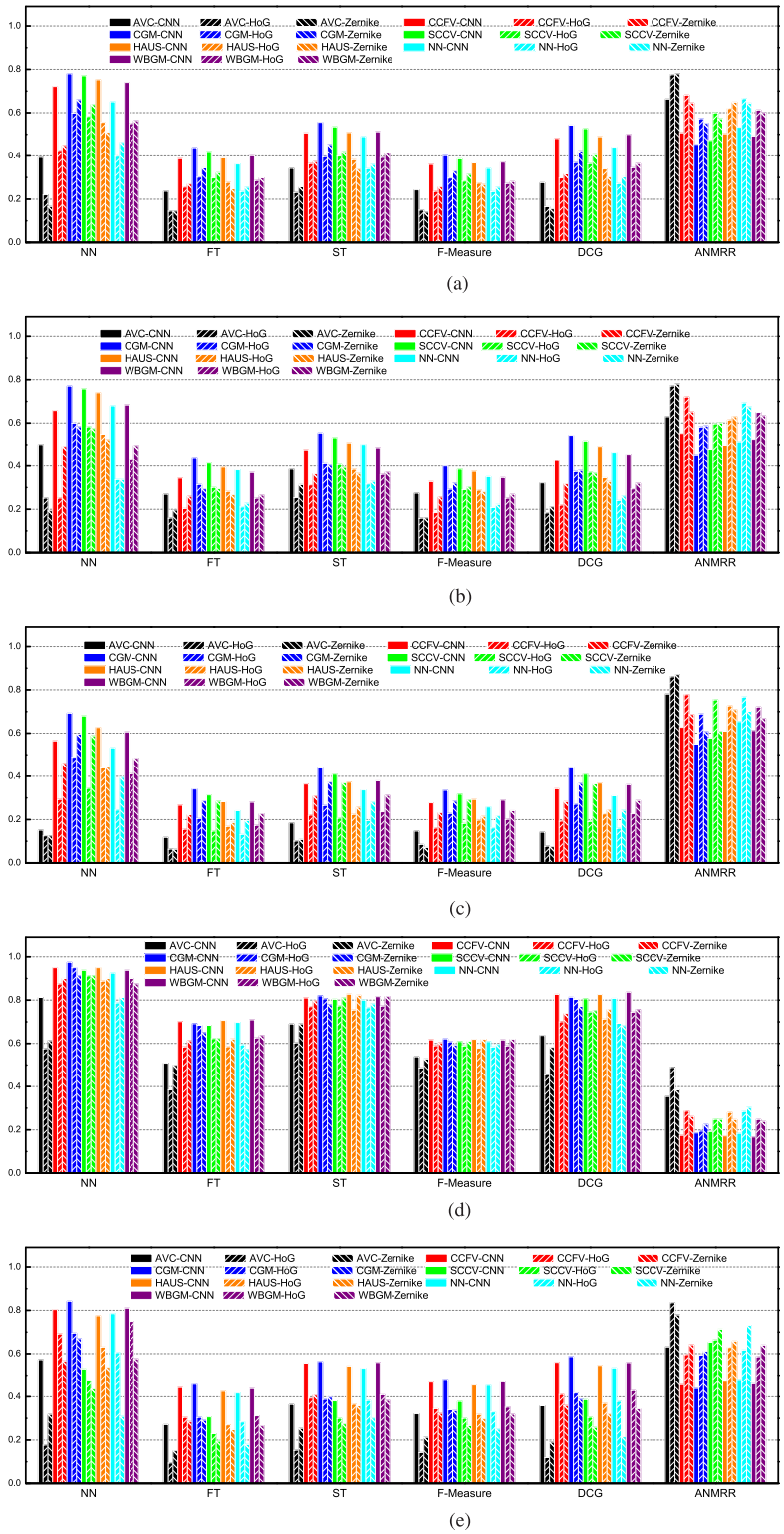


Fig. 6. Performance comparison on five datasets. (Z: Zernike moments; C: CNN features; H: HoG features; (a)–(e) correspond to the performances on NTU60, NTU216, PSB, ETH, and MV-RED, respectively.) (a) Comparison@NTU60. (b) Comparison@NTU216. (c) Comparison@PSB. (d) Comparison@ETH. (e) Comparison@MV-RED.

smallest standard deviation represented by the error bar shown in Fig. 7(a), (c), (e), (g), and (i). Consequently, this kind of methods is more stable than the others.

2) Both distance-based method and statistical model-based method highly depend on the multiview images of each 3-D model. Since the latter can adaptively learn the view-invariant characteristics of each model with probabilistic model learning, it can outperform the former on most of datasets (NTU60, NTU216, PSB, and MV-RED). Since ETH is only a small scale dataset containing 80 models and 41 views with significant visual variance per model, it is not suited for statistical model learning

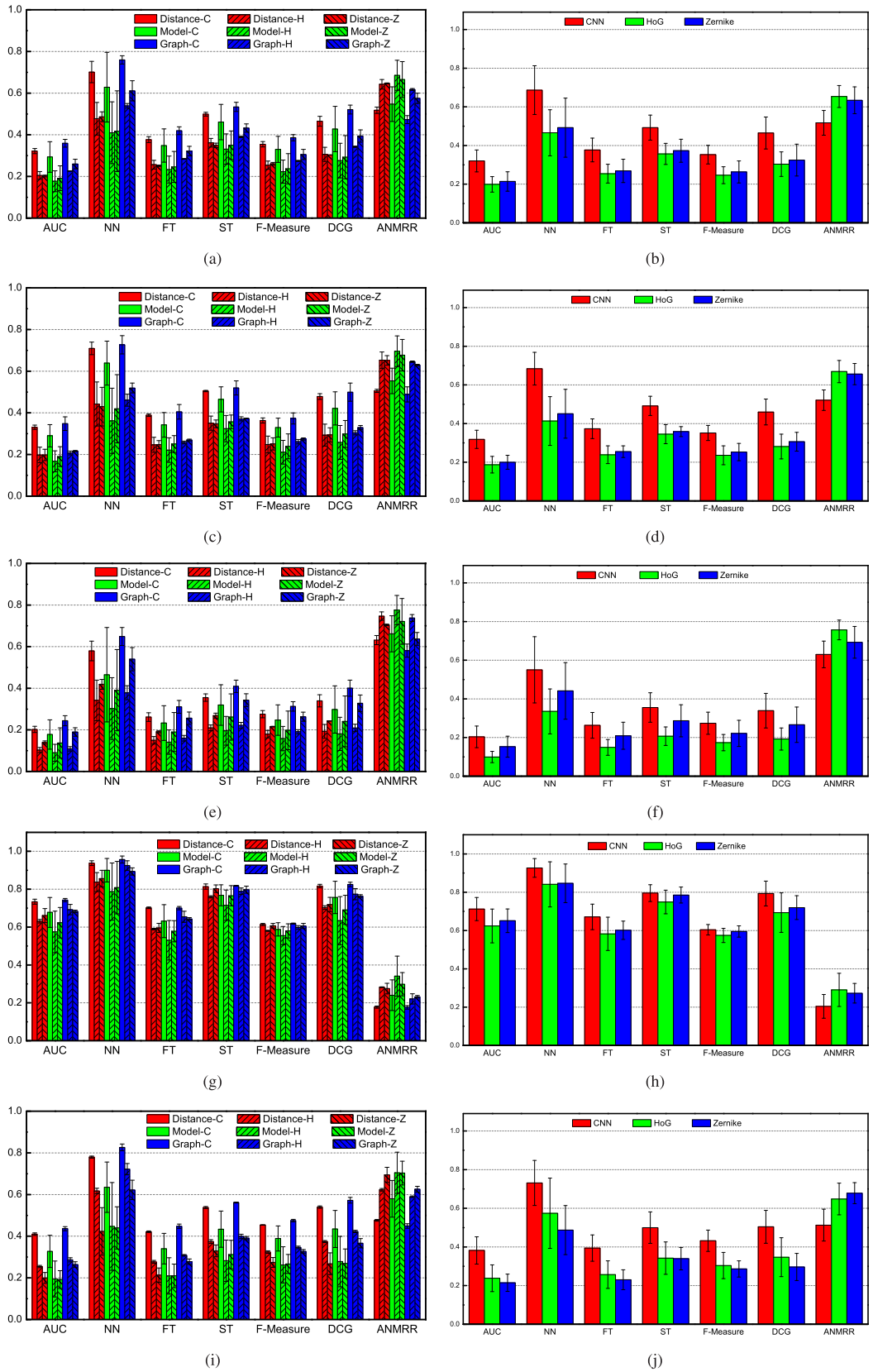


Fig. 7. Statistical analysis of the performances by (a), (c), (e), (g), and (i) different methods and (b), (d), (f), (h), and (j) different features (rows 1–5 correspond to the performances on NTU60, NTU216, PSB, ETH, and MV-RED, respectively).

on ETH. Consequently, the distance-based methods can outperform the statistical model-based methods as shown in Fig. 7(g).

3) The methods with CNN features can consistently outperform those with Zernike moments and the HoG feature since the CNN model can extract a hierarchy

of increasingly complex features with the trainable filters and local neighborhood pooling [23]. On NTU60, the methods with CNN features can achieve a gain of 30%, 29%, 36%, 32%, 23%, and 41% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 21% in terms of ANMRR compared to Zernike moments. CNN features can achieve a gain of 61%, 47%, 48%, 38%, 43%, and 52% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 21% in terms of ANMRR compared to HoG. On NTU216, the methods with CNN features can achieve a gain of 71%, 44%, 57%, 48%, 36%, and 37% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 29% in terms of ANMRR compared to Zernike moments. CNN features can achieve a gain of 70%, 66%, 56%, 42%, 49%, and 64% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 22% in terms of ANMRR compared to HoG. On PSB, the methods with CNN features can achieve a gain of 34%, 23%, 26%, 21%, 18%, and 26% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 9% in terms of ANMRR compared to Zernike moments. CNN features can achieve a gain of 105%, 64%, 77%, 72%, 58%, and 76% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 17% in terms of ANMRR compared to HoG. On ETH, the methods with CNN features can achieve a gain of 7%, 12%, 10%, 1%, 1%, and 9% in terms of AUC, NN, FT, ST, F, and DCG, and also achieve a decline of 26% in terms of ANMRR compared to Zernike moments. CNN features can achieve a gain of 23%, 10%, 15%, 6%, 5%, and 14% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 30% in terms of ANMRR compared to HoG. On MV-RED, the methods with CNN features can achieve a gain of 63%, 43%, 39%, 41%, 31%, and 63% in terms of AUC, NN, FT, ST, F, and DCG, and also achieve a decline of 28% in terms of ANMRR compared to Zernike moments. CNN features can achieve a gain of 60%, 27%, 53%, 46%, 42%, and 45% in terms of AUC, NN, FT, ST, F, and DCG, and achieve a decline of 21% in terms of ANMRR compared to HoG.

4) From Fig. 3, we can achieve the optimal performances (AUC:0.65, NN:0.87, FT:0.61, ST:0.75, F:0.57, DCG:0.73, and ANMRR:0.27) on ETH since this dataset with limited samples is not as challenging as the others. However, the performances on the other four datasets is far from satisfaction since AUC, FT, ST, F, and DCG on them are usually below 0.4 and NN and ANMRR on them are around 0.6. Therefore, ETH is not a challenging dataset for 3-D model retrieval for future evaluation.

C. Speed Analysis

The statistical analysis of the speeds by different methods and different features on five datasets are, respectively, shown in Fig. 4. All methods were tested on the PC with single core (CPU: 3.1 GHz; RAM: 8GB). According to [18], [19], and [41], the step of model training can be completed offline. Here only the time for retrieval is considered for fair comparison. By comparison, we have the following observations.

1) Generally speaking, the graph matching-based methods usually cost the most time for retrieval since solving (6) for WBGM and (8) for CGM is quite time-consuming. Moreover, CGM needs to select the characteristic cliques to keep the local characteristics from different views and consequently cost more than WBGM.
2) The statistical model-based methods usually spend much time on characteristic views selection by view clustering. Comparatively, the distance-based methods can straightforwardly implement similarity measure between pairwise views. Therefore, the latter usually has the higher speed.
3) When fixing the method of similarity measure for individual dataset, the speed is proportional to the dimension of features. Therefore, with the same method of similarity measure, the speed of the Zernike (49-D)-based method can consistently outperform the CNN(4096-D) and HoG (144-D)-based methods.
4) When fixing the methods for both feature representation and similarity measure, the speed for individual dataset is directly affected by the sample number of individual dataset and the view number of individual sample. Therefore, it is critical to leverage the advanced algorithms for multimedia retrieval [4], [67], [70], [72] to speed up the current methods for real application with 3-D big data.

VII. Conclusion

In this paper, we systematically evaluate the performance of the state-of-the-art methods on 3-D model retrieval. This large-scale evaluation can facilitate better understanding of current methods and provide a platform to integrate new algorithms and features. Based on the comparison experiments and observations, we highlight two essential components for this task. From the viewpoint of similarity measure, graph matching-based methods can preserve both local and global characteristics for similarity measure. The strict constraints by graph matching can usually benefit achieving a stable similarity measure and consistently outperform the distance-based method and the statistical model-based method. From the viewpoint of feature representation, deep features with hierarchical complex features can benefit visual representation and characteristic view extraction and consequently boost the performance. Improving both components will further advance the research in this field. Moreover, in the era of big data, it is also important to develop large-scale 3-D model datasets with multiview and multimodal information. Our future work will focus on extending the dataset (MV-RED) with more samples and multimodal information while developing and integrating novel algorithms for systematic evaluation.

References

[1] T. F. Ansary, M. Daoudi, and J.-P. Vandeborre, "A Bayesian 3-D search engine using adaptive views clustering," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 78–88, Jan. 2007.

[2] S. Bai, X. Bai, W. Liu, and F. Roli, "Neural shape codes for 3D model retrieval," *Pattern Recognit. Lett.*, vol. 65, pp. 15–21, Nov. 2015.

[3] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 223–232, Sep. 2003.

[4] L. Chen, D. Xu, I. W.-H. Tsang, and X. Li, "Spectral embedded hashing for scalable image retrieval," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1180–1190, Jul. 2014.

[5] Z. Cheng, X. Li, J. Shen, and A. G. Hauptmann, "Which information sources are more effective and reliable in video search," in *Proc. SIGIR*, Pisa, Italy, 2016, pp. 1069–1072.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[7] P.-E. Danielsson, "Euclidean distance mapping," *Comput. Graph. Image Process.*, vol. 14, no. 3, pp. 227–248, Nov. 1980.

[8] P. Daras and A. Axenopoulos, "A 3D shape retrieval framework supporting multimodal queries," *Int. J. Comput. Vis.*, vol. 89, nos. 2–3, pp. 229–247, Sep. 2010.

[9] H. Dutagaci *et al.*, "Shrec 2009—Shape retrieval contest of partial 3D models," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Munich, Germany, 2009, pp. 1–8.

[10] T. Furuya and R. Ohbuchi, "Fusing multiple features for shape-based 3D model retrieval," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., Sep. 2014, pp. 1–12.

[11] T. Furuya and R. Ohbuchi, "Hashing cross-modal manifold for scalable sketch-based 3D model retrieval," in *Proc. 2nd Int. Conf. 3D Vis. 3DV*, vol. 1. Tokyo, Japan, Dec. 2014, pp. 543–550.

[12] K. Gao *et al.*, "Visual stem mapping and geometric tense coding for augmented visual vocabulary," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3234–3241.

[13] Y. Gao and Q. Dai, *View-Based 3-D Object Retrieval*. Holland: Morgan Kaufmann, 2014.

[14] Y. Gao, Q. Dai, M. Wang, and N. Zhang, "3D model retrieval using weighted bipartite graph matching," *Sig. Process. Image Commun.*, vol. 26, no. 1, pp. 39–47, 2011.

[15] Y. Gao, Q. Dai, and N. Zhang, "3D model comparison using spatial structure circular descriptor," *Pattern Recognit.*, vol. 43, no. 3, pp. 1142–1151, 2010.

[16] Y. Gao *et al.*, "Shrec'15 track: 3D object retrieval with multimodal views," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Zürich, Switzerland, 2015, pp. 129–136.

[17] Y. Gao *et al.*, "Shrec'16 track: 3D object retrieval with multimodal views," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Lisbon, Portugal, 2016, pp. 129–136.

[18] Y. Gao *et al.*, "Camera constraint-free view-based 3-D object retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2269–2281, Apr. 2012.

[19] Y. Gao, M. Wang, R. Ji, X. Wu, and Q. Dai, "3-D object retrieval with Hausdorff distance learning," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2088–2098, Apr. 2014.

[20] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[21] Y. Gao *et al.*, "Less is more: Efficient 3-D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, Oct. 2011.

[22] Z. Gao, H. Zhang, G. P. Xu, Y. B. Xue, and A. G. Hauptmann, "Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition," *Signal Process.*, vol. 112, pp. 83–97, Jul. 2015.

[23] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 580–587.

[24] A. Godil *et al.*, "Range scans based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Zürich, Switzerland, 2015, pp. 153–160.

[25] A. Goyal and E. Walia, "Variants of dense descriptors and Zernike moments as features for accurate shape-based image retrieval," *Image Video Process.*, vol. 8, no. 7, pp. 1273–1289, 2014.

[26] X.-H. Han, Y.-W. Chen, and G. Xu, "High-order statistics of Weber local descriptors for image representation," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1180–1193, Jun. 2015.

[27] X. He, M. Gao, M.-Y. Kan, and D. Wang, "Birank: Towards ranking on bipartite graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 57–71, Jan. 2017.

[28] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[29] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[30] H.-U. Jang, D.-K. Hyun, D.-J. Jung, and H.-K. Lee, "Fingerprint-PKI authentication using Zernike moments," in *Proc. ICIP*, Paris, France, 2014, pp. 5022–5026.

[31] R. Ji, L.-Y. Duan, J. Chen, T. Huang, and W. Gao, "Mining compact bag-of-patterns for low bit rate mobile visual search," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3099–3113, Jul. 2014.

[32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[33] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.

[34] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. CVPR*, vol. 2. 2003, pp. 409–415.

[35] B. Leng, J. Zeng, M. Yao, and Z. Xiong, "3D object retrieval with multitopic model combining relevance feedback and LDA model," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 94–105, Jan. 2015.

[36] B. Li *et al.*, "Shrec'13 track: Large scale sketch-based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Girona, Spain, 2013, pp. 89–96.

[37] B. Li and H. Johan, "3D model retrieval using hybrid features and class information," *Multimedia Tools Appl.*, vol. 62, no. 3, pp. 821–846, Feb. 2013.

[38] B. Li *et al.*, "Shrec'16 track: 3D sketch-based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Lisbon, Portugal, 2016, pp. 47–54.

[39] B. Li, Y. Lu, and H. Johan, "Sketch-based 3D model retrieval by viewpoint entropy-based adaptive view clustering," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Girona, Spain, 2013, pp. 49–56.

[40] M. Liang, H. Min, R. Luo, and J.-H. Zhu, "Simultaneous recognition and modeling for learning 3-D object models from everyday scenes," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2237–2248, Oct. 2015.

[41] A.-A. Liu, W.-Z. Nie, Y. Gao, and Y.-T. Su, "Multi-modal clique-graph matching for view-based 3D model retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2103–2116, May 2016.

[42] A. Liu, Z. Wang, W. Nie, and Y. Su, "Graph-based characteristic view set extraction and matching for 3D model retrieval," *Inf. Sci.*, vol. 320, pp. 429–442, Nov. 2015.

[43] A. Liu *et al.*, *Spatial Context Constrained Characteristic View Extraction for 3D Model Retrieval* (Lecture Notes in Electrical Engineering), vol. 322. Cham, Switzerland: Springer, 2015, pp. 695–703.

[44] X. Liu, M. Wang, B.-C. Yin, B. Huet, and X. Li, "Event-based media enrichment using an adaptive probabilistic hypergraph model," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2461–2471, Nov. 2015.

[45] K. Lu, N. He, J. Xue, J. Dong, and L. Shao, "Learning view-model joint relevance for 3D object retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1449–1459, May 2015.

[46] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 593–601, 2001.

[47] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua, "Oracle in image search: A content-based approach to performance prediction," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 1–23, 2012.

[48] W. Nie, Q. Cao, A. Liu, and Y. Su, "Convolutional deep learning for 3D object retrieval," *Multimedia Syst.*, vol. 23, pp. 1–8, Oct. 2015.

[49] W.-Z. Nie, A.-A. Liu, Z. Gao, and Y.-T. Su, "Clique-graph matching by preserving global & local structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 4503–4510.

[50] W.-Z. Nie, A.-A. Liu, and Y.-T. Su, "3D object retrieval based on sparse coding in weak supervision," *J. Vis. Commun. Image Represent.*, vol. 37, pp. 40–45, May 2016.

[51] W. Nie, A. Liu, Z. Wang, and Y. Su, "Effective 3D object detection based on detector and tracker," *Neurocomputing*, vol. 215, pp. 63–70, Nov. 2016.

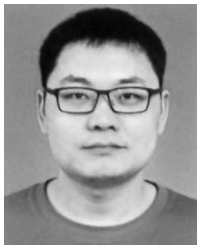
[52] I. Pratikakis *et al.*, "Shrec'16 track: Partial shape queries for 3D object retrieval," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Lisbon, Portugal, 2016, pp. 1–8.

[53] M. Savva *et al.*, "Shrec'16 track: Large-scale 3d shape retrieval from shapenet core55," in *Proc. Eurograph. Workshop 3D Object Retrieval*, Lisbon, Portugal, 2016, pp. 1–11.

[54] K. Sfikas, T. Theoharis, and I. Pratikakis, "3D object retrieval via range image queries in a bag-of-visual-words context," *Vis. Comput. Int. J. Comput. Graph.*, vol. 29, no. 12, pp. 1351–1361, Dec. 2013.
[55] S. Sharma and P. Khanna, "Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM," *J. Digit. Imag.*, vol. 28, no. 1, pp. 77–90, Feb. 2015.
[56] J.-L. Shih, C.-H. Lee, and J. T. Wang, "A new 3D model retrieval approach based on the elevation descriptor," *Pattern Recognit.*, vol. 40, no. 1, pp. 283–295, Jan. 2007.
[57] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton shape benchmark," in *Proc. SMI*, Genova, Italy, 2004, pp. 167–178.
[58] C.-W. Tan and A. Kumar, "Accurate iris recognition at a distance using stabilized iris encoding and Zernike moments phase features," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3962–3974, Sep. 2014.
[59] S. Tang, Y.-T. Zheng, Y. Wang, and T.-S. Chua, "Sparse ensemble learning for concept detection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 43–54, Feb. 2012.
[60] X. Tian, L. Yang, Y. Lu, Q. Tian, and D. Tao, "Image search reranking with hierarchical topic awareness," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2177–2189, Oct. 2015.
[61] D. V. Vranić and D. Saupe, "3D model retrieval," in *Proc. SCCG*, Bratislava, Slovakia, 2004, pp. 3–6.
[62] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1875–1883.
[63] X. Wang and W. Nie, "3D model retrieval with weighted locality-constrained group sparse coding," *Neurocomputing*, vol. 151, pp. 620–625, Mar. 2015.
[64] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
[65] J. Xuan, J. Lu, G. Zhang, and X. Luo, "Topic model for graph mining," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2792–2803, Dec. 2015.
[66] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2431–2442, Dec. 2014.
[67] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 767–779, Apr. 2015.
[68] H. Zhang, X. Shang, H. Luan, M. Wang, and T.-S. Chua, "Learning from collective intelligence: Feature learning using social images and tags," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, pp. 1–23, 2016.
[69] S. Zhang, H.-S. Wong, Z. Yu, and H. H. S. Ip, "Hybrid associative retrieval of three-dimensional models," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1582–1595, Dec. 2010.
[70] Y. Zhen, Y. Gao, D.-Y. Yeung, H. Zha, and X. Li, "Spectral multimodal hashing and its application to multimedia retrieval," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 27–38, Jan. 2016.
[71] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2756–2769, Dec. 2015.
[72] L. Zhu, J. Shen, L. Xie, and Z. Cheng, "Unsupervised topic hypergraph hashing for efficient mobile image retrieval," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–14, Oct. 2016.

An-An Liu (M'10) received the B.Eng. and Ph.D. degrees from Tianjin University, Tianjin, China.

He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 2008 to 2009, where he researched with Prof. T. Kanade. He was a Visiting Scholar with the School of Computing, National University of Singapore, Singapore, in 2016, where he researched with Prof. M. Kankanhalli. His current research interests include computer vision and machine learning.



Wei-Zhi Nie received the Ph.D. degree from Tianjin University, Tianjin, China.

He is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with the NExT Center, National University of Singapore, Singapore, where he researched with Prof. T.-S. Chua. His current research interests include computer vision, machine learning, and social networks.

Yue Gao (SM'14) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China.

He is an Associate Professor with the School of Software and TNList, Tsinghua University.

Yu-Ting Su received the B.Eng. and Ph.D. degrees from Tianjin University, Tianjin, China.

He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with Case Western Reserve University, Cleveland, OH, USA. His current research interests include multimedia content analysis and security.