Multi-Modal Clique-Graph Matching for View-Based 3D Model Retrieval

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

***Abstract*— Multi-view matching is an important but a challenging task in view-based 3D model retrieval. To address this challenge, we propose an original multi-modal clique graph (MCG) matching method in this paper. We systematically present a method for MCG generation that is composed of cliques, which consist of neighbor nodes in multi-modal feature space and hyper-edges that link pairwise cliques. Moreover, we propose an image set-based clique/edgewise similarity measure to address the issue of the set-to-set distance measure, which is the core problem in MCG matching. The proposed MCG provides the following benefits: 1) preserves the local and global attributes of a graph with the designed structure; 2) eliminates redundant and noisy information by strengthening inliers while suppressing outliers; and 3) avoids the difficulty of defining high-order attributes and solving hyper-graph matching. We validate the MCG-based 3D model retrieval using three popular single-modal data sets and one novel multi-modal data set. Extensive experiments show the superiority of the proposed method through comparisons. Moreover, we contribute a novel real-world 3D object data set, the multi-view RGB-D object data set. To the best of our knowledge, it is the largest real-world 3D object data set containing multi-modal and multi-view information.**

***Index Terms*— 3D model retrieval, graph matching, image set, multi-modal.**

I. Introduction

THE use of the recently advanced 3D model retrieval technique [1]–[4] is becoming mandatory in diverse domains, 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 3D technologies for modeling, reconstruction, printing and so on, has resulted in an increasing number of 3D models. It is one of the most active areas of research in both computer vision and graphics, and multiple approaches and datasets for this task have been developed [5]–[8].

A. Motivation and Overview

3D model retrieval aims to find the relevant models from the 3D object database for a given query model [9]. Generally, the related methods can be grouped into two categories: model-based and view-based methods. Model-based methods directly utilize 3D model data as the query and extract the statistics-, volume-, and surface-geometry-based low-level features for retrieval [10]. Because 3D models might not be available in many practical applications and 3D model reconstruction usually requires high computational costs, the usage of model-based methods is severely limited. In recent years, researchers have been actively engaged in view-based methods [11]–[13]. In view-based methods, the multi-view representation of 3D models is leveraged without explicit requirement of 3D model reconstruction, and model retrieval can be accomplished by measuring the similarity between different models using multi-view data [11]. Bu *et al.* [14] extracted high-level shape features learned via deep belief networks for view representation and significantly augmented the performance of 3D model retrieval. Lu *et al.* [15] jointly explored view-based and model-based relevance in a graph-based framework for 3D object retrieval. The literature reports that view-based methods can usually achieve better performance than model-based methods [16]. Although the current methods have demonstrated superior performance for this task, two challenges still remain: a) The current methods usually extract representative views from a view pool [9], [17], and several methods further amend an incremental view selection process with relevant feedback from users. This key step is usually accomplished simply by view clustering and center selection with visual features. However, these methods are not robust and effective enough to eliminate redundant and noisy data. Both the missed inliers (the eliminated but related views) and the existence of outliers (the kept but non-related views) will have a negative influence on model matching. b) The graph matching-based methods can usually outperform the statistical model-based method because of the structural constraints of graph matching. However, the performance of the current graph matching-based methods is usually restricted either by the limited ability of preserving the global geometrical structure of the graph by the node-to-node mapping of the bipartite graph-based methods [18], [19] or by the difficulty of high-order attribute discovery and hyper-graph matching of the hyper-graph-based methods [11].

To address these problems, we propose a multi-modal clique-graph matching method for view-based

Manuscript received April 20, 2015; revised November 23, 2015 and January 11, 2016; accepted February 11, 2016. Date of publication March 10, 2016; date of current version March 29, 2016. 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. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kiyoharu Aizawa. *(Corresponding authors: An-An Liu and Wei-Zhi Nie.)*

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

Y. Gao is with Tsinghua University, Beijing 100084, China (e-mail: kevin.gaoy@gmail.com).

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/TIP.2016.2540802

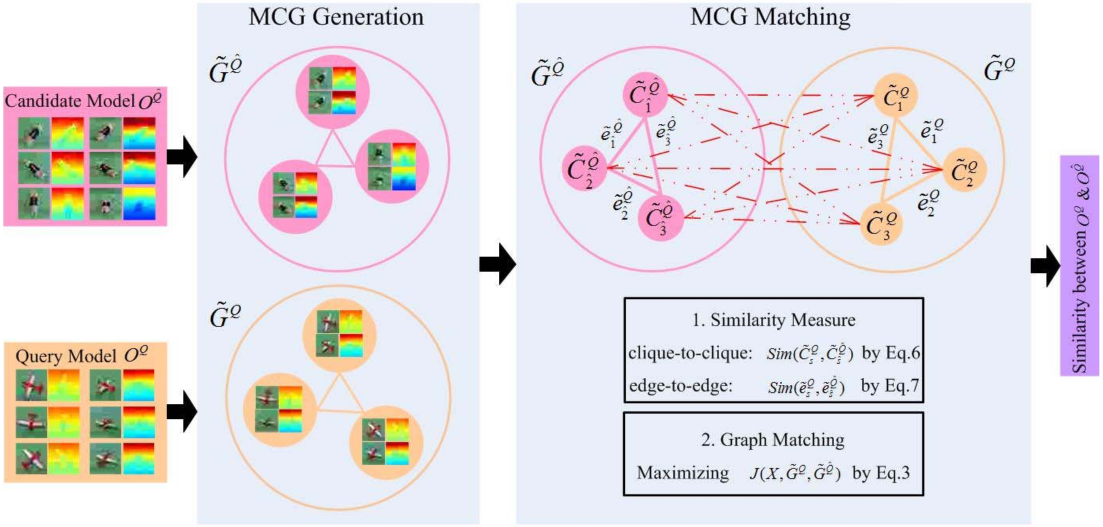
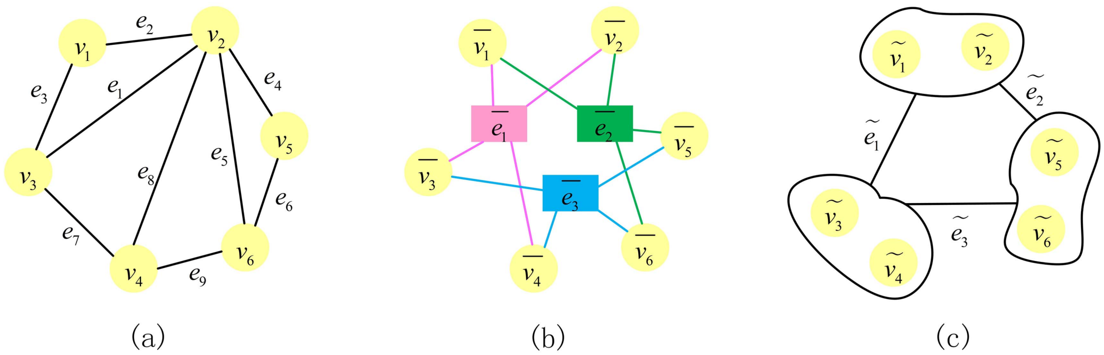


Fig. 1. The framework of the proposed MCG-based 3D model retrieval method.


v_1 v_2 v_3 v_4 v_5 v_6
e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8 e_9
$\bar{v}_1$ $\bar{v}_2$ $\bar{v}_3$ $\bar{v}_4$ $\bar{v}_5$ $\bar{v}_6$
$\bar{e}_1$ $\bar{e}_2$ $\bar{e}_3$
$\tilde{v}_1$ $\tilde{v}_2$ $\tilde{v}_3$ $\tilde{v}_4$ $\tilde{v}_5$ $\tilde{v}_6$
$\tilde{e}_1$ $\tilde{e}_2$ $\tilde{e}_3$

(a) (b) (c)

Fig. 2. Visualization of three kinds of graph structures: (a) classic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; (b) hyper-graph $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$; and (c) MCG $\tilde{\mathcal{G}} = (\tilde{\mathcal{C}}, \tilde{\mathcal{E}})$.

3D model retrieval. Fig. 1 illustrates the framework for the proposed method. We propose an original multi-modal clique graph (MCG) to represent an individual 3D model with multi-view and multi-modal information. The proposed MCG can preserve both the local and global structure of a graph. We further systematically present methods for MCG generation and matching. In particular, we propose an image set-based clique/edge-wise similarity measure to address the issue of the set-to-set distance measure, which is the core problem of MCG. Consequently, the proposed method can aid in redundant and noisy data elimination by strengthening inliers while suppressing outliers in a data-driven manner and furthermore avoid the difficulty of defining high-order attributes and solving hyper-graph matching with mapping constraints. The proposed MCG matching method is applied to view-based 3D model retrieval. The extensive experiments conducted demonstrate its superiority through comparisons.

B. Contributions

The main contributions of the proposed method are summarized as follows:

- This paper proposes an original clique-based graph matching method that considers both multi-modal and multi-view information for 3D model retrieval. Theoretically, we propose the method for multi-modal clique graph (MCG) generation and matching, which can be considered as the generalized form of a classic graph (Fig. 2(a)) and hyper-graph (Fig. 2(b)). The proposed MCG can simultaneously preserve both local arbitrary order attributes conveyed by the cliques (hyper-node) and global attributes conveyed by the structure of the MCG.
- The proposed method is extensively evaluated using three popular single-modal datasets and one novel multi-modal dataset. We discuss the effects of the fusion of RGB and depth modalities, the clique numbers of MCG, and the sparsity coefficients on the performance. We compare the proposed method with the representative methods using both single-modal and multi-modal data and further explore and compare their performances by varying the view numbers of each model.
- We contribute a real-world 3D object dataset, the Multi-view RGB-D Object Dataset (MV-RED).[1] MV-RED consists of 505 objects from 60 categories. For each object, both RGB and depth information were recorded simultaneously using 3 Microsoft Kinect sensors along 3 directions. To the best of our knowledge, it is the largest real-world 3D object dataset with multi-modal and multi-view information. MV-RED

[1] media.tju.edu.cn/mvred/dataset.html

has been leveraged as the evaluation dataset of the track of 3D Object Retrieval with Multimodal Views in the 3D Shape Retrieval Contest 2015 held in Eurographics 2015 [20].

The rest of the paper is structured as follows. In Section II, we present the related work. Then, the multi-modal clique graph matching method for view-based 3D model retrieval and the similarity measure method are detailed in Sections III and IV, respectively. Section V analyzes the computational complexity. The experimental settings and results are separately introduced in Sections VI and VII. Section VIII concludes the paper.

II. Related Work

Generally, 3D model retrieval methods are mainly classified into two categories, model-based methods and view-based methods [21].

Early approaches usually belong to model-based methods, which require an explicit 3D model data for retrieval. Popular model-based methods usually leverage geometric moments [22], surface distributions [23], voxel-based features [24], shape descriptors [25], and Fourier descriptors [26], among others, for 3D model representation. Ankerst *et al.* [27] proposed utilizing 3D shape histograms as an intuitive and powerful similarity model for 3D objects. Meanwhile, a particular flexibility was allowed by using quadratic form distance functions to account for errors of measurement. Osada *et al.* [28] proposed a novel retrieval method based on the shape feature of 3D models. They constructed the shape distribution sampled from the 3D model as the digital signature of an object and further utilized it to compute the similarity between different models. Hilaga *et al.* [29] leveraged graph models to represent 3D models based on 3D shape information. These graph models represented the skeletal and topological structure of 3D shapes at various levels of resolution. Then, they utilized a coarse-to-fine strategy to compute the similarity between different multi-resolutional graph models to handle 3D model retrieval. Sundar *et al.* [30] proposed a 3D model retrieval method based on skeletal information. They encoded the geometric and topological information in the form of a skeletal graph. They then used some popular graph matching methods to compute the similarity between different models. Model-based methods can employ all of the 3D patterns from the model for retrieval [31] and classification [32]. When no model information is available, a 3D model construction procedure is required to generate the virtual model using a collection of images. However, 3D model reconstruction is computationally expensive, and its performance is highly restricted by the sampled images. Therefore, the practical applications of model-based methods are seriously limited.

View-based methods have attracted much more attention in recent years because they are independent of 3D models and can be realized simply with the multi-view representation of models [33], [34]. Moreover, this approach can be directly extended to the retrieval of real objects, which have promising applications in e-business and location-based mobile applications, among others [35]. Chen *et al.* [36] proposed a visual similarity-based 3D model retrieval system. Zernike moments and Fourier descriptors were extracted from each view image. Then, the nearest neighbor method was used for the similarity measure between different models. Shih *et al.* [37] proposed a novel feature descriptor, *elevation descriptor*, for 3D model representation, which was invariant to translation, rotation and scaling of 3D models. Ansary *et al.* [17] proposed a Bayesian 3D object retrieval method, which utilized X-means to select representative views and applied Bayesian models to compute the similarity between different models. Gao *et al.* [9] proposed a general framework for 3D object retrieval, which was independent of camera array restriction. Each object was represented by a free set of views. 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 3D object database. Daras and Axenopoulos [16] proposed novel compact multi-view descriptors (CMVDs) for 3D model representation. Camera arrays were set at the 18 vertices of a 32-hedron to capture the CMVD, where multiple views were uniformly distributed. The two 3D models were compared by the match between the selected CMVD views. Wang and Nie [38] proposed applying group sparse coding to handle this task. They selected the view set of the individual candidate model as a dictionary to reconstruct the query model. The reconstruction error was utilized as the similarity measure for retrieval.

Recently, graph matching was widely leveraged for this task because the multi-view image set of a 3D model conveys the spatial context, which will benefit 3D model retrieval. In [18], the weighted bipartite graph was built with the representative views, and the matching results were used to measure the similarity between two 3D models. To explore the higher-order relationships between objects, Gao *et al.* [11] further proposed a hyper-graph analysis approach by avoiding the estimation of the pairwise distance between objects. Although these graph matching-based methods can achieve improvement in the retrieval performance to some extent, the current image-wise similarity measure is not robust enough and can easily have a 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 [39] for more effective 3D model retrieval. However, to the best of our knowledge, research concerning both aspects for this task is limited, especially by leveraging multi-modal and multi-view information.

III. Multi-Modal Clique Graph Matching

In this section, we introduce the proposed multi-modal clique graph (MCG) matching method for view-based 3D model retrieval. To achieve this goal, this method involves two consecutive steps, MCG generation and matching.

A. MCG Generation

Given a 3D model represented by a group of multi-view 2D images, as shown in Fig. 1, it is intuitive to design a specific graph structure to represent the visual characteristics and spatial context of 3D models and then leverage graph matching to calculate the similarity between pairwise models. A classical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of the node set $\mathcal{V} = \{v_i\}_{i=1}^{I}$ and the edge set $\mathcal{E} = \{e_j\}_{j=1}^{J}$ as shown in Fig. 2(a). Given two graphs, graph matching aims to determine the correct correspondences. This is usually realized by leveraging the unary attribute with respect to the individual node and the pairwise attribute with respect to the individual edge [40], [41]. Because pairwise relations are not enough to incorporate the entire geometrical structure information of the nodes, the hyper-graph is designed and hyper-graph matching is proposed to overcome the limitations of classical graph matching [42], [43]. A hyper-graph $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$ also consists of the node set $\bar{\mathcal{V}} = \{\bar{v}_i\}_{i=1}^{I}$ and the edge set $\bar{\mathcal{E}} = \{\bar{e}_j\}_{j=1}^{J}$. Different from $\mathcal{G}$, a hyper-graph leverages the hyper-edge, which encloses a subset of nodes, to represent the high-order attributes with respect to different scales of local structures as shown in Fig. 2(b). Although both graph structures can be applied to 3D model representation, there are two serious problems: 1) The construction of both the classic graph [18] and hyper-graph [11] highly depends on the characteristic view extraction. However, it is quite challenging to select characteristic views from multiple similar views. Moreover, the image-wise similarity measure is not robust enough when outliers exist. 2) For the hyper-graph case, defining the high-order attributes is non-trivial, and it is extremely challenging to solve the hyper-graph matching with mapping constraints [44], [45].

To tackle these problems, we propose the construction of multi-modal clique graphs for view-based 3D model representation and retrieval. On the one hand, it can effectively benefit redundant and noisy data elimination by strengthening inliers while suppressing outliers with the proposed image set-based clique/edge-wise similarity measure. On the other hand, different from previous work advancing the design of hyper-edges to convey local structures while treating individual nodes for matching, we propose regarding each clique as the basic object for matching, which can be considered as a hyper-node consisting of a set of neighbor nodes in specific feature spaces and can explicitly convey arbitrary-order attributes with the specific structure. Because each node can be represented in multiple feature spaces, we originally propose the concept of the multi-modal clique-graph (MCG) and further present the method for MCG matching in Section III-B.

As shown in Fig. 2(c), an MCG $\tilde{\mathcal{G}} = \{\tilde{\mathcal{C}}, \tilde{\mathcal{E}}\}$ is composed of two elements, the clique (hyper-node) set $\tilde{\mathcal{C}} = \{\tilde{C}_s\}_{s=1}^{S}$ and the hyper-edge set $\tilde{\mathcal{E}} = \{\tilde{e}_t\}_{t=1}^{T}$. The classical graph (Fig. 2(a)) can be transformed into MCG as follows:

- Clique: Given one model $\mathcal{O}$ with its N-view set $\{o_i\}_{i=1}^{N}$, clique discovery can be realized by utilizing a hierarchical agglomerative clustering (HAC) method [46]. Because the multi-modal representation of individual views is taken into consideration, the multi-modal features can be concatenated into the visual representation when performing HAC. Consequently, the view clusters achieved correspond to the cliques in MCG, which can explicitly preserve the consistency in feature spaces and implicitly convey the spatial correlations. The order of one clique, $\delta(\tilde{C}_s)$, equals the number of the views in it. Supposing one model $\mathcal{O} = \{o_i\}_{i=1}^{N}$ can be clustered into S cliques, the model $\mathcal{O}$ can be further represented by the clique set $\tilde{\mathcal{C}} = \{\tilde{C}_s\}_{s=1}^{S}$, where $\tilde{C}_s = \{\tilde{v}_i\}_{i=1}^{\delta(\tilde{C}_s)}$. Each node $\tilde{v}_i$ can be represented by the M-modality feature set $F_{\tilde{v}_i} = \{f_{\tilde{v}_i m}\}_{m=1}^{M}$, and each column vector $f_{\tilde{v}_{im}} \in \mathrm{R}^{D_m}$ denotes the D_m-dimension feature representation of node $\tilde{v}_i$ in the m^{th} modality. Likewise, each clique $\tilde{C}_s$ can also be represented by the M-modality feature set $F_s = \{F_{sm}\}_{m=1}^{M}$, where $F_{sm} = \{f_{\tilde{v}_i m}\}_{i=1}^{\delta(\tilde{C}_s)}$ denotes the feature representation in the m^{th} modality of all nodes in $\tilde{C}_s$. The multi-modal model $\mathcal{O}$ can be represented by the M-modality feature set $F = \{F_s\}_{s=1}^{S} = \{F_m\}_{m=1}^{M}$, where $F_m = \{F_{sm}\}_{s=1}^{S}$ denotes the feature representation in the m^{th} modality of S cliques in the model $\mathcal{O}$.
- Hyper-edge: Because each clique can already convey high-order attributes with complicated local structures, we can simplify the hyper-edge in the hyper-graph (Fig. 2(b)). As shown in Fig. 2(c), there will exist one edge between two cliques if the node-to-node edge number between them exceeds a certain threshold ϑ. To avoid the complex computation caused by the dense graph, we set ϑ as the mean of the node-to-node edge numbers between all pairwise cliques. The edges in MCG can still be considered as hyper-edges because they can represent high-order attributes by linking two cliques.

The proposed MCG provides the following benefits: 1) increasing the complexity to convey high-order attributes of hyper-nodes by replacing individual nodes with only unary attributes by cliques; 2) decreasing the complexity of hyper-edges by replacing the classic hyper-edges, which link multiple nodes to represent high-order attributes, by the proposed hyper-edges, which only link two cliques and can be easily represented by pairwise attributes. In this way, the MCG can preserve local and global structural attributes while avoiding the difficulty of defining the high-order attributes and solving hyper-graph matching with mapping constraints, which will be further explained in Section III-B.

B. MCG Matching

The task of 3D object retrieval requires computing the similarity scores between the query model and individual candidate model. Given two MCGs, $\tilde{\mathcal{G}}^{Q} = \{\tilde{\mathcal{C}}^{Q}, \tilde{\mathcal{E}}^{Q}\}$ and $\tilde{\mathcal{G}}^{\hat{Q}} = \{\tilde{\mathcal{C}}^{\hat{Q}}, \tilde{\mathcal{E}}^{\hat{Q}}\}$, we aim to achieve the optimal similarity measure, $J(\bar{X}, \tilde{\mathcal{G}}^{Q}, \tilde{\mathcal{G}}^{\hat{Q}})$, by considering both the structure characteristics $(\tilde{\mathcal{G}}^{Q}/\tilde{\mathcal{G}}^{\hat{Q}})$ and the clique-to-clique correspondence $(\bar{X})$.

Because each clique consisting of a view set can convey arbitrary order attributes to preserve the local structure, we only consider the clique-wise and edge-wise attributes for representing the global graph structure for MCG matching. Given a pair of graphs $\tilde{\mathcal{G}}^{Q}\&\tilde{\mathcal{G}}^{\hat{Q}}$, two affinity matrices, $K^{C} \in R^{S_1 \times S_2}$ and $K^{E} \in R^{T_1 \times T_2}$, are computed to measure the similarity of the pairwise cliques and edges, respectively. More specifically, $k^{C}_{s_1 \hat{s}_1} = \Phi_C(\tilde{C}^{Q}_{s_1}, \tilde{C}^{\hat{Q}}_{\hat{s}_1})$ measures the similarity between clique $\tilde{C}^{Q}_{s_1}$ in $\tilde{\mathcal{G}}^{Q}$ and clique $\tilde{C}^{\hat{Q}}_{\hat{s}_1}$ in $\tilde{\mathcal{G}}^{\hat{Q}}$, and $k^{E}_{t\hat{t}} = \Phi_{E}(\tilde{e}^{Q}_{t}, \tilde{e}^{\hat{Q}}_{\hat{t}})$ measures the similarity between edge $\tilde{e}^{Q}_{t}$ in $\tilde{\mathcal{G}}^{Q}$ and edge $\tilde{e}^{\hat{Q}}_{\hat{t}}$ in $\tilde{\mathcal{G}}^{\hat{Q}}$. The formulations of both are detailed in Section IV. In this way, the problem of MCG matching can be formulated for finding the correspondence between the cliques of $\tilde{\mathcal{G}}^{Q}\&\tilde{\mathcal{G}}^{\hat{Q}}$ that maximizes the following score function for global consistency:

$$\begin{aligned} &J(\bar{X}, \tilde{\mathcal{G}}^{Q}, \tilde{\mathcal{G}}^{\hat{Q}}) \\ &\quad = \sum_{s_1,\hat{s}_1} x_{s_1\hat{s}_1} k^{C}_{s_1\hat{s}_1} + \sum_{\substack{s_1 \neq s_2, \hat{s}_1 \neq \hat{s}_2, \\ h^{Q}_{s_1 t}.h^{Q}_{s_2 t} = 1, \\ h^{\hat{Q}}_{\hat{s}_1\hat{t}}.h^{\hat{Q}}_{\hat{s}_2\hat{t}} = 1}} x_{s_1\hat{s}_1} x_{s_2\hat{s}_2} k^{E}_{t\hat{t}}, \end{aligned} \tag{1}$$

where matrix $\bar{X} \in \{0,1\}^{S_1\times S_2}$ denotes the clique correspondence, i.e., $x_{s_1\hat{s}_1} = 1$ if the s_1^{th} clique in $\tilde{\mathcal{G}}^{Q}$ corresponds to the $\hat{s}_1^{th}$ clique in $\tilde{\mathcal{G}}^{\hat{Q}}$. Considering that one clique in $\tilde{\mathcal{G}}^{Q}$ can only be linked with one clique in $\tilde{\mathcal{G}}^{\hat{Q}}$ at most, Eq.1 should be constrained by one-to-one matching with affinity constraints, i.e, $X \cdot 1_{S_2} \leq 1_{S_1}$, and $X^{T} \cdot 1_{S_1} \leq 1_{S_2}$.

$J(\bar{X}, \tilde{\mathcal{G}}^{Q}, \tilde{\mathcal{G}}^{\hat{Q}})$ is equivalent to the quadratic form, $X^{T}KX$, where $X \in \{0,1\}^{S_1 S_2}$ (the vectorization of matrix $\bar{X}$) is an indicator vector and $K \in R^{S_1 S_2 \times S_1 S_2}$ can be computed by:

$$k_{s_1 s_2 \hat{s}_1 \hat{s}_2} = \begin{cases} k^{C}_{s_1\hat{s}_1}, & if \quad s_1 = s_2 \ \ \& \ \ \hat{s}_1 = \hat{s}_2 \\ k^{E}_{t\hat{t}}, & if \quad \begin{array}{l} s_1 \neq s_2 \ \ \& \ \ \hat{s}_1 \neq \hat{s}_2 \ \ \& \\ h^{Q}_{s_1 t}.h^{Q}_{s_2 t}.h^{\hat{Q}}_{\hat{s}_1\hat{t}}.h^{\hat{Q}}_{\hat{s}_2\hat{t}} = 1 \end{array} \\ 0, & otherwise. \end{cases} \tag{2}$$

With these notations, MCG matching can be converted to optimize the following Integral Quadratic Programming problem (IQP) with affinity constraints:

$$\begin{aligned} &\max_{X} J(X, \tilde{\mathcal{G}}^{Q}, \tilde{\mathcal{G}}^{\hat{Q}}) = X^{\top} K X, \\ &s.t. \ \ X \leq 1 \ \ and \ \ X \in \{0,1\}^{S_1 S_2} \end{aligned} \tag{3}$$

For MCG matching in Eq.3, multiple state-of-the-art IQP methods, such as semidefinite programming [47], graduated assignment [48], and spectral matching [49], can be leveraged. However, solving the IQP with affinity constraints would prove difficult. In our work, we implement the Rayleigh Quotient Maximization with affinity constraints [50]. The objective function in Eq. 3 is first transferred into the Rayleigh Quotient with linear constraints, and then, approximate solutions are implemented by continuous relaxation. At last, discretization by the greedy procedure [48] is performed to tighten the approximation result of the one-to-one matching.

Supposing we can measure the similarity of pairwise cliques and edges ($k^{C}_{s_1\hat{s}_1}$ & $k^{E}_{t\hat{t}}$), the similarity score between the query model and candidate models can be achieved by computing $J(X, \tilde{\mathcal{G}}^{Q}, \tilde{\mathcal{G}}^{\hat{Q}})$. Then, the ranking list with the similarity scores in descending order will be return for retrieval. We now consider the similarity measure between pairwise cliques/edges of two MCGs, which is detailed in Section IV.

IV. Similarity Measure

Different from the previous graph matching method focusing on the first & second order or the high-order based distance measure, we propose the image set-based pariwise clique/edge distance measure for two reasons: 1) compared to the node-to-node method, computing the similarity by leveraging image set, which implicitly conveys complicated local structural attributes, can be more robust, and 2) compared to the hyper-edge-to-hyper-edge method, the difficulty in designing and computing high-order hyper-edge similarity measures can be avoided. In the following sub-sections, we detail the formulation and related optimization method of the proposed image set-based pariwise clique/edge similarity measure.

A. Formulation

Given a query model $\mathcal{O}^{Q}$ represented as $\tilde{\mathcal{G}}^{Q} = \{\tilde{\mathcal{C}}^{Q}, \tilde{\mathcal{E}}^{Q}\}$ and one candidate model $\mathcal{O}^{\hat{Q}}$ represented as $\tilde{\mathcal{G}}^{\hat{Q}} = \{\tilde{\mathcal{C}}^{\hat{Q}}, \tilde{\mathcal{E}}^{Q}\}$, we need to compute the similarity between pairwise cliques of both models, $Sim(\tilde{C}^{Q}_{s}, \tilde{C}^{\hat{Q}}_{\hat{s}})$, and the similarity between pairwise edges of both models, $Sim(\tilde{e}^{Q}_{t}, \tilde{e}^{\hat{Q}}_{\hat{t}})$ (edge $\tilde{e}^{Q}_{t}$ links the cliques $\tilde{C}^{Q}_{s_1}\&\tilde{C}^{Q}_{s_2}$ in $\mathcal{O}^{Q}$, and edge $\tilde{e}^{\hat{Q}}_{\hat{t}}$ links the cliques $\tilde{C}^{\hat{Q}}_{\hat{s}_1}\&\tilde{C}^{\hat{Q}}_{\hat{s}_2}$ in $\mathcal{O}^{\hat{Q}}$). Because each clique/hyper-edge is essentially an image set or a group of image sets, the task of computing the similarity measure between pairwise cliques/edges can be converted into the problem of computing the set-to-set distance measure. To address this problem, we propose the image set-based pairwise clique/edge distance measure method by considering both the local characteristics and global context for the similarity measure.

Consider a clique $\tilde{C}^{Q}_{s}$ with the feature set $F_s = \{F_{sm}\}_{m=1}^{M}$ from the query model and multiple cliques $\{\tilde{C}^{\hat{Q}}_{\hat{s}}\}_{\hat{s}=1}^{\hat{S}}$ with the feature sets $\hat{F} = \{\hat{F}_m\}_{m=1}^{M}$ from a candidate model. We model F_s in the m^{th} modality as a hull, $F_{sm}a$, where a is the coefficient vector, and we then formulate the convex objective function, $\Omega(F_s, \hat{F}, a, b)$, to reconstruct $F_{sm}a$ with $\hat{F}_m$, inspired by the collaboration representation theory [51], [52]:

$$\begin{aligned} \Omega(F_s, \hat{F}, a, b) &= \sum_{m=1}^{M} \phi_m ||F_{sm}a - \hat{F}_m b||^2 + Reg(a, b) \\ &s.t. \sum_{i=1,\ldots,\delta(\tilde{C}^{Q}_{s})} a_i = 1, \end{aligned} \tag{4}$$

where a_i is the i^{th} coefficient in a and $\sum_{i=1,\ldots,\delta(\tilde{C}^{Q}_{s})} a_i = 1$ is required by the hull formulation and can also avoid the trivial solution $a = b = 0$; b is the coefficient for reconstruction and can be decomposed as $b = [b_1, \ldots, b_{\hat{s}}, \ldots, b_{\hat{S}}]$, where $b_{\hat{s}}$ is the sub-vector of coefficients associated with the clique $\tilde{C}^{\hat{Q}}_{\hat{s}}$ in the candidate model; and ϕ_m denotes the weight of the m^{th} modality. In Eq. 4, the hull $F_{sm}a$ of the query set F_s in the m^{th} modality is collaboratively represented over the cliques in the m^{th} modality of the candidate model. The coefficients in a will lead to the samples in F_{sm} being treated differently in the representation. By minimizing the distance between $F_{sm}a$ and $\hat{F}_m b$, the outliers in both F_{sm} and $\hat{F}_m$ will be assigned very small coefficients. Therefore, the impact of the outliers can be reduced greatly. The objective function above is composed of two parts:

- Fidelity: The first term penalizes the sum-of-squares difference between $F_{sm}a$ and $\hat{F}_m b$. Assuming that each clique in the query model can be constructed well by the feature set of one candidate model, the fidelity term should produce a small residual.
- Regularization: Two important cases should be taken into consideration for the regularization term: 1) because $\hat{F}$, consisting of cliques both related and unrelated to F_s, contains more bases for collaboration representation, strengthening the coefficients of related cliques while significantly suppressing those of non-related cliques is intuitive. Therefore, a sparsity penalty should be implemented on b. 2) Outliers might exist in the clique F_s output by clustering. Therefore, a sparsity penalty is also required for a. The Lasso penalty is well known for imposing a sparsity penalty for decomposition. Therefore, $|| \cdot ||_1$ (L1 norm) is implemented for both $a\&b$.

The proposed convex objective function can therefore be formulated as follows:

$$\Omega(F_s, \hat{F}, a, b) = \sum_{m=1}^{M} \phi_m ||F_{sm}a - \hat{F}_m b||^2 + \gamma_1 ||a||_1 + \gamma_2 ||b||_1$$
$$s.t. \quad \sum_{i=1,\ldots,\delta(\tilde{C}_s^Q)} a_i = 1, \tag{5}$$

where γ_1 & γ_2 are sparsity coefficients. By minimizing Eq. 5, we can obtain the optimal coefficient vectors a^* and b^*. b^* can be rewritten as $b^* = [b_1^*, \ldots, b_{\hat{s}}^*, \ldots, b_{\hat{S}}^*]$, where $b_{\hat{s}}^*$ is the sub-vector of coefficients associated with the clique $\tilde{C}_{\hat{s}}^{\hat{Q}}$ in the candidate model.

Using the collaborative representation of F_s and $\hat{F}$, the clique/edge-wise similarity can be defined as follows:

- The residual of clique $\tilde{C}_s^Q$ constructed by individual clique $\tilde{C}_{\hat{s}}^{\hat{Q}}$ can be leveraged for the clique-to-clique similarity:

$$k_{s\hat{s}}^C = Sim(\tilde{C}_s^Q, \tilde{C}_{\hat{s}}^{\hat{Q}})$$
$$= exp\{-\sum_{m=1}^{M} \phi_m ||F_{sm}a^* - \hat{F}_{\hat{s}m} b_{\hat{s}}^*||^2\} \tag{6}$$

- Because each edge can be considered as a pair of image sets, each of which corresponds to the image set of one clique, the edge-to-edge similarity can be likewise formulated by considering 2 possible matches between two edges:

$$k_{t\hat{t}}^E = Sim(\tilde{e}_t^Q, \tilde{e}_{\hat{t}}^{\hat{Q}})$$
$$= exp\{-\sum_{m=1}^{M} \phi_m \sum_{\substack{s \in \{s_1, s_2\} \\ \hat{s} \in \{\hat{s}_1, \hat{s}_2\}}} ||F_{sm}a^* - \hat{F}_{\hat{s}m} b_{\hat{s}}^*||^2/2\} \tag{7}$$

Because each hyper-edge in the traditional hyper-graph can be considered as a group of image sets, each of which corresponds to one clique, the hyper-edge-to-hyper-edge similarity can be easily extended from the edge-to-edge similarity measure in Eq. 7 by updating the edges (s and $\hat{s}$) with the hyper-edges. Therefore, the proposed MCG matching can be easily extended to hyper-graph matching by replacing each node with one clique and utilizing the classic objective function for hyper-graph matching as used in [44]. We now develop the solution of Eq. 5, which is detailed in the next section.

B. Optimization

For Eq. 5, we have the following augmented Lagrangian function:

$$L(a, b, \lambda) = \sum_{m=1}^{M} \phi_m ||F_{sm}a - \hat{F}_m b||^2 + \gamma_1 ||a||_1 + \gamma_2 ||b||_1$$
$$+ < \lambda, ea - 1 > + \frac{\tau}{2} ||ea - 1||_2^2 \tag{8}$$

where λ is the Lagrange multiplier, $< ., . >$ is the inner product, $\tau > 0$ is the penalty parameter, and e is a row vector whose elements are 1.

Because Eq. 8 is not jointly convex with $a\&b$, we suitably adapt the alternating minimization theory [53] for the optimal solution, which is very efficient for solving multiple variable optimization problems. a and b can be optimized alternatively with the other one fixed:

- Optimizing a with the fixed b^t:

$$a^{t+1} = \arg\min_a L(a, b^t, \lambda^t)$$
$$= \arg\min_a f(a) + \frac{\tau}{2} ||ea - 1 + \lambda^t/\tau||_2^2$$
$$= \arg\min_a ||\mathcal{F}a - \hat{\mathcal{F}}||_2^2 + \gamma_1 ||a||_1 \tag{9}$$

where $f(a) = \sum_{m=1}^{M} \varphi_m ||F_{sm}a - \hat{F}_m b^t||^2 + \gamma_1 ||a||_1$, $\mathcal{F} = [\varphi_1^{1/2} F_{s1}; \ldots; \varphi_M^{1/2} F_{sM}; (\tau/2)^{1/2} e]$, $\hat{\mathcal{F}} = [\varphi_1^{1/2} \hat{F}_1 b^t; \ldots; \varphi_M^{1/2} \hat{F}_M b^t; (\tau/2)^{1/2} (1 - \lambda^t/\tau)]$.
The subproblem in Eq. 9 can be easily solved using the representative l_1-minimization approaches, such as LARS [54].

- Optimizing b with the updated a^{t+1}:

$$b^{t+1} = \arg\min_b L(a^{t+1}, b, \lambda^t)$$
$$= \arg\min_b ||\mathcal{F}' - \hat{\mathcal{F}}' b||_2^2 + \gamma_2 ||b||_1 \tag{10}$$

where $\hat{\mathcal{F}}' = [\varphi_1^{1/2} \hat{F}_1; \ldots; \varphi_M^{1/2} \hat{F}_M]$, $\mathcal{F}' = [\varphi_1^{1/2} F_{s1} a^{t+1}; \ldots; \varphi_1^{1/2} F_{sM} a^{t+1}]$. The subproblem Eq. 10 can also be solved by LARS.

Once both a^{t+1} and b^{t+1} are updated, λ can be updated by

$$\lambda^{t+1} = \lambda^t + \gamma (ea^{t+1} - 1) \tag{11}$$

The algorithm for the image set-based similarity measure (ISBSM) is summarized in Algorithm 1. It has been shown in [55], for the general convex problem, that the alternating minimization approach would converge to the correct solution.

Algorithm 1 ISBSM Algorithm

```
Input:
  feature set $F_s$ of clique $\tilde{C}_s^Q$;
  feature set $\hat{F}$ of one candidate model;
  $\gamma_1$ ; $\gamma_2$; max iteration number $T_{max}$.
Output:
  $k_{s\hat{s}}^C$; $k_{t\hat{t}}^E$ .
  Initialization $b^0$, $\lambda^0$, $t \longleftarrow 0$.
while $t < T_{max}$ do
    Step 1: Update $a$ by Eq.9;
    Step 2: Update $b$ by Eq.10;
    Step 3: Update $\lambda$ by Eq.11;
    Step 4: $t \longleftarrow t + 1$.
Compute $k_{s\hat{s}}^C$ by Eq.6;
Compute $k_{t\hat{t}}^E$ by Eq.7.
```

V. Complexity Analysis

According to the algorithms introduced in Sections III & IV, we analyze the computational cost of the proposed method. The computational cost of view clustering is $O(N^2 \log N)$, where N is the number of views in an individual 3D model. According to [9], this step is an offline procedure. The similarity measure process costs $O(Sd_f^2 T_{max}(\bar{N}_C^{\varrho} + \bar{N}^{\varrho}))$, where S is the clique number, d_f is the feature dimension, T_{max} is the number of maximal iterations in Algorithm 1, $\bar{N}_C$ is the average number of views in each clique, and $\varrho \geq 1.2$ [52], [56]. The MCG matching process costs $O(T^2)$ [50], where T is the number of hyper-edges in MCG. Therefore, the computational cost of the MCG-based method scales as $O(\mathcal{N}(Sd_f^2 T_{max}(\bar{N}_C^{\varrho} + \bar{N}^{\varrho}) + T^2))$, where $\mathcal{N}$ is the number of 3D models in the database.

VI. Experimental Methods

The proposed method was quantitatively evaluated for view-based 3D model retrieval in single/multiple modalities. In this section, we describe the datasets for testing, experimental settings, methods for comparison, and evaluation criteria. The source code will be made public after the patent is approved.

A. Dataset

Three popular datasets with single modality were utilized for the evaluation. Moreover, we prepared a novel dataset, Multi-view RGB-D Object Dataset (MV-RED), to evaluate the proposed method in the multi-modal scenario. The four datasets are briefly introduced as follows:

- NTU60/NTU216 [36]: The NTU dataset contains 500 models from 50 categories. Because this dataset only provides 3D models, which cannot be directly used for view-based methods, we capture views of individual models by setting up a virtual camera array consisting of 60 cameras. The cameras are set on the vertices of a polyhedron with the same structure as that of Buckminsterfullerene (C60). Consequently, individual 3D 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. They all form 6 faces of each 3D model, and each face contains 36 views. Therefore, an individual model contains 216 views. We term NTU in this setting as NTU216.

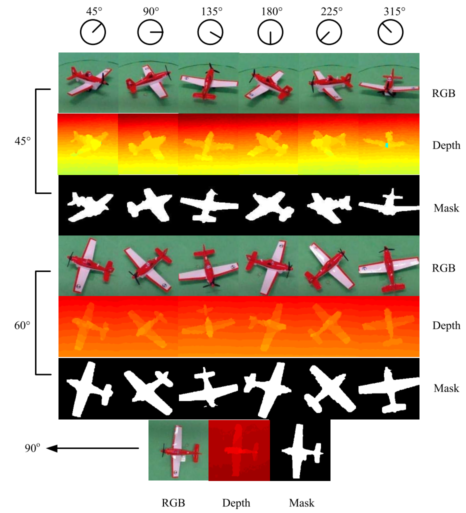


Fig. 3. Samples of MV-RED.

- PSB [57]: PSB contains 1,814 models from 161 classes and only contains 3D model information. Similar to NTU, we generate a view set representation for each model with a virtual camera array containing 60 cameras so that PSB is suitable for view-based methods.
- ETH [58]: ETH contains 80 objects from 8 categories and provides each object with 41 views, which are captured using the camera array spaced evenly over the upper viewing hemisphere and where the position for each camera is set by subdividing the faces of an octahedron to the third recursion level.
- MV-RED: We contribute a real-world 3D object dataset with multimodal views, the Multi-view RGB-D Object Dataset (MV-RED) (Fig. 3). The MV-RED dataset consists of 505 objects from 60 categories. For each object, information on both the RGB and depth was recorded simultaneously by 3 Microsoft Kinect sensors along 3 directions. This dataset is recorded under two different recording settings: 1) 202 objects were recorded with 0°, 45° and 90° cameras, and 2) 303 objects were recorded with 45°, 60° and 90° cameras. Camera 0°(setting1)/45°(setting2) and Camera 45°(setting1)/60°(setting2) respectively captured 360 RGB and depth images by uniformly rotating the table controlled by the step motor. Camera 90° only captured one RGB image and one depth image in the top-down view. In this way, each object has 721 RGB images and 721 depth images in total. The resolution of the RGB/depth image is 640 × 480. Foreground segmentation was implemented for the dataset, and masks were provided. To reduce the redundant information and simplify the computational complexity, we uniformly sampled the images from the first and the second views every 10 degrees and provided the compact version with 73 images. The difference between these two settings lies in the directions for view capturing, which is set to increase the difficulty in view matching.

B. Experimental Setting

We extracted the Zernike moments as the visual features in the RGB modality, which have been widely used in view-based 3D model retrieval [6], [9], [11] due to their robustness to view scaling and rotation. Similar to [60] and [61], we extracted the histogram of the oriented gradient in the depth modality because it can represent the shape characteristics well. The proposed method is validated under five scenarios:

- The proposed method is extensively compared to several representative methods in 6 scenarios: 1) 3 datasets with single modality (NTU60/NTU216, PSB, ETH); 2) RGB data from MV-RED (R-MV-RED), depth data from MV-RED (D-MV-RED), and both RGB and depth data from MV-RED (MV-RED).
- The proposed method is compared to several competing methods by varying the view numbers, N, of each model with the N-view set ($\{O_i\}_{i=1}^{N}$) to evaluate its effect on the performance.
- We vary the weights ($\phi_1\&\phi_2$ in Eq. 4) to explore the importance of RGB and depth modalities on the performance.
- We vary the clique number, S, of MCG ($\tilde{\mathcal{C}} = \{\tilde{C}_s\}_{s=1}^{S}$) according to individual models to evaluate its effect on the performance.
- We vary the sparsity coefficients ($\gamma_1\&\gamma_2$) to evaluate the effect on the performance.

C. Evaluation

For the evaluation of each dataset, each 3D model is selected as the query once for retrieval. To evaluate the 3D model retrieval performance, the following popular criteria are employed as the measures of the retrieval performance.

1) The Precision-Recall Curve (PR-Curve) [17] can comprehensively demonstrate the retrieval performance and illustrates the precision and recall measures by varying the threshold for distinguishing relevance and irrelevance in model retrieval.
2) The Nearest Neighbor (NN) evaluates the retrieval accuracy of the first returned result.
3) The 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) The Second Tier (ST) is defined as the recall of the top 2κ results.
5) The F-measure (F) jointly evaluates the precision and the recall of the top returned results. In our experiments, the top 20 retrieved results are used for calculation.
6) The Discounted Cumulative Gain (DCG) [61] 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) The Average Normalized Modified Retrieval Rank (ANMRR) [62] 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 are ranked at the top positions.

D. Competing Methods

The proposed MCG-based method is termed as MCG. Several popular methods are implemented for comparison.

- Adaptive view clustering (AVC) [17]: AVC selects the optimal 2D characteristic views of a 3D model based on the adaptive clustering algorithm and then utilizes a probabilistic Bayesian method for 3D model retrieval.
- Camera constraint free view (CCFV) [9]: For each query object, all query views are first grouped to generate view clusters, which are used to represent the query model. Then, a positive matching model and a negative matching mode are individually trained with positive and negative matched samples, and a CCFV model is generated on the basis of the query Gaussian models by combining the positive matching model and the negative matching model. CCFV removes the constraint of the setting of the static camera array for view capture and can be applied to any view-based 3D object database.
- Weighted Bipartite Graph Matching (WBGM) [18] & Spectral Matching (SM) [49] & Reweighted Random Walks Matching (RRWM) [43]: Representative views are first selected from the query model and the candidate model. The initial weights of the representative views are initialized and further updated based on the correlations among them. WBGM builds the weighted bipartite graph only with the attributes of individual 2D views. In comparison, SM and RRWM build the graph with both unary and pairwise attributes of the representative views and implement different algorithms for graph matching. The matching score of the individual method is utilized to measure the similarity between two 3D models.
- Hausdorff distance (HAUS) & Nearest Neighbor (NN): We implement HAC [46] for view clustering and select the one with the smallest distance from the rest of the views in each cluster as the representative view. The Hausdorff distance is used to measure the maximum distance between a set and its nearest point in the other set. The nearest neighbor-based method is similar to HAUS. The only difference between NN and HAUS is that NN leverages the Euclidean distance between the characteristic views of two models as the similarity measure.

VII. Experimental Results and Discussions

A. MCG vs. State of the Arts

The proposed method is extensively compared to the representative methods. The proposed method is executed in two ways: 1) single modality: Because NTU60/NTU216, PSB, ETH, R-MV-RED. and D-MV-RED only contain RGB/depth data, the proposed method can be implemented by setting the weight of the corresponding modality as 1 and setting the other as 0; 2) multiple modality: for MV-MED, the weights for the RGB and depth modalities are set as 0.2 and 0.8, respectively. The clique number in MCG is set as 6. The sparsity coefficients in Eq. 5, $\gamma_1\&\gamma_2$, are set as 0.03 and 0.01, respectively. The selection of these parameters is detailed in Section VII-C-VII-E.

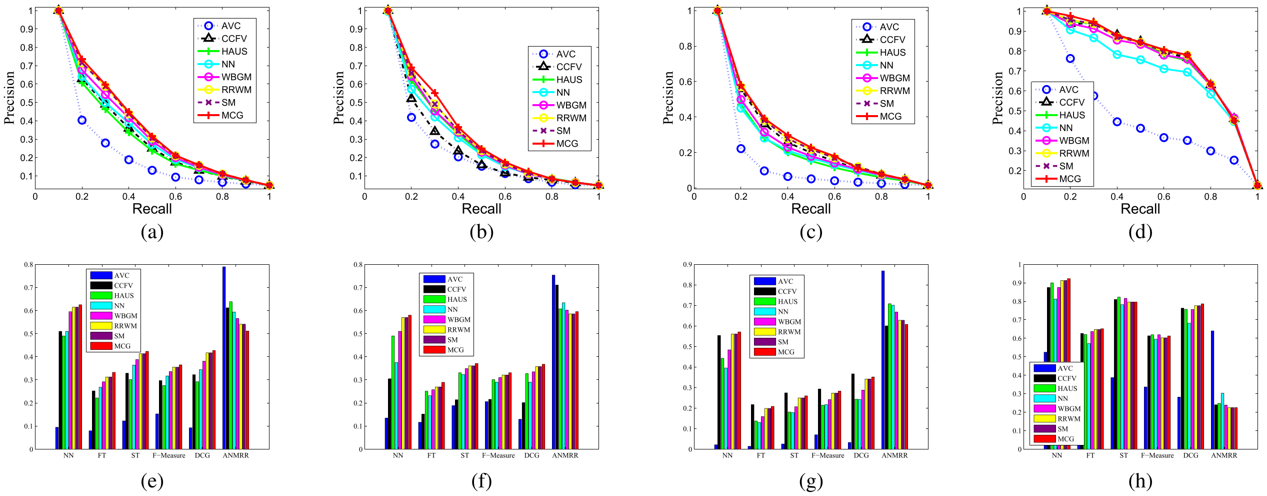


Fig. 4. Comparison on NTU60, NTU216, PSB, ETH. (a) PR Curve@NTU60. (b) PR Curve@NTU216. (c) PR Curve@PSB. (d) PR Curve@ETH. (e) Comparison@NTU60. (f) Comparison@NTU216. (g) Comparison@PSB. (h) Comparison@ETH.

The PR-Curve and the quantitative evaluation of NTU60/NTU216, PSB, and ETH are shown in Fig. 4. With these comparisons, we make the following observations.

- MCG can consistently outperform all the others because it can discover the cliques to preserve the local structures and further leverage clique-wise and edge-wise similarities for clique graph matching to preserve the global correspondence. In the NTU60, the proposed method can achieve a gain of $4\% - 500\%$, $2.6\% - 233\%$, $6\% - 255\%$, $10\% - 113\%$, and $7.5\% - 32\%$ in terms of NN, FT, ST, F-measure, and DCG, respectively, and also achieve a decline of $4.1\% - 31\%$ in terms of ANMRR compared with AVC, CCFV, NN, HAUS, WBGM, SM, and RRWM. For NTU-216, the proposed method can achieve a gain of $11\% - 392\%$, $4.8\% - 118\%$, $3.9\% - 97\%$, $3.2\% - 40\%$, and $4\% - 146\%$ in terms of NN, FT, ST, F, and DCG, respectively, and achieve a decline of $3.3\% - 23.3\%$ in terms of ANMRR. For PSB, the proposed method can achieve a gain of $3.5\% - 1733\%$, $5\% - 600\%$, $9.1\% - 380\%$, $6.7\% - 350\%$, and $5.9\% - 700\%$ in terms of NN, FT, ST, F, and DCG, respectively, and achieve a decline of $4.1\% - 31\%$ in terms of ANMRR. For ETH, the proposed method can achieve a gain of $2\% - 73\%$, $3\% - 94\%$, and $3\% - 115\%$ in terms of NN, FT, and DCG, respectively, and achieve a decline of $1\% - 61\%$ in terms of ANMRR.
- Generally speaking, the non-parametric methods, MCG & SM & RRWM & WBGM, can outperform the parametric methods, including HAUS, NN, CCFV, and AVC, on most of datasets, except PSB. The non-parametric methods usually select the representative views or cliques to identify the characteristics from different views and then leverage graph matching for the similarity measure. The strict constraints of graph matching can usually yield a relatively stable similarity measure. In comparison, the parametric methods usually learn a statistical model to represent a cluster of a view set (HAUS/NN) or a model (CCFV/AVC). Because the number of samples in NTU60, NTU216, and ETH is limited, the parametric methods do not have a high generalization ability, which is echoed in Fig. 4. The experiment on PSB shows that the parametric method (CCFV) can be improved to outperform WBGM when the dataset (PSB) is large enough.
- WBGM/SM/RRWM only keep the representative views of one model, which can be regarded as the members of the corresponding cliques in MCG. However, it is usually difficult to select representative views from multiple similar views. Moreover, the image-wise similarity measure is not robust enough when outliers exist. Compared to WBGM/SM/RRWM, MCG replaces the image-wise node with the clique, consisting of a set images with similar visual patterns. Although outliers might exist, the image set-wise similarity measure can significantly enhance inliers while suppressing outliers. Therefore, MCG can theoretically and experimentally outperform WBGM/SM/RRWM.

The proposed MCG can be naturally utilized for multimodal 3D model retrieval. We empirically leverage the feature-level fusion of RGB and depth features with the same weights as MCG for the competing methods and compare MCG to them in 3 scenarios, including R-MV-RED/D-MV-RED/MV-RED. The PR-Curve and the quantitative evaluation are shown in Fig. 5. With these comparisons, we make the following observations.

- From Fig. 5(a-c)(e-g), MCG can consistently outperform all the others by preserving both local structures and global correspondence in all three cases. In the R-MV-RED, the proposed method can achieve a gain of $2.9\% - 520\%$, $6.6\% - 150\%$, $3\% - 167\%$, $3.9\% - 106\%$, and $3\% - 356\%$ in terms of NN, FT, ST, F, and DCG, respectively, and achieve a decline of $2\% - 31.6\%$ in terms of ANMRR. In the D-MV-RED, the proposed method can achieve a gain of $8.6\% - 91.5\%$, $6.2\% - 60.9\%$, $6\% - 18.0\%$, $5.5\% - 32.1\%$, and $5\% - 78.6\%$ in terms of NN, FT, ST, F, and DCG, respectively, and achieve

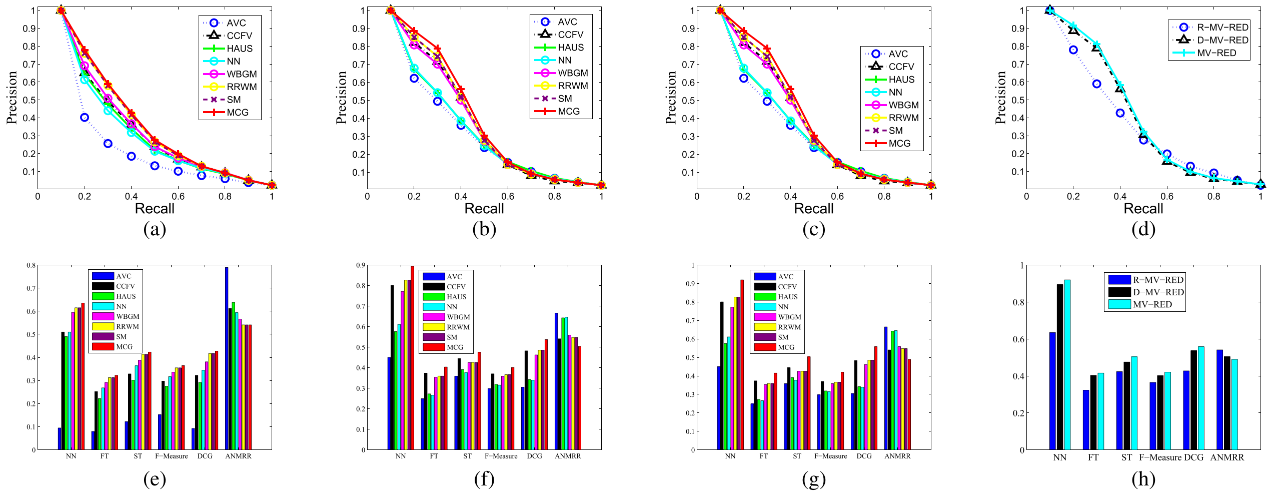


Fig. 5. Comparison on R-MV-RED, D-MV-RED, MV-RED. (a) PR Curve@R-MV-RED. (b) PR Curve@D-MV-RED. (c) PR Curve@MV-RED. (d) PR Curve by MCG in 3 cases. (e) Comparison@R-MV-RED. (f) Comparison@D-MV-RED. (g) Comparison@MV-RED. (h) Comparison by MCG in 3 cases.

a decline of 4.5% − 20.6% in terms of ANMRR. In the MV-RED, the proposed method can achieve a gain of 5.7% − 113%, 12.5% − 65.8%, 9% − 25%, 15% − 45.1%, and 11% − 70.3% in terms of NN, FT, ST, F, and DCG, respectively, and achieve a decline of 9% − 30.7% in terms of ANMRR.

- The fusion of multiple modalities can be advantageous for 3D model retrieval. Fig. 5(d, h) shows that MCG with multimodal information can achieve a gain of 33.2%, 46%, 31%, 23%, and 14% in terms of NN, FT, ST, F, and DCG, respectively, and achieve a decline of 36% in terms of ANMRR compared to MCG with only RGB information and obtain a gain of 11.5%, 23%, 22.3%, and 10% in terms of FT, ST, F, and DCG, respectively, and achieve a decline of 9% in terms of ANMRR against MCG with only depth information. It also shows that the depth modality is more important than the RGB modality for this task because we can extract more discriminative visual features in the depth modality, which can directly benefit 3D model retrieval.

B. Comparison by Varying the View Numbers

For the real application, it is always expected that 3D model retrieval is conducted with as few view images as possible. To further demonstrate the robustness of the proposed method, we compare MCG with other representative methods by varying the view numbers of individual 3D models. The view number, N, is varied from 10 to 70 with a step size of 10 to evaluate its effect on the performance. The RGB, depth weights, clique number, and sparsity coefficients are set to the same values as those in Section VII-A. The competing methods, including AVC, CCFV, HAUS, NN, WBGM, SM, and RRWM, are implemented by following the original problem statement with the feature-level fusion explained in Section VII-A. The quantitative result with respect to a specific view number is achieved by averaging 10 random trials.

From Fig. 6, we have the following observations.

- The performances of all the methods can be improved when the view number is increased. Similar to the comparison in Section VII-A, the non-parametric methods can outperform the parametric methods, which is reasonable because a greater number of view images can convey more appearance and structural characteristics of 3D models. Consequently, we can obtain better performance for real applications.
- Fig. 6 shows that MCG can consistently outperform the competing methods in terms of NN, FT, ST, F-measure, DCG and ANMRR when varying N from 10 to 70. MCG can achieve an average gain of 12.1%, 9.2%, 10.2%, 11.3%, and 9.3% in terms of NN, FT, ST, F-measure, and DCG, respectively, and achieve a decline of 8.2% in terms of ANMRR when N is tuned from 10 to 70. It can outperform the second best method with an average gain of 3.2%, 5.1%, 4.7%, 3.8%, and 5.2% in terms of NN, FT, ST, F-measure, and DCG, respectively, and achieve a decline of 4.1% in terms of ANMRR. In particular, MCG with 50 views can outperform all the competing methods with 70 images in terms of all criteria.

C. Effect of Weights on RGB and Depth Fusion

The weights ($\phi_1 \& \phi_2$) for RGB and depth fusion are varied to explore the importance of RGB and depth modalities on the performance. The RGB weight, ϕ_1, is tuned from 0 to 1, and the sum of both is kept as 1. The clique number in MCG is set to 6, which is detailed in Section.VII-D. The sparsity coefficients in Eq. 5, $\gamma_1 \& \gamma_2$, are respectively set to 0.03 and 0.01, which is explained in Section.VII-E. We implemented the proposed method on the MV-RED dataset. The PR-Curve and the quantitative evaluation are shown in Fig. 7.

From Fig. 7, MCG achieves optimal performance when $\phi_1 = 0.2 \& \phi_2 = 0.8$. The performance will degrade when ϕ_1 either increases or decreases. Moreover, the left sides of the peaks are usually higher than the right sides in terms

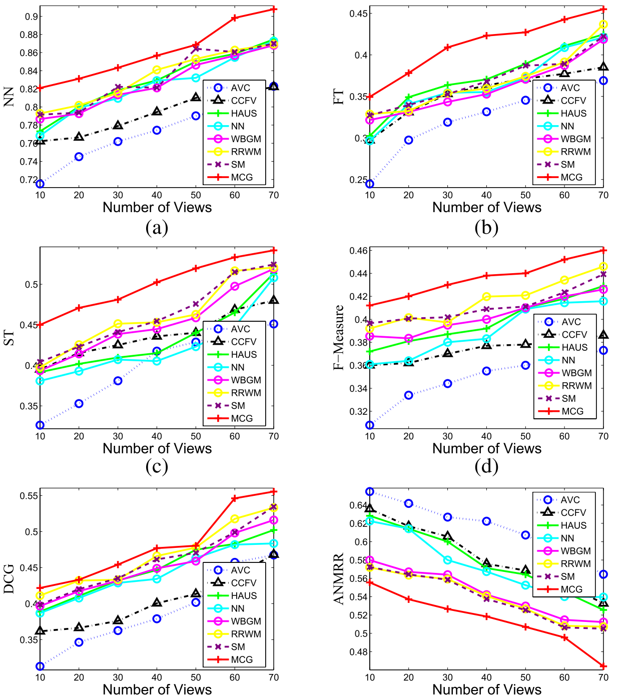


Fig. 6. Comparison by varying view numbers. (a) NN. (b) FT. (c) ST. (d) F-Measure. (e) DCG. (f) ANMRR.

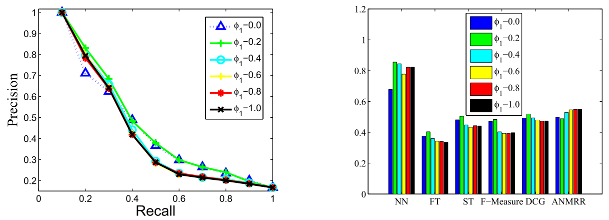


Fig. 7. Experiment by varying the weights of the RGB and depth modalities. (a) PR curve. (b) quantitative Comparison.

of FT, ST, F-Measure, and DCG, and vice versa for ANMRR. This observation denotes that the depth modality is more important than the RGB modality because the visual appearance of the RGB is usually more complicated than the depth appearance, and it is easier to extract more discriminative visual features in the depth image.

D. Effect of the Clique Number

The clique number in the MCG of individual models is varied from 2 to 10 with a step size of 2 to evaluate its effect on the performance. $\phi_1\&\phi_2$ are respectively set to 0.2&0.8 as stated in Section VII-C. The sparsity coefficients in Eq. 5, $\gamma_1\&\gamma_2$, are respectively set to 0.03 and 0.01, which is explained in Section.VII-E. The experiment is implemented on the MV-RED dataset. The PR-Curve and the quantitative evaluation are shown in Fig. 8.

From Fig. 8, the upper bound performance is obtained when the optimal clique number (6) is selected. The performance will degrade when the clique number either increases or decreases after the optimal clique number is achieved, which denotes that too few cliques in the under-division case and too many cliques in the over-division case will have negative influences on the performance. Theoretically, the performance will reach the lower bound when the clique number is assigned to either the smallest or largest allowed for a specific problem.

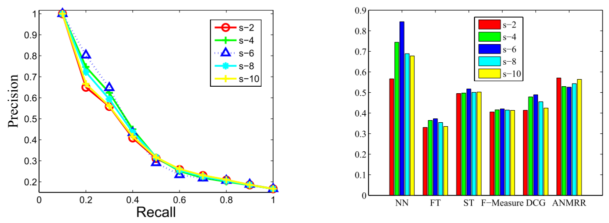


Fig. 8. Experiment by varying the clique numbers. (a) PR curve. (b) quantitative Comparison.

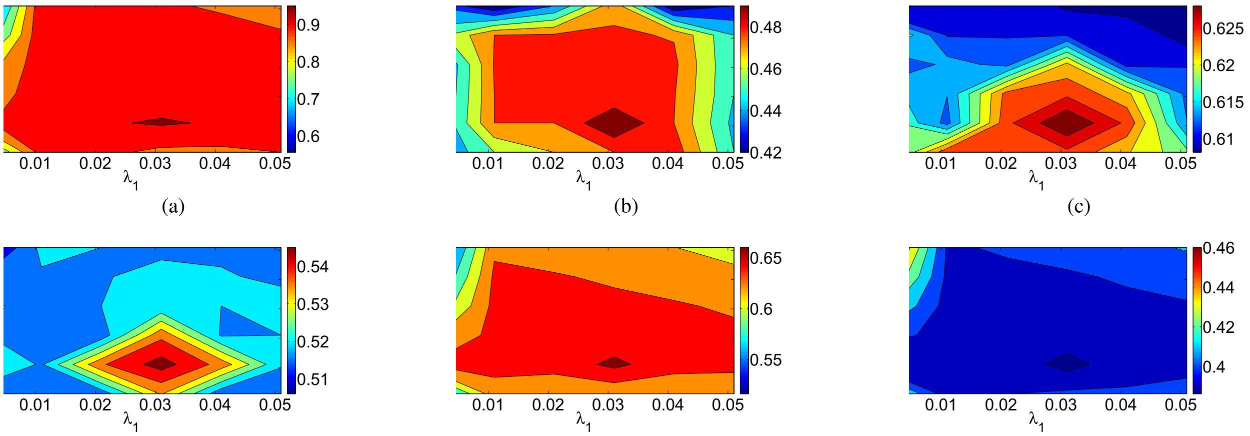
E. Effect of the Sparsity Coefficients

The sparsity coefficients ($\gamma_1\&\gamma_2$) are respectively tuned from 0.005 to 0.05 to evaluate their effect on the performance. As stated above, the clique number in MCG is set as 6, and the RGB and depth weights are respectively set to 0.2&0.8. The PR-Curve and the quantitative evaluation are shown in Fig. 9. The performances in terms of NN, FT, ST, F score, DCG and ANMRR show that the proposed method is able to achieve a steady performance with respective to the wide ranges of both sparsity coefficients. The best performance can be obtained when $\gamma_1\&\gamma_2$ are respectively set to around 0.03 and 0.01.

F. Comparison of the Computational Cost

The computational cost is measured by implementing all the methods with the Zernike moments on NTU-60. All the competing methods are run on a PC with Intel i3-2350 CPU (2.30GHz) and 6 GB of RAM. The speeds of AVC, CCFV, HAUS, NN, WBGM, SM, RRWM, and MCG are 8.1,6.8, 1.8, 1.8, 2.1, 4.5, 2.9, and 2.4 seconds, respectively. Several observations are explained below.

- AVC and CCFV belong to the statistical model-based method category and require the stage of model learning. Therefore, they usually have the highest computational complexity.
- HAUS and NN belong to the distance-based method and can easily realize similarity measures between pairwise models by directly calculating the distances between characteristic view images. Therefore, they usually have the lowest computational complexity.
- WBGM, SM, RRWM and MCG belong to the graph matching-based method. WBGM only considers the node-wise attribute without the edge-wise attribute, and its processing time is only a little longer than that of HAUS and NN. MCG and RRWM take advantage of both node-wise and edge-wise attributes, and consequently, their processing time is a little longer than that of WBGM. SM usually costs much more time than WBGM/MC/RRWM because it requires more iterations to reach convergence during optimization.



(a) (b) (c)

(d) (e) (f)

Fig. 9. Experiment on different sparsity coefficients. (a) NN. (b) FT. (c) ST. (d) F Score. (e) DCG. (f) ANMRR.

VIII. Conclusion and Future Work

In this paper, we propose a multi-modal clique graph matching (MCG) method for view-based 3D model retrieval. We leverage MCG to represent individual 3D models with multi-view and multi-modal information. The methods for MCG generation and matching are detailed in this paper. In particular, we propose the image set-based clique/edge-wise similarity measure to deal with the set-to-set distance measure, which is the core problem of MCG matching, and we suitably adapt the alternating minimization approach for optimal solution. The MCG-based 3D model retrieval is extensively validated on three popular single-modal datasets (NTU60/NTU216, PSB, ETH) and one novel multi-modal dataset (MV-TJU). We discuss the effects of the fusion of RGB and depth modalities, the clique numbers in MCG, and the sparsity coefficients on the performances. We compare the proposed method against the representative methods (AVC, CCFV, HAUS, NN, WBGM, SM, RRWM) with both single-modal and multi-modal data and further explore and compare their performances by varying the view numbers of each model. The experiments show that the proposed method can preserve local and global attributes of one graph well to achieve a robust similarity measure between 3D models and consistently outperform the competing methods even with a lower number of view images.

Because MCG matching is a general graph-matching method and can be widely utilized to solve the set-to-set matching problem, we would like to apply it in multiple research topics, such as feature point correspondence [43], multiple object tracking [63], and human action recognition [64], [65], in our future work.

References

[1] C. B. Akgül, B. Sankur, Y. Yemez, and F. Schmitt, "3D model retrieval using probability density-based shape descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 1117–1133, Jun. 2009.

[2] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 3, pp. 193–207, Jul. 2006.

[3] A. Del Bimbo and P. Pala, "Content-based retrieval of 3D models," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 20–43, 2006.

[4] S. Jayanti, Y. Kalyanaraman, N. Iyer, and K. Ramani, "Developing an engineering shape benchmark for CAD models," *Comput.-Aided Design*, vol. 38, no. 9, pp. 939–953, 2006.

[5] P. Papadakis, I. Pratikakis, T. Theoharis, and S. Perantonis, "PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval," *Int. J. Comput. Vis.*, vol. 89, nos. 2–3, pp. 177–192, Sep. 2010.

[6] 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.

[7] B. Leng and Z. Xiong, "ModelSeek: An effective 3D model retrieval system," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 935–962, 2011.

[8] K. Lai, L. Bo, X. Ren, and D. Fox, "RGB-D object recognition: Features, algorithms, and a large scale benchmark," in *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. London, U.K.: Springer, 2013, pp. 167–192.

[9] 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.

[10] B. Leng, S. Guo, X. Zhang, and Z. Xiong, "3D object retrieval with stacked local convolutional autoencoder," *Signal Process.*, vol. 112, pp. 119–128, Jul. 2015.

[11] 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.

[12] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. V. Vranic, "Feature-based similarity search in 3D object databases," *ACM Comput. Surv.*, vol. 37, no. 4, pp. 345–387, 2005.

[13] W. Nie, Q. Cao, A. Liu, and Y. Su, "Convolutional deep learning for 3D object retrieval," *Multimedia Syst.*, vol. 1, no. 1, pp. 1–8, 2015.

[14] S. Bu, Z. Liu, J. Han, J. Wu, and R. Ji, "Learning high-level feature by deep belief networks for 3-D model retrieval and recognition," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2154–2167, Dec. 2014.

[15] 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.

[16] 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, 2010.

[17] 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.

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

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

[20] *3D Shape Retrieval Contest 2015*, accessed on May 3, 2015. [Online]. Available: http://www.projects.science.uu.nl/shrec/

[21] 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.

[22] L. Yang and F. Albregtsen, “Fast and exact computation of Cartesian geometric moments using discrete Green’s theorem,” *Pattern Recognit.*, vol. 29, no. 7, pp. 1061–1073, 1996.

[23] K. Lu, Q. Wang, J. Xue, and W. Pan, “3D model retrieval and classification by semi-supervised learning with content-based similarity,” *Inf. Sci.*, vol. 281, pp. 703–713, Oct. 2014.

[24] A. D. Papoiu *et al.*, “Voxel-based morphometry and arterial spin labeling fMRI reveal neuropathic and neuroplastic features of brain processing of itch in end-stage renal disease,” *J. Neurophysiol.*, vol. 112, no. 7, pp. 1729–1738, 2014.

[25] P. Polewski, W. Yao, M. Heurich, P. Krzystek, and U. Stilla, “Detection of fallen trees in ALS point clouds of a temperate forest by combining point/primitive-level shape descriptors,” in *Proc. Gemeinsame Tagung*, 2014, pp. 1–12.

[26] E. Persoon and K.-S. Fu, “Shape discrimination using Fourier descriptors,” *IEEE Trans. Syst., Man, Cybern.*, vol. 7, no. 3, pp. 170–179, Mar. 1977.

[27] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, “3D shape histograms for similarity search and classification in spatial databases,” in *Advances in Spatial Databases*. Berlin, Germany: Springer, 1999, pp. 207–226.

[28] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, “Matching 3D models with shape distributions,” in *Proc. Int. Conf. Shape Modeling Appl. (SMI)*, 2001, pp. 154–166.

[29] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, “Topology matching for fully automatic similarity estimation of 3D shapes,” in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, 2001, pp. 203–212.

[30] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson, “Skeleton based shape matching and retrieval,” in *Proc. Shape Modeling Int.*, 2003, pp. 130–139.

[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, May 2014.

[32] B. Leng, C. Du, S. Guo, X. Zhang, and Z. Xiong, “A powerful 3D model classification mechanism based on fusing multi-graph,” *Neurocomputing*, vol. 168, pp. 761–769, Nov. 2015.

[33] 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 2015.

[34] B. Leng, X. Zhang, M. Yao, and Z. Xiong, “A 3D model recognition mechanism based on deep Boltzmann machines,” *Neurocomputing*, vol. 151, pp. 593–602, Mar. 2015.

[35] W. Nie, X. Li, A. Liu, and Y. Su, “3D object retrieval based on Spatial+LDA model,” *Multimedia Tools Appl.*, vol. 1, no. 1, pp. 1–14, 2015.

[36] 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.

[37] 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, 2007.

[38] X. Wang and W. Nie, “3D model retrieval with weighted locality-constrained group sparse coding,” *Neurocomputing*, vol. 151, pp. 620–625, Mar. 2015.

[39] A. Liu, K. Li, and T. Kanade, “A semi-Markov model for mitosis segmentation in time-lapse phase contrast microscopy image sequences of stem cell populations,” *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 359–369, Feb. 2012.

[40] F. Zhou and F. De la Torre, “Deformable graph matching,” in *Proc. CVPR*, 2013, pp. 2922–2929.

[41] F. Zhou and F. De la Torre, “Factorized graph matching,” in *Proc. CVPR*, 2012, pp. 127–134.

[42] R. Zass and A. Shashua, “Probabilistic graph and hypergraph matching,” in *Proc. CVPR*, 2008, pp. 1–8.

[43] M. Cho, J. Lee, and K. M. Lee, “Reweighted random walks for graph matching,” in *Proc. ECCV*, 2010, pp. 492–505.

[44] J. Lee, M. Cho, and K. M. Lee, “Hyper-graph matching via reweighted random walks,” in *Proc. CVPR*, 2011, pp. 1633–1640.

[45] O. Duchenne, F. Bach, I. S. Kweon, and J. Ponce, “A tensor-based algorithm for high-order graph matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2383–2395, Dec. 2011.

[46] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in *Proc. KDD Workshop Text Mining*, 2000, pp. 1–2.

[47] C. Schellewald and C. Schnörr, “Probabilistic subgraph matching based on convex relaxation,” in *Proc. 5th Int. Workshop Energy Minimization Methods Comput. Vis. Pattern Recognit.*, 2005, pp. 171–186.

[48] S. Gold and A. Rangarajan, “A graduated assignment algorithm for graph matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 4, pp. 377–388, Apr. 1996.

[49] M. Leordeanu and M. Hebert, “A spectral technique for correspondence problems using pairwise constraints,” in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Beijing, China, Oct. 2005, pp. 1482–1489.

[50] T. Cour, P. Srinivasan, and J. Shi, “Balanced graph matching,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 313–320.

[51] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?” in *Proc. ICCV*, 2011, pp. 471–478.

[52] P. Zhu, W. Zuo, L. Zhang, S. C.-K. Shiu, and D. Zhang, “Image set-based collaborative representation for face recognition,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1120–1132, Jul. 2014.

[53] A. Gunawardana and W. Byrne, “Convergence theorems for generalized alternating minimization procedures,” *J. Mach. Learn. Res.*, vol. 6, pp. 2049–2073, Dec. 2005.

[54] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[55] U. Niesen, D. Shah, and G. W. Wornell, “Adaptive alternating minimization algorithms,” *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1423–1429, Mar. 2009.

[56] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale l_1-regularized least squares,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[57] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, “The Princeton shape benchmark,” in *Proc. Shape Modeling Appl.*, 2004, pp. 167–178.

[58] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, “A mobile vision system for robust multi-person tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[59] K. Lai, L. Bo, X. Ren, and D. Fox, “Detection-based object labeling in 3D scenes,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1330–1337.

[60] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view RGB-D object dataset,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 1817–1824.

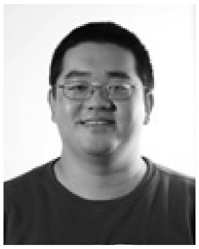
[61] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[62] *Description of Core Experiments for MPEG-7 Color/Texture Descriptors*, document ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819, 1999.

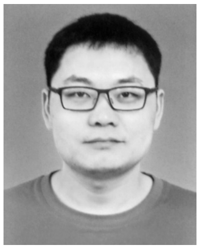
[63] W. Nie *et al.*, “Single/cross-camera multiple-person tracking by graph matching,” *Neurocomputing*, vol. 139, pp. 220–232, Sep. 2014.

[64] A. A. Liu, Y. T. Su, P. P. Jia, Z. Gao, T. Hao, and Z. X. Yang, “Multipe/single-view human action recognition via part-induced multitask structural learning,” *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

[65] 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.



An-An Liu (M’10) received the B.Eng. and Ph.D. degrees from Tianjin University, Tianjin, China. He was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, where he was with Prof. T. Kanade. He is currently an Associate Professor with the School of Electronic Information Engineering, Tianjin University. His research interests include computer vision and machine learning.



Wei-Zhi Nie received the Ph.D. degree from Tianjin University, Tianjin, China. He was a Visiting Scholar with the NExT Center, National University of Singapore, where he was with Prof. T.-S. Chua. He is currently an Assistant Professor with the School of Electronic Information Engineering, Tianjin University. His 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.

Yu-Ting Su received the B.Eng. and Ph.D. degrees from Tianjin University, Tianjin, China. He was a Visiting Scholar with Case Western Reserve University. He is currently a Professor with the School of Electronic Engineering, Tianjin University. His research interests include multimedia content analysis and security.