Learning Semantic Signatures for 3D Object Retrieval

Boqing Gong, *Student Member, IEEE*, Jianzhuang Liu, *Senior Member, IEEE*, Xiaogang Wang, *Member, IEEE*, and Xiaoou Tang, *Fellow, IEEE*

***Abstract*—In this paper, we propose two kinds of semantic signatures for 3D object retrieval (3DOR). Humans are capable of describing an object using attribute terms like "symmetric" and "flyable", or using its similarities to some known object classes. We convert such qualitative descriptions into attribute signature (AS) and reference set signature (RSS), respectively, and use them for 3DOR. We also show that AS and RSS can be understood as two different quantization methods of the same semantic space of human descriptions of objects. The advantages of the semantic signatures are threefold. First, they are much more compact than low-level shape features yet working with comparable retrieval accuracy. Therefore, the proposed semantic signatures require less storage space and computation cost in retrieval. Second, the high-level signatures are a good complement to low-level shape features. As a result, by incorporating the signatures we can improve the performance of state-of-the-art 3DOR methods by a large margin. To the best of our knowledge, we obtain the best results on two popular benchmarks. Third, the AS enables us to build a user-friendly interface, with which the user can trigger a search by simply clicking attribute bars instead of finding a 3D object as the query. This interface is of great significance in 3DOR considering the fact that while searching, the user usually does not have a 3D query at hand that is similar to his/her targeted objects in the database.**

***Index Terms*—3D object retrieval, semantic signature, attribute, reference set, user-friendly interface.**

I. Introduction

AN explosive increase in 3D data has been witnessed in recent years. As a result, 3D object retrieval (3DOR) becomes an active research topic attracting researchers from different areas, such as computer vision, CAD, and graphics. Some experimental search engines have been developed [1], [2], and there is an annual 3D shape retrieval contest (SHREC) to evaluate the effectiveness of different algorithms [3].

Initially, the problem is studied as retrieving generic 3D objects from a database potentially consisting of any kinds of objects, such as plants, animals, and buildings. An object is often represented by "mesh soup" which is not necessarily oriented or watertight. The dominant line of work to tackle this generic 3DOR problem is to use global features. We roughly classify them into four categories:

- statistics-based methods such as shape distributions [4] and enhanced shape functions [5],
- volume-based methods such as the ray-based approach with spherical harmonic representation [6] and the exponentially decaying Euclidean distance transform [7],
- view-based methods like the *LightField* descriptor (LFD) [8] and PANORAMA [9], and
- graph-based methods [10]–[13].

From the experimental results of these methods on different databases, a consensus seems achieved that view-based ones outperform the others under most conditions [9], [14]. We refer readers to [14] with a representative database, the Princeton shape benchmark (PSB), for the details.

Later, 3DOR is pursued with different emphases. Instead of generic 3D objects, McGill benchmark [13] focuses on objects with articulating parts. The partial retrieval problem is raised in [15]. 3D face recognition and identification [16] can been seen as a subproblem of the broad 3DOR. It is also worth noting that the SHREC contest makes a great effort to refine the problem into some sub-problems, such as "mesh soup" *vs.* watertight mesh and range data in terms of object representation, generic objects *vs.* architecture data in terms of database diversity, and self-created objects *vs.* Google warehouse in terms of data source collection. Obviously, global features are not able to work as well in all the cases as in retrieving generic objects. As a result, local feature based methods [17], [18] draw the attention of some researchers. A more comprehensive survey can be found in [19]. In this paper, we focus on generic 3DOR and compare global features.

From the study of these approaches, we find that most of them stem from the analysis of low-level 3D shape features, which have no high-level semantic interpretations. These low-level 3D shape features may not be able to capture users' search intention and distinguish objects of different classes. Take Fig. 1 as an example. The upper row shows top-ranked retrieval results with a 3D object "human_arms_out" as the query from the PSB database using the state-of-the-art 3D shape feature PANORAMA. It is not surprising that the planes are ranked top by the feature because they are indeed similar to the "human_arms_out" object in shape.

Manuscript received July 14, 2011; revised January 04, 2012; accepted July 15, 2012. Date of publication November 30, 2012; date of current version January 15, 2013. This work was supported in part by the Natural Science Foundation of China, under Grant 61070148, and in part by the Science, Industry, Trade, Information Technology Commission of Shenzhen Municipality, China, under Grant JC201005270378A. It was also supported through Introduced Innovative R&D Team of Guangdong Province "Robot and Intelligent Information Technology". The associate editor coordinating the review of this manuscript and approving it for publication was Nicu Sebe.

B. Gong is with the Department of Computer Science, University of Southern California, Los Angeles, CA 90095 USA (e-mail: boqinggo@usc.edu).

J. Liu is with the Media Laboratory, Huawei Technologies Co., Ltd., Shenzhen 518129, China, and also with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: liu.jianzhuang@huawei.com).

X. Wang is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: xgwang@ee.cuhk.edu.hk).

X. Tang is with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, and also with Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China (e-mail: xtang@ie.cuhk.edu.hk).

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/TMM.2012.2231059

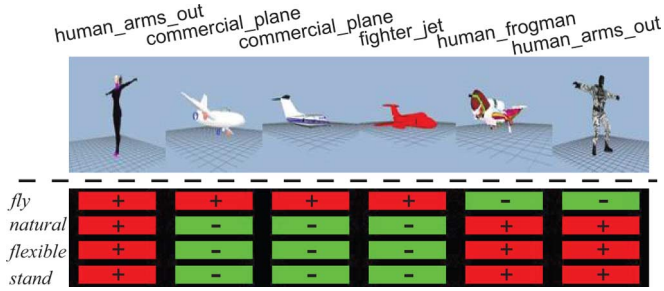


Fig. 1. Upper row: retrieved objects with the first object as the query using the latest 3D shape features PANORAMA. Lower row: responses of our *attribute detectors* to the objects where red blocks $(+)$ stand for positive outputs and green ones $(-)$ for negative. While the "*fly*" detector considers both the "human_arms_out" and the planes as positive, the "*natural*", "*flexible*", and "*stand*" detectors successfully distinguish the planes from the humans.

Human beings are capable of describing an object using attribute terms, such as "is the object flexible or rigid?" and "is the object symmetric or not", or using its similarities with some known object classes. These descriptions have semantic meanings and also some capability to distinguish objects. We propose to map 3D shape features to these descriptions, which are called semantic signatures and are used for 3DOR.

The first semantic signature is to emulate the human description of objects in terms of attributes. An attribute a_i describes a specific object characteristic (see Fig. 2 for example). It can be shared by different object classes, and most objects in the same class share the same attributes. We use a learning algorithm to train attribute detectors for automatically detecting the presence of attribute a_i on a 3D object x with a confidence measure $p(a_i|x)$. These measures are concatenated to form our *attribute signature (AS)*. Look at the lower row in Fig. 1 for instance. The attribute detectors detect whether an object is *natural*, *flexible*, able to *fly*, and able to *stand*. A red block $(+)$ stands for a positive output $(p(a_i|x)>0.5)$ of a detector and a green one $(-)$ stands for a negative output $(p(a_i|x)<0.5)$. While the "*fly*" detector considers both the "human_arms_out" and the planes as positives, all the others ("*natural*", "*flexible*", and "*stand*") successfully capture the differences of them.

The second is to describe an object by comparing it with known object classes. Suppose that there is a reference set with data classified into classes $\{c_j\}_{j=1}^n$. In the retrieval scenario, we cannot assume that each object in the database must belong to some c_j because of limited training data. Nonetheless, $\{c_j\}_{j=1}^n$ builds a context in which an object x can be represented through a similarity-based feature $(s_1(x), s_2(x), \ldots, s_n(x))^T$ called *reference set signature (RSS)*, where $s_j(x)$ stands for the similarity between x and the known class c_j. This similarity-based representation has been applied to image classification in [20] and [21]. However, in the retrieval scenario, we show that the direct application is not appropriate due to an interesting and exclusive problem regarding users' search intention (see Section III-B2). As a result, we extend the reference set to a hierarchical structure to satisfy different users' search intentions in terms of object classification granularity.

The proposed semantic signatures of 3D objects have the capability to distinguish objects with performances comparable to state-of-the-art low-level shape descriptors in 3DOR. To the best of our knowledge, this is the first work using semantic representations for 3DOR. The advantages are threefold:

1) The low dimensional AS (with 11 dimensions) and RSS (with 90 dimensions) are particular suitable for large scale and efficient 3DOR, since they require much less storage space and comparing time than low-level 3D shape features whose dimensions are usually between 100 and 10000 (see Section V).
2) They are complementary to low-level shape descriptors and can hence improve the performance of previous retrieval algorithms. We obtain the best retrieval results on two popular benchmarks (see Section V).
3) They enable us to build a user-friendly interface with which the user can search the database by simply choosing some attribute terms instead of finding or sketching a 3D object as the query.

In the next section we discuss some related work to this paper. The details of AS and RSS are developed in Sections III-A and Section III-B, respectively. After that we analyze the semantic essence of the two signatures in Section IV, and point out that AS and RSS can be understood as two different quantization methods of the same semantic space. Section V shows the experimental results of AS and RSS boosted 3DOR, and Section VI presents our 3DOR user interface built upon the signatures. The paper is concluded in Section VII.

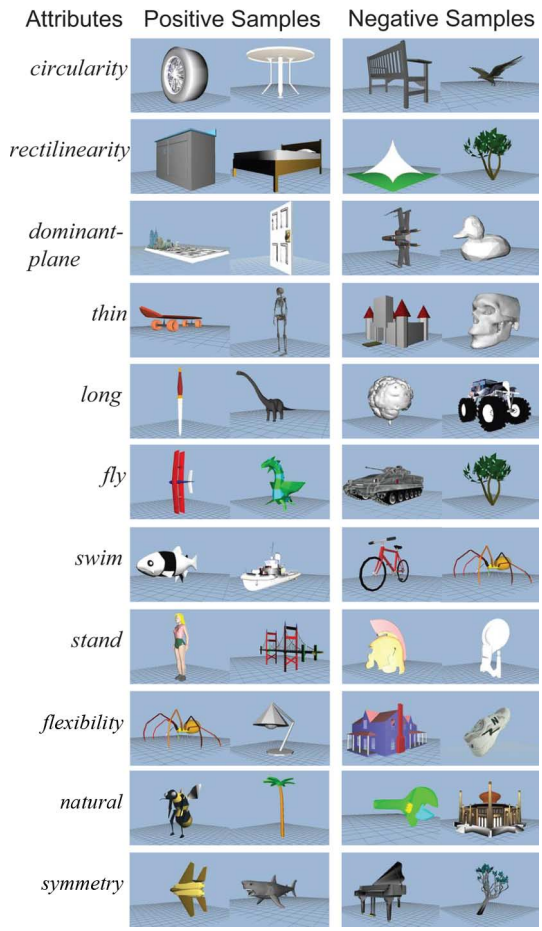


Fig. 2. Some positive and negative samples of the attributes.

II. Related Work

This section reviews some recent work related to using higher-level knowledge to do retrieval and object recognition.

3D Object Retrieval: Some previous work can be considered related to ours though they do not use the term of *attribute* explicitly. Kazhdan *et al.* propose to boost general 3D shape features (e.g., EDT [7]) by symmetry in [22]. Flexibility is defined to evaluate the ability to bend a part of an object and has been used in both 2D and 3D object retrieval [23], [24]. In [25], a measure of the rectilinearity of a 3D object is designed and added to other features. Our approach is in line with boosting general 3D shape features by descriptive attributes. However, 1) we replace the handcrafted computation methods to measuring attributes by a uniform learning stage, and 2) the attributes in previous work describe only the shape characteristic of an object, while we embed richer semantic information such as " *fly* " and " *natural* " into the attributes (see Section III-A for the details).

Semantic-oriented 3DOR using relevance feedback: Relevance feedback is an effective method to improve the performance of a retrieval system by involving user input in the retrieval loop. In [26], Leifman *et al.* pursue semantic-oriented 3DOR using latent discriminative analysis and biased discriminative analysis for relevance feedback. Our method differentiates from relevance feedback in that we embed semantic information into the signatures, instead of employing user input in the retrieval process.

Concept-Based Video Retrieval: Our approach has some connection to the concept-based video retrieval [27]. Automatic detection in videos of presence of semantic concepts, such as " *person* " or " *outdoor* ", are studied extensively. However, the concepts in video retrieval mainly refer to people, scenes, and events, and cannot be used in 3DOR. It is also worth noting that some image retrieval works [28], [29] share similar spirit.

2D Object Recognition: Attributes are also used in object recognition from images. On one hand, they serve as an intermediate layer between low-level image features and high-level object recognition tasks. They bridge a between-class transfer to detect unseen objects in [30] and [31]. Wang and Forsyth [32] develop a joint detector of visual attributes and object classes. On the other hand, inferring attributes becomes (part of) the core problem of recognition in [33] and [34]. The attributes in these papers are diversified, from color and texture to animal habitats, but cannot be applied to 3DOR since most 3D objects are pure shapes without any texture or color information. In [20] and [31], reference sets are used as a knowledge transfer model for 2D object recognition. They have been used in face recognition [35] and image retrieval [28]. Our work shows that reference sets are also very useful for 3DOR. In addition, we advance this idea by proposing a hierarchical reference set to achieve better performance.

III. Semantic Signatures

A. Learning Attribute Signature

In this section, we begin with the attributes used in our work, then present a learning method to measure these attributes, and finally define our first semantic signature AS.

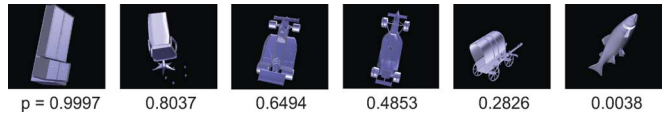


Fig. 3. Some measurements of *rectilinearity* on several 3D objects.

1) Attributes: The first question is: what attributes are *proper* to represent a 3D object? Lampert *et al.* introduce 85 attributes to describe animals for image-based object (animal) recognition [30], but few of them can be used here since 1) we deal with all kinds of objects including not only animals but also plants and man-made objects, and 2) we treat each 3D object as pure shape without textual or texture information. Therefore, the motivation is to choose attributes discriminative enough for broad object categories and closely related to the underlying 3D shape.

In previous works, some shape characteristics such as 1) *symmetry* [22], 2) *flexibility* [24], and 3) *rectilinearity* [25] have been proposed and measured by handcrafted methods. We include them in our attribute set but measure them by learned detectors (see Section III-A2). In line with the three shape characteristics, we also use 4) *circularity*, 5) *dominant-plane*, 6) *long*, and 7) *thin*. In addition, we emphasize that some higher-level semantic attributes can also be inferred from 3D shape. For example, an object being able to *swim* usually has a fish-like shape. The higher-level attributes used in our work are: 8) *swim*, 9) *fly*, 10) *stand with leg(s)*, and 11) *natural*. In total, we use eleven attributes in our work as shown in Fig. 2. Compared with the three existing ones in the 3DOR literature, this set of eleven attributes is much richer and it can be enriched further.

2) Attribute Detectors: When the number of attributes becomes large, it is not realistic to handcraft specific computation methods for each of them. Besides, the detection method for the higher-level semantic attributes, such as *swim* and *fly*, is difficult to be handcrafted. We use a uniform learning approach instead to obtain their measurements. The aim here is to train for attribute a a detector $\mathcal{C}^a$ with which we can measure the presence of a on a 3D object x with some confidence measurement (probability) $p(a|x)$. Since we use binary (presence or absence) attributes, any two-class classification algorithm with probability output can be used here. For each attribute a, the training data are 3D objects with labels of either 1 (presence of a) or -1 (absence of a). At the testing stage, given a 3D object x, the trained classifier $\mathcal{C}^a$ outputs $p(a|x)$ which is the attribute measurement.

In this paper, we use LIBSVM [36] with an RBF kernel to train a binary classifier (detector) for each attribute. The parameters in LIBSVM are determined through 5-fold cross validation, and the probability output is obtained by pairwise coupling [37]. LIBSVM requires vectorial representations of 3D objects as input. We use three kinds of complementary shape features in our algorithm: depth buffer descriptor (DBD) [6], wavelet transform of a 3D object's panorama [9] which is generated after the object is normalized by the continuous PCA [6], and the mutual absolute-angle distance (AAD) histogram [5]. They are concatenated to form a 1378-dimensional feature vector as the input to the detectors. Fig. 3 shows some outputs $p(a_{rec}|x)$ of detector $\mathcal{C}^a_{rec}$ for attribute *rectilinearity*.

3) Attribute Signature (AS): Given a 3D object x, we concatenate the outputs of all the attribute detectors to form AS,

$$AS(x) = (p(a_1|x), p(a_2|x), \ldots, p(a_{11}|x))^T, \quad (1)$$

which converts human descriptions of a 3D object with respect to the attributes into a vectorial representation. Next, we define a dissimilarity measure $d^a(x, y)$ between two 3D objects x and y based on AS,

$$d^a(x, y) = \sum_{i=1}^{11} \omega_i f(p(a_i|x), p(a_i|y)), \quad (2)$$

where $f(p(a_i|x), p(a_i|y))$ is the dissimilarity between 3D objects x and y in terms of attribute a_i, $1 \leq i \leq 11$, and ω_i is the weight to balance a_i's contribution.

We have tested various dissimilarity measures for $f(\cdot, \cdot)$ and found that the symmetric KL-divergence and χ^2-distance perform the best. In the rest of the paper, we use the former one, defined as

$$f(p_1, p_2) = d_{KL}(\vec{p_1} \| \vec{p_2}) + d_{KL}(\vec{p_2} \| \vec{p_1}), \quad (3)$$

where $\vec{p_1} = (p_1, 1-p_1)^T$, $\vec{p_2} = (p_2, 1-p_2)^T$, and d_{KL} is the KL-divergence [38].

B. Learning Reference Set Signature

In this section, we define our second semantic signature RSS based on a reference set, and then extend it to a hierarchical structure.

1) Reference Set Signature (RSS): Besides the attributes introduced in the previous section, human beings often describe a new object by comparing it with known object classes. We convert this kind of description into RSS. Let $S = \{c_j\}_{j=1}^n$ denote the reference set of n available classes. The RSS of a 3D object x is then defined as

$$RSS(x) = (s_1(x), s_2(x), \ldots, s_n(x))^T, \quad (4)$$

where $s_j(x)$, $1 \leq j \leq n$, stand for the similarities between the object x and the known classes $\{c_j\}_{j=1}^n$.

In this paper, we also compute $s_j(x)$ using a LIBSVM classifier trained with the same three shape features mentioned in the last section. The difference is that only one multi-class classifier $\mathcal{C}^r$ is trained based on $\{c_j\}_{j=1}^n$, whose outputs denote the similarities between x and the classes $\{c_j\}_{j=1}^n$ (or the probabilities of x belonging to the classes). We have also tested Torresani *et al.*'s "classemes" [21] which are binary classifiers trained for each class individually, and found that the multi-class classifier outperforms the batch of classemes in our experiments. It is probably due to the fact that some of our training classes are very small (with less than 5 objects) and hence lead to bad classemes.

Suppose that a 3D object x in the database belongs to class c_{j_x}. Note that the class c_{j_x} can either exist in the reference set (case 1) or be a novel class (case 2). Fig. 4 shows the RSSs of 4 biplanes corresponding to case 1, and the RSSs of 4 ants and 4 churches corresponding to case 2.

To compare two objects x and y by RSS, the following distance metric is used:

$$d^r(x, y) = f(RSS(x), RSS(y)), \quad (5)$$

where $f(\cdot)$ is the symmetric KL-divergence (see (3)).

2) Hierarchical Reference Set: Intuitively, any classified 3D object data set can be used as the reference set. However, there is an interesting issue regarding the granularity of object classes. For example, should we put helicopters and fighter jets into the same class named *plane* or two different classes? Suppose that we have two reference sets $S1 = \{\text{helicopter, fighter jet, other classes}\}$ and $S2 = \{\text{plane, other classes}\}$, and a database $D = \{\ldots, x_h, x_f, \ldots\}$ where x_h is a helicopter and x_f a fighter jet, and that the user submits a fighter jet to query the database. If an ideal classifier $\mathcal{C}^r$ is assumed to compute the RSSs, the system adopting $S1$ as the reference set will rank x_f top but may rank x_h at a low position mixed with other non-plane objects. The system adopting $S2$ will consider x_h and x_f equally similar to the query and may rank x_f after x_h. This example tells us that reference sets with different granularities of object classes may lead to different retrieval results.

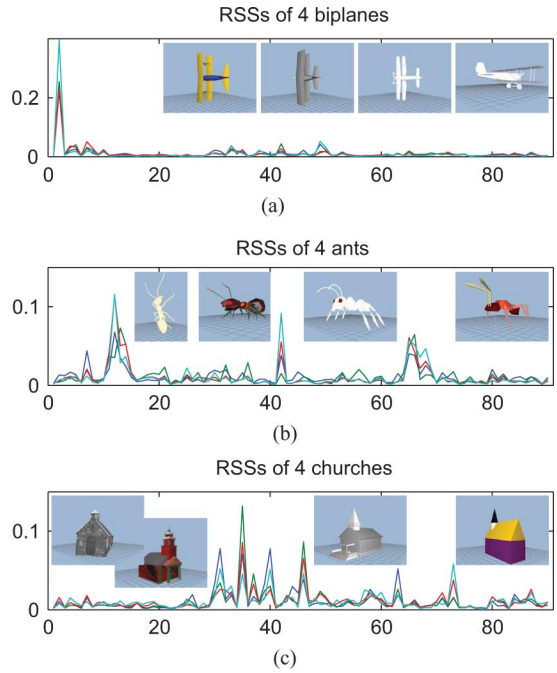


Fig. 4. (a) RSSs of 4 biplanes whose class is in the reference set. (b) RSSs of 4 ants. (c) RSSs of 4 churches. The ant class and church class are not in the reference set. The horizontal axis denotes the reference classes.

We ought to choose a granularity which properly reflects users' search intention: $S1$ satisfies the user if he/she wants to search for fighter jets, and $S2$ is proper if the user wants roughly all kinds of planes. However, it is not easy to infer the "correct" granularity (perhaps the "correct" granularity does not exist at all). Our solution is to use a hierarchical structure that gives different levels of classification granularity simultaneously:

$$S_{\mathcal{H}} = \{\{c_{1j}\}_{j=1}^{n_1}, \{c_{2j}\}_{j=1}^{n_2}, \ldots, \{c_{Lj}\}_{j=1}^{n_L}\},$$

where $\{c_{lj}\}_{j=1}^{n_l}$ ($n_l < n_{l-1}$, $1 < l \leq L$) denotes the n_l classes of the reference set on the lth level, and are generated by merging some classes on its preceding level for $l > 1$. We can hence obtain a hierarchical RSS set $RSS_{\mathcal{H}} = \{RSS_l(x)\}_{l=1}^L$ for a 3D object x with each $RSS_l(x)$ computed from $\{c_{lj}\}_{j=1}^{n_l}$. The dissimilarity between two objects x and y is now defined as

$$d^r_{\mathcal{H}}(x, y) = \sum_{l=1}^{L} \frac{n_l}{n_1} \cdot d^r_l(x, y), \quad (6)$$

where $d_l^r(x, y)$ is computed with (5). The largest weight $(= 1)$ is given to the finest class level.

Now let us turn back to the example at the beginning of this section. With (6), if the query is a fighter jet, the system can always rank x_f above x_h and x_h above non-plane objects. This ranking list satisfies the user whether he/she wants exactly fighter jets or roughly planes.

C. Building the Training Set

In order to learn AS and RSS, we use the standard training set of PSB (PSB-train) [14] which includes 907 3D objects. It has four-level classifications with 90, 42, 7, and 2 classes according to the classification scheme defined in [14], respectively. It has no attribute labels and we tag the labels to the training objects. For each attribute a, we label its presence or absence on the objects. If we are not sure about how to assign a label to some objects, these objects are not used to train the detector of a. Fig. 2 shows some of the labeling results of the attributes. This procedure is conducted independently by three students, and the final attribute labels are determined by majority voting.

D. Comparison of AS and RSS

Both AS and RSS represent a 3D object with respect to some semantic and descriptive characteristics: AS maps the qualitative attributes to a quantitative signature; RSS quantizes the similarities of a 3D object to a set of known classes. It is therefore interesting to examine the relationship between them. Given a fixed set of attributes or known classes, the computation costs of both AS and RSS are linear with the size of the set. Nonetheless, should the underlying attribute or known class set changes, AS can be extended more easily than RSS. From (1), AS can be expanded or shrunk by adding or removing some attribute classifiers. On the other hand, we have to re-train our model to obtain RSS.

In addition to the computation cost, we compare AS and RSS further from the aspect of their semantic essences in the next section, demonstrating that they are visually two different quantization methods of the same semantic space.

IV. Quantizing Semantic Space by AS/RSS

After describing the details of AS and RSS, we discuss their semantic essence in this section.

A. AS vs. Basis

Suppose that there exists a semantic space $\mathbb{S}$ of human descriptions of objects (not necessarily 3D objects). Since human beings describe an object using different attributes, the attributes can be intuitively seen as the basis of the space $\mathbb{S}$, and their binary observations (i.e., presence or absence of the attributes on 3D objects) quantize $\mathbb{S}$ into different quadrants (see Fig. 5(a)). Therefore, each object is mapped by AS to a point in $\mathbb{S}$. With no *a priori* knowledge about the dimension of $\mathbb{S}$, the best way to model it would be to find as many useful attributes as possible.

Our experiments also confirm that the more attributes, the better results. We evaluate the effectiveness of individual attributes by the leave-one-out strategy. First, we run an experiment using all the 11 attributes, and then examine each attribute by leaving it out and conducting retrieval using the 10 remaining

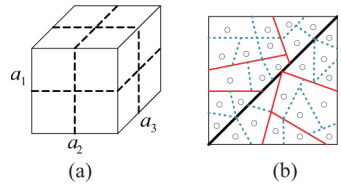


Fig. 5. Two methods to quantize the semantic space $\mathbb{S}$ of human descriptions of objects, (a) attributes and (b) a hierarchical reference set.

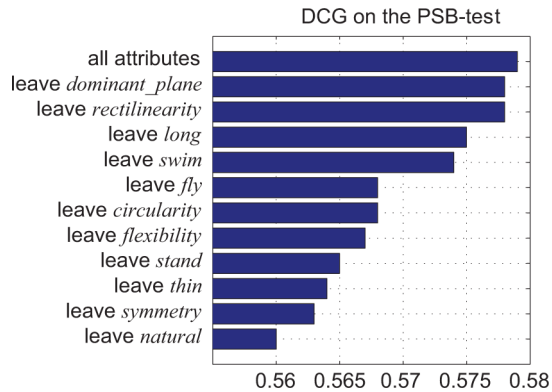


Fig. 6. Comparison of each attribute's effectiveness by the leave-one-out strategy evaluated on the Base level of the PSB-test.

ones. The experiments are carried out on the testing set of PSB (PSB-test) [14]. Each object in PSB-test is taken as the query once and the results are evaluated by discounted cumulative gain (DCG) [14]. Fig. 6 shows the DCGs of the 12 experiments, from which we can see different DCG reductions when one attribute is left out. The larger the DCG decreases, the more effective the corresponding attribute is. This figure indicates that all the attributes are useful and the last seven contribute more.

B. RSS vs. Vector Quantization

Similar to AS, RSS can also be understood from the viewpoint of the semantic space $\mathbb{S}$ of human descriptions of objects. Fig. 5(b) shows the reference classes in $\mathbb{S}$ by small circles. These classes quantize $\mathbb{S}$ in a similar way to vector quantization (VQ). VQ divides a large set of points (vectors) into groups and represents each group by its centroid. Similarly, each reference class c_{lj} serves as a representative centroid in the semantic space $\mathbb{S}$, and the extended hierarchical reference set $S_{\mathcal{H}}$ provides a coarse-to-fine quantization of $\mathbb{S}$. Accordingly, the learned $RSS_{\mathcal{H}}$ embeds multi-scale information and is hence more versatile in representing 3D objects than the RSS_1 learned from a single-level reference set. We present experiments to verify the advantage of $RSS_{\mathcal{H}}$ below.

We learn the single-level RSS_1 and the hierarchial $RSS_{\mathcal{H}}$ using the PSB-train with the Base level and the given 4-level classifications, respectively, and compare their performance on PSB-test. PSB-test contains 907 objects classified into 92 classes each with 4 to 50 objects on the "Base" level, 42 classes on the "Coarse1" level, 7 classes on the "Coarse2" level, and 2 classes (natural and man-made objects) on the "Coarse3" level. The two kinds of RSSs are compared on each of the four classification levels of the PSB-test. Fig. 7 shows the DCG values of the results. It is clear to see that the hierarchical RSS performs better in all the cases.

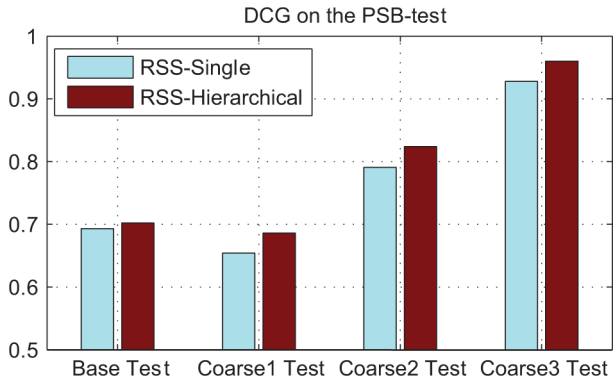


Fig. 7. Comparison of hierarchical and single-level reference set based RSSs.

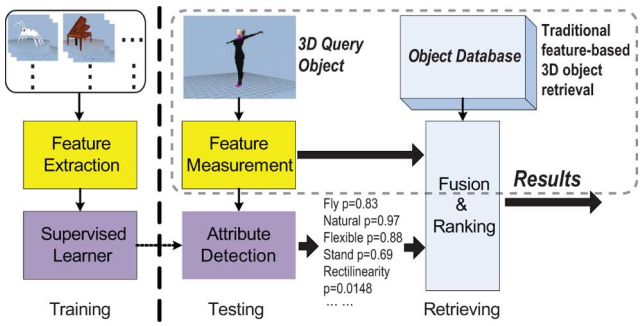


Fig. 8. Framework of AS boosted 3DOR.

V. Boosting 3D Object Retrieval

Recall the example shown in Fig. 1. The semantic signatures and low-level shape features are mutually complementary since they represent 3D objects from different aspects. We can therefore use AS and RSS to boost existing 3DOR algorithms.

We use a linear combination of normalized dissimilarities computed from the semantic signatures and 3D shape features to be the final dissimilarity between the query x and an object y in the database:

$$d(x,y) = \frac{\alpha d^f(x,y)}{max_{(u,v)}\{d^f(u,v)\}} + \frac{(1-\alpha)d^s(x,y)}{\max_{(u,v)}\{d^s(u,v)\}}, \quad (7)$$

where $d^f(\cdot,\cdot)$ is the dissimilarity computed from some 3D shape feature, $d^s(\cdot,\cdot)$ is calculated by either AS $(d^a(\cdot,\cdot))$ or by RSS $(d^r_{\mathcal{H}}(\cdot,\cdot))$, (u,v) denotes a pair of objects in the database, and α is the balance weight. The normalization numbers (denominators) do not cause extra retrieval time because they are computed off-line. Some other combination methods [39] are applicable too. Nonetheless, by the simple linear combination, significant improvements have already been achieved as shown in the experiment section.

Fig. 8 shows the framework of our approach with AS for example. At the training stage, we train a detector for each attribute. The detector is used at the testing stage to generate a 3D object's AS which is then combined with existing low-level 3D shape features to do retrieval. RSS is utilized in the same framework with the attribute detection replaced by reference set comparison.

A. Experiments

Two popular benchmarks are used to test how AS and RSS can boost 3DOR algorithms that are based on shape features. One is the PSB-test mentioned in Section IV, and the other contains watertight models in SHREC 2007 (WM-SHREC) [40] with 400 objects divided equally into 20 classes. Note that the training data (PSB-train) for AS and RSS are not in the test databases. As mentioned in Section I, mainly global features are used to do generic 3DOR. Thus, the following state-of-the-art global shape features are compared in the experiments:

- D2 [4], a representative 3D shape descriptor of statistics-based methods,
- view-based *LightField* descriptor (LFD) [8],
- DSR [6], a hybrid descriptor of view-based and volume-based techniques, and
- view-based Panorama (PAN) descriptor [9], which achieves the best performance in the comparative experiments in [9].

The results are evaluated by the standard criteria used in the annual SHREC contest: Nearest Neighbor (1-NN), First Tier (FT), Second Tier (ST), E-Measure (EM), and Discounted Cumulative Gain (DCG). The higher their values are, the better the performance of the retrieval algorithm. We refer the reader to [14] for the details of the criteria.

The rowwise parts 1 and 8 in Table I show the results using AS and RSS on the two databases respectively. The DCG values of the low-level shape features (D2, LFD, DSR, and PAN) on PSB-test are 0.462, 0.643, 0.663, and 0.733, respectively, while AS's DCG value is 0.579 and RSS's is 0.702. We can see that the high-level semantic signatures alone have comparable performance to the low-level shape features for 3DOR. The same conclusion can be drawn on the WM-SHREC database. The dimensions of the features and the signatures are shown in the parentheses in Table I. We can observe that the signatures alone achieve similar retrieval results with much fewer dimensions compared to the low-level shape features. This makes the semantic signatures particularly suitable for efficient large scale 3DOR. Another observation is that AS performs not as well as RSS overall. Considering that there are 90 classes on the base level for training RSS while only 11 attributes for AS, it is likely that the inferiority of AS is caused by the insufficient attributes. It is part of our future work to explore more useful attributes for 3DOR.

In each of parts 2–5 and 9–12, the performance of a 3D shape feature is presented in the first row, and its combinations (see (7)) with AS in the second row and with RSS in the third row. The underlined numbers denote the best results in their corresponding parts for different criteria. We can see that the semantic signatures significantly boost the performance of the low-level shape features. One may argue that the improvement is due to feature combination, instead of the semantic information. However, if we combine the best shape feature PAN with other shape descriptors to test if they can also have similar significant improvements, only slight gains are observed: the combination of PAN and LFD increases the DCG of PAN by about 0.015, and all the other combinations have a gain of less than 0.005. The fact that AS and RSS improve PAN much more than

TABLE I
RETRIEVAL RESULTS BY AS, RSS, THE FOUR SHAPE FEATURES, AND THEIR COMBINATIONS

PSB-test (evaluated on the Base level with 92 classes)						
Part	Method	1-NN	FT	ST	EM	DCG
1	AS (11D)	0.475	0.307	0.455	0.263	0.579
	RSS (90D)	0.665	0.465	0.596	0.357	0.702
	D2 (512D)	0.364	0.187	0.272	0.161	0.462
2	D2+AS	0.564	0.353	0.496	0.296	0.621
	D2+RSS	0.680	0.462	0.601	0.358	0.701
	LFD (≈1000D)	0.655	0.379	0.486	0.279	0.643
3	LFD+AS	0.698	0.449	0.575	0.331	0.699
	LFD+RSS	0.741	0.520	0.648	0.376	0.744
	DSR (472D)	0.658	0.404	0.513	0.294	0.663
4	DSR+AS	0.699	0.466	0.601	0.344	0.718
	DSR+RSS	0.724	0.512	0.643	0.370	0.741
	PAN (>10000D)	0.764	0.490	0.615	0.344	0.733
5	PAN+AS	0.776	0.516	0.655	0.385	0.756
	PAN+RSS	0.764	0.540	0.677	0.385	0.763
6	PAN+Both	**0.770**	**0.543**	**0.684**	**0.385**	**0.767**
	PAN_{LRF}	0.761	0.536	0.664	0.368	0.756
7	$(\text{PAN+Both})_{\text{LRF}}$	**0.783**	**0.585**	**0.717**	**0.400**	**0.786**
	$(\text{PAN+Both})_{\text{LRF+R}}$	**0.786**	**0.593**	**0.722**	**0.400**	**0.789**
WM-SHREC						
Part	Method	1-NN	FT	ST	EM	DCG
8	AS (11D)	0.735	0.502	0.684	0.478	0.782
	RSS (90D)	0.892	0.607	0.754	0.535	0.849
	D2 (512D)	0.735	0.381	0.535	0.366	0.713
9	D2+AS	0.855	0.572	0.73	0.518	0.831
	D2+RSS	0.91	0.635	0.775	0.555	0.864
	LFD (≈1000D)	0.92	0.521	0.657	0.465	0.824
10	LFD+AS	0.913	0.634	0.771	0.551	0.872
	LFD+RSS	0.938	0.661	0.785	0.566	0.881
	DSR (472D)	0.918	0.535	0.671	0.475	0.824
11	DSR+AS	0.935	0.628	0.764	0.547	0.869
	DSR+RSS	0.942	0.651	0.770	0.555	0.876
	PAN (>10000D)	0.967	0.672	0.784	0.569	0.896
12	PAN+AS	0.965	0.704	0.818	0.595	0.908
	PAN+RSS	0.960	0.707	0.821	0.594	0.907
13	PAN+Both	0.965	0.715	0.825	0.597	0.910
	PAN_{LRF}	0.960	0.732	0.831	0.608	0.916
14	$(\text{PAN+Both})_{\text{LRF}}$	**0.965**	**0.771**	**0.871**	**0.635**	**0.928**
	$(\text{PAN+Both})_{\text{LRF+R}}$	**0.962**	**0.793**	**0.888**	**0.647**	**0.935**

other shape features is reasonable, because the semantic signatures are high-level object descriptions that are important complements to the shape features, while all these shape features themselves represent 3D objects based on low-level shape details. Parts 6 and 13 in Table I are the results of the combination of PAN with both AS and RSS $(\text{PAN} + \text{Both})$.

In [9], the authors use a scheme called local relevance feedback (LRF) to improve PAN's performance. In parts 7 and 14 of Table I, we also give the results obtained by PAN with LRF, denoted by PAN_{LRF}, and the results by PAN+Both with LRF, denoted by $(\text{PAN} + \text{Both})_{\text{LRF}}$. This scheme works well in our experiments. In fact, we can further improve the results by a re-ranking scheme that re-ranks retrieved objects with the assumption that objects of the same class have the same manifold structure [41], [42]. The third rows of parts 7 and 14 show the results obtained by $\text{PAN} + \text{Both}$ with LRF and re-ranking, denoted by $(\text{PAN} + \text{Both})_{\text{LRF+R}}$. We can see more gains are achieved.

The bold numbers in Table I denote the results obtained by PAN+Both, $(\text{PAN} + \text{Both})_{\text{LRF}}$, and $(\text{PAN} + \text{Both})_{\text{LRF+R}}$ on PSB-test, and the results obtained by $(\text{PAN} + \text{Both})_{\text{LRF}}$ and $(\text{PAN} + \text{Both})_{\text{LRF+R}}$ on WM-SHREC. To the best of our knowledge, they are better than the best results in the previous works for all the criteria (except 1-NN) in Table I. Another point we want to emphasize is that the AS and RSS learnt from PSB-train work well not only on PSB-test but also on WM-SHREC. The data structures in the two databases are quite different. PSB contains objects represented by "mesh soup" that may have outlier faces and holes, while objects in WM-SHREC are represented by watertight meshes.

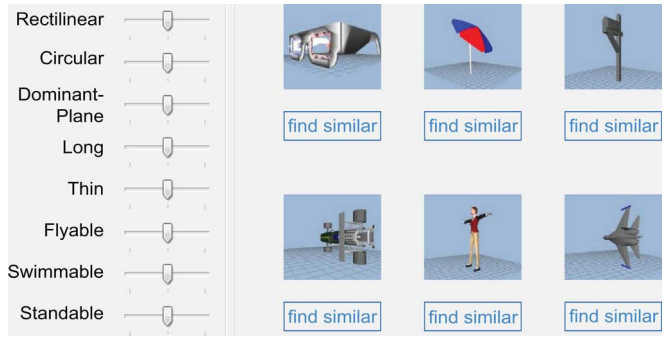


Fig. 9. A 3D object retrieval interface. On the left part of the interface, the user clicks the attribute bars to browse the database. On the right, the user can click the "find similar" button under an object to search for similar objects in the database.

VI. A USER-FRIENDLY INTERFACE FOR 3DOR

One key issue for a user-friendly search engine is how to deliver user search intention to the system. This problem is even more important in the 3DOR scenario due to two facts: 1) Text-based query can only search for a small part of existing 3D objects because currently most 3D objects are pure shapes without textual descriptions or tags. 2) Content-based retrieval, i.e., using a 3D object as the query, causes great trouble for the user who does not have a 3D query that is similar to his/her targeted objects in the database. To circumvent this problem, some systems allow the user to form a query by sketching the object's silhouette or skeleton [2], [43], [44]. The main drawback of these systems is that a 2D sketch has only a small part of shape information of a 3D object, leading to lower retrieval accuracy.

In this section, we develop an alternative solution toward building a user-friendly 3DOR interface. In our system, the user searches for a targeted 3D object *in mind* by simply clicking attribute bars (see Fig. 9), without requiring the user to have a similar 3D query object *at hand*. This is a natural way of delivering search intentions to the system for common users. With the input from the user, we obtain the attribute signature of the targeted 3D object x, and can hence carry out retrieval using (2). In addition, the user can select an object displayed currently on the interface as the query. The whole search procedure contains two steps:

1) Browse the database by clicking the attribute bars on the left part of the interface. Each attribute takes three possible values, 0 (no this attribute), 1 (with this attribute), and 0.5 (not sure). With the targeted object x in mind, the user inputs its attribute values to the system with the bars naturally. Our system then ranks objects in the database by their similarities to x computed using (2).

2) Search the database by clicking "find similar" buttons. If the user finds some object $\tilde{x}$ similar to x on the interface, he/she can click the "find similar" button under object $\tilde{x}$. In this situation, the system ranks objects by their similarities to $\tilde{x}$ using (7).

The above two steps can be carried out multiple times to search for more objects.

A. User Study

We conduct user study on this interface. Ten users are asked to find out some specific objects from the database, and the numbers of their mouse clicks are recorded during the search. When the targeted object is displayed on the interface where 20 top ranked objects are shown, the search is done successfully. Take PSB-test as the test database which has 907 objects. We find that our interface is both effective and efficient: about 600 objects can be found out within 5 mouse clicks, and more than 700 can be found out within 11 clicks. The failures caused by the other objects are mainly due to two reasons: 1) the users cannot well judge the presence of attributes on an object, and 2) there are too few objects in the database which are similar to the targeted object—there are 27 out of 92 classes in PSB-test each with less than 5 objects. A demo of the interface can be found from our supplementary material.

B. Discussion

To build a practical 3DOR system, it is important to improve user satisfaction of the system, which consists of several factors such as query formulation, retrieval accuracy, and real-time response. Since common users are often lack of 3D objects, our approach of query formulation is of great convenience for them: the user can "describe" the query by simply clicking the attribute bars. In terms of retrieval accuracy, our signatures improve the performance of traditional low-level 3D shape descriptors. We can see that the semantic signatures are versatile and can potentially lead to better user satisfaction than traditional methods.

VII. Conclusions and Future Work

We have proposed two semantic signatures, attribute signature (AS) and reference set signature (RSS), to boost 3D object retrieval methods that base on shape features. Both signatures map low-level shape features to high-level semantic descriptions of objects, and are applicable to new object classes that are not in the training set. To learn AS, we have systematically studied a set of eleven attributes that are suitable to describe 3D objects, and developed a uniform learning algorithm to train the attribute detectors. To obtain effective RSS, a hierarchical reference set has been proposed to meet the search intentions of different users in terms of classification granularity. The extensive experiments indicate that AS and RSS improve previous methods significantly with the best retrieval results obtained, revealing that the semantic signtures and the low-level shape features are mutually complementary. Our experiments also show that the combination of different low-level shape features only gives limited gains. The proposed semantic signatures are also used to develop a novel user-friendly 3DOR interface.

In the future work, we will study more attributes to describe 3D objects. The knowledge mining methods in [31] may be helpful to find suitable attributes. We plan to extend the current binary attribute labels to continuous labels. The selection of a proper reference set is also worth exploring.

References

[1] 3D Search Engine. [Online]. Available: http://3d-search.iti.gr/3Dsearch.

[2] Princeton 3D Model Search Engine. [Online]. Available: http://shape.cs.princeton.edu/search.html.

[3] 3D Shape Retrieval Contest (Shrec). [Online]. Available: http://www.aimatshape.net/event/SHREC.

[4] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Trans. Graph.*, vol. 21, no. 4, pp. 807–832, 2002.

[5] R. Ohbuchi, T. Minamitani, and T. Takei, "Shape-similarity search of 3D models by using enhanced shape functions," *Int. J. Comput. Appl. Technol.*, vol. 23, no. 2, pp. 70–85, 2005.

[6] D. V. Vranic, "3D model retrieval," Ph.D. dissertation, Univ. Leipzig, Leipzig, Germany, 2004.

[7] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Proc. Eurographics Symp. Geometry Processing*, 2003.

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

[9] 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. Vision*, vol. 89, no. 2, pp. 177–192, 2010.

[10] T. Tung and F. Schmitt, "The augmented multiresolution reeb graph approach for content-based retrieval of 3D shapes," *Int. J. Shape Model.*, vol. 11, no. 1, pp. 91–120, 2005.

[11] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. Strintzis, "Combining topological and geometrical features for global and partial 3-D shape retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 819–831, 2008.

[12] A. Agathos, I. Pratikakis, P. Papadakis, S. Perantonis, P. Azariadis, and N. Sapidis, "3D articulated object retrieval using a graph-based representation," *Visual Comput.*, vol. 26, no. 10, pp. 1301–1319, 2010.

[13] J. Zhang, K. Siddiqi, D. Macrini, A. Shokoufandeh, and S. Dickinson, "Retrieving articulated 3D models using medial surfaces and their graph spectra," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition Workshop of Energy Minimization Methods*, 2005.

[14] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton shape benchmark," *Proc. Shape Modeling Int.*, 2004.

[15] Y. Liu, H. Zha, and H. Qin, "Shape topics: A compact representation and new algorithms for 3D partial shape retrieval," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2006.

[16] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1858–1870, 2010.

[17] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. v. Gool, "Hough transform and 3D surf for robust three dimensional classification," in *Proc. Eur. Conf. Comput. Vision*, 2010.

[18] M. Ovsjanikov, A. Bronstein, M. Bronstein, and L. Guibas, "Shape Google: A computer vision approach to invariant shape retrieval," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition Workshop of Non-Rigid Shape Anal. and Deformable Image Alignment*, 2009.

[19] J. Tangelder and R. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools Appl.*, vol. 39, no. 3, pp. 441–471, 2008.

[20] E. Bart and S. Ullman, "Single-example learning of novel classes using representation by similarity," in *Proc. British Machine Vision Conf.*, 2005.

[21] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vision*, 2010.

[22] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Symmetry descriptors and 3D shape matching," in *Proc. Eurographics Symp. Geometry Processing*, 2004.

[23] C. Xu, J. Liu, and X. Tang, "2D shape matching by contour flexibility," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 180–186, 2008.

[24] B. Gong, C. Xu, J. Liu, and X. Tang, "Boosting 3D object retrieval by object flexibility," in *Proc. ACM Int. Conf. Multimedia*, 2009.

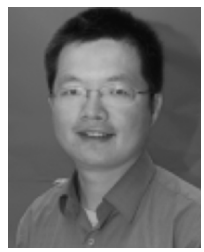
[25] Z. Lian, P. Rosin, and X. Sun, "Rectilinearity of 3D meshes," *Int. J. Comput. Vision*, vol. 89, no. 2, pp. 130–151, 2010.

[26] G. Leifman, R. Meir, and A. Tal, "Semantic-oriented 3D shape retrieval using relevance feedback," *Visual Comput.*, vol. 21, no. 8, pp. 865–875, 2005.

[27] C. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retriev.*, vol. 2, no. 4, pp. 215–322, 2009.
[28] X. Wang, K. Liu, and X. Tang, "Query-specific visual semantic spaces for web image re-ranking," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2011.
[29] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang, "Intentsearch: Capturing user intention for one-click internet image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1342–1353, 2012.
[30] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2009.
[31] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where-and why? semantic relatedness for knowledge transfer," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2010.
[32] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. Int. Conf. Comput. Vision*, 2009.
[33] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2009.
[34] A. Endres and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2010.
[35] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2011.
[36] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm.
[37] T. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.
[38] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, pp. 79–86, 1951.
[39] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. Int. Conf. Comput. Vision*, 2009.
[40] R. Veltkamp and F. t. Haar, Shrec2007: 3D Shape Retrieval Contest, Dept. Inf. Comput. Sci., Utrecht Univ., 2007.
[41] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," *Proc. Neural Inf. Process. Syst.*, 2003.
[42] R. Ohbuchi and T. Shimizu, "Ranking on semantic manifold for shape-based 3D model retrieval," *Proc. ACM Multimedia Inf. Retriev.*, 2008.
[43] L. Cao, J. Liu, and X. Tang, "3D object retrieval using 2D line drawing and graph based relevance reedback," in *Proc. ACM Int. Conf. Multimedia*, 2006.
[44] S. Yoon, M. Scherer, T. Schreck, and A. Kuijper, "Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours," in *Proc. ACM Int. Conf. Multimedia*, 2010.

Boqing Gong (S'12) received the B.E. degree from the University of Science and Technology of China in Electrical Engineering and Information Science in 2008, and the M.Phil. degree from Chinese University of Hong Kong in Information Engineering in 2010. He is pursuing the Ph.D. degree in Computer Science at the University of Southern California. His research interests include computer vision and machine learning.

Jianzhuang Liu (M'02–SM'02) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a research fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a postdoctoral fellow, then an assistant professor, and then an adjunct associate professor with The Chinese University of Hong Kong. He joined Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, as a professor, in 2011. He is currently a chief scientist with Huawei Technologies Co. Ltd., Shenzhen, China. He has published more than 100 papers, most of which are in prestigious journals and conferences in computer science. His research interests include computer vision, image processing, machine learning, multimedia, and graphics.



Xiaogang Wang (S'03–M'10) received the B.S. degree from the Special Class for Gifted Young at University of Science and Technology of China in Electrical Engineering and Information Science in 2001, and the M.Phil. degree from Chinese University of Hong Kong in Information Engineering in 2004. He received the Ph.D. degree in Computer Science from the Massachusetts Institute of Technology. He is currently an assistant professor in the Department of Electronic Engineering at the Chinese University of Hong Kong. He was the area chair of IEEE International Conference on Computer Vision (ICCV) 2011. He is the associate editor of the Image and Visual Computing Journal. His research interests include computer vision and machine learning.

Xiaoou Tang (S'93–M'96–SM'02–F'09) received the B.S. degree from the University of Science and Technology of China, Hefei, in 1990, and the M.S. degree from the University of Rochester, Rochester, NY, in 1991. He received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996.

He is a Professor in the Department of Information Engineering and Associate Dean (Research) of the Faculty of Engineering of the Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at the Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing.

Dr. Tang received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009. He is a program chair of the IEEE International Conference on Computer Vision (ICCV) 2009 and an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI) and *International Journal of Computer Vision* (IJCV).