Query Adaptive Multiview Object Instance Search and Localization Using Sketches

Sreyasee Das Bhattacharjee, *Member, IEEE*, Junsong Yuan, *Senior Member, IEEE*, Yicheng Huang, Jingjing Meng, *Member, IEEE*, and Lingyu Duan, *Member, IEEE*

***Abstract*—Sketch-based object search is a challenging problem mainly due to three difficulties: 1) how to match the primary sketch query with the colorful image; 2) how to locate the small object in a big image that is similar to the sketch query; and 3) given the large image database, how to ensure an efficient search scheme that is reasonably scalable. To address the above challenges, we propose leveraging object proposals for object search and localization. However, instead of purely relying on sketch features, we propose fully utilizing the appearance features of object proposals to resolve the ambiguities between the matching sketch query and object proposals. Our proposed query adaptive search is formulated as a subgraph selection problem, which can be solved by the maximum flow algorithm. By performing query expansion, it can accurately locate the small target objects in a cluttered background or densely drawn deformation-intensive cartoon (Manga like) images. To improve the computing efficiency of matching proposal candidates, the proposed Multi View Spatially Constrained Proposal Selection encodes each identified object proposal in terms of a small local basis of anchor objects. The results on benchmark datasets validate the advantages of utilizing both the sketch and appearance features for sketch-based search, while ensuring sufficient scalability at the same time.**

***Index Terms*—Sketch Based Search, object localization, object recognition, object retrieval, multi-view proposal selection, transductive clustering.**

I. Introduction

THE task of the object instance search is to retrieve and localize all similar objects in the database images. An enormous amount of image and visual data being available via several web based resources like Flickr, Facebook etc., an effective search module can support automatic annotation of multimedia contents and help content-based retrieval. Although sufficient works have been reported to efficiently explore image/object level similarities [1], [2] for various application scenarios, obtaining precise image example sufficing the user specification may not be always handy and in such cases sketch can be an alternative solution to initialize the search. Although the hand drawn sketch may not be precise, if drawn with care, it can still provide sufficient amount of object details to achieve an effective instance search [3]–[6]. Despite previous work of sketch-based image retrieval, object instance search model needs to address three main challenges: (1) Sketches are far from being complete in terms of the object information that would be critical for a reliable search performance. For example, if a user is looking for some 'pyramid' images, only drawing a 'triangle' is not sufficiently discriminative to uniquely resemble the pyramids. On the other hand, the precise image example sufficing the user specification may also not be handy in every instance. Like, in a public gathering, when you notice a stranger carrying a fashionable handbag which you would desperately want to buy, taking a photograph is always not very decent. Instead, drawing a sketch of its shape or at least its displayed logo (which may not be among your known brands) with fingers in the smartphone will probably be easier. Therefore, with the advent of touch screen devices, sketch based query input is indeed a viable and more effective option for the present generation of users. (2) Accurately matching and locating the small objects of interest in a big image of significantly cluttered background is still challenging. To the best of our knowledge, such a localization problem is not fully explored in the previous works of sketch based image retrieval. (3) The challenges continue to become more critical with the ever increasing database size in time, as the system is expected to remain sufficiently scalable to maintain its effectiveness.

The proposed graph-based search optimization framework is to enable query adaptive object instance search using sketches. The first challenge is that the quality of the sketch provided by a random user may not be satisfactory always, which makes the performance deteriorate. Therefore, instead of purely relying on the sketch features, we also explore the appearance based similarity among database images in a graph regularization framework to improve the search quality. Object proposals are used to identify a small number of candidate object regions irrespective of its sizes, which enable to evaluate object level similarities to ensure a reliable localization performance as well. Fig. 1 describes the entire framework in details.

Manuscript received October 14, 2017; revised January 8, 2018 and February 11, 2018; accepted February 11, 2018. Date of publication March 9, 2018; date of current version September 18, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61661146005 and Grant U1611461, in part by the Key Research and Development Program of Beijing Municipal Science and Technology Commission (No. D171100003517002), and in part by the PKU-NTU Joint Research Institute through the Ng Teng Fong Charitable Foundation, and start-up grants of the University at Buffalo. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei. *(Corresponding author: Sreyasee Das Bhattacharjee.)*

S. D. Bhattacharjee is with the Department of Computer Science, University of North Carolina, Charlotte, NC 28223 USA (e-mail: sreya.iitm@gmail.com).

J. Yuan and J. Meng are with the Department of Computer Science & Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: jsyuan@ntu.edu.sg; jingjing.meng@ntu.edu.sg).

Y. Huang and L. Duan are with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100080, China (e-mail: anorange0409@gmail.com; lingyu@pku.edu.cn).

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.2018.2814338

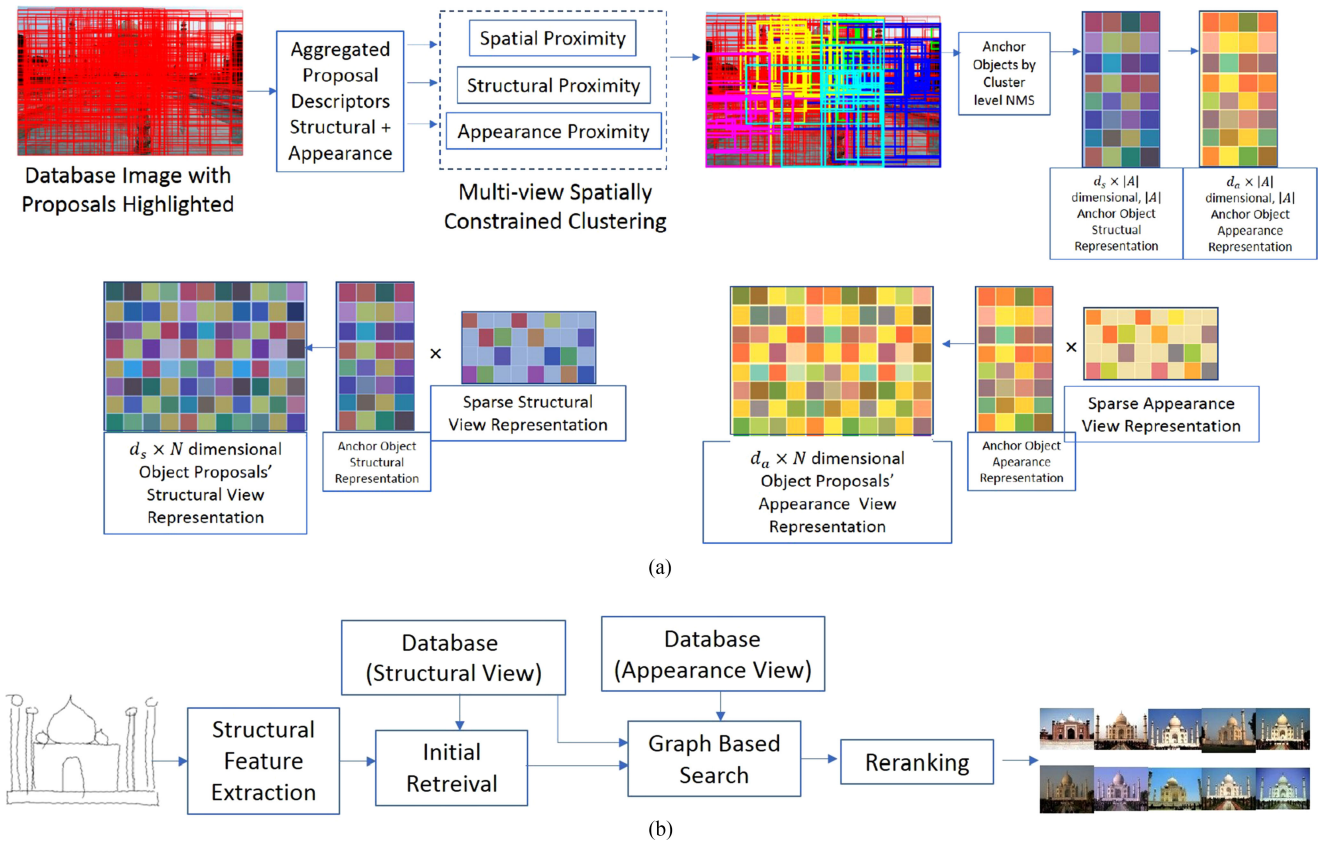


Fig. 1. The Overview of the proposed sketch based object search framework. Figure (a) describes the feature extraction Module from a given database image. Figure (b) shows the proposed retrieval method.

Each database image is segmented in terms of a collection of image sub-regions called object proposals to capture the category independent potential object regions, each of which is then matched with the query sketch. Important to note that such class invariant object hypotheses are typically highly overlapped and thus redundant in nature, the processing of which require a large amount of storage and affects with reduced computational efficiency. Unlike the typical clustering methods attempting to eliminate only the overlapped proposals, the proposed multi-view proposal selection scheme is designed to identify a set of shortlisted proposals which are simultaneously more distinct from the spatial, structural and appearance views. This ensures a more comprehensive grouping performance at the image level. Given the initial retrieval obtained based on structural similarities, the proposed query adaptive finetuning is formulated as a graph regularization problem. Each object proposal corresponds to a graph node with its prior weight depending on its shape similarity extent with the query sketch. While the query-to-image structural proximity can be obtained only based on their corresponding shape-similarity, edge strength among the proposals (each representing a node in the graph) however, depend on their mutual similarities of both the shape and appearance features. A joint optimization is then performed via maximum flow to select the subset of object proposals that are similar to the query in terms of their sketches, and also mutually similar to each other in terms of both sketches and appearances. The finally selected object proposals can well capture all the object instances at different images. An extensive set of experiments performed on multiple benchmark image and video datasets demonstrates its superior performance compared with existing methods. The datasets used here include Flickr15K [7], EITZ [8], a subset of Flickr3M [9] called FlickrLarge, NTU-VoI [10], eBDtheque Comic collection [11] and the very recent Manga109 [12] Japanese comics collections.

The advantages of our query adaptive object instance search can be summarized as follows:

- The proposed multi-view spatially constrained proposal selection strategy enables to identify a smaller set of selected proposals more unique (within the set) from multiple view points like spatial, structural and appearance. While the spatial information helps to minimize the inclusion of overlapped proposals, structural and appearance proximity information capture their content level similarities. This approach is thus found to be more suitable for images with many co-existing objects of interest like comic images, which are typically very dense often containing multiple objects in a close spatial proximity of each other.
- Unlike traditional Sketch Based Image Retrieval (SBIR), the proposed Sketch Based Object Search (SBOS) is designed to perform both search and localization and can perform more reliably by exploring the appearance similarities among the object proposals capturing the image

regions from the database. The comprehensive query expansion strategy can help outlier elimination via graph selection. Thanks to object proposals, it can also accurately locate the object instance in the cluttered background.

- Our subgraph selection formulation which fuses multiple cross modal features, is sufficiently generic to address the task of object instance search problem to various other application scenarios, e.g., comic character retrievals or object search in video as shown in our experiments.

The rest of the paper is organized as follows: Section II briefly describes some related works. The detailed description of the proposed search approach is given in Section III. Section IV presents the experimental results. Finally, the conclusion is in Section V.

II. Related Work

Our work is related to the problem of Sketch Based Image Retrieval (SBIR) , which uses only a single hand drawn sketch without having used any additional information such as text keywords, user-click etc.. Based on the underlying features in use, the entire spectrum of SBIR methods can be partitioned into two categories: methods using (1) global descriptors and (2) local descriptors.

Global descriptors have been very handy in the traditional SBIR methods. For example, Park *et al.* [13] use frequency histogram of the edge orientations or Chalechale *et al.* [3] use a histogram representing the distribution of edge pixels as the descriptors. Bimbo & Pala [14] propose Elastic contours to define a parametric curve that is deformed appropriately to fit with the object boundary. Similarity invariant Zarnike moments [15] have also been used for matching sketches to the image. Cao *et al.* [16] propose Edgel index, an indexing scheme using a Chamfer Distance based descriptors for such purpose. In fact, various methods [4], [16] use Oriented Chamfer Matching (OCM) in some form for faster search. However, the Chamfer Matching based methods are not very robust to translation etc. Chen *et al.* [17] use a freehand sketch along with some text labels to search from the Internet images. Contour consistency filtering is performed using Shape Context descriptor. Cao *et al.* [18] propose SYM-FISH which incorporates the symmetry information into the shape context descriptor. However, such global descriptors [15], [19] are not very suitable for this task as they are more sensitive to deformations, occlusions and transformations etc. and thus not suitable for many generic real life problem scenarios.

In order to address these, various sophisticated local descriptors have been proposed in the recent literature. Eitz *et al.* [5] use the Structure Tensor descriptor. BoF based methods [20] have also been adopted for this scenario. Hu *et al.* [7] use a gradient field (GF) image to represent the sketches, which is then used to compute a multi-scale Histogram of Oriented Gradient (HoG) descriptor in a BoF model. Some other variants of HoG descriptors [21], [22] have also been proposed for the retrieval problem scenario. Sousa *et al.* [23] segment the sketch into a number of parts, where topological and geometric information is utilized for matching purpose. Furuya & Ohbuchi [6] define visual saliency weighting to highlight the object of interest. In order to address the sparse nature of the sketch representatives, Saavedra *et al.* [24] propose Histogram of Edge Local Orientations (HELO) to estimate the local orientations by grouping pixels into a fixed number of cells and assign just a single orientation estimate for each cell. Wang *et al.* [9] treat the query and the edgemap as a collection of edge-segments. Histogram of Line Relationship (HLR) descriptor captures the relations between these line segments within the descriptor.

While majority of the literature relies on handcrafted feature, they remain effective for the near-planer objects with limited view angle variations, so that sufficient correspondences can be established between query and the database entries. In order to attain a more reliable classification performance, only a few attempts have been reported to design the deep models [25], [26] for the task of recognizing the hand-drawn sketches against a sketch database. In a few recent works [27], [28], deep models have been proposed to learn the cross domain features from sketch and image with end-to-end frameworks.

As mentioned earlier that the database images typically contain a single dominating object capturing a significant portion of it. Therefore, precise localization is not a big concern in such cases. However, in a real-life image or video, objects of interest may also be small occupying only a small portion of the entire image content, which (in case of a video) may also vary in its position and appearance from frame to frame. Therefore, an accurate localization is still a critical issue to address in a generic scenario.

This motivates us to design a framework which is equally adept at tracing small objects in a database image. In order to evaluate, therefore, we test the proposed algorithm for searching a specific comic character in a comic album database, where the characters are typically hand-drawn and very much sketch like. In fact, a larger range of deformation along with a dense placement of the sequential story snapshots observed, make the localization problem even more challenging. Some works have been reported for Manga (Japanese Comic) like comic data collections; vectorization [29], coloring [30], generating/recognizing layouts [31], manga-like rendering [32], detecting text balloons [33], segmentation [34], retrieval [35] etc.. Matsui *et al.* [35] capture the local spatial information by defining a histogram of edge orientation for a local area. A set of recent works [36], [37] dealt with the problem of manga frame (or character) detection. While some commercial products like search engines have already attempted to address this problem, their solution heavily relies on the text based query related information provided by the user, the scope of the search process is therefore pretty limited.

By using multiple automatically generated deep learnt features, specialized in capturing shape and appearance information within a single framework, enables the proposed method to handle various issues faced by the typical global and local features used in the community. While most of the recent search methods having their main focus on retrieval than localization (specifically for small objects) except some [7], [35], [38], the proposed search strategy shows an equal expertize in retrieval and localization simultaneously. A handful of top-ranked Edgebox

generated proposals are used to identify the candidate object regions, which are then investigated thoroughly by the proposed search methodology for attaining an impressive matching and localization performance.

III. Proposed Approach

Given an image database $\mathcal{D}_{img} = \{\mathcal{I}_i\}_{i\in 1,..,M}$, the ultimate task in this paper is to identify the subset $\{\mathcal{I}_i\}$ of $\mathcal{D}_{img}$ containing images having the similar instances as the object of interest whose single hand-drawn sketch Q is provided as query. This also involves localizing the object's position within each gray level (or color) database image $\mathcal{I}_i$. In order to achieve this goal, the first task is to identify a set of interest candidate regions called 'proposals', at which the probability of an object's presence is high.

A. Image Representation and Query Matching

EdgeBox by Zitnick & Dollar [39] is used to pick out a smaller set of candidate object regions in an image. Any proposal with a skew ratio (defined as the ratio of the skew of the proposal and an average skew of the logo categories in the database) beyond a certain range is discarded. Due to its sole dependence on the sparse yet informative edge-based representation, EdgeBox is simultaneously efficient and more accurate in spotting a smaller set of image interest regions. Given each such region identified using a bounding box, the associated 'objectness measure' quantifies its likelihood to contain an object.

Each image in the database is presented as a pool of N object proposals, $\mathcal{I} = \{P_j\}_{j=1}^{N}$, where each proposal P_j is represented by a high dimensional (d_0) feature vector $\mathbf{p}_j \in \mathbb{R}^{d_0}$. The whole image dataset is thus presented as a big pool of object proposals $\mathcal{D}_{img} = \{\mathcal{I}\}_i^M = \{P_j\}_{j=1}^{MN}$. Given a query object sketch Q which is treated as an equivalent to a proposal extracted from a database image and represented by a d_0 dimensional vector $\mathbf{q} \in \mathbb{R}^{d_0}$, our goal is to rank the proposals in the database $\mathcal{D}_{img}$ such that the higher ranked proposals can find the similar object instances. In this work, the automatically learnt DNN features obtained from various layers of Sketch-a-Net model are used as the proposal descriptors. and the L2 distance is used to compute the dissimilarity between two object regions represented by $\mathbf{p}_i$, $\mathbf{p}_j$, i.e., $d(\mathbf{p}_i, \mathbf{p}_j) = \|\mathbf{p}_i - \mathbf{p}_j\|_2$.

Given the matching scores of object proposals in $\mathcal{D}_{img}$, the dissimilarity score between query Q and any image $\mathcal{I}$ ($\in \mathcal{D}_{img}$) is defined in terms of its best matched proposal score : $d(Q, \mathcal{I}) = \min_{P_j \in \mathcal{I}}[d(\mathbf{q}, \mathbf{p}_j)]$. Structured Edge detector [40] is used to initially compute the required edge map in an image. A small set of edges with each image pixel p assigned with an edge magnitude m_p and an orientation θ_p creates this image edgemap. A multi-scale variant of the approach enables us to estimate an approximate scale for every pixel at which the edge response is maximized. In order to eliminate some spurious edges, only the pixels with an edge magnitude $m_p > 0.1$ is defined as the edge pixels. EdgeBoxes use this edge map to generate a set of object bounding box proposals along with their respective 'objectness' scores. Due to sparse yet informative representation of edges, EdgeBoxes are computationally efficient and also more accurate to identify a small set of salient image regions simultaneously. One bottleneck of using such initial object hypotheses like Edgebox lies in the fact that these proposals are pretty redundant and largely overlapping in nature. However, in order to capture a maximum amount of information in a clutter intensive image/videoframe (like identifying small objects occupying only a very small image region) it is still important to investigate a large number of object proposals per image. In order to address these mutually orthogonal issues, we propose a multi-view spatially constrained clustering approach to identify a smaller number of object proposals (called anchor objects) which capture a set of more mutually exclusive information, working as the basis for the entire image specific proposal collection. This entire set of proposals originating from a single image forms a image specific dictionary for the underlying image.

B. Multi-View Spatially Constrained Proposal Selection (MVSCPS)

In order to identify the smaller set of anchor objects in each of these image specific dictionaries, we use a multi-view clustering approach to combine the spatial, structural and appearance (each representing a different view) based affinities among the proposals via a Multi-View Spectral and Transductive Clustering (MVSTC) [41]. The basic philosophy behind MVSTC is to generalize the usual single view normalized cut to the multi-view cut, where obtained cut is close to optimal for each graph representing the affinity matrix corresponding to each view. The multi-view normalized cut is approximately optimized via a real-valued relaxation, which does not lead to convex combination of graph Laplacians, but results in a vertex-wise mixture of Markov chains associated with multiple graphs.

In this paper, MVSTC having been utilized to group the similar proposals at the image level, offers a more compact object level description scheme which takes into consideration of similarities from three different views: (1) Spatial, (2) Structural and (3) Appearance, all within a single formulation. In the proposed Multi-View Spatially Constrained Proposal Selection (MVSCPS) strategy, the spatial affinity between two proposals is evaluated using their pairwise overlap ratio in terms of the Union over Intersection (UoI). Sketch-a-Net ($\mathbf{p}$) and CNN ($\mathbf{c}$) feature respectively represent the proposal's structure and appearance information. The Structural and Appearance similarity are computed using the cosine kernels of two proposals' corresponding representative descriptors. For the sake of ease in discussion, we will use $\mathbf{p}^v$ to denote the vth-view specific descriptor of a proposal P. In order to obtain the view specific sparse representations for both descriptors Sketch-a-Net ($\mathbf{p}$) and CNN ($\mathbf{c}$), we repeat the following transformation scheme for each of these views separately.

Sparse Representation: Given the collection of N proposals representing an image $\mathcal{I}$, we cluster them to categorize into c groups of the similar proposals. Within each such cluster $\mathcal{C}_t$ ($\forall t \in \{1, \ldots, c\}$) of similar proposals obtained from $\mathcal{I}$, a non-maxima suppression reduces the information redundancy by eliminating overlapped proposals with $UoI > 0.5$ while retaining only one of them with highest objectness score. This set

of shortlisted proposals constitutes the cluster representative set of Anchor proposals $\mathcal{A}_t = \{a_{t,j}\}_j$. The set of all such Anchor objects $\mathcal{A} = \cup_{t=1}^{c} \mathcal{A}_t$, constitutes an image level dictionary for $\mathcal{I}$. In order to achieve an efficient sparse representation scheme, the task is to further learn the view specific coding matrices $\mathbf{H}^v \in \mathbb{R}^{|\mathcal{A}| \times N} = [\mathbf{h}_1^v, \ldots, \mathbf{h}_N^v]$ satisfying the following criteria:

$$\min_{\mathbf{h}^v \in \mathbb{R}^{|\mathcal{A}|}} ||\mathbf{p}^v - \mathbf{A}^v \mathbf{h}^v||_2 \text{ such that } \mathbf{1}^T \mathbf{h}^v = 1 \tag{1}$$

where, $\mathbf{A}^v = [\mathbf{a}_1^v, \ldots, \mathbf{a}_{|\mathcal{A}|}^v]$ and $\mathbf{a}_j^v$ represents the v-view descriptor of $a_j (\in \mathcal{A})$. However, in order to ensure sparsity, we represent each proposal descriptor $\mathbf{p}^v$ using a code obtained from (1) by searching only against the local base $\mathcal{A}_l^v = \cup_{t \in \mathcal{N}_{\mathbf{p}^v}} \mathcal{A}_t$, where $\mathcal{N}_{\mathbf{p}^v}$ consists only of k_0-nearest cluster neighbors of $\mathbf{p}^v$ in the view specific feature space. This restricts the search domain of (1) against $\mathcal{A}_l^v$ and requires solving the following sub-problem, which ensures the resulting $\mathbf{h}^v$ to lie in a much smaller $|\mathcal{A}_l^v|$ dimensional subspace of $\mathcal{A}$ [42].

$$\min_{\mathbf{h}_l^v \in \mathbb{R}^{|\mathcal{A}_l^v|}} ||\mathbf{p}^v - \mathbf{A}_l^v \mathbf{h}_l^v||_2 \text{ such that } \mathbf{1}^T \mathbf{h}_l^v = 1 \tag{2}$$

where $\mathbf{A}_l^v$ represents a submatrix of $\mathbf{A}$, consisting of columns describing the v-view of the anchors from $\mathcal{A}_l^v$. It is important to note that the solution for (2) can be derived analytically by:

$$\begin{aligned} \tilde{\mathbf{h}}_l^v &= (\mathbf{A}_l^v - \mathbf{p}^v \mathbf{1}^T)(\mathbf{A}_l^v - \mathbf{p}^v \mathbf{1}^T)^T \setminus \mathbf{1} \\ \mathbf{h}_l^v &= \tilde{\mathbf{h}}_l^v / (\mathbf{1}^T \tilde{\mathbf{h}}_l^v) \end{aligned} \tag{3}$$

where $\setminus$ represents a left-matrix division and $/$ performs a right-matrix division[43]. Therefore, each $\mathbf{p}^v \approx \mathbf{A}^v \mathbf{h}_l^v$. Thus the large number of proposals per image being represented in terms of some sparse linear codes reduce the memory requirement and computation cost significantly, while achieving a nearly comparable search precision at the same time.

Computational Efficiency of the Proposal Selection Scheme: The proposed proposal selection scheme proves to be handy in efficient storage management. Since typically $c \sim 0.2N$ to obtain a reasonable retrieval performance, the number ($|\mathcal{A}_l^v|$) of nonzero terms in $\mathbf{h}_l^v$, is even lesser and usually $\mathcal{A}_l^v \sim 0.1N$, the storage requirement due to this encoding reduces significantly. More importantly in many real life applications including ours, $\mathcal{A}_l^v < 0.01 d_0$, where d_0 is the dimension of the proposal descriptors. which is mostly very high.

C. Initial Retrieval using Sketch-a-Net

The deep neural network (DNN), Sketch-a-Net [25] learnt on the TU-Berlin sketch dataset [21], is used to represent the edgemap of each database proposal (resized to a pre-defined size) in terms of a d dimensional feature $\mathbf{p}$. Sketch-a-Net has five convolutional layers each with rectifier (ReLU) units and the first, second and fifth layers followed by max pooling. Then added three Fully Connected Layer (FC) with a dropout regularization applied to each of the first two. For more details on the architecture, we refer the interested readers to [25]. We have used three kinds of compact Sketch-a-Net deep learnt features for representation: the 512 dimensional outputs of first (and second) fully connected layers, called L6 (and L7) obtained after dropout regularization and the 250 dimensional output of the third fully connected layer (called L8). More details on their individual performances will be discussed in Section IV.

As shown in the region cropping of Fig. 1, 5 sub-regions are cropped from each resized proposal. Given a database proposal, the entire set of 10 n dimensional descriptors representing its 5 cropped regions (and their corresponding horizontally flipped versions) are concatenated together to obtain its $(10 \times n)$ dimensional Sketch-a-Net representatives.

D. Re-ranking

Give a query Q the initial retrieval set is denoted as $\mathcal{N}$. While Sketch-a-Net generated features work well compared to other state-of-the-art shape descriptors, due to the lack of an equivalent amount of information, it is unfair to expect a performance competitive to the typical deep features designed for gray (or color) images. Now, it is important to note that, not all of these retrievals are actually similar to Q. Moreover, although Q provides only a sketch based query information, an efficiently chosen subset of $\mathcal{N}$ actually serves as a more salient expanded representations for Q. The subset has a smaller collection of some more reliable matches. Therefore, choosing the subset $\mathcal{E}$ $(\subset \mathcal{N})$ that holds a handful of good matches for Q is important. At the same time, the entry of false positives is minimized. In fact, $\mathcal{E}$ is designed to provide a more insightful and complete representation (in terms of both its structure and gray/color level appearances) for Q. In contrast to the Average Query Expansion (AQE) [44], which uses just a collection of few top retrievals as the expanded query representation or typical clustering algorithms like k-means or spectral clustering, which focus only on the mutual similarity but does not consider the prior weight of individual retrievals, the proposed graph-based re-ranking method considers both prior of retrievals and their pairwise similarities within a single formulation to obtain a set of more reliable entries in $\mathcal{E}$.

As such, AQE has been proved to be effective as a re-ranking scheme in a generic object search scenario, where appearance based features are typically used for the representation purpose. However, in this problem scenario, where query information is incomplete and thus all the initial top-retrievals are not equally reliable in general, the proposed graph-based re-ranking scheme holds a better promise. More details on this comparative study are presented in the Section IV.

In order to capture the sufficient appearance information within the descriptor, the first step is to choose the effective feature representative. While sketch-a-net feature can capture the structural information, the appearance information is represented using two effective representation schemes described below.

1) Extracting Appearance Features: In the re-ranking phase, CNN learnt deep feature is used to capture the proposal level appearance information. In order for comparative study, we have also used Compact Descriptor for Visual Search (CDVS) [45], [46] as a hand-crafted feature for the purpose. An improved

performance using CDVS as a replacement of CNN during fine-tuning proves the effectiveness of the method. However, CNN being the best performer in this task, we have set CNN as the default choice in all our experiments.

CNN Features: Convolutional Neural Network (CNN) activations are used to represent each proposal in the database. Each proposal captures some dominant object region in an image. A CNN-based feature descriptor, therefore, represents a very generic object level representation for the underlying image. The 4096 dimensional SPP-net [47] activation ($\mathbf{c}_i$), which is an aggregated descriptor obtained from the collection of deep features (generated using a fast model by Zeiler and Fergus [48]) is used to represent each proposal in $\mathcal{D}_{img}$.

As can be seen in the Section IV that a combination of both these appearance information works excellent to achieve a promising retrieval performance, while the proposed re-ranking still capable of attaining a competitive performance exclusively with CNN learnt features capturing the mutual similarity among the database proposals.

2) Query Expansion via Graph Based Reranking: Given a query Q, the set of initial top-K retrieved proposals in $\mathcal{N}$ is expressed in terms of a graph structure, where each graph node represents one of these top-K proposals. The edges and their corresponding weights are defined using their pairwise mutual appearance based similarity.

In other words, $G = (\mathcal{V}, E, S)$, $|\mathcal{V}| = K$ and each $v \in \mathcal{V}$ represents one of the top-K proposals in $\mathcal{N}$. there is an edge $e(i,j) \in E$ between two vertices v_i and v_j in G if and only if the corresponding proposals (P_i and P_j) are reciprocal K-neighbors to each other, i.e., P_i and P_j both are the top-K neighbors of each other. Appearance based SPP-feature descriptor is used for representation and L2 distance is computed to evaluate the pairwise dissimilarity in this purpose.

Edge-strength between $v_i, v_j \in \mathcal{V}$ is computed as:

$$S(v_i, v_j) = \begin{cases} \text{sim}(\mathbf{p}_i, \mathbf{p}_j) sim(\mathbf{c}_i, \mathbf{c}_j) & \text{if } e(i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{p}_i$ and $\mathbf{c}_i$ respectively represent the Sketch-a-Net and SPP-feature of a proposal represented by v_i in G, $sim(,)$ computes the cosine similarity between two vectors.

Important to note that the set of initial matching scores used to identify $\mathcal{N}$ works as a prior in our re-ranking approach. The similarity extent of P_i (representing $v_i \in \mathcal{V}$) with Q and defined as:

$$T_i = sim(\mathbf{q}, \mathbf{p}_i) \tag{5}$$

here $\mathbf{q}$ and $\mathbf{p}_i$ respectively represent the Sketch-a-Net feature of the query Q and a proposal $P_i \in \mathcal{N}$.

Given this G, the proposed re-ranking approach selects a maximal subset of the pairwise similar database proposals through graph regularization. The objective function [49] is defined as follows:

$$\underset{\mathbf{d} \in \{0,1\}^K}{\arg\min} \left[\lambda \mathbf{d}^t W \mathbf{d} - \mathbf{T}^t \mathbf{d} + \eta ||\mathbf{d}||_0 \right] \tag{6}$$

where $\mathbf{T} = [T_1, ..., T_K]^t \in \mathbb{R}^K$, $W = D - S$ and $D \in \mathbb{R}^{K \times K}$ a diagonal matrix with $D(i,i)$ representing the degree of v_i in G and $\mathbf{d} \in \{0,1\}^K$ is an indicator vector specifying the inclusion/exclusion of a node in the resulting subgraph.

The second term $\mathbf{T}^t\mathbf{d}$ of Eq. (6) plays prior in the proposed re-ranking scheme. Given W defined as the Laplacian, the first term $\mathbf{d}^t W \mathbf{d}$ attempts to emphasize on the subgraph with higher edge strength, while eliminating the nodes with higher degrees. Thus the less discriminative proposals which typically match with many others, are pruned from the list. Finally, the third term $||\mathbf{d}||_0$ acts as a regularizing factor that ensures a certain amount of sparsity of $\mathbf{d}$. $\eta \in [0\ 1]$ controls the effect of sparsity. The set of the resulting shortlisted matches ($\mathcal{E}$) corresponding to the nodes in this maximal subgraph, is used as a more salient expanded query representation for Q. The sparsity factor η controls the size of $\mathcal{E}$.

Computational Efficiency of the Graph Search Process: In our implementation for the graph-based search, we use the Boykov Kolmogorov algorithm [50] for this purpose. With an order $O(n_c K^2 |E|)$, where n_c is the size of the minimum cut, the resulting maximum flow identifies a small subgraph of maximally connected nodes.

3) Optimization: In order to prove the convergence of the objective function (6), the first thing to notice is that the binary indicator vector $\mathbf{d}$, $||.||_0$ and $||.||_1$ are equivalent, which ensures convexity of the third term.

By definition, $0 \leq S(i,j) \leq 1\ \forall v_i$ and $v_j \in G$. $D(i,i) = \sum_j S(i,j)$. The second term in the objective function of (6) can be re-written as :

$$\begin{aligned} \mathbf{d}^t W \mathbf{d} &= \sum_{i=1}^{K} d_i \left(D(i,i) - \sum_{j=1}^{K} S(i,j) d_j \right) \\ &= \sum_{i=1}^{K} d_i \left(\sum_j S(i,j) - \sum_{j=1}^{K} S(i,j) d_j \right) \\ &= \sum_{i=1}^{K} \sum_{j=1}^{K} S(i,j) d_i \left(1 - d_j\right) \end{aligned} \tag{7}$$

and the function in (6) turns out to be:

$$\underset{\mathbf{d} \in \{0,1\}^K}{\arg\min} \left[\lambda \sum_{i=1}^{K} \sum_{j=1}^{K} S(i,j) d_i \left(1 - d_j\right) + \sum_{i=1}^{K} d_i (\eta - T_i) \right] \tag{8}$$

The first term is thus a cut-function for the graph G. In order to format the second term in terms of a cut-function, we define a hand-crafted supergraph $G^{Aug} = G \cup \{s\} \cup \{t\}$ of G with two additional vertices s and t and a new set of induced edges with known weights defined as follows:

$$S(s,i) = \begin{cases} (T_i - \eta) & \text{if } T_i > \eta \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Fig. 2. Some example search results from the Flickr15K dataset.

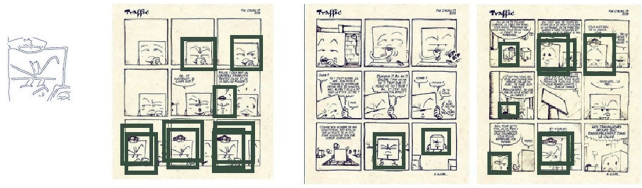


Fig. 3. Given the comic query edge map shown in the left column, is extracted to be used as an input to the system. The right column shows the top-20 image level retrieved result.

and

$$S(i,t) = \begin{cases} (\eta - T_i) & \text{if } T_i < \eta \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Which means that the new node s will connect only those v_i's in G, which are more similar to Q (i.e., $T_i > \eta$) and t will connect other v_i's in G which are relatively less similar to Q (i.e., $T_i < \eta$), for a chosen $\eta > 0$. Given an indicator vector $\mathbf{d}$, we define $\mathcal{R}$ as the set of all indices i, for which the corresponding $d_i \neq 0$. Then, the second term in Eq. (8) can be rewritten as:

$$\begin{aligned} &\sum_{i=1}^{K} d_i(\eta - T_i) \\ &= \sum_{\substack{i \in \mathcal{R} \\ T_i < \eta}} d_i(\eta - T_i) + \sum_{\substack{i \in \mathcal{V} \\ T_i > \eta}} d_i(\eta - T_i) - \sum_{\substack{i \notin \mathcal{R} \\ T_i > \eta}} d_i(\eta - T_i) \\ &= \sum_{i \notin \mathcal{R}} S(s,i) d_s (1 - d_i) + \sum_{i \in \mathcal{R}} S(i,t) d_i (1 - d_t) \\ &\quad + \underbrace{\sum_{\substack{i \in \mathcal{V} \\ T_i > \eta}} d_i(\eta - T_i)}_{\text{a Constant Term}} \end{aligned} \tag{11}$$

where $d_s = 1$ and $d_t = 0$. It shows that the second term can also be represented in terms of a cut-function in G^{Aug}. Therefore, finding an optimized solution to Eq. (8) is equivalent to obtaining a minimum cut solution in G^{Aug}. The maximal flow algorithm is used for this purpose.

E. Matching with Fusion

Given the query Q and $\mathcal{N}$ as the initial set of its top retrievals, $\mathcal{E}$ is used as an expanded representation for Q and its resulting matching cost with $P_i \in \mathcal{N}$ is defined as:

$$d_{rank}(Q, P_i) = d(\mathbf{q}, \mathbf{p_i}) \frac{\sum_{j \in \mathcal{E}} d(\mathbf{c}_j, \mathbf{c}_i)}{|\mathcal{E}|} \tag{12}$$

where $\mathbf{q}$ represents the Sketch-a-Net feature of Q, $\mathbf{c}_i$ represents the 4096 dimensional SPP feature of P_i and $d(,)$ computes the L2 distance between two vectors. The entire set of proposals in $\mathcal{N}$ is reranked using $d_{rank}(,)$ to obtain the final search result. Some example localization results are shown in Figs. 2 and 3.

Computational Efficiency for the Entire Search: Depending on the database size, given a query, retrieving the top-1000 matches requires approximately 3–5 seconds on average in a stand-alone 3.2 GHZ PC with 8 GB memory.

IV. Experiment

The proposed Sketch Based Object Search (SBOS) framework is used for mainly two types of search related tasks: (1) given only its rough user-drawn sketch as a query, search for an object's instances in the real life gray/color images and video treated as a densely sampled frame sequence and (2) comic character search, where given a single image of the query comic character, the task is to count/localize the same character within the entire comic album. While SBOS is difficult due to the huge difference in the content of the database image and the query, the challenge for the comic character search is attributed to the fact that the instances of a comic character are often small in the very densely drawn comic pages. Moreover, they are mostly hand drawn, having a wider range of shape and appearance deformation compared to the objects seen in a real-life image. A successful evaluation for these two tasks on the varied datasets therefore experimentally proves the generalization capability of the proposed SBIR framework.

Dataset and their Accuracy Measures: The proposed SBOS method in this paper, is evaluated in six different datasets containing images of a wide range of objects with diverse structural characteristics: Flickr15K [7], EITZ [8], a subset of Flickr3M [9] called FlickrLarge, NTU-VoI [10], eBDtheque Comic collection [11] and the very recent Manga109 [12] Japanese comics dataset.

Flickr15K [7] is a large-scale dataset containing around $15k$ photographs sampled from Flickr under Creative Commons license. It consists of a set of 330 sketch queries drawn by 10 random users. The collection of sketch queries are partitioned into 33 shape categories like "circular", "heartshape", etc. The database images are taken from 60 semantics categories (e.g., "pyramid", "bicycle", etc.). A query can represent objects from multiple semantic categories, for example, the "circular" shape belongs to three different semntic categories ("moon", "fire-balloon", and "london-eye") of database images. Following the standard protocol, we use mean Average Precision (mAP) as a metric for evaluation.

EITZ [8] is another large scale dataset, consisting of 31 user drawn sketches outlining various objects and scenarios. Each sketch query is related with 40 database images and the authors provide the corresponding pairwise similarity along with a score on a 7 point scale, where the lowest value 1 means similar and the highest value 7 means dissimilar. Including 100K distractors, the entire database has a total $101,240$ images. Given each sketch query, the generated ranks for its 40 related database images are compared with the ground truth ranks, provided with the

dataset. The resulting correlation value measures the retrieval accuracy for the specific query sketch in consideration. The average correlation score over all the 31 queries defines the final performance measure on the dataset. While for EITZ dataset a standard practice has been to use this score for evaluation, recently Qian *et al.* [51] have proposed another measure called precision under depth n (denoted as $Precision@n$) to measure the quality of the performance, defined as:

$$\text{Precision@n} = \frac{1}{z}\sum_{q=1}^{z}\frac{1}{n}\sum_{i=1}^{n} R_q(i) \tag{13}$$

where $R_q(i)$ represents the confidence of the ith result for the qth query, $q \in [1, 2, .., z]$ and $i \in [1, 2, .., n]$. z is the total number of query used in the experiments. $R_q(i) = 1$ if the ith retrieval is relevant to the qth query, otherwise set as $R_q(i) = 0$ In order for a fair comparison, we have reported the performance measure using both these measures.

FlickrLarge is a subset of the original Flickr3M [9], which is a very recent dataset containing objects from 80 different categories with 20 object images per category and 3M distracters. Five sketches per category, with a total of 400 object drawings are used as the queries for experiments. Due to resource constraints, we have used a subset of the dataset, which consists of the 1600 object images and 200 K distractors and thus the entire dataset contains a total $201,600$ images. The mAP scores computed over all the sketch queries are used for evaluation.

NTU-VoI [10] consists of 146 video clips captured by mobile cameras or downloaded from YouTube, with frame-wise bounding box annotations of object instances. In total 33,018 frames are annotated. The VOI dataset covers diverse scenes, most of which are significantly cluttered with the target objects occupying only a small portion of the frame. $Precision@100$ as defined in Eq. (13), is calculated for the evaluation purpose. In order to explore the effectiveness of the proposed approach, $Precision@1000$ is also reported here.

In order to evaluate the performance of the task of comic character retrieval, eBDtheque [11] and Manga109 [12] datasets are used in the experiments. eBDtheque [11] is a corpus of 100 pages (72 of which are colored) of comic pages from 25 different albums, Franco-Belgian, American comics and Mangas. Important to note that, except Rigaud *et al.* [52], this dataset is popularly used for evaluating the task of comic document analysis, where the goal is to extract panels, balloons, tails, texts, comic characters etc. In contrast, we use this dataset to assess the localization performance of the proposed framework. Following [52], Object level precision, recall measures are used for the evaluation purpose. Given a query object image, the set of top-100 retrieved proposals are treated as the system output. A matched proposal is considered as correctly detected if it overlaps with the query's ground truth. Recall (R) computes the number of the correctly detected object divided by the number of objects to detect. Precision (P) measures the number of the correctly detected objects divided by the number of detected objects. In our experiments, we use the top-100 retrieved proposals as the system output and compute precision/recall measure on this set.

Fig. 4. Some example proposal retrieval results from the EITZ dataset.

TABLE I
PERFORMANCE OF THE VARIOUS SKETCH-A-NET FEATURES USING MAP SCORES

Features \ Dataset	Flickr15K	FlickrLarge
f6 (5120 dim. L6 layer output)	0.27	0.29
f7 (5120 dim. L7 layer output)	0.25	0.30
f8 (2500 dim. L8 layer output)	0.23	0.25

Manga109 [12] dataset has 109 Manga titled albums. Each album includes 194 pages on average, with a total of 21,142 pages. The average size of images is 833×1179. Each image is segmented into multiple panels using Manga panel extraction method by Pang *et al.* [53]. Each Panel is then treated as a separate album image in our work. With an average of 3 panels per image, each album therefore typically contains around 582 comic images in total. As such, the dataset is very new and still under process, as communicated by the authors. Therefore, the ground truth details for the entire dataset being unavailable, we use the Precision under depth n as defined in Eqn 13.

A. *Evaluating for the SBIR task in the natural images*

Flickr15K, EITZ, FlickrLarge and NTU-VoI datasets are used for this part of the experiments. Since these databases are larger and the image backgrounds are cluttered, 100 top-ranked Edge-Box proposals are extracted from each image in the EITZ, Flickr15K and NTU-VoI dataset. The object backgrounds being relatively cleaner, for FlickrLarge we have chosen top-20 proposals per image. All the proposals extracted from the entire collection of images are re-sized to 256×256. Fig. 2, 4 and 6 show some visual results of the top-20 retrievals obtained by the proposed SBOS framework. Typically the retrieved objects are small in a given database image. Therefore, in order for clarity in visualization, we have chosen to display the proposal level retrievals in the figures. As shown in Fig. 2, the proposed search process actually identifies these retrieved proposals as a subregion of the image, whose locations are known a-priory.

Comparative Evaluation of the Sketch-a-Net Features: Three kinds of Sketch-a-Net deep features have been used for the experiments: 5120 dimensional L6 layer activation output (**f6**), 5120 dimensional L7 layer activation output (**f7**) and 2500 dimensional L8 layer activation output (**f8**). Table I shows their comparative performance in the Flickr15K and FlickrLarge dataset. The mAP scores reported in the table are calculated from the collection of initial retrievals, (without re-ranking) obtained after step [A] in Fig. 1. As seen in Table I, the 5120 dimensional **f6** and **f7** are found to be more effective as the descriptors compared to the more compact 2500 dimensional

TABLE II
EFFECT OF RERANKING USING **f7** AS THE PROPOSAL DESCRIPTORS

Method \ Dataset	Flickr15K	FlickrLarge
Initial Retrieval ($c = 20$)	0.241	0.277
Initial Retrieval ($c = 30$)	0.259	0.295
Initial Retrieval ($c = 100$)	0.27	0.30
Initial retrieval ($c = 100$) + AQE	0.236	0.397
Full System ($c = 100$)	**0.319**	**0.418**
Full System ($c = 30$)	0.294	0.402

The Performance is Evaluated Using the mAP Scores. The table reports performance with the choices of different cluster numbers (i.e., $c = |\mathcal{C}|$) for MVSCPS in Section III-B.

f8. While both **f6** and **f7** features report a nearly comparable mAP scores on these two datasets, we have used **f7** for the rest of the experiments in the paper.

Effect of Reranking: Table II shows the performance improvement contributed by the proposed graph based reranking in both Flickr15K and FlickrLarge dataset. In our experiments, we select the top-K initial retrievals to further explore their mutual similarity within the proposed graph-based method and identify a smaller subset $\mathcal{E}$, which acts as the expanded representation of the query. In our experiments, different values for K were chosen from the range $[50, 200]$. While the search performance obtained by varying the K values (or the size of $\mathcal{E}$) are nearly similar, we just report the results using $K = 100$. The other factor affecting the structure of $\mathcal{E}$ is the sparsity factor η. As seen in the experiments, a value of η in the range [0.5 0.8] typically ensures extracting some genuine matches while preserving the desired size constraint $|\mathcal{E}| < 20$ at the same time. In case $|\mathcal{E}| > 20$, we choose the best 20 matches within $\mathcal{E}$ for expanded query representation. As can be observed from the table, the proposed graph based reranking scheme contributes to around 6% gain in mAP performance score on average. In contrast, AQE deteriorates the performance in a scenario where query sketches are not of very good quality and therefore some false positives do exist in the set of initial top retrievals. Although AQE has performed considerably well in improving the search performance of the methods using appearance based feature information which are more complete in nature, in a SBOS based framework where a sketch provides only a rough information about the query, the proposed graph based re-ranking scheme holds a better promise. The other effective component of the proposed framework is its multi-view proposal clustering scheme, which combines the spatial, structural and appearance based affinities among the top retrieved proposals within a single formulation and as seen in the table, this is to ensure a very computationally efficient search mechanism with a competitive performance details. The proposed MVSCPS helps to achieve a nearly equivalent performance (in contrast to choosing all the top 100 proposals from an image) by using only 30 selected anchor objects, which work as an efficient basis for the image level representation task.

Table IV also reports an effective performance gain by the proposed reranking in the EITZ dataset under varying number of clusters chosen for proposal selection task in Section III-B. Fig. 5 displays the performance improvement observed by the proposed re-ranking method over the Sketch-a-Net based initial retrieval. We will resume a detailed discussion on the EITZ dataset results again later in this section.

Fig. 5. Given the sketch query in the left, the top row shows the initial top-10 retrievals and the bottom shows the top-10 retrieved proposals obtained after re-ranking.

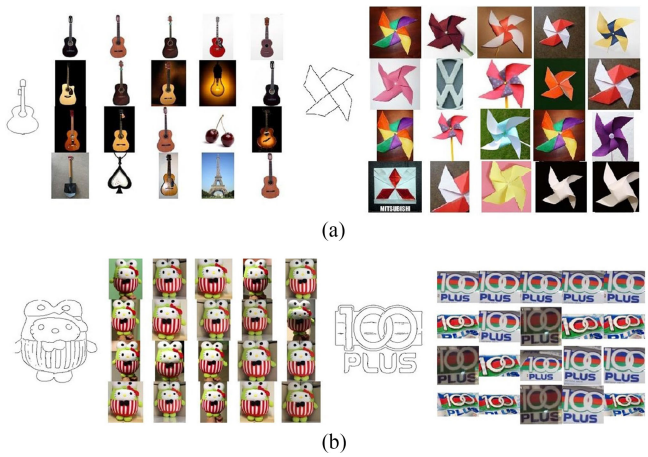


Fig. 6. Some example proposal retrieval results from the (a) FlickrLarge dataset and (b) NTU-VoI dataset.

TABLE III
PERFORMANCE OF THE PROPOSED FRAMEWORK VS OTHER STATE-OF-THE-ART METHODS ON THE FLICKR15K DATASET USING MAP MEASURE. * MARKED METHODS ARE LEARNING BASED

Methods	mAP
Bag of Features [8]	0.131
SIFT [54]	0.091
Self Similarity [55]	0.096
Shape Context [56]	0.081
Structure Tensor [5]	0.080
HoG [57]	0.109
GF-HoG [7]	0.139
PH-HoG [58]	0.20
HLR descriptor [9]	0.171
Color Gradient [59]	0.19
HELO [60]	0.0967
S-HELO [61]	0.1236
RST S-HELO [62]*	0.2002
Learnt Key Shapes [63]*	0.245
Spatial Pyramid Pooling Deep Feature [47]	0.0472
Object level Sketch-a-Net descriptor + Reranking	**0.319**

Comparative Study: Table III reports the performance achieved by the proposed framework on the Flickr15K dataset. As seen in the table, GF-HoG [7] descriptor reports a better

TABLE IV
PERFORMANCE OF THE PROPOSED FRAMEWORK VS BASELINE REPORTED BY EITZ *et al.* [8] ON THE EITZ DATASET USING RANK CORRELATION MEASURE

Methods	mAP
Shape Context	0.18
BoF [8]	0.277
Key Shapes [64]	0.289
Sketch-a-Net descriptor ($\mathbf{f}7$) ($c = 100$)	**0.317**
L7 layer Sketch-a-Net descriptor ($\mathbf{f}7$) ($c = 30$)	0.292
Initial retrieval using $\mathbf{f}7$ + AQE, ($c = 100$)	0.298
Proposed Framework ($c = 100$)	**0.335**
Proposed Framework ($c = 30$)	0.322

The table also reports performance with the choices of different cluster numbers (i.e., $c = |\mathcal{C}|$) for MVSCPS in Section III-B.

performance compared to other five representative features, namely Shape Context [56], Structure Tensor [5], HoG [57], Self Similarity [55] and SIFT [54]. Recently, Wang *et al.* [9] propose a new line segment-based descriptor named histogram of line relationship (HLR) which treats sketches and extracted edges of photo-realistic images as a series of piece-wise line segments and captures the relationship between them. While HLR descriptor [9] outperforms GF-HoG by showing about 4% improvement on its resulting mAP score, PH-HoG [58], a variant of HoG further improves the performance. HELO (histogram of edge local orientations) and its multiple variants have also been proposed to improve the scenario. In order to achieve a better system, Saavedra & Barrios [63] relies on learning a set of key shapes through an intensive offline processing phase and we report their result in the table for completion purpose, such supervised methods are not directly comparable to our unsupervised scenario. Spatial Pyramid Pooling Deep Feature [47] is not found effective for this purpose. Thus, the detailed comparative evaluation of the proposed Sketch-a-Net based SBOS framework shows a promising insight against some of the established literatures [7]–[9], [58], [59], [61], [63] using a set of handcrafted features. The object level Sketch-a-Net descriptors along with the proposed query adaptive reranking scheme offers a dominating performance over other state-of-the-art methods.

In a similar line, Table IV also shows a favorable performance comparison of the proposed method against the baseline reported by Eitz *et al.* [8] on the EITZ dataset. [8] uses a Bag of Features methodology with a variant of the histogram of gradient descriptor and a codebook with 1000 codewords. Saavendra and Bustos [64] propose an alternative with keyshape which does not need to be trained and offers a slight improvement over the baseline. In contrast, as we show in the experiment that the object level L7 layer Sketch-a-Net feature $\mathbf{f}7$ can provide a rank correlation score 0.317. In contrast to AQE, the proposed graph-based re-ranking scheme can improve the performance further to attain an impressive 0.342 score. While rank correlation score has been standard evaluation method for EITZ dataset, we also report $Precision@n$ (as defined in 13) for comparing the performance with Qian *et al.* [51]. Compared to the $Precision@n$ scores 0.20, 0.17 and 0.15 as attained by Qian *et al.* [51], the proposed method attains the impressive $Precision@n$ scores of 0.43 0.38 and 0.31 at $n = 10$, 25 and 50 respectively.

TABLE V
COMPARATIVE PERFORMANCE STUDY OF THE PROPOSED FRAMEWORK IN NTU-VOI DATABASE

Features \ Number of Top Ret. (n)	10	20	30	40	50
GF-HoG [7] (500 words Vocab)	0.43	0.41	0.41	0.36	037
GF-HoG [7] (1000 words Vocab)	0.46	0.46	0.44	0.42	0.42
GF-HoG [7] (1500 words Vocab)	0.44	0.45	0.43	0.39	0.36
SIFT [54] (500 words Vocab)	0.39	0.40	0.36	0.35	0.33
SIFT [54] (1000 words Vocab)	0.41	0.38	0.38	0.37	0.32
SIFT [54] (1500 words Vocab)	0.41	0.37	0.37	0.356	0.31
Sketch-a-Net Only	**0.81**	**0.79**	**0.74**	**0.71**	**0.64**
The Proposed Framework	**0.88**	**0.86**	**0.80**	**0.73**	**0.68**

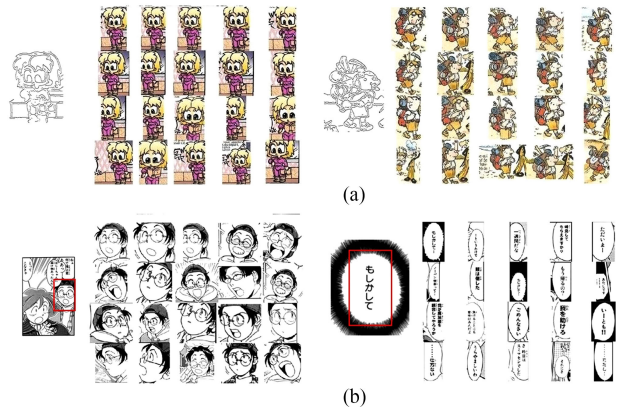


Fig. 7. Some example proposal retrieval results from the eBDtheque [11] and Manga109 [12] dataset.

Table V shows the performance of the proposed framework against GF-HoG [7] and SIFT [54] in the NTU-VoI dataset. While the proposed re-ranking with the help of object level Sketch-a-Net descriptor works excellent in searching the generic object instances, emphasizing the system shows significant stability by reporting an impressive $Precision@1000$ of 0.4349 obtained for the sketch-a-net based initial task, which is further improved by the proposed graph based reranking to 0.5708 is worthful.

B. Evaluating for the Comic Character Retrieval

In this section, the proposed method is evaluated on two available comic album datasets: eBDtheque [11] and Manga109 [12]. Some proposal level retrieval results from Manga109 dataset are shown in Fig. 7.

Experimental Setup: In these experiments the users were shown the pages from a whole album. Each of them was asked to choose a page at random and identify his/her query region (or object of interest) with a bounding box cropping. The user expertise was expected to be standard, typically having no specialized knowledge of the specific research domain, e.g., the University administrative stuffs. For each participating user, this process of query collection was iterated multiple times to select only one randomly chosen instance of all the prominent queries in a given comic album. Total 43 queries were used for evaluation

TABLE VI
PERFORMANCE OF THE PROPOSED FRAMEWORK (USING SPP [47] AND CDVS [45] AS THE APPEARANCE DESCRIPRTORS) VS BASELINE REPORTED BY RIGAUD *et al.* [52] ON THE EBDTHEQUE DATASET USING PRECISION/RECALL AS THE ACCURACY MEASURES

Comic Album (No. of Pages/No. of Char. used for exp.)	Rigaud et al.[52]	Sketch-a-Net ($\mathbf{f}8$)	Sketch-a-Net ($\mathbf{f}7$)	Full System ($\mathbf{f}7$+SPP)	Full System ($\mathbf{f}7$+ CDVS)
MIDAM GAMEOVER (10/2)	40.9/92.05	75.0/100.0	71.0/100.0	87.0/100.0	85.0/100
CYB COSMOZONE (5/4)	48.07/97.77	53.33/89.33	56.33/96.0	72.0/100.0	67.0/94.44
CYB MAGICIENLOOSE (1/1)	85.7/85.7	84/100	86/100	92/100	95.0/100
CYB MOUETTEMAN (4/2)	51.4/75.0	61.5/86.5	69.0/86.5	93.5/92.5	87.0/91.67
LAMISSEB LESNOEILS1 (5/2)	20.0/85.9	63.0/95.0	67.0/95.0	74.0/95.0	77.00/91.08
TRONDHEIM LES TROIS CHEMINS 005 (2/4)	31.25/91.28	75.0/75.0	73.5/72.5	90.0/100	88.5/84.49
MIDAM KIDPADDLE7 (2/2)	43.25/90.9	27.5/54.99	27.5/54.99	66.0/89.47	73.0/84.21
7 Album Average	45.79/88.37	62.76/85.83	64.33/86.42	**82.07/97.02**	81.79/92.27
SAINTOGAN_PROSPER* (5/5)	-	42.0/100	38.0/80.0	65.0/93.25	-
ROUDIER* (4/2)	-	41.0/61.43	40.0/65.0	63.0/85.71	-
CYB_TRAFFIC* (3/2)	-	51.5/70.0	57.0/73.33	73.5/73.33	-
10 Album Average	-	**57.38**/82.87	**58.23**/82.33	**77.69/93.14**	-

The asterisk (*) marked albums are not very rich in color and the results in these albums are not reported by Rigaud *et al.* [52].

process, a few queries which were not cropped properly or not capturing an object region (fully/partially) were removed from the query collection.

The proposed framework was evaluated on 41 comic pages from 10 different albums of eBDtheque representing 26 comic characters appearing more than 400 times in total. In order to evaluate the Sketch-a-Net object descriptor, we report the performances at 3 levels, e.g., the retrieval performance of the $\mathbf{f}8$, $\mathbf{f}7$ and finally the performance of the full system. Table VI compares the details of our experimental findings against Rigaud *et al.* [52] treated as the baseline. In an ideal case, we expect a higher precision at a higher recall. As can be seen from the table that the $\mathbf{f}8$ Sketch-a-Net descriptor achieves the precision/recall scores 62.76/85.83, which is like gaining around 17% in precision at the cost of just 3% drop in the recall, as compared to 45.79/88.37 obtained by Rigaud *et al.* [52]. Compared to $\mathbf{f}8$, $\mathbf{f}7$ improves both the precision and recall scores on average. Finally, using $\mathbf{f}7$ as the feature, the entire proposed framework is able to attain an impressive 82.21/97.02 precision/recall scores averaged over all the 7 albums. Important to note that the proposed sketch based retrieval framework remains pretty effective also for those albums where the color information are minimum and color based feature descriptors proposed by Rigaud *et al.* [52] will not be applicable. The third example in Fig. 3 shows the localization performance on the ROUDIER_LESTERRESCREUSEES album.

Unlike eBDtheque, Manga109 albums have even lesser amount of color features present within its images, which are mainly black/white sketch-like in nature. We have experimented on 9 albums from this collection and the results are shown in Table VII. As such GF-HoG [7], one of the recent proposed descriptors for Sketch based search, does not show a lot of promise in such scenarios. Compared to GF-HoG, SIFT [54] works somewhat better. In such cases, query objects also having a similar content (unlike matching a pure hand-drawn sketch query in the database of natural images as typically handled in a SBIR problem) to that of the database images, SPP shows good results. Finally, a combination of Sketch-a-Net with SPP works even better.

TABLE VII
PERFORMANCE EVALUATION USING AVERAGE $Precision@n$ SCORES COMPUTED FOR 9 ALBUMS IN MANGA109 DATASET

Features \ Number of Top Ret. (n)	10	20	30	40	50
GF-HoG [7] (500 words Vocab)	0.37	0.35	0.37	0.34	0.34
GF-HoG [7] (1000 words Vocab)	0.31	0.28	0.29	0.28	0.28
GF-HoG [7] (1500 words Vocab)	0.38	0.34	0.30	0.34	0.33
SIFT [54] (500 words Vocab)	0.55	0.51	0.52	0.47	0.48
SIFT [54] (1000 words Vocab)	0.52	0.46	0.41	0.41	0.38
SIFT [54] (1500 words Vocab)	0.52	0.52	0.48	0.46	0.45
CDVS Global [45]	0.86	0.81	0.62	0.58	0.56
Spatial Pyramid Pooling Deep Feature [47]	0.87	0.83	0.68	0.70	0.65
Proposed Framework (Sketch-a-Net and SPP)	**0.94**	**0.91**	**0.87**	**0.82**	**0.7**

V. CONCLUSION

In this work we address the problem of sketch (or sketch like object) based object search in a real life image or video. A multi-view clustering approach to identify a shortlist of spatially distinct proposals is also capable of handling the content level (spatial and structural) differences among proposals within a single formulation. The proposed graph-based re-ranking strategy that utilizes both structural and appearance based similarity information shows much improved discriminative characteristics in terms of identifying the outliers. Object Regions identified by 'proposals' are leveraged to localize objects of varying sizes within a clutter intensive real-life image. Unlike traditional methods which explore a coarse image-level pairwise similarity, the search is designed to exploit the similarity measures at the object level for a salient localization performance. Its generalization ability to handle the search task in several databases from different domains within a single framework shows significant promise. We look forward to evaluate its applicability in more intensive multi-modal scenarios with missing information.

REFERENCES

[1] C. Yan *et al.*, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 284–295, Jan. 2018.

[2] C. Yan *et al.*, "Effective uyghur language text detection in complex background images for traffic prompt identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 220–229, Jan. 2018.

[3] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-based image matching using angular partitioning," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 35, no. 1, pp. 28–41, Jan. 2005.

[4] X. Sun, C. Wang, C. Xu, and L. Zhang, "Indexing billions of images for sketch-based retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 233–242.

[5] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "A descriptor for large scale image retrieval based on sketched feature lines," in *Proc. 6th Eurographics Symp. Sketch-Based Interfaces Model.*, 2009, pp. 29–36.

[6] T. Furuya and R. Ohbuchi, "Visual saliency weighting and cross-domain manifold ranking for sketch-based image retrieval," in *Proc. Int. Conf. MultiMedia Model.*, 2014, vol. 8325, pp. 37–49.

[7] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Comput. Vis. Image Understanding*, vol. 117, no. 7, pp. 790–806, 2013.

[8] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[9] S. Wang, J. Zhang, T. X. Han, and Z. Miao, "Sketch-based image retrieval through hypothesis-driven object boundary selection with HLR descriptor," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1045–1057, Jul. 2015.

[10] J. Meng, J. Yuan, J. Yang, G. Wang, and Y. P. Tan, "Object instance search in videos via spatio-temporal trajectory discovery," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 116–127, Jan. 2016.

[11] C. Guérin *et al.*, "Ebdtheque: A representative database of comics," in *Proc. Int. Conf. Document Anal. Recognit.*, 2013, pp. 1145–1149.

[12] Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *CoRR*, vol. abs/1510.04389, 2015.

[13] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proc. ACM Workshops Multimedia*, 2000, pp. 51–54.

[14] A. D. Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 121–132, Feb. 1997.

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

[16] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proc. Comput. Vis. Pattern Recognit.*, 2011, pp. 761–768.

[17] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graphics*, vol. 28, no. 5, 2009, Art no. 124.

[18] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin, "SYM-fish: A symmetry-aware flip invariant sketch histogram shape descriptor," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 313–320.

[19] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[20] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[21] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *in Proc. SIGGRAPH, ACM Trans. Graphics*, 2014, vol. 31, no. 4, Art no. 44.

[22] Y.-L. Lin, C.-Y. Huang, H.-J. Wang, and W. Hsu, "3D sub-query expansion for improving sketch-based multi-view image retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3495–3502.

[23] P. M. A. Sousa and M. J. Fonseca, "Sketch-based retrieval of drawings using spatial proximity." *J. Visual Lang. Comput.*, vol. 21, no. 2, pp. 69–80, 2010.

[24] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *Proc. DAGM Conf. Pattern Recognit.*, 2010, pp. 432–441.

[25] Y. Qian, Y. Yongxin, S. Yi-Zhe, X. Tao, and T. Hospedales, "Sketch-a-net that beats humans," in *Proc. British Mach. Vis. Conf.*, 2015, pp. 7.1–7.12.

[26] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1875–1883.

[27] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graphics*, vol. 35, no. 4, Jul. 2016, Art no. 119.

[28] Y. Qi, Y. Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *Proc. IEEE Conf. Image Process.*, 2016, pp. 2460–2464.

[29] J. Kopf and D. Lischinski, "Digital reconstruction of halftoned color comics," *ACM Trans. Graphics*, vol. 31, no. 6, 2012, Art no. 140.

[30] D. Sýkora, J. Dingliana, and S. Collins, "LazyBrush: Flexible painting tool for hand-drawn cartoons," *Comput. Graphics Forum*, vol. 28, no. 2, pp. 599–608, 2009.

[31] K. Hoashi, C. Ono, D. Ishii, and H. Watanabe, "Automatic preview generation of comic episodes for digitized comic search," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1489–1492.

[32] Y. Qu, W.-M. Pang, T.-T. Wong, and P.-A. Heng, "Richness-preserving manga screening," in *Proc. SIGGRAPH ACM Trans. Graphics,* 2008, Art no. 155.

[33] C. Rigaud, J. C. Burie, J. M. Ogier, D. Karatzas, and J. V. D. Weijer, "An active contour model for speech balloon detection in comics," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, 2013, pp. 1240–1244.

[34] Y. Aramaki, Y. Matsui, T. Yamasaki, and K. Aizawa, "Interactive segmentation for manga," in *Proc. SIGGRAPH ACM Trans. Graphics,* 2014, pp. 2386–2391.

[35] Y. Matsui, K. Aizawa, and Y. Jing, "Sketch2manga: Sketch-based manga retrieval," in *Proc. IEEE Conf. Image Process.*, 2014, pp. 3097–3101.

[36] F. S. Khan *et al.*, "Color attributes for object detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 3306–3313.

[37] M. Wang, R. Hong, X. T. Yuan, S. Yan, and T. S. Chua, "Movie2comics: Towards a lively video content presentation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 858–870, Jun. 2012.

[38] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *Proc. IEEE Conf. Image Process.*, 2010, pp. 1025–1028.

[39] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[40] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1841–1848.

[41] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 1159–1166.

[42] T. Yu, Y. Wu, S. D. Bhattacharjee, and J. Yuan, "Efficient object instance search using fuzzy objects matching," in *Proc. Conf. Artif. Intell.*, 2017, pp. 4320–4326.

[43] G. Paola *et al.*, *Scientific Computing with MATLAB and Octave*. Berlin, Germany: Springer-Verlag, 2014.

[44] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit.*, 2007, pp. 1–8.

[45] L. Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 179–194, Jan. 2016.

[46] L. Y. Duan, J. Lin, J. Chen, T. Huang, and W. Gao, "Compact descriptors for visual search," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 30–40, Jul./Sep. 2014.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[48] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp 818–833.

[49] S. D. Bhattacharjee, J. Yuan, Y. P. Tan, and L. Y. Duan, "Query-adaptive small object search using object proposals and shape-aware descriptors," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 726–737, Apr. 2016.

[50] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 359–374, Sep. 2001.

[51] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 195–208, Jan. 2016.

[52] C. Rigaud, J.-C. Burie, J.-M. Ogier, and D. Karatzas, "Color descriptor for content-based drawing retrieval," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst.*, 2014, pp. 267–271.

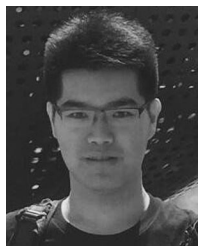
[53] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan, "A robust panel extraction method for manga," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1125–1128.

[54] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

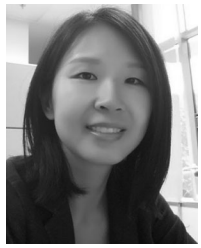
[55] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[56] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.
[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[58] K. Bozas and E. Izquierdo, "Horizontal flip-invariant sketch recognition via local patch hashing," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 1146–1150.
[59] T. Bui and J. Collomosse, "Scalable sketch-based image retrieval using color gradient features," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 1012–1019.
[60] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *Proc. DAGM Conf. Pattern Recognit.*, 2010, pp. 432–441.
[61] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (S-HELO)," in *Proc. IEEE Conf. Image Process.*, 2014, pp. 2998–3002.
[62] J. M. Saavedra, "RST-SHELO: Sketch-based image retrieval using sketch tokens and square root normalization," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 931–951, 2017.
[63] J. M. Saavedra and J. M. Barrios, "Sketch based image retrieval using learned keyshapes (LKS)," in *Proc. British Mach. Vis. Conf.*, 2015, pp. 164.1–164.11.
[64] J. M. Saavedra and B. Bustos, "Sketch-based image retrieval using keyshapes," *Multimedia Tools Appl.*, vol. 73, no. 3, pp. 2033–2062, 2014.

Sreyasee Das Bhattacharjee (M'10) received the M.Tech. degree from the Indian Institute of Technology Delhi, New Delhi, India, in 2005, and the Ph.D. degree from the Indian Institute of Technology Madras, Chennai, India, in 2013. She is currently a Researcher with the Department of Computer Science, University of North Carolina Charlotte (UNCC), Charlotte, NC, USA. Before joining UNCC, she was a Research Fellow with the School of Electrical and Electronics Engineering, Nanyang Technological University (NTU), Singapore. Prior to joining NTU, she was a Scientist with Defence Research and Development Organization (DRDO), Government of India. Her current research interests include visual object search and retrieval, natural scene understanding, machine learning, natural language processing, and multimodal data analytics. Dr. Bhattacharjee was the recipient of the National Scholarship by Govt. of India in 2000, the IBM Ph.D. fellowship in 2008, and the Australia Endeavour Research Fellowship in 2007.

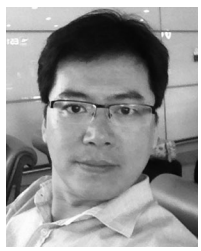
Junsong Yuan (M'08–SM'14) received the Graduate degree from the Special Class for the Gifted Young, Huazhong University of Science and Technology, Wuhan, China, in 2002, the M.Eng. degree from the National University of Singapore, Singapore, and the Ph.D. degree from Northwestern University, Evanston, IL, USA. He is currently an Associate Professor with the Department of CSE, State University of New York at Buffalo, Buffalo, NY, USA. He was an Associate Professor with Nanyang Technological University (NTU), Singapore. His research interests include computer vision, video analytics, gesture and action analysis, and large-scale visual search and mining. Dr. Yuan was the recipient of the Best Paper Award from the International Conference on Advanced Robotics (ICAR'17), the 2016 Best Paper Award of the IEEE TRANSANCTIONS ON MULTIMEDIA, the Doctoral Spotlight Award of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), the Nanyang Assistant Professorship from NTU, and the Outstanding EECS Ph.D. Thesis Award from Northwestern University. He is currently a Senior Area Editor for the *Journal of Visual Communications and Image Representations*, an Associate Editor for the IEEE TRANSANCTIONS ON IMAGE PROCESSING, the IEEE TRANSANCTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and was a Guest Editor for the *International Journal of Computer Vision*. He is the Program Co-Chair of ICME'18 and VCIP'15, and the Area Chair of ACM MM'18, ICPR'18, CVPR'17, ICIP'18'17, ACCV'18'14, etc.



Yicheng Huang received the Bachelor's degree in computer science and technology in 2015 from Peking University, Beijing, China, where he is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science. His current research interests include large-scale image retrieval and fast nearest neighbor search.



Jingjing Meng (M'09) received the B.E. degree from the Huazhong University of Science and Technology, Wuhan, China, the M.S. degree from Vanderbilt University, Nashville, TN, USA, and the Ph.D. degree from Nanyang Technological University, Singapore. She is currently a Teaching Assistant Professor with the Department of CSE, State University of New York at Buffalo, Buffalo, NY, USA. Her current research interests include big image and video data analytics, computer vision, and human–computer interaction. Prof. Meng was the recipient of Best Paper Award of the IEEE TRANSACTION ON MULTIMEDIA in 2016. She is an Associate Editor for *The Visual Computer* journal and was the Financial Chair of the IEEE Conference on Visual Communications and Image Processing (VCIP'15).



Lingyu Duan (M'06) received the M.Sc. degrees in automation and computer science from the University of Science and Technology of China, Hefei, China, and the National University of Singapore, Singapore, in 1999 and 2002, respectively, and the Ph.D. degree in information technology from The University of Newcastle, Callaghan, NSW, Australia, in 2008. He is a currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China, and has been the Associate Director of the Rapid-Rich Object Search Laboratory, a joint laboratory between Nanyang Technological University, Singapore, and PKU since 2012. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, video analytics, etc. Dr. Duan was the recipient of the *EURASIP Journal on Image and Video Processing* Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, the China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was a Co-Editor of MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13), and is serving as a Co-Chair of MPEG Compact Descriptor for Video Analytics (CDVA). He is an Associate Editor for the *ACM Transactions on Intelligent Systems and Technology* and the *ACM Transactions on Multimedia Computing, Communications, and Applications*.