Query Adaptive Instance Search using Object Sketches

Sreyasee Das Bhattacharjee[1], Junsong Yuan[1], Weixiang Hong[1], Xiang Ruan[2]
[1] School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
[2] Tiwaki Co. Ltd. Japan
dbhattacharjee, jsyuan, wxhong@ntu.edu.sg, ruanxiang@tiwaki.com

ABSTRACT

Sketch-based object search is a challenging problem mainly due to two difficulties: (1) how to match the binary sketch query with the colorful image, and (2) how to locate the small object in a big image with the sketch query. To address the above challenges, we propose to leverage object proposals for object search and localization. However, instead of purely relying on sketch features, e.g., Sketch-a-Net, to locate the candidate object proposals, we propose to fully utilize the appearance information to resolve the ambiguities among object proposals and refine the search results. Our proposed query adaptive search is formulated as a sub-graph selection problem, which can be solved by maximum flow algorithm. By performing query expansion using a smaller set of more salient matches as the query representatives, it can accurately locate the small target objects in cluttered background or densely drawn deformation intensive cartoon (Manga like) images. Our query adaptive sketch based object search on benchmark datasets exhibits superior performance when compared with existing methods, which validates the advantages of utilizing both the shape and appearance features for sketch-based search.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Image Processing, Computer Vision**]: [Scene Analysis, object recognition]

Keywords

Mobile Visual Search, Sketch-Based Object Recognition, Localization, Graph-based Search

1. INTRODUCTION

Given a query image, the goal of the object instance search is to retrieve and localize all similar objects in the database images. With the ever increasing amount of image and video data through Flickr, Facebook etc., an effective object instance search can support automatic annotation of multimedia contents and help content-based retrieval. Considering precise image example sufficing the user specification may not be always handy, 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 [1, 2, 3, 4]. Despite previous work of sketch-based image retrieval, still there are two challenges to apply it for object instance search. First of all, it is a difficult problem to match a sketch which largely abstracts the object from an original image that exhibits much richer set of information. For example, if a user is looking for some 'pyramid' images, only drawing a 'triangle' is not sufficiently discriminative to uniquely resemble the pyramids. Thus it is important to fully utilize the original image information. Second, for object instance search, it is a non-trivial problem to accurately match and locate the small objects of interest in a big image of cluttered background. Such a localization problem is not fully explored in the previous works of sketch based image retrieval.

To address the above two challenges, we propose a novel graph-based optimization framework to enable query adaptive object instance search using sketches. To match the sketch with the RGB image, we leverage recent deep learning feature such as Sketch-a-Net [5] to provide robust matching. However, considering the quality of the sketch provided by a random user may not be satisfactory, instead of purely relying on the sketch features, we also

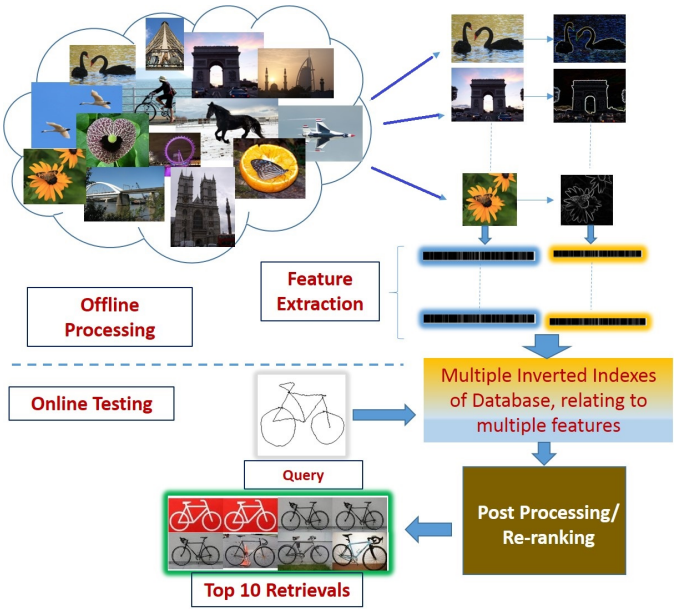

Figure 1: The outline of the Sketch-based Image Retrieval system.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2964317

explore the appearance features among images in the database to improve the search quality.

The entire process proposed in this work can be illustrated in Fig. 1. To find the object instance, each database image is represented by a collection of object proposals, which are then matched with the query sketch. After extracting the edge maps, Sketch-a-Net model is used to extract shape features while Convolution Neural Networks (CNN) applies to the original image to extract the appearance features. Our proposed query adaptive search is then formulated as a graph selection problem. Each object proposal corresponds to a graph node with its prior weight depending on the shape matching quality with the query sketch. The edges among the proposals, however, depend on the similarities of both the sketch and appearance features. A joint optimization is 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 then can well capture all the object instances at different images.

An extensive evaluation is performed on multiple benchmark datasets including Flickr15K [6], a subset of Flickr3M [7] called FlickrLarge, and eBDtheque Comic collection [8]. The superior performance compared with existing methods validates the advantages of our query adaptive object instance search, which can be summarized as follows:

- Even with a simple sketch, by exploring the appearance similarities among the object proposals in the database, it still can obtain satisfactory search results by simultaneously performing query expansion and outlier elimination via graph selection. Thanks to object proposals, it can also accurately locate the object instance in the cluttered background.
- Our graph selection formulation seamlessly fuses deep sketch and appearance features to achieve robust object instance search. As a generic formulation, it can be easily extended to other object search problems, e.g., comic character retrievals as shown in our experiments.

The rest of the paper is organized as follows: Section 2 briefly describes some related works. The detailed description of the proposed search approach is given in Section 3. Section 4 presents the experimental results. Finally the conclusion is in Section 5.

2. 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. [9] use frequency histogram of the edge orientations or Chalechale et al. [2] use a histogram representing the distribution of edge pixels as the descriptors. Bimbo & Pala [10] propose Elastic contours to define a parametric curve that is deformed appropriately to fit with the object boundary. Similarity invariant Zarnike moments [11] have also been used for matching sketches to the image. Cao et al. [12] propose Edgel index, an indexing scheme using a Chamfer Distance based descriptors for such purpose. Contour consistency filtering is performed using Shape Context descriptor. Cao et al. [1] propose SYM-FISH which incorporates the symmetry information into the shape context descriptor. However, such global descriptors [11, 13] are not very suitable for this task as they are more sensitive to deformation, occlusion and transformations etc. and thus not suitable for many generic real life problem scenario.

In order to address these, various sophisticated local descriptors have been proposed in the recent literature. Hu et al. [6, 14] use a gradient field (GF) image to represent the sketches, which is then used to compute a multi-scale HoG descriptor in a BoF model. Some other variants of HoG descriptors [15, 16] have also been proposed for the retrieval problem scenario. Wang et al. [7] 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 the majority of these literatures relies on hand crafted 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. 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, objects of interest may also be small occupying only a small portion of the entire image content. 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, we therefore 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 in the instances make the problem even more challenging. Matsui et al. [17, 18] capture the local spatial information by defining a histogram of edge orientation for a local area. 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 have their main focus on retrieval than localization (specifically for small objects) except some [19, 6, 14, 17], the proposed search strategy shows an equal expertize in retrieval and localization simultaneously. A handful of top-ranked Edgebox generated proposals identified as the candidate object regions are then investigated thoroughly by the proposed search methodology for attaining an impressive matching and localization performance.

3. 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 $\{\mathcal{I}_g\}$ of images containing 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 $\mathcal{I}_g$. 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.

3.1 Image Representation and Query Matching

EdgeBox by Zitnick & Dollar [20] is used to pick out a smaller set of candidate object regions in an image. Due to its sole dependence on the sparse yet informative edge-based representation,

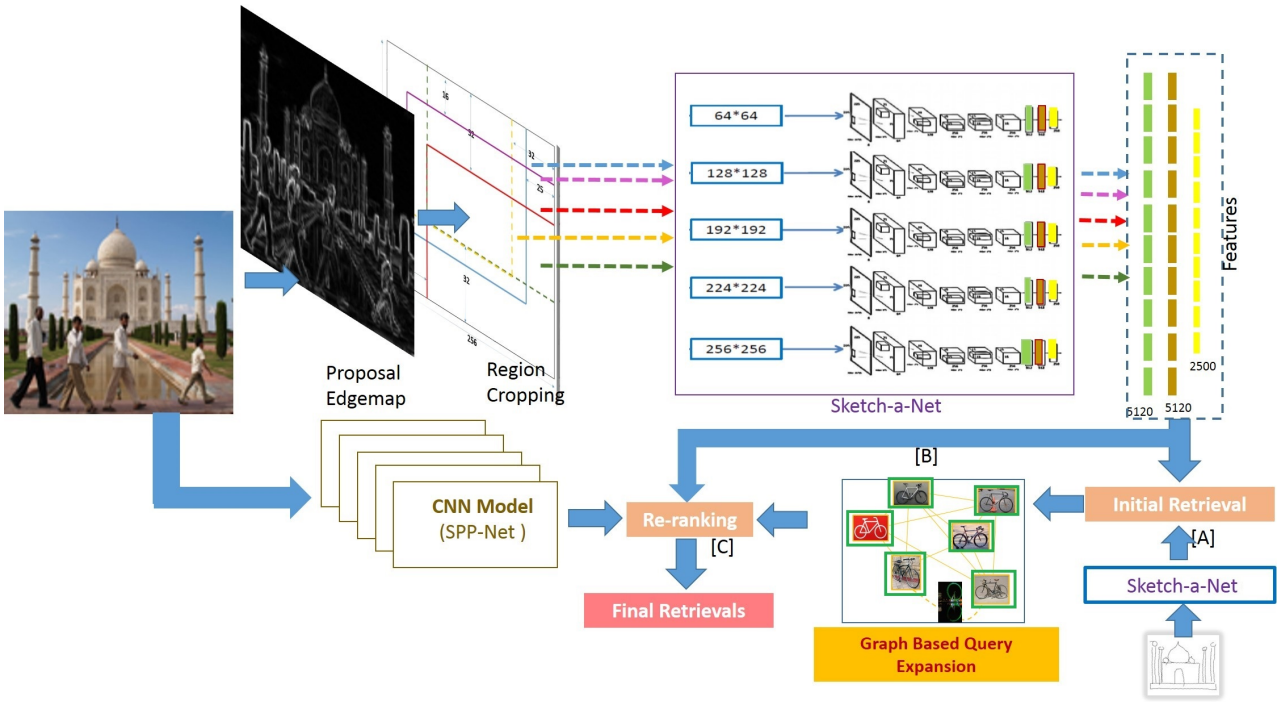

Figure 2: The Overview of the Proposed SBIR framework. Nodes highlighted with 'green' rectangles in the 'Graph Based Query Expansion' stage represents the selected nodes by our method.The steps A, B and C are the online parts of the framework, while the others can be made offline.

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 feature vector $\mathbf{p}_j \in \mathbb{R}^n$, e.g., Sketch-a-Net features. The whole image dataset is thus presented as a big pool of object proposals $\mathcal{D} = \{\mathcal{I}_i\}_{i=1}^{M} = \{P_j\}_{j=1}^{M \times N}$. Given a query object sketch Q represented by $\mathbf{q} \in \mathbb{R}^n$, our goal is to rank the proposals in the database $\mathcal{D}$ 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. $d(\mathbf{p}_i, \mathbf{p}_j) = \|\mathbf{p}_i - \mathbf{p}_j\|$.

Given the matching scores of object proposals in $\mathcal{D}$, the dissimilarity score between query Q and any image $\mathcal{I}$ is defined in terms of its best matched proposal score : $d(Q, \mathcal{I}) = \min_{P_j \in \mathcal{I}}[d(\mathbf{q}, \mathbf{p}_j)]$.

3.2 Initial Retrieval using Sketch-a-Net

The deep neural network (DNN), Sketch-a-Net [5] learnt on the TU-Berlin sketch dataset [15], is used to represent the edgemap of each database proposal (resized to a pre-defined size) in terms of a multi dimensional feature $\mathbf{p}$. Sketch-a-Net has five convolution 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 [5]. We have used three kinds of Sketch-a-Net 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 4.

As shown in the region cropping of Figure 2, 5 sub-regions are cropped from a resized proposal. Given a database proposal, the entire set of 10 d dimensional descriptors representing its 5 cropped regions (and their corresponding horizontally flipped versions) are concatenated together to obtain its $(10 \times d)$ dimensional Sketch-a-Net representative $\mathbf{p}$. In addition to the better representation scheme, another advantage of using Sketch-a-Net feature is that it can process the variability of the structural abstraction automatically with an ensemble of multi-scale model. Each module in the 5 network ensemble independently learns a model by backdrops at varying coarseness by blurring (down-sampling and up-sampling) the original image to different degrees.

3.3 Re-ranking

Give a query Q, we first find the initial retrieval set is denoted as $\mathcal{N}_Q$. 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 $\hat{K}$ retrievals are actually similar to Q. Moreover, although Q provides only a sketch based query information, an efficiently chosen subset of $\mathcal{N}_Q$ consisting of a smaller collection of some more reliable matches, can actually serve as an expanded representation for Q. Therefore, it is important to choose the subset E_Q $(\subset \mathcal{N}_Q)$, which can serve as the effective representatives for Q. In fact if properly chosen, E_Q can provide a more insightful and complete representation (in terms of both its structure and gray/color level appearances) for Q. In contrast to the common practice [21] of

choosing just a collection of handful few of top retrievals as the expanded query representation, we propose a graph based re-ranking scheme to obtain a set of more reliable entries in E_Q.

In order to capture the sufficient appearance information within the descriptor, the first step is to obtain the effective representation schemes for the database proposals.

3.3.1 Extracting Appearance Features

In the re-ranking phase, two types of appearance features are used : (1) Gabor Descriptor, (2) CNN learnt feature. While, Gabor descriptor is efficient to handle the local object structural details under varying conditions, CNN features are specialized to capture the high level semantic information. Using both these complementary features within an unified framework ensures an improved performance.

Gabor Descriptor: Each proposal in $\mathcal{D}$ is represented using their corresponding Gabor filter based response in the spatial domain. Due to its invariance to similarity transformation and various lighting conditions [22], Gabor filter is found to be advantageous for our purpose. In the spatial domain, the two dimensional Gabor filter is a Gaussian kernel function modulated by a complex sinusoid and computed as:

$$G(x,y) = \frac{f^2}{\pi\gamma\eta} exp(-\frac{x_0^2+\gamma^2 y_0^2}{2\sigma^2}) exp(j2\pi x_0 + \phi) \quad (1)$$

where $x_0 = xcos(\theta)+ysin(\theta)$ and $y_0 = -xsin(\theta)+ycos(\theta)$. f represents the sinusoid frequency, θ is the orientation, ϕ is the phase offset, σ the standard deviation and finally γ is the spatial aspect ratio which specifies the ellipticity of the support of the Gabor function. Forty Gabor filters are employed at five scales and eight uniformly sampled orientation angles. Each proposal is resized to a size 120×120 and represented in terms of a $120 \times 120 \times 8 \times 5 = 576000$ dimensional descriptor. As the image pixels in a very closeby neighborhood are usually very highly correlated, each feature map is down-sampled by a factor of 8 and the resulting reduced sized proposal representative ($\mathbf{g}_i$) has a size of $\frac{576000}{(8\times 8)} = 9000$, which are then normalized to zero mean and unit variance.

CNN Features: Convolution 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 [23] activation ($\mathbf{c}_j$), which is an aggregated descriptor obtained from the collection of deep features (generated using a fast model by Zeiler and Fergus [24]) is used to represent each proposal in $\mathcal{D}$.

3.3.2 Query Expansion via Graph Based Reranking

Given a query Q, the set of initial top-K retrieved proposals $\mathcal{N}_K \subset \mathcal{N}_Q$ can now be represented in terms of a query adaptive graph $G_Q = (\mathcal{V}_Q, E_Q, W_Q)$, where each $v \in \mathcal{V}_Q$ represents one proposal in $\mathcal{N}_K$. Given a pair of nodes $v_i, v_j \in \mathcal{V}_Q$, there is a connecting edge between them, if and only if the corresponding proposals (represented by v_i and v_j respectively) are the K-reciprocal neighbors to each other, i.e. $|\mathcal{N}_K^i \cap \mathcal{N}_K^j| \neq 0$ and $\mathcal{N}_K^i$ represents the set of top-K retrievals using P_i (represented by v_i) as query. In order to compute the reciprocal neighbors, the proposals P_i and P_j (representing $v_i, v_j \in \mathcal{V}_Q$) in $\mathcal{N}_K$ are represented in terms of their corresponding SPP-features $\mathbf{c}_i$ and $\mathbf{c}_j$. The required pairwise appearance dissimilarity between two proposals is computed using L2 distance. Each edge-weight between $v_i, v_j \in \mathcal{V}_Q$ is computed as:

$$W_Q(v_j, v_i) = \begin{cases} sim(\mathbf{p}_j, \mathbf{p}_i) sim(\mathbf{c}_j, \mathbf{c}_i) & |\mathcal{N}_K^i \cap \mathcal{N}_K^j| \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

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

Each node $v_i \in \mathcal{N}_K$ represents one of the top-K retrieved database proposals P_i using Q as a query and is tagged with a utility measure. This represents the similarity extent of P_i with Q and defined as:

$$U_i^Q = sim(\mathbf{q}, \mathbf{p}_i) \quad (3)$$

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

Given this G_Q, we propose to use a graph based re ranking scheme that can select a maximal subset of the pairwise similar proposals through graph regularization using an objective function [25], defined as follows:

$$\underset{\mathbf{b}\in\{0,1\}^K}{\arg\max} \left[\mathbf{U}_Q^T \mathbf{b} - \lambda \mathbf{b}^T L_Q \mathbf{b} - \eta ||\mathbf{b}||_0 \right] \quad (4)$$

where $\mathbf{U}_Q = [U_1^Q, ..., U_K^Q]^T \in \mathbb{R}^K$, $L_Q = D - W_Q$ and $D \in \mathbb{R}^{K\times K}$ a diagonal matrix with $D(i,i)$ representing the degree of the node i in G_Q and $\mathbf{b} \in \{0,1\}^K$ is an indicator vector specifying the inclusion/exclusion of a node in the resulting subgraph.

The first term $\mathbf{U}_Q^T\mathbf{b}$ of Eq. (4) aims at maximizing the cumulative similarity score of the selected subgraph to the query. Given the structure of the Laplacian L_Q, the focus of the second term $\mathbf{b}^T L_\mathbf{q} \mathbf{b}$ is on minimizing the outliers by emphasizing more on the edge connections with higher weights, while at the same time attempting to eliminate the nodes with larger degrees. Thus, the system is designed to reject those generic images having similarity to many others. Parameter λ controls the effect of this connectivity constraint. Finally the third term $||\mathbf{b}||_0$ acts as a regularizing factor that ensures a certain amount of sparsity of $\mathbf{b}$. η controls the effect of sparsity. In order to obtain an optimized solution of Eqn. (4), it is possible to represent both the terms in the equation in terms of the cut functions. In our implementation, we use the Boykov Kolmogorov algorithm [26] for this purpose. With an order $O(K^2en)$, where e represents the number of edges in the graph and n is the size of the minimum cut, this effectively serves our purpose. The resulting maximum flow identifies a small number of graph nodes, maximally similar within themselves, which are then re-ranked based on their assignment scores to obtain a salient expanded query representation for Q, denoted as E_Q, with a size constraint $|E_Q| < k$. Typically, k is small compared to K, e.g. $\frac{k}{K} = 0.2$.

3.4 Matching with Fusion

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

$$d_{rank}(Q, P_i) = w_i d(\mathbf{q}, \mathbf{p_i}) [\min_{j\in E_Q} [d(\mathbf{c}_j, \mathbf{c}_i)] \quad (5)$$

where $w_i = \frac{1-s(\mathbf{g},\mathbf{g}_i)}{2}$, $\mathbf{g}$ and $\mathbf{q}$ respectively represent the Gabor feature and Skatch-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 and thus w_i works only as a weighting factor

Figure 3: Given the sketch query in the left, top-10 image level retrievals using the proposed framework are shown in the right. For the second comic query, edge map of the query is extracted to be used as an input to the system.

(lying in the range $[0, 1]$) in the function above. Based on the user requirement, one can choose to omit the Gabor similarity component in the Eqn. (5) above and use $w_i = 1$. In section 4, we will discuss more on this.

The entire set of proposals in $\mathcal{N}_Q$ is reranked using $d_{rank}(,)$ to obtain the final search result. Some example localization results are shown in Figure 3.

4. EXPERIMENT

The proposed Sketch Based Image Retrieval (SBIR) 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 (2) comic character retrieval, 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 SBIR 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: SBIR being a relatively new problem, there are comparatively lesser number of benchmark results available for any dataset used in the community. The proposed SBIR method in this paper, is evaluated in three different datasets containing images of a wide range of objects with diverse structural characteristics: Flickr15K [6], a subset of Flickr3M [7] called FlickrLarge and eBDtheque Comic collection [8].

Flickr15K [6] 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.

FlickrLarge is a subset of the original Flickr3M [7], 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 200K 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.

In order to evaluate the performance for the task of comic character retrieval, eBDtheque [8] dataset is used in the experiments. eBDtheque [8] 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. [27], 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 [27], Object level precision, recall measures are used for the evaluation purpose. Given a query object image, the set of top-100 retrievals 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 correctly detected object divided by the number of objects to detect. Precision (P) measures the number of 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.

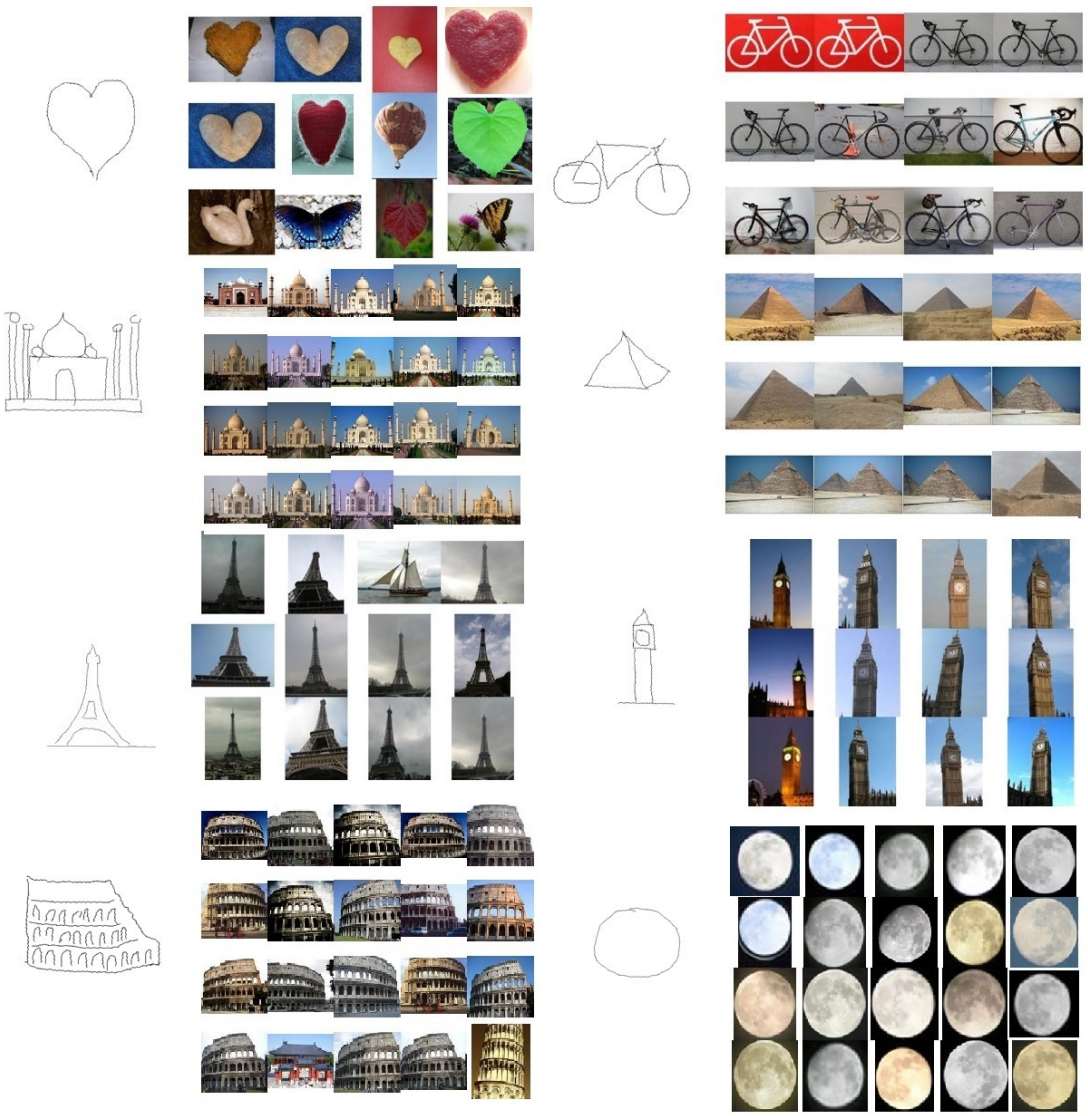

Figure 4: Some example proposal retrieval results from the Flickr15K dataset [6].

4.1 Evaluating for the SBIR task in the natural images

Flickr15K and FlickrLarge datasets are used for this part of the experiments. Since these databases are larger and the image backgrounds are cluttered, 100 top-ranked EdgeBox proposals are extracted from each image in the Flickr15K dataset. The object backgrounds being relatively cleaner, for FlickrLarge we have chosen top -20 proposals per image. For all the proposals extracted from the entire collection of images are re-sized to 256×256. Figure 4 and 5 show some visual results of the top-20 retrievals obtained by the proposed SBIR 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 Figure 3, 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 1 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, obtained after step [A] in the figure 2. As seen in the table 1, the 5120 dimensional **f6** and **f7** are found to be more effective as the descriptors compared to the more compact 2500 dimensional **f8**.

Table 1: 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

Figure 5: Some example proposal retrieval results from the FlickrLarge dataset.

Effect of Reranking: Table 2 shows the performance improvement contributed by the proposed graph based reranking in both Flickr15K and FlickrLarge dataset. It is important to note that, unlike Average Query Expansion [21] where all the top-20 retrievals are used as a representative of the query, the proposed graph based selection process aims to shortlist only those reliable matches, which are highly probable to contain the true instances of the object whose sketch is actually provided as the query. In our experiments we select the top-K (K=100) initial retrievals to further explore their mutual similarity within the proposed graph-based method and identify a smaller subset of at the most k matches ($k = 20$) , which act as the expanded representation of the query. We have used the **f7** feature as the descriptors for this set of experiments. As can be observed from the table that the proposed graph based reranking scheme contributes of around 6% gain in mAP performance score on average. The Gabor descriptors help to improve the performance a little.

Comparative Study: Table 3 reports the performance achieved by the proposed framework on the Flickr15K dataset. As seen in the table, GF-HoG [6] descriptor reports a better performance compared to other five representative features, namely shape context, structure tensor, HoG, self similarity and SIFT. Recently, Wang et al. [7] 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

Table 2: Effect of Reranking using **f6** as the proposal descriptors. The performance is evaluated using the mAP scores.

Mathod \ Dataset	Flickr15K	FlickrLarge
Initial Retrieval	0.27	0.30
Initial retrieval + Reranking (using $w_i = 1$ in Eqn. 5)	0.319	0.418
Initial retrieval + Reranking (Eqn. 5)	0.323	0.431

Table 3: 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 [28]	0.131
SIFT [29]	0.091
Self Similarity [30]	0.096
Shape Context [31]	0.081
Structure Tensor [32]	0.080
HoG [33]	0.109
GF-HoG [6]	0.139
PH-HoG [34]	0.20
HLR descriptor [7]	0.171
Color Gradient [35]	0.19
HELO [36]	0.0967
S-HELO [37]	0.1236
RST S-HELO [38]*	0.2002
Learnt Key Shapes [39]*	0.245
Spatial Pyramid Pooling Deep Feature [23]	0.0472
Object level Sketch-a-Net descriptor + Reranking	**0.323**

HLR descriptor [7] outperforms GF-HoG by showing about 4% improvement on its resulting mAP score, PH-HoG [34], 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 [39] 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 [23] is not found effective for this purpose. Thus, the detailed comparative evaluation of the proposed Sketch-a-Net based SBIR framework shows a promising insight against some of the established literatures [28, 6, 7, 39, 34, 35, 37] using a set of hand crafted features. The object level Sketch-a-Net offers a reliable description scheme, by showing a dominating performance over other state-of-the-art methods.

4.2 Evaluating for the Comic Character Retrieval

In this section, the proposed method is evaluated on the comic album dataset eBDtheque [8]. Some visual results are shown in Figure 6(a) and (b).

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 39 queries were used for evaluation process.

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 4 compares the details of our experimental findings against Rigaud et al. [27] 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. [27]. 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. [27] will not be applicable.

5. CONCLUSION

This paper addresses the problem of sketch (or sketch like object) based object search in a real life scenario. A multi-stage graph based matching strategy using both shape (Sketch-a-Net) and appearance information (CNN) extracted by its deep descriptors, can ensure a more impressive retrieval and object localization performance while still remaining discriminative to the outliers. Object Regions identified by 'proposals' are leveraged to localize small objects occupying only a small fraction of an image. Unlike popular methods exploring a coarse image-level pairwise similarity, the search is designed to exploit the similarity measures at the proposal level for a salient localization performance. Its generalization ability to deal with both the clutter intensive gray level as well as the deformation intensive sketch-like comic images within a single framework is promising. Further extensions may include its application to object instance search in video inputs.

6. ACKNOWLEDGMENTS

This research is supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2015-T2-2-114 and was carried out at the Rapid-Rich Object Search (ROSE) Lab in the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

7. REFERENCES

[1] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin, "Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor," in *Proceedings of the International Conference on Computer Vision*, 2013.

[2] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-based image matching using angular partitioning," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 1, pp. 28–41, 2004.

[3] T. Furuya and R. Ohbuchi, "Visual saliency weighting and cross-domain manifold ranking for sketch-based image retrieval," in *Proceedings of the International Conference on MultiMedia Modeling - Volume 8325*, 2014.

[4] J. M. Saavedra and B. Bustos, "Sketch-based image retrieval using keyshapes," *Multimedia Tools Appl.*, vol. 73, no. 3, pp. 2033–2062, 2014.

[5] Q. Yu, Y. Yang, Y. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net that beats humans," in *British Machine Vision Conference*, 2015.

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

Figure 6: Some example proposal retrieval results from the eBDtheque dataset [8] . To evaluate the proposed SBIR, only the edge maps of all the queries are used as an input to the system.

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

[8] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel, "ebdtheque: a representative database of comics," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2013.

[9] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proceedings of the 2000 ACM Workshops on Multimedia*, 2000.

[10] A. D. Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 121–132, 1997.

[11] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 489–497, May 1990.

[12] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proceedings of the Computer Vision and Pattern Recognition*, 2011.

[13] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[14] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *IEEE Conference on Image Processing*, 2010.

[15] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, pp. 44:1–44:10, 2012.

[16] Y.-L. Lin, C.-Y. Huang, H.-J. Wang, and W. Hsu, "3d sub-query expansion for improving sketch-based multi-view

Table 4: Performance of the proposed framework Vs baseline reported by Rigaud et al. [27] on the eBDTheque dataset using Precision/Recall as the accuracy measures. The asterisk (*) marked albums are not very rich in color and the results in these albums are not reported by Rigaud et al. [27].

Comic Album (No. of Pages/No. of Char. used for exp.)	Rigaud et al.[27]	Sketch-a-Net ($\mathbf{f8}$)	Sketch-a-Net ($\mathbf{f7}$)	Full System using $\mathbf{f7}$
MIDAM GAMEOVER (10/2)	40.9/92.05	75.0/100.0	71.0/100.0	87.0/100.0
CYB COSMOZONE (5/4)	48.07/97.77	53.33/89.33	56.33/96.0	73.0/100.0
CYB MAGICIENLOOSE (1/1)	85.7/85.7	84/100	86/100	92/100
CYB MOUETTEMAN (4/2)	51.4/75.0	61.5/86.5	69.0/86.5	93.5/92.5
LAMISSEB LESNOEILS1 (5/2)	20.0/85.9	63.0/95.0	67.0/95.0	74.0/95.0
TRONDHEIM LES TROIS CHEMINS 005 (2/4)	31.25/91.28	75.0/75.0	73.5/72.5	90.0/100
MIDAM KIDPADDLE7 (2/2)	43.25/90.9	27.5/54.99	27.5/54.99	66.0/91.67
7 Album Average	45.79/88.37	**62.76**/85.83	**64.33**/86.42	**82.21/97.02**
SAINTOGAN_PROSPER_ET_LE_MONSTRE_MARIN* (5/5)	-	42.0/100	38.0/80.0	65.0/93.25
ROUDIER_LESTERRESCREUSEES* (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**

image retrieval," in *Proceedings of the International Conference on Computer Vision*, 2013.

[17] Y. Matsui, K. Aizawa, and Y. Jing, "Sketch2manga: Sketch-based manga retrieval," in *IEEE Conference on Image Processing*, pp. 3097–3101, 2014.

[18] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," in *Proceedings of the Computer Vision and Pattern Recognition*, 2004.

[19] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 124:1–124:10, 2009.

[20] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proceedings of the European Conference on Computer Vision*, 2014.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the Computer Vision and Pattern Recognition*, 2007.

[22] J.-K. Kamarainen, V. Kyrki, and H. KaÌĹlviaÌĹinen, "Invariance properties of gabor filter-based features-overview and applications," *Image Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1088–1099, 2006.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*, 2014.

[25] 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 Transactions on Multimedia*, vol. 18, no. 4, pp. 726–737, 2016.

[26] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 359–374, 2001.

[27] C. Rigaud, J.-C. Burie, J.-M. Ogier, and D. Karatzas, "Color descriptor for content-based drawing retrieval," in *IAPR International Workshop on Document Analysis Systems (DAS)*, 2014.

[28] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 11, pp. 1624–1636, 2011.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[30] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proceedings of the Computer Vision and Pattern Recognition*, 2007.

[31] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509–522, Apr. 2002.

[32] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "A descriptor for large scale image retrieval based on sketched feature lines," in *Proceedings of the Eurographics Symposium on Sketch-Based Interfaces and Modeling*, 2009.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the Computer Vision and Pattern Recognition*, 2005.

[34] K. Bozas and E. Izquierdo, "Horizontal flip-invariant sketch recognition via local patch hashing," in *Proc. of the Int. Conf. on Acoustics Speech and Signal Processing*, pp. 1146–1150, 2015.

[35] T. Bui and J. Collomosse, "Scalable sketch-based image retrieval using color gradient features," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015.

[36] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *Proceedings of the DAGM Conference on Pattern Recognition*, 2010.

[37] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (S-HELO)," in *IEEE Conference on Image Processing*, 2014.

[38] J. M. Saavedra, "Rst-shelo: sketch-based image retrieval using sketch tokens and square root normalization," *Multimedia Tools and Applications*, pp. 1–21, 2015.

[39] J. M. Saavedra and J. M. Barrios, "Sketch based image retrieval using learned keyshapes (lks)," in *British Machine Vision Conference*, 2015.