Fine-Grained Sketch-Based Image Retrieval by Matching Deformable Part Models

Yi Li
yi.li@qmul.ac.uk

Timothy M. Hospedales
t.hospedales@qmul.ac.uk

Yi-Zhe Song
yizhe.song@qmul.ac.uk

Shaogang Gong
s.gong@qmul.ac.uk

Queen Mary University of London
London, UK, E1 4NS

Abstract

An important characteristic of sketches, compared with text, rests with their ability to intrinsically capture object appearance and structure. Nonetheless, akin to traditional text-based image retrieval, conventional sketch-based image retrieval (SBIR) principally focuses on retrieving images of the same category, neglecting the *fine-grained* characteristics of sketches. In this paper, we advocate the expressiveness of sketches and examine their efficacy under a novel *fine-grained* SBIR framework. In particular, we study how sketches enable *fine-grained* retrieval within object categories. Key to this problem is introducing a mid-level sketch representation that not only captures object pose, but also possesses the ability to traverse sketch and image domains. Specifically, we learn deformable part-based model (DPM) as a mid-level representation to discover and encode the various poses in sketch and image domains independently, after which graph matching is performed on DPMs to establish pose correspondences across the two domains. We further propose an SBIR dataset that covers the unique aspects of *fine-grained* SBIR. Through in-depth experiments, we demonstrate the superior performance of our SBIR framework, and showcase its unique ability in *fine-grained* retrieval.

1 Introduction

Sketches are incredibly intuitive to humans and descriptive in nature. They provide a convenient and intuitive way to specify object appearance and structure. As a query modality, they offer a degree of precision and flexibility that is missing in traditional text-based image retrieval – a sketch speaks for a 'hundred' words. Closely correlated with the explosion in the availability of touch-screen devices, sketch-based image retrieval (SBIR) [2, 3, 11, 12, 13, 17, 18, 19, 24] has become an increasingly prominent research topic in recent years. However, to date the main focus has been on retrieving images of the same category, overlooking an important property of sketches — they can capture *fine-grained* variations of objects such as pose (standing vs. sitting) and iconic pattern (textures on a cow's body). By further leveraging this descriptive power of sketches, in this paper, for

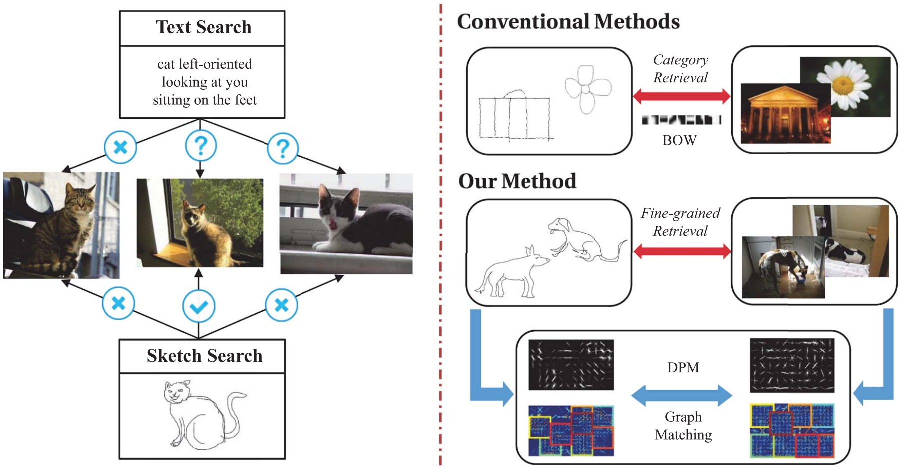


Figure 1: Comparison of traditional text-based image retrieval, conventional SBIR, and the proposed *fine-grained* SBIR framework.

the first time we introduce *fine-grained* SBIR. That is to study how sketches can be used to differentiate *fine-grained* variations of objects for retrieval, specifically pose variations. We examine how *fine-grained* knowledge extracted from sketches can be used to rank images from the same object category according to pose similarity. Figure 1 contrasts text-based image retrieval and conventional SBIR with our proposed *fine-grained* SBIR.

Key challenges for conventional SBIR include but are not limited to: (i) sketches and images are from inherently heterogeneous domains, e.g., sparse black and white line drawings versus dense colour pixels; (ii) sketches are often highly abstract in representation compared with images, e.g., image of a person can be drawn as a stick-man; (iii) cluttered backgrounds commonly captured in natural images that are not exhibited in sketches; and more importantly, moving towards *fine-grained* SBIR, (iv) a representation is needed that captures semantic *fine-grained* details such as object pose across the two domains.

Most SBIR solutions [2, 3, 11, 12, 13, 18, 19, 24] mainly focus on the first challenge. They usually proceed by first converting images into edge maps that are then directly compared against sketches, e.g., via a bag-of-words [21] representation. Few have addressed the abstractness challenge by introducing higher-level representations. Moreover, most existing work simplifies the problem by working with images having plain background with dominant objects in the center, thus reducing its breadth of applicability to realistic images. This paper aims to address all four challenges, placing particular focus on the last two. We argue that (i) object detection is necessary to address the cluttered background, (ii) a mid-level representation that encodes object parts and their geometric relationships is mandatory for pose alignment.

We propose a *fine-grained* SBIR framework that addresses the identified challenges by first learning a mid-level semi-semantic representation independently in each domain, and then learning a flexible cross-domain correspondence at this level. In contrast to previous approaches that project both domains into a common low-level representation, the mid-level correspondence approach allows us to exploit geometric/topological and appearance similarity but without requiring implausibly detailed pixel-level correspondence. This allows the user to naturally specify a *fine-grained* variation of interest (e.g., viewpoint, body configuration) in sketch domain for retrieval in image domain.

To realize our framework, we use deformable part-based model (DPM) both as an object detector and mid-level representation with which to bridge the two domains. Pose alignment is performed via graph matching, taking account both geometry and appearance information encoded in the DPMs. Specifically, we train sketch-image retrieval by first training per-category DPMs independently for each domain, then aligning DPM-mixture components across the domains to obtain a component correspondence via graph matching. At retrieval time, we use the trained DPMs to detect both the probe sketch and all gallery images, and use the learned component alignment mapping to rank the images for the first round. Then we perform finer pose alignment on the DPM detections via graph matching to rank the image for the second round. Intuitively, the component-level matching ensures retrieved objects are in broadly the same pose/appearance as the sketch. The detection-level matching enables matching *fine-grained* details such as body configuration (e.g., limb position) attributes (e.g., fat), and individual part-features help match detailed aspects of appearance (e.g., visible claws).

We demonstrate our proposed system's performance quantitatively and qualitatively against previous bag-of-words [12, 19] and spatial-pyramid [15] based methods. To perform the evaluation, we create the first SBIR dataset for *fine-grained* retrieval by sampling sketches from the 20,000 sketch dataset [20] and images from corresponding categories in the PASCAL VOC dataset [8]. Ground-truth for sketch-image pairwise similarities within each category is carefully labeled according to four criteria for *fine-grained* similarity on a portion of the proposed dataset used for testing. This ground-truth then provides overall criterion for performance evaluation.

2 Related work

Sketch-based Image Retrieval The power of sketch to differentiate *fine-grained* variations more precisely than text could potentially lead to beneficial applications, yet is not stressed in previous studies. Most prior works [2, 3, 11, 12, 18, 19, 24] assume images with dominant objects in the center with plain background, and expect or require that the sketch object and image object have rigid location correspondence. However, this is normally not the case for realistic images and sketches. The bag-of-words (BOW) representation combined with some form of edge detection (e.g. Canny edge detector), are often employed to bridge the feature gap. Although the BOW model is effective and scalable, it is weak at distinguishing *fine-grained* pose variations as it does not represent any semantic information. [13] started to work with more practical images by proposing a bag-of-regions scheme that is essentially a hierarchal structure of detected objects. Yet inside each region the same BOW model is employed again. [17] proposed to use synthesized multi-view (view is a coarse pose) sketches to boost SBIR performance. Nevertheless, they do not emphasize on explicit sketch-image pose correspondence in the retrieval step therefore neglecting sketches' pose discrimination power. Besides, they still utilize BOW model for retrieval and hold the same assumption for images. Very recently, one study [14] exploited sketch's power to describe pose, yet is engineered for a very specific domain of humans and "stick man" sketches, which has a predefined drawing style. Therefore, their method is not easily applicable to more complicated sketches. In this work, we evaluate our framework on more challenging PASCAL VOC dataset and the 20,000 sketch dataset. The images often have cluttered backgrounds and the main object is often not central and dominant, while the sketches have more complex structure and more variations. Previous SBIR methods do not perform well on this extremely challenging dataset, but our proposed method achieves encouraging performance.

Deformable Part-based Model To bridge the sketch-image semantic gap, we employ DPM as the representation to encode pose and basic appearance in each domain. The deformable part-based model (DPM) [9] is designed for object detection and obtains state-of-the-art performance on the challenging PASCAL VOC dataset. [23, 26] have used strongly supervised DPM for human pose estimation. However, their methods need a pre-defined pose model for each specific category and extensive part annotations are mandatory, which make them non-scalable in the general case for numerous diverse categories. Therefore, we adopt the original DPM [9] to encode the poses in two domains. To bridge the DPMs from different domains, we further propose an effective graph matching method to measure the cross-domain similarity of DPMs.

Graph Matching Graph matching is widely used in computer vision applications such as object categorization [7], face recognition [25] and tracking [22]. Graph matching has the advantage of flexibly encoding topological object structure, and coping with relatively large structural deformations. There has been a great body of research to date on graph matching. Cho et al. [4] establish matches by performing random graph walk on an association graph whose nodes represent candidate matches, which is later extended to cope with node progression by iteratively examining homography projection errors [16]. Very recently, supervised learning techniques have also shown prominence towards graph matching [5, 10]. Despite offering state-of-the-art results on standard datasets, they require explicit training a priori.

3 Methodology

We start this section with introducing basic notations for deformable part-based model, followed by the formulation of our graph matching method. Given those, we finally illustrate our overall framework in detail.

3.1 Deformable Part-based Model and Notations

To use deformable part-based model (DPM), a mixture of DPM is trained from a set of images, which comprises several components and is used for detection. During detection, only one component will be triggered for one object in the image, and a corresponding DPM detection is obtained for that object. Both DPM components and detections are in the form of a two-layer structure composed of a root filter and a set of N part filters connected as a star graph (part filter represents a small portion of the root filter and has twice the resolution of the root filter; all part filters have the same size). We denote this two-layer structure as $M = (\mathbf{r}, G)$ and refer it as DPM, where $\mathbf{r} = (w, h, f)$ specifies the width w, height h and global appearance feature (HOG [6] is employed) of the root filter; and $G = (V, E, A)$ represents the star graph composed of the part filters. For the star graph G, V represents a set of nodes, E, edges, and A, attributes. More specifically, $V = \{v_i\}_{i=1}^{N} \cup c$ represents all N parts v_i and the center c that is the center of $\mathbf{r}$. Each node v_i has an associated attribute $a_i \in A$ describing appearance feature (also HOG) of v_i, and an associated edge $e_{ic} \in E$ describing the geometrical relationship between the center of v_i and c in terms of relative coordinate offset.

3.2 Graph Matching for Deformable Part-based Model

The key challenge for matching sketch with images in our approach is the computation of the distance metric between the DPMs across domains, including both DPM model components and DPM image detections. In this section, we introduce our similarity measure $S(M^R|M^T)$ between two DPMs, M^R and M^T.

Our matching objective accounts for both appearance and geometric information encoded in the DPMs, as well as both layers of representation, i.e., root filter $\mathbf{r}$ and part filter star graph G. The similarity function is defined as:

$$S(M^R|M^T) = \gamma * S_{root}(M^R|M^T) + (1-\gamma) * S_{part}(M^R|M^T) \tag{1}$$

where S_{root} is the root similarity and S_{part} is the part similarity; γ is a weighting factor balancing root and part similarities.

Root Similarity (S_{root}) Given that all part filters of a DPM share a common size, differences in root size and aspect ratio implicitly reflects pose variations. Therefore, we introduce a term to represent root filter similarity based on appearance features, sizes and aspect ratios of the root filters of M^R and M^T. We denote the root filters as $\mathbf{r}^R$ and $\mathbf{r}^T$, the widths as w^R and w^T, the heights as h^R and h^T, and the appearance features as f^R and f^T respectively. Then, the root similarity metric can be written as:

$$S_{root}(M^R|M^T) = \delta * (f^R \cdot f^T) + (1-\delta) * \exp\left(-|\frac{w^R}{h^R} - \frac{w^T}{h^T}| \cdot \frac{\max(h^R,h^T)}{\min(h^R,h^T)}\right), \tag{2}$$

where the first term represents appearance similarity (dot product is inherited from [9]), and the second term accounts for size and aspect ratio variations of the root filters. δ is a linear weighting factor balancing the significance of both terms. The appearance feature f^R and f^T are extracted after normalizing $\mathbf{r}^R$ and $\mathbf{r}^T$ to the same size.

Part Similarity (S_{part}) The part-level similarity between two DPMs depends on the unknown mapping of the parts from one DPM to another. We achieve this by finding the mapping that maximizes the overall geometrical and appearance consistency between the two DPMs' part filters. Since the part filters are organized as a star graph, we formalize this mapping task as a graph-matching problem between the part filter star graphs.

Given two DPMs M^R and M^T, their part filters are represented as star graphs $G^R = (V^R, E^R, A^R)$ and $G^T = (V^T, E^T, A^T)$. We are going to find out a set of one-to-one matchings from all the nodes in V^R to all the nodes in V^T that maximizes the overall geometrical and appearance consistency of G^R and G^T. The mutual consistency of geometrical and appearance attributes between one pair of matching candidates (v_i^R, v_a^T) and (v_j^R, v_b^T) can be described by an affinity function $W_{ia;jb} = f(a_i^R, a_j^R, e_{ic}^R, e_{jc}^R, a_a^T, a_b^T, e_{ac}^T, e_{bc}^T)$. It follows that we can construct an affinity matrix $\mathbf{W}$, whose non-diagonal element $W_{ia;jb}$ contains a pair-wise affinity between two matching candidates (v_i^R, v_a^T) and (v_j^R, v_b^T) and whose diagonal element $W_{ia;ia}$ denotes a unary affinity of one matching candidate (v_i^R, v_a^T).

If the number of parts of DPM is N, the correspondence between the parts of two DPMs can be represented by an assignment matrix $\mathbf{X} \in \{0,1\}^{N \times N}$, where $\mathbf{X}_{ia} = 1$ states that node v_i^R corresponds to node v_a^T. It can then be further substituted by its column-wise vectorized replica $\mathbf{x} \in \{0,1\}^{N \cdot N}$. Finally, the graph matching problem can be formulated as seeking an assignment $\mathbf{x}^*$ that maximizes the quadratic score function:

$$\begin{aligned} &\mathbf{x}^* = \arg\max(\mathbf{x}^T \mathbf{W}\mathbf{x}) \\ &s.t.\ \ \mathbf{x} \in \{0,1\}^{N \cdot N}, \forall i \sum_{a=1}^{N} x_{ia} \leq 1, \forall a \sum_{i=1}^{N} x_{ia} \leq 1, \end{aligned} \tag{3}$$

where the two-way constrains define a one-to-one matching from G^R to G^T. It follows that the part similarity can be calculated by :

$$S_{part}(M^R|M^T) = \mathbf{x}^{*T}\mathbf{W}\mathbf{x}^* \tag{4}$$

where $\mathbf{W}$ is the affinity matrix given by:

$$W_{ia;jb} = \max(s_{app}(m_{ia}) * s_{geo}(m_{ia}) + s_{app}(m_{jb}) * s_{geo}(m_{jb}), 0) \tag{5}$$

where $m_{ia} = (a_i^R, a_a^T, e_{ic}^R, e_{ac}^T)$ and $m_{jb} = (a_j^R, b_b^T, e_{jc}^R, e_{bc}^T)$ represent matching pair (v_i^R, v_a^T) and (v_j^R, v_b^T), respectively. $W_{ia;jb}$ denotes the overall similarity between such pairs, in which $s_{app}(m_{ia})$ denotes feature similarity, $s_{geo}(m_{ia})$ represents geometrical similarity, and they can be computed as follows:

$$s_{app}(m_{ia}) = a_i^R \cdot a_a^T \tag{6}$$

$$s_{geo}(m_{ia}) = \exp(-(e_{ic}^R - e_{ac}^T)^T S_D^{-1}(e_{ic}^R - e_{ac}^T)) \tag{7}$$

where S_D is a constant covariance matrix controlling the allowed deviation of the matched cross-domain parts and is empirically set to the normalized side length of the part of DPM. $s_{geo}(m_{jb})$ and $s_{app}(m_{jb})$ can also be calculated as above.

In principle, any graph matching algorithm that is capable of solving a binary quadratic maximisation function can be used to solve Equation 3. In this paper, we employ the method of [4] that delivers good performance for our purpose.

3.3 Algorithm Overview

The desired input of our proposed method is a sketch probe S with known category, and the output is a sequence of images from the same category ordered by their similarities with the probe S in terms of pose/appearance details. Achieving this *fine-grained* SBIR requires two major steps: (i) Training: DPM training and component alignment; (ii) Retrieval: *fine-grained* retrieval based on matching a probe sketch DPM detection with image DPM detections. Below, we refer to DPM component as M^c, and DPM detection as M^d.

DPM Training and component alignment: At this step, a mixture DPM is learned from each domain, comprising several components. We denote the mixture DPM for sketch as $L^s = \{M_i^c\}_{i=1}^U$, and mixture DPM for image as $L^p = \{M_j^c\}_{j=1}^V$. For each M_i^c, its similarities with $\{M_j^c\}_{j=1}^V$ are calculated with Eq. (1). And $\{M_j^c\}$ are rearranged in descending order of the similarities into $\{M_j^c\}_i$, which is preserved for the next step. As each component represents a coarse pose category (e.g., left, right or 45° views), this step will establish a consistent coarse pose mapping across domains.

Fine-grained Retrieval: Given the query sketch S, the mixture DPM L^s is used to generate a detection M_s^d for the sketch S, while all corresponding image detections $\{M_k^d\}_{k=1}^W$ are generated by L^p. Supposing the sketch is detected by M_i^c and the images are grouped into V groups $\{G_j\}_{j=1}^V$ according to which component M_j^c detected it (each group $G_j =$

$\{M^d_{k_j}\}^{W_j}_{k_j=1}, \sum^V_{j=1} W_j = W$), we sort $\{G_j\}$ into the same order of $\{M^c_j\}_i$ obtained in the component alignment. Graph matching is then performed again within each group G_j, to rank $\{M^d_{k_j}\}$'s similarities with M^d_s via Eq. (1), and this will ensure the consistency of the detailed part shape and appearance.

4 Experiments

In this section, we first introduce a challenging SBIR dataset with human labels that enables *fine-grained* SBIR performance to be quantified. We then use this dataset to evaluate performance of the proposed *fine-grained* SBIR framework compared to conventional baselines [2, 3, 11, 12, 13, 19, 24] employing bag-of-words (BOW) and spatial pyramid (SP).

4.1 SBIR Dataset and Annotation

We create our SBIR dataset by intersecting 14 common categories from the 20,000 sketch dataset and PASCAL VOC dataset, resulting in a new dataset of $14 \times 80 = 1,120$ sketches and $7,267$ images (made up of $14 \times n_i$ images, where n_i is the total number of images in the corresponding PASCAL category).

We divide the whole dataset into testing and training sets of the equal size. To enable quantitative evaluation, we manually annotate a subset of the testing set with exhaustive pairwise similarity ground-truth. Specifically, 6 sketches and 60 images from each category are sampled from the full testing set, and sketch-image pair has its similarity manually annotated. For each sketch-image pair ($14 \times 6 \times 60 = 5,040$ pairs in total), we score their similarity in terms of four independent criteria: (i) viewpoint (V), e.g., left or right, (ii) zoom (Z), e.g., head only or whole body; (iii) configuration (C), e.g., position and shape of the limbs; (iv) body feature (B), e.g., fat or thin. For each criterion, we annotate ($5,040 \times 4 = 20,160$ annotations in total) three levels of similarity: 0 for not similar, 1 for similar and 2 for very similar. The results in Figure 3 include some example annotations.

4.2 Experimental Settings

We compare our framework to HOG Bag-of-Words and Spatial Pyramid baselines. The settings for each model are given as follows.

Bag-of-Words Following common practice [19, 20], to compute the BOW representation, images are first converted into edge maps using Canny edge detector [1]. Both images and sketches are then scaled into a fixed size of 256×256 pixels. HOG features are generated from sketch/image patches of the size 90×90 pixels. A 51×51 grid is applied to each sketch/image, and the patches are centered in the grid intersections. A large set of n features are randomly sampled from all HOG features extracted (including both sketch and image features). Afterwards, K-means clustering is employed to cluster those n features into M clusters. A code book $V = \{u_i\}^M_{i=1}$ is formed using the mean values of the clusters. After obtaining the codebook, a feature f is represented by its distance to all the words u_i. The distance is measured by Gaussian kernel with parameter σ. We set $n = 1,000,000$, $M = 2000$, $\sigma = 0.1$ for our experiments.

Spatial Pyramid The spatial pyramid strategy [15] aims to encode the geometrical structure of BOW by partitioning the image into increasingly finer equal sub-regions (i.e., in level 1 the image has 1×1 region, and in level 2 the image has 2×2 regions) and compute the

Table 1: SBIR performance comparison for top $K = 5, 10$ retrievals: Ours, Spatial Pyramid (SP) and Bag-of-Words (BOW).

(a) $K = 5$

Top 5	Ours	SP	BOW
airplane	22.00	20.33	18.83
bicycle	11.67	13.83	13.67
standing bird	14.67	13.50	11.33
bus	24.67	10.50	10.50
car (sedan)	18.83	14.50	13.50
cat	12.17	7.67	7.50
chair	20.00	20.33	19.50
cow	19.67	14.00	13.17
table	8.67	3.33	4.33
dog	9.50	6.83	5.50
horse	31.67	7.33	4.67
motorbike	22.50	9.00	11.50
sheep	17.67	5.00	6.17
train	12.50	10.33	11.50
Average	**17.58**	11.18	10.83

(b) $K = 10$

Top 10	Ours	SP	BOW
airplane	48.17	34.00	32.33
bicycle	25.50	26.67	25.00
standing bird	26.33	25.83	25.50
bus	37.67	19.17	20.00
car (sedan)	36.50	27.00	26.33
cat	20.33	16.17	15.17
chair	38.50	33.50	31.67
cow	27.17	26.50	25.33
table	12.33	9.00	9.33
dog	20.33	11.17	11.00
horse	57.33	14.50	13.33
motorbike	38.17	20.17	20.50
sheep	23.67	11.50	12.33
train	26.67	25.33	23.50
Average	**31.33**	21.46	20.81

BOW for each sub-region. The final representation is a concatenated vector of weighted BOW from all the sub-regions. In our experiment, we use 2 levels of pyramid, and adopted the implementation of [15].

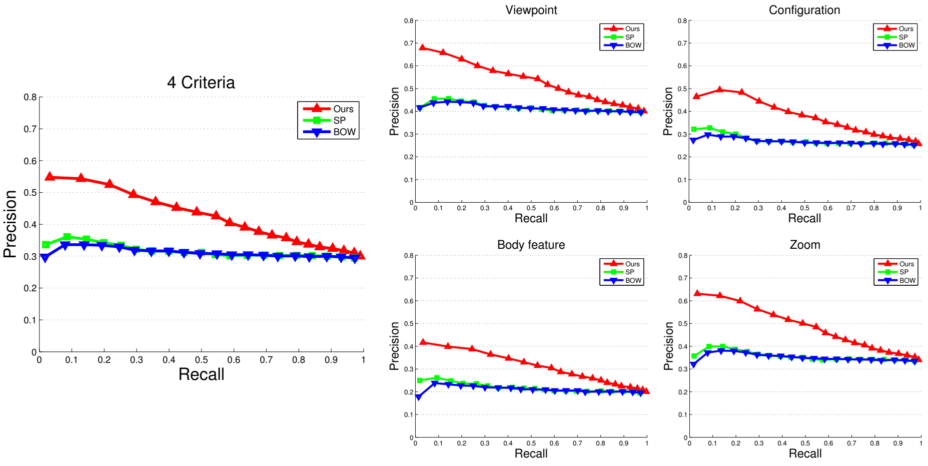
DPM training and detection We train DPMs in each domain on the full training set of sketches/images for each category, using the implementation of [9]. Each DPM is set to 3 mixture components and 8 parts per component. For each category, the sketches/images of that category are used as positive training examples while those from all remaining categories are employed as negative examples. During training, bounding boxes provided by PASCAL VOC are used to crop image objects, and sketch bounding box is extracted from the borders of the sketch object. During detection, we choose the DPM detection with the largest probability in each image.

Graph Matching Our graph matching works both on the obtained DPM components and detections. Two parameters, the root-part weight γ and the root filter appearance-geometry weight δ, are optimized by searching among $[0,1]$ with interval of 0.1 on half of the annotated dataset, and applied to the other half upon testing.

4.3 SBIR Performance Evaluation

We perform quantitative evaluation on the ground-truth dataset previously introduced in Section 4.1. Given a probe sketch, we retrieve K images, and accumulate the ground-truth similarity scores of those K images as the performance metric (the larger the better). Table 1 summarizes our results when $K = 5$ and $K = 10$. The per-category score is the average over all 3 query sketches in that category. It can be seen that our method significantly outperforms the conventional alternatives on most categories.

In Figure 2 we offer precision-recall curves computed over the full available range $K = 1:60$, utilizing all four criteria combined (Figure 2(a)) and each criterion alone (Figure 2(b)). Given an image with retrieval score S, we compute its precision as $p = S/N$, where N is the maximum score an image can have (8 in our case), and recall as $r = S/M$, where M is the accumulative image score of the entire category. The results show that our SBIR framework provides the biggest margin in its ability to perform at high-precision, suggesting that it has a much better chance of retrieving the most relevant images in the first few results. Moreover,



(a) (b)

Figure 2: Precision-recall curves comparing bag-of-words (BOW), spatial pyramid (SP), and our method (Ours), using: (a) all 4 criteria, (b) criterion viewpoint, configuration, body feature, zoom separately.

our framework is more effective at *fine-grained* SBIR under all individual criteria.

Qualitative retrieval results with ground-truth annotation are provided in Figure 3. It can be seen that our SBIR framework generally retrieves images having the same pose as the sketch query. This is because the DPM training has summarized and encoded the representative poses in the category as components, and our matching has corresponded similar representative poses from two domains.

To provide further insight into the mechanism of our model, in particular graph matching, we also demonstrate retrieval using only root similarity versus both root and part similarities. Figure 4 shows a qualitative comparison, in this case querying the entire test set rather than just the subset with ground-truth similarity annotation, as more sufficient images available for evaluation. Part-level graph-matching is illustrated in the second row by way of color coding the parts based on their sketch-image correspondence. Part similarity helps our method retrieve images with more similar *fine-grained* details (e.g., the bent legs of the running horse). Although not all the parts are perfectly aligned, their cumulative impact still helps to retrieve better matches than using the root similarity alone.

5 Conclusion

In this paper, we propose the *fine-grained* SBIR problem for the first time. It importantly recognizes the descriptive power of sketches over text and conventional SBIR where retrieval is performed at category-level only. DPMs are introduced as a novel mid-level representation strategy that captures pose information at an abstract level suitable for cross-domain mapping. Graph matching is utilized to perform pose alignment and upon retrieval to rank images. By constructing a carefully annotated cross-domain ground-truth dataset, we clearly demonstrated our system's effectiveness over conventional SBIR approaches. In the future, this work can be extended in many directions, e.g., pose discovery, retrieval metric optimisation, etc. We hope that this line of work will lead towards more practical SBIR systems suitable for realistic data and in particular for *fine-grained* retrieval: where SBIR can provide a qualitative advantage over conventional tag-based indexing and querying.

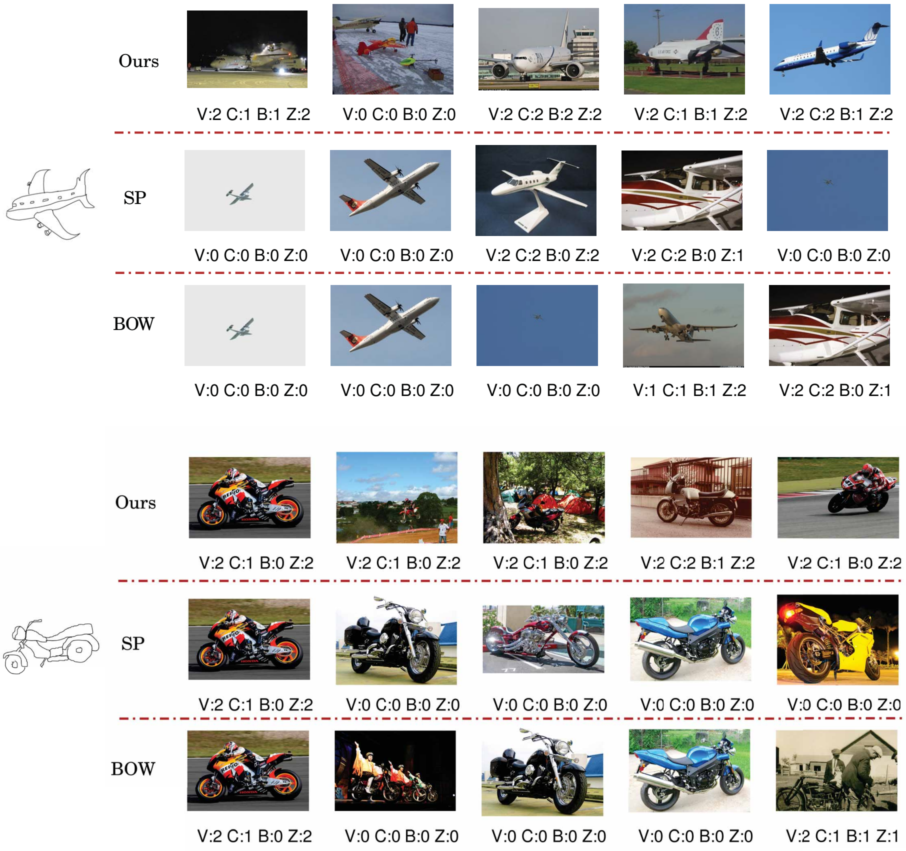


Figure 3: Two example retrievals of our method (Ours), spatial pyramid (SP) and bag-of-words (BOW). Ground truth similarity is also illustrated with the decomposition of viewpoint (V), configuration (C), body (B) and zoom (Z).

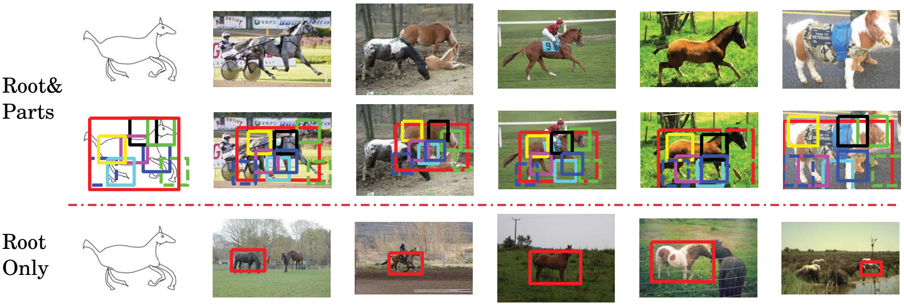


Figure 4: Comparison of retrievals using root similarity only (Root Only) and root and part similarities (Root&Parts) in graph matching.

References

[1] J. Canny. A computational approach to edge detection. *PAMI*, 1986.

[2] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *International Conference on Multimedia*, 2010.

[3] Y. Cao, C. Wang, L. Zhang, and L. Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, 2011.

[4] M. Cho, J. Lee, and K. Lee. Reweighted random walks for graph matching. In *ECCV*, 2010.

[5] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*, 2013.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *ICCV*, 2011.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[10] N. Hu, R.M. Rustamov, and L.J. Guibas. Graph matching with anchor nodes: A learning approach. In *CVPR*, 2013.

[11] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 2013.

[12] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, 2010.

[13] R. Hu, T. Wang, and J. Collomosse. A bag-of-regions approach to sketch based image retrieval. In *ICIP*, 2011.

[14] S. James, M. Fonseca, and J. Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In *ICMR*, 2014.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] K. Lee. Progressive graph matching: Making a move of graphs via probabilistic voting. In *CVPR*, 2012.

[17] Y. Lin, C. Huang, C. Wan, and W. Hsu. 3D sub-query expansion for improving sketch-based multi-view image retrieval. In *ICCV*, 2013.

[18] E. Mathias, H. Kristian, B. Tamy, and A. Marc. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 2010.

[19] E. Mathias, H. Kristian, B. Tamy, and A. Marc. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 2011.

[20] E. Mathias, H. James, and A. Marc. How do humans sketch objects? *ACM TOG (Proceedings SIGGRAPH)*, 2012.

[21] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[22] Y. Song, C. Li, L. Wang, P. Hall, and P. Shen. Robust visual tracking using structural region hierarchy and graph matching. *Neurocomputing*, 2012.

[23] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV, 2011.

[24] C. Wang, Z. Li, and L. Zhang. Mindfinder: image search by interactive sketching and tagging. In *Proceedings of the 19th international conference on World wide web*, 2010.

[25] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 1997.

[26] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.