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

Yi Li
yi.li@qmul.ac.uk
Timothy Hospedales
t.hospedales@qmul.ac.uk
Yi-Zhe Song
y.song@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Queen Mary University of London
London, UK

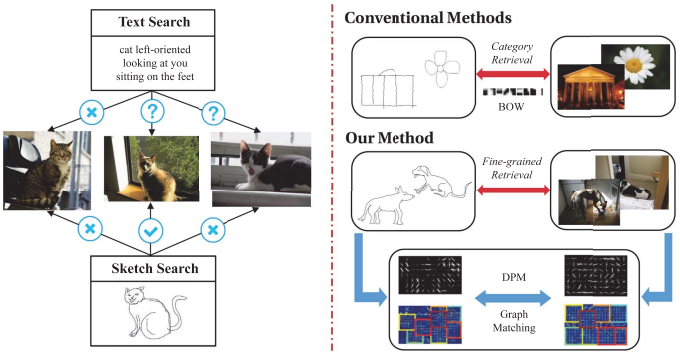

Figure 1: Comparison of traditional text-based image retrieval, conventional SBIR, and the proposed *fine-grained* SBIR framework.

Introduction Sketches are known to be able to capture object appearance and structure more intuitively and precisely than bare texts. However, to date the main focus of sketch-based image retrieval (SBIR) has been on retrieving photos 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 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. Figure 1 contrasts text-based image retrieval and conventional SBIR with our proposed *fine-grained* SBIR.

Methodology 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 photo domains. Specifically, we learn deformable part-based model (DPM) [3] to discover and encode the various poses and parts in sketch and image domains independently, and employ graph matching [1] to establishing the correspondence between DPMs from different domains. The DPM is a two-layer structure, composed of root filter and part filters. We denote DPM as $M = (\mathbf{r}, G)$, where $\mathbf{r} = (w, h, f)$ specifies the width w, height h and global appearance feature 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. Our matching objective for DPM accounts for both appearance and geometric information encoded in DPM, as well as both layers of representation, i.e., root filter $\mathbf{r}$ and part filter star graph G. Given two DPMs M^R and M^T, the similarity function is defined as:

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

where S_{root} is the root similarity and S_{part} is the part similarity; γ is a weighting factor balancing root and part similarities. The root filter similarity is generated considering appearance features, sizes and aspect ratios of the root filters, while the part similarity is solved as a graph matching problem on the part filter star graphs. 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.

Experiment We propose an SBIR dataset by intersecting 14 common categories from the 20,000 sketch dataset [4] and PASCAL VOC dataset [2]. 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. For each sketch-image pair, we score their similarity in terms of four independent criteria: (i) viewpoint (V),(ii) zoom (Z),(iii) configuration (C), (iv) body feature (B). For each criterion, we annotate 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. We compare our method with conventional bag-of-words and spatial pyramid methods, both quantitative results (Figure 2) and qualitative results (Figure 3) have demonstrated our superior performance.

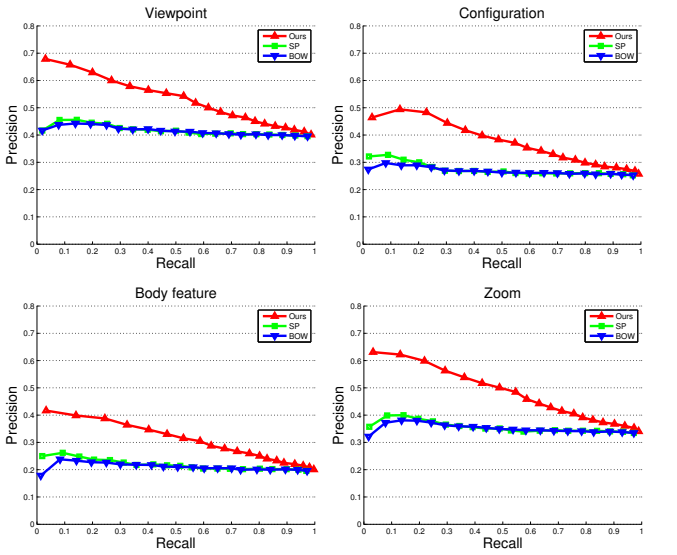

Figure 2: Precision-recall curves comparing bag-of-words (BOW), spatial pyramid (SP), and our method (Ours), using criterion: viewpoint, configuration, body feature, zoom separately.

Figure 3: 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 feature (B) and zoom (Z).

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

[2] 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.

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

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