Fine-Grained Sketch-Based Image Retrieval: The Role of Part-Aware Attributes

{Ke Li[1&2], Kaiyue Pang[1&2], Yi-Zhe Song[2], Timothy Hospedales[2], Honggang Zhang[1]}
[1]Beijing University of Posts and Telecommunications. [2] Queen Mary University of London.

Introduction

We study the problem of *fine-grained* sketch-based image retrieval (SBIR), which embodies a timely and a practical application, particularly with the ubiquitous availability of touchscreens. To address this, we propose to detect visual attributes at **part-level**, which not only captures *fine-grained* characteristics but also traverses across visual domains.

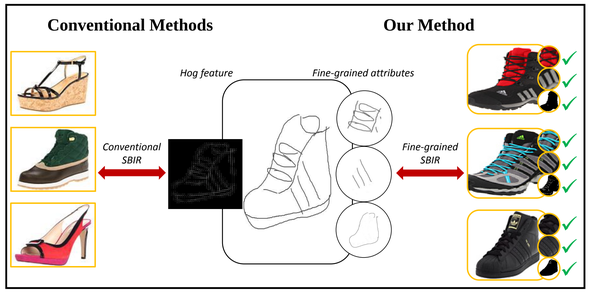


Figure 1: Conventional SBIR operates at **category-level**, but *fine-grained* SBIR requires more scrutiny at subtle details on an **instance-level** basis.

Contribution

The overall contributions of our work are:

- We propose a *fine-grained* SBIR shoe dataset with free-hand human sketches and photos, as well as *fine-grained* attribute annotations.
- We propose a **part-aware** paradigm that allows *fine-grained* attribute detection.
- We propose a synergistic low-level + mid-level + high-level feature representation that proves to be crucial to improve the performance of *fine-grained* SBIR.

Methodology

Feature and attribute extraction

- **Low-level feature** Histogram of Oriented Gradients (HOG) is extracted from shoes in both image and sketch domains.
- **Mid-level feature** A bank of relative coordinates derived from fully-connected graph model are used to represent shoe structural information.
- **Part-based attributes by Strongly-supervised DPM (SS-DPM) Model [1]** Once individual parts have been detected, these can be used to further improve attribute detection process through off-the-shelf localized attribute annotations.

Generating a synergistic combined representation

- **Three-view CCA** Inspired by [2], we use three-view CCA to lean a new space that integrates all of these cues.

$$\min_{W_1,W_2,W_3} \sum_{i,j=1}^{3} \left\| X_i W_i - X_j W_j \right\|_F^2$$
$$subject\ to \quad W_i^T \Sigma_{ii} W_i = I, \quad w_{ik}^T \Sigma_{ij} w_{jl} = 0, \quad i,j = 1, \quad \ldots, 3, i \neq j, \quad k,l = 1, \ldots, c, \quad k \neq l \tag{1}$$

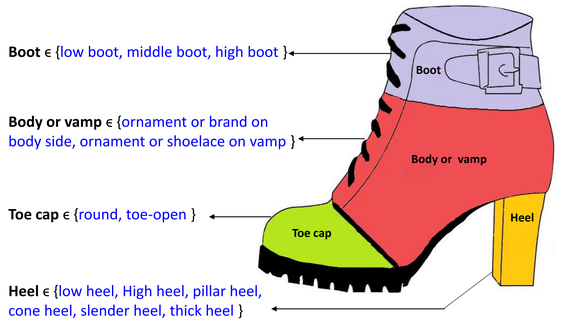
- **Using representation for *fine-grained* SBIR** Once our new robust and domain invariant representation is obtained for both sketches and images, matching a sketch $\mathbf{x}^s$ against a image dataset $D = \{\mathbf{x}_i^p\}_{i=1}^N$ is performed by nearest neighbor with L2 distance:

$$i^* = \underset{i}{\operatorname{argmin}} \left| R^s(\mathbf{x}^s) - R^p(\mathbf{x}_i^p) \right| \tag{2}$$

Fine-grained SBIR Dataset

The main contribution of our proposed *fine-grained* SBIR dataset is:

- The dataset has 304 images and 912 free-hand human sketches, with each image having three corresponding sketches.
- We define a taxonomy of 13 *fine-grained* attributes to describe each image/sketch, with each attribute associated with a localized shoe part.
- Each image/sketch has comprehensive part, attribute and bounding box annotations.



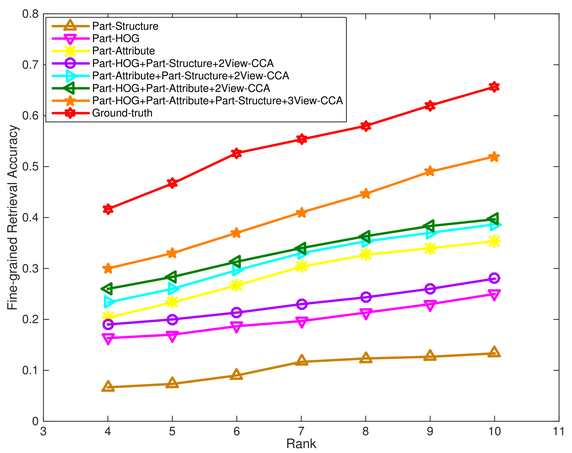
Experiments

- Attribute Detection

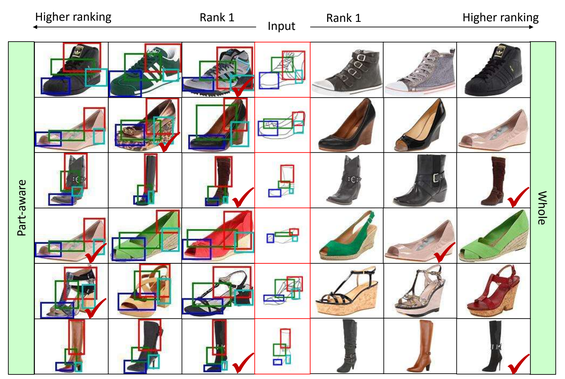
Attribute	Whole-Image	WS-Decor [3]	WS-DPM	SS-Decor [3]	Ours	Ground-truth part
Round	90.33%	88.93%	90.96%	92.08%	**93.92%**	94.25%
Toe-open	90.33%	88.93%	90.96%	92.08%	**93.92%**	94.25%
Ornament or brand on body	65.45%	61.13%	66.39%	67.47%	**70.32%**	73.85%
Shoelace or ornament on vamp	63.03%	65.38%	64.10%	64.87%	**65.98%**	70.89%
Low heel	73.72%	70.74%	74.89%	73.11%	**75.44%**	77.25%
High heel	71.19%	77.60%	72.21%	**78.90%**	73.70%	76.72%
Pillar heel	82.64%	70.91%	82.50%	72.08%	**85.13%**	88.44%
Cone heel	63.71%	69.46%	64.11%	**74.85%**	67.53%	74.64%
Slender heel	82.76%	85.29%	84.02%	**88.24%**	86.54%	89.63%
Thick heel	88.24%	76.34%	88.89%	79.97%	**91.38%**	92.83%
Low boot	96.67%	90.94%	95.42%	95.82%	**97.08%**	98.04%
Middle boot	94.39%	87.91%	92.26%	91.67%	**95.78%**	96.92%
High boot	89.10%	88.98%	86.89%	91.41%	**91.15%**	93.23%
Average	80.89%	78.66%	81.05%	81.72%	**83.68%**	86.19%

Attribute	Whole-Image	WS-Decor [3]	WS-DPM	SS-Decor [3]	Ours	Ground-truth part
Round	80.80%	78.93%	80.14%	80.30%	**81.22%**	81.96%
Toe-open	80.80%	78.93%	80.14%	80.30%	**81.22%**	81.96%
Ornament or brand on body	54.91%	53.31%	56.81%	52.95%	**60.12%**	62.34%
Shoelace or ornament on vamp	73.02%	66.90%	**74.45%**	70.96%	72.99%	73.89%
Low heel	**66.45%**	63.20%	64.89%	64.21%	66.15%	74.29%
High heel	80.46%	79.86%	79.55%	**81.24%**	75.68%	83.29%
Pillar heel	69.86%	70.91%	67.89%	72.07%	**76.00%**	77.10%
Cone heel	59.79%	60.62%	60.12%	**64.07%**	63.10%	71.66%
Slender heel	78.51%	85.95%	76.87%	**87.38%**	79.71%	88.53%
Thick heel	69.93%	71.79%	65.21%	**74.73%**	70.60%	78.83%
Low boot	**92.51%**	87.49%	87.45%	87.70%	90.87%	94.04%
Middle boot	78.11%	77.74%	72.48%	79.65%	**84.03%**	85.51%
High boot	88.65%	86.32%	84.51%	**88.98%**	84.94%	90.32%
Average	74.91%	74.00%	73.12%	75.73%	**75.89%**	80.29%

- CMC Curve



- Illustration of *fine-grained* SBIR result with and without our proposed part-aware method



References

[1] Hossein Azizpour and Ivan Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, pages 836–849. 2012.

[2] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, pages 210–233, 2014.

[3] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014.

[4] Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[5] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. Cross-modal face matching: Beyond viewed sketches. In *ACCV*, pages 210–225. 2014.