# Style Finder: Fine-Grained Clothing Style Recognition and Retrieval

Wei Di[2], Catherine Wah[1], Anurag Bhardwaj[2], Robinson Piramuthu[2], and Neel Sundaresan[2]

[1]Department of Computer Science and Engineering, University of California, San Diego
[2]eBay Research Labs, 2145 Hamilton Ave. San Jose, CA
[1]cwah@cs.ucsd.edu, [2]{wedi,anbhardwaj,rpiramuthu,nsundaresan}@ebay.com

## Abstract

*With the rapid proliferation of smartphones and tablet computers, search has moved beyond text to other modalities like images and voice. For many applications like Fashion, visual search offers a compelling interface that can capture stylistic visual elements beyond color and pattern that cannot be as easily described using text. However, extracting and matching such attributes remains an extremely challenging task due to high variability and deformability of clothing items. In this paper, we propose a fine-grained learning model and multimedia retrieval framework to address this problem. First, an attribute vocabulary is constructed using human annotations obtained on a novel fine-grained clothing dataset. This vocabulary is then used to train a fine-grained visual recognition system for clothing styles. We report benchmark recognition and retrieval results on Women's Fashion Coat Dataset and illustrate potential mobile applications for attribute-based multimedia retrieval of clothing items and image annotation.*

## 1. Introduction

Recent advances in mobile computing have redefined the dynamics of e-commerce. The ability to shop anywhere and anytime has allowed consumers to bridge the gap between offline and online stores. It has also led to recent shopping trends where users will browse for goods in offline stores and use online stores to find the best deals.

However, the effort involved in searching for a desired product among a massive collection of items remains a major bottleneck for the online shopping experience. While recent work in text retrieval has addressed some of these issues, domains such as fashion continue to present significant challenges. These challenges include the following:

- Most of the items in these domains lack useful product specifications that can be used for indexing.
- The notion of relevance in these domains is mostly visual, exposing the limitation of textual queries.
- The mobile shopping experience requires a fast and memory-efficient solution both for indexing and search.

Figure 1: Examples from the *Women's Fashion: Coat* (WFC) dataset to demonstrate the diversity of the clothing images. These items were also found to be challenging for Mechanical Turk workers to annotate with attribute labels.

Online clothing retail sites occasionally will provide image tags pertaining to attributes such as color or pattern. However, given the typically small size of their attribute vocabularies, it is difficult to sufficiently characterize the visual diversity of clothing.

Beyond colors and patterns, style is an important dimension for describing clothing. For instance, shoppers may be looking for a particular type of clothing ("peacoat") or have a certain style in mind ("a peacoat with gold buttons"). They may even have an image of a product and wish to find similar-looking products ("a coat or jacket that looks like

this"). Without style-related attribute indexing, shoppers are given few options to quickly target the desired item and are forced to continue browsing. In this work, we address the recognition and retrieval of such fine-grained clothing types and styles. We focus on a single category of clothing: women's coats and jackets. As such, the categories within this domain are fine-grained in nature, and they differ based on the presence of visual style elements or attributes.

Attribute-based approaches are attractive in a visual recognition setting, given their generalizability to different objects and efficient representation—a set of binary attributes can represent an exponential number of classes. As such, attributes are powerful high-level features that can aid in object description and characterization, their inherent semantic interpretation makes them amenable to user-centric applications such as image retrieval, where having human-understandable components for composing multimedia search queries is necessary.

Describing stylistic clothing attributes is challenging for multiple reasons, especially since clothing styles change over time and vary by season or geographic location. Several examples of images of coats and jackets are shown in Figure 1. Not only is it challenging to agree on how to describe the individual articles of clothing, it is also difficult to tell which attributes are shared among them. This problem is compounded by the fact that relevant style attributes vary across clothing domains. For example, attributes pertaining to swimming suits could greatly differ from the attributes that are used to describe blouses.

In this work, we train attribute classifiers on fine-grained clothing styles, formulating image retrieval as as a classification problem. Given an input query, which may either be text-based (a set of desired attributes), image-based, or multimodal (image with text), our proposed retrieval system returns a ranked list of related items that contains the same visual attributes as the input. This enables us to provide attribute-oriented search results that can help navigate the user towards more relevant items.

Our contributions are three-fold. First, we present the new WFC dataset of fine-grained clothing styles, focusing on the category of coats and jackets, that have been harvested from online clothing retail sites. Second, we propose an attribute-based search and retrieval schema for mobile clothing shopping. Lastly, we obtain a classification benchmark on the WFC dataset using various types of visual features. The proposed schema has multiple potential mobile applications including style-based retrieval and navigation, as well as automatic style tagging for query images.

The rest of the paper is organized as follows. In Section 2, we review related work. In Section 3, we discuss the data collection stage and how the style attribute vocabulary was constructed. In Section 4, we provide classification and retrieval benchmarks for clothing attributes. We also present examples of attribute-based image retrieval and automated image annotation as two potential mobile applications of our system. We conclude in Section 5.

| **Attribute type** | **Attributes** |
|---|---|
| Material | Fur, Denim, Leather/is leather-like, Shiny, Wool/is woolen or felt-like |
| Fastener | Zip, Button, Open, Has belt |
| Fastener style | Symmetrical (single-breasted), Asymmetrical (single-breasted), Asymmetrical (double-breasted) |
| Length | Short, Medium, Long |
| Cut | Fitted, Loose |
| Pocket | Chest pocket, Side pocket |
| Collar | V-neck collar, Round collar, Turtle neck, V-neck shirt collar, Round shirt collar, Notched collar, Shawl collar, Peak collar |

Table 1: Coat and jacket attributes in the WFC attribute vocabulary. Attribute labels are obtained using MTurk.

| **Attribute** | **Index** | **POS** | **NEG** |
|---|---|---|---|
| (Material) Fur | 1 | 250 | 1824 |
| (Material) Denim | 2 | 50 | 2024 |
| (Material) Leather/is leather-like | 3 | 249 | 1825 |
| (Material) Shiny | 4 | 177 | 1897 |
| (Material) Wool/is woolen/felt-like | 5 | 291 | 1783 |
| (Fastener) Zip | 6 | 493 | 1581 |
| (Fastener) Button | 7 | 841 | 1233 |
| (Fastener) Open | 8 | 170 | 1904 |
| (Fastener) Has belt | 9 | 305 | 1769 |
| (Fastener Style) Symmet./single | 10 | 970 | 1104 |
| (Fastener Style) Asymmet./single | 11 | 328 | 1746 |
| (Fastener Style) Asymmet./double | 12 | 265 | 1809 |
| (Length) Short | 13 | 180 | 1894 |
| (Length) Medium | 14 | 589 | 1485 |
| (Length) Long | 15 | 502 | 1572 |
| (Fit) Slim/Fitted | 16 | 413 | 1661 |
| (Fit) Loose | 17 | 332 | 1742 |
| (Pocket) Chest | 18 | 78 | 1996 |
| (Pocket) Side | 19 | 554 | 1520 |
| (Collar) V-neck collar | 20 | 66 | 1313 |
| (Collar) Round collar | 21 | 175 | 1204 |
| (Collar) Turtle neck | 22 | 261 | 1118 |
| (Collar) V-neck shirt collar | 23 | 204 | 1175 |
| (Collar) Round shirt collar | 24 | 121 | 1258 |
| (Collar) Notched collar | 25 | 190 | 1189 |
| (Collar) Shawl collar | 26 | 141 | 1238 |
| (Collar) Peak collar | 27 | 221 | 1158 |

Table 2: Positive/negative splits for the annotated attributes.

## 2. Related Work

Recent work focusing on computer vision methods applied to fashion and clothing [4, 6, 11] address the problem of clothing parsing and similarity retrieval, where the goal is to retrieve similar fashion imagery given a query. Others have studied clothing recognition in surveillance videos [12]. However, these systems generally focus on recognition or retrieval at the category level (*e.g.* suit, dress, sweater). In addition, similarity is usually measured based on distance in image feature space, rather than alignment of attributes that capture high-level semantic information regarding the image.

Some work [5, 7] has primarily focused on analyzing the relationships between general clothing attributes, with respect to human attractiveness and occasion. Bossard et


(a) (b) (c)

Figure 2: Examples of labeling interfaces used to collect attribute annotations on Mechanical Turk.

al. [2] presented a classification pipeline for categorizing upper-body clothing by appearance by utilizing visual attributes such as color, patterns and look. They also manually define general style attributes related to occasion, season, culture and lifestyle, whereas we define and use fine-grained visual style attributes.

Others [3, 9] have focused on predicting human occupations and generating descriptions from photos by observing clothing features. For example, [4] aims to address context-aware applications such as shopping recommendations and person identification by utilizing clothing information. We instead build a style-related vocabulary that represents clothing composition, rather than identifying mappings between clothing and social status.

We also note that to the best of our knowledge, there does not currently exist a suitable dataset for investigating clothing style recognition at the fine-grained level [2]. The WFC dataset we present in this work consists of standard merchandise images that are labeled with fine-grained style attributes; additional details are described in Section 3.

## 3. *Women's Fashion: Coat* Dataset (WFC)

In creating the WFC dataset, we first obtained the coats/jackets subcategories from eBay, which sells a notably diverse variety of items and organizes clothing types by style tags. We then harvested clothing images from various online retail sites which had well-aligned product images. We used the subcategory labels as search terms within the retail sites, along with other similar search terms (e.g., "moto jacket" and "biker jacket" were also used to identify motorcycle jackets).

Since clothing is highly deformable in general and can have different shapes based on how it is displayed or photographed in the image, we chose standard merchandise images with clean backgrounds, similar to images used in [1], in order to focus on style detection and attribute identification. We also filtered out images with human models and automatically cropped a tight bounding box over the clothing region. This helped to improve the alignment of images.

We discarded some eBay categories due to the lack of images found on general online retail sites, while adding the category blazer, which occurred frequently on those sites. The final WFC dataset contains 2092 images total, each labeled as one of 12 coat/jacket categories: blazer, cape, jean, military, motorcycle, parka, peacoat, poncho, puffer, raincoat, trench, vest. Each category includes a minimum of 44 images.

### 3.1. Style Attribute Vocabulary

Unlike other attribute-labeled clothing datasets [4], we are interested in attributes that are specific to coat and jacket styles. By leveraging domain-specific knowledge, we manually identified 27 binary attributes pertaining to clothing style (see Table 1 for a complete list). To our knowledge, this is the first such dataset within the fashion domain with fine-grained attribute annotations.

The attribute vocabulary is organized into groupings, and image-level attribute annotations are obtained with Amazon Mechanical Turk [1] (see examples of the labeling interfaces used in Figure 2). Each image is labeled by 5 Turkers, and the presence or absence of an attribute is determined by majority annotator agreement. The numbers of positive and negative examples for the attributes in the vocabulary are shown in Table 2. We note that the clothing images each have on average 3-4 positive attributes associated with it.

### 3.2. Attribute Cooccurrences

Figure 3 shows the cooccurrence matrix for a subset of the binary style attributes. We note that some pairs of attributes have higher cooccurrence rates. For example, *loose* clothing is more likely to co-occur with attribute *open* than with *slim cut* clothing. In addition, clothing that is *leather/leather-like* tends to also have *zip fasteners*, versus *button fasteners*. They are also more likely to be *short* or *medium* in length, rather than *long*. This attribute description of *short leather/leather-like* jackets with *zip fasteners* is characteristic of many motorcycle jackets in the dataset.

Similarly, *symmetrical/single-breasted* coats/jackets often use a *zip fastener* (*e.g.* motorcycle jackets), whereas

[1] http://www.mturk.com

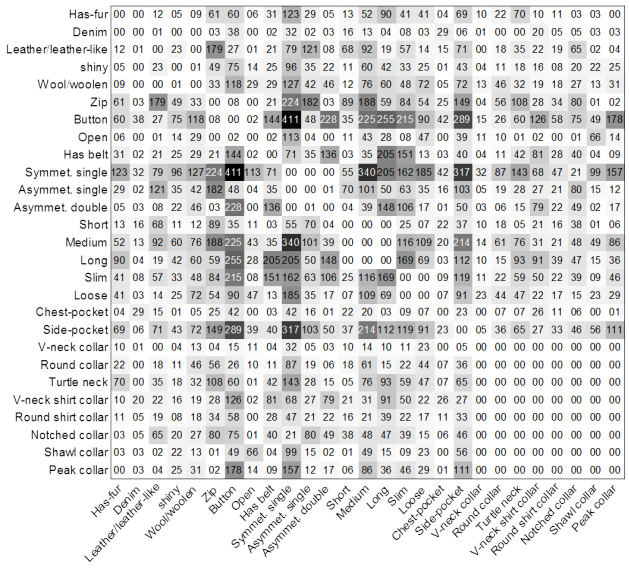


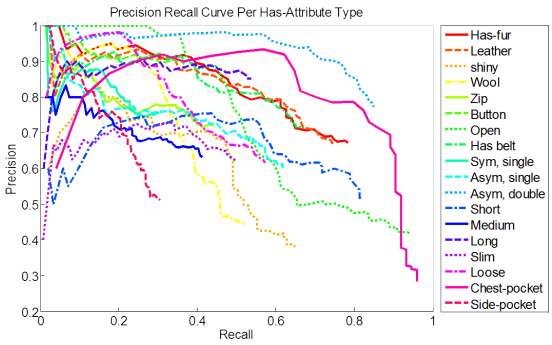

Figure 3: Attribute-attribute cooccurrence statistics.

| Attribute | PHOW | PHOW-Color | HOG | GIST | LBP | PHOW-LBP |
|---|---|---|---|---|---|---|
| has-fur | 86.7 | 86.7 | 86.7 | 63.3 | 68.3 | 80 |
| denim | 40 | 38.3 | 38.3 | 25 | 30 | 36.7 |
| leather | 80 | 80 | 60 | 53.3 | 63.3 | 83.3 |
| shiny | 70 | 66.7 | 56.7 | 60 | 56.7 | 65 |
| wool | 58.3 | 58.3 | 51.7 | 61.7 | 55 | 58.3 |
| zip | 76.7 | 75 | 70 | 73.3 | 68.3 | 76.7 |
| button | 66.7 | 60 | 61.7 | 68.3 | 50 | 66.7 |
| open | 78.3 | 83.3 | 78.3 | 68.3 | 71.7 | 83.3 |
| belt | 83.3 | 81.7 | 78.3 | 73.3 | 76.7 | 78.3 |
| sym-single | 60 | 55 | 61.7 | 53.3 | 73.3 | 63.3 |
| asym-single | 63.3 | 63.3 | 60 | 70 | 58.3 | 65 |
| asym-double | 76.7 | 80 | 75 | 71.7 | 73.3 | 76.7 |
| short | 83.3 | 85 | 85 | 80 | 80 | 88.3 |
| medium | 55 | 60 | 65 | 56.7 | 53.3 | 60 |
| long | 70 | 75 | 78.3 | 63.3 | 71.7 | 80 |
| slim | 73.3 | 75 | 65 | 53.3 | 55 | 73.3 |
| loose | 73.3 | 70 | 66.7 | 55 | 63.3 | 66.7 |
| chest pocket | 55 | 55 | 60 | 58.3 | 58.3 | 53.3 |
| side pocket | 56.7 | 63.3 | 68.3 | 58.3 | 58.3 | 66.7 |
| v-shape | 48.3 | 41.7 | 53.3 | 30 | 46.7 | 40 |
| round | 83.3 | 83.3 | 71.7 | 50 | 70 | 81.7 |
| turtle | 83.3 | 78.3 | 71.7 | 60 | 75 | 76.7 |
| v-shirt | 73.3 | 73.3 | 61.7 | 61.7 | 53.3 | 65 |
| round shirt | 56.7 | 60 | 70 | 55 | 61.7 | 63.3 |
| notched | 78.3 | 78.3 | 76.7 | 70 | 71.7 | 76.7 |
| shawl | 83.3 | 83.3 | 85 | 78.3 | 81.7 | 80 |
| peak | 86.7 | 85 | 83.3 | 78.3 | 80 | 83.3 |

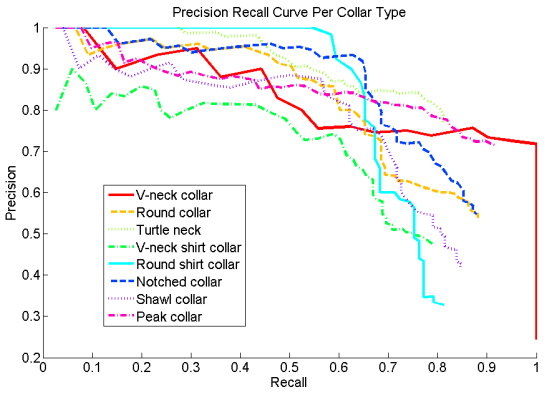
Table 3: Classification accuracy for each attribute for various feature types and feature combinations.

*asymmetrical/double-breasted* coats/jackets are more likely to have *buttons* (*e.g.* trench coats).

Although the presence of different clothing materials (shown in the top left $5 \times 5$ submatrix of Figure 3) is not mutually exclusive, the attributes still have low cooccurrence rates. Among the different materials observed, *leather/leather-like* and *shiny* are observed to have higher correlation, which is reasonable given that leather usually has a sheen to it.




(a) Binary attributes excluding collar types




(b) Collar attributes

Figure 4: Precision Recall curves for each attribute for the top 250 queries.

## 4. Experiments

In this section, we describe our experiments for producing a benchmark on attribute recognition (Section 4.1) and demonstrating various applications (Section 4.2).

### 4.1. Attribute Recognition

We performed experiments on a number of feature types: dense grayscale/color SIFT (PHOW/PHOW-Color), local binary patterns (LBP), Histogram of Oriented Gradients (HOG), and GIST [8], using the publicly available classification framework VLFeat [10]. For SIFT features, we used a bag-of-words model, training feature vocabularies with 1000 words.

We did not include color-based features such as color histograms. We note that certain coat categories may be dominated by particular color, *e.g.* most jean jackets are blue. However, color can provide only limited discriminative information in categorizing different coat styles. Because we wished to focus primarily on style-related features such as shape, cut, etc., we extracted features that relate to image texture and shape information.

We trained binary linear SVMs ($C = 10$) for each at-


| Image | Automatically Generated Description |
|---|---|
| (a) | Button, has-belt, asymmetrical and double-breasted, long, slim, v-neck shirt collar, round shirt collar |
| (b) | Wool/is woolen or felt-like, zip, loose, round collar |
| (c) | Has-fur, v-neck-collar |
| (d) | Symmetrical and single-breasted, shawl collar |
| (e) | Shiny, short |
| (f) | Has-fur, leather/is leather-like |
| (g) | Open, loose, round collar |
| (h) | Button, short, chest pocket, v-neck shirt collar |

Table 4: Examples of automatic image attribute tagging.

tribute in our attribute vocabulary; up to 60 positive examples and 60 negative examples for each attribute were randomly selected for training. We performed classification on individual feature types, as well as on joint features by combining individual features through concatenation of their histograms.

Results are reported in Table 3, which lists the average classification accuracy for the various attributes using individual features and the joint feature PHOW-LBP. Overall, we found that concatenating features did not significantly improve classification accuracy. We note that the attribute *denim* has the lowest average accuracy among all attribute classifiers. This is possibly due to the fact that *denim* refers to a detailed texture pattern which can be affected by the image resolution.

## 4.2. Applications

### 4.2.1 Attribute-Based Image Retrieval

Our proposed system enables attribute-oriented image retrieval. The user can either submit a query image and specify a target attribute group to steer the search (*e.g.* "Search for jackets with *collars* like this"), or the user can search for a particular set of requested attributes (*e.g.* "Search for jackets with a *v-shaped collar*"). This type of directed attribute-based image search enables a user to find clothing with specific requested style elements.

Given the input query, we evaluate our attribute classifiers on the query image to detect the presence of attributes in the image. The detected attributes are then used as query terms and the search engine returns a ranked list of similar images from the dataset that possess the same attributes. For example, a user can ask the system to return a set of items with the same "material" that is present in the query image. The system will first classify the material type and then retrieve items with the same type of material. The similarity metric used in ranking the images derives from the attribute classifier scores.

We present example retrieval results using multimedia (image and text) queries in Figure 5 for a subset of the clothing attributes, omitting the *denim* attribute due to lack of examples. Precision-Recall curves computed from the top 250 queries for the different attributes are reported in Figure 4.

### 4.2.2 Automatic Image Tagging

Another e-commerce application of our system is automated image annotation. Given a submitted image, the attribute classifiers can be used to identify attributes to automatically generate a suitable image description or suggest tags for indexing images. This can be a useful application for consumer-to-consumer e-commerce sites to assist sellers in creating listing descriptions. Table 4 lists the automatically generated attribute descriptions for the example images shown in Figure 1.

## 5. Conclusion

In this work, we present a new dataset of fine-grained clothing categories. We define a set of style-related visual attributes and learn individual classifiers for each style attribute. The use of attribute classifiers enables attribute-based multimedia retrieval applications, thereby allowing users to search for clothing items by attribute/stylistic element. In addition, such attribute classifiers can be used to automatically suggest image tags and generate product descriptions based on detected attributes.

Future research directions can include using transfer learning to improve the generalizability of the proposed approach for other fashion categories or for different domains. We are also interested in leveraging multiple binary visual attributes for category-level classification and developing inference methods for that purpose.

## 6. Acknowledgments

We are grateful to the anonymous reviewers who provided valuable suggestions for improving the paper.

## References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[2] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. V. Gool. Apparel classification with style. In *ACCV*, 2012.


(a) Attribute "Material": Fur, Leather, Shiny, Wool


(b) Attribute "Fastener": Zip, Button, Open, Has-belt

Figure 5: Example retrieval results for "material" and "fastener" attributes. The first image on the left is the query image, and the top 12 retrieved results are shown to the right of the query.


[3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.

[5] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM MM*, 2012.

[6] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[7] T. V. Nguyen, S. Liu, and B. Ni. Sense beauty via face, dressing and/or voice. In *MM*, 2012.

[8] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[9] Z. Song, M. Wang, X. sheng Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011.

[10] A. Vedaldi and B. Fulkerson. VLFeat. http://www.vlfeat.org/, 2008.

[11] K. Yamaguchi, H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[12] M. Yang. Real-time clothing recognition in surveillance videos. In *ICIP*, 2011.