Tagging Products using Image Classification

Brian Tomasik, Phyo Thiha, and Douglas Turnbull
Dept. of Computer Science, Swarthmore College
Swarthmore, PA 19081
btomasi1@alum.swarthmore.edu, pthiha1@alum.swarthmore.edu,
turnbull@cs.swarthmore.edu

ABSTRACT

Associating labels with online products can be a labor-intensive task. We study the extent to which a standard "bag of visual words" image classifier can be used to tag products with useful information, such as whether a sneaker has laces or velcro straps. Using Scale Invariant Feature Transform (SIFT) image descriptors at random keypoints, a hierarchical visual vocabulary, and a variant of nearest-neighbor classification, we achieve accuracies between 66% and 98% on 2- and 3-class classification tasks using several dozen training examples. We also increase accuracy by combining information from multiple views of the same product.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—*Object recognition*

General Terms: Algorithms, Experimentation

Keywords: Content-based tagging, image classification, bag of visual words

1. INTRODUCTION

Online shoppers benefit from being able to filter their search results according to product characteristics. For instance, in browsing through the women's high-heeled shoe section on Amazon.com, a consumer may wish to view only shoes with a pointy toe. This ability to refine search results is a prominent feature of specialty product-search websites such as Like.com.

Manual labeling of product traits could be expensive for large numbers of items. In order for a company like Amazon to add currently unlabeled descriptions to its products, it would likely need an automated classifier, one of whose inputs could be images of the product. The computer-vision community has made impressive advances in supervised image classification over the past several years (e.g., [1]), so we investigate how well a standard implementation of such techniques would perform on the product-tagging task. In addition to assessing raw classification accuracy, we explore how many manually annotated training examples are necessary and whether performance can be improved by using multiple image views of a single product.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

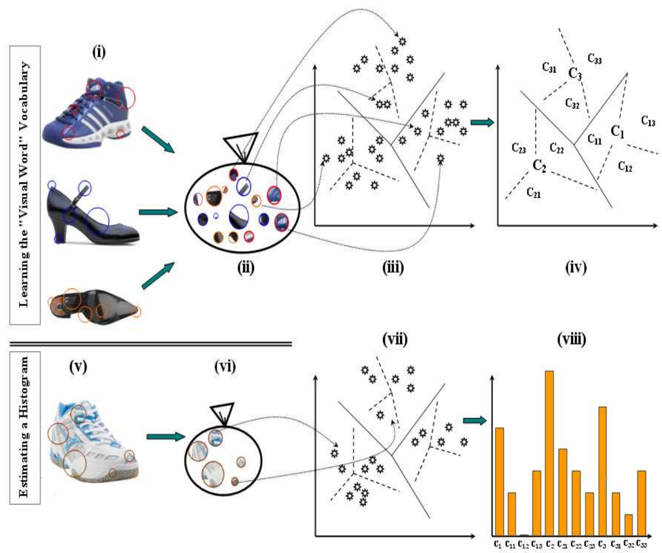

Figure 1: Illustration of the *bag of visual words* approach that we use for classification. The first row shows the process of learning a vocabulary of visual words by (i) selecting keypoints from each image, (ii) - (iii) computing SIFT descriptor vectors at those keypoints, and (iv) clustering the entire collection of SIFT descriptors into groups whose centers will define the visual words. We cluster into k groups ($k = 3$ shown, $k = 100$ used) and then recursively cluster each of those groups to create a tree of cluster centers. The second row shows how we use the visual-word tree. (v) Given an image, we (vi) again compute SIFT descriptors at keypoints and then (vii) walk each descriptor down the vocabulary tree using the closest cluster centers. Each time a descriptor walks through a cluster center, we increment the frequency count for that visual word. (viii) The result is a histogram of visual-word counts.

2. METHODS

We use a standard "bag of visual words" image classifier [2], as implemented in A. Vedaldi's open-source Matlab package [3]. The feature-extraction process is illustrated in Figure 1. In particular, we extract 10,000 SIFT features [4] from each image. We collect a subset of these features from each training image and apply hierarchical k-means clustering to construct a tree of cluster centers in SIFT-feature space [5]. Each of these vectors can be thought of as a "visual word" that characterizes an image in some way. Using this tree, we transform each of our images into a "bag of words" by associating each of the image's SIFT vectors with the words in the tree to which it is closest. The result is a histogram of frequency counts for each word, to which we can apply standard information-retrieval techniques like *term frequency-inverse document frequency* (TF-IDF) weighting and cosine similarity [2]. (For further details on the classifier, including parameter settings, see [6].)

We classify test images using a distance-weighted variant of k-nearest neighbor, in which each training image "votes" for its own category label in proportion to how much closer it is to the test image than the average training image. When we have multiple views of a test product image from different angles, as is commonly available for items like shoes on Amazon.com, we compute distances from each of these views to the training images separately and then apply our distance-weighted nearest-neighbor classifier to the entire resulting set of distances at once. Views of the product that are more informative in the sense of having smaller average distances to the training images have a bigger influence on the classification decision because of the distance weighting.

3. DATA AND RESULTS

We collected approximately 3,500 training images from the shoe and men's shirt departments of Amazon.com and manually labeled them with characteristics visible from the product images alone.[1] We created five classification problems: velcro vs. laced sneakers, pointy vs. nonpointy high-heeled shoes, short- vs. long-sleeved shirts, ballet vs. boating vs. baseball shoes, and collar vs. v-neck vs. crew shirts. For each, we report class-size-adjusted accuracies, i.e., the within-class accuracy rates averaged over each class.

Figure 2 shows mean accuracies over 5 folds of cross-validation. Performance improves with increasing numbers of product training examples from each category and with increasing numbers of image views of each product. (Where not experimentally manipulated, the number of training images and product views used was maximal.) The problems clearly vary in their difficulty, with some being relatively easy; for instance, we can distinguish short- vs. long-sleeved shirts to 90% accuracy with only 5 training images.

4. FUTURE DIRECTIONS

Given training sets containing labeled images of several dozen consumer-product items, we achieved accuracies between 66% and 98% on 2- and 3-class classification tasks. Our classifier was relatively simple, and we expect that higher accuracies could be achieved with more advanced methods (e.g., [7]). In particular, going "beyond bags of features" [8] by using local image information could be helpful, especially for items like pointy-toed shoes where the relevant visual information is contained in a small region of the image. Finally, we note that our classification performance reflects a relatively clean test set, in which each image belonged to one of the two or three main categories. A commercial system would likely need to handle miscellaneous items that don't fit any of the training-set labels.

[1]If interested in using this data set, please see `http://www.sccs.swarthmore.edu/users/09/btomasi1/tagging-products.html`

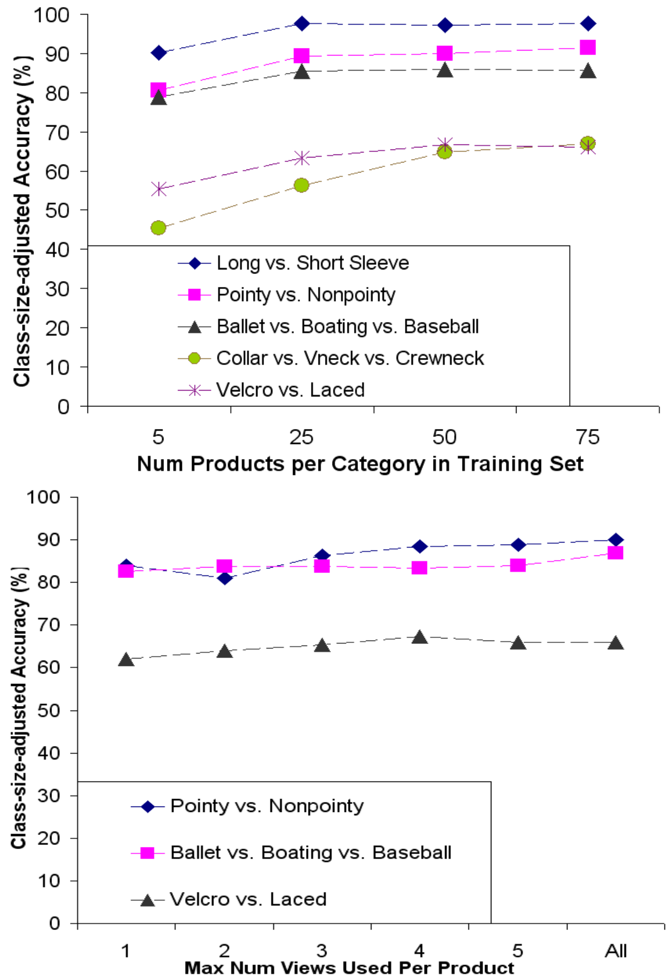

Figure 2: Effects of per-category training-set size and number of image views of each product.

Acknowledgments

We thank Lorrin Nelson and Wally Tseng at Amazon, Andrea Vedaldi, and the members of Swarthmore College Fall 2008 Senior Conference.

5. REFERENCES

[1] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical report, Caltech, 2007.

[2] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03*.

[3] A. Vedaldi. Bag of features. http://www.vlfeat.org/∼vedaldi/code/bag/bag.html

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV '04*, 60(2):91–110.

[5] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR '06*, pages 2161–2168.

[6] B. Tomasik, P. Thiha, and D. Turnbull. Tagging products using image classification. Technical report, Swarthmore College, 2009.

[7] A. Bosch, A. Zisserman, and X. Munoz. Image classification using rois and multiple kernel learning. *IJCV '08*.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06*, volume 2.