Tagging Products Using Image Classification

Brian Tomasik, Phyo Thiha and Douglas Turnbull

{btomasi1,pthiha1}@alum.swarthmore.edu, turnbull@cs.swarthmore.edu

Computer Science Department, Swarthmore College, Swarthmore, PA 19081

1. Introduction

Goal

- Develop a system to automatically **annotating products with labels**

Approach

- “Bag of visual words” image classifier
- Scale Invariant Feature Transform (SIFT)
- Hierarchical visual vocabulary
- Variant of nearest-neighbor classification

Tasks

- Classifying product images
  - Velcro vs. Lace shoes
  - Collar vs. V-Neck vs. Crew-Neck shirts
- Investigate the effect of
  - Numbers of product training examples
  - Multiple views of products in classification

2. Data Collection

- Shoe and shirt images
- ~ **3500 images** from Amazon.com and other online stores
- Labelled images with **category** and **viewpoint**

Category	Velcro	Laces	Pointy	Nonpointy	Short Sleeve	Long Sleeve
Sample image						
Num. products	75	75	77	63	150	147
Num. individual images	299	390	414	360	152	152
Avg. views per product	~ 4	~ 5	~ 5	~ 6	~ 1	~ 1

Category	Ballet	Boating	Baseball	Collar	V-Neck	Crew
Sample image						
Num. products	50	98	66	100	97	100
Num. individual images	53	453	205	100	99	105
Avg. views per product	~ 1	~ 5	~ 3	~ 1	~ 1	~ 1

Table 1: The product categories collected. The vertical lines separate the classification tasks carried out in our experiments.

3. The *Bag of Visual Words* Approach

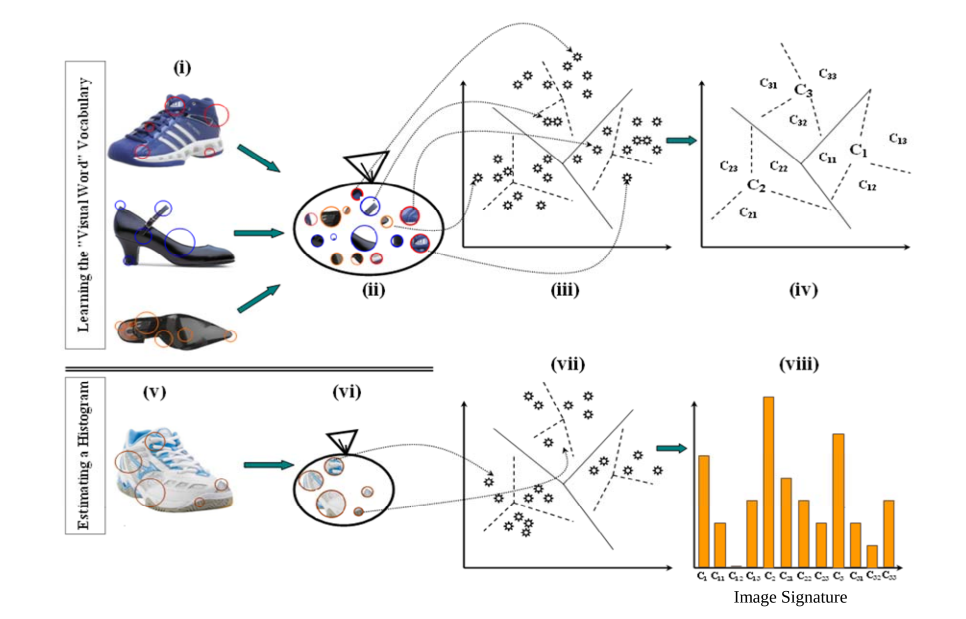

Figure 1: The first row shows the process of learning a vocabulary of **visual words** by **(i)** selecting **keypoints** from each image, **(ii) - (iii)** computing **SIFT descriptor vectors** at those keypoints, and **(iv) clustering** the entire collection of SIFT descriptors into groups whose centers 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,** namely **image signatures**.

4 (i). Image Classification

- TF-IDF for Image Signatures, s_w
- Cosine-similarity for distance measures between s_{wi} and s_{wj}
- Variant of *k*-NN namely *Z-Score Voting* where weight, w_i , is:

$$w_i = -zscore(d(s_t, s_i)) = \frac{\mu_t - d(s_t, s_i)}{\sigma_t}$$

- Keypoints 10,000 with different selection methods: Canny edge detection, Random and combined

4 (ii). Multiple Views

- Each product has multiple associated images, corresponding to multiple views of the product
- Some viewpoints available are not helpful—e.g., underside for laced vs. velcro shoes
- Solution: Use all views available by calculating distances from each view of a product from all views of other products

5 (i). Training Set Size

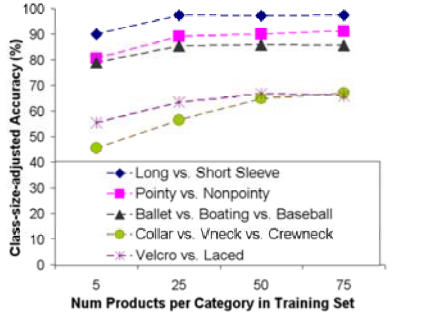

Figure 2. Class-size-adjusted accuracies saturated as we increased the number of images used in the training process for each product beyond 50.

5 (ii). Number of Views

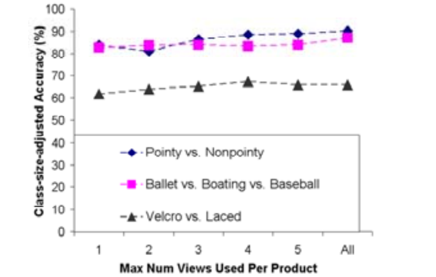

Figure 3. Class-size-adjusted accuracies improved as we increased the number of views used for some products.

References

- J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03*.
- D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV '04*, 60(2):91-110.
- D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR '06*, pages 2161-2168.
- A. Vedaldi. Bag of features.

  http://www.vlfeat.org/~vedaldi/code/bag/bag.html.
- B. Tomasik, P. Thiha, and D. Turnbull. Tagging products using image classification. Technical report, Swarthmore College, 2009. http://www.sccs.swarthmore.edu/users/09/btomasi1/tagging-products.html