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 in 2- and 3-class classification problems
- Investigate the effect of numbers of product training examples as well as that of using multiple views of products in classification

2. Data Collection

- Shoe and shirt images
- Approximately 3500 images mainly from Amazon.com and other online stores
- Labeled images in accordance with category, viewpoint and image number

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

(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 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.

Figure 1: Illustration of the Bag of Visual Words approach that we used 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 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.

4. The Bag of Visual Words Approach

(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 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.

Figure 1: Illustration of the Bag of Visual Words approach that we used 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 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.

5. (i) Image Classification

- TF-IDF for Image Signatures, s_w
- Cosine-similarity for distance measures between s_w and s
- Variant of k-NN namely Z-Score Voting where:w = \frac{Z_{score}(d(s, s_w))}{\mu - d(s, s_w)}
- Keypoints 10,000 with different selection methods: Canny edge detection, Random and combined

5. (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 of laced and velcro shoes
- Solution: Use all views available by calculating distances from each view of a product from all views of other products

6. (i) Results

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

6. (ii) Results

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

References