Food brand image (Logos) recognition

Ritobrata Sur(rsur@stanford.edu), Shengkai Wang (sk.wang@stanford.edu)
Mentor: Hui Chao (huichao@qti.qualcomm.com)
Final Report, March 19, 2014.

1. Introduction

Food label brand image (Logos) recognition has many useful applications such as displaying nutritional information and advertisement. Various methods have been proposed using local and global feature matching, however, it continues to be a challenging area. Unlike nature scene images, which are rich in texture details, brand images often lack texture variation, and therefore provide fewer key feature points for matching. In addition, reference and test images may be acquired with different resolution, size, quality, and illumination conditions. These factors combine to make logo detection more challenging.

To illustrate the basic idea let us look at the Fig. 1. Suppose a customer goes to a grocery store to pick up some food items. On the shelves there are numerous items and the person has certain dietary needs / restrictions. The customer takes a picture using their phone. The program is supposed to identify the food labels and quickly guide the user as to where their desired items are located. This program can also be used by the grocery stores to keep track of the misplaced items.

Fig. 1. Example illustrating the concept of logo recognition

The logo recognition problem has become popular from the 1990s, and ever since then there have been mainly five methods, falling into two categories: local and global. The local methods compute a number of statistical and morphological shape features for each connected component of an image foreground and background. This category includes graphical distribution features [1] and local invariants [2,3]. More recently, scale-invariant feature transform [4] (SIFT) has been used for more

effective feature extraction and mapping, and was successfully applied to vehicle logo recognition [5]. The global methods try to identify the structure of the logo image as a whole. This category covers the wavelet decomposition method [6], and neural-network-based architecture recognition [7].

Currently, for the purposes of this class, we trained our program to identify 6 different classes (logos) viz. Coca Cola, Pepsi, Dole, Quaker, Special K and Nestle. The code-words of the images were identified using the difference of Gaussian (DoG) detectors. The code-words were described using a Scale-Invariant Feature Transform (SIFT) descriptor. Then a dictionary of code-words was created by using two alternate methods: Bag of Words (BoW) and Spatial Pyramid Matching (SPM). The results using BoW was poor with smaller number of training examples but it improved significantly by increasing the training set size. The object identification was done by two distinct methods: sliding window method and keypoint matching by homography transformation and using RANSAC to remove outliers. The following sections describe the methods in greater detail.

2. Methods

The algorithm involves three main steps (Fig. 2): (a) Feature extractor, (b) Logo classifier and (c) Automatic detector. The training step uses presorted pictures corresponding to different logos to train classifiers. During detection, the features are mapped to classifiers to obtain matches within the training set. The functions and methods behind each of these steps are listed in the following sections.

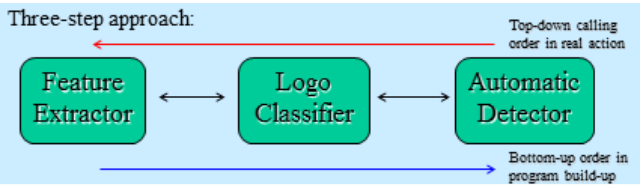

Fig. 2. The logo identification algorithm has the three main steps

(a) Feature extractor

The code-word blobs were detected using the difference of Gaussian model. The code-words thus obtained were described using SIFT [4] descriptors. An online library for SIFT detector and descriptor, "vl_sift" [8] was utilized for this project. A comprehensive and detailed description of the methods used in the library can be found in the website [8]. A brief discussion is presented in this report to guide the reader about the specific parameters used and the reasoning behind it.

As discussed in D. Lowe's article [4], the stable keypoint locations in the scale space can be efficiently and reliably extracted by using the difference of Gaussian convolution of the images in nearby scales. The same procedure is used in the "vl_sift" function [8]. Once the features are extracted, the low contrast points are filtered out below a threshold of the values of the difference of Gaussian values at the local extremum points [4, section 4]. The best value for this factor was found by trial and error. For identification of meaningful keypoints, it was also very crucial to increase the

threshold for the "r" factor for setting the edge threshold as described in the section 4.1 of D. Lowe's article [4]. This is because for logos, it is very common to have a large principal curvature across the edge but a small one in the perpendicular direction, which is otherwise rejected by the edge threshold. But this factor had to be optimized to remove noisy keypoints. A sample of the keypoint extraction is shown in Fig. 3.

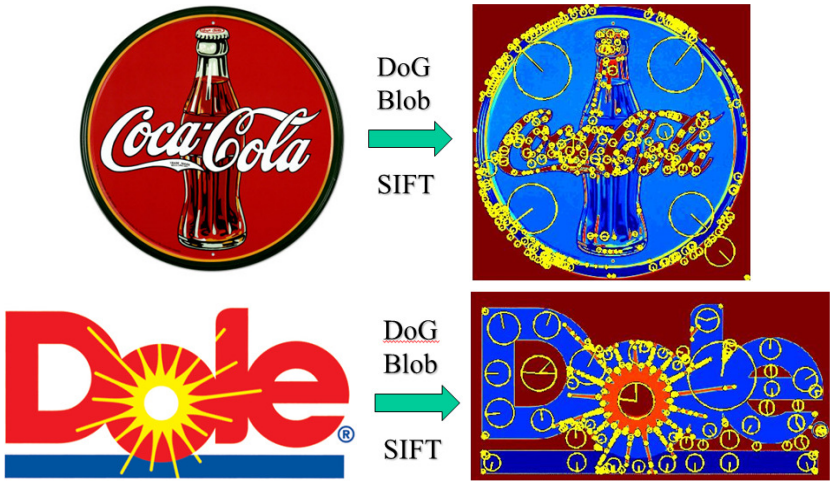

Fig. 3. Sample SIFT blob detection. The circles indicate the blob the line shows the dominant direction in the SIFT descriptor.

The SIFT descriptor [4] uses the spatial gradients in 8 directions in 4 X 4 pixel bins corresponding to the extracted features. The histograms for each of the 8 X 4 X 4 = 128 values forms the SIFT descriptor. The magnification factor (determines the size of the code word corresponding to the blob size) and the Gaussian window size (for reducing the influence of the gradients farther from the keypoint center) were key in determining the optimum parameters for describing the code-words effectively. The same descriptor parameters were used in training and test phases.

(b) Logo classifier

For a given class, to construct a dictionary of code-words was constructed by i. Bag of words and ii. Spatial Pyramid Matching.

i. ***Bag of words (BoW):*** A BoW dictionary containing 256 code-words (e.g. Fig. 4) is trained from over 30000 features extracted from 150 training images of 6 different logo classes by using K-means clustering. The size of the dictionary was optimized for speed and accuracy.

Fig. 4. Constructing a bag of words dictionary for logo identification

For classification, the following steps are implemented to identify specific logo classes:

1. Extract SIFT features
2. Project each SIFT feature onto the dictionary space to get code-word
3. Populate a histogram of feature by counting the matched code-words
4. Run classification using the histograms, either nearest neighbor (NN) or SVM

It was found that the BOW model made it extremely necessary to generate a large training set to remove spurious features. Using 5 sample images for each logo, cross-validation (leave one out) accuracy was only 33%. However when trained with 25 images for each class, a cross-validation accuracy was improved to **93%**.

ii. Spatial Pyramid Matching (SPM):

The BoW model completely ignores the spatial location of the code-words in relation to each other. This can be vital in identifying complete logos. This problem is tackled by using spatial pyramid matching technique [9] where the picture is divided into successively smaller grids and the descriptor histogram is calculated for a weighted sum of these features (Fig. 5).

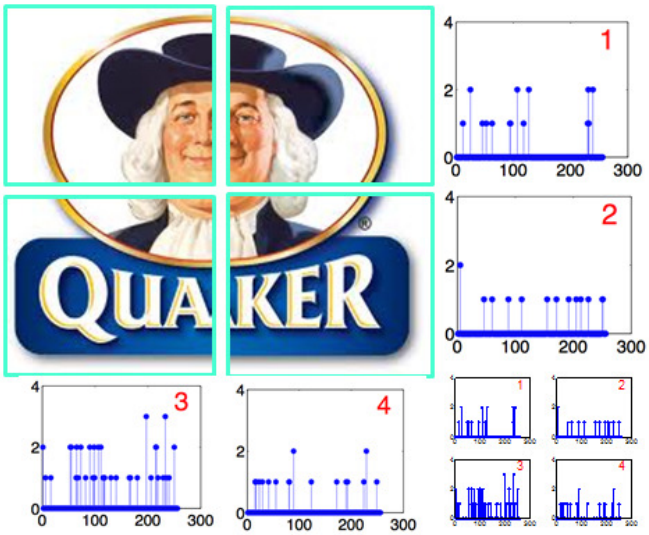

Fig. 5. Spatial pyramid matching for generating code-words

This method significantly improves the performance of the training step even with a smaller number of examples. However, it introduces dependence on the orientation of the images and slows the performance of the algorithm. Therefore it is necessary to generate artificially perturbed set of images [10] in terms of orientation, aspect ratio, blur, shake, illumination, etc. (Fig. 6) to train a robust set of code-words in all its variations.

Fig. 6. Training data orbit for robust training

(c) Automatic detector

The object identification was done by two distinct methods: i. keypoint matching by homography transformation and using RANSAC to remove outliers and ii. sliding window method.

i. Keypoint matching by homography transformation and using RANSAC [10]:

A fast method for matching logos is to directly calculate the homography pairs with the closest training image keypoints and map the training image on to the test image. Then use RANSAC to remove spurious keypoints by maximizing the number of matching pairs between the two images. Then we calculate the logo bounding box in the test image using Hough voting from the matched keypoints. This method is fast but it becomes erroneous in the presence of multiple instances of the same logo. An example is shown in Fig. 7.

Fig. 7. Logo matching using homography mapping and RANSAC filtering

ii. Sliding window [11]:

To detect multiple logos from the same image, the sliding window method was used. As the name suggests, a rectangular box is slid to scan across the extents of the test image. At each window, the keypoints are classified by using the codewords within the box. The match is determined by a score within the window. This is demonstrated in Fig. 8. This method is extremely slow in finding the match as it also iterates over the box size. Sometimes due to close location of the multiple logos of the same class, it may be hard to pin-point the location of the specific occurrence of the logo.

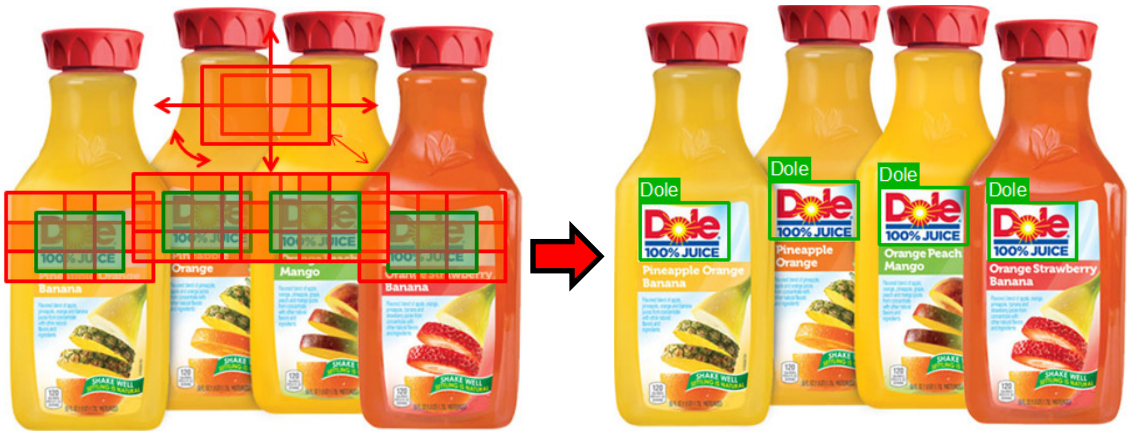

Fig. 8. Logo matching using sliding window method

3. Performance analysis

The different variations of the algorithm yield different results as shown in Fig. 9. The general trend is a better success rate with a bigger training set and improved classification scheme (eg. SPM vs

BoW). For training set size larger than 10, traditional BoW works fairly well. But for smaller training set, improvements like SPM and training sample orbit are necessary.

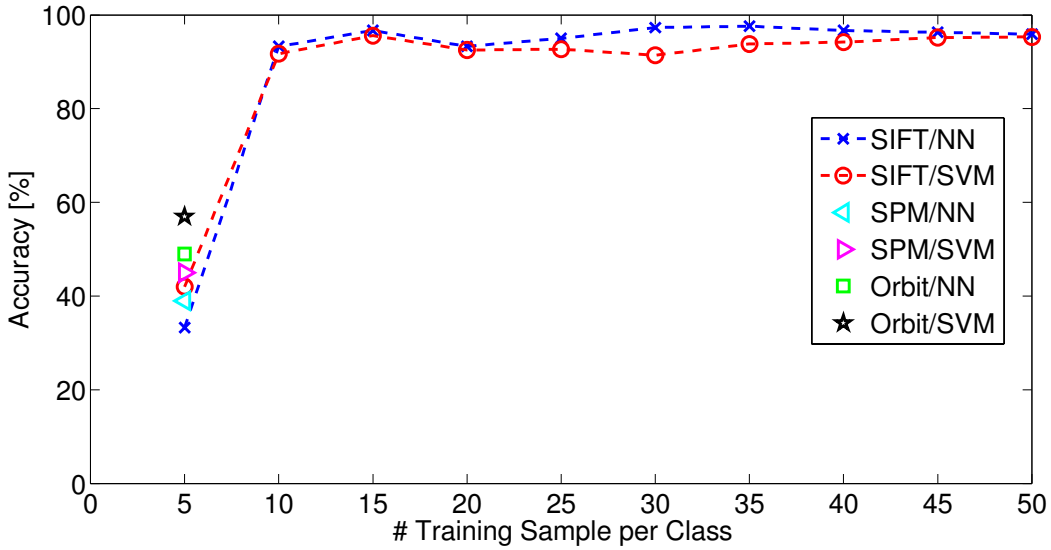

Fig. 9. Accuracy of different classification methods as a function of training set size

4. Conclusions

We experimented on developing a pipeline program for automatic food logo detection and classification in photo images. A three-step approach was proposed and implemented, using DoG + SIFT for feature extraction, variations of BoW for classification, sliding windows and Hough Voting with RANSAC pre-filtering for detection. Despite of the limited size of the the current training data, very promising results have been achieved.

5. References

[1] Kato, T., 1992. Database architecture for content-based image retrieval. Proceedings of the SPIE, Image Storage and Retrieval Systems, Vol. 1662, San Jose, CA, February 1992, pp. 112–123

[2] Doermann, D., Rivlin, E., Weiss I., Applying algebraic and differential invariants for logo recognition, Machine Vision and Applications, 9 (2) (1996), pp. 73–86

[3] Kliot, M., Rivlin, E., Shape retrieval in pictorial databases via geometric features. Technical Report CIS9701, Technion—IIT, Computer Science Department, Haifa, Israel, 1997

[4] Lowe D. G., Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

[5] Psyllos, A. P., Anagnostopoulos, C. N., Kayafas, E. (2010). Vehicle logo recognition using a SIFT-based enhanced matching scheme. Intelligent Transportation Systems, IEEE Transactions on, 11(2), 322-328.

[6] Jaisimha, M.Y., Wavelet features for similarity based retrieval of logo images. Proceedings of the SPIE, Document Recognition III, Vol. 2660, San Jose, CA, January 1996, pp. 89–100

[7] Cesarini, F., Fracesconi, E., Gori, M., Marinai, S., Sheng, J.Q., Soda. G., 1997. A neural-based architecture for spot-noisy logo recognition. Proceedings of the Fourth International Conference on Document Analysis and Recognition, Ulm, Germany, August 1997, pp. 175–179
[8] http://www.vlfeat.org/matlab/vl_sift.htm
[9] Lazebnik S., Schmid C., and Ponce J., Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories., 2006.
[10] Constantinopoulos C., Meinhardt-Llopis E., Liu Y., and Caselles V., A robust pipeline for logo detection, IEEE International Conference on Multimedia and Expo (ICME), 2011.
[11] Viola P., and Jones M., Rapid Object Detection using a Boosted Cascade of Simple Features, Computer vision and pattern recognition, 2001.