The CarLogo-51 Dataset

Lingxi Xie, Qi Tian, Wengang Zhou and Bo Zhang

January 1, 2014

Abstract

This paper presents the CarLogo-51 dataset used to evaluate the ImageWeb algorithm in [1]. The dataset is collected from the Internet, composed of 51 categories of images containing famous car logos, and could be combined with any sets of distractor images for large-scale near-duplicate Web image search. There are in total 11903 images in the dataset, all of which contain a unique kind of recognizable car logos, and the minimal number of images in one category is 200. We claim that we simulate a Web environment that every concept contains many instances, so that we could adopt affinity propagation methods, such as ImageWeb, to improve the image search quality significantly.

1 Introduction

With more than twenty years' efforts, content-based image retrieval (CBIR) has become a successful application in computer vision. It provides an effective way of bridging the intent gap by analysing the actual contents of the query image, rather than the metadata such as keywords, tags, and/or descriptions associated with the image. With compact image representation, it is possible for the state-of-the-art Web image search engines such as Google and Bing to handle billions of images and process each query with real-time response.

To search among a large corpus of images, the Bag-of-Visual-Words (BoVW) model [2] is widely adopted. The BoVW-based image search framework contains two major stages, *i.e.*, offline indexing and online searching. At the offline stage, local descriptors [3] are extracted on the crawled images, quantized onto a large visual vocabulary [4], and indexed into the inverted structure [2]. At the online stage, local descriptors are also extracted on the query image, and quantized into visual words to access the corresponding entries in the inverted index. Finally, all the retrieved inverted lists are aggregated as ranked search result.

Despite the simplicity, efficiency and scalability of the BoVW-based image search framework, the search results often suffer from the unsatisfied precision and recall. The main reasons arise from the limited descriptive power of low-level descriptors and the considerable information loss in the quantization step. In fact, the accurate matching between local features could be highly unstable especially in the cases of manual editing and geometric deformation or stretching, meanwhile there also exist a number of incorrect feature matches between some totally irrelevant images. This may cause some relevant images to be ranked after the irrelevant ones.

In our paper [1], we investigate the image search problem from a graph-based perspective, and discover a natural way of re-ranking the initial search results without using handcrafted tricks. We propose ImageWeb, a novel data structure to capture the image-level context properties. Essentially speaking, ImageWeb is a sparse graph in which each image is represented by a node. There exist an edge from node $\mathbf{I}_a$ to node $\mathbf{I}_b$ if and only if image $\mathbf{I}_b$ appears among the top of the initial search result of image $\mathbf{I}_a$. Since the links in ImageWeb actually imply the recommendation such as "$\mathbf{I}_a$ thinks $\mathbf{I}_b$ is relevant", it is straightforward to adopt the query-dependent link analysis algorithms, say, HITS [5], to re-rank the initial search results by propagating affinities through the links. We verify that, with the efficient discovery of image contexts, it is possible to achieve very accurate search results.

The CarLogo-51 dataset is therefore collected to evaluate the ImageWeb algorithm. The dataset is composed of 11903 images coming from 51 categories. Each category contains images with one famous car logo, with at least 200 images. We guarantee that each image contains exactly one kind of car logo(s) with recognizable vision features. We aim at constructing a dataset to simulate the Web environment that every concept contains a number of instances. Experimental results using the ImageWeb algorithm have verified our hypothesis, that affinity propagation methods could greatly improve the performance of near-duplicate image search.

2 The Dataset

This section introduces the CarLogo-51 dataset. We list every details in the dataset as well as principles followed in the image collection process.

2.1 Overview

A brief overview of the 51 categories could be found in Figure 1. We have collected 11903 images from 51 text queries in the Google image search engine. The text queries are composed of the logo name and the word "Logo" (for example, "Acura Logo").

The number of images varies from category to category, but there are at least 200 images in one category. We believe that such setting (one concept contains many entities) is similar to the real-world Web environments.

2.2 Image Collection

Not all images are as standard as the sample ones shown in Figure 1. Since the images are crawled from the Web, some of them could be very noisy or even

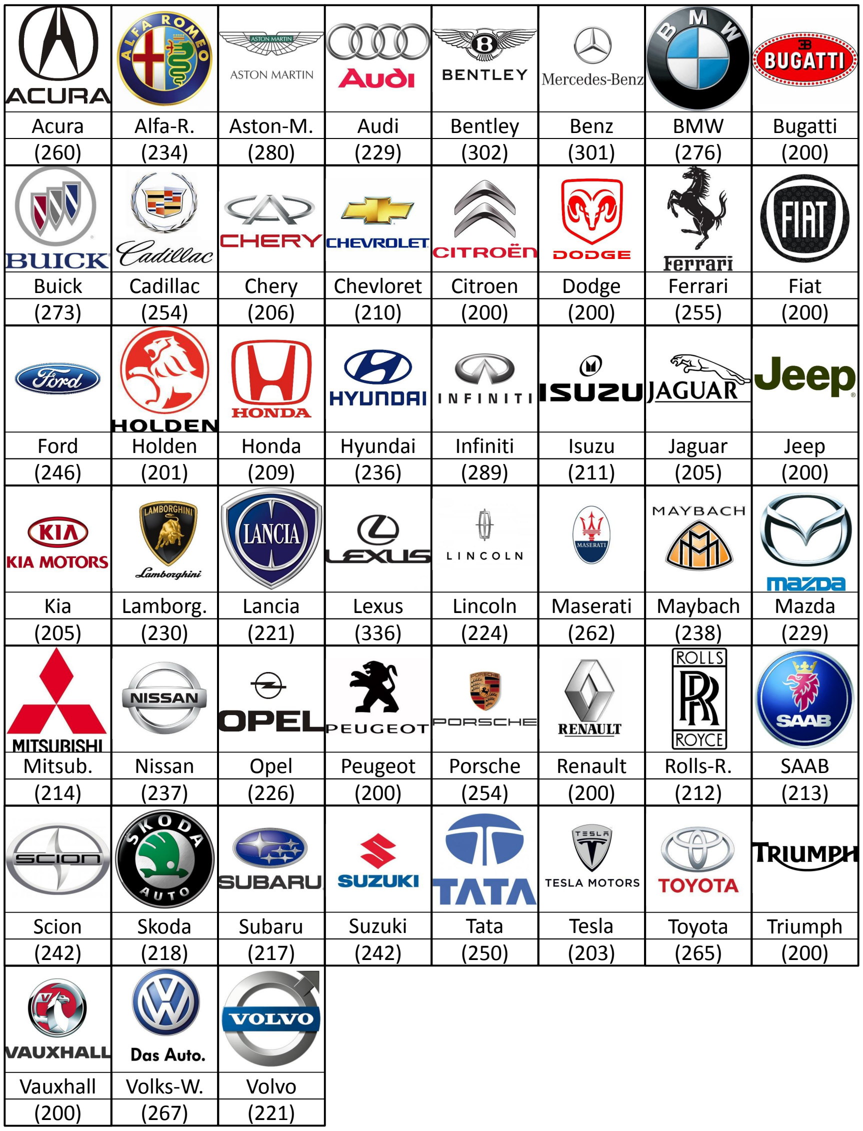

Figure 1: A brief overview of all the categories in the CarLogo-51 dataset. We list the logo names we have used to query the search engine, and one representative image for each logo name. Numbers in brackets indicate the counts of images in the corresponding categories.

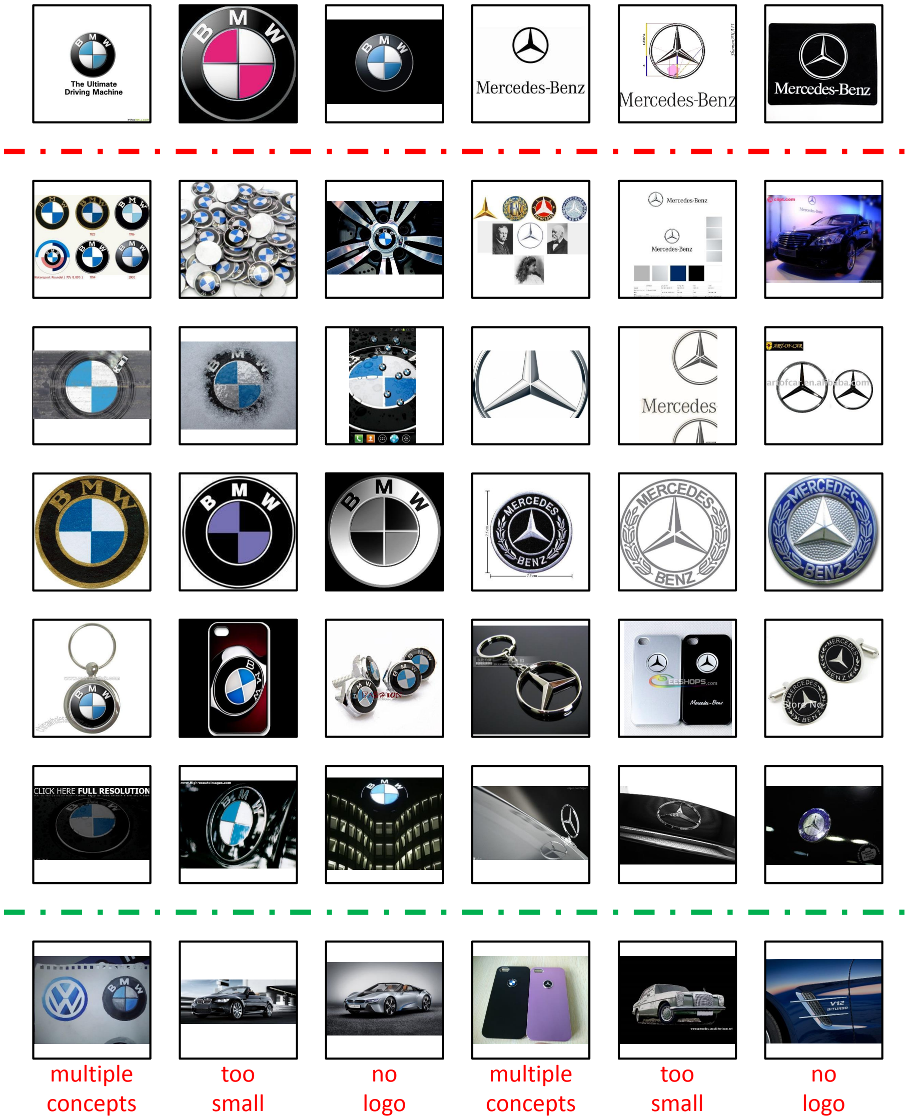

Figure 2: Examples of excellent (above the red line), fair (middle) and poor (below the green line) images. We have shown samples of BMW and Benz logos for illustration. In the middle part, we list the fair samples of the following cases, one per line, top to bottom: small logos, partial missing or occluded logos, other versions of logos, logos on common objects and logos with heavy geometric or illuminative transformations. The reasons that we could not accept the poor images are marked below the images.

unacceptable. We have manually checked the quality of the images, and verified that all the images are above fair quality.

In practise, we define an image *excellent*, if it contains a complete, high-resolution logo concept. All the sample logos in Figure 1 are excellent. The excellent logo images should be correctly analysed by most search engines.

An image is *fair*, if one or more of the following conditions happen.

- The logo is very small.
- Some element of the logo is missing.
- The logo is partial occluded or clipped.
- The logo in the image is another version of the commonly used one.
- The logo is painted on some common articles such as licence plates.
- The logo is heavily impacted by geometric transformation or illumination.

The fair logo images should be used to test the robustness of the image search engines.

An *poor* image, which is unacceptable, is related with at least one of the following conditions.

- The logo is extremely small, that we could not recognize them when we resize the image to 300 × 300.
- There exist more than one categories of logos in the image. Please note that we allow an image containing more than one logos of the specified brand, including older-version logos and/or sub-brand logos.
- The image simply contains other concepts related to the search query.

We list several samples of excellent, fair and poor quality in Figure 2. Only excellent and fair images are collected in the dataset. In spite of this, the dataset is very challenging especially for the large-scale image search task. One can check the baseline search performance in Figure 3, which is very low (mAP value around 0.2) without using the ImageWeb algorithm.

3 Experiments

We use Scalar Quantization (SQ) [6] as the baseline system. Based on the initial search results provided by SQ, we construct ImageWeb for post-processing. To make fair comparison, we keep the same settings as the baselines.

- **Descriptor extraction.** We use the SIFT descriptors [3] calculated on the Regions of Interest detected by DoG operators [3]. All the images are greyscale, and resized so that the larger axis size is 300.

- **Descriptor quantization.** We use Scalar Quantization (SQ) [6] formulation to encode each 128-D SIFT descriptor into a 256-bit binary code.
- **Indexing.** The first 32 out of 256 bits of each visual word are taken as the indexing address. Image ID as well as the remaining 224-bit binary codes are stored in the inverted index. We remove the features which appear in more than $N^{1/3}$ images where N is the number of images in the whole corpus.
- **Online searching.** We follow the basic searching process of Scalar Quantization [6] to obtain the initial scores. The **codeword expansion threshold** d and the **Hamming threshold** κ is 0 and 16, respectively. The HITS algorithm is then performed on a pre-constructed ImageWeb with depth $R = 10$ and breadth $K = 20$.
- **Accuracy evaluation.** We use the mean average precision (mAP) to evaluate the accuracies of all methods.

We compare our algorithm with the following popular methods:

1. **HVVT** [4] is the original BoVW-based framework with **Hierarchical Visual Vocabulary Tree**. We train a codebook with one million leaf codewords (6 layers, at most 10 branches at each node).
2. **HE** [7] uses **Hamming Embedding** to filter the candidate features which are quantized to the same codeword but have large Hamming distances to the query feature. The threshold for Hamming distance is selected as 20 for the best performance.
3. **SA** [8] exploits **Soft Assignment** to identify a local descriptor with a representation of nearby codewords. For the accuracy-efficiency tradeoff, we set the error bound in the k-d tree as 5.
4. **SQ** [6] gives an efficient image search flowchart based on codebook-free **Scalar Quantization**. Next, the features are indexed by their first 32 out of 256 quantized bits. Following [6], we select the codeword expansion threshold $d = 2$ and Hamming threshold $\kappa = 24$.

The mAP values are plotted in Figure 3. Our algorithm beats all the competing algorithms significantly, and enjoys a surprisingly 102% relative improvement of mAP value (0.426) over SQ (0.211), the best candidate search system, with one million distractors.

4 Conclusions

This paper is a supplementary material of the ImageWeb paper [1]. We expound the technical details we have used to construct the dataset at length, and demonstrate the promising experimental results on this dataset with the ImageWeb algorithm. For more details, please refer to the ImageWeb paper [1].

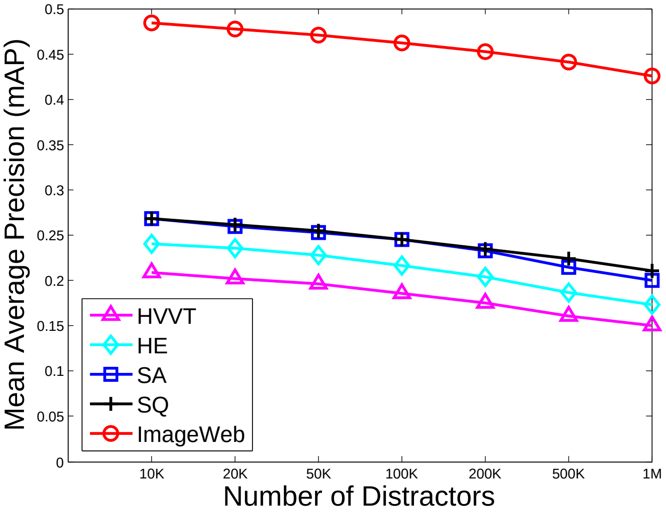

Figure 3: Performance comparison with different numbers of distractor images.

References

[1] Xie, L., Tian, Q., Zhou, W., Zhang, B.: Fast and Accurate Near-Duplicate Image Search with Affinity Propagation on the ImageWeb. Computer Vision and Image Understanding (2014)

[2] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. International Conference on Computer Vision (2003)

[3] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal on Computer Vision (2004)

[4] Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. Computer Vision and Pattern Recognition (2006)

[5] Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM (1999)

[6] Zhou, W., Lu, Y., Li, H., Tian, Q.: Scalar Quantization for Large Scale Image Search. In: ACM Multimedia. (2012)

[7] Jegou, H., Douze, M., Schmid, C.: Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In: European Conference on Computer Vision. (2008)

[8] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In: Computer Vision and Pattern Recognition. (2008)