Extremely Fast Unsupervised Codebook Learning for Landmark Recognition

Yilin Guo and Wanming Lu
HTC Beijing Advanced Technology and Research Center
Beijing, P.R. China
{guoyilin1987,louiselu0214}@gmail.com

Abstract. Traditional landmark recognition methods work by using local image features, k-means vector quantization and classifiers like SVM to recognize landmarks. However, the inefficient codebook learning by k-means constraints the possibility of using high-dimensional feature spaces, large numbers of image descriptors and large codebooks which are needed for good results. In this paper we introduce a fast unsupervised codebook learning - Extremely Random Projection Forest (ERPF), which is an ensemble of random projection tree with randomly splitting direction. We evaluate our approach on two public datasets and ERPF significantly outperforms other spatial tree methods and k-means.

Keywords: Landmark Recognition, Random Projection Tree, Codebook Learning.

1 Introduction

Recently, the massiveness of sharing photos on the Internet and the prevalence of mobile camera harness the research in landmark recognition. It is practically useful to recognize an unknown image when a user captures a landmark picture. Landmark recognition has brought great commercial potential in geo-location [1,2] and tourist guide [3].

In this paper, we treat landmark recognition as a classification task. The typical image classification pipeline is composed of the following three steps: (1) Image feature Extraction (e.g., SIFT [4], SURF [5]). (2) Codebook learning and vector quantization of each image (e.g., histogram of image descriptors). (3) Classification (e.g., SVM, Logistic Regression and Boosting). Given an image classification model on a pre-defined landmark list, which has low memory and storage costs, the image query time is very short.

Due to the image background clutter, object occlusion, changes of illumination, various scales and orientations, it is a great challenge to extract image feature efficiently while keeping the discrimination between categories and invariance inside each class. Broadly speaking, previous studies on landmark recognition ignore the impact of different local features especially the binary features [6]. Recently, the limited computational power and storage space in mobile devices
have driven the researchers to propose binary feature which has impact representation. In this paper, we firstly provide a comprehensive empirical analysis on different local features involving the recent wave of binary feature descriptors.

Conventional image classification usually uses k-means to learn visual codebook and bag-of-words model for image representation[7]. However, k-means method is computationally expensive owing to its high $O(tNK)$ processing time ($t$ is the number of iterations, $N$ is the number of data points, $K$ is the number of clusters) and the linear cost of assigning local descriptors to each visual word to find the nearest neighbor during training and testing.

Other unsupervised codebook learning methods, such as Gaussian Mixture Models[8] and mean-shift[9] have been generally used in codebook learning. But the cost of finding nearest visual words for each visual descriptor is expensive. Instead of these flatten codebook learning methods, spatial trees that involve hierarchical k-means tree[10], PCA-tree[11] and Random Projection Tree (RP tree)[11], not only efficiently encode visual descriptors into discrete codes, but also take logarithmic time to traverse the tree.

Recently, we have witnessed spatial trees applied in face recognition[12] and music similarity search[13]. To our knowledge, we have not seen a thorough analysis of its impact on landmark recognition. In this paper, we not only comprehensively evaluate these methods, but also firstly introduce the idea of random forest[14] by extending RP tree as an ensemble of trees, which we call Random Projection Forest (RP Forest). Our experiments show that RP Forest is more effective than other methods in codebook construction. Furthermore, in order to choose a projection direction promptly, we adopt the method of [13], which selects a direction maximizing the projected diameter. However, the complexity of $O(mN)$ ($m$ is number of directions) in each node is still costly when applying to large-scale data. Consequently, we randomly generate a direction without sacrificing accuracy in our experiment, which we call Extremely Random Projection Forest (ERPF). Our ERPF only needs $O(N)$ time for splitting node. Then the total forest construction time is $O(KN)$.

We evaluate our approach on two public datasets: PKUBench[15] and Landmark-3D benchmark[1]. Experiments show that ERPF outperforms k-means, ensembles of PCA-tree, RP tree and 2-means tree in time and classification performance.

2 Related Work

Most previous works consider landmark recognition as image retrieval[3,16,17] or classification problems [12,18,19]. Although image retrieval method is very fast to return similar landmark images when users input a query, it is costly to store the inverted index like vocabulary tree[10] in memory. In this paper, we treat our problem as a classification task.

The bag-of-words model[7] is proved to be efficient in image classification. Most of previous works have focused on the building of visual vocabulary[9,20,21], which is called codebook learning. The existing methods could be divided into two categories: generative and discriminative. The derived vocabularies by discriminative