International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering Vol. 2, No. 2, 2015, pp. 46-53.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

A Content Based Image Retrieval System Using K-means and K-means++

Hassan Ketabi[a] , Mohammadreza Asghari. Oskoei[b*]

[a] *Faculty of Mathematics & Computer Science (MCS) AllamehTabataba'i University (ATU), Tehran, Iran*.

[b] *Faculty of Mathematics & Computer Science (MCS) AllamehTabataba'i University (ATU), Tehran, Iran*.

Abstract

In recent years, there is an increasing research focus on Bag-of-Words based visual search. The state-of-the-art in visual search systems are built based upon a visual vocabulary model with an inverted indexing structure. This search paradigm is achieved by systems that are inspired by text retrieval. Visual words are image descriptors selected from a discrete vocabulary. These words are mapped from high-dimensional descriptors characterizing local regions of images.

In this paper, we address problems of visual vocabulary constructing by comparing k-means with k-means++ algorithms. As expected, k-means++ is clearly superior to the classical algorithm k-means both in terms of quality and running time.

Keywords: Content Based Image Retrieval, Visual search, k-means

* Corresponding Author

Introduction:

Coming with the popularity in the visual search, recent years have witnessed an increasing research focus on problem of robust image understanding with numerous applications in mobile location search, mobile product search, video copy detection and web image retrieval etc. Finding similarity between pictures, image annotation algorithms, clustering images, and annotating them are all in purview of content-based image retrieval or CBIR. There has been a renewed interest in the media about application of visual search and image analysis technologies. Nowadays hundreds of millions images are hosted in photo-sharing websites and video sharing hosts have received many requests for sharing multimedia clips.

In general, state-of-the-art visual search systems are built based upon a visual vocabulary model with an inverted indexing structure [1], in which local features of query and reference image are quantized into visual words. Each reference image is then represented as a Bag-of-Words histogram and is invert indexed by quantized words of local features in the image. Good robustness against photographing variances in occlusion, illumination, viewpoint, scale and background are the benefits of Bag-of-Words representation. The standard tf-idf weighting scheme [2] is used, which down-weights the contribution that commonly occurring, and therefore less discriminative, words make to the relevance score. Visual vocabulary based search systems work in a client-server paradigm as follows: the user from client side sends a query image to the server. At the server side, searching for similar reference images works in three steps:

1. **Feature Extraction:** Extracting local features from a query image.
2. **Quantization:** Quantizing local features into a Bag-of-Words histogram;
3. **Ranking: Ranking** similar images in the inverted indexing of all nonempty words, where the linear scanning of all reference images can be avoided in similarity ranking.

Visual Vocabulary Building: Construction of visual vocabulary often resorts to unsupervised vector quantization, which subdivides the local feature space into discrete regions each corresponds to a visual word. An image is represented as a Bag-of-Words (BoW) histogram, where each word bin counts how many local features of this image fall in the corresponding feature space partition of this word. K-means as a clustering method has been used for vector quantization and then building visual vocabulary. K-means [3], Hierarchical k-means (Vocabulary Tree) [4], Approximate k-means [1], and their well-known variants [1, 5, 6] are proposed to this end.

Quantization problems can be broadly divided into two groups: [7]: information about the original descriptor is lost, and equivalent descriptors may be assigned to different visual words. J´egou et al. have tried to address the first of these problems to keep a more accurate approximation of the original descriptor by Hamming embedding or product quantization [6]. Second problem have also been extensively studied in [1, 6], including soft-assignment [8, 9] and explicitly learning quantization variability approaches [10].

In this paper, we address quantization problems by comparing k-means with k-means++ algorithm. A visual image search system is developed to do this comparison using different constructed visual vocabularies. The goal is to retrieve all images containing a specific object in a large scale image dataset, given a query image of that object. The dataset used for the mentioned evaluation is Ukbench[11] which is an object recognition benchmark set. As expected, k-means++ is clearly superior to the classical algorithm k-means both in terms of quality and running time. Comparing k-means++ with k-means algorithm, we find that on this dataset the quality of the visual vocabulary constructed by algorithm k-means++ is on a par with or even better than the visual vocabulary obtained by algorithm k-means.

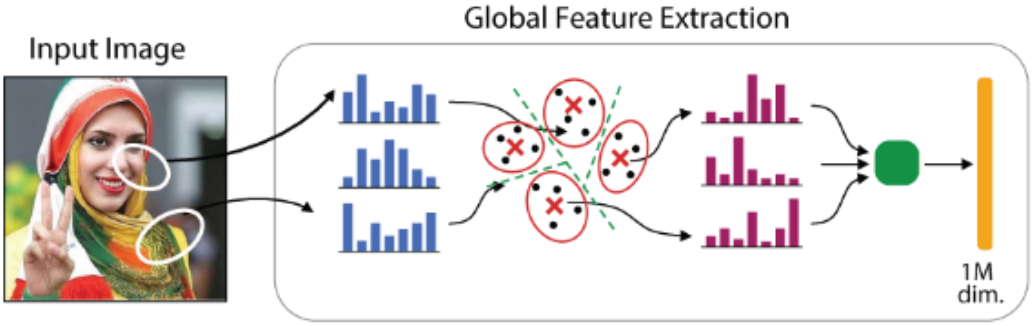

Figure (1) Here, we show a diagram for the preprocessing of image data performed before query time to build a visual vocabulary containing 1M words. This diagram illustrates extracting image descriptors and then quantizing or clustering them.

Visual Search

Visual search or content-based image retrieval systems (CBIR) tackle the problem of retrieving similar images from a large database of images. Our system uses vector-space model of information retrieval systems. It is a model for representing and searching documents. It is possible to apply this model to any kind of objects, including images. This vector contains the number of occurrences of every visual word at the position matching to that word. Recently, this approach have been adopted by the majority of image retrieval systems and also called Bag-of-Words approach, since we ignore order of the location of words. We also follow this approach and use the standard tf-idf weighting scheme where the term frequency of a word w in document *d* is:

$$tf_{w,d} = n_w / \sum_j n_j$$

where the number of occurrences of w in d is showed by nw. This is divided by the total number of occurrences of all words in the document for normalization. The inverse document frequency is

$$idf_{w,d} = \log (| D | / |\{ d : w \in d \}|)$$

where |D| is the number of documents in the corpus D and the denominator is the number of documents d in D containing w. Multiplication of tf and idf results in the tf-idf weight which is then the elements in v.

A. Visual words

First step in applying text mining techniques to images is to create the visual equivalent of words. This is usually done using local descriptors and we have implemented our own retrieval system with SIFT descriptors [12]. We quantize the descriptor space into a number of characteristic instances and assign each descriptor in the image to one of those instances. Analyzing a training set of images determines these typical samples which can be considered as visual words and the set of all words is then a visual vocabulary (sometimes called a visual codebook). This vocabulary can be made specifically for a given problem or type of images or just try to represent visual content overall.

Figure (1) Sift descriptors of a sample image. These descriptors are illustrated visually.

By using some clustering algorithm applied to the feature descriptors extracted from a large set of images, the visual words are built. K-means is the most common choice, which is what we will use in this paper. In the given feature descriptor space visual words are a collection of vectors and in the case of using k-means as a clustering method they are the cluster centroids. An image is then represented as a bag of visual words, a histogram of visual words.

B. k-Means Clustering Algorithms

The most prominent and widely used clustering algorithm is Lloyd's algorithm sometime referred to as k-means. Lloyd's algorithm is based on two remarks:

(1) Given a fixed set of centers, we acquire the best clustering by assigning each point to the nearest center.

(2) Given a cluster, the best center of the cluster is the center of mean of its points.

Lloyd's algorithm uses these two local optimizations phases many times to the in progress solution, until no more enhancement is possible. The algorithm is a heuristic that converges to a local optimum, and the quality of the computed solution is sensitive to the choice of starting centers. The main benefit of Lloyd's algorithm is its simplicity and its foundation on analysis of variances.

From a theoretical viewpoint, in terms of efficiency or quality k-means is not a good clustering algorithm: the running time can be exponential in the worst case [13] and even though the final solution can be very distant from the global optimum (even under repeated random initializations). Nonetheless, in practice the speed and simplicity of k-means cannot be defeat. Therefore, recent researches has focused on improving the initialization procedure: deciding on a better way to initialize the clustering dramatically changes the performance of the Lloyd's algorithm, both in terms of quality and convergence properties.

Figure (2) Some examples of images from the Ukbench (University of Kentucky object recognition benchmark) data set.

1) k-MEANS++

Arthur and Vassilvitskii proposed the k-MEANS++ algorithm [14], which is a seeding procedure for Lloyd's k-means algorithm. They showed a simple procedure that both leads to good theoretical guarantees for the quality of the solution, and, as a result of a good starting point, improves upon the running time of Lloyd's iteration in practice. K-means++ selects only the first center uniformly at random from the data. Each subsequent center is selected with a probability proportional to its contribution to the overall error given the previous selections. In other words, choosing each point as the next center is done with probability proportional to its squared distance from the nearest center already chosen. Note that, for a given set of centers, the squared distance of a point from its nearest center corresponds to the current contribution of this point to the total k-means clustering cost.

Comparing Clustering Methods: We compare k-means against k-means++ using exactly the same descriptors and experimental methodology as [15]. K-means++ is clearly superior.

Experimental Results

Dataset: Our proposed system is evaluated by Ukbench dataset. This challenging dataset contains images from the University of Kentucky object recognition benchmark set (The full set, reported benchmarks and some supporting code can be found at [11]). The Ukbench set contains many sets of four images, each of the same scene or object (stored consecutively so that 0...3, 4...7, etc. belong together). Some examples from the dataset is shown in Figure 3. Another dataset that have been used in object detection challenges is Oxford 5K dataset. Table 1 shows these datasets and their descriptor counts. Number of images and size of descriptors are showed in this table.

We report the results of our approaches based on using different clustering methods. As we said before our system uses BOW model and it requires the definition of a codebook of K visual words which is centroids usually obtained by k-means.

Table 1- The number of images, descriptors and size of Descriptors for two dataset.

Dataset	# images	# descriptors	Size of descriptors
Ukbench	10200	9023408	2.3 GB
Oxford	5063	16344970	4.66 GB
Total	**15263**	**25368378**	**6.9 GB**

Evaluation scheme. The accuracy measurement we used to evaluate the results of designed system is a score function that returns the average number of correct images on the top four result of queries. The fact that the correct images are consecutive in groups of four is being used here. An ideal result gives a score of 4, nothing right gives a score of 0 and only retrieving the identical images gives a score of 1. Finding the identical image together with two of the three other images gives a score of 3.

Our goal is to evaluate the retrieval performance of visual vocabularies built using the two clustering algorithms described above. This is to establish how much performance is lost by using a less accurate method at detecting centroids. Experimental results demonstrate that choosing k-means++ significantly outperforms visual search on most datasets.

Figure (3) confirms this outperforming and depicts using k-means++ in 1K centers is similar to standard k-means, differing in evaluation score by less than 0.3 units. But in 5K centers k-means++ is clearly superior to the classical algorithm k-means.

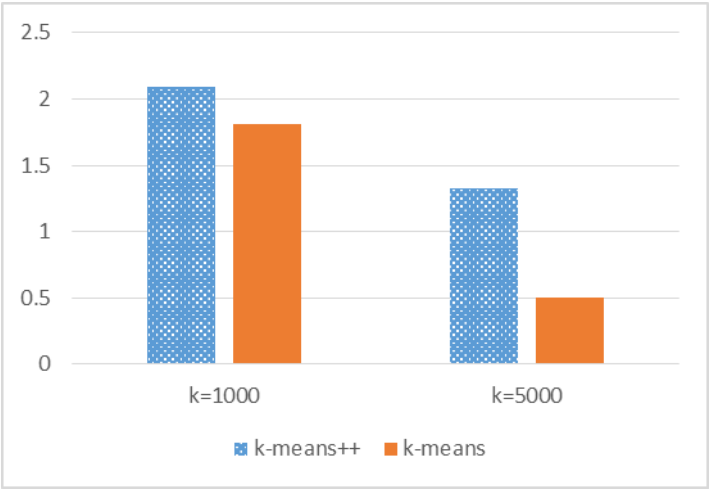

Figure (3) A comparison of k-means and k-means++ on the Ukbench dataset using the 1K and 5K vocabularies. The numbers are evaluation scores described in evaluation scheme section.

Conclusions And Future works

In this paper we proposed a visual search system to compare two different k-means algorithms for visual vocabulary building to show how much important it is to retrieve better results. We first built visual vocabulary and then indexing images was done afterwards means extracting descriptors from the images, converting them to visual words using the vocabulary and storing the visual words and word histograms

with information about which image they belong to. After indexing images, search for similar images was possible. The bag-of-word representation was used for the whole image which groups together local descriptors.
Results showed that visual vocabulary built using k-means++ presents better to build a visual vocabulary. It shows that image descriptors that is lost in quantization step have a big effect on performance of system and choosing a more exact algorithm for this step can increase the results of queries.
Content based image retrieval is a challenging problem, sitting at the intersection of computer vision, information retrieval and machine learning. Scaling the system to handle millions of images for object retrieval is hard, and there is huge number of possible direction for future work – scaling the system and using other clustering method.

Acknowledgements. We are grateful for support from Yooz.ir, Iranian search engine that facilitated computing environment.

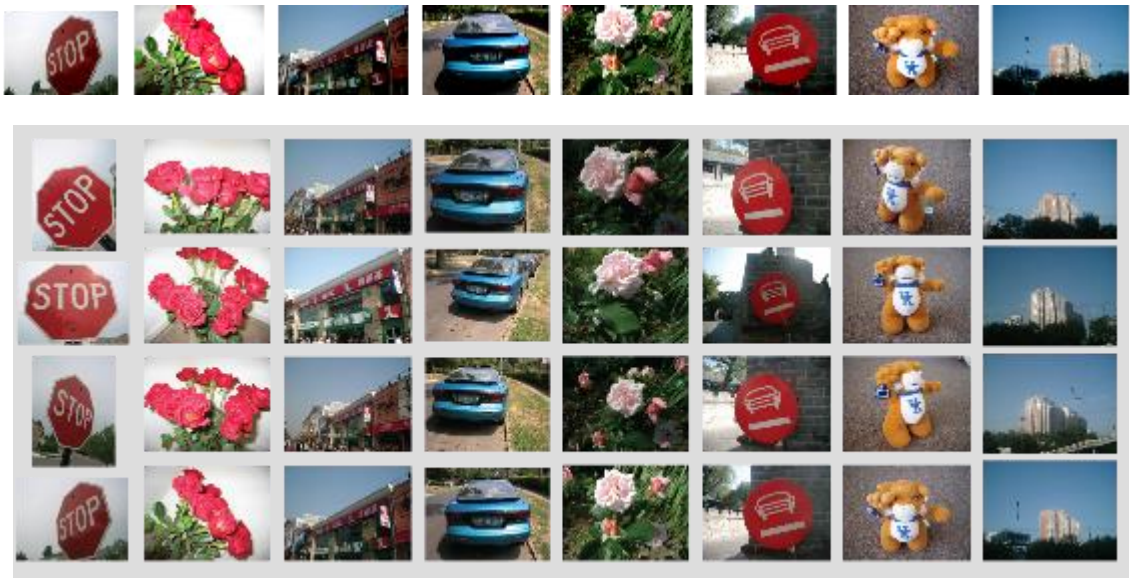

Figure (4) Some example search results on the Ukbench data set. The query image is shown on the far top followed by the top four retrieved images.

References:

A. Makadia. Feature tracking for wide-baseline image retrieval. In Proc. ECCV.

D. Arthur and S. Vassilvitskii. How slow is the k-means method? In SOCG, p., 2006.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In SODA, p., 2007.

D. Blei, A.N., and M. Jordan, "Latent dirichlet allocation,"J. Mach. Learn. Res., vol. 3, no. 4/5, pp. 993–1022, 2003.

D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

D. Nister and H. Stewenius, S.r.w.a.v.t., ” inProc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

D. Nister H. Stewenius. Scalable recognition with a vocabualry tree. In proceedings of IEEE Conference Vision and Pattern Recognition.

F. Jurie and B. Triggs, C.e.c.f.v.r., ” inProc. Int. Conf. Computer Vision, 2005.

G. Salton and C. Buckley, T.-w.a.i.a.t.r., ”Inf. Process. Manage., 1988.

G.SchindlerandM.Brown, C.-s.l.r., ” inProc. IEEE Conf. Computer Vision and Pattern Recognition, 2007.

H. J´egou, M.D., and C. Schmid. Improving bag-of-features for large scale image search. IJCV, 87(3):316–336, 2010.

http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/.

http://www.vis.uky.edu/˜stewe/ukbench.

J. Philbin, O.C., M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabulary and fast spatial matching,” inProc. IEEE Conf. Computer Vision and Pattern Recognition, 2007.

J. Sivic and A. Zisserman, V.G.A.t.r.a.t.o.m.i.v., ” in Proc. Int. Conf. Computer Vision, 2003.

R. Arandjelovi ´ c and A. Zisserman. Three things everyone should know to improve object retrieval. In Proc. CVPR.