Bag of Words – quality issues of near-duplicate image retrieval

Mariusz Paradowski, Mariusz Durak, Bartosz Broda
Institute of Informatics, Wrocław University of Technology, Poland

Abstract. This paper addresses the problem of large scale near-duplicate image retrieval. Issues related to visual words dictionary generation are discussed. A new spatial verification routine is proposed. It incorporates neighborhood consistency, term weighting and it is integrated into the Bhattacharyya coefficient. The proposed approach reaches almost 10% higher retrieval quality, comparing to other recently reported state-of-the-art methods.

Key words: spatial verification, vector space model, visual words, clustering.

1. Introduction

In recent ten years lots of research effort has been put to the problem of large scale near-duplicate image retrieval. Databases containing thousands or even millions of images are successfully processed by efficient retrieval algorithms. The basic and well known technique for efficient near-duplicate retrieval is *vector space model* [1, 5, 15]. Feature representation used in vector space model is called *bag of words* (BoW). Bag of words is a histogram (one-dimensional in most cases) representing a single image and constructed from many *visual words* [5]. A *visual word* is a single feature vector compressed to a single numeric value (group, cluster). Image similarity measurement is usually done by histogram comparison. Application of various histogram similarity measures [9] is possible. Efficient image retrieval systems are constructed on top an of inverted index structure [8]. Data related to visual words is stored in *inverted index files*. Files are read on demand, thus memory usage is kept low. Computational complexity of the retrieval process is less than $O(n)$ per image pair, where n is the number of visual words of the image.

1.1. Research background

Plain bag of words histogram comparison is not sufficient in many cases. Visual words context has to be taken into account (similarly as in Natural Language Processing) using spatial (context) analysis. Spatial analysis methods come in two major categories: global and local. Global methods find image transformations (usually affine or perspective) between images. Major approaches are: RANSAC [2] based methods for single transformation detection (e.g. [18]), Hough transformation for multiple object detection

Machine GRAPHICS & VISION 23(1/2):83–96, 2014. DOI: 10.22630/MGV.2014.23.1.5 .

(e.g. [7, 17]) and non-linear approaches (e.g. [16]). Local methods usually verify visual word neighborhood consistency, e.g. [4].

Spatial image analysis is a time consuming process, much slower comparing to simple histogram comparison. An efficient, spatially enforced retrieval scheme is hierarchical. First, a complete database of images is processed using histogram similarity measurement. Later on, a subset of best results is analyzed using spatial methods. In the paper we address both issues: efficient histogram comparison and spatial analysis of images.

1.2. Contribution

The contribution of the paper is the following. We propose an extended version of simple spatial verification routine [4]. We show that the commonly used *cosine* bag of words similarity measure [10,18] is outperformed by *Bhattacharyya coefficient* and χ^2 distances. We discuss bag of words dictionary generation. We point out a problem in a commonly used experimental protocol.

The paper is organized as follows. The second section presents the complete retrieval scheme, together with our proposals. The third section demonstrates the experiments performed. The last section concludes our work.

2. Image retrieval method

The discussed retrieval method is based on the *vector space model* [1]. The vector space model assumes the existence of discrete terms (words), which describe both the database and the query. The model has been originally used in Natural Language Processing, where terms are naturally related to words or concepts. The model can be effectively used in image processing [5]. However the existence of discrete terms is not obvious. *Visual terms* also known as *visual words* have to be generated from visual features, which are usually continuous in nature. The process of visual terms generation can be modeled as a grouping problem and performed using variants of *k-means* method. Representation of a single image consists of many high dimensional feature vectors (key points, key regions) [6]. Each vector is converted into a visual term and thus the image representation becomes a bag of words.

One of the elementary issues in vector space model is a proper construction of *visual term dictionary.* Construction of the dictionary is an unsupervised grouping problem. There are many grouping approaches, but classic *k-means* have gained most popularity. Application of *k-means* is not accidental – it minimizes the total distance between cluster centers and data points. To speed up the grouping routine some researches have applied various modifications of k-means, including: *hierarchical k-means* and *approximate k-means* [10]. We revert to the original k-means with much success.

Bag of words representing a single image may be defined both in terms of vectors and probability distributions, thus both vector similarity and PDF similarity measures may be used. For efficient retrieval the similarity function has to be calculated on sparse vector representation in $O(n+m)$ computational complexity, where n and m are number of non-zero elements in sparse *BoW* vectors. Only a subset of similarity measures follow this criterion. The widely recognized and used one is the *cosine similarity*:

$$cos(x, y) = \sum_{i \in V} x_i y_i, \tag{1}$$

where: x and y are L2 normalized vectors representing weighted image BoW's, V is the set of visual words, x_i and y_i are BoW values of the i-th bin.

After the generation of the initial bag of words ranking, spatial verification takes place. It is a key issue in successful near-duplicate retrieval [10]. There are many approaches to spatial verification, including:

1. geometry-based approaches that reconstruct various transformations between two images, e.g., RANSAC [2,10], Hough Transform [17];
2. topology-based approaches focused on various local pseudo-invariants, e.g. [3,4].

Further post-processing methods are also available, such as *query expansion*. They are outside the scope of our research and we do not address them in the paper. However, it is worth noting that the presented approach is compatible with these routines.

2.1. Alternative bag of words similarity measurement

Let us first address the similarity measurement of two images in a vector space model. Many similarity measures may be used in the vector space model [9]. We have analyzed several of them. Two of them proven to be worth of interest.

A measure that is rarely used in image retrieval, but has proven to be interesting, is *Bhattacharyya coefficient* of two probability distributions. In this case each *BoW* representation becomes a discrete PDF (L1 normalized vector). Distribution similarity is defined as follows:

$$BC(x, y) = \sum_{i \in V} \sqrt{x_i y_i}, \tag{2}$$

Another measure worth noting is χ^2 histogram distance. There are several variants of χ^2 [9], we choose the symmetric one:

$$\chi^2(x, y) = \sum_{i \in V} \frac{(x_i - y_i)^2}{x_i + y_i}. \tag{3}$$

Both of the above similarity measures can be used in inverted–index retrieval approach. Usage of $BC(x, y)$ is straightforward, because only corresponding non-zero elements influence the final result. χ^2 distance integration is slightly more difficult. Let us assume that x is the query image. In such case $V_x \subset V$ is the set of visual words belonging to x through which we iterate:

$$\begin{aligned}\chi^2(x,y) &= \sum_{i\in V}\frac{(x_i-y_i)^2}{x_i+y_i} = \sum_{i\in V}\frac{x_i^2}{x_i+y_i} + \sum_{i\in V}\frac{y_i^2}{x_i+y_i} - 2\sum_{i\in V}\frac{x_i y_i}{x_i+y_i} = \\ &= \sum_{i\in V_x}\frac{x_i^2}{x_i+y_i} + \left(\sum_{i\in V}\frac{y_i^2}{y_i} - \sum_{i\in V_x}\frac{y_i^2}{y_i} + \sum_{i\in V_x}\frac{y_i^2}{x_i+y_i}\right) - 2\sum_{i\in V_x}\frac{x_i y_i}{x_i+y_i}.\end{aligned} \tag{4}$$

The above equation shows that only one component per database image has to be precomputed ($\sum_{i\in V} y_i$). All other are either equal to 0 or may be calculated only from $V_x \subset V$, during inverted index lookup.

2.2. Proposed spatial validation routine

Let us now describe the main research contribution of this paper. The proposed method is a spatial validation routine. The basic idea of the approach may be traced back to early works of Mohr and Schmid [3, 4], later incorporated into the vector space model by Sivic [5]. The main idea is to check spatial consistency of neighboring pairs. The higher the spatial consistency the better. We extend this idea using other well known techniques in the following ways:

1. neighboring pairs are *tf-idf* weighted instead of simple counting to incorporate their importance in vector space model (e.g. [12]),
2. neighborhood size is dynamic and is relative to key points size ratio,
3. neighboring pairs consistency is relative to key point size ratio and key point distance ratio,
4. the routine is integrated into *Bhattacharyya coefficient* giving a re-weighted BoW histogram.

First let us define a standard *tf-idf* weighted *Bhattacharyya coefficient*:

$$BC_{tf\,idf}(x,y) = \sum_{i\in V}\sqrt{tf(x_i)\cdot idf_i \cdot tf(y_i)\cdot idf_i}. \tag{5}$$

Term frequencies $tf(x_i)$ and $tf(y_i)$ represent histogram bin data. In spatial verification routine they have to be replaced by the neighborhood consistency function $SV(x_i, y_i)$:

$$BC_{SV}(x,y) = \sum_{i\in V}\sqrt{idf_i^2 \cdot SV(x_i,y_i)}. \tag{6}$$

Spatial verification function $SV(x_i, y_i)$ measures the consistency of all key point pairs belonging to the i-th bin. Let us define a function $N(a,b)$ which measures the consistency of a key point pair (a,b). Consistency value has to be normalized, thus maximum possible consistency $M(a,b)$ has to be defined. Functions $N(a,b)$ and $M(a,b)$ are defined further, by eqs. (9) and (10). Basic spatial consistency is defined as a ratio of measured consistency and maximum consistency (e.g. [12,14]):

$$SV_{basic}(x_i,y_i) = \sum_{(a,b)\in P_i} \frac{|N(a,b)|}{|M(a,b)|}, \tag{7}$$

where: $P_i = x_i \times y_i$ – set of key point pairs belonging to i-th BoW bin, i.e., a Cartesian product of key point sets x_i and y_i.

The down side of such spatial consistency is that all neighboring pairs have the same contribution. As we know from the vector space model, some key points are more important than other ones. Their importance is measured by *tf-idf*. Following this idea, spatial verification with key point pair importance measurement is defined as:

$$SV_{tfidf}(x_i,y_i) = \sum_{(a,b)\in P_i} \frac{\sum_{(\alpha,\beta,k)\in N(a,b)} idf_k}{\sum_{(\alpha,\beta,k)\in M(a,b)} idf_k}, \tag{8}$$

where: (α,β) is the neighboring key point pair and it belongs to cluster k.

To address *scale-invariance* problem, neighborhood size for key point a and key point b should not be set equal. Useful information about scale can be extracted from square roots r_a and r_b of the areas of key points a and b. Thus neighborhood functions $N(a,b)$ and $M(a,b)$ are defined as:

$$N(a,b) = \{(\alpha,\beta)\in P : ||a,\alpha|| < \epsilon \wedge r_b||b,\beta|| < r_a\epsilon\}, \tag{9}$$

and

$$M(a,b) = \{(\alpha,\beta)\in P : ||a,\alpha|| < \epsilon \vee r_b||b,\beta|| < r_a\epsilon\}. \tag{10}$$

where: P is the set of all key point pairs, $||\cdot,\cdot||$ stands for Euclidean distance and ϵ is the distance limit. The above definition of $N(a,b)$ and $M(a,b)$ ensures that $|N(a,b)| \leq |M(a,b)|$ is always true.

Yet another information about quality of each key point pair can be extracted from the relative size of key points. Given that pair (a,b) has size ratio $\frac{r_a}{r_b}$, distances between key points ($||a,\alpha||$ and $||b,\beta||$) should follow the same ratio. Distance ratio agreement $ratio(a,b,\alpha,\beta)$ can be defined as:

$$ratio(a,b,\alpha,\beta) = \min\left(\frac{||a,\alpha||r_a}{||b,\beta||r_b}, \frac{||a,\alpha||r_b}{||b,\beta||r_a}\right). \tag{11}$$

Thus, final spatial verification routine $SV(x_i,y_i)$ is defined as:

$$SV(x_i,y_i) = \sum_{(a,b)\in P_i} \frac{\sum_{(\alpha,\beta,k)\in N(a,b)} idf_k \min\left(\frac{||a,\alpha||r_a}{||b,\beta||r_b}, \frac{||a,\alpha||r_b}{||b,\beta||r_a}\right)}{\sum_{(\alpha,\beta,k)\in M(a,b)} idf_k}. \tag{12}$$

Finally, *Bhattacharyya coefficient* with spatially weighted histogram bins is defined as:

$$BC_{SV}(x,y) = \sum_{i \in V} \sqrt{ idf_i^2 \sum_{(a,b) \in P_i} \frac{\sum_{(\alpha,\beta,k) \in N(a,b)} idf_k \min\left(\frac{||a,\alpha||r_a}{||b,\beta||r_b}, \frac{||a,\alpha||r_b}{||b,\beta||r_a} \right)}{\sum_{(\alpha,\beta,k) \in M(a,b)} idf_k} }. \tag{13}$$

Let us now present the retrieval quality verification of all the discussed ideas.

3. Experimental verification

The proposed retrieval approach has been experimentally verified according to well established image retrieval protocols. Quality is measured using *mean average precision.* Reference queries and reference results are predefined. Precision–recall curves are shown. Two widely recognized image datasets are used in the process: Oxford5K [10] and Paris6K [11]. Our experiments address the following aspects of retrieval quality:

1. influence of number of *k-means* iterations during visual words dictionary generation,
2. influence of *BoW* similarity measurement,
3. a small flaw in the widely used experimental protocol,
4. proposed spatial verification routine,
5. cross-database visual words dictionary use.

3.1. Clustering and number of k-means iterations

The first of the addressed issues deals with the quality of BoW representation. It has been shown by researchers that high dimensional BoW gives better retrieval quality comparing to low dimensional ones. A standard approach based on *approximate k-means* have been used in several state-of-the-art papers [10, 18]. In our experiments we show that reverting to standard k-means can lead to higher quality. We also show that only a few k-means iterations are necessary to get satisfying clusters quality.

Modern hardware allows highly efficient implementation of k-means, using vector CPU and GPU processing (e.g. [13]). Thus, computational complexity of k-means method is no longer a problematic issue. Despite large increase in speed, the generation of visual words dictionary still takes some time. One iteration with 32 simultaneous distance calculations (CPU, 8 threads and 4 values in SIMD instructions) takes from several minutes to few hours for the presented data. Thus, we would like to know if it is worth iterating until k-means converges or it is possible to stop earlier.

Obviously, in each iteration of k-means the total distance between cluster centers and data points decreases. This decrease leads to cluster centers improvements and in result, to better retrieval quality. Achieved results are presented in Tab. 1. Our experiments show that it is sufficient to perform only a few k-means iterations.

Tab. 1. Retrieval quality for various number of k-means iterations.

iteration	k-means	distance	RootSIFT descriptor [18]		
number	distance	change	$\cos(x,y)$	$BC(x,y)$	$\chi^2(x,y)$
	Oxford5K image dataset, 500000 clusters				
0	1,732,112,755	–	0.740	0.761	0.760
1	1,511,378,391	220,734,364	0.751	0.773	0.772
2	1,486,011,493	25,366,898	0.758	0.783	0.783
5	1,465,147,423	20,864,070	0.764	0.786	0.786
10	1,459,041,178	6,106,245	0.763	0.787	0.786

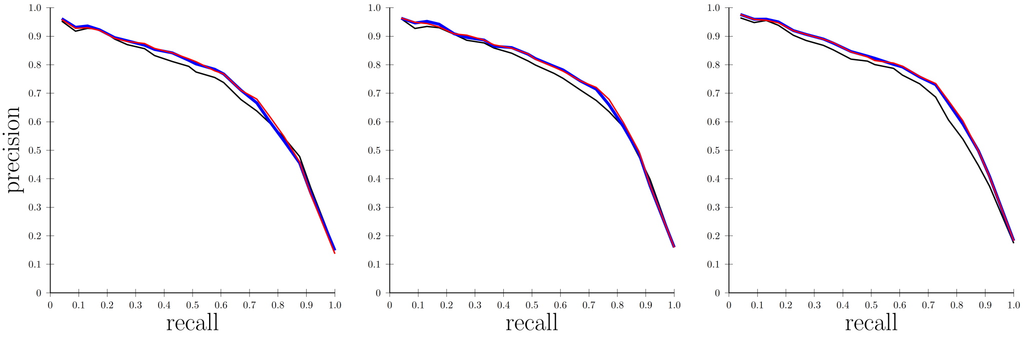
Tab. 2. Retrieval quality for various vocabulary sizes and similarity measures.

vocabulary	SIFT descriptor [6, 7]			RootSIFT descriptor [18]		
size	$\cos(x,y)$	$BC(x,y)$	$\chi^2(x,y)$	$\cos(x,y)$	$BC(x,y)$	$\chi^2(x,y)$
	Oxford5K image dataset					
50000	0.673	0.661	0.659	0.708	0.692	0.696
100000	0.688	0.688	0.684	0.736	0.741	0.741
200000	0.720	0.738	0.733	0.760	0.771	0.770
500000	0.726	0.756	0.748	0.762	**0.787**	**0.787**
1000000	—	—	—	0.746	0.781	0.779
	Oxford5K image dataset – original clusters reference					
1000000	0.636 [10]	—	—	0.683 [18]		
	Paris6K image dataset					
200000	—	—	—	0.742	0.753	**0.759**

3.2. Bag of words similarity measures

A standard framework for large scale image retrieval uses *cosine similarity* of image BoW representation. Our experiments have shown that better results may be obtained with other similarity measures. Table 2 compares BoW retrieval results obtained for: *cosine similarity*, *Bhattacharyya coefficient* and χ^2 *histogram distance. Cosine similarity* turns to be the least effective because bag of words is in fact a histogram. Similar conclusions have been drawn by Zisserman [18] regarding SIFT features, when RootSIFT features were designed.

Comparison of retrieval quality is presented in Tab. 2. Precision–recall curves are shown in Fig. 1. Retrieval quality using *Bhattacharyya coefficient* and χ^2 *histogram distance* is higher in 7 out of 10 tested cases. These 7 cases are the important ones, because they have the highest overall quality. Achieved quality values are higher by up to 4%. Yet another interesting result is the comparison with state-of-the-art reference



(a) 100000 (b) 200000 (c) 500000

Fig. 1. Precision–recall curves for various vocabulary sizes and similarity measures (cos, BC, χ^2); Oxford5K database, RootSIFT descriptor.

results. Using the proposed approach, the achieved results quality is 8% higher for RootSIFT features and 9% higher for SIFT features.

3.3. Discussion on experimental protocol

In this section we would like to point out a small flaw in the broadly accepted experimental protocol. As pointed in [10] all 16.7M feature vectors are used in visual words dictionary generation. This approach does not seem to be valid in terms of machine learning quality estimation, because it is positively biased. The value of the bias increases with the size of the dictionary, because there are less and less vectors assigned to each group. Given *Oxford5k* database and 1M clusters, there are only about 17 vectors per cluster.

Comparison of clustering with and without query vectors is presented in Tab. 3. A decrease in retrieval quality is clearly visible in all cases.

Tab. 3. Retrieval quality for visual words dictionary generation with and without query vectors.

vocabulary	grouping with all vectors			grouping without query vectors		
size	$\cos(x,y)$	$BC(x,y)$	$\chi^2(x,y)$	$\cos(x,y)$	$BC(x,y)$	$\chi^2(x,y)$
	Oxford5K image dataset					
50000	0.708	0.692	0.696	0.646	0.650	0.652
100000	0.736	0.741	0.741	0.717	0.715	0.712
200000	0.760	0.771	0.770	0.731	0.743	0.741

There are two conclusions. First, more care should be taken when experimental protocols are constructed, because flaws like this one may lead to incorrect conclusions. Second, despite the drop in quality, *Bhattacharyya coefficient* and χ^2 *distance* are still better than *cosine similarity*, especially when the number of clusters increases.

3.4. Proposed spatial verification

Next performed experiment addresses the proposed spatial verification. Retrieval is organized in a hierarchical way [10, 18]. First, BoW similarity is calculated for the entire database. Results are ordered according to similarity (most similar come first). Top n images processed using spatial verification ($n = 1000$, according to the accepted protocol). After the verification, the subset of images is resorted once again.

First, we present baseline results of the spatial verification. The results of the following approaches are shown in Tab. 4:

- plain BoW with BC similarity measure (reference),
- spatial verification without BC integration (no square-root and L2 distance as norm),
- BC_{SV} variant without *tf-idf* weighting and without relative size of neighborhood,
- BC_{SV} variant without relative size of neighborhood,
- BC_{SV} variant without *tf-idf* weighting,
- full BC_{SV} proposed approach.

The proposed BC_{SV} outperforms all the other tested approaches. The following order of contribution arises out of the presented result:

- BC integration plays the key role and contributes most,
- relative key point size weighting is secondary,
- *tf-idf* contribution is least significant, but still permanent.

It is worth emphasizing that simple neighbors counting without the proposed extensions achieves worse results than plain bag of words retrieval (see Tab. 4, first and second rows). Let us now present the experimental setup of the parameter ϵ (see eq. (9)

Tab. 4. Retrieval quality comparison of BoW and spatial verification variants.

retrieval approach	*tf-idf neighbors*	*relative size weight*	*BC BoW integration*	dictionary size 100K	200K	500K
BoW, BC	–			0.741	0.771	0.787
partial BC_{SV}	no	no	no	0.726	0.744	0.749
partial BC_{SV}	no	no	yes	0.775	0.795	0.809
partial BC_{SV}	yes	no	yes	0.779	0.797	0.810
partial BC_{SV}	no	yes	yes	0.783	0.804	0.819
BC_{SV}	yes	yes	yes	**0.785**	**0.807**	**0.820**

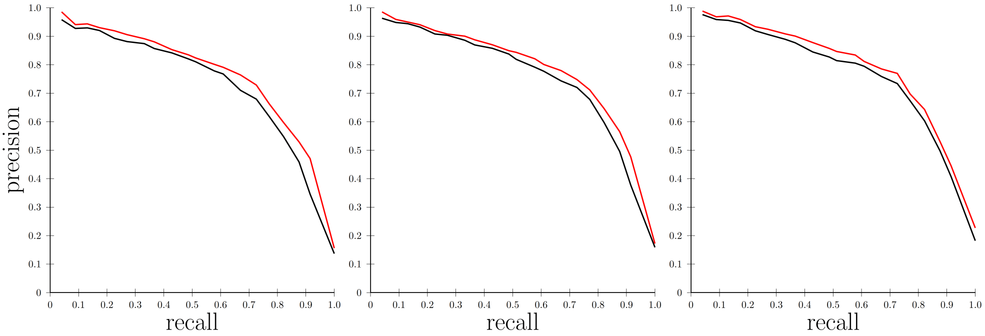
Tab. 5. Retrieval quality for various ϵ parameter values in spatial verification.

dictionary	ϵ parameter value				
size	0.05	0.10	0.15	0.20	0.25
Oxford5K database, RootSIFT features					
100000	0.777	0.792	0.785	0.784	0.781
200000	0.791	0.809	0.807	0.805	0.803
500000	0.803	0.817	**0.820**	0.818	0.817

Tab. 6. Retrieval quality comparison for BoW and proposed spatial verification.

dictionary	method			
size	BoW	BC_{SV}	BC_{SV} - BoW	gain(BC_{SV}, BoW)
Oxford5K database, RootSIFT features				
50000	0.692	0.754	0.062	8.9%
100000	0.741	0.785	0.044	5.9%
200000	0.771	0.807	0.036	4.6%
500000	0.787	**0.820**	0.033	4.2%
1000000	0.781	0.799	0.018	2.3%
reference [18]	0.683	0.720	0.037	5.4%
Oxford5K database, RootSIFT features, no queries in clusters				
50000	0.650	0.713	0.063	9.6%
100000	0.715	0.769	0.054	7.5%
200000	0.743	**0.786**	0.043	5.7%
Oxford5K database, SIFT features				
50000	0.661	0.719	0.058	8.8%
100000	0.688	0.745	0.057	8.3%
200000	0.738	0.785	0.047	6.4%
500000	0.756	0.787	0.031	4.1%
reference [18]	0.636	0.672	0.036	5.7%
Paris6K database, RootSIFT features				
200000	0.753	0.783	0.030	4.1%

and (10)) specifying the neighborhood size. Tab. 5 presents retrieval quality with various values of the parameter. Best retrieval quality is reached for $\epsilon \in \langle 0.10, 0.20 \rangle$. Suggested value of ϵ is 0.15. This value is used in all the presented experiments. Detailed results of the final BC_{SV} approach are presented in Tab. 6. Precision–recall curves for various vocabulary sizes are presented in Fig. 2. Retrieval with spatial verification clearly dominates bag of words retrieval in all the tested cases.



(a) 100000 (b) 200000 (c) 500000

Fig. 2. Precision–recall curves for retrieval with and without spatial verification (BoW, SV), Oxford5K database, RootSIFT descriptor, various vocabulary sizes.

Due to differences in BoW retrieval quality, comparison with reference approaches is done using $gain(x,y)$ measure for a given quality measure q:

$$gain(x,y) = \frac{q(x) - q(y)}{q(y)}. \tag{14}$$

This approach works in favor of the reference approach. Reference BoW quality is lower that achieved BoW quality. Thus it is more difficult for the proposed method to get a similar quality increase. Presented results show that the proposed retrieval scheme outperforms reference approach both in absolute values and in gain values.

3.5. Cross-database retrieval

The last presented test addresses cross-database retrieval. Visual words dictionary is built on one database but tests are performed on a different database. In the presented tests we use Oxford5K and Paris6K databases. This is a very important test because it minimizes the problem of cluster over-training.

Achieved retrieval quality is much lower comparing to single database tests (see Tab. 7). Visual words dictionary is not fine tuned for the retrieved data. One can observe large differences in cluster quality directly estimated using the distance criterion which *k-means* minimizes (Tab. 7, first column).

Interestingly, relations between quality of tested approaches are similar to those in single database tests. *Cosine* similarity has worst results, spatial verification on *Bhattacharyya coefficient* achieves highest quality. Gain values for spatial verification routine on cross-database retrieval are larger than those for a single database test. For Oxford5K database it reaches 9.7%, for Paris6K database it is equal to 6.9%.

Tab. 7. Retrieval quality for cross database visual dictionary generation, 200000 clusters.

k-means	retrieval approach				
distance	$\cos(x, y)$	$BC(x, y)$	$\chi^2(x, y)$	BC_{SV}	gain(BC_{SV}, BoW)
Oxford5K image dataset, Paris6K clusters					
1,617,336,879	0.587	0.603	0.607	0.662	9.7%
Oxford5K image dataset, Oxford5K clusters					
1,538,261,883	0.760	0.771	0.770	0.807	4.8%
Paris6K image dataset, Oxford5K clusters					
1,879,399,544	0.608	0.619	0.630	0.662	6.9%
Paris6K image dataset, Paris6K clusters					
1,774,815,215	0.742	0.753	0.759	0.783	4.1%

3.6. Concluding remarks

Taking into account the above experiments we found the following conclusions. Quality of visual words dictionary has a large impact on the retrieval quality. The above statement confirms the results and conclusions reached by other researchers. However, the usage of exact k-means instead of approximate or hierarchical versions seems to be a better choice. Modern hardware (both CPU and GPU) makes usage of exact k-means no longer a blocker, as it was several years ago. We have also found out that only few iterations of k-means is sufficient to get satisfying retrieval quality.

The second conclusion addresses a small flaw in the widely used experimental protocol [10]. We state that clustering all data, together with query vectors, breaks the principles of machine learning. We show that when query vectors are removed from the clustering process, the retrieval quality degrades.

The third conclusion focuses on the retrieval process itself. We suggest that *Bhattacharyya coefficient* and χ^2 distances should replace *cosine similarity.* They give better retrieval quality (see Tabs. 1, 2, 3 and 7). We have also presented an alternative to RANSAC–based spatial verification routine. The proposed approach combines neighborhood consistency with term weighting and bag of words similarity measurement. Measured gain values over standard bag of words approach are highest for cross-database retrieval. In the presented scenario they reach 9.7%.

4. Summary

This paper shows that the potential of simple image retrieval approaches have not been fully explored. Vector space model is a well known, researched and established technique.

However, our approach has outperformed the recently reported results [18] by 10% using the same experimental protocol. The reasons of such an improvement are:

- better visual words dictionary generation using *k-means* instead of *approximate k-means*,
- application of less popular *Bhattacharyya coefficient* or χ^2 distance instead of *cosine similarity*.

The third reason and research contribution of the paper is the spatial verification. It integrates neighborhood consistency with *tf-idf* neighbor weighting and key point size ratio weighting. It is formulated as histogram weighting routine for *Bhattacharyya coefficient*. Measured gain values for the proposed spatial verification are between 2.3% and 9.7%.

Acknowledgment

The work is financed from the European Union structural funds, Innovative Economy Operational Program, POIG.01.01.02-14-013/09.

References

1975

[1] G. Salton, A. Wong and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, vol. 18(11), 1975, pp. 613–620.

1996

[2] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. 4th European Conference on Computer Vision (ECCV'96), Cambridge (UK), 1996, pp. 683–695.

[3] C. Schmid and R. Mohr. Object recognition using local characterization and semi-local constraints. Technical report, INRIA, 1996.

1997

[4] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5), 1997, pp. 530–535.

2003

[5] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03), vol. 2, 2003, pp. 1470–1477.

2004

[6] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. International Journal of Computer Vision, vol. 60, 2004, pp. 63–86.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, 2004, pp. 91–110.

2006

[8] J. Zobel and M. Alistair. Inverted files for text search engines. ACM computing surveys (CSUR), vol. 38(2), 2006.

2007

[9] S. H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences, vol. 1(4), 2007, pp. 300–307.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. IEEE Conference on Computer Vision and Pattern Recognition, 2007.

2008

[11] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman. Lost in quantization: improving particular object retrieval in large scale image databases. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[12] O. Chum, J. Philbin and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. Proceedings of the British Machine Vision Conference, 2008, pp. 812–815.

[13] R. Farivar, D. Rebolledo, E. Chan and R. H.Campbell. A parallel implementation of k-means clustering on GPUs. The 2008 International Conference on Parallel and Distributed Processing Techniques and Applications, 2008, pp. 340–345.

2009

[14] O. Chum, M. Perdoch and J. Matas. Geometric min-hashing: finding a (thick) needle in a haystack. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 17–24.

2010

[15] H. Jegou, M. Douze and C. Schmid. Improving bag-of-features for large scale image search. International Journal of Computer Vision, vol. 87, 2010, pp. 316–336.

[16] D. D. Yang and A. Sluzek. A low-dimensional local descriptor incorporating TPS warping for image matching. Image and Vision Computing, Vol. 28(8), August 2010, pp. 1184–1195.

2011

[17] M. Paradowski and A. Sluzek. Local keypoints and global affine geometry: triangles and ellipses for image fragment matching. Innovations in Intelligent Image Analysis (eds. H. Kwasnicka, L. Jain), Springer Verlag, vol. 339, 2011, pp. 195–224.

2012

[18] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2911–2918.