Bag-of-colors for improved image search
Christian Wengert, Matthijs Douze, Hervé Jégou

To cite this version:

HAL Id: inria-00614523
https://inria.hal.science/inria-00614523
Submitted on 12 Aug 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Bag-of-colors for improved image search

Christian Wengert
kooaba AG
Förrlibuckstrasse 178
CH-8005 Zürich

Matthijs Douze
INRIA Grenoble
655 avenue de l’Europe
F-38334 Saint Ismier Cedex

Hervé Jégou
INRIA Rennes
Campus de Beaulieu
F-35042 Rennes Cedex

ABSTRACT
This paper investigates the use of color information when used within a state-of-the-art large scale image search system. We introduce a simple yet effective and efficient color signature generation procedure. It is used either to produce global or local descriptors. As a global descriptor, it outperforms several state-of-the-art color description methods, in particular the bag-of-words method based on color SIFT. As a local descriptor, our signature is used jointly with SIFT descriptors (no color) to provide complementary information. This significantly improves the recognition rate, outperforming the state of the art on two image search benchmarks. We will provide an open source package of our signature.

1. INTRODUCTION
We consider the problem of content-based image retrieval, for applications such as object recognition or similar image retrieval. This problem requires to produce a description of the image, typically a fixed-size vector. This description must be discriminative, but sufficiently invariant to handle the transformations the image may have undergone (cropping, different viewpoints, rotation, etc).

Most of the recent state-of-the-art large scale image search systems [7,3] rely on local descriptors, in particular the SIFT descriptors [4] and its variants. They are extracted from regions of interest using a scale-invariant detector [4,6,5]. These descriptors are often used jointly with the bag-of-words (BOW) framework [11], which produces a single vector from a set of local descriptors. Surprisingly and despite its assessed relevance for image classification [13], most of the recent papers on large scale image search ignore color information. The best results reported for this task on popular benchmarks, such as the UKB [7] or Holidays datasets [1], which both contain colorful images, are reported for techniques based on grey-level SIFT descriptors only.

Related work on color descriptors: The body of literature dealing with color descriptors is vast. Some of these descriptors have been normalized in the MPEG7 standard. The color version of the GIST descriptor [8] is widely used for scene recognition. For the sake of conciseness, two color signatures will be used as reference, namely color GIST [8] and color SIFT [13]. The color SIFT descriptor is the counterpart of SIFT [4] with color integrated. Its best variant [13] is the OpponentSIFT [14]. Similar to SIFT, it describes regions of interest extracted by a detector. It has mainly been considered for image and video classification tasks. For this reason, it is used in conjunction with the BOW aggregation framework [11]. However, this paper shows that much better results can be achieved using these descriptors by combining them with the state-of-the-art local descriptor matching method of [3].

The main contribution of our paper is a simple color description method, which is used to produce either global or local descriptors. In the first case (global), computing the image signature is much faster than producing a BOW vector with [13]: the main bottleneck is the computation of a color histogram, which is fast. Most importantly, improved search results are obtained on the evaluation datasets compared to other global descriptors.

The paper is organized as follows. Section 2 introduces our color signature generation procedure and analyzes the impact of several parameters and variants. As a global descriptor, our signature outperforms other global descriptors, including BOW generated from SIFT or color SIFT descriptors. Section 3 explains how the signature is used to describe local patches. Our color signature used jointly with SIFT descriptors outperforms the state of the art for the same memory usage and complexity. It is also better than color SIFT used within a comparable search system.

2. BAG-OF-COLORS SIGNATURES
In this section, we propose a method to extract a color signature, called bag-of-colors (BOC). This signature is extracted either for the whole image, producing a global descriptor, or from a patch (extracted by a region detector [6,5]), in which case we obtain local color descriptors.

Since the method relies on computing a color histogram, the proposed signature is conceptually close to the work of [12]. However, it departs from this method in several ways that drastically improve the recognition results. First, instead of using a regular color partition as in [12], the color codebook is learned on a set of real-world images. Most importantly, the histogram is constructed and normalized in a way inspired by the successful ingredients of the BOW framework [11] and recent improvement proposed for the Fisher kernel [10], but here applied to color components. As a result, the signature is more robust and avoids the typical problem of [12], where the most frequent colors dominate the other colors in the final representation and deteriorate the performance.

Hereafter, we introduce the signature generation procedure, analyze its results, and finally give a comparison with a BOW approach which shows that our signature provides consistently better results.

2.1 Color descriptor pipeline
Our color signature generation pipeline includes several steps. Different choices of parameters are investigated and evaluated.

Color space: Similar to most color descriptors, we use the CIE-
Lab\(^1\) color space in all our methods. This space is preferred, e.g., the opponent space because it is more consistent with the Euclidean space structure, in particular the color dictionary generation procedure, which is based on k-means.

Color codebook: Our method needs a color dictionary \(\mathcal{C} = \{c_1, \ldots, c_{k_c}\}\), defined as a set of \(k_c\) colors (often referred to as a palette). We first consider building a codebook by quantizing each \(L\), \(a\) and \(b\) components separately and regularly in 4 or 8 bins, leading to \(k_c = 64\) or 512 colors in total.

Then, we also learn a color dictionary that is more adapted to natural images. It is obtained by selecting and clustering typical colors from real-world images. In order to reduce the impact of some images containing large uniformly colored areas, we propose the following procedure in order to collect typical and unique colors for producing the color dictionary.

1) Collect 10000 random images from Flickr.
2) Resize each image to 256\(\times\)256 pixels, convert it to CIE-Lab and split it in blocks of 16\(\times\)16 pixels (i.e., 256 blocks in total).
3) For each block, find the most occurring color. Ties are randomly resolved. If this color corresponds to less than 5 occurrences (out of 256), then select an arbitrary color from the block.
4) At this point, we have extracted 256 Lab colors per image. The set of 256\(\times\)10000 colors from all images is clustered using a k-means algorithm, producing \(k_c\) Lab colors palette.

The palettes learned by our procedure are shown in Figure 1 for different values of \(k_c\). Some colors (e.g., dark grey/brown) are more represented than others in the palette, due to the unbalanced color statistic of natural images.

Color signature: The first step of our color signature method consists in computing the histogram of colors in the image for the fixed color codebook \(\mathcal{C}\). This is done by first resizing (isotropically scaling) the image to a fixed number (\(\approx 16384\)) of pixels. For each pixel \(p\) we select the closest color (w.r.t. the Euclidean distance) in the codebook \(\mathcal{C}\), and increment the corresponding bin of the output \(k_c\)-dimensional histogram.

Inverse document frequency is a successful technique in BOW. It downweights the contribution of the most common visual words and increases the importance of rare visual words by multiplying the frequency component with its corresponding \(idf\) weighting term. Similar to this technique, we update the histogram components by applying color \(idf\) weighting terms.

Normalization: power-law and normalization. Two techniques are considered. First, the “power-law” transformation proposed in [10] for Fisher vectors is applied component-wise. For a vector \(x = (x_1, \ldots, x_d)\), it consists in updating each component as \(x_i := \sqrt{x_i}\). Although an Euclidean normalization is often adopted in the BOW framework (e.g., in [11, 1]), the L1 normalization provides consistently better results for our color signatures.

Comparison metric: Signatures are compared to find the nearest neighbors of the query image in a database of signatures. The choice of the comparison metric is therefore critical. We compare the simple Manhattan (L1) and Euclidean (L2) distances, as well as distances specifically used for histograms: the fast Earth-Mover-Distance of [9], the \(\chi^2\) distance, the Shannon-Jenson divergence (SID). The histogram intersection (HI) which were used in [12] for color histograms are equivalent to the L1 distance if the vectors are L1-normalized.

2.2 Evaluation

To motivate our final method, we analyze the impact of the choices involved in the bag-of-colors generation process. The search quality is evaluated as the mAP on the INRIA Holiday dataset [1], computed by the evaluation package available online. In this section, for the sake of exposition, our analysis is conducted on global descriptors only. The results for local descriptors obtained with the same signature generation method are analyzed in Section 3.

Dictionary, normalization and comparison metric have an impact that is summarized in Table 1. First observe that a learned vocabulary gives much better results that the original scheme by Swain which regularly splits the \(L\), \(a\) and \(b\) axes (leaving many bins empty, as the Lab colorspace is not a cube). Second, the power-law transform has an important impact: an improvement of at least +8% of mAP is achieved by this simple component-wise preprocessing, showing that it is important not to consider dominant colors only. Third, the \(idf\) weighting provides a slight improvement of 1 point of mAP, comparable to what is typically observed for the standard

---

\(^1\)More precisely, we used the ITU-R recommendation BT.709 using the D65 white point reference.
Table 2: Impact of the palette size on accuracy (mAP on Holiday), signature extraction time and search time for querying a dataset of 1 million vectors (one processor core).

<table>
<thead>
<tr>
<th>mAP</th>
<th>ms/signature</th>
<th>ms/search</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.2</td>
<td>22</td>
<td>31</td>
</tr>
<tr>
<td>53.8</td>
<td>25</td>
<td>59</td>
</tr>
<tr>
<td>57.1</td>
<td>27</td>
<td>112</td>
</tr>
<tr>
<td>61.0</td>
<td>47</td>
<td>223</td>
</tr>
<tr>
<td>63.8</td>
<td>65</td>
<td>432</td>
</tr>
<tr>
<td>64.9</td>
<td>128</td>
<td>851</td>
</tr>
<tr>
<td>64.9</td>
<td>220</td>
<td>1692</td>
</tr>
</tbody>
</table>

Table 3: Comparison with the state of the art. quality: mAP on Holiday and score/4 on UKB (higher is better). All schemes are compared without the spatial verification re-ranking scheme. N/A means that the image is not represented by a single vector of fixed length \( D \) but by a set of descriptors’ signatures (typically requires 30kB/image in memory).

<table>
<thead>
<tr>
<th>Method</th>
<th>base descriptor</th>
<th>( D )</th>
<th>Holidays</th>
<th>UKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW [3]</td>
<td>SIFT</td>
<td>20,000</td>
<td>46.9</td>
<td>2.88</td>
</tr>
<tr>
<td>BOW [5]</td>
<td>SIFT</td>
<td>200,000</td>
<td>57.2</td>
<td>2.95</td>
</tr>
<tr>
<td>HE [3]</td>
<td>SIFT</td>
<td>N/A</td>
<td>74.5</td>
<td>3.30</td>
</tr>
</tbody>
</table>

Proposed approach

| BOC | Lab pixels | 256 | 63.8 | 3.34 |
| LBOC | SIFT+BOC | N/A | 78.9 | 3.50 |

In this section, our color signature is applied directly on patches and used in conjunction with SIFT descriptors [4] to provide a rich combination of both color and texture, which was shown successful in the recent color extension of SIFT descriptors [13]. It is integrated within our own implementation of the image matching method of [3], called Hamming Embedding (HE), which is a state-of-the-art extension of BOW.

3.1 Binary signatures

In [3], the SIFT descriptors are assigned to visual words, but are additionally described by a binary signature of 64 bits. Each binary signature is computed from the SIFT descriptor, complementing the information given by the visual word. The Hamming distance between binary signatures is used to remove non-matching feature pairs and assign a voting score to the selected matches.

Since this distance removes mismatches, it offers an alternative to the dramatic increase in visual vocabulary sizes required by other approaches (up to 1M visual words in [7]). It is also an elegant way of taking into account the density variations in descriptor space. Indeed, the Hamming distance is implicitly normalized with respect to the local density in the BOW quantization cell. This property facilitates the early fusion of heterogeneous binary signatures, in our case the binary color signature described below.

3.2 Color binary signatures

Our approach follows the guidelines of [3], but instead of using the 64 bits to describe the SIFT, some of the bits are dedicated to represent color information. A local BOC descriptor is extracted from the elliptical region on the image that is used to compute the SIFT descriptor. We use a palette of \( k_c=256 \) colors. The vector is then binarized into a short binary signature of fixed size. The Hamming distance between two binary color signature should, at best, reflect the similarity between the original BOC descriptors.

The 256-dimensional BOC vector needs to be reduced to a small number of bits \( b_c \) (\( b_c = 0.64 \), depending on the balance between SIFT and color). This descriptor tends to be sparse because it is computed on small regions: on average 80% of its components are 0. We therefore adopt a dimensionality reduction inspired by [2], which was shown effective to reduce sparse BOW vectors. We first generate a sparse binary matrix \( B \) of size \( k_c \times k_c \) such that each row and each column of \( B \) sum up to 8, and keep only the first \( b_c \) rows, producing a sparse binary matrix \( B' \) of size \( b_c \times k_c \).

The color binary signature is obtained from a given local BOC

\( k_c \quad 16 \quad 32 \quad 64 \quad 128 \quad 256 \quad 512 \quad 1024 \)

\( \text{mAP} \quad 41.2 \quad 53.8 \quad 57.1 \quad 61.0 \quad 63.8 \quad 64.9 \quad 64.9 \)

\( \text{ms/signature} \quad 22 \quad 25 \quad 27 \quad 47 \quad 65 \quad 128 \quad 220 \)

\( \text{ms/search} \quad 31 \quad 59 \quad 112 \quad 223 \quad 432 \quad 851 \quad 1692 \)
that color is more important than SIFT on this benchmark. The trade-off between color and texture is shown by the large scale experiments and 2.4% points of mAP, respectively. The comparison with the state of the art is shown by merging the reference method (SIFT+HE) and color SIFT by 4.4% best results are clearly obtained with our LBOC approach, outperforming the SIFT descriptor. Therefore one should not conclude that color is more important than SIFT on this benchmark.

Figure 2: Trade-off between SIFT and color on Holiday and UKB. Each detected region is represented by 64 bits in total, shared by the representation of the SIFT and the LBOC descriptor. The performance scale on the left is for the Holiday dataset, on the right it is for UKB.

Vector $x$ by first computing the projected vector $x' = B' \times x$. Second, each component of $x'$ is compared to the median value for this component (learned on a training set), which produces a bit. Note: From our experiments, this strategy is better than using a projection obtained by principal component analysis. This might be due to the fact our sparse projection is more consistent with the L1 distance, which is clearly better than L2 for BOC vectors. It is also better than binarizing a BOC vector generated by a palette comprising $k_c = b_c$ colors.

3.3 Experiments

As in Section 2.2, our evaluation is performed on Holiday and UKB. To obtain a fair comparison, we have adjusted the threshold of the color SIFT detector so that, on average, it approximately produces the same number of descriptors per image, as shown below:

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Holiday</th>
<th>UKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hessian-Affine + SIFT</td>
<td>2988</td>
<td>1048</td>
</tr>
<tr>
<td>color SIFT</td>
<td>3244</td>
<td>1066</td>
</tr>
</tbody>
</table>

The trade-off between color and texture is shown by Figure 2 for a fixed signature of 64 bits, where $b_c$ bits are used for color and the remaining $64 - b_c$ bits are the ones defined for SIFT by HE. The importance of color clearly depends on the dataset: SIFT alone performs better than the LBOC descriptor on Holiday, but this observation is reversed for UKB\(^2\). Overall, the optimum is clearly obtained when sharing the bits between these descriptors, with approximately 70 to 75% of the signature dedicated to SIFT and the rest used to represent color.

The comparison with the state of the art is shown by Table 3. Abusing the notation, the method that uses jointly SIFT and our binarized color signature is denoted by LBOC (local bag-of-colors). The color SIFT descriptor provides better results than SIFT on Holiday. However on UKB this descriptor suffers from not being invariant to rotation. The results of HE are significantly better than those obtained with BOW, as reported in the literature. The best results are clearly obtained with our LBOC approach, outperforming the reference method (SIFT+HE) and color SIFT by 4.4% and 2.4% points of mAP, respectively.

Large scale experiments have been performed on Holiday merged with 100K images from Flickr. With SIFT+LBOC, we obtained mAP=65.3% against 61.2% for HE with SIFT only. Querying the inverted file takes 0.56 s using a single processor core (2.5Ghz). This timing and the memory usage, about 30 kB/image, do not depend on the input descriptor (SIFT, LBOC or a combination).

4. CONCLUSION

We have proposed a color signature inspired by the bag-of-words framework, and used it either as a global or a local descriptor. Our experiments show that color information significantly improve the retrieval quality in large scale image search, which is often ignored in this context. Our global descriptor exhibits very competitive performance compared to more sophisticated methods. Used as a local descriptor, the results are better than those obtained by a state-of-the-art system based on SIFT or color SIFT descriptors.

Acknowledgments:

This work was realized as part of the Quaero Project, funded by OSEO, French State agency for innovation.

5. REFERENCES