Improving the Discriminative Power of Bag of Visual Words Model

Achref Ouni[1], Thierry Urruty[1(✉)], and Muriel Visani[2]

[1] XLIM, UMR CNRS 7252, University of Poitiers, Poitiers, France
{achref.ouni,thierry.urruty}@xlim.fr
[2] Laboratory L3i, University of La Rochelle, La Rochelle, France
muriel.visani@univ-lr.fr

Abstract. With the exponential increase of image database, Content Based Image Retrieval research field has started a race to always propose more effective and efficient tools to manage massive amount of data. In this paper, we focus on improving the discriminative power of the well-known bag of visual words model. To do so, we present n-BoVW, an approach that combines visual phrase model effectiveness keeping the efficiency of visual words model with a binary based compression algorithm. Experimental results on widely used datasets (UKB, INRIA Holidays, Corel1000 and PASCAL 2012) show the effectiveness of the proposed approach.

Keywords: Bag of visual words · Visual phrases · Image retrieval

1 Introduction

Content Based Image Retrieval (CBIR) has been an active field in the last decades. The massive amount of data available today has highlighted the needs for efficient and effective tools to manage this data. One important topic from CBIR field is the construction of the *image signature*. Indeed, image signature is at the core of any CBIR system. An accurate and discriminative signature will improve the precision of the retrieval process. It will also help to bridge the well-known *semantic gap* issue between low-level features and the semantic concepts a user perceived in the image.

Among the numerous state-of-the-art approaches that have tried to "narrow down" the semantic gap, some of them have improved the descriptive power of visual features [4,16], improving gradually the existing local and global image descriptors, while others have proposed effective ways to use, mix and optimize the use of these features. Among them, the bag of visual words model (BoVW) [3,15] has become a reference in CBIR. The BoVW model represents images as histograms of visual words, enhancing the retrieval efficiency without losing much accuracy.

More recently, some researchers have stated that the BoVW discriminative power was not enough and have proposed to construct visual phrases or bags

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 245–256, 2017.
DOI: 10.1007/978-3-319-51814-5_21

of bags of words by structuring visual words together using different means. However, Bag of visual phrases models [11,14,18] are computationally expensive. In this paper, we present a novel framework, called n-BoVW, to increase the discriminative power of the BoVW model. n-BoVW uses the idea of visual phrases by selecting multiple visual words to represent each key-point but keeps the efficiency of BoVW model with a binary based compressing algorithm. Two methodologies are proposed and combined with the BoVW model to obtain our final image representation. Our experimental results on different datasets highlight the potential of our proposal.

The remainder of this article is structured as follows: we provide a brief overview of bag of visual words and phrases related works in Sect. 2. Then, we explain our different proposals in Sect. 3. We present the experiments on 3 different datasets and discuss the findings of our study in Sect. 4. Section 5 concludes and gives some perspectives to our work.

2 State of the Art

We present in this section a brief overview of the literature of CBIR field that is linked to the BoVW model proposed by Csurka et al. [3]. Its inspiration comes from the Bag of Words model [5] of the Information Retrieval domain. BoVW model contains four main parts in its retrieval framework. For all images, feature detection and extraction has to be done. These two steps detect a list of key-points with rich visual and local information and convert this information into a vector. Many visual descriptors have been created, among them the Scale Invariant Feature Transform (SIFT) [9] and Speeded-up Robust Features (SURF) [2] became two of the most popular descriptors. Then, an off-line process extracts the visual vocabulary, a set of visual words, using a clustering algorithm on the set of visual features. Finally, each key-point of each image is assigned to the closest visual word of the vocabulary. Thus, each image is represented by a histogram of visual word frequencies, i.e. the image signature.

Inspired by the BoVW model, Fisher Kernel [12] or Vector of Locally Aggregated Descriptors (VLAD) [7] have met with great success. The first approach proposed by Perronnin and Dance [12] applies Fisher Kernels to visual vocabularies represented by means of a Gaussian Mixture Model (GMM). VLAD has been introduced by Jégou et al. [7] and can be seen as a simplification of the Fisher kernel. The idea of VLAD is to assign each key-point to its closest visual word and accumulate this difference for each visual word.

Recently, some researchers have focused on improving the discriminative power of the BoVW model. Thus, they have proposed to construct visual phrases or groups/bags of bags of words. Among them, we can cite the work of Yang and Newsam [18] or Alqasrawi et al. [1] who have used the spatial pyramid representation [8] to construct visual phrases from words spatially close or co-occurring in the same sub-region. They obtained good results for classification purposes. Ren et al. [14] have extended the BoVW model into Bag of Bags of Visual Words. They have proposed an Irregular Pyramid Matching with the

Normalized Cut methodology to subdivide the image into a connected graph. Other researchers have chosen to mix several vocabularies with different image resolutions as Yeganli et al. [19].

Most of those methodologies, by considering more meaningful words combinations, reach a better effectiveness than the original BoVW model. However, this improved performance can be reached only at the cost of a lower efficiency, as the processes for extracting/matching word combinations are generally quite costly.

3 Approach

In this section, we first describe our global framework before we detail our contributions. Our main objective is to improve the BoVW model discriminative power without losing much efficiency in the retrieval process. Thus, we use a common CBIR framework without any filtering on image or refining process on the used visual features nor the constructed vocabularies. It insures the reproducibility of our results. As most CBIR systems using the BoVW model are similar, we have a standard off-line learning process to construct the visual vocabulary on a separate dataset.

Figure 1 presents the different steps of our global framework. In the top part of Fig. 1, we find the detection and extraction steps for each image of the dataset. Then, using the visual vocabulary constructed previously, we proposed three different ways to construct the image signature. First "line" is the standard BoVW model that gives for each image a histogram of visual word frequencies as signature (which will be binarized to be combined). The second and third "lines" is our first contribution, an approach we denote n-Bag of Visual Words (n-BoVW). n-BoVW selects n visual words from the vocabulary to represent each detected key-point by a visual phrase. Two different methodologies are studied: (i) selecting the n closest visual words from a key-point (second "line" in the image) and (ii) clustering n nearest key-points together in the visual feature space to obtain a list of n visual words, one word by key-point inside the small cluster. For both proposals, our second contribution is a binary based compression process used to ensure an efficient retrieval. Thus, both methodologies also represent each image by a histogram of frequencies. A final combining step is also proposed to construct the final signature of the image. This step mixes the three obtained histograms to improve the discriminative power of the image signature. The following subsections detail these proposals.

3.1 *n*-Bag of Visual Words Methodologies

As visual phrases group visual words together to be more discriminative, the first contribution of this paper presents two different methodologies to better describe or represent each key-point by n visual words. Note that visual phrase models from the literature take usually n words from different key-points with the objective to better represent the near sub-region. Our approach differs as

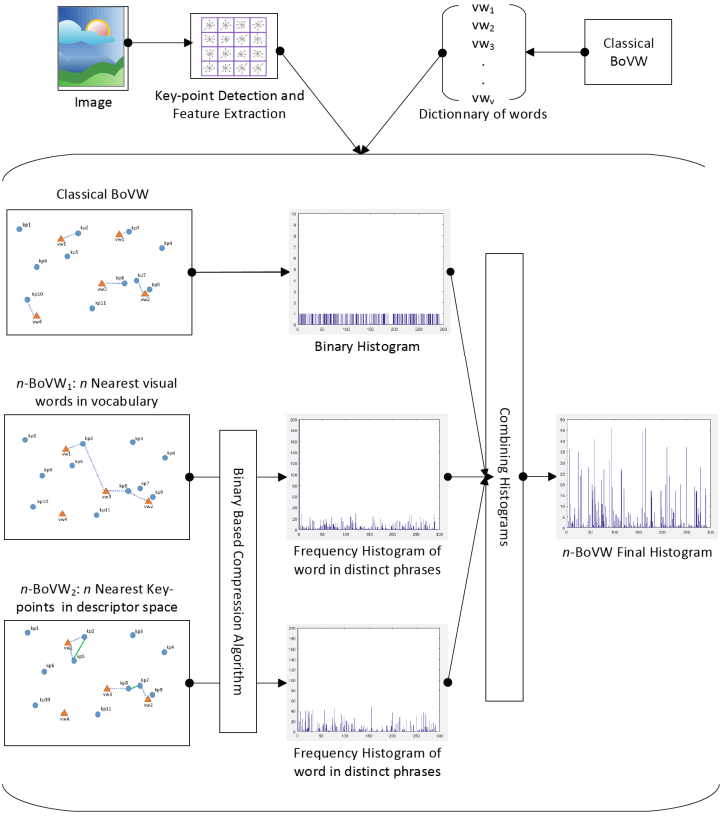


Fig. 1. Global framework

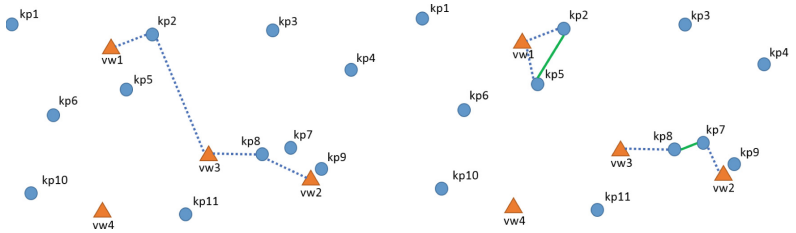
we aim at providing a more precise description of each key-point using a small vocabulary size.

The first methodology we propose is to select n visual words from the same visual vocabulary to represent each key-point of an image, referred as n-BoVW$_1$. Let W denote the vocabulary of v visual words $vw_1, \ldots, vw_v$ constructed using an offline process on a separate dataset. Let KP_i be the set of key-points extracted by the detection step for image i, with kp_{ip} the p-th key-point of image i. For each kp_{ip}, we compute the Euclidean distance ($dL2$) between the key-point and each visual word vw_j from the vocabulary W.

$$dL2(kp_{ip}, w_j) = \sqrt{\sum_{d=1}^{dim} (f_{ip_d} - vw_{j_d})^2}, \tag{1}$$

where f_{ip_d} is the d-th value of the extracted visual feature f of dimension dim for kp_{ip}.

Then, W is sorted according to these distances in order to pick the n nearest visual words from kp_{ip}. Thus, for each key-point kp_{ip}, we obtain a visual phrase $vp1_{ip}$, i.e. a set of n distinct visual words $vw1_{1_{ip}}, \ldots, vw1_{n_{ip}}$. An example is given Fig. 2(a) with $n = 2$. We can see $kp2$ two nearest visual words are $vw1$ and $vw3$, thus $kp2$ is represented by the visual phrase $(vw1, vw3)$, similarly, $kp8$ is represented by $(vw2, vw3)$.



(a) n-BoVW$_1$ (b) n-BoVW$_2$

Fig. 2. Examples for n-BoVW$_1$ and n-BoVW$_2$

With this first methodology, we ensure the description of a key-point by a visual phrase of n distinct words. However, it never takes into account the possibility that the key-point could be better represented by only one and unique word.

Our second proposal, referred as n-BoVW$_2$, is based on this non-possibility we mention and also the fact that a key-point description could be a bit noisy, thus it is interesting to look at his surrounding directly in the descriptor (or visual feature) space. A bit as strong clustering algorithm works, we gather the nearest key-points in the visual feature space to form a strong choice of visual words. Each of those selected key-points is then linked to only one visual word. Note that, the probability to have nearest visual features represented by similar visual words is high. So, it allows the possibility to have only one representative visual word for the small cluster of key-points.

For each key-point kp_{ip}, we compute the Euclidean distances with KP_i, i.e. the other key-points of image i. These distances are then sorted in order to retrieve the n nearest key-points in the visual feature spaces (including the current key-point itself). This set of nearest key-points NKP_{ip} is then used to select the representative visual words. For each key-point of NKP_{ip}, its nearest visual word is calculated using the L2-distance. At the end, for each key-point kp_{ip} we also obtain a visual phrase $vp2_{ip}$, i.e. a set of n visual words $vw2_{1_{ip}}, \ldots, vw2_{n_{ip}}$ with a high probability of duplicates. An example is given in Fig. 2(b). We can see $kp2$ nearest key-point is $kp5$, both key-points are represented by $vw1$, so we have only $vw1$ to represent $kp2$ which is different from first method. However, $kp8$ (with the link to its nearest neighbor $kp7$) is still represented by the same visual phrase $(vw2, vw3)$.

Algorithm 1. Global approach

procedure CREATEIMAGESIGNATURE
  $hisBoVW_i \leftarrow init()$; ▷ BoVW histogram
  $hisNBoVW1_i \leftarrow init()$; ▷ n-BoVW$_1$ histogram
  $hisNBoVW2_i \leftarrow init()$; ▷ n-BoVW$_2$ histogram
  $FinalHisNBoVW_i \leftarrow init()$; ▷ image i signature
  $vp1_i \leftarrow init()$ ▷ set of visual phrases from n-BoVW$_1$
  $vp2_i \leftarrow init()$ ▷ set of visual phrases from n-BoVW$_2$
  for kp_{ip} in KP_i **do** ▷ for all key-points of an image
    $hisBoVW_i \leftarrow hisBoVW_i + computeNearestVisualWords(kp_{ip}, W, 1)$;
    $vp1_i \leftarrow vp1_i + computeNearestVisualWords(kp_{ip}, W, n)$;
    $NKP_{ip} \leftarrow computeNearestKeypoints(kp_{ip}, KP_i, n)$;
    for nkp_{ip} in NKP_{ip} **do**
      $vp2_i \leftarrow vp2_i + computeNearestVisualWords(nkp_{ip}, W, 1)$;
  $hisNBoVW1_i \leftarrow BinaryBasedCompression(vp1_i)$;
  $hisNBoVW2_i \leftarrow BinaryBasedCompression(vp2_i)$;
  $FinalHisNBoVW_i \leftarrow CombHis(hisNBoVW1_i, hisNBoVW2_i, hisBoVW_i)$;

3.2 Binary Based Compression

The main disadvantage of having such visual phrases from our two proposed methodologies is the number of phrase possibilities, i.e. $\frac{v!}{(v-n)!n!}$ which will be computationally too high for a retrieval system. To deal with this phenomenon, we propose a binary based compression algorithm that is used for both proposals.

We first noticed in literature approaches that visual phrases of only 2 words give better performance [1,18]. So, for each key-point visual phrase vp_{ip} of n words, we construct all possible combinations of 2 visual words. Then, we also observed that for BoVP model approaches with a high number of phrases, only the presence or the absence of the visual phrase is enough to be discriminative. Thus, we decide to binarize the presence of visual phrases in one image. The final step of our compression methodology sums the presence of a word in distinct visual phrases.

The results of our proposal is an histogram of v bins as image signature which is similar to the BoVW model. However, in our approach the histogram contains the frequencies of a word appearing in distinct visual phrases extracted at each key-point.

Algorithms 1 and 2 give the global approach with the binary based compression method. Some part of those algorithms are more detailed to be easier to understand but are obviously optimized in our real code. Of course, the information gathered from both methodologies described previously is different, even from the standard BoVW model. Thus, it is relevant to try and combine the histograms from BoVW model and both n-BoVW methodologies. Out of the different solutions we have tried to combine these histograms, adding the occurrences of visual phrases together before going through the binary based compression process has given the best results. Our exhaustive experimental results are discussed in the next section.

Algorithm 2. Binary Based Compression Algorithm

```
function BinaryBasedCompression(VP)
    hisVP ← init();                    ▷ histogram of presences of words in phrases
    binaryVP ← init();                 ▷ binarized visual phrases of 2 words
    for vw_j in VP do
        for vw_k in VP, k >= j do
            tempVP ← (vw_j, vw_k);
            if tempVP is not in binaryVP then
                binaryVP ← binaryVP + tempVP;
    for v1 in W do
        for v2 in W, v1 >= v2 do
            tempVP ← (v1, v2);
            if tempVP in binaryVP then
                hisVPv1++
                hisVPv2++
    return hisVP
```

4 Experimental Results

In this section, we present the experiments done to highlight the potential of our approach. To evaluate our different propositions, 2 low level visual features, Speeded-up Robust Features (SURF) and Color Moment Invariant (CMI), and 3 datasets were considered:

University of Kentucky Benchmark which has been proposed by Nistér and Stewénius [10] is referred as UKB to simplify the reading. UKB contains of 10200 images divided into 2550 groups, each group consists of 4 images of the same object with different conditions (rotated, blurred...). The score is the mean precision over all images for 4 nearest neighbors.

INRIA Holidays [6], referred as Holidays, is a collection of 1491 images, 500 of them are query images, and the remaining 991 images are the corresponding relevant images. The evaluation on Holidays is based on mean average precision score (mAP) [13].

Corel1000 or Wang [17], referred as Wang, is a collection of 1000 images of 10 categories. The evaluation is the average precision of all images for first 100 nearest neighbors.

To construct the initial visual vocabulary, we used the *PASCAL VOC 2012* [4] containing 17225 heterogeneous images categorized into 20 object classes. We use a visual vocabulary of 500 words for each descriptor.

4.1 Performance of n-BoVW

First, we study the effect of the parameter n. Figure 3 shows the performance of the retrieval on the 3 datasets with SURF descriptor. It clearly indicates that $n > 2$ has very little interest ($n = 3$ is better only on Holidays for SURF). Similar observations have also been noticed with CMI descriptor for both methodologies: sometimes, for $n > 2$, the precision is stable, sometimes it drops little by little.

This results is similar to literature visual phrases results where most approaches construct visual phrases of 2 words [1,18]. Thus, we decide to focus the following experiments with $n = 2$ even if $n = 3$ could give small improvement with a specific dataset and descriptor.

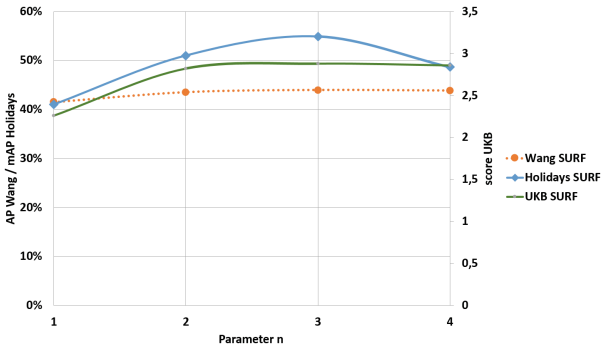


Fig. 3. Study of the effect of parameter n, number of visual words in phrases, for SURF

The next experiments evaluate the performance of the two proposed methodologies. Table 1 presents the performance of using the 2 nearest visual words to represent one key-point, referred as 2-BoVW$_1$. Note that CMI.SURF denotes the concatenation at the end of the process of the two final histograms for retrieval, and 2-BoVW$_1$ + BoVW denotes the addition of BoVW frequencies before the binary based compression step. As one can observe, 2-BoVW$_1$ methodology outperforms the BoVW model in almost all scenarios and when we add the BoVW histogram, the performance is even higher. For example, a score of 3.50 (out of 4) on UKB using 2-BoVW$_1$ + BoVW with the concatenation of both descriptor histograms is very high compared to BoVW (+37%).

Table 1. 2-BoVW$_1$ performance

Dataset	Descriptor	BoVW	2-BoVW$_1$	2-BoVW$_1$ + BoVW
Wang	CMI	40%	48%	48%
	SURF	42%	43%	45%
	CMI.SURF	48%	55%	57%
UKB	CMI	2.52	3.01	3.18
	SURF	2.26	2.82	2.92
	CMI.SURF	2.55	3.41	3.50
Holidays	CMI	41%	51%	52%
	SURF	53%	56%	56%
	CMI.SURF	44%	64%	64%

Table 2 presents the good performance of using the nearest neighbor keypoints to obtain visual phrases of $n = 2$ words, referred as 2-BoVW$_2$. We observe that 2-BoVW$_2$ has almost similar results than 2-BoVW$_1$, with only a small decrease on UKB dataset. These two tables clearly highlight the interest of our proposals.

Table 2. 2-BoVW$_2$ performance

Dataset	Descriptor	BoVW	2-BoVW$_2$	2-BoVW$_2$ + BoVW
Wang	CMI	40%	46%	47%
	SURF	42%	42%	44%
	CMI.SURF	48%	54%	56%
UKB	CMI	2.52	3.08	3.04
	SURF	2.26	2.73	2.81
	CMI.SURF	2.55	3.30	3.41
Holidays	CMI	41%	52%	53%
	SURF	53%	56%	56%
	CMI.SURF	44%	65%	65%

Performance Combining Methodologies. As the two proposed methodologies present good performance but similar, we try to mix both obtained histograms together in order to check if the performance of the system could benefit from this combination. On Table 3, we observe that on UKB with single descriptor, the precision has increased. Note that we highlight in bold, the precision scores that are strictly above n-BoVW$_1$ or n-BoVW$_2$. However, it is important to notice that combining histograms never decreases the results.

Table 3. Performance combining 2-BoVW$_2$, 2-BoVW$_1$ and BoVW histograms

Dataset	Descriptor	BoVW	2-BoVW	2-BoVW + BoVW
Wang	CMI	40%	48%	48%
	SURF	42%	44%	45%
	CMI.SURF	48%	55%	57%
UKB	CMI	2.52	**3.14**	**3.20**
	SURF	2.26	**2.90**	**3.02**
	CMI.SURF	2.55	3.41	3.50
Holidays	CMI	41%	52%	53%
	SURF	53%	**57%**	**57%**
	CMI.SURF	44%	65%	65%

4.2 Discussion

The observed results show the interest of n-BoVW with the 2 methodologies we have proposed combined with the BoVW model. The precision of the retrieval is clearly higher than the BoVW alone. Most of literature approaches have indeed improved the BoVW model but needed some indexing structure to decrease the loss in efficiency for the retrieval. Constructing the image signature with our framework is obviously more complex than the BoVW model: for one image, 3 histograms are created and combined. The most complex one is the second methodologies n-BoVW$_2$ because it needs to sort all image key-points to pick n nearest ones. Constructing this histogram takes 5 times more longer than BoVW histogram. However, as we obtain an image signature of the same size (vocabulary size) than the BoVW one, the increase in complexity has little effect in the global retrieval process. Extracting the descriptor, and searching for nearest neighbors in the dataset are still preponderant processes.

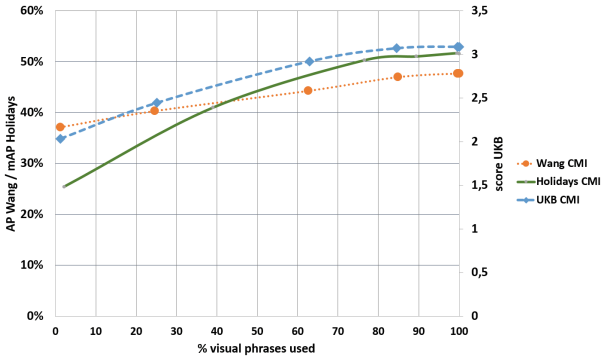


Fig. 4. Performance of n-BoVW with respect to the % of visual phrases used for CMI

Table 4. n-BoVW vs. other methods

Method	UKB score	Wang AP	Holidays mAP
BoVW [3]	2.95	48%	53%
Fisher [12]	3.07	na	69.9%
VLAD [7]	3.17	na	53%
n-Grams [11]	na	34%	na
n-BoVW	3.50	57%	65%

Another point of discussion we highlight is the possibility that for more than one word to describe a key-point could add noise in the image description. Thus, we have tried to put a distance threshold in our algorithm. The visual

phrases constructed with words too "far" should not be taken into consideration, replacing the visual phrase by only one word. Figure 4 presents the observed results on the 3 datasets with respect to the percentage of visual phrases used for CMI descriptor. Note that SURF results are similar. The results are a bit surprising because best results are achieved with a percentage of visual phrases close to 100%. Thus, we may conclude all visual phrases are needed in n-BoVW even if results still improve when combining with BoVW. Finally, we compare our approach against few state-of-the-art methods in Table 4. We give here results given by authors when available (*na* when not available). We observe easily that our proposed approach mostly outperforms other recent methods.

5 Conclusion

This paper presents a more discriminative BoVW framework called n-BoVW. Two different methodologies based on visual phrases model were proposed with for both, results outperforming the BoVW model on all test datasets and with two different descriptors. Mixing these methods together with the BoVW model also improves greatly the performance. Another contribution of this paper is the proposed binary based compression method. It allows the proposed framework to have a similar computational cost than the BoVW model for retrieval. Our perspective will focus first on the notion of distance from a key-point to a visual word discussed in the previous section. We believe it could be useful to adapt automatically the parameter n for each key-point. Thus, different lengths of visual phrases could represent each key-point of the image. A study of the effect of number of visual words in the starting vocabulary would also be interesting even if increasing this number will decrease the efficiency of the retrieval framework.

Acknowledgments. This research is supported by the Poitou-Charentes Regional Founds for Research activities and the European Regional Development Founds (ERDF) inside the e-Patrimoine project from the axe 1 of the NUMERIC Program.

References

1. Alqasrawi, Y., Neagu, D., Cowling, P.I.: Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. Sig. Image Video Process. **7**(4), 759–775 (2013)
2. Bay, H., Tuytelaars, T., Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). doi:10.1007/11744023_32
3. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results (2012). http://www.pascal-network.org/challenges/-VOC/voc2012/workshop/index.html

5. Harris, Z.: Distributional structure. Word **10**(23), 146–162 (1954)
6. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_24
7. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pp. 3304–3311, San Francisco, United States. IEEE Computer Society (2010)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, USA, 17–22 June 2006, pp. 2169–2178 (2006)
9. Lowe, D.G.: Object recognition from local scale-invariant features. Int. Conf. Comput. Vis. **2**, 1150–1157 (1999)
10. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR) **2**, 2161–2168 (2006)
11. Pedrosa, G., Traina, A.: From bag-of-visual-words to bag-of-visual-phrases using n-grams. In: 2013 26th SIBGRAPI - Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 304–311, August 2013
12. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, Minnesota, USA, 18–23 June 2007. IEEE Computer Society (2007)
13. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)
14. Ren, Y., Bugeau, A., Benois-Pineau, J.: Bag-of-bags of words irregular graph pyramids vs spatial pyramid matching for image retrieval. In: 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6, October 2014
15. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, pp. 1470–1477, October 2003
16. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2010)
17. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
18. Yang, Y., Newsam, S.D.: Spatial pyramid co-occurrence for image classification. In: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V. (eds.) IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011, pp. 1465–1472. IEEE Computer Society (2011)
19. Yeganli, F., Nazzal, M., Özkaramanli, H.: Image super-resolution via sparse representation over multiple learned dictionaries based on edge sharpness and gradient phase angle. Sig. Image Video Process. **9**, 285–293 (2015)