A machine learning approach for image retrieval tasks

Achref Ouni

▶ **To cite this version:**

Achref Ouni. A machine learning approach for image retrieval tasks. International Conference on Image and Vision Computing New Zealand (IVCNZ), Nov 2020, Wellington, New Zealand. hal-03103965

HAL Id: hal-03103965
https://uca.hal.science/hal-03103965

Submitted on 8 Jan 2021

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.

A machine learning approach for image retrieval tasks

1st Achref Ouni
Université Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal
F-63000 CLERMONT-FERRAND, FRANCE
Achref.EL_OUNI@uca.fr

***Abstract*—Several methods based on visual methods (BoVW, VLAD, ...) or recent deep leaning methods try to solve the CBIR problem. Bag of visual words (BoVW) is one of most module used for both classification and image recognition. But, even with the high performance of BoVW, the problem of retrieving the image by content is still a challenge in computer vision. In this paper, we propose an improvement on a bag of visual words by increasing the accuracy of the retrieved candidates. In addition, we reduce the signature construction time by exploiting the powerful of the approximate nearest neighbor algorithms (ANNs). Experimental results will be applied to widely data sets (UKB, Wang, Corel 10K) and with different descriptors (CMI, SURF).**

***Index Terms*—ANNs, bag of visual words , bag of visual phrases , supervised classification , descriptors, Image retrieval**

I. Introduction

Content Based Image retrieval (CBIR) has made the recent years of big strides to look for images visually close to some image request, or to find a specific object in an image. However, these systems are still under performance when it comes to research semantics of images of textual query. One of the reasons of comes from the way whose images are described in computer systems.

Bag of visual words (BoVW) among the powerful algorithm of computer vision. This approach is composed of three main steps: (i) Detection and Feature extraction (ii) Codebook construction (iii) Vector quantization. Many works have been proposed for improving BoVW model such as VLAD Fisher BoVP. In this paper, we interest to improve the BoVW model by transforming to bag of visual phrase (BoVP). The main concept of visual phrase is to improve the visual description by linking two or more visual words.

To do, we used approximate nearest neighbor search algorithms (ANNs) to create a discriminative visual phrases that means a maximum approximation error bound, Thus allowing the user to control the trade off running time. Second use of ANNs is to reduce the signature construction time, especially for large databases. Also, we solve the problem of noise because the exponential numbers of visual words in region or sliding window.

Our main contributions are composed of two parts

Part I :

- Transformation of the classical model bag of visual words to bag of visual phrases
- Improve the quality of phrases without noise using recent implementation of approximate nearest neighbors(ANNs).
- Reducing the time query that means phrase will construct fastly and more robust than classical methods.
- We reduce the complexity to O(n) instead of O(n^2) during the phrase construction step by proposing a new algorithm based compression.

Part II :

- Comprehensive experimental study of state-of-the-art ANNs methods and evaluating all the methods using widely data sets with different descriptors(Surf,Cmi).
- We compare and study the algorithms performance (search time complexity, search quality, robustness).

Our work is divided into four sections: At first we provide a brief overview of bag of visual words and phrases and their mechanism in Section 2. We explain our proposals in Section 3. We present the experiments on 3 different datasets and discuss and comparison with others methods in Section 4. Section 5 concludes and gives some perspectives to our work.

II. State of the Art

Content based image retrieval(CBIR) is one of the techniques that allow to look for images by visual characteristics. In the field of search for image by the contents the detection and extraction of the characteristics since an image is made by a local descriptor such as SIFT, SURF. Last year bags of visual word[2] (BoVW) (also called bags of features or bags of key-points) have been very widely used in the computer vision community for classification and image recognition. Bow treated as follows. For each image, the visual features detected, then extracted using a visual descriptor such as SIFT [**?**]. This step will be repeated in a recursive way on all image datasets until collecting all visual descriptors datasets. Then a clustering step using K-MEANS [**?**] will be applied on the descriptors to build the visual vocabulary (visual words) from the center of each cluster. In order to obtain the visual words, the features query replaced by the index of the visual words that consider the nearest using euclidean distance. Finally, the image described as a histogram of the frequency of

978-1-7281-8579-8/20/$31.00 ©2020 European Union

the visual words exist in the image. But even the discriminative power of BoVW is still weak against new challenge.

Recently, some researchers have focused on improving BoVW model. First work presented by [3] group each key point with its nearest neighbor. The distance of the pairs of the key points is the multiplication of the distance L1. [4] group each visual word with its nearest K neighbors. Choosing the most common subsets of these groups as visual phrase. [5] use a hierarchical dictionary to quantify the descriptors. Visual phrases are constructed by grouping each descriptor with its closest neighbors. By using n-gram notion, we find [6] phrase formed by a sequence of n-consecutive words using L1 as metric between words. The recent work proposed by [7] the phrases is constructed by brute force that means nearest neighbor between values of key-point using L2. Also, we find the second contribution presented by double bag of words that mean each phrase constructed by double words instead of one. Sliding window of the bow has been introduced by [20] groups visual words in pairs, choose the pairs of visuals words satisfy the conditions of distance and frequency. Several improvements of the principle of visual phrases create by a sliding window [21] use a sliding window of a fixed radius according to the lengths of the axes of each elliptical region. For improving search by content, a lot of research focus on improving feature descriptors for robust presentation image [8-13]. Fisher Kernel[14] and VLAD[15] have met with great success. Many new learning algorithms and architectures are currently being developed using deep neural networks. AlexNet[17-19] and VGG16/VGG19 among the most used and efficient architectures of object detection and classification.

In a parallel way many search focus on speed up the search in high dimensional space when the number of descriptors is very important. State of the arts mention's five categories of algorithm ANNs: (1) Lsh based methods,(2) Encoding based methods,(3)Tree based space partition methods (4) Neighborhood based methods and (5) Diversified proximity graph. The more important of ANNs is the high precision and rapidity search in high dimensions. We introduce by locality- sensitive hashing (LSH)[22], which maps a high-dimensional point to a low-dimensional point via a set of appropriately chosen random projection functions. Encoding based methods is also a powerful algorithms for ANN likes Selective Hashing [23], Anchor Graph Hashing [24], Scalable Graph Hashing [25], Neighborhood Approximation index, and Optimal Product Quantization .

III. Approach

The proposed framework named ANN-BOVP based on the standard method BOVW[2] and ANNs[1]. In this paper, we describe an image as Bag-of-Visual-Phrases created by the most powerful ANN algorithms. So we have applied ANN algorithms on bag of visual words for creating a discriminative phrases. The power given by the incorporation between ANNs and bag of visual words improve the image retrieval and gives better results.

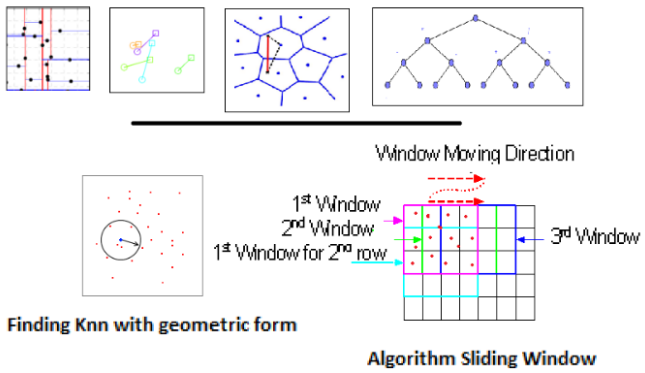

Fig. 1. ANNs algorithms vs classical methods for searching NN

Our global framework starts by detecting and extracting the features from the dataset. Then, using the visual vocabulary constructed using K-means, we create the BoVW for each image. Once we obtained the visual representation, we linked the visual words together using ANNs to create a powerful phrase.

A. Create BoVP using approximate nearest-neighbour search algorithms

Approximate nearest-neighbor search is a set of algorithms composed of five big families as described in state of the art. In this section we select three algorithms from different categories and we will apply the most powerful ANNs on bag of visual words for creating a robust and discriminative phrases.

- The first algorithm named NNP[29](Nearest neighboring particle search) from Encoding based methods category. The algorithm used is all pairs search. The distance between a concerned point and all the points is calculated and points are sorted out depending on the given length.
- The second algorithm named ClusterNN[30] from Space Partitioning-based Methods category. This algorithm uses a clustering method for finding the nearest neighbor inside clusters.
- The third algorithm named Ball-tree[31] from Tree-based space partition methods category. This algorithm uses the data as tree and nodes. Euclidean distance is the metric used for searching nearest neighbor in space partition.

Figure 1 presents the different search methods of searching to nearest neighbor for creating a bag of visual phrases. In the top part of Fig. 1, we find some example of an algorithm for searching in high dimensions. These algorithms are based on model to set up the points in space. Tree based space partition methods create a tree of values and in the end, we find the similar couples. Lsh use a hash function family such as that close points in the original space have a high probability of having the same hash value. In the bottom of figure 1 we find classical method of constructing bag of visual phrases based on geometric or sliding window to find nearest neighbors. The

principal disadvantage of the classical methods is the absence of precision which linked the visual words with an exponential number of vocabulary. Using ANNs we reduce time query that mean phrase will construct faster and robuster than classical methods with high time and complexity.

B. Algorithm based compression

In computer vision the images are represented by digital contents. However the images have several particular characteristics which intervene in their analysis. The image presented in a matrix of numeric values. Using BoVW method we can transform the keypoints to a vector of occurrence of visual words. But the transformation from BoVW to BOVP transform once again the vector to matrix. Then, we obtained a matrix for each image which most elements are zero. Most solutions for sparse matrix do not keep the size of all images in database and the new size depends on the values that different to zero. Also, learning on the histograms of size n^2 is quite complex and expensive, especially when the dimensions of the matrix are too high. This situation requires to use a powerful device, especially for massive databases with thousands of images.

We have therefore developed a compression algorithm to accelerate the comparison between the images. To do this, we transform the histogram of size n^2 into a vector of size n using standard deviation function on each occurrence of a visual word as figure 3. The occurrence of phrases not taken into consideration we memories only the presence of phrases not redundancy.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \tag{1}$$

(1) is the formula of STD Where S= the standard deviation of a simple $\sum$ means "sum of,"

- X = each value in the data set,
- $\bar{x}$ = mean of all values in the data set,
- N= number of values in the data set,

Algorithm 1 Algorithm based compression

```
Function CreateCompressionHistogram
For i=1:Number-keypoint do
    For j=1:Dim-descriptor do
    (Pos-response,value)=ANNs(Query_i,DataBase)
    (VW_1,VW_2)=Find-VW(Query_i,Pos-response);
    if(Matrix(VW_1,VW_2)==0) Matrix(VW_1,VW_2)==1;
    end
end
For Line=1:n do
    For Column=1:n do
        tab-line(Column)=Matrix(Column,Line)
        tab-column(Column)=Matrix(Line,Column)
    end
Vector(i) = STD(tab-line)+STD(tab-column)
end
Return Vector
```

This transformation has five major advantages :

- Reduce comparison time between images
- Reduce the complexity O(n) with the vector instead of O(n^2) with matrix
- Better image representation
- Solve the sparse matrix problem
- Save time

IV. Experimental Results

In this section we will show the details of dataset, algorithms used for retrieval, and evaluation results.

A. Experimental Dataset

In the work reported in this paper, retrieval tests are conducted on three different databases (Wang, Corel 10K, UKB) and results are presented separately. To evaluate our framework (ANN-BOVP) we used two different descriptors. The first is Speeded-up Robust Features (SURF) and the second is Color Moment Invariant (CMI) and three data sets:

- Wang ,[30] consists of images of natural scenes. It is composed of 1000 images divided into 10 classes, 100 images in each class.
- Corel-10k dataset [29] contains 100 categories, and there are 10,000 images from diverse contents such as sunset, beach, flower, building, car, horses, mountains, fish, food, door, etc. Each category contains 100 images of size 192*128 or 128*192 in the JPEG format.
- University of Kentucky Benchmark "UKB" [31] contains 10200 images divided into 2550 groups, each group consists of 4 images of the same object with different conditions (rotated, blurred...).

B. Performance Evaluation

In this section we present the experiments done to highlight the potential of our approach. To evaluate our different propositions we used two different descriptors Speeded-up Robust Features (SURF) and Color Moment Invariant (CMI), and three datasets. To validate the robustness of our work we used for supervised classifier algorithm. All tests with option K-Fold Cross Validation [32] option on our work we used the default value of K=5. We have used four different supervised classifiers to show the power of our framework (ANN-BOVP) .The algorithms used are SVM (support vector machine), LDA (linear discriminant analysis), KNN (K nearest neighbor). For SVM, we tested with linear kernel. The experiments evaluate the performance of our framework(ANN-BOVP) proposed . Table 1,2,3 presents the performance of NNP(Nearest neighboring particle search) , ClusterNN(Cluster nearest neighbor) and Ball-tree. As one can observe the output results are high. For example, a score of 0.93 for Wang (out of on Wang with the concatenation of both descriptor histograms is very high compared to BoVW). Svm and LDA have a better result than others KNN. As the proposed frameworks(ANN-BOVP) present good performance with different image representations, we mix the obtained histograms together.

Databases	Method	Cmi	Surf	Cmi.Surf
Wang	SVM	0.81	0.79	0.93
	LDA	0.74	0.70	0.83
	Knn	0.75	0.73	0.85
Corel 10K	SVM	0.42	0.34	0.61
	LDA	0.37	0.32	0.51
	Knn	0.28	0.18	0.40
UKB	SVM	0.84	0.76	0.91
	LDA	0.88	0.83	0.94
	Knn	0.70	0.78	0.82

TABLE I
RESULTS FOR NNP METHOD

Databases	Method	Cmi	Surf	Cmi.Surf
Wang	SVM	0.83	0.77	0.92
	LDA	0.73	0.69	0.83
	KNN	0.79	0.71	0.86
Corel 10K	SVM	0.40	0.35	0.59
	LDA	0.29	0.32	0.50
	KNN	0.26	0.20	0.37
UKB	SVM	0.83	0.66	0.92
	LDA	0.86	0.82	0.92
	KNN	0.71	0.61	0.71

TABLE II
RESULTS FOR CLUSTERNN METHOD

Databases	Method	Cmi	Surf	Cmi.Surf
Wang	SVM	0.79	0.78	0.92
	LDA	0.70	0.68	0.83
	KNN	0.75	0.73	0.85
Corel 10K	SVM	0.43	0.34	0.59
	LDA	0.36	0.31	0.50
	KNN	0.28	0.21	0.36
UKB	SVM	0.79	0.50	0.81
	LDA	0.85	0.79	0.90
	KNN	0.66	0.55	0.80

TABLE III
RESULTS FOR BALL-TREE METHOD

Our aim here is to evaluate the possible benefits of using a mixing strategy between the different image representations. We observe that concatenating the histograms tends to increase the performance of our approach. The dimension of the histogram is higher, and consequently more discriminative.

C. Discussion

The observed results show the interest of Our framework(ANN-BOVP) with the 3 different algorithms for searching in high dimensional. The precision of the retrieval is clearly higher than the BoVW alone. The framework(ANN-BOVP) proposed is capable of increasing the accuracy and speed up the search with low complexity compared with other methods.

Databases	Method	Score
Corel 1K	BoVW[16]	0.48%
	n-Grams [6]	0.34%
	AlexNet [19]	0.87%
	n-BoVW[7]	0.60%
	ANN-BOVP(best)	**0.93%**
Corel 10K	Ri-HOG[8]	0.52%
	HOG [9]	0.33%
	Gabor [10]	0.29%
	EHD[11]	0.32%
	MSD[12]	0.45%
	MTH[13]	0.41%
	ANN-BOVP(best)	**0.61%**
UKB	BoVW[16]	2.95
	Fisher [14]	3.07
	VLAD [15]	3.17
	R-CNN [17]	3.34
	AlexNet [19]	3.40
	SPoC [18]	3.63
	n-BoVW [7]	3.50
	ANN-BOVP(best)	**3.76**

TABLE IV
ANN-BOVP VS. OTHER METHODS

Finally, we compare our approach against few state-of-the-art methods in Table 4. We present here results given by authors . For Nister the score calculated per the mean precision the top score is 4 that means each image has four similar images so 1 equal 25% . So we converted the percentage to a score to compare with the results.

How created phrases	Methods	time	Complexity	setting
BoVP with classical methods	N1-BoVW[7]	0.07s	O(nlog n)	-
	N2-BoVW[7]	65s	O(n^3)	-
	S-Window[20]	41s	O(n^3)	-
	N-Gram[6]	87s	O(n^2)	-
BoVP with ANNs	NNP	0.12s	O(nlog n)	1st
	Ball-Tree	0.26s	O(n($logn^2$))	3rd
	ClusterNN	0.12s	O(nlog n)	2nd

TABLE V
COMPARISON TIME BETWEEN CLASSICAL METHODS AND BOVP WITH ANNS PER SECOND

In table 5 we study and compare the performance of each algorithm by complexity, time and parameter setting.The methods [20-7-6] are very expensive in time. The results show the importance of using ANNs for improving BoVW model.

V. CONCLUSION

In this paper, we present a new content-based image retrieval framework(ANN-BOVP) based on approximate nearest-neighbor search algorithms and bag of visual words model. A powerful phrases was created using ANNs able to improve the search by content and reduce time query. Our results show the effectiveness of our work. Mixing the histograms together also improves greatly the performance. Another contribution is that this work presents a new algorithm based compression. Experiments result shows the interest of our approach compared to deep learning based features and classical methods for Bag of visual phrases .

REFERENCES

[1] https://github.com/erikbern/ann-benchmarks, 2016 EE Bernhardsson Benchmarking nearest neighbor

[2] Csurka C. Dance LX Fan J. Willamowski C Bray (2004). "Visual categorization with bags of keypoints". Proc of ECCV International Workshop on Statistical Learning in Computer Vision.

[3] Josef Sivic and Andrew Zisserman. Video google : A text retrieval approach to object matching in videos. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pages 14701477. IEEE, 2003.

[4] Yuan, J. Wu Y. Yang, M. (2007, June). Discovery of collocation patterns: from visual words to visual phrases. In Computer Vision and Pattern Recognition 2007. CVPR'07. IEEE Conference on (pp. 1-8). IEEE.

[5] Zhang, Jianguo, et al. "Local features and kernels for classification of texture and object categories: A comprehensive study." International journal of computer vision 73.2 (2007): 213-238.

[6] Bespalov D. Bai B. Qi Y. Shokoufandeh, A. (2011, October). Sentiment classification based on supervised latent n-gram analysis. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 375-382). ACM.

[7] Ouni A. Urruty T. and Visani M. (2017, January). Improving the Discriminative Power of Bag of Visual Words Model. In International Conference on Multimedia Modeling (pp. 245-256). Springer, Cham.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 886893, Jun. 2005.

[9] G.-H. Liu Z.-Y. Li, L. Zhang, and Y. Xu. Image retrieval based on micro-structure descriptor. Pattern Recognition, 44(9):2123 2133, 2011.

[10] G.-H. Liu, L. Zhang, Y.-K. Hou, Z.-Y. Li, and J.-Y. Yang. Image retrieval based on multi-texton histogram. Pattern Recognition, 43(7):2380 2389, 2010.

[11] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. IEEE Trans. PAMI, 18(8):837842, Aug. 1996.

[12] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. IEEE Trans. Circ. Syst . Video Tech., 11(6):703715, Jun. 2001.

[13] Chen J. Nakashika, T. Takiguchi, T. Ariki, Y. (2015, June). Content-based image retrieval using rotation-invariant histograms of oriented gradients. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 443-446). ACM.

[14] 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)

[15] Jegou, H., Douze, M., Schmid, C., Perez, 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)

[16] 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)(BVW)

[17] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus,

[18] A. Babenko and V. S. Lempitsky. Aggregating local deep features for image retrieval. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 1269–1277. IEEE Computer Society, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097– 1105. Curran Associates, Inc., 2012.

[20] Sun, H., Sun, X., Wang, H., Li, Y., Li, X. (2012). Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. IEEE Geoscience and Remote Sensing Letters, 9(1), 109-113.

[21] Naeem A Bhatti and Allan Hanbury. Co-occurrence bag of words for object recognition. In Proceedings of the 15th Computer Vision Winter Workshop, pages 2128. Citeseer, 2010.

[22] Datar, M., Immorlica, N., Indyk, P., Mirrokni, V. S. (2004, June). Locality-sensitive hashing scheme based on p-stable distributions. In Proceedings of the twentieth annual symposium on Computational geometry (pp. 253-262). ACM.

[23] Gao, J., Jagadish, H. V., Ooi, B. C., Wang, S. (2015, August). Selective hashing: Closing the gap between radius search and k-nn search. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 349-358). ACM.

[24] Kim, S., Choi, S. (2013, May). Multi-view anchor graph hashing. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 3123-3127). IEEE.

[25] Siva Srinivas Kolukula, https://sites.google.com/site/kolukulasivasrinivas/

[26] Bashir, S., Doolan, D., Petrovski, A. (2015, December). Clusternn: A hybrid classification approach to mobile activity recognition. In Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia (pp. 263-267). ACM.

[27] Bhatia, N. (2010). Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085.

[28] Bengio, Y., Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research, 5(Sep), 1089-1105.

[29] 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)

[30] Guang-Hai Liu, Jing-Yu Yang, Content-Based Image retrieval using color difference histogram, Pattern Recognition, 46(1) (2013)188-198

[31] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR) 2, 2161–2168 (2006)