An Unsupervised IR Approach Based Density Clustering Algorithm

Achref Ouni[a]

Laboratoire LIMOS, CNRS UMR 6158, Universite Clermont Auvergne, 63170 Aubiere, France

Keywords: Image Retrieval, BoVW, Descriptors, Classification.

Abstract: Finding the most similar images to an input query in the database is an important task in computer vision. Many approaches have been proposed from visual content have proven its effectiveness in retrieving the most relevant images. Bag of visual words model (BoVW) is one of the most algorithm used for image classification and recognition. Even the discriminative power of BoVW, the problem of retrieving the relevant images from the dataset is still a challenge.

In this paper, we propose an efficient method inspired by the BoVW algorithm. Our key idea is to convert the standard BoVW model into a BoVP (Bag of Visual Phrase) model based on a density-spatial clustering algorithm. We show experimentally that the proposed model is able to perform better than classical methods. We examine the performance of the proposed method on four different datasets.

1 INTRODUCTION

Content based image retrieval (CBIR) is the problem of finding from a database the images that are similar to a query image. This field is a key step for many applications in computer vision as pose estimation, virtual reality, medical analysis. CBIR is based on the robustness of the signature image. The CBIR system is composed by three key steps: (1) Features extraction (2) Signature construction (3) Retrieved images. Two major state-of-the-art works proposed the best performing signature images: Bag of Visual Words (BoVW) and deep learning based convolutional neural network (CNN). In a CNN, the signature image consists of N floating-point vectors extracted from feature layers within the architectural network. In BoVW, signature images are computed based on the frequency of visual words in the image. After signature images are created for a query and dataset, the selection of candidates considered similar to the input query is determined by the spacing between signatures using a specific metric.

We focus on this work to enhance BoVW because it is not necessary to train the algorithm on a set of images. In addition, BoVW is a robust algorithm with low complexity and signature creation is faster than CNN approaches. So, our aim in this paper is to increase the BoVW accuracy by transforming it to bag of visual phrase (BoVP).A visual phrase is a set of related visual words that aim to more robustly encode the visual features in image. The transformation can be applied with different ways (Pedrosa and Traina, 2013) (Ren et al., 2013) (Chen et al., 2014). We proposed in this paper a robust transformation based density clustering. In this case the image signature is defined by a matrix with size $M*M$ where M is the number of visual words. After building the images' signatures for both queries and datasets, the candidates will be selected based on the euclidean distance. We consider the images with low scores to be similar to an input query. We test this approach on three different datasets and two different visual descriptors (SURF, KAZE). We show experimentally that the proposed approach achieve a better results in terms of accuracy compared to the state of the art methods.

The paper is structured as follows: We give a brief overview of the work related to metaphor words in Section 2. We detail our proposition in Section 3. We show the experimental part on 4 datasets in Section 4. Conclusion in Section 5.

We first discuss the state-of-the-art in both of our contribution fields: Bag of Visual Words model and Density-Based Spatial Clustering approach.

1.1 Bag of Visual Words Model

Bag of Visual Words proposed by (Csurka et al., 2004) is one of the most common model used to classify the images by content. This approach is com-

[a] https://orcid.org/0000-0002-1197-253X

Ouni, A.
An Unsupervised IR Approach Based Density Clustering Algorithm.
DOI: 10.5220/0011892800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 491-496
ISBN: 978-989-758-634-7; ISSN: 2184-4321

posed of three main steps: (i) Detection and Feature extraction (ii) Codebook construction (iii) Vector quantization. Detecting and extracting features in an image can be performed using extractor algorithms. Many descriptors have been proposed to encodes the images into a vector. Scale Invariant Feature Transform (SIFT) (Lindeberg, 2012) and Speeded-up Robust Features (SURF) (Bay et al., 2006) are the most used descriptors in CBIR. Interesting work from Arandjelović and Zisserman (Arandjelovic and Zisserman, 2013) introduces an improvement by upgrading SIFT to RootSift. On the other side, binary descriptors have proven useful (Rublee et al., 2011) proposes ORB (Oriented FAST and Rotated BRIEF) to speed up the search. An other work (Leutenegger et al., 2011) combines two aspects: precision and speed thanks to BRISK (Binary Robust Invariant Scalable Keypoints) descriptor. (Iakovidou et al., 2019) present a discriminative descriptor for image dependent on both contour and color information. In (Chatzichristofis and Boutalis, 2008), the authors present descriptors both color and edge information. Due to the limit of bag of visual words model many improvement have been proposed for more precision. Bag of visual phrases (BoVP) (Pedrosa and Traina, 2013) is a high-level description using a more than word for representing an image. formed phrases using a sequence of n-consecutive words regrouped by L2 metric. (Ren et al., 2013) Build an initial graph then split the graph into a fixed number of sub-graphs using the N-Cut algorithm. Each histogram of visual words in a sub graph forms a visual phrases. (Ouni et al., 2021) present a method based on grid clustering algorithm for linking the visual words. (Chen et al., 2014) Groups visual words in pairs using the neighbourhood of each point of interest. The pair words are chosen as visual phrases. Perronnin and Dance (Perronnin and Dance, 2007) applies Fisher Kernels to visual words represented by means of a Gaussian Mixture Model(GMM). Similar approach, introduced a simplification for Fisher kernel. Similar to bag of visual words, vector of locally aggregated descriptors (VLAD) (Jégou et al., 2010) assign each feature or keypoint to its nearest visual word and accumulate this difference for each visual word. (Mehmood et al., 2016) proposes a framework between local and global histograms of visual words. in other side, CNN based approach are widely used in CBIR context (Krizhevsky et al., 2012) (Szegedy et al., 2017) (Simonyan and Zisserman, 2014). In many instances, CNN supersedes local detectors and descriptors. The main idea is to train the network on set of images. The next step use the pretrained CNN for extracting the signatures image.

1.2 Density-Based Spatial Clustering

The notion of the " density of a cell", defined relative to the number of objects contained in the cell. Density clustering algorithms are based on a similar notion, complemented by other fundamental concepts such as the neighborhood of an object, kernel object, accessibility, or connection between objects. The complexity of this algorithm is O(nlogn) which makes it a rather inexpensive method. In addition, the clusters obtained can be of various forms. However the algorithm has a major disadvantage: the choice of parameters ε and M. Even if the authors of the algorithm propose a heuristic to automatically determine these parameters, this choice remains difficult in practice. The data are not generally distributed identically and these parameters should be able to vary according to the regions of the space.

DBSCAN (Schubert et al., 2017) is an algorithm of $O(nlogn)$ complexity which makes it a fairly inexpensive method. It is a density-based algorithm in the measure that relies on the estimated density of the clusters to perform partitioning. DBCLASD (Xu et al., 1998)(Distribution-Based Clustering of Large Spatial Databases) algorithm proposes a distributional approach to manage this problem of variation in local densities. Same authors propose OPTICS (Ankerst et al., 2008) algorithm (Ordering Points To Identify the Clustering Structure). OPTICS defines an order on the objects which can then be used by DBSCAN, in the cluster expansion phase.

2 BAG OF VISUAL WORDS BASED DENSITY CLUSTERING: DBoVW

We aim in this section to improve the image representation. As described in previous section, bag of visual words is the model that can describe the visual features in image in a robust way. However, even the new representation, the given outcomes do not fulfill the ideal need. Therefore we attempt to improve the BoVW model by changing to a bag of visual phrases based on density clustering approaches.

The goal of using a clustering strategy is to represent an image as a set of clusters, each cluster containing at least two keypoints and representing one or more visual sets. For this case, each cluster takes on a visual representation when its size is >2. Then, from the acquired groups, we consolidated the visual words inside each cluster to fabricate a discriminative signature denoted "bag of visual phrase". Also, we test and study the distinction between the density cluster-

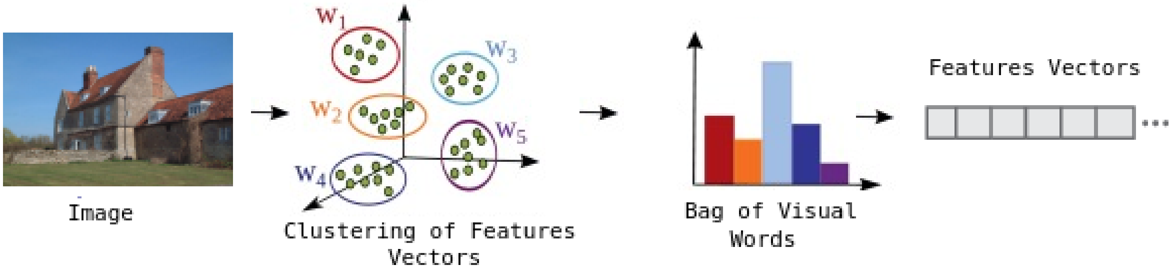

Figure 1: Bag of visual words model.

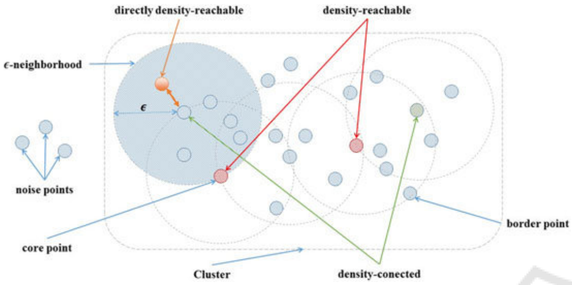

Figure 2: Density-based spatial clustering (Schubert et al., 2017) (DBSCAN).

ing techniques (DBSCAN, Optics) at the degree of execution, productivity and which one has better improvement the image representation.

In the first part of our framework, we start by recognizing and extracting of the keypoints from the input query. After using the offline-created visual words or vocabulary, we use the L2 metric to match each keypoint detected in the image with the corresponding visual vocabulary.

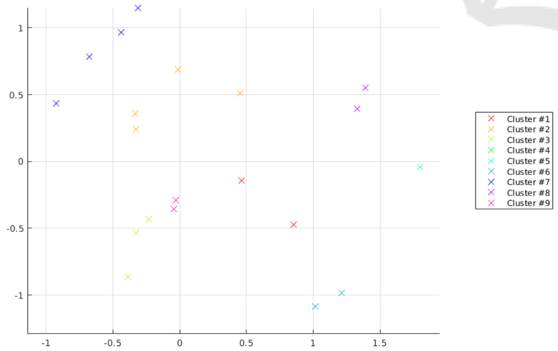

Figure 3: Clustering Key-points using Density clustering(DBSCAN).

Figure 3 presents the clusters obtained by density clustering approach. A cluster is the maximum set of connected points(Key-points in our case). The algorithm has a major disadvantage: the choice of parameters ε and Min ε. In our case we propose a heuristic to automatically determine these parameters. We fixed Minε= 1 and we begin from ε=0.001 and in each iteration we increase the value of ε until the number of cluster will be equal or close to the half size of keypoints. As shown in figure 3 we obtained 9 clusters using 20 key-point. We succeeded in finding the ideal number of clusters in relation to the number of keys-points. The number of key-points in each cluster is different to others. We consider each cluster as a set of visual phrases.

The image signature size is $M * M$ where M is the number of visual words. We initialize the matrix to zero. Next, we fill the matrix with the indices of the visual phrases. For example, if the pair visual words are VW_5 and VW_{30} then we increment the element at index (5,30). Finally, the similarity is computed by the euclidean distance between the matrix query and the matrix from dataset.

3 RESULTS

In this section, we present the potential of our approach on large datasets. Our goal is to increase the CBIR accuracy and reduce the execution time. To evaluate our proposition we test on the following datasets:

• Corel 1K or Wang (Wang et al., 2001) is a dataset of 1000 images divided into 10 categories and each category contains 100 images. The evaluation is computed by the average precision of the first 100 nearest neighbors among 1000.

• The University of Kentucky Benchmark proposed by Nister, abbreviated as UKB for ease of reading. UKB contains 10200 images divided into 2550 groups, each group contains 4 images of the same object with different conditions (rotation, out-of-focus...). The score is the average accuracy across all images of the 4 nearest neighbors.

• INRIA Holidays, referred to as Holidays, is a collection of 1491 candidates, of which 500 are query pictures and the remaining 991 are corresponding related candidates. Public holidays are evaluated based on mean precision (mAP) [29].

• MSRC v1 (Microsoft Research in Cambridge) sug-

Figure 4: The 10 categories of Corel-1K dataset.

Figure 5: Example of images from UKB dataset.

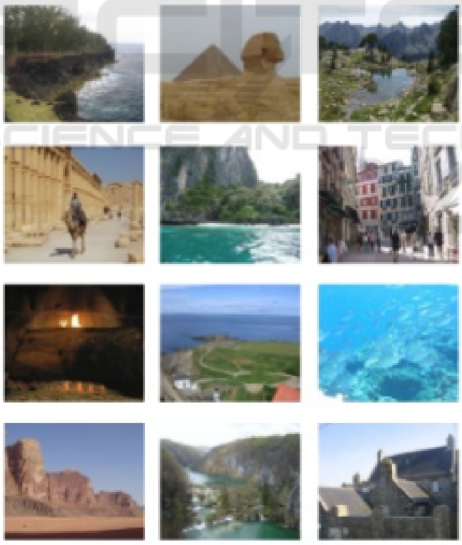

Figure 6: Example of images from holidays dataset.

gested by Microsoft Research Team. MSRC v1 contains 241 images divided into 9 categories. Scoring for MSRC v1 is based on MAP score (mean precision)

In this section we present the experiments of our approach based density clustering approach. To test the proficiency of our proposed strategies, we led the experimentation on four datasets(MSRC V1, UKB, Holidays, Wang) and two descriptors(SURF, KAZE). In tables 1, 2, we separated the outcomes into two principle parts. Above, we build the signature utilizing DBSCAN approach. Down, the signature constructed using Optics algorithm. The outcomes got utilizing DBSCAN approach are superior to the outcomes got utilizing Optics. We perform the results by concatenating the GBOP$_{surf}$ and GBOP$_{kaze}$.

Figure 7: Example of images from MSRC v1 dataset.

Table 1: Bag of visual phrase based DBSCAN algorithm.

Datasets	GBoVP$_{surf}$	GBoVP$_{kaze}$	GBoVP$_{surf \cdot kaze}$
MSRC v1	0.53	0.57	0.68
UKB	3.09	3.11	3.55
Holidays	0.55	0.56	0.68
Wang	0.55	0.52	0.63

Table 2: Bag of visual phrase based OPTICS algorithm.

Datasets	GBoVP$_{surf}$	GBoVP$_{kaze}$	GBoVP$_{surf \cdot kaze}$
MSRC v1	0.57	0.6	0.69
UKB	3.1	3.18	3.61
Holidays	0.56	0.56	0.65
Wang	0.59	0.55	0.65

Finally, we compare our approach against two different state-of-the-art methods in Table3. The first is methods based on keypoints for building the image signatures and the second is methods based on deep leaning approaches. As show the aftereffects of our proposed present great presentation for all datasets compared to methods based on keypoints.

Table 3: Comparison of the accuracy of our approach with methods from the state of the art.

Methods	MSRC v1	Wang	Holidays	UKB
BoVW (Csurka et al., 2004)	0.48	0.41	0.50	2.95
SaCoCo(Iakovidou et al., 2019)	-	0.51	0.76	3.33
CEDD (Chatzichristofis and Boutalis, 2008)	-	0.54	0.72	3.24
VLAD (Jégou et al., 2010)	-	-	0.53	3.17
N-Gram (Pedrosa and Traina, 2013)	-	0.34	-	-
Grid (Ouni et al., 2021)	0.67	0.62	0.64	3.57
Fisher (Perronnin and Dance, 2007)	-	-	0.69	3.07
Ours	0.69	0.65	0.68	3.61

4 CONCLUSION

In this paper, we present an effective bag of visual phrase model dependent on grouping approach. We show that the utilization of density clustering approach joined with BoVW model increment the CBIR precision. Utilizing two descriptors (KAZE, SURF), our methodology accomplish a superior outcomes compared to the state of the art methods.

REFERENCES

Ankerst, M., Breunig, M., Kriegel, H., Ng, R., and Sander, J. (2008). Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD*, volume 99.

Arandjelovic, R. and Zisserman, A. (2013). All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Chatzichristofis, S. A. and Boutalis, Y. S. (2008). Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *International conference on computer vision systems*, pages 312–322. Springer.

Chen, T., Yap, K.-H., and Zhang, D. (2014). Discriminative soft bag-of-visual phrase for mobile landmark recognition. *IEEE Transactions on Multimedia*, 16(3):612–622.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

Iakovidou, C., Anagnostopoulos, N., Lux, M., Christodoulou, K., Boutalis, Y., and Chatzichristofis, S. A. (2019). Composite description based on salient contours and color information for cbir tasks. *IEEE Transactions on Image Processing*, 28(6):3115–3129.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE.

Lindeberg, T. (2012). Scale invariant feature transform.

Mehmood, Z., Anwar, S. M., Ali, N., Habib, H. A., and Rashid, M. (2016). A novel image retrieval based on a combination of local and global histograms of visual words. *Mathematical Problems in Engineering*, 2016.

Ouni, A., Royer, E., Chevaldonné, M., and Dhome, M. (2021). Robust visual vocabulary based on grid clustering. In *Intelligent Decision Technologies*, pages 221–230. Springer.

Pedrosa, G. V. and Traina, A. J. (2013). From bag-of-visual-words to bag-of-visual-phrases using n-grams. In *2013 XXVI Conference on Graphics, Patterns and Images*, pages 304–311. IEEE.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Ren, Y., Bugeau, A., and Benois-Pineau, J. (2013). Visual object retrieval by graph features.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Wang, J. Z., Li, J., and Wiederhold, G. (2001). Simplicity:

Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on pattern analysis and machine intelligence*, 23(9):947–963.

Xu, X., Ester, M., Kriegel, H.-P., and Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings 14th International Conference on Data Engineering*, pages 324–331. IEEE.