A COMPARISON OF BAG-OF-WORDS METHOD AND NORMALIZED COMPRESSION DISTANCE FOR SATELLITE IMAGE RETRIEVAL

Shiyong Cui, Mihai Datcu

Remote Sensing Technology Institute (IMF)
German Aerospace Center (DLR)
Münchener Straße 20, 82234 Wessling
shiyong.cui, mihai.datcu@dlr.de

ABSTRACT

Recently, two improved methods have shown their advantages in browsing Earth Observation (EO) dataset. The first method is the Bag-of-Words (BoW) feature extraction method and the second is the Normalized Compression Distance (NCD) for assessing image similarity. However, they have not been compared so far for satellite image retrieval, which motivates this paper. Two retrieval experiments have been performed on a freely available optical image dataset and a SAR image dataset. Through these two experiments, we conclude that the BoW method performs generally better than NCD. Although it is a parameter-free solution for data mining, NCD only performs well for images with repetitive patterns like some homogeneous classes. In contrast, BoW method performs much far beyond that of NCD. In addition, NCD is computationally very expensive, which makes it infeasible to be applied in real applications. In contrast, BoW method is more realistic in practical applications in terms of both accuracy and computation.

Index Terms— Normalized compression distance (NCD), Bag-of-Words (BoW), Satellite image retrieval.

1. INTRODUCTION

Nowadays in Earth Observation (EO), the data volume increases rapidly beyond the users' capability to access the information content of the data. This makes fast browsing and automatic interpretation of a large data volume challenging. Thus, content based image retrieval has been developed since years to solve this problem, such as the Knowledge-driven Information Mining (KIM) system [1] and the Geospatial Information Retrieval and Indexing (GeoIRIS) system [2].

In theory, content based image retrieval is a trivial nearest neighbor search problem. Given a database and a query, we first loop over all images in the database and compute a similarity of the query to each image. Then we find all similar images by sorting the similarity values of all images to the query. To this end, we have to solve two fundamental problems. The first is to find a discriminative representation of image content. The second is to compute a similarity between the query and each image in the database using the selected image content representation.

Since years, the first problem, which is referred to as feature extraction in the literature, has been continuously developed. A large variety of feature representations of image content have been developed. Recently, the Bag-of-Words (BoW) [3, 4] representation has been demonstrated more discriminative than many other methods. Another line of relevant research in information theory is to develop parameter-free method [5, 6] to address content based image retrieval. One prominent method of this kind is the Normalized Compression Distance (NCD). The advantage of this kind of methods is no need to find a discriminative image representation. Both of these two kinds of methods have shown promising performance in image retrieval. However, to our knowledge, they have not been compared with each other for content based image retrieval, which motivates this paper. In this paper we perform a systematic comparison of BoW method and NCD for content based satellite image retrieval. Through this study, we try to explain from a point view of information theory why BoW method is so powerful in image representation.

2. BAG-OF-WORDS METHOD

The framework of BoW feature extraction shown Fig.1 is composed of five steps, which are feature detection, local feature extraction, dictionary learning, feature coding, and feature pooling. Assume we have a dataset of N images $I_i, i = 1, ..., N$, the first step is to sample a collect of patches from the images in the database. This can be done by dense sampling or sparse detection. The second step is to extract local descriptors $\mathbf{x}_i^j \in \mathbb{R}^D, j = 1, ..., M$ from all images. The third one is to learning a dictionary $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_K) \in \mathbb{R}^{D \times K}$ with K words using all local features. Normally, this is done by time consuming unsupervised learning method, such as k-means and gaussian mixture model. Thus the elements $\mathbf{d}_i$ in a dictionary are the centers of the clusters. The next step is to find a more discriminative

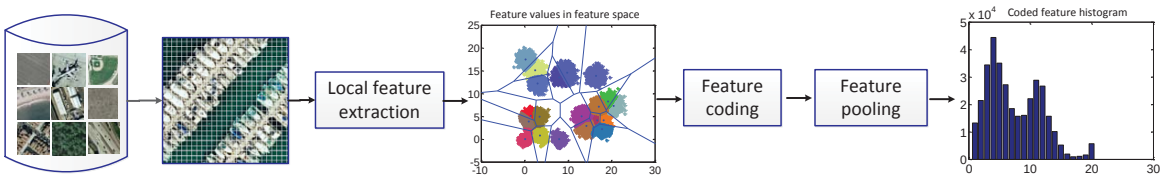


Fig. 1. The framework of the Bag-of-Words model.

representation $\mathbf{v} = [v_1, ..., v_K]$ for each local descriptor $\mathbf{x}$. This can be done using hard feature assignment or soft assignment. Hard assignment assigns a label, the index of the nearest neighbors in the dictionary, to each local descriptor $\mathbf{x}$. Formally, it is defined as:

$$v_i(\mathbf{x}) = \begin{cases} 1 & \text{if } k = \min \|\mathbf{x} - \mathbf{d}_i\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Thus, the final descriptor representation $\mathbf{v} = [v_1, ..., v_K]$ has only one non-zero element. The last step is to take the sum-pooling [1] of all local descriptors extracted from one image $\mathbf{v}_i = \text{sum}(\mathbf{v}_i^j, ..., \mathbf{v}_i^j)$. Based on the BoW feature representation, image retrieval can be achieved by selecting a distance measure, such as Euclidean distance. In this paper, χ^2 distance is used for image retrieval.

3. NORMALIZED COMPRESSION DISTANCE

Normailzed Compression Distance (NCD) [7] based on a universal lossless data compressor $Z(x)$, defined by (2), is a general distance measuring the similarity of two objects, which can be images, documents, letters, etc.

$$NCD_Z(x, y) = \frac{Z(x, y) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}} \quad (2)$$

$Z(x)$, $Z(y)$ and $Z(x, y)$ denote the binary length of the single image x, y and the concatenation of image x and y. The fundamental idea of NCD is two objects will be similar if they can be jointly compressed significantly. The value of NCD is a nonnegative number $0 < r < 1 + e$ denoting how similar two images are. A smaller NCD value represents that the two images are very similar because they can be jointly compressed significantly. Since NCD is a similarity metric, it can be applied to address many problems in information retrieval and data mining. In earth observation, NCD has been successfully employed in [6] to address clustering, classification, artifact detection, and image time series mining. In this paper, our goal is not to evaluate NCD but to compare it with the BoW method for image retrieval.

[1] Sum-pooling is equivalent to computing the histogram.

4. EXPERIMENTS AND DISCUSSION

In this section, we present the datasets we used for comparison and the results of our experiments.

4.1. Datasets

4.1.1. SAR Image Dataset

The first dataset is composed of 15 classes of altogether 3434 TerraSAR-X sub-scenes with a size of 160×160 pixels and a pixel spacing of about 3 m (cf. Fig. 2). The sub-scenes are cut from radiometrically enhanced high resolution Stripmap TerraSAR-X images with good signal-to-noise ratios. This dataset was compiled interactively using an active learning system [8]. We could discriminate 15 classes; among them there are 7 classes of urban areas, which is sufficient for comparison.

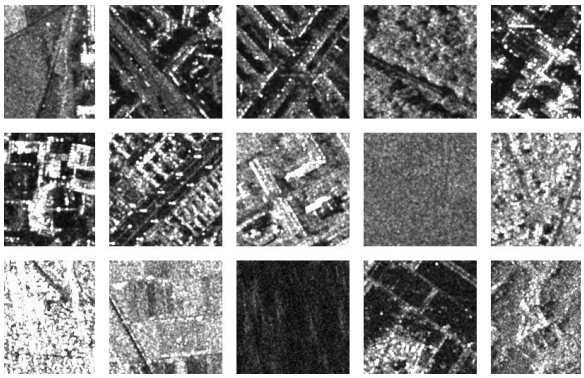


Fig. 2. Example images of the SAR image dataset.

4.1.2. UC Merced Land Use Dataset

The second dataset is the UCMerced land use dataset [9][2]. The images were manually extracted from large images existing in the USGS national map urban area imagery collection covering various urban areas around the country. The pixel resolution of this public domain imagery is 1.0 foot. The dataset comprises 21 classes, namely *agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection,*

[2] The data is available at `http://vision.ucmerced.edu/datasets/landuse.html`

medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, tennis court, and each class has 100 images with a size of 256×256 pixels. Example images from each class are shown in Fig. 3.

In extracting BoW features, we used the raw pixel values and k-means clustering is used for codebook learning. The coodbook size we used is 200. For NCD, we used a publicly available open-source downloadable software tool CompLearn, which can be found from `http://www.complearn.org`. The compressor we applied is the zlib algorithm for image compression.

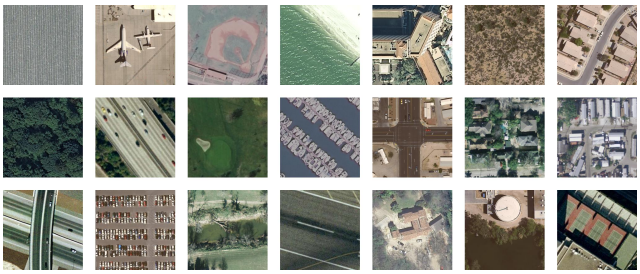


Fig. 3. example images of the UC Merced land use dataset.

4.2. Results and discussion

For evaluation, we use each image as a query and search for the similar images in the remaining images. The accuracy measures we used for evaluation are *precision* and *recall*, *Precision* is the fraction of retrieved images that are relevant to the search and *recall* is the fraction of the images that are relevant to the query and that are successfully retrieved. For each query, we compute the precision and recall curve. Since the precision and recall curve has a distinctive saw-tooth shape, interpolated average precision and recall cures are used. The area under this curve is computed and used for comparison.

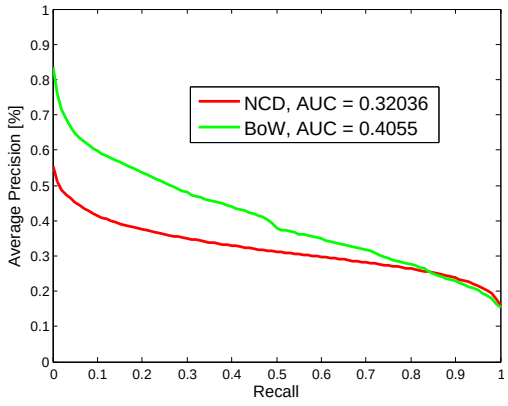
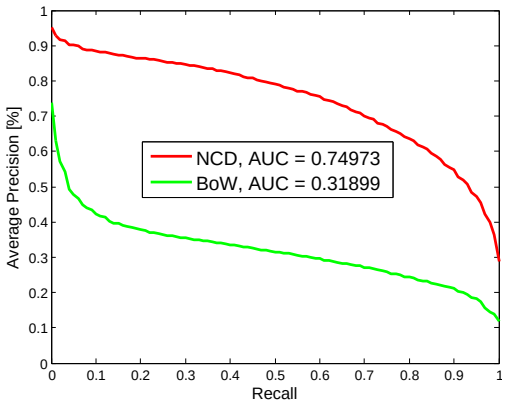
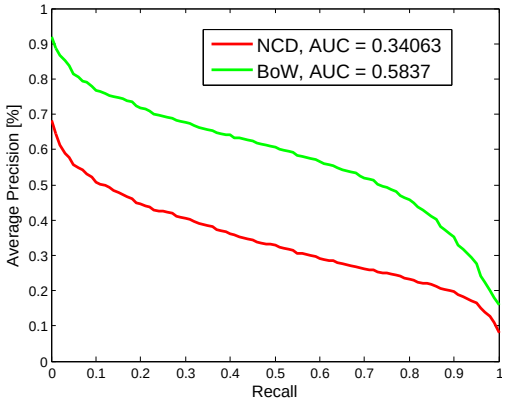


Fig. 4. Average precision of NCD and BoW on the SAR dataset.



(a)



(b)

Fig. 5. Average precision of NCD and BoW for (a) a homogeneous class *grass* and (b) a heterogenous class *flooded field*.

The average precisions of NCD and BoW on the SAR image dataset is shown in Fig.4. We can see clearly that BoW performs much better than NCD. To compare their performance on individual class, the average class-wise precisions are shown in Table 1. From table, we see that NCD only performs well for homogeneous classes, such as mountain and grass, etc. For homogeneous class *grass*, NCD performs much better than BoW. However, for heterogeneous class *flooded fields*, BoW is much better than NCD.

The results on the second dataset are shown in Table 2. The average precisions of NCD and BoW over all classes are shown in Fig. 6. From these results, we can observe that BoW is totally superior than NCD for all classes, which is almost three times better. The same as the observation in the first experiment, NCD has a similar performance as boW only for homogeneous class. Most classes do not have repetitive patterns that is the main reason that BoW performs better than NCD in this experiment: Another disadvantage of NCD is that its computation is very prohibitively slow, which make it infeasible to be applied in real applications.

Table 1. Average AUC [%] of NCD and BoW on the SAR dataset.

Methods	flooded	mountain	Skyscrap.	small	forest	field	agricult.	house
NCD [%]	0.34	0.57	0.08	0.09	0.22	0.43	0.35	0.17
BoW [%]	0.58	0.39	0.27	0.15	0.17	0.30	0.44	0.33
Methods	building	highway	industrial	sea	resit.2	resid.	grass	average
NCD [%]	0.17	0.20	0.10	0.97	0.19	0.17	0.75	0.3204
BoW [%]	0.34	0.44	0.62	0.96	0.33	0.45	0.32	0.4055

Table 2. Average AUC [%] of NCD and BoW on the UCMerced dataset.

Methods	golf.	build.	beach	freeway	base.	runway	m. resid.	park.	river	tennis.	plane
NCD [%]	0.11	0.09	0.15	0.08	0.08	0.06	0.07	0.07	0.07	0.07	0.08
BoW [%]	0.13	0.14	0.36	0.17	0.68	0.12	0.15	0.19	0.18	0.18	0.12
Methods	chaparral	tank	agricult.	overpass	harbor	s. resid.	forest	d. resid.	inter.	park	aver.
NCD [%]	0.06	0.06	0.07	0.07	0.10	0.07	0.10	0.08	0.08	0.09	0.08
BoW [%]	0.34	0.20	0.18	0.15	0.21	0.39	0.17	0.14	0.19	0.15	0.22

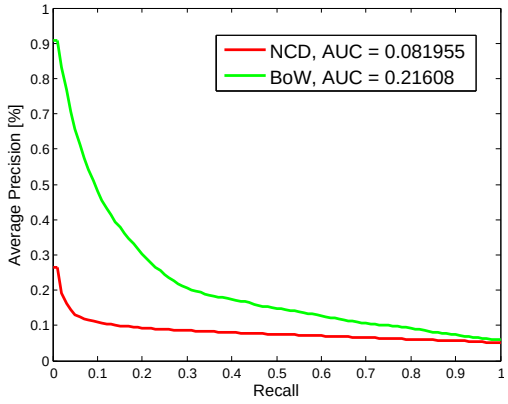


Fig. 6. Average precision of NCD and BoW on the UCMerced dataset.

5. CONCLUSION

In this paper, we compare the Normalized Compression Distance (NCD) and the Bag-of-Words (BoW) method for satellite image retrieval. Two experiments using optical dataset and SAR image dataset are performed. They are compared in terms of precision and recall, as well as the area under the precision-recall curve. Through this study, we found that in many cases BoW performs better than NCD for both optical and SAR image retrieval. Although it is a parameter-free solution for data mining, NCD only performs well for homogeneous class with repetitive patterns. In addition, NCD is computationally very expensive, which makes it infeasible to be applied in real applications. In contrast, BoW method is more realistic in practical applications in terms of both accuracy and computation.

6. REFERENCES

[1] M. Datcu and K. Seidel, "Human-centered concepts for exploration and understanding of earth observation images," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 601–609, March 2005.

[2] Chi-Ren Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial information retrieval and indexing system — content mining, semantics modeling, and complex queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 839–852, 2007.

[3] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *Proc. IEEE International Conference on Computer Vision, ICCV*, Washington, DC, USA, Oct 2003, vol. 2, pp. 1470–1477.

[4] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Jun. 2005, vol. 2, pp. 524–531.

[5] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards Parameter-free Data Mining," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2004, number 10 in KDD '04, pp. 206–215.

[6] D. Cerra, A. Mallet, L. Gueguen, and M. Datcu, "Algorithmic information theory-based analysis of earth observation images: An assessment," *Geoscience and Remote Sensing Letters, IEEE*, vol. 7, no. 1, pp. 8–12, Jan. 2010.

[7] M. Li, X. Chen, X.Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3250–3264, December 2004.

[8] S. Cui, C. O. Dumitru, and M. Datcu, "Semantic annotation in Earth observation based on active learning," *International Journal of Image and Data Fusion*, vol. 5, no. 2, pp. 152–174, 2013.

[9] Y. Yang and S. Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," in *Proc. 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10*, New York, NY, 2010, pp. 270–279.