Exploiting indexing structures for large scale Remote Sensing Image Classification

Eduardo de A. Tavares
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
Email: eteduardotavares@gmail.com

Jefersson Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
Email: jefersson@dcc.ufmg.br

***Abstract*—The rapid increase on the volume of data generated by remote sensing systems boosted by the evolution of satellites and the popularization of their imagery has enabled a wide range of new Earth Observation applications. At the same time, it created the challenge of how to efficiently deal with these collections of data. In this work we evaluate the use of indexing techniques for speeding up remote sensing image retrieval aiming automatic large scale geographical mapping in the future. Three CNNs are employed as feature extractors and compared to three low-level features on retrieval tasks performed on a dataset of aerial images with the LSH algorithm. Preliminary results showed a recall level of almost 50% when only roughly 5% of the samples of the evaluated dataset needed to be considered.**

I. Introduction

With the increase of the capacity of digital storage systems, the creation of large databases containing huge amounts of data has been exploited in many domains, including that of Remote Sensing.

Remote sensing is defined in [1] as the science of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation. In remote sensing, energy reflected or emanating from the surface of the planet is measured using a sensor mounted on an aircraft or spacecraft platform. The measurement of this energy is then used to construct an image of the landscape beneath the platform [2].

As storage systems had their capacities increased, the amount of data generated by remote sensing systems also grew. While in the past the use of satellites was restricted to military applications, civilians now have widespread access to satellite imagery. Such images are in some cases publicly available and/or available for research purposes (e.g. CBERS-4, Sentinell), or more commonly available for purchase from a multitude of specialized companies. Furthermore, the popularization of devices like drones equipped with cameras enable an even wider range of possible applications.

Large-scale remote sensing (RS) image search and retrieval are an active field of research due to the rapid evolution of satellite systems, which resulted in an ever increasing number of image archives, with higher spectral and spatial resolution [3]. The data provided by these satellites usually has hundreds of spectral bands, which poses difficulties for traditional image processing and data handling techniques [2].

While large-scale databases create many opportunities for novel applications, the large volume poses unique challenges for the retrieval of similar objects. How to measure similarity is also a challenge, as the same dataset may enable a wide range of different applications. Land cover classification, agro-ecological classification and land-use classification are categories of tasks related to the classification of land area that are enabled by the use of satellite imagery, with examples including plant species identification [4], terrain classification [5], natural disasters analysis [6] and ecosystems monitoring [7].

Each of these diverse applications will require the retrieval of different kinds of data, which will involve the processing of huge datasets. Finding accurate nearest neighbours efficiently is still a challenge, especially for large databases and computationally expensive underlying distance measures [8]. An exhaustive search through linear scan from such archives is time demanding and not scalable in operational applications. This limitation enticed the development of methods to speed up systems that rely on such kind of databases. The use of indexing techniques to organize remote sensing data in a structured way is expected to enable the development of even more applications by speeding up its access.

In the scenario of classification tasks applied to large scale automated cartography based on satellite imagery, the motivations for the study of indexing techniques may be grouped in two phases:

- *Training*: in this phase, indexing may help with the balancing of the dataset by aiding with the choice of instances located in different buckets after hashing, in essence, diverse instances. Better accuracy in classification is also expected with the reduction of redundant samples and the maintenance of representative ones.
- *Testing*: in the testing phase, the indexing module may act as a preliminar filter capable of easily discarding instances that strongly differ from a query. Discarding such instances preliminarily will save computation time by avoiding the use of expensive classifiers that shall be only fed with instances whose class is very similar to that of the query.

The rest of this work is organized as follows: **Section II** provides general background on hashing-based nearest neighbour search methods; **Section III** presents the methodology employed in this work while **Section IV** details the features and the dataset explored; finally **Section V** exposes the preliminary results obtained and **Section VI** gives conclusions, recommendations for future research, and final remarks.

II. Related Work

Due to its high efficiency in terms of storage and computational cost, hashing has become a popular method for nearest neighbor search in large-scale image retrieval [3]. This chapter presents background knowledge on hashing based methods for image retrieval.

A. *Search with hashing*

Hashing is a clever way to address the challenges for large-scale similarity search [9]. In hashing, each database item is represented by a compact binary code that is constructed such that similar items have similar binary codes. Binary codes are storage efficient and computing Hamming distance can be performed extremely fast with few machine instructions, so much so that millions of items can be compared to a query in less than a second [10].

A hash function is a function H which has, as a minimum, the following two properties [11]:

- compression – H maps an input x of arbitrary finite bit length, to an output H(x) of fixed bit length n.
- ease of computation – given H and an input x, H(x) is easy to compute.

Briefly speaking, hashing has as its goal mapping an original D-dimensional data space R^D to a binary Hamming space B^K, where each data point is represented by a binary hash code (i.e., a K-bit hash key) and the entire data set is mapped to a table with hash keys as entries, namely a hash table [12]. In this way, approximate nearest neighbor search can be efficiently and accurately performed using query items and possibly a small subset of the data space.

Given a sample point $x \in R^D$, one can employ a set of hash functions $H = \{h_1, ..., h_k\}$ to compute a *K*-bit binary code $\mathbf{y} = \{y_1, ..., y_K\}$ for $\mathbf{x}$ as $\mathbf{y} = \{h_1(x), ..., h_2(x), ..., h_K(x)\}$ where the k^{th} bit is computed as $y_k = h_k(\mathbf{x})$. The hash function performs the mapping as $h_k : R^D \mapsto B$.

There are two basic strategies for using hash codes to perform nearest (near) neighbor search: hash table lookup and hash code ranking [13].

1) Hash Table Lookup: Hash table lookup accelerates the search by reducing the number of the distance computations. The hash table (i.e., a form of inverted index) is composed of buckets with each bucket indexed by a hash code, while each item **x** from the database is placed into a bucket h(**x**) [13].

While conventional hashing in computer science avoids collisions (i.e., avoids mapping two items into a same bucket) the hash table approach aims to maximize the probability of collision of near items while minimizing that of items that are far away colliding.

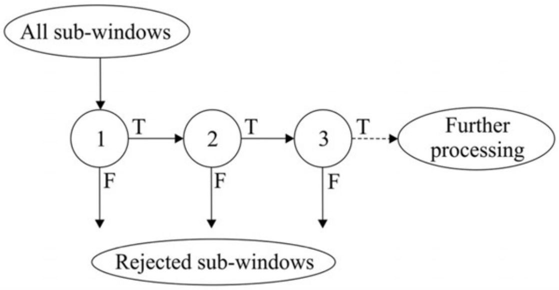

Fig. 1. The structure of the Viola-Jones cascade classifier. Extracted from [14].

Given a query **q**, items lying in the bucket h(**q**) are retrieved as candidates of the nearest items of **q**, usually followed by a reranking step: the retrieved candidates are reranked according to the true distances computed using the original features and attain the nearest neighbors.

2) Hash Code Reranking: Hash code ranking performs an exhaustive search comparing the query with each database item by fast evaluating their distance, retrieving the database items with the smallest distances as the candidates of nearest neighbors. Usually this is followed by a reranking step: rerank the retrieved nearest neighbor candidates according to the true distances computed using the original features and attain the nearest neighbors [13]. This strategy exploits one main advantage of hash codes: the distance using hash codes is efficiently computed and the cost is much smaller than that of the computation in the original input space.

III. Methodology

An idea successfully applied to face [14] and hand detection [15] is that of a cascade of classifiers, whose key insight is that simpler, and therefore more efficient, boosted classifiers can be constructed, rejecting many of the negative patches fed while detecting almost all positive instances. Only after the rejection of the majority of patches, more complex classifiers are used in order to achieve low false positive rates, which saves considerable computation time [14]. The architecture of such system is depicted in Fig. 1.

Based on the idea of cascade classifiers, this work proposes a similar structure with the addition of a module responsible for the indexing of the images that constitute large databases used in Remote Sensing applications. By indexing the database, it is expected that a great speed up in usage will be achieved, as the index will allow the retrieval of specific objects in constant time. Once these objects are retrieved, only the portions of the images containing them need to be accessed, avoiding the necessity of allocating huge images into main memory. The proposed structure is depicted in Fig. 2.

IV. Experimental Validation

A. *Features*

The preliminary experiments were performed using the LSH technique [16]. Based on the fact that several recent works revealed that deep convolutional neural networks (CNNs) are

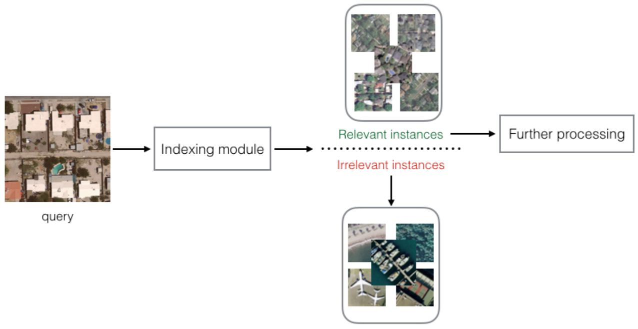

Fig. 2. Structure of the proposed method. Irrelevant instances are discarded before they are fed to complex classifiers, which will save computation time.

capable of learning rich mid-level representations effective for tasks like image classification, object detection, and semantic segmentation [17]–[20], the activations of final layers from pre-trained CNNs were extracted to serve as features representing the images from the evaluated dataset.

It is believed that these deep CNN architectures trained on huge datasets of numerous categories can be transferred to new domains by employing them as feature extractors [21] on other tasks including recognition and retrieval, providing better performance than handcrafted features [21] such as GIST [22] and HOG [23]. Three low-level features were employed as to serve as a baseline for comparison with the deep features used, namely: Color Auto Correlogram (CAC), Local Binary Patterns (LBP) and Histogram of oriented gradients (HOG).

The three CNNs from which the features were extracted are detailed below.

1) AlexNet: AlexNet has 60 million parameters and 650,000 neurons, with five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final softmax. It was proposed by Krizhevsky et al. [17], and was the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [24] in 2012.

AlexNet was used as a feature extractor by extracting the activation values from the last and second to last fully-connected layers, which results in feature vectors of 4096 and 1000 dimensions, respectively.

2) VGG networks: VGG_{16} and VGG_{19} are the two more successful networks out of a set proposed in [18] which won the localization and classification tracks of the ILSVRC-2014 competition. VGG_{16} has 13 convolutional layers, 5 pooling ones and 3 fully-connected ones (considering the softmax). VVG_{19} has a similar architecture, differing only by the fact that it has 19 weight layers while VGG_{16} has 16.

These networks were used as feature extractors by extracting the activations from the last and second to last fully-connected layer, resulting in feature vectors of 4096 and 1000 dimensions, respectively.

B. Datasets

Preliminary experiments were performed using the UCMerced Land-use dataset [25]. It is composed of 2,100 aerial images of size 256×256 obtained from the United States Geological Survey (USGS) National Map, portraying different US locations for the sake of providing diversity to the dataset. They are divided into 21 land-use categories: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Classes like “dense residential”, “medium residential” and “sparse residential” have very similar instances, which mainly differ in the density of structures.

Further experiments shall be performed on images obtained with the SENTINEL 2 satellite [26] encompassing the State of Minas Gerais. The selection criteria for choosing the images included the preference for images with a presence of clouds below 5%; images not cropped and without noise; and images obtained in the second semester of 2016. Four main goals are associated with this dataset: i)the mapping of productive and unproductive land properties; ii)the mapping of primary and local roads; iii)the mapping of water bodies (e.g., lakes, dams) and water wells; iv)prediction of the Human Development Index by city (*IDHM - Índice de Desenvolvimento Humano Municipal*) both on urban and rural areas [27].

V. Preliminary results

Table I presents the average precision obtained with the 10, 20, 30, 50 and 100 nearest neighbours returned. The average was obtained with all 2100 samples of the UCMerced Land-use dataset being used as query, with the activation values of different layers from pre-trained CNNs used as features, the concatenation of the fc8 layer of all CNNs, as well as the low-level features CAC, LBP and HOG, and their concatenation, while Fig. 3 depicts the $precision \times recall$ curve obtained.

Feature	P@10	P@20	P@30	P@50	P@100
CAC	20.49%	15.87%	13.95%	12.61%	11.37%
HOG	45.87%	36.52%	31.63%	26.24%	19.85%
LBP	55.53%	43.54%	37.18%	30.00%	21.35%
Concatenation	44.80%	36.60%	32.36%	27.54%	21.48%
AlexNet fc7	73.20%	64.93%	59.71%	52.54%	41.80%
AlexNet fc8	71.48%	63.37%	58.55%	51.80%	41.66%
VGG-16 fc7	73.20%	64.93%	59,71%	52.54%	43.75%
VGG-16 fc8	69.91%	61.40%	56.36%	49.86%	39.98%
VGG-19 fc7	72.68%	64.27%	58,89%	51.87%	41.09%
VGG-19 fc8	69.54%	60.86%	56.09%	49.88%	40.02%
Concatenation fc8	**75.56%**	**67.53%**	**62.50%**	**55.94%**	**45.45%**

TABLE I
Precision@K obtained for each feature, for K = 10,20,30,50,100.

Considering that each class of the evaluated dataset has 100 instances, the P@100 values show that when the 100 nearest neighbours of a query are returned, roughly half of the true nearest neighbours are contained in this set when the concatenation of the fc8 layers of the CNNs was used. The higher the precision obtained, the less irrelevant samples of the dataset are needed, which will result in the expected saving of computation time as only relevant samples are to be processed.

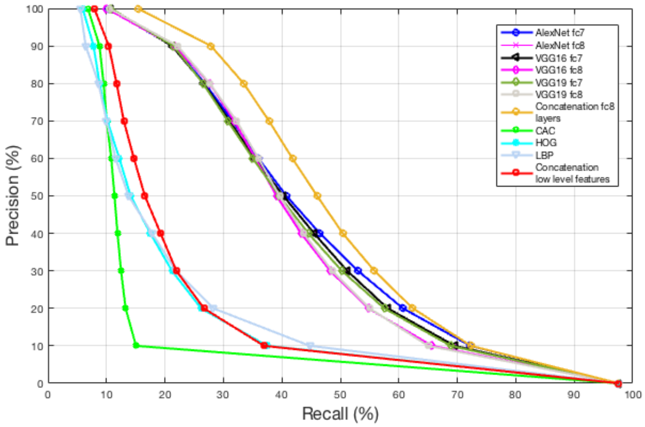

Fig. 3. Interpolated 21-points precision-recall curves of the selected features.

VI. Conclusion

This work evaluated the feasibility of using indexing techniques to speed up the access time of remote sensing imagery datasets by employing the LSH algorithm and using pre-trained CNNs as feature extractors. In comparison with classic low-level features previously used in such task, the use of deep features resulted in sharp improvements on the accuracy of the experiments performed. The next steps of this work incude fine tuning pre-trained CNNs on images from the RS domain and use the resulting features as visual descriptors, as well as the evaluation of other nearest neighbour search algorithms. Since the random nature of the unsupervised data-agnostic approach of the LSH algorithm causes its resulting hashing codes not to be optimal, special emphasis will be given to methods based on supervised approaches that construct hash functions as a latent layer in deep networks, performing the joint learning of image representations and the hash codes.

Acknowledgment

This work was financed by CNPq (grant 449638/2014-6), CAPES, and Fapemig (APQ-00768-14 and APQ-00449-17). We also thank Microsoft Azure for Research grant.

References

[1] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2014.

[2] J. A. Richards and J. Richards, *Remote sensing digital image analysis*. Springer, 1999, vol. 3.

[3] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 2, pp. 892–904, Feb 2016.

[4] J. Almeida, J. A. dos Santos, B. Alberton, L. P. C. Morellato, and R. d. S. Torres, "Phenological visual rhythms: Compact representations for fine-grained plant species identification," *Pattern Recognition Letters*, vol. 81, pp. 90–100, 2016.

[5] H. Liu, Q. Min, C. Sun, J. Zhao, S. Yang, B. Hou, J. Feng, and L. Jiao, "Terrain classification with polarimetric sar based on deep sparse filtering network," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2016, pp. 64–67.

[6] S. Radhika, Y. Tamura, and M. Matsui, "Application of remote sensing images for natural disaster mitigation using wavelet based pattern recognition analysis," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2016, pp. 84–87.

[7] Z. S. Zhou, P. Caccetta, N. C. Sims, and A. Held, "Multiband sar data for rangeland pasture monitoring," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2016, pp. 170–173.

[8] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios, "Learning euclidean embeddings for indexing and classification," DTIC Document, Tech. Rep., 2004.

[9] J. Wang, S. Kumar, and S. F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 3424–3431.

[10] H. YU, "Learning compact hashing codes for large-scale similarity search," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2015.

[11] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, *Handbook of Applied Cryptography*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 1996.

[12] J. Wang, W. Liu, A. X. Sun, and Y.-G. Jiang, "Learning hash codes with listwise supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3032–3039.

[13] J. Wang, W. Liu, S. Kumar, and S. Chang, "Learning to hash for indexing big data - A survey," *CoRR*, vol. abs/1509.05472, 2015. [Online]. Available: http://arxiv.org/abs/1509.05472

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[15] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, May 2004, pp. 889–894.

[16] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, ser. STOC '98. New York, NY, USA: ACM, 1998, pp. 604–613. [Online]. Available: http://doi.acm.org/10.1145/276698.276876

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[21] H. F. Yang, K. Lin, and C. S. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1011139631724

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 886–893. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.177

[24] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-fei, "Imagenet: A large-scale hierarchical image database," in *In CVPR*, 2009.

[25] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '10. New York, NY, USA: ACM, 2010, pp. 270–279. [Online]. Available: http://doi.acm.org/10.1145/1869790.1869829

[26] ESA. (2017) Sentinel online. [Online]. Available: https://sentinel.esa.int/web/sentinel/home

[27] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016. [Online]. Available: http://science.sciencemag.org/content/353/6301/790