PathoSpotter-Search: A Content-Based Image Retrieval Tool for Nephropathology

Ellen Aguiar
Laboratório de Computação de Alto Desempenho
Universidade Estadual de Feira de Santana
Feira de Santana, BA – Brazil
Email: chalegreaguiar@gmail.com

Rodrigo Calumby
Departamento de Ciências Exatas
Universidade Estadual de Feira de Santana
Feira de Santana, BA – Brazil
Email: rtcalumby@uefs.br

Luciano Oliveira
Intellligent Vision Research Lab
Universidade Federal da Bahia
Salvador, BA – Brazil
Email: lrebouca@ufba.br

Washington dos Santos
Instituto Gonçalo Moniz
Fundação Oswaldo Cruz
Salvador, BA – Brazil
Email: washington.santos@fiocruz.br

Angelo Duarte
Departamento de Tecnologia
Universidade Estadual de Feira de Santana
Feira de Santana, BA – Brazil
Email: angeloduarte@uefs.br

***Abstract*—Nephropathologists typically organizes their repository of digital images of kidney biopsies in such a way that it is difficult to retrieve cases that have images similar to a picture under analysis. Having this in mind, we initiated the development of PathoSpotter-Search, a Content-Based Image Retrieval system for images of kidney biopsies. The system operates as a cloud service to avoid the need to install any software on the pathologist's computer. Our approach combines a feature extractor followed by a similarity score calculator. We evaluated convolutional network (CN) architectures (VGG-16 (original and fine-tuned) and Inception-ResNet, and a network used in the proprietary classifier for glomerular hypercellularity), combined with Cosine and Euclidean distances as similarity scores. The first results have shown that the CN of the VGG-16 combined with cosine distance yielded the best performance (precision $\approx 53\%$). To assess the usability and functionality of the PathoSpotter-Search as a cloud service, the system was tested by nephropathologists and proved to be useful as a tool for retrieving similar images from their local repositories. Currently, we are working to improve the system precision to at least 70%, and evaluating strategies to retrieve similar images based on segments or tiles of the query image.**

I. Introduction

In digital pathology, to retrieve images from a repository is useful for supporting diagnostic tasks, comparing the progression of a given patient against others or even creating scientific illustrations [1], [2]. Unfortunately, the typical way a pathologist stores those images (often saving them in some local storage) makes it difficult to index and search them according to their similarity.

A possible solution for this problem is content-based image retrieval (CBIR) systems [3], which are computational tools used to retrieve relevant images from a repository based on their visual content. CBIR systems have been used in the medical field [4]–[7], and their performance greatly varies according to the field of application, tending to be low when the images from different subjects or categories have subtle differences between them, as frequently occurs in nephropathology. For example, in kidney biopsies, depending on the type of lesion of interest, the images from a healthy glomerulus may be just slightly different from the images of lesioned ones. Such similarities become an additional challenge for building an effective CBIR system for nephropathology. Having this in mind, we initiated the development of PathoSpotter-Search, a CBIR system that will be part of the set of tools of the PathoSpotter[1] project, which is dedicated to creating tools and knowledge for aiding the diagnosis and facilitating other tasks in nephropathology, as annotation of structures in Whole Slide Images and retrieval of similar cases based on images of glomeruli.

To develop PathoSpotter-Search, we used a typical CBIR architecture composed of a feature extractor followed by a similarity-score calculator. The system will operate as a web service, in which pathologists use the CBIR engine on their proprietary images without the need to upload them to any external server. This paper presents the first results obtained by the system and the future steps to improve its performance.

II. Methods

A CBIR system performs two essential tasks to be able to retrieve the most relevant images. The first is to extract and organize a set of descriptors from the images. It is typically represented as a feature vector or embedding, with the lowest possible dimension that still assures a robust differentiation among classes. The second task is ranking the images according to a similarity score between the descriptors, usually computing the distance between the vectors and ordering them accordingly.

The performance of a CBIR system is inherently constrained by the features adopted to represent the images [8]. Traditional approaches build the feature vector, combining filters and transformations to extract the features of the image. Such

[1] http://pathospotter.bahia.fiocruz.br

perspectives have to deal with several specificities, such as variations in scale and rotation. Several authors used techniques like SIFT [9], SURF [10], and HOG [11] to face such problems. Particularly for medical purposes, the development of good feature extractors have been considered a challenging problem [12], [13].

In the last years, Convolutional Networks (CN) have attenuated some problems of feature engineering by automatizing the learning and extraction of invariant features [14]. The combination of a CN with a Neural Network (NN), creates a Convolutional Neural Network (CNN), which has proven to be a successful architecture for several computer vision tasks, particularly image classification. This has motivated the use of Convolutional Networks as feature extractors for CBIR systems [13], [15]–[18].

Using an approach similar to the works presented by Swati [19] and Kumar [12], we chose the best PathoSpotter-Search configuration assessing the precision achieved by combinations of CN modules of two well known CNN architectures, VGG-16 [20] and Inception-ResNet [21], which have presented good results in several classification tasks, and also a third CN, a proprietary convolution network used in a classifier for hypercellularity [22]. We also evaluated two functions to compute similarity scores: cosine and Euclidean distances.

We started by using the approach proposed by Tajbakhsh et al. [23], which indicates that a pre-trained CNN with adequate fine-tuning outperformed or, in the worst case, performed as well as a CNN trained from scratch. Even so, because of the computational resources to fine-tune the Inception network, at this moment we just have finished the fine-tuning of VGG-16. The fine-tuning of VGG-16 was made using the public Kimia Path960 dataset [24] with 960 histopathology images of 20 different classes.

III. Experiments

We performed experiments for assessing the performance of the system in two different contexts, both using the architecture depicted in Figure 1. In one context, evaluate the performance of the system using a dataset with different types of glomerular lesions (Experiment 1). For this task, different lesions produce images with artifacts that are relatively easy to differentiate. In the other context, we evaluated the performance of the system using a dataset with variants of the hypercellularity lesion, in which the differences between images are more subtle (Experiment 2). In total, we assessed eight CBIR configurations based on four feature extractors (VGG-16 original and fine-tuned, Inception and Proprietary) and two similarity functions (cosine and Euclidean distances).

In Experiment 1, we used the dataset described in Table I, which is composed of images from three types of glomerular lesions (membranous thickness, sclerosis, hypercellularity) and images from normal glomeruli. In Experiment 2, we used 811 images of glomeruli separated into four classes: normal glomeruli and three types of lesion for hypercellularity (endocapillary, endomesangial and mesangial). The distribution of images per class is presented in Table II.

TABLE I
ARRANGEMENT OF THE FIRST DATASET.

Class	**Images**
Membranous thickness	869
Sclerosis	759
Hypercellularity	507
Normal	465
Total	2600

TABLE II
ARRANGEMENT OF THE SECOND DATASET.

Class	**Images**
Endocapillary	90
Endomesangial	179
Mesangial	238
Normal	304
Total	811

Provided that the images of the datasets are grouped into classes, to assess the precision of each CBIR configuration, we counted the number of images correctly retrieved using all images of the dataset as a query image in each test. An image was considered as correctly retrieved if it belonged to the same class as the query image. For example, if the query image presented a sclerosis, the k-retrieved images must present sclerosis. The precision for each class was calculated dividing the number of correct images retrieved by the total number of images in the class. We are aware that this is not the standard way to assess the performance of a CBIR system, which is the calculation of the mean average precision (mAP). Nevertheless, our interest was to gain information about the performance of this simple approach, in order to define a baseline to the system.

IV. Results

The results of the Experiment 1 and are available in Table III and Table IV, and Table V and Table VI, respectively. Before the experiments, we hypothesized that the feature extractor (convolutional network) of the proprietary classifier would achieve the best performance, since it was trained with glomeruli images. However, that hypotheses was not confirmed by the results. Furthermore, neither one of the tested networks performed significantly better than the others, nor the networks presented a stable behavior between classes. Likewise, none of the distance methods generated a noticeably superior performance in the results.

TABLE III
PRECISION FOR EXPERIMENT 1 USING COSINE DISTANCE.

Class	**Prop.**	**VGG-16**	**Inception**	**VGG-16(ft)**
thickness	51.4%	**52.7%**	44.3%	51.9%
sclerosis	42.9%	40.1%	**45.0%**	42.5%
hypercell.	47.6%	**61.5%**	53.6%	58.9%
normal	**42.1%**	41.0%	41.9%	34.4%
average	46.5%	**48.7%**	45.9%	47.4%

Although there is no defined standard in literature, we established that a reasonable precision would be at least 70%.

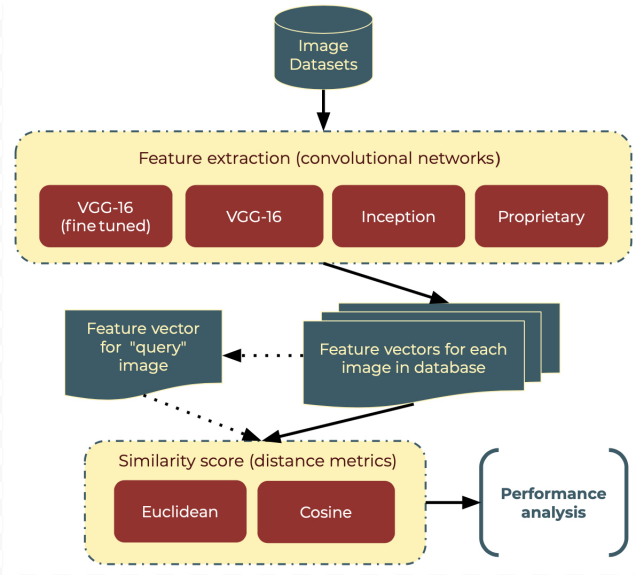

Fig. 1. Architecture of the experiments. For each convolution network, a set of feature vectors is calculated for all images in the datasets. Then, the similarity scores between each image in the dataset (working as a "query" image) and the remaining images are computed. The performance analysis evaluates how many of the **K** more similar images belongs to the class of the query image.

TABLE IV
PRECISION FOR EXPERIMENT 1 USING EUCLIDEAN DISTANCE.

Class	Prop.	VGG-16	Inception	VGG-16(ft)
thickness	52.3%	**59.0%**	43.7%	57.4%
sclerosis	41.3%	33.0%	**42.3%**	40.0%
hypercell.	46.9%	48.3%	49.9%	**52.2%**
normal	40.2%	**46.2%**	41.4%	36.6%
average	45.9%	47.0%	44.1%	**47.6%**

In Experiment 1 using images of different glomerular lesions, the VGG-16 architecture yielded the best average precision, although its performance was inconsistent between classes. This same architecture also outperformed its counterparts in Experiment 2, in which the images tend to be more similar to each other since they are related to the same glomerular lesion (hypercellularity). It can be noted that the average precision of the Experiment 2 was greater than the ones from Experiment 1 (except for the proprietary network), that may indicate that both, VGG and Inception, were capable to compute better discriminant features for hypercellularity than for other lesions.

It was somewhat a surprise that the feature extractor of the proprietary classifier performed poorly when compared to others. We initially hypothesized that the CN of this classifier would be able to generate good discriminative features for the classes. Although the classifier yielded an accuracy above 80% for hypercellularity multi-class classification [22], it is clear that is not generating features discriminative enough to yield a good performance for a generic CBIR system.

TABLE V
PRECISION FOR EXPERIMENT 2 USING COSINE DISTANCE.

Class	Prop.	VGG-16	Inception	VGG-16(ft)
endocapillary	14.9%	19.4%	15.4%	**20.7%**
endomesangial	27.4%	**40.3%**	40.0%	40.1%
mesangial	40.7%	42.7%	42.0%	**45.3%**
normal	61.2%	**79.7%**	72.3%	71.9%
average	42.6%	**53.4%**	50.0%	51.4%

TABLE VI
PRECISION FOR EXPERIMENT 2 USING EUCLIDEAN DISTANCE.

Class	Prop.	VGG-16	Inception	VGG-16(ft)
thickness	14.6%	18.3%	13.2%	**23.4%**
sclerosis	28.1%	24.1%	**39.9%**	28.0%
hypercell.	39.4%	35.7%	**40.7%**	40.4%
normal	61.8%	**87.7%**	69.8%	79.8%
average	42.6%	**52.9%**	48.4%	52.7%

A. *PathoSpotter-Search Cloud Service*

Using the configuration with the best results (VGG-16 without fine-tuning and cosine distance), we publish the system as a cloud service, so it could be tested by the pathologists of the PathoSpotter project. Although the performance is still below our 70% goal, the system received positive reviews for its usability and functionality, indicating its usefulness as a tool to aid nephropathologists in their daily tasks.

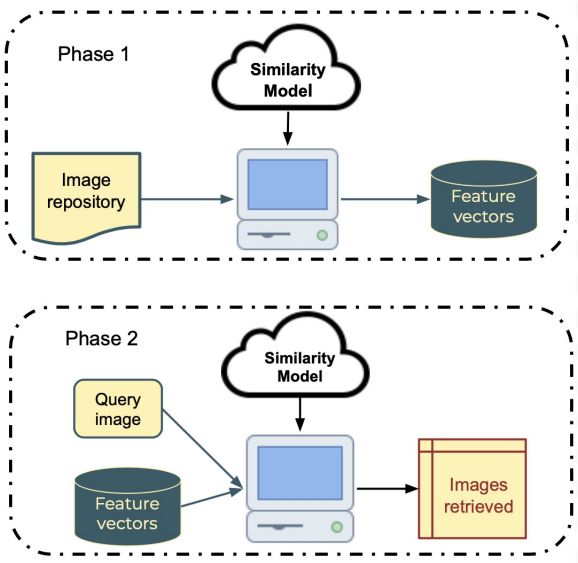

Fig. 2. Cloud service operation. In Phase 1 the system creates a database with feature vectors of the images in the repository. In Phase 2, the system retrieves K images from the repository according to their similarity with the query image.

The system operates as depicted in Figure 2. In Phase 1 (repository characterization), the pathologist indicates the folder with the repository of images of interest. Then the system creates a database of the feature vectors of those images. Finally, the pathologist saves this database for reusing it to retrieve images from this repository. It is worthy to note that, to assure the confidentiality of the information, no image upload is required for the system to operate, as all the computation of the feature vectors is done locally.

In Phase 2 (retrieving), the pathologist indicates the database of the feature vectors for the images of interest (built previously). Then the user loads the query image and informs how many (K) similar images the system must retrieve, and the system returns K images from higher to lower similarity scores.

V. Conclusion

This paper presents the initial results of the PathoSpotter-Search system, a CBIR system for digital images of kidney biopsies. As far as we know, PathoSpotter-Search is the first approach to build a CBIR system specific for nephropathology images. Using a traditional CBIR configuration, we evaluated the performance of four architectures of convolutional networks (original and fine-tuned VGG-16, Inception and a proprietary network) combined with two distance metrics (Euclidean and cosine) and tested them over two different data sets. Best result was obtained with VGG-16 original convolutional network as feature extractor, associated with cosine distance.

We are investigating why using fine-tuning in VGG-16 did not lead to a better performance for the system. It is also an open question why the proprietary CN used in the PathoSpotter-Classifier did not get higher results. The final CBIR was tested by nephropathologists using it as a cloud service, and proved to be useful as a tool to help in collecting similar images from datasets. Currently, we are improving the system performance to reach at least 70% of mAP, by evaluating other feature extractors and analyzing feature sensitivity. We also are evaluating strategies to retrieve similar images based on segments or tiles of the query image.

Acknowledgment

The PathoSpotter is partially sponsored by the Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), grants TO-P0008/15 and TO-SUS0031/2018, and by the Inova FIOCRUZ grant. Ellen Aguiar received a scholarship from FAPESB, grant 3627/2020. Washington dos Santos and Luciano Oliveira are research fellows of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants 306779/2017 and 307550/2018-4, respectively.

References

[1] N. A. M. Zin, R. Yusof, S. A. Lashari, A. Mustapha, N. Senan, and R. Ibrahim, "Content-based image retrieval in medical domain: A review," *Journal of Physics: Conference Series*, vol. 1019, p. 012044, jun 2018. [Online]. Available: https://doi.org/10.1088/1742-6596/1019/1/012044

[2] S. K. Sundararajan, B. Sankaragomathi, and D. S. Priya, "Deep belief cnn feature representation based content based image retrieval for medical images," *Journal of medical systems*, vol. 43, no. 6, pp. 1–9, 2019.

[3] D. A. Kumar and J. Esther, "Comparative study on cbir based by color histogram, gabor and wavelet transform," *International Journal of Computer Applications*, vol. 17, no. 3, pp. 37–44, 2011.

[4] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 8–20, 2017.

[5] K. Zheng, "Content-based image retrieval for medical image," in *2015 11th International Conference on Computational Intelligence and Security (CIS)*, 2015, pp. 219–222.

[6] A. H. Pilevar, "CBMIR: Content-based Image Retrieval Algorithm for Medical Image Databases," *Journal of medical signals and sensors*, vol. 1, no. 1, pp. 12–18, jan 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22606654https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317765/

[7] L. R. Long, S. Antani, T. M. Deserno, and G. R. Thoma, "Content-Based Image Retrieval in Medicine: Retrospective Assessment, State of the Art, and Future Directions," *International journal of healthcare information systems and informatics : official publication of the Information Resources Management Association*, vol. 4, no. 1, pp. 1–16, jan 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20523757https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2879660/

[8] X. S. Zhou and T. S. Huang, "Cbir: from low-level features to high-level semantics," in *Image and Video Communications and Processing 2000*, vol. 3974. International Society for Optics and Photonics, 2000, pp. 426–431.

[9] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[12] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data," *Journal of digital imaging*, vol. 26, no. 6, pp. 1025–1039, 2013.

[13] A. M. Rinaldi and C. Russo, "A content based image retrieval approach based on multiple multimedia features descriptors in e-health environment," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2020, pp. 1–6.

[14] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*. IEEE, 2010, pp. 253–256.

[15] A. Shah, R. Naseem, S. Iqbal, M. A. Shah *et al.*, "Improving cbir accuracy using convolutional neural network for feature extraction," in *2017 13th International Conference on Emerging Technologies (ICET)*. IEEE, 2017, pp. 1–5.

[16] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 36–45.

[17] K. Kruthika, H. Maheshappa, A. D. N. Initiative *et al.*, "Cbir system using capsule networks and 3d cnn for alzheimer's disease diagnosis," *Informatics in Medicine Unlocked*, vol. 14, pp. 59–68, 2019.

[18] A. Sezavar, H. Farsi, and S. Mohamadzadeh, "Content-based image retrieval by combining convolutional neural networks and sparse representation," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20 895–20 912, 2019.

[19] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, and J. Lu, "Content-based brain tumor retrieval for mr images using transfer learning," *IEEE Access*, vol. 7, pp. 17 809–17 822, 2019.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[22] P. Chagas, L. Souza, I. Araújo, N. Aldeman, A. Duarte, M. Angelo, W. L. Dos-Santos, and L. Oliveira, "Classification of glomerular hypercellularity using convolutional features and support vector machine," *Artificial intelligence in medicine*, vol. 103, p. 101808, 2020.

[23] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[24] M. D. Kumar, M. Babaie, S. Zhu, S. Kalra, and H. R. Tizhoosh, "A comparative study of cnn, bovw and lbp for classification of histopathological images," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–7.