Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images

Jorge A. Vanegas, Juan C. Caicedo, Jorge E. Camargo, Raul Ramos-Pollán, and Fabio A. González

Bioingenium Research Group
Universidad Nacional de Colombia
{javanegasr, jccaicedoru, jecamargom, rramosp, fagonzalezo}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. This paper describes the participation of the Bioingenium research group of Universidad Nacional de Colombia in the ImageCLEF2012 Medical Retrieval challenge, specifically in the ad-hoc image-based retrieval task. The methods used for solving textual and visual queries with which we submitted uni-modal runs are described. They were ranked 1st and 3rd respectively. These results have been obtained by using our own implementation of Okapi-BM25 weighting scheme for text retrieval, and by adding spatial layouts to the CEDD descriptors for visual retrieval. We also used these uni-modal features to learn multimodal representations using matrix factorization for solving visual queries. Despite the potential of multimodal indexes to improve the quality of visual queries, these experiments were not as successful as uni-modal indexes. We discuss the main findings of all these experiments.

Keywords: Image Retrieval, Medical Images, Multimodal Indexing.

1 Introduction

This paper describes the participation of the Bioingenium research group of Universidad Nacional de Colombia in the 2012 version of the Medical Image Retrieval challenge at ImageCLEF [7]. Our first motivation was to investigate the extent to which textual and visual indexes may be improved for searching in the collection of medical images, using keywords and visual examples separately. We aimed at designing suitable textual and visual representations by extending models that were successful in previous years, and preparing these representations for subsequent multimodal analysis.

For text indexing, we developed our own implementation of Okapi-BM25, which allows to determine limits on the number of terms used in the vector representation, by pruning irrelevant terms and keeping the most informative ones. For visual indexing, we introduced a spatial pyramid of CEDD features, making recursive partitions of the image and computing descriptors in each subregion. This representation extends the popular CEDD descriptor with spatial information, which results in an improved performance. We also implemented spatial extensions for bag-of-features histograms.

Our second motivation was to build an enhanced image index using both modalities, but for searching with visual examples only. The goal was to learn a multimodal representation that incorporates textual and visual information in the database, and then predict the multimodal representation for queries using visual features. This represents a very challenging problem since the medical image collection, with more than 300K images, constitutes a very large training set that poses computational difficulties for most learning algorithms. Other problems arise from this large multimodal image collection, such as the high dimensionality of textual and visual representations, and the presence of noise.

The results obtained with uni-modal strategies were successful in the general pooling, which ranked first in the case of textual queries among 54 other submissions, and third in the case of visual queries among 36 experiments. We consider that the multimodal indexing submission was not successful since it did not improve upon our own visual indexing strategy, which was the original goal. However, further experiments conducted off competition demonstrate interesting improvements. We believe that further research in this front may help to design more accurate image search systems working with the query-by-visual-example paradigm.

The structure of this paper is as follows: Section 2 briefly describes the medical image collection, Section 3 describes our text indexing approach, Section 4 presents the visual indexing strategies, Section 5 discusses the multimodal indexing approach for visual queries, and finally, conclusions and future works are outlined in Section 6.

2 The Data Set

The medical retrieval task of ImageCLEF 2012 is based on a subset of PubMed Central papers, containing 305,000 images extracted from biomedical articles. Participants have access to the selected images as well as all content of the corresponding articles. This year, a set of 22 topics was released for evaluation of the retrieval systems, where each one is composed of a variable number of images and associated text in 4 languages.

3 Text Indexing

Images in the collection belong to a medical article, so they can be indexed using the surrounding text content. Our goal was to build a term-document matrix using a vector space model with the Okapi-BM25 weighting scheme [6]. We developed an indexing tool using the Natural Language Toolkit for Python [1], which provides a clean API and extensive functionalities for common text processing tasks.

The text representation adopted in this work included information from the title of the paper and the image caption, which can be found in the XML file corresponding to each image in the data set. With that, a text corpus for the image collection was built, and standard text processing operations were applied,

including tokenization, stemming, and stop-word removal. These operations determined the initial list of indexing terms.

We designed a prunning criterion to discard irrelevant terms from the initial list, thus, preserving only the most informative ones. A limit for the number of terms was established depending on their document frequency. If it is outside of a predefined interval, the term is removed from the indexing list. The thresholds were computed according to a minimum and maximum number of documents in which a term is allowed to occur. The criterion is as follows:

$$keep(t) = \begin{cases} true & \text{if } min < \mathrm{df}_t < max \\ false & \text{otherwise} \end{cases} \tag{1}$$

where df_t is the number of documents that contain the term t, and min and max are parameters that define the minimum and maximum number of documents in which the term should appear. The definitive list of indexing terms is obtained by applying this rule, which is very useful to limit the dimensionality of the resulting vector space for indexing.

The term-document matrix is built using term frequencies in each document, and Okapi-BM25 [6] is used to highlight the importance of the most relevant terms. Usually, BM25 is used as a ranking function that involves different factors including: term frequencies, inverse document frequencies, and the length of both, the document and the query. However, in our approach, we wanted to use the ideas of BM25 as a term weighting scheme so that we can apply further processes to the term-document matrix (such as multimodal fusion). The following equation describes the BM25-based term weighting used:

$$\mathrm{weight}(t,d) = \left[log \frac{N}{\mathrm{df}_t} \right] \cdot \left[\frac{(k_1+1)\,\mathrm{tf}_{t,d}}{k_1 \left((1-b) + b \left(\frac{L_d}{L_{avg}} \right) \right) + \mathrm{tf}_{t,d}} \right] \tag{2}$$

where $\mathrm{tf}_{t,d}$ is the frequency of term t in document d, L_d and L_{avg} are the length of document d and the average document length in the collection, respectively, and, k_1 and b are positive tuning parameters to calibrate the term frequency scaling. We fixed $k_1 = 1.5$ and $b = 0.75$ according to the suggestions presented by Manning et al. [6]. For queries, we only used term frequencies without weighting, and the dot product similarity score was employed for document ranking.

3.1 Results

We submitted 2 textual runs using the indexing strategy described above, with the goal of evaluating the difference in performance after pruning the list of indexing terms. In both cases, we set a minimum frequency of 20 documents in which a term should be present to keep it in the list. The first experiment used 20,000 documents as maximum frequency, and the second experiment used 5,000 documents. These parameters resulted in vector spaces with approximately 28,000 and 18,000 dimensions, respectively.

Table 1. Retrieval performance of the submitted runs in the Medical Ad-hoc Image-Based Retrieval Task, using textual queries.

Run	Position	MAP	P@10
unal.text.bm25.20000	1	0.2182	0.3409
unal.text.bm25.5000	14	0.2045	0.2955
Extra (50,000)	N.A.	0.1991	0.3318

An additional experiment was evaluated after the competition finished to assess the contribution of the pruning strategy with respect to a longer list of terms. This experiment used 50,000 documents as maximum frequency, and produced a list of 29,000 terms. Notice how the number of indexing terms is controlled by the use of these two parameters, which remove rare terms as well as too common terms. We used this property to control the dimensionality of the resulting term-document matrix for further analysis, as is described later in Section 5.

Table 1 reports performance measures for the three experiments, the two first submitted to the official pooling and a third experiment run after the challenge. These results show the impact of the pruning strategy in the precision of the retrieval task, showing how the performance decreases by keeping or removing the wrong terms. The best response was obtained by the index limited by a frequency of 20,000 documents. This result ranked first in the category of textual experiments, and is the second best performance overall in the poolings for adhoc image-based medical retrieval.

Our second submission used 18,633 indexing terms, resulting in a significant dimensionality reduction, but also an important reduction in performance. This difference dropped the MAP performance in about 6%, leaving this experiment in the position number 14. However, notice that keeping more terms than those actually needed, can hurt the general retrieval precision even more. The additional experiment shows that a slight increase in the number of terms resulted in a decrease in performance of about 9%.

4 Visual Indexing

Our research group is currently leading an initiative to develop a framework for large scale image analysis for academic and scientific applications. The framework, named BIGS[8], is implemented in the Java programming language and integrates a wide variety of image processing tools, including feature extraction and learning algorithms. One of the most remarkable characteristics of BIGS is that it can easily run in a distributed environment with heterogeneous computing resources, from laptop and desktop computers to high-performance servers.

Obtaining a good quality representation for image contents in large databases is a challenging task, and BIGS was used to tune up image indexes by conducting experiments on the ImageCLEFmed 2011 data set. The experiments were run on different servers scattered throughout our lab, using BIGS to process all images

stored on an HBase NoSQL database[1]. In spite of the large size of the image collection, having an lightweight experimental lifecycle as provided by BIGS was key to be able to gain understanding on how to better tune up image indexes.

As a result, we designed two indexes for content-based image retrieval for this year's data set, focusing on including spatial information in the representation, since it can help to better discriminate medical image arrangements. The Color and Edge Directivity Descriptor (CEDD) [3] was used as basic low-level characteristic in both indexes, since it has demonstrated good performance in image retrieval tasks, while keeping a small and compact representation.

4.1 Spatial Pyramid CEDD

The CEDD descriptor is a compact representation of the image content, consisting in a histogram of 144 bins to codify information of colors and edges. The small size of this descriptor makes it an excellent choice for indexing large scale image collections. This descriptor has been previously evaluated in the context of medical image retrieval at ImageCLEF, exhibiting a competitive performance due to the variety of image modalities and visual configurations in this data set.

We extended this representation by computing the CEDD descriptor in a recursive partition of the image in quadrants, forming a pyramid of spatially organized regions [5]. We employed a configuration using the full image plus 2 pyramid levels, which results in 21 spatially distributed regions, ending up in a visual representation with 3,024 features. These descriptors were computed from high-resolution images, i.e., as they are distributed in the ImageCLEFmed data set.

This descriptor was computed by the BIGS framework using 40 workers deployed in several computers at our lab. The total time required to index the full image collection of 305,000 images using this strategy was 37 minutes. Finally, the similarity between two images is calculated on this descriptor using the Tanimoto coefficient. Assuming that x and y are vector representations of the spatial pyramids for two images, this is computed as:

$$T_D = t(x, y) = \frac{x^T y}{x^T x + y^T y - x^T y} \tag{3}$$

4.2 Spatial Bag-of-Features

An image index using the bag-of-features representation [4] was introduced in our experiments as well. The bag-of-features methodology is comprised of 3 main procedures: extraction of local features from images, construction of a dictionary of visual words, and the computation of the histogram for each image. Spatial layouts can be added to enhance the representation with the relative position of words in the image plane. In that sense, this representation incorporates local,

[1] http://hbase.apache.org/

low-level information of images as well as global, spatially distributed arrangements.

For local features, we extracted blocks of 32×32 pixels on a regular grid and the CEDD descriptor is computed in these patches. The k-means algorithm is used to cluster a large sample of patches extracted from the collection, for building a dictionary of 5,000 visual terms. The histogram is constructed by counting the occurrence of dictionary words in each image. Besides the global counting of visual patterns, each image is also split in 3 horizontal, non-overlapping strips, and an additional histogram is computed there to estimate the spatial distribution of visual words. This results in four bag-of-features histograms that are bounded together in a single image descriptor with 20,000 features.

This representation was also computed using the BIGS framework with 40 workers deployed in several computers at our lab. The total time required to extract this representation for all images in the collection was 116 minutes, which is less than one hour and a half. The similarity measure computed for this representation is the histogram intersection, for two images with histograms x and y:

$$K_{HI}(x, y) = \sum_{i=1}^{n} min\{x_i, y_i\} \tag{4}$$

4.3 Visual Queries with Multiple Images

The topics proposed for this year's challenge included 22 different queries with multiple images, some of them with 6 or even 7 example images. Since a single ranking is required for queries with multiple image examples, a similarity integration rule was employed. The similarity score for a database image d with respect to a multi-image query $q = \{q_1, q_2, ..., q_n\}$, is obtained as follows:

$$score(d, q) = \sum_{k=1}^{n} similarity(d, q_k) \tag{5}$$

4.4 Results

We submitted two runs, one using the spatial pyramid of CEDD features and another with the spatial bag-of-features. The results are reported in Table 2, and shows that the spatial pyramid obtains a significantly better performance than the bag-of-features, both in general precision (MAP) and early precision (P@10 and P@30). The difference can also be observed in the positions obtained by these experiments in the general poolings, the spatial pyramid was ranked 3rd, whereas the bag-of-features was ranked 14th.

The spatial pyramid extension for the CEDD descriptor demonstrated to be an effective representation to discriminate more relevant images in this task. In addition, computing the spatial pyramid did not result in an excessive load of both, computational effort and representation length. This representation is still

very light to compute with respect to the bag-of-features and keeps a compact descriptor with about 3,000 features.

In our preliminary experiments, we observed that adding a spatial layout on the image representation improves the performance of the medical image retrieval task. The two visual representations proposed in this work include spatial information using recursive computations of the same descriptor in partitions of the image. One of the reasons the spatial pyramid CEDD presented better performance than the bag-of-features is because of the level of granularity in the recursive partition, that allows to introduce more spatial details. This can be achieved because of the short length of the original CEDD descriptor, as opposed to the large dictionary of visual features that we employed in these experiments.

Table 2. Performance measures of the submitted runs in the Medical Ad-hoc Image-Based Retrieval Task for visual queries.

Run	Position	MAP	P@10	P@30
unal.visual.pyramidal.cedd.tanimoto	3	0,0073	0,0636	0,05
unal.visual.spatial.bof.3x1	14	0,0033	0,0455	0,0364

5 Multimodal Indexing for Visual Queries

One of our motivations to design textual and visual indexes for medical image collections is to develop a multimodal framework to integrate both modalities in a common representation. We focus our attention to the specific case of enhancing visual search functionalities by introducing available text information into the visual index. Thus, the goal is to improve the retrieval response using multimodal information even when users search with example images only.

In this work, we employed a multimodal latent factors model proposed in [2] for learning the relationships between visual features and text terms. The method is based on a matrix factorization algorithm, that proceeds with a multimodal decomposition of the visual and text matrices on a training data set. The matrix factorization problem is defined as follows:

$$\min_{P,Q,H} \frac{1}{2} \left(\|V - PH\|_F^2 + \|T - QH\|_F^2 + \lambda \left(\|P\|_F^2 + \|Q\|_F^2 + \|H\|_F^2 \right) \right) \qquad (6)$$

where $V \in \mathbb{R}^{n \times \ell}$ is the matrix of n visual features for ℓ training examples, $T \in \mathbb{R}^{m \times \ell}$ is the matrix of textual information with m terms, $P \in \mathbb{R}^{n \times r}$ is the transformation from the visual space to a multimodal space with r factors, $Q \in \mathbb{R}^{m \times r}$ is the transformation from the textual space to the multimodal space, and $H \in \mathbb{R}^{r \times \ell}$ is the multimodal latent representation for the training images. λ is a regularization parameter for this learning problem.

The solution to this problem presented in [2] is an online matrix factorization algorithm that can be scaled up to large data sets. This is specially useful for the

ImageCLEFmed 2012 collection, which has a large number of images that can be used for learning multimodal relationships between visual and textual information. When the linear transformation functions P and Q have been learned, new images can be projected to the multimodal space using the following equation:

$$h = \left(P^T P + \xi Q^T Q\right)^{-1} P^T v \tag{7}$$

where h is the multimodal representation for an image with visual features v, and ξ is a regularization parameter. The purpose of using these algorithms is to obtain a multimodal latent factor representations for all images, even if they do not have available text annotations, as may be the case of the queries. Using the multimodal representation, the ranking of images is computed using the dot product similarity measure, which indicates the extent to which two images share the same latent factors.

This strategy has demonstrated to be an effective method to learn multimodal relationships from image collections with attached texts, resulting in a data-driven representation for images that incorporates both modalities. Previous studies have shown important performance gains for these approaches, since visual features are complemented by the semantics of text descriptions, providing an enhanced mechanism of representing images.

5.1 Results

To construct a multimodal index for image search, we employed the matrices of text terms and visual features described in Sections 3 and 4, respectively. More specifically, we used the term-document matrix with 18,000 terms weighted with Okapi-BM25, and the visual matrix with 3,024 spatial pyramid CEDD features. One of the reasons we were interested in designing textual and visual indexes with bounded dimensionality is to reduce the computational cost of learning multimodal relationships.

The online multimodal matrix factorization (OMMF) algorithm was trained with the full collection of images in this challenge, i.e., using the 305,000 images with their corresponding text annotations. An implementation of the algorithm in the Java programming language was employed, which decomposed the matrices of 18,000 rows for text data, and 3,024 rows for visual data, with 305,000 columns in both cases, in 131 minutes in average. This algorithm has been designed to learn from as many examples as possible in a short time.

To tune up the learning algorithm and determine appropriate parameters for the factorization, experiments were conducted with the ImageCLEFmed 2011 collection. We found good parameters to solve queries in the previous year's challenge, that included 600 multimodal latent factors and other regularization parameters as needed. The criteria to select parameters for this algorithm is to observe improvements with respect to the direct visual matching, i.e., with respect to the visual indexing methods presented in Section 4, since the queries used in this experiment are also based on example images only.

Table 3. Performance results for multimodal indexing to solve visual queries. The first row reports the baseline method based on visual features only. The second row presents the results of the run submitted to the official poolings. The third row reports the result of an additional experiment run off competition.

Run	Position	MAP	Improvement	P@10	Rel-Ret
unal.visual.pyramidal.cedd.tanimoto	3	0,0073	N.A.	0,0636	117
unal.cedd.factorization.600	19	0,0024	-67.1%	0,0091	45
Additional experiment	N.A.	0.0087	+19.2%	0.0182	137

With the parameters that showed improvements in the 2011 collection, we prepared and submitted a run to the official poolings. Table 3 reports the results of this submission, as well as two other experiments for comparison. The first experiment in the Table is our baseline method, based on direct matching of visual features. The second result is the performance of the prepared run that has shown a decrease in performance with respect to the baseline. This loss is mainly explained by the use of parameters tuned to improve the performance in the 2011 challenge.

There are several differences between the challenge of 2011 and 2012. First, the nature of the proposed topics varied significantly, as this year's queries included more example images per topic, in average. Second, the size of the collection was increased, which resulted in bigger matrices in both dimensions. Third, this year's visual queries seem to be more difficult to answer, judging by the relative decrease in MAP observed in the results from 2011 to 2012. All these aspects may require a different configuration for the learning algorithm, in order to make it effective to retrieve more relevant results.

The results reported in the third row of Table 3 present the performance measures for an additional experiment run off competition to estimate the potential of the OMMF algorithm to improve upon the baseline. This result was obtained by tuning the algorithm parameters more appropriately for this year's task, and shows an important relative improvement.

The main goal of a multimodal algorithm in this context is to extract meaningful relationships between visual features and text terms. An additional challenge that makes the multimodal indexing strategy difficult to setup correctly, is attributed to the properties of the textual modality, which is very noisy and unstructured. Extracting semantic information useful for image analysis in this condition is still a very interesting research problem that requires further analysis.

6 Conclusions And Future Work

This paper presented the participation of the Bioingenium research group of Universidad Nacional de Colombia in the ad-hoc image-based medical retrieval task at ImageCLEF 2012. We submitted 5 runs: 2 textual and 3 visual, from which one was ranked first in the text modality and another was ranked third in the vi-

sual modality. These results were obtained by incorporating simple and effective extensions to well-known strategies for this task. We also explored multimodal indexing to answer visual queries, which is a very challenging and interesting research problem, that still requires further analysis. We believe that this is a promising research direction for improving image search systems, and the study of these models are the focus of our future research.

One of the main difficulties of this year's challenge was the size of the database, which required efficient computational tools to process and index the collection. In this work, we supported all of our visual indexing experiments on a distributed computing framework for large scale image analysis, named BIGS [8]. This framework allowed us to accelerate the exploration of visual indexing strategies, and investigate new image representation designs, such as the spatial pyramid CEDD that ranked third among 36 other experiments. We also used online learning algorithms for extracting multimodal relationships efficiently by training with the full collection of medical images in short execution times.

Acknowledgments

This work was partially funded by the project Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, number 110152128803 and project Sistema para la Recuperación de Imágenes Médicas Utilizando Indexación Multimodal, number 110152128767 by Convocatoria Colciencias 521 de 2010.

References

1. Steven Bird. Nltk: The natural language toolkit, 2002.
2. Juan C. Caicedo and Fabio A. Gonzalez. Online matrix factorization for multimodal image retrieval. In *17th Iberoamerican Congress on Pattern Recognition, CIARP*, 2012.
3. Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems*, ICVS'08, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.
4. G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
5. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
6. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
7. Jayashree Kalpathy-Cramer Dina Demner Fushman Sameer Antani Ivan Eggel Müller, Alba Garcia Seco de Herrera. Overview of the imageclef 2012 medical image retrieval and classification tasks. 2012.

8. Raul Ramos-Pollán, Fabio A. González, Juan C. Caicedo, Angel Cruz-Roa, Jorge E. Camargo, Jorge A. Vanegas, John E Arévalo, Paola K Rozo, Santiago A Pérez, Jose D Bermeo, and Juan S Otálora. BIGS: A framework for large-scale image processing and analysis over distributed and heterogeneous computing resources. In *IEEE International Conference on eScience. To appear*, 2012.