Image Retrieval: Importance and Applications

João Augusto da Silva Júnior
Universidade Federal de Goiás - CAC
Computer Science Department
Catalão - GO, Brazil
jhonaugustjunior@gmail.com

Rodiney Elias Marçal
Universidade Federal de Goiás - CAC
Computer Science Department
Catalão - GO, Brazil
rod@wgo.com.br

Marcos Aurélio Batista
Universidade Federal de Goiás - CAC
Computer Science Department
Catalão - GO, Brazil
marcos@catalao.ufg.br

***Abstract*—This paper presents a brief analysis of the main techniques used for image retrieval, while pointing out the importance of this emerging technology. Due to the alarming growth of the Internet and the high volume of data, we emphasize the technique of CBIR - Content-Based Image Retrieval. For the case study, we developed one application of this technique over an image database using the Euclidean Distance metric.**

Keywords - content-based image retrieval; euclidean distance, indexing, feature vector;

I. Introduction

Advances in data storage and image acquisition technologies have enabled the creation of large image datasets. In this scenario, it is necessary to develop appropriate information systems to efficiently manage these collections. The commonest approaches use the so-called CBIR - Content-Based Image Retrieval systems.

Basically, CBIR systems try to retrieve images similar to a user-defined specification or pattern (e.g., shape sketch, image example). Their goal is to support image retrieval based on content properties (e.g., shape, color, texture), usually encoded into feature vectors. One of the main advantages of the CBIR approach is the possibility of an automatic retrieval process, instead of the traditional keyword-based approach, which usually requires very laborious and time-consuming previous annotation of database images. The CBIR technology has been used in several applications such as fingerprint identification, biodiversity information systems, digital libraries, crime prevention, medicine, historical research, among others [6].

This paper aims to introduce the importance and applications in the field of image retrieval. In particular, we focus in the technique known as Content-Based Image Retrieval. A simple implementation of this technique is then presented so that we can analyse the results.

This article introduces the motivation for image retrieval systems and its importance in Section II. Section III presents some techniques used in image retrieval, while Section IV discusses distances metrics. Some applications that take advantage of the CBIR technologies are discussed in Section V. Section VI explains the algorithm used in the implementation of a basic CBIR system. Next, we discuss experiments and results involving our CBIR system. Section VIII states our conclusions.

II. Motivation

Due to the increase of online users on the Internet, the amount of collections of digital images have grown continuously during this period, for example, in web applications that allows adding images and digital albums [1]. Also is important to note that the images are globally used. The influence of television, old photographs and games has contributed to this growth as well. Images are increasingly used to convey information, whether one local information, weather, advertising, etc. In this context, it is necessary the development of appropriate systems to manage effectively these collections [6].

Another problem was the complexity of image data, and these data can be interpreted in various ways, thus raising the question of how to work in order to manipulate these data and represent or establish policies to its content. This motivated the birth of the image retrieval area whose goal is try to solve those problems.

III. Techniques for Image Retrieval

In this section we present two common methods for image retrieval: TBIR - Text-Based Image Retrieval and CBIR - Content-Based Image Retrieval.

A. Text-Based Image Retrieval

It is an old method, starting in 1970s [8]. This technique requires a text as input to search for image. Example of queries would be "search results for flowers or even "search results for flowers added on 2014-10-05.". So, the keyword may be by image name, date of adding, deleting, modifying and others.

Main problems of the query by text:

- Unexpressed feelings, emotions;
- Many ways of saying the same thing;
- Synonyms and homonyms;
- Misspellings.

B. Content-Based Image Retrieval

Different from the previous one, *Content-Based Image Retrieval* take as input a query image and the goal is search similar images by color, texture or form, as a comparison. Example of queries of this techniques would be something like "search for results similar to that image containing flowers ".

So, the user owns an image of a flower and the search will return similar images to that query image.

A CBIR system is composed of a query interface for the acquisition of the query image, databases for storing indexing data and metrics, similarity and retrieval system. Figure 1 illustrates these schema:

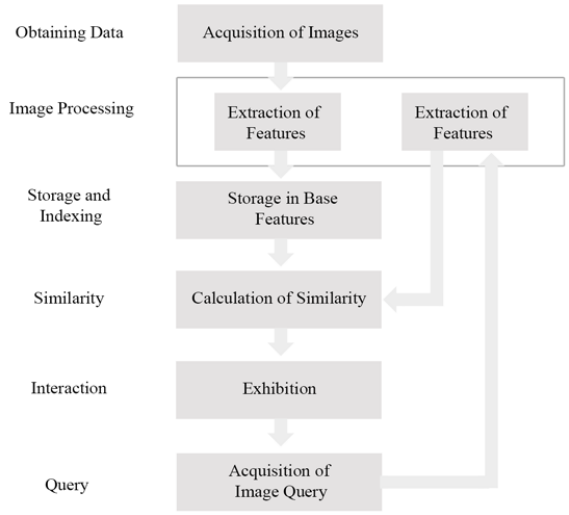

Fig. 1. Basic Steps of CBIR system.

It may be noted that CBIR is a process that may require long processing time due to the amount of images to be analyzed in a database, so comparison between images is made using a set of features extracted from the images [6].

The implementation of such a system requires the extraction and storing of the image features to be compared with the features of the query image. With this flow, the implementation process is more dynamic, since all features have already been stored somewhere.

IV. Distance Metrics

The calculation of similarity is given by the shortest distance between the feature vectors.

There are several functions responsible for calculating this distance, such as Minkowsky function and Euclidean distance. In practice, the latter is the most widely used measure of dissimilarity.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In this equation x_i and y_i are elements of features vectors x and y. Therefore, the vectors must have the same dimensions. The shortest distance vector gives the result of this calculation, that is, the smaller the distance between a vector and another, higher is the similarity for the analyzed image.

The Minkowsky function can be considered a generalization of the Euclidean distance. Its mathematical equation is given by:

$$D(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^P)^{\frac{1}{P}}$$

.

In this equation, the value of P can vary between 1 or 2. We have Euclidean distance in the case when P assumes the value of 2.

V. Applications

Content-Based Image Retrieval has been used in several applications, such as medicine, fingerprint identification, biodiversity information systems, digital libraries, crime prevention, historical research, among others.

A. Medical Applications

The number of medical images produced by digital devices has increased more and more. For instance, a medium-sized hospital usually performs procedures that generate medical images that require hundreds or even thousands of gigabytes within a small space of time. The task of taking care of such huge amount of data is hard and time-consuming. That's one of the reasons that has motivated research in the field of Content-Based Image Retrieval. In fact, the medical domain is frequently mentioned as one of the main areas where Content-Based Image Retrieval finds its application. [7].

B. Biodiversity Information Systems

Biologists gather many kinds of data for biodiversity studies, including spatial data, and images of living beings. Ideally, Biodiversity Information Systems (BIS) should help researchers to enhance or complete their knowledge and understanding about species and their habitats by combining textual, image content-based, and geographical queries. An example of such a query might start by providing an image as input (e.g., a photo of a fish) and then asking the system to *"Retrieve all database images containing fish whose fins are shaped like those of the fish in this photo"*. [6].

C. Digital Libraries

There are several digital libraries that support services based on image content. One example is the digital museum of butterflies, aimed at building a digital collection of Taiwanese butterflies. This digital library includes a module responsible for content-based image retrieval based on color, texture, and patterns. [6].

In a different image context, [5] present a content-based image retrieval digital library that supports geographical image retrieval. The system manages air photos which can be retrieved through texture descriptors. Place names associated with retrieved images can be displayed by cross-referencing with a Geographical Name Information System (GNIS) gazetter.

VI. Implementation

The Python programming language was chosen for developing the Content-Based Image Retrieval System that we mention in this work.

Python is a high-level programming language which supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles.

Python was conceived in the late 1980s by Guido van Rossum and has a large standard library, providing tools suited to many tasks, such as scientific computing, text and image processing.

The procedure used in the implementation can be summarized in the following topics:

1) Convert all images in a collection C to grayscale;
2) For each image I, apply a function that does pixels intensity counting. This function returns an feature vector where each of its positions contains the amount of intensity referring to each pixel in grayscale.
3) For each i in the feature vector produced in the previous step, apply a normalization function. There are a number of ways for computing such normalization. The one chosen here divide each value of the vector by the highest value of the same vector. This way, the resulting vector only contains values between 0 and 1.
4) Let Q be an query image. Apply to Q all the very same previous functions. Then, apply the Euclidean Distance between the vector of Q and each vector of the images I. The images with shorter distances mean that they are the most similar to the query image. We use the Euclidean Distance because it is easy to calculate and it's the most common technique.

Taking in consideration that the dataset for the images may be large and still contains high-definition images, the process for obtaining the feature vectors ends up being infeasible every time a query is requested. Ideally, such a process should be performed only once every time the dataset suffers an update. So, the feature vectors should be stored in a data structure in order to facilitate and further optimize performance and time for the retrieving process. Having this ideia in mind, we decided to divide the implementation in two phases. Each phase correspond to the execution of a python script:

- *generate.py*: this script is responsible for creating and storing in the file *db.txt* all the feature vectors from the images in our dataset. This file *db.txt* is just a regular text file containing two cólon separated columns: the first column has the name of the image being characterized; the second has the normalized feature vector itself.
- *query.py*: Using the Euclidean distance metric, the task of this script is to find the images from the database that are closer (i.e., images that have the lowest distances) to the query image. To accomplish this task, this script uses the data file *db.txt*. The script output is a report in HTML *(HyperText Markup Language)* that is used for easy exhibition of the results. The results consist in 16 images that are closer to the input image (or query image).

VII. Experiments and Results

In our Content-Based Image Retrieval system experiments, we used the *Corel* [4] image database. This database has 10,800 images belonging into 80 concept groups, e.g., autumn, aviation, bonsai, castle, cloud, dog, elephant, iceberg, primates, ship, stalactite, steam-engine, tiger, train, and waterfall. All the images have an approximated size of 5KB and dimensions 120 x 80 pixels or 80 x 120 pixels.

The first step was to use the script *generate.py* to generate the feature vectors of all the images of our *Corel* database, thus producing the textual file *db.txt*. The time taken for this task using a Intel Corel i5 2.5GHz 16GB RAM machine was approximately 1 minute. Soon after this task, we used the script *query.py* to generate a report in *HTML* format containing the 16 closest images of the image input by using Euclidean distance metric.

Figure 2 shows the query image and figure 3 shows, left to right, all the results in descending order of similarity.

Fig. 2. Query or input image.

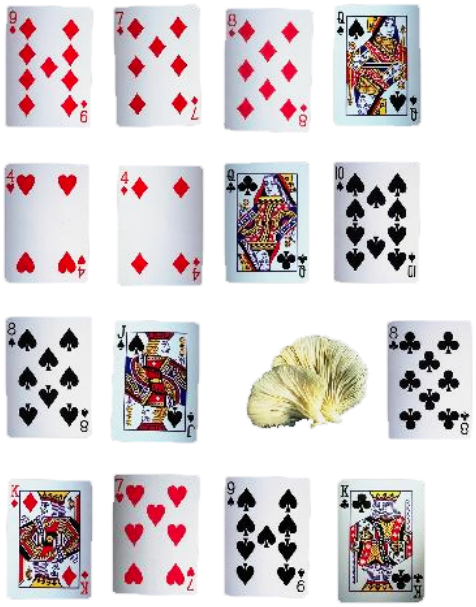

Fig. 3. From left to right: images most similar to the query image by using the Euclidean distance.

The second experiment used the query image shown in figure 4. The results can be seen in figure 5.

Fig. 4. Query image from the second experiment.

Fig. 5. From left to right: images most similar to the query image used in the second experiment.

Third experiment used the query image shown in figure 6 and its results are show in figure 7.

Fig. 6. Query image of the third experiment.

In all experiments, the response time for the retrieval was about 2 seconds.

The results demonstrate that the ranking of the returning images really take in consideration the lowest distances computed based in the intensity counting of the pixels in grayscale. Also, the first image in the result set is the same as the query image. This happens because we have included the query image in our image collection just for the checking if there would be an exact match. This means, then, that the distance computed for identical images is null.

The experiments revealed that, although the technique we used was simple, our Content-Based Image Retrieval system still may return good results when the query image and the imagens in the dataset have a homogeneous aspect.

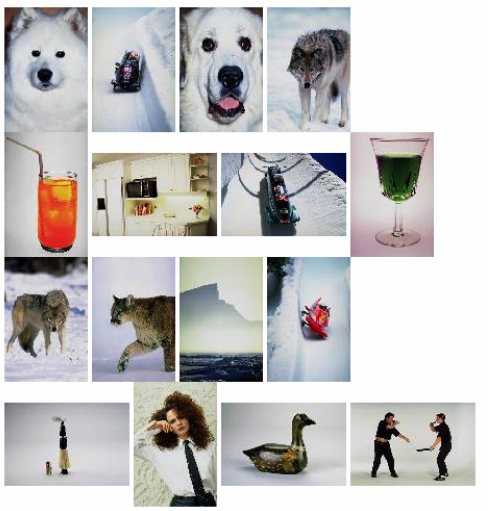

Fig. 7. From left to right: images most similar to the query image of the third experiment.

VIII. Conclusions

Advances in data storage and image acquisition technologies have enabled the creation of large image collections. Thus, we need appropriate information systems able to efficiently manage such collections. Systems that address this task are commonly known as *Content-Based Image Retrieval Systems* whose operation is basically trying to retrieve images similar to an image sample. For this purpose, parameters such as shape, color, texture, etc. are used, usually encoded in a feature vector.

This paper presented an overview of the field of image retrieval. In special, we emphasized the *Content-Based Image Retrieval Systems*. Concepts, applications and experiments using the method of *Euclidean Distance* for calculating similarity between two images served to elucidate aspects of this area that is gaining more and more importance each day.

As already mentioned in this paper, to optimize the response time of a search system and make it truly scalable on a collection of large images, it is necessary to use a data structure for efficient representation and indexing of the feature vector. In addition, there are several ways to calculate the similarity distances between two images. As a future work, we propose a comparative study of the various measures of distances, as well as the techniques of extraction and indexing of the feature vector.

Acknowledgments

The authors would like to thank CNPq and FAPEG for the financial support.

REFERENCES

[1] Santos, Ana Paula de Oliveira - Recuperação de Imagens Mamográficas Baseada em Conteúdo. Trabalho de Conclusão de Curso. Fundação de Ensino Eurípides Soares da Rocha, 2006.

[2] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[3] John Eakins and Margaret Graham. Content-based Image Retrieval - University of Northumbria at Newcastle

[4] Wang, James Z. and Li, Jia and Wiederhold, Gio. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947-963, 2001.

[5] B. Zhu, M. Ramsey, and H. Chen. Creating a Large-Scale Content-Based Airphoto Image Digital Library. IEEE Transactions on Image Processing, 9(1):163-167, January 2000.

[6] R.S. Torres and A.X. Falcão. Content-Based Image Retrieval: Theory and Applications. Revista de Informática Teórica e Aplicada, nro 2, volume 13, 2006, pags 161–185.

[7] Castanon, Cesar Armando Beltran. Recuperação de imagens por conteúdo através de análise multiresolução por Wavelets. São Carlos : Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2003. Dissertação de Mestrado em Ciências de Computação e Matemática Computacional.

[8] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng - Fundamentals of Content-Based Image Retrieval