



Content-Based Information Retrieval System for dermatological issues

17/03/2022

Carlos Alcoba
Javier Jiménez
Javier Pozo

M. Sc. in Health and Medical Data Analytics (UPM)



Table of contents

| | |
|--|----------|
| Code and data availability | 1 |
| Introduction and Motivation | 1 |
| Technical proposal | 2 |
| Implementation of the retrieval system | 2 |
| Principal descriptor: Gabor filter and Canberra distance | 3 |
| Smart Histogram Descriptor: Color histogram and Euclidean distance | 3 |
| Implementation of CBIR processes | 3 |
| Conclusions | 3 |
| Bibliography | 4 |

Code and data availability

All produced code is openly available in this [GitHub Repository](#). It includes a notebook that shows how the non-textual retrieval system works with a toy subset of images extracted from this [Kaggle Dataset](#), but it is designed to work with whatever number of images that are of the same size as those included there. In addition, the auxiliary Python script that computes the descriptors and distances, as well as the README.md containing the instructions on how to run the code, are also available there.

Introduction and Motivation

Even nowadays, dermatological injuries with very different causes are difficult to distinguish, which is one of the main causes of misdiagnosis – or late diagnosis – in this field. Luckily, in the current context, the amount of information available in image format from the Dermatology domain is huge enough to train very complex Machine Learning models which can constitute an important advantage in this field. Therefore, one of the most urgent tasks in experimental medicine consists of discovering new procedures to sift through all available dermatoscopic images to create appropriate databases for each specific research project.


For this reason, the aim of this work is to develop a CBIR that will automatically retrieve the most similar images to those the researchers are interested in; which will allow them to automatically prepare customized training or testing datasets whose applications are extremely wide.

However, this is not the only possible application of this system. For instance, it can also support the clinicians in their diagnosis by analyzing images. From clinical studies, it is known that colour and texture are two of the main features of this kind of analysis. For this reason, in this work, we will use these two characteristics to obtain the more similar images present in a database with respect to a query image.

Technical proposal

The methodology proposed for this work consists of two steps that are related to the low-level features that are more important in the dermatological field: colour and texture. In these steps, two different approaches are followed to retrieve the most similar images. These methods are based on the state of the art explained in (Shereena & Davis, 2014).

In the first step, the images will be ranked based on their texture similarity. For this purpose, a Gabor filter will be applied and a feature vector consisting of the means and



variance of each kernel will be created. With this feature vector, the Canberra distance will be computed between each of the images in the database and the query image. Our proposal is to design 16 different kernels combining 4 orientations and 4 frequencies.

The second step consists of the application of the colour histogram in the RGB domain for the selected images. In this step, one histogram is computed for each of the channels of the images corresponding to one colour. This histogram is compared with the euclidean distance to the histogram of the query image and ranked according to it. To avoid very long distances compared with those based on texture similarity, the result will be divided by the number of pixels present in the image which will give a percentage.

The process of searching is based on the idea of filtering and sorting: the Gabor filter retrieves the 10 most similar images regarding texture and the histogram sorts them attending to colour resemblance.

Implementation of the retrieval system

The implementation of the methodology proposed in the previous section has been done using the provided Python code with a toy set of approximately 80 images randomly selected among those available in the Kaggle *Derma Diseases* repository (Ripamonti, 2022). No preprocessing has been done to the images as they all have the same resolution. However, not all the interesting key points of the injuries are centred, so we have avoided any kind of similarity measure that prioritizes the central part of the images.


As has been explained in the previous section, two different descriptors with their corresponding similarity measures are implemented and combined to obtain the final ranking of all images, which is the output of the CBIR. Below, a deeper explanation of each of the steps is provided.

Principal descriptor: Gabor filter and Canberra distance

First of all, several functions are designed to both compute a set of *Gabor kernels* – with different orientations and frequencies – and to convolute them with all the images in the *data* folder to obtain, for each image, as many $\mu\text{-}\sigma$ (mean and standard deviation) pairs as the number of filters. Then, the distances among each of these pair vectors and the one of the query image is computed using the *Canberra Distance*, which provides our first criteria to sort the images regarding their texture similarity with the query.

Smart Histogram Descriptor: Color histogram and Euclidean distance

The selected smart histogram is an RGB one because of the importance of colour in distinguishing different kinds of dermal injuries. As commented before, we have avoided computing different partial histograms for each image because the key points of these



injuries are not typically located in the same places, so doing it could lead to a biased system. After separating the images in each of the three colours and computing their three correspondent histograms, all of them are compared with those of the query image via euclidean distance. This provides a ranking containing all of the images that sort them attending to their colour similarity to the query one.

Implementation of CBIR processes

First of all, the system requires the user to include the query image in the data folder and to indicate its index in the notebook line prepared for this purpose. Once the query is identified, the Gabor filters provide a first ranking of images from where the 10 more similar ones are selected. Then, those ten images are sorted regarding the results of the smart histogram to return the final ranking of images as the desired result of the CBIR in the form of an indexed Python list. The result after this *search and sort* process is, then, 10 retrieved images sorted by an overall similarity measure that considers both colour and texture.

Conclusions

The toy-CBIR that has been built is functional enough to provide a ranking of images that is, from a human (informal) point of view, correctly ordered in the sense that those images that are more similar to the query tend to occupy the first positions. Therefore, even if retrieval systems do not accept purely objective performance metrics, it is possible to conclude that it works in a satisfactory way.

In regards to future updates of this system, it is important to remark that it has been developed in a way that ensures the possibility of generalizing it to be used with images of different fields. The only requirement is that all the images in the data folder need to have the same size, which could be avoided using an automatized reshape function. In addition, if any other low-level feature is of the interest of the researcher, a filter to take it into account (e. g., a HOG for the shape) could be easily added to the auxiliary file. However, it is important to take into account in what order these features are going to be used to compute the ranking because it can affect the final result.

Bibliography

- V B, Shereena & M.David, Julie. (2014). Content Based Image Retrieval: A Review. Computer Science & Information Technology. 4. 65-77. 10.5121/csit.2014.4906.
- Ripamonti, P. (2022). Derma Diseases. Kaggle.com. Retrieved 16 March 2022, from <https://www.kaggle.com/paoloripamonti/derma-diseases>.