NON-TEXTUAL DATA EXTRACTION

Information retrieval, extraction and integration

GARCÍA DE VIEDMA PÉREZ LUCAS

CATALÁN GRIS LUCÍA

EIT Digital Data Science Master

TABLE OF CONTENTS

1 MOTIVATION

Rice ranks second in the most consumed primary aliments in the world. Majorly known for its use in Asia, China herself consumed around 31% of the 509.87 million tons consumed last year. It is also of great importance in other regions like Brazil or India, where it is considered to be the staple food. However, like other crops, rice has numerous diseases that may affect the production to the point of not getting grains from the plants when harvested.

Thus, we propose the development of a CBIR that receives an image and finds similar images from a dataset that contains images of three of the main illnesses rice suffers from bacterial blight, brown spot and leaf smut. Together with a tool that takes pictures from the crops periodically, this system could allow the automatization of rice diseases diagnosis, which could prevent losses in the harvest by taking the appropriate measures.

2 STATE OF THE ART

The study [1] focuses their work on developing a system for recognizing paddy plant diseases using image processing. The system should classify images of infected leaves between three classes of diseases: brown spot, bacterial blight and leaf blast disease. Their proposed methodology consisted on pre-processing the input image, extracting 144 Histogram Oriented Gradients features from the diseased region of the leaf and, finally, using these features and the SVM classifier to recognize the accurate disease based on maximum distance value. The results of the study show that the model is able to anticipate the sickness with a precision of 97.73% utilizing SVM. It can be concluded that the use of Histogram Oriented Gradients had a great success in summarizing the features of the paddy plant images, so it is going to be used as the histogram descriptor in this work.

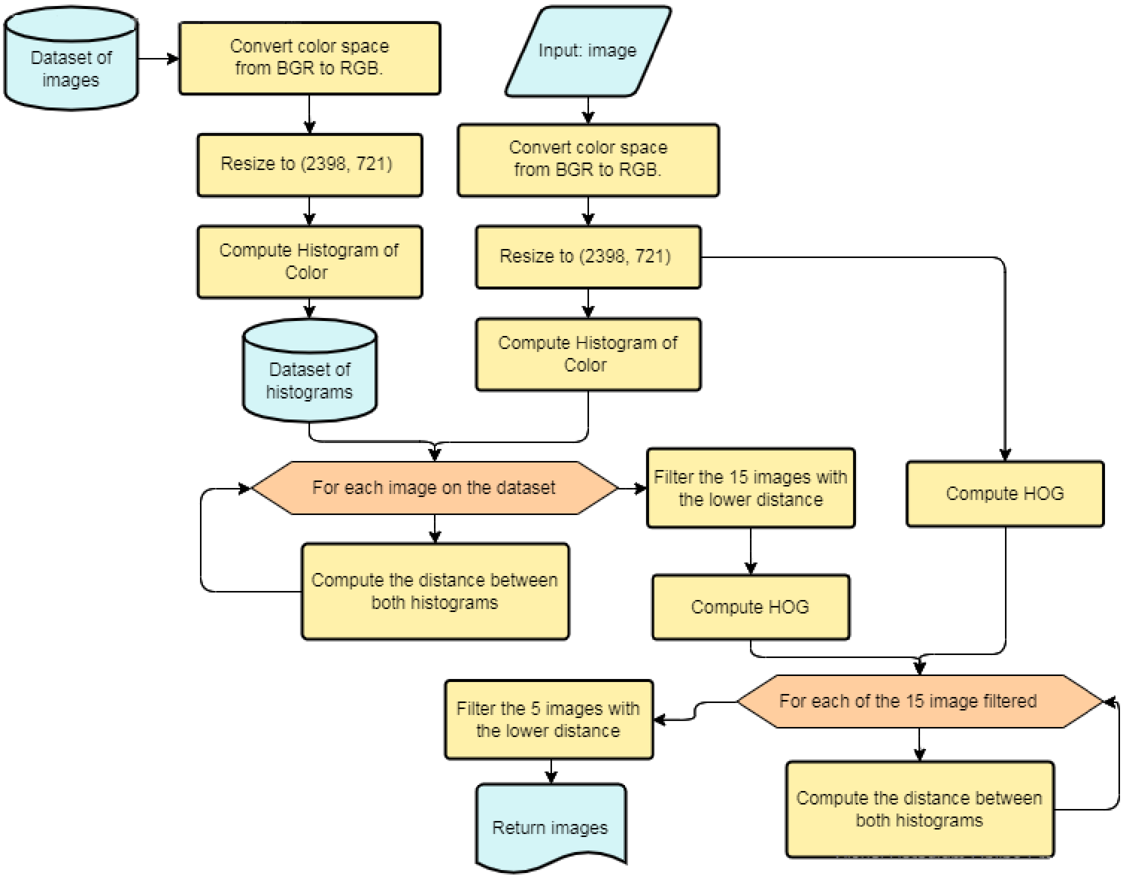
Since the study [1] focuses on image classification, which is not the purpose of this assignment, an alternative distance is going to be selected. In the conference transcription [2], they concluded that the Bhattacharyya distance was successful in identifying the presence of two types of diseases in rice crops, so it could be a good option for this work. Thus, we will compare the results with the Chi-square, Bhattacharyya and Euclidean distance.

Additionally, Histograms of Colour are going to be used as a complementary tool to help in the ranking process. It was chosen because in the paper [3] it is presented a method based on the Histograms of Colour for identifying plant disease based on colour, edge detection and histogram matching.

3 IMPLEMENTATION OF THE TOY CBIR

The CBIR was designed and implemented using a dataset of 120 images of paddy plant leaves that presented 3 different illnesses: 40 images of leaves which have bacterial leaf blight, 40 with brown spot and 40 with leaf smut. The original dataset was obtained from Kaggle through this link. **However**, as will be said in the following paragraph, we have set the background of the images to black to improve the performance. Since it took quite a few minutes to execute the whole process, we thought it was more convenient to store the altered dataset and it can be downloaded **HERE.**

The process followed by the CBIR can be seen in the following graph:



Dataset of images
Convert color space from BGR to RGB.
Input: image
Resize to (2398, 721)
Convert color space from BGR to RGB.
Compute Histogram of Color
Resize to (2398, 721)
Dataset of histograms
Compute Histogram of Color
For each image on the dataset
Filter the 15 images with the lower distance
Compute HOG
Compute the distance between both histograms
Compute HOG
Filter the 5 images with the lower distance
For each of the 15 image filtered
Return images
Compute the distance between both histograms

Figure 1: Process followed in the assignment

In addition, the performance of the system was also evaluated setting the background of all the images of the dataset to black, to avoid possible errors in the histograms of colour. In order to do so, we converted the images to HSV, applied a filter depending on their “value” value (pixels with a “value” higher than 220 would be set to 0. I.e., black colour), reconverted them to RGB and store them to avoid the burden of calculating them on every execution.

4 RESULTS

The results obtained with both the distance metrics, the different histograms and the two datasets are the following:

Used dataset	Used histograms	Result euclid. *	Result ChiSqr *	Result Bhatta. *
Original	HOG	0.66	0.48	0.67
Original	Histogram of Color	0.67	0.73	0.79
Original	Histogram of Color + HOG	0.68	0.71	0.75
Fixed background	HOG	0.62	0.46	0.60
Fixed background	Histogram of Color	0.77	0.81	0.82
Fixed background	Histogram of Color + HOG	0.74	0.70	0.81

* The values in these columns refer to the average percentage of plants that have the same disease as the one used as input among the best 5 ranked by the system. The average has been obtained with the percentage obtained for every single image of the dataset.

Surprisingly, the process that obtained the best results is the one where only the histogram of colour was used, together with the fixed black background and the Bhattacharyya distance (marked in darker green). Either way, we have included the three systems marked in green in the source code folder to allow testing them separately.

It's important to highlight that, in the combined system (HCL + HOG), we first filter with the histogram of colour because the statistics showed a better performance than using the HOG first. Also, we can observe how in the fixed background options, the HOG obtains worse results, because we completely remove the texture from the background.

5 REFERENCES

[1] K. Jagan Mohan and M. Balasubramanian, “Recognition of Paddy Plant Diseases Based on Histogram Oriented Gradient Features,” *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 5, no. 3, 2016.

[2] M. R. Tejonidhi, B. R. Nanjesh, J. G. Math and A. G. D'sa, “Plant disease analysis using histogram matching based on Bhattacharya's distance calculation,” in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016.

[3] S. Bankar, A. Dube, P. Kadam and S. Deokule, “Plant Disease Detection Techniques Using Canny Edge Detection & Color Histogram in Image Processing,” *(IJCSIT) International Journal of Computer Science and Information Technologies,* vol. 5, no. 2, pp. 1165-1168, 2014.