Statistical Analysis Of The Features And Classification Of Coffee Beans In Three Maturation Stages

Jose Alfredo Palacio-Fernández, William Orozco, Bayardo Cadavid

Abstract: This article presents a statistical analysis of the features of RGB, HSV, Wavelet and the relation of coffee axes based on the root square mean value, the standard deviation and the Wavelet approximation coefficients' average for the images obtained from three types of coffee beans with different maturation states. By means of a statistical analysis, the relations between the features were obtained and, three main components were selected. These were subjected to a Bayesian classifier, which allowed to determine a full classification of the three types of grains, using the two main components and, two other combinations of the features, mainly color in the second Wavelet transformation filtering level.

Index Terms: features, classifier, main components, coffee, wavelet, image

1. INTRODUCTION

The coffee's organoleptic quality, as well as its economic performance, are some of the reasons for approaching the study on improved coffee production, analyzing genetic aspects of species that provide soft or strong characteristics, typically organic production techniques, selection and drying among others. In terms of selection, it is important to analyze the grains' health characteristics, as well as the ripping state. When the grain is in a low ripping state, it requires more pulping and the grain has a lower weight; also at the end, it may present part of the pulp adhered to the final grain, requiring classifying it as Pasilla (Pasilla is the moldy, overcooked, or shriveled leftovers from commercial coffee production). It suggests low quality grains [1]. A semi-mature grain presents a lower quality state than a ripe grain but its weight is better compared to green grains. A grain of better quality provides the coffee grower a better income. This is how manual selection can be made if the crop is small; for large production volumes, the automatic selection is suggested. A selection process allowing the separation of three main stages of maturation (green, semi-mature and mature) is required. Due to the fact that coffee plants bloom at different seasons, on the same plant, grains with different development stages can be found; when harvested and subject to the milling process, these cause several problems, from the pulping stage on (due to lack of uniformity in size because thegreen grains do not reach their final growth size since the green grains do not have a well-developed mucilage; this is what allows a more efficient pulping [2]. In [3], perform a coffee beans classification from different sources but without the pulp, using neural networks. In [4], the ripping state selecting process starts representing each state as a class, for example, this work presents a segmentation into three main classes defined as mature grains, semi-mature grains and green grains.

The color is most frequently used feature for image recognition.

- *Jose Alfredo Palacio, Institución Universitaria Pascual Bravo, josealpa@pascualbravo.edu.co*
- *William Orozco Murillo, Institución Universitaria Pascual Bravo, william.orozco@pascualbravo.edu.co*
- *Bayardo Emilio Cadavid, Institución Universitaria Pascual Bravo, b.cadavid@pascualbravo.edu.co*

The colour is most frequently used feature for image recognition. Colour has significant advantages over other features like high frequency ease of extraction, invariant to size, shape and orientation and independent to background complication [5]. The RGB (Red-Green-Blue) components, taken from each grain's image, provides characteristics that, with a linear classifier, could separate these types of grains into the R (red) and G (green) components; but the semi-mature grain provides information on each component causing errors in the classification. Another type of representation that provides information inherent to each type of grain is the transformation H (hue), S (saturation) and V (value). This representation shows the different nuances of the image color; the lack of color or the closeness to the gray tone and the closest proximity to black or white. The HSV color space is related to the human vision; it tests the change of shade of a blurred image and it is invariant under certain circumstances. This feature leads to recover blurry images when dealing with brightness and saturation. The conversion from RGB to HSV is as follows [6] authors are suggested to present their articles in the section structure:

$$\begin{Bmatrix} 0 & si\ C = 0 \\ \frac{G-B}{C} + 6 & si\ M = R \\ \frac{B-R}{C} + 2 & si\ M = G \\ \frac{R-G}{C} + 4 & si\ M = B \end{Bmatrix}$$

$$H = 60°H' \quad , \quad V = M, \quad S = \begin{Bmatrix} 0 & si \quad V = 0 \\ \frac{c}{v} & Otherwise \end{Bmatrix}$$

Where M = max (RGB), C = M-m and m = min (RGB)

The HSV method for greenness identification was proposed by [7]. The method was proposed to identify plants from maize seedling images acquired outdoors. The greenness identification results of the visible spectralindex based methods are seriously affected by the image brightness. So a simple idea for resolve this problem is to find a new color space in which the color is not correlated with brightness. The HSV (Hue, Saturation and Value) color space naturally has this property. In the HSV color space, the color distribution of a

single-colored object is invariant with respect to brightness variationAnother transformation that was implemented and that allows to eliminate certain details of the image (focusing on its general component) is the 2D Wavelet transform; it is generally used as an image compression technique. It was used in this work to determine the statistical behavior of the characteristics that are deduced from it. Wavelets are small wave functions that focus on time and the frequency around a certain point. The Fourier Transforms only deal with the frequency component in a signal, while the temporal detail is not available. It is appropriate for non-stationary signals and varies both, for the frequency range and the spatial range. The Discrete Wavelet Transform uses a mother Wavelet. This research work used a Daubechies Wave, which gives an efficient result compared to other wavelets. The process involved filtering the image through low-pass and high-pass filters [8]. To both, the original RGB images, as well as to the transformed ones, an extraction of characteristics RMS, average, standard deviation is performed (besides the particular case of the characteristic that presents each grain's axe ratio). A method to determine the proximity between the clusters, formed by the data of each characteristic, is the Dendrogram of variables. This method can establish the closeness or similarity between variables [9]. To observe the general effects of different characteristics, a principal component analysis (PCA) was carried out, such as the one implemented by [10] for the total data set (20 samples per class and 19 variables). The relation between the PCA and the original variables can be represented in a Biplot, which is a 2D graph, where both loads represent variable information and the scores represent information from observations. In order to distinguish between the two sources of information, the loads are represented by arrows from the origin and points for the plot. The Biplot is generally built on the basis of the first two main components [11]. The best characteristics obtained through the PCA are used as the basis of the training stage of a Bayesian classifier; it is one of the most used classification techniques in data mining and automatic learning. The Bayesian classifier, works based on the Bayesian rule and probability theorems [12].

2. MATERIALS AND METHODS

Initially, it was used a database of 60 images divided into three groups (green, semi-mature and mature) to which a statistical analysis is performed using functions of Matlab®, as well as the classification process by the Bayesian analysis implicit where the same a priori probability is defined. The grains used have the following RGB representation (Figure 1)

Figure 1. *RGB Representation of the employed grains*

And the HSV representation of the same types of grains are presented in Figure 2.

Figure 2. *HSV Transformation of three grains (green, semi-mature and mature)*

The Wavelet Transform of the original images was made using the Matlab® Wavedec function for each grain, using a level 2 decomposition and an order 2 Daubechies Waveletder. The mean values of each image were extracted as well as the standard deviation (Eq. 1) and the RMS value (Eq. 2), as well as each grain's axe ratio, as shown in Figure 3.

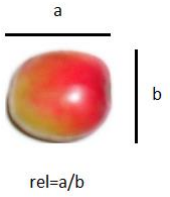

Figure 3. *Principal axes relation*

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{1}$$

$$X_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_i|^2} \tag{2}$$

At the end, 19 characteristics are obtained to which a projection is made; this provides the main components that represent the three most important classes. The three main components obtained, as well as the other characteristics, were tested through a Bayesian classifier, assuming an equal a priori probability for the three types of grains. A training, using 20 grains of each type, is carried out and it is evaluated in the distribution functions (Eq. 3) generated for each class or grain type in the training stage.

$$p(x) = \frac{1}{(2\pi)^{1/2}|S|^{1/2}} e^{-\frac{1}{2}(x-m)^T S^{-1}(x-m)} \tag{3}$$

3. RESULTS

From the characteristics Dendrogram (in total 19) (Figure 4) and the Biplot diagram (Figure 6), a noticeable relation between the mean's variables of both, the R component of the original RGB image (meanRGBR), as well as the Wavelet Transform signal for the same component (meanWDR) are observed. The same occurs between the V component of the HSV transformation in the RMS characteristic, as well as the mean (rms_hsv3 and mean_hsv3). In general, the relation that originates the lower branches of the Dendrogram can be observed. In the Biplots, this behavior is observed but with the angle´s similarity of some characteristics on the axes of the two main components obtained by the projection of the same original variables.

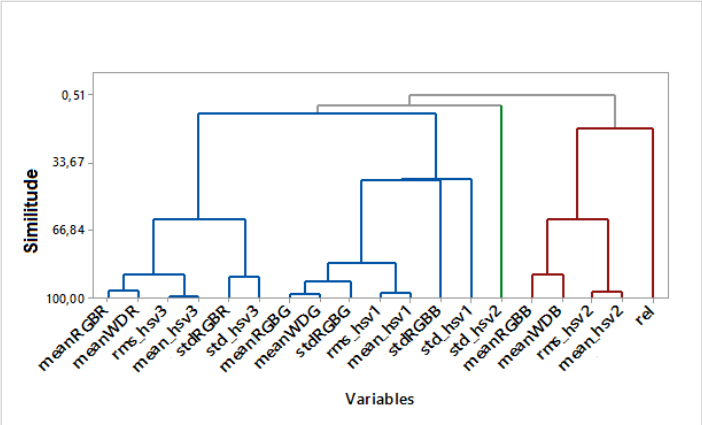

***Figure 3**. Characteristics Dendogram based on the relative correlation coefficient's distance.*

From the conglomerate in Figure 4, the similarity in the first 11 characteristics is perceived. Also, it can be noted that the rel characteristic is away from the others' behavior due to its spatial nature, not its color. PCA is a multivariate statistical method. Its main idea is to get the principal components through linear transformation, which is obtained from the eigenvalues, which is shown in Figure 5 ordered descending from which explains more the variance of the data to which less explanation is given, multiplied by the original data.

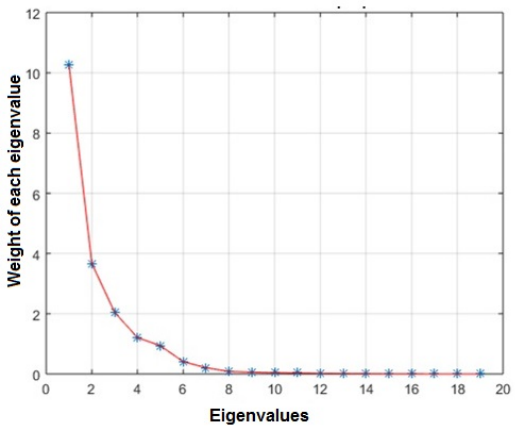

***Figure 4**. Descending order of the eigenvalues*

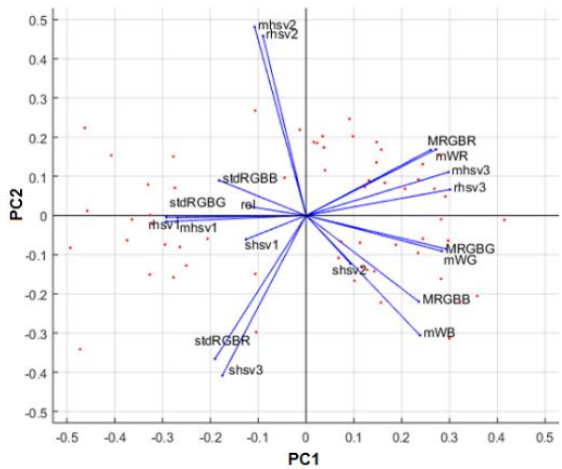

***Figure 5.** Biplot between the 2 main components and the original features with a Level 2 Wavelet.*

The characteristic's value defined by each RGB component (MRGB)'s average, as well as the (mWR, mWG and mWB) Wavelet's mean for the three RGB components is being separated as a higher level of Wavelet decomposition is chosen; this can be observed in Figure 6 and Figure 7. However, increasing the decomposition level reduces the weight on each main component (shsvx and stdRGBx correspond to standard deviations with x = 1,2,3).

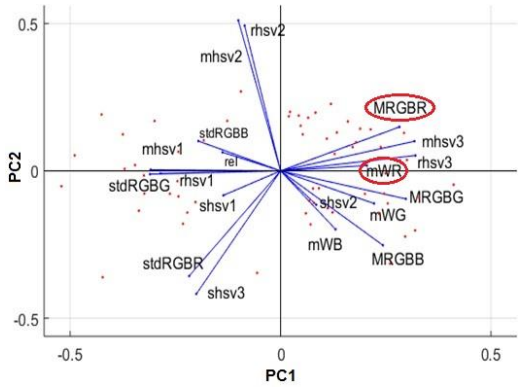

***Figure 7.** Biplot between the 2 Principal components and the original characteristics with wavelet level 4*

At the end, the obtained components generate classes with certain probability distributions (Figure 8) and its projection on the PC1-PC2 axis (Figure 9).

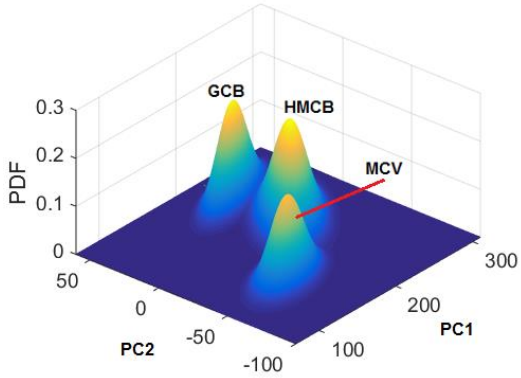

***Figure 6.** Gaussian Distributions for each class.*

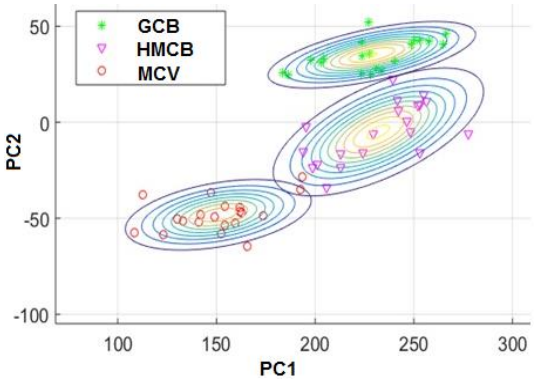

***Figure 7.** Projection of probability distribution functions PDF*

The point cloud of each type of grain related to each of the three main components (Figure 10) presents a major separation in the training grains.

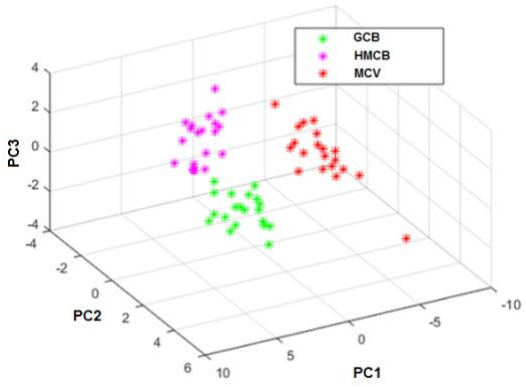

Figure 8. *Point cloud of the three types of grains related to the three main components.*

And those with the worst performance (Figure 11)

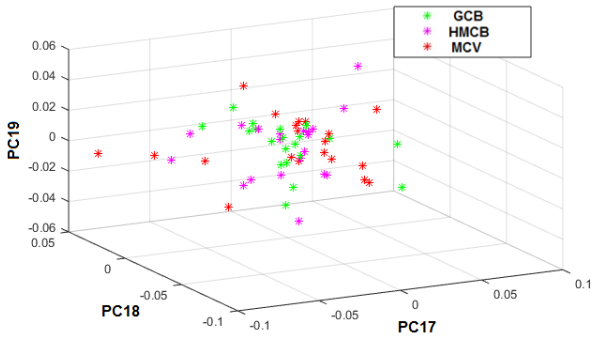

Figure 9. *PC with worse performance*

A summary of the best combinations of two characteristics that generate different scopes of classification can be observed in Figure 12. In the image, the mean prefixes represent an average value; numerals 1 in the HSV transformation correspond to H, numerals 2 correspond to S and numerals 3 correspond to V. The best performance combination (besides those obtained by main components) were the combination of the component obtained from Wavelet component's average for the red component combined with the RMS value of the H component in the HSV transformation and, the average value of the green component with Wavelet transformation.

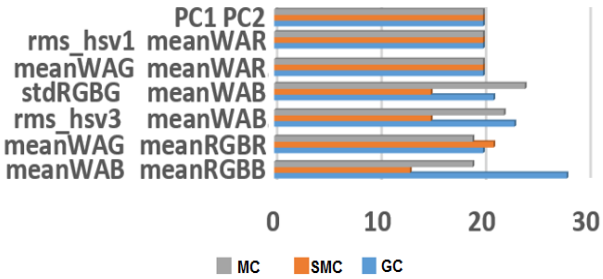

Figure 10. *Classification of MC green grains, SMC semi-mature and mature GC grains.*

4. CONCLUSIONS

The analysis of the biplot, can give an idea of the redundancy of characteristics. The dendrogram of variables allows to observe, part of the relationship between variables as found in biplot. The classifier has good performance, although the characteristics used can present very inherent relationships to each type of grain, as is the case of the R component for the mature grain and G for the green grain. The diameter ratio delivers very poor classification since there is no direct relationship with the degree of ripening of the coffee bean. Only a variation of axes in the green grain and very similar axes ratio for the pinon and mature grains are seen with the naked eye.

REFERENCES

[1] G. I. Puerta Quintero, "Rendimientos y calidad de coffea arabica L., según el desarrollo del fruto y la remoción del mucílago," Cenicafé, vol. 1, no. 61, pp. 67-89, 2010.

[2] G. Puerta Q., "Influencia de los granos de café cosechados verdes, en la calidad física y organoléptica de la bebida," Cenicafé, vol. 2, no. 51, pp. 136-150, 2000.

[3] E. R. Arboleda, . A. C. Fajardo and R. P. Medina, "Classification of Coffee Bean Species Using Image Processing, Artificial Neural Network and KNearest Neighbors," 2018 IEEE International Conference on Innovative Research and Development , pp. 1-5, 2018.

[4] R. H. M. H. J. H. C. P.-Z. C. E. G.-C. J. C. &. B.-C. C. A. Condori, "Automatic classification of physical defects in green coffee beans using CGLCM and SVM," in XL Latin American Computing Conference (CLEI), 2014.

[5] K. Hameed, D. Chai and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques," Image and Vision Computing, vol. 80, pp. 24-44, 2018.

[6] T. Zhang, H.-M. Hu and B. Li, "A Naturalness Preserved Fast Dehazing Algorithm Using HSV Color Space," IEEE access, vol. 6, pp. 10644-10649, 2018.

[7] W. W. S. Z. X. Z. J. &. F. J. Yang, "Greenness identification based on HSV decision tree," Information Processing in Agriculture, vol. 2, no. 3-4, pp. 149-160, 2015.

[8] A. Nazir, R. Ashraf, T. Hamdani and N. Ali, "Content Based Image Retrieval System by using HSV Color Histogram, Discrete Wavelet Transform and Edge Histogram Descriptor," in International Conference on Computing, Mathematics and Engineering Technologies, Taitung, 2018.

[9] C. . Y. Chong , . S. P. Lee and T. C. Ling, "Efficient software clustering technique using an adaptive and preventive dendrogram cutting approach," Information and Software Technology, pp. 1994-2012, 2013.

[10] W. Dong, R. Hu, Z. Chu and J. Zha, "Effect of different drying techniques on bioactive components, fatty acid composition, and volatile profile of robusta coffee beans," Food Chemistry, pp. 121-130, 2017.

[11] K. Hron, M. Jelínková, P. Filzmoser and R. Kre, "Statistical analysis of wines using a robust compositional biplot," Talanta, pp. 46-50, 2012.

[12] A. H. &. T. M. Jahromi, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in Artificial Intelligence and Signal Processing Conference (AISP) , 2017.