Open source iterative bayesian classifier algorithm for quality assessment of processed coffee beans

Fernando Ferreira Lima dos SANTOS1*, Lucas Arthur de Almeida TELLES2, Jorge Tadeu Filho ROSAS1, Amanda Pereira Assis GOMES1, Rodrigo Nogueira MARTINS1, Emanoel Di Tarso dos Santos SOUSA1

1Programa de Pós-Graduação em Engenharia Agrícola (Máquinas e Mecanização Agrícola), Universidade Federal de Viçosa, Viçosa, MG, Brasil. (ORCID: *; 0000-0002-3244-4816; 0000-0002-8912-2690; 0000-0002-0265-0889; 0000-0002-7699-1942; 0000-0002-8290-899X)
2Programa de Pós-Graduação em Ciência Florestal, Universidade Federal de Viçosa, Viçosa, MG, Brasil. (ORCID: 0000-0002-5310-5442)
*E-mail: Fernando.flsantos@gmail.com (ORCID: 0000-0002-0946-6005)

ABSTRACT: The selection of coffee beans plays a key role in the product's final quality. After processing, coffee beans are classified according to their quantity of defects. Traditionally this classification is performed manually, which makes the process laborious and time-consuming. This problem can be solved with digital image processing techniques since defective grains have unique visual characteristics. Considering the difficulty of manual classification of the defects, this study aimed to elaborate a Bayesian classifier algorithm to identify these defects in benighted coffee beans, based on its shape and color. To do so, 630 grains of arabica coffee were used, composing eight images in total. The algorithm aimed to classify four classes, which were: regular beans, normal broken beans, black beans, and black broken beans. In order to evaluate the accuracy of the classifier algorithm, it was calculated the global accuracy and the Kappa coefficient, which allows inferring if the classifier is better than a random classification. It was concluded that the developed algorithm presented a global accuracy of 76% and kappa equals to 0.6. Also, the proposed methodology showed great potential for application in the quality evaluation of other products, whose shape and spectral parameters are relevant in evaluating its quality.

Keywords: coffee beans quality; digital image processing; Jupyter Notebook; supervised classification.

RESUMO: A seleção de grãos de café desempenha um papel fundamental na qualidade final do produto. Após o processamento, os grãos de café são classificados de acordo com a quantidade de defeitos. Tradicionalmente, essa classificação é executada manualmente, o que torna o processo trabalhoso e demorado. Este problema pode ser resolvido com técnicas de processamento digital de imagens, uma vez que os grãos defeituosos possuem características visuais únicas. Considerando a dificuldade de classificação manual dos defeitos, este trabalho teve como objetivo elaborar um algoritmo classificador bayesiano para identificar esses defeitos em grãos de café beneficiados, com base em sua forma e cor. Para tal, foram utilizados 630 grãos de café arábica, somando oito imagens ao todo. O algoritmo objetivou classificar quatro classes, que foram: grãos normais, grãos normais quebrados, grãos pretos e grãos pretos quebrados. Para avaliar a precisão do algoritmo classificador, calculou-se a exatidão global e o coeficiente Kappa, o que permite inferir se o classificador é melhor que uma classificação aleatória. Concluiu-se que o algoritmo desenvolvido apresentou uma precisão global de 76% e kappa igual a 0,6. Além disso, a metodologia proposta mostrou grande potencial para aplicação na avaliação da qualidade de outros produtos, cujos parâmetros de forma e espectrais são relevantes na avaliação de sua qualidade.

Palavras-chave: qualidade de grãos de café; processamento digital de imagens; Jupyter Notebook; classificação supervisionada.

1. INTRODUCTION

Nowadays, Brazil is the largest coffee producer in the world, with about 2.2 million hectares cultivated of Coffea arabica and Coffea canephora species, which the first is cultivated on a larger scale. Among the producer states, the ones that produce the most are Espírito Santo and Minas Gerais, responsible for approximately 75% of the national production (DURAN et al., 2017). According to the International Coffee Organization (ICO, 2018), the Brazilian coffee production estimated for the 2018/2019 harvest is about 56.5 million bags, corresponding to an increase of 25.6% in relation to the 2017/2018 harvest.

Among the stages of coffee processing, bean selection plays a key role in the final quality of the product (PIMENTA
et al., 2018). Traditionally, the separation of defective coffee beans in Brazil is done manually. This manual classification by visual inspection is laborious, time-consuming, inconsistent and even with a certain degree of imprecision since it depends on the judgment of a professional involved in this task (OLIVERI et al., 2019).

Studies have identified that defective coffee beans have a different chemical composition from healthy beans. Depending on the quantity and type of defect, it leads to a significant reduction in the quality of the entire sample (COELHO; PEREIRA, 2002; MORais et al., 2007; PIMENTA et al., 2018). As a consequence, the final value of the product is affected.

The coffee beans classification by color is often used to remove defective beans that were not removed during the processing or coffee peeling. This removal is based on diseases, defects in shape and color. As a result, these physical defects of color, shape, and size are characteristics that help to have criteria for grading the quality of the coffee bean from a small sample (DE OLIVEIRA et al., 2016).

Defective coffee beans can be classified in distinct groups; some of them are black, burnt, green, broken, sour and impurities (extrinsic defects). This classification of the physical defects can indicate coffee beans quality and add value to the product, directly interfering with the producer's profitability (DE OLIVEIRA et al., 2016; SANO et al., 2003).

There are currently a large number of studies about digital image processing techniques and methodologies for classification of nutritional and physical defects of various parts of the plant, such as root, stem, flower, fruit and leaf (BARESEL et al., 2017; LEME et al., 2019; PATEL et al., 2012; VASSALLO-BARCO et al., 2017). However, most of the research is focused on maize, rice and vegetable crops, with a lack of information in coffee production (MAJUMDAR et al., 2014). These studies have sought to reduce the subjectivity resulting from manual classification of diseases and defects in plants (MAJUMDAR et al., 2014).

Considering the difficulty of manual classification of defects in processed coffee beans, this study was aimed at the elaboration of an open source Bayesian algorithm for identification of defects in processed coffee beans, related to the shape and color of the beans.

2. MATERIAL AND METHODS

2.1. Experimental Proceedings

The experiment was conducted at the Laboratory of Artificial Vision Machine Designs of the Department of Agricultural Engineering at the Universidade Federal de Viçosa. For this study, eight images of processed arabica coffee beans were obtained. In total, 630 grains were used.

Image acquisition was performed using a Smartphone Moto G5 model. It has a Sony IMX258 Complementary Metal-Oxide Semiconductor (CMOS) sensor that captures images with a resolution of 13 megapixels. The acquisition of the images was performed in an environment with a fully controlled illumination, for which a totally enclosed room was used to prevent interference of any external illumination. The artificial lighting was provided by a halogen lamp Gx 6.35 (Bi-pin) 300/650W, coupled to a light diffuser.

During the acquisition of the images, the Smartphone device was attached to a metal support, and it was established a vertical distance between the coffee beans sample and the camera lens of 0.255 m. The camera was set to auto focus and auto white balance modes. The bottom surface was covered with an A4 sheet of paper to increase the contrast between the beans and the image background. The scheme of the image acquisition system is shown in Figure 1.

2.2. Analysis and Determination of Coffee Beans Quality

Several authors state that the quality of coffee beans can be determined by evaluating its shape, size, and color (CARRILLO; PEÑALOZA, 2009; FARIDAH et al., 2011; SMRKE et al., 2015). For the present study, two parameters for processed coffee beans were considered: shape descriptors and spectral descriptors. Considering this, the defects were classified according to such parameters into different classes, which were: classes I, II, III and IV (Figure 2).

Class I represents the whole regular beans; class II represents the whole black beans; class III represents the broken regular beans; and class IV represents the broken black beans. Therefore, the developed algorithm aimed to classify coffee beans defects according to their respective class.

2.3. Analysis and Digital Image Processing

The image processing and the classifier development were made using the online open source app Jupyter Notebook. For this application, the OpenCV package was installed. First, it was necessary to segment the coffee beans from the background. This step was performed using the thresholding method, which despite being one of the most
used segmentation methods, this still needs to be followed by another method to improve its accuracy (HE et al., 2016).

Therefore, after converting the RGB image (red, green, blue) into a grayscale image, an inverse binary thresholding iterative code was implemented, followed by a logarithmic transformation or self-scaling (at the discretion of the user) and finally, the noise was removed by a median blur filter (Figure 3b). Then, the coffee beans labeling was performed (Figure 3c). After this process, the shape descriptors of each coffee bean were calculated, which were: area, perimeter, and length. For the spectral descriptors, means of each color component intensity of pixels in each coffee bean original sample image was calculated, and then it was named as Rmean, Gmean, and Bmean.

In possession of the training and validation data, the Naive Bayes classifier algorithm (VENORA et al., 2009), which in machine learning is a probabilistic and supervised classifier was used to perform the data classification.

2.4. Classifier Elaboration

For the data collection and pre-classification stage, an iterative sampling algorithm was developed. In this step, the coffee beans of the original image were manually selected and classified by the user, according to their respective class (I, II, III or IV). In order to ease the operation, as the coffee beans were being selected and pre-classified, markers were added to the image (Figure 4).

After sampling, the information regarding the coffee bean identification number and its respective shape and spectral descriptors information were exported in the xlsx worksheet format. After this process, the user must choose the purpose of the exported file samples, either for the classifier training or validation. For the present study, approximately 2/3 of the data collected were used for training, as proposed by Zhang et al. (2016).

2.5. Statistical Analysis

In order to evaluate the accuracy of the classifier, performance indexes as the global accuracy (GA) and the Kappa coefficient were calculated. The GA of the classifier indicates the percentage of the algorithm correctness in relation to the real class of the coffee beans, chosen in the pre-classification stage. The GA index was calculated using equation 1.

\[
GA = \frac{\sum_{i} x_{ii} \times 100}{N}
\]  
(Eq. 1)

where: \(x_{ii}\) = Number of coffee beans per cell in the confusion matrix; and \(N\) = Total number of coffee beans.

The Kappa coefficient is a concordance coefficient between the evaluation of two classifiers (CONGALTON, 1991). In this case, the two classifiers would be the user and the classification algorithm. The Kappa coefficient was calculated using Equation 2. According to Landis; Koch (1977), the result of this coefficient can be interpreted as presented in Table 1.

\[
\hat{K} = \frac{n \sum_{i=1}^{c} x_{ii} - \sum_{i=1}^{c} (x_{ii} \times x_{ii})}{n^2 - \sum_{i=1}^{c} (x_{ii} \times x_{ii})}
\]
(Eq. 2)

where: \(\hat{K}\) = Kappa coefficient estimation; \(x_{ii}\) = value in line i and column i (diagonal) of the error matrix; \(x_{ii} = \) total in line i; \(x_{ii} = \) total in column i; \(n = \) total number of samples; and \(c = \) total number of zones.

| Kappa | Interpretation |
|-------|----------------|
| < 0   | Poor           |
| 0 – 0.2 | Slight       |
| 0.21 – 0.4 | Fair      |
| 0.41 – 0.6 | Moderate  |
| 0.61 – 0.8 | Substantial |
| 0.81 - 1 | Almost Perfect |

Table 1. Kappa coefficient interpretation

Tabela 1. Interpretação do coeficiente Kappa

Source: Adapted from Landis; Koch (1977)
In order to verify the significance of the Kappa coefficient, the Z test at a significance level of 5% was performed. This coefficient was calculated using Equation 3.

\[
Z = \frac{|\hat{K}_i - \hat{K}_j|}{\sqrt{\text{Vár}(\hat{K}_i) + \text{Vár}(\hat{K}_j)}}
\]  
(Eq. 3)

where \(\hat{K}_{i,j}\) = Kappa’s to be compared, dimensionless; and \(\text{Vár}(\hat{K}_{i,j})\) = variances of Kappa’s under analysis.

In order to analyze the weight of all descriptors in the classification algorithm, the average, coefficient of variation (CV), minimum and maximum values of the spectral and shape descriptors were calculated.

3. RESULTS

The descriptive statistics variables for all the coffee beans is presented in Table 2. As a result, the present algorithm obtained a global accuracy rating of 76% and a kappa coefficient equal to 0.6, which was significant according to the z test (11.95) at a 5% significance level.

Table 2. Descriptive analysis of the coffee beans’ shape and spectral descriptors

| Descriptors       | Mean  | Max. | Min. | CV (%) |
|-------------------|-------|------|------|--------|
| Perimeter (pixel) | 81.93 | 112.81 | 40.63 | 15.95  |
| Area (pixel²)     | 516.46 | 925.00 | 105.50 | 30.64  |
| Length (pixel)    | 22.15 | 30.41 | 8.00  | 20.44  |
| R mean            | 52.05 | 70.67 | 30.69 | 13.13  |
| G mean            | 72.55 | 89.62 | 41.41 | 14.56  |
| B mean            | 96.00 | 116.42 | 51.75 | 14.50  |
| Gray mean         | 157.36 | 193.74 | 89.24 | 14.11  |

With the confusion matrix (Table 3) generated for the supervised classification, it is possible to observe the distribution of the classification performed by the algorithm. For class I, 98% of the observations were correctly identified, and 74% of the observations classified as class I were actually class I; for class II, only 49% of the observations were correctly identified and 78% of the observations classified as class I were actually class II; for class III, 65% of the observations were correctly identified, and 88% of the observations classified as class III were actually class III; for class IV, only 45% of the observations were correctly identified, and 62% of the observations classified as class IV were actually class IV.

Table 3. Confusion matrix of the coffee beans classification

| Class | Total | UA (%) |
|-------|-------|--------|
| I     | II    | III   | IV    |       |
| 84    | 18    | 9     | 3     | 114   | 73.7  |
| 0     | 21    | 3     | 3     | 27    | 77.8  |
| 2     | 1     | 22    | 0     | 25    | 88.0  |
| 0     | 3     | 0     | 5     | 8     | 62.5  |
| 86    | 43    | 143   | 11    | -     | -     |
| CA (%)| 97.7  | 48.8  | 64.7  | 45.4  | -     |

CA: Classifier’s accuracy; UA: User’s accuracy

4. DISCUSSION

Regarding the weight of all descriptors in the classification algorithm, Table 2 points to a significant importance of the variables area and length, as their coefficient of variation values were the highest among the variables. For the classification process, the higher the variation of a descriptor is, the better the classifier can separate it in different classes.

Similar studies in the literature achieved approximated results when compared to the present methodology. Carrillo; Peñazo (2009) elaborated a classifier that used the average color of coffee beans as the input parameter to the classifier and reported as a result 90% accuracy of whole coffee beans defects detection (excluding one of the defects classes that presented low accuracy). Faridah et al. (2011) proposed a methodology based on feature extraction from parameters of coffee beans image samples in terms of texture and color by implementing artificial neural network algorithm. As a result, these authors obtained a global accuracy rating of, approximately, 74%, for seven distinct classes of coffee beans.

Regarding the algorithm’s performance, the best results were obtained for classes I and III (whole regular beans and broken regular beans), which both classifier and user accuracy showed the highest results according to Table 2. On the other hand, the lowest accuracy results were obtained for classes II and IV (whole black beans and broken black beans) (Table 3).

The reasons for this performance could be the paucity of faulty coffee beans in the samples and/or the low number of samples. In addition, possibly, using only three shape descriptors (area, perimeter, and length) was not enough to correctly classify the coffee beans.

Although the results can be considered as moderate according to the Kappa performance (0.6) (Table 1), better results could be obtained by increasing the number of descriptors and coffee beans samples to discriminate between defective and non-defective coffee beans. Also, the use of iterative classifiers and computer vision has proven to be a promising approach for application in the quality evaluation of products, such as cereals and fruits, whose shape and spectral parameters are relevant in the evaluation of their quality (CHAUQUELE; MALI, 2014; LIMING; YANCHAO, 2010; MEBATSION et al., 2013).

Finally, as stated by Bhargava; Bansal (2018), there are still various challenges to be beaten to improve the defects detection, which includes surface assessment, lengthy time exhausting of acquisition and processing for spectral images and different defects discrimination. Also, other descriptors should also be used to improve the classifier’s accuracy.

5. CONCLUSIONS

This study presents an iterative classifier algorithm for quality assessment of coffee beans using shape and spectral descriptors as input data. The algorithm presented a global accuracy of 76%, and kappa equals to 0.6, which indicates that the classification was moderate. The best results were obtained for classes I and III, representing the regular beans and the broken regular beans. On the other hand, the lowest accuracies were obtained for classes II and IV (whole black beans and broken black beans).

Nativa, Sinop, v. 8, n. 1, p. 118-123, jan./fev. 2020.
Finally, as a recommendation for future studies, the developed algorithm can be used for quality analysis of other agricultural products. Also, other geometric shape (diameter and elongation), color (CIE L*a*b* color space), and textural descriptors (contrast, coarseness, roughness, and regularity) should be tested to improve the outcomes.

6. ACKNOWLEDGEMENT

The authors thank the Coordination for Improvement of Higher Personnel Education (CAPES) and the Foundation for Research Support of the State of Minas Gerais (FAPEMIG) for the financial support, and the Agricultural Mechanization Laboratory of the Universidade Federal de Viçosa (UFV).

7. REFERENCES

BARESEL, J. P.; RISCHBECK, P.; HU, Y.; KIPP, S.; BARMEIER, G.; MISTELE, B.; SCHMIDHÄLTER, U. Use of a digital camera as alternative method for non-destructive detection of the leaf chlorophyll content and the nitrogen nutrition status in wheat. Computers and Electronics in Agriculture, New York, v. 140, p. 25-33, 2017. DOI: https://dx.doi.org/10.1016/j.compag.2017.05.032

BHARGAVA, A.; BANSAL, A. Fruits and vegetables quality evaluation using computer vision: A review. Journal of King Saud University - Computer and Information Sciences, p. 1-15, 2018. DOI: https://dx.doi.org/10.1016/j.jskuci.2018.06.002

CARRILLO, E.; PEÑALOZA, A. A. Artificial vision to assure coffee-excelso beans quality. In: Proceedings of the 2009 Euro American Conference on Telematics and Information Systems: New Opportunities to increase Digital Citizenship; 2009. Annals... 2009. DOI: https://dx.doi.org/10.1145/1551722.1551757

CHAGUDE, A.; MALI, S. N. Evaluation of texture and shape features for classification of four paddy varieties. Journal of Engineering, v. 18, p. 1-8, 2014. DOI: http://dx.doi.org/10.1155/2014/617263

COELHO, K. F.; Pereira, R. G. F. A. Influência de grãos defeituosos em algumas características químicas do café cru e torrado. Ciência e Agrotecnologia, Lavras, v. 26, n. 2, p. 375-384, 2002.

CONGALTON, R. G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, Amsterdam, v. 37, n. 1, p. 35-46, 1991. DOI: https://dx.doi.org/10.1016/0034-4257(91)90048-B

DE OLIVEIRA, E. M.; LEME, D. S.; BARBOSA, B. H. G.; RODARTE, M. P.; PEREIRA, R. G. F. A. A computer vision system for coffee beans classification based on computational intelligence techniques. Journal of Food Engineering, Essex, v. 171, p. 22–27, feb. 2016. DOI: https://dx.doi.org/10.1016/j.jfoodeng.2015.10.009

DURÁN, C. A.; TSUKUI, A.; SANTOS, F.; MARTINEZ, S. T.; BIZZO, H. R.; REZENDE, C. M. Café: Aspectos Gerais e seu Aproveitamento para além da Bebida. Revista Virtual de Química, v. 9, n. 1, p.107-134. 2017. DOI: https://dx.doi.org/10.21577/1984-6835.20170010

FARIDAH, F.; PARIKESIT, G. O.; FERDIANJAH, F. Coffee bean grade determination based on image parameter. TELKOMNIKA - Telecommunication Computing Electronics and Control, v. 9, n. 3, p.547-554, 2011. DOI: https://dx.doi.org/10.12928/telkomnika.v9i3.747

HE, H. J; ZHENG, C.; SUN, D. W. Image Segmentation Techniques. In: SUN, D. W. (Ed.). Computer vision technology for food quality evaluation. ACADEMIC PRESS, 2016. p. 45-63.

INTERNATIONAL COFFEE ORGANIZATION. Coffee Market Report – January 2018. Available in: http://www.ico.org/documents/cy2017-18/cmre0118-e.pdf. Access: 20 June 2018.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. Biometrics, Washington, v. 31, n. 1, p. 159-174, mar. 1977. DOI: https://dx.doi.org/10.2307/2529310

LEME, D. S.; DA SILVA, S. A.; BARBOSA, B. H. G.; BOREM, F. M.; PEREIRA, R. G. F. A. Recognition of coffee roasting degree using a computer vision system. Computers and Electronics in Agriculture, New York, v. 156, 312-317. 2019. DOI: 10.1016/j.compag.2018.11.029

LIN, X.; YANCHAO, Z. Automated strawberry grading system based on image processing. Computers and Electronics in Agriculture, New York, v. 71, p. 32-39, 2010. DOI: https://dx.doi.org/10.1016/j.compag.2009.09.013

MAJUMDER, D.; KOLE, D. K.; CHAKRABORTY, A.; MAJUMDER, D. D. Review: Detection & Diagnosis of plant leaf disease using integrated image processing approach. International Journal of Computer Engineering and Applications, v. 6, n. 3, p.1-16, 2014.

MEBATSION, H. K.; PALIWAL, J.; JAYAS, D. S. Automatic classification of non-touching cereal grains in digital images using limited morphological and color features. Computers and electronics in agriculture, New York, v. 90, p. 99-105, jan. 2013. DOI: https://dx.doi.org/10.1016/j.compag.2012.09.007

MORAIS, S. A. L.; AQUINO, F. J. T.; CHANG, R.; NASCIMENTO, E. A.; OLIVEIRA, G. S.; SANTOS, N. C. Análise química de café arábica (Coffea arabica L.) e grãos pretos, verdes e ardidos (PVA) submetidos a diferentes graus de torrefação. Coffee Science, Lavras, v. 2, n. 2, p. 97-111, July-dec. 2007.

OLIVERI, P.; MALEGORi, C.; CASALE, M.; TARTACCIA, E.; SALVATORI, G. An innovative multivariate strategy for hsi-nir images to automatically detect defects in green coffee. Talanta, London, v. 199, n. 1, p. 270-276, 2019. DOI: https://dx.doi.org/10.1016/j.talanta.2019.02.049

PATEL, K. K.; KAR, A.; JHA, S. N.; KHAN, M. A. Machine vision system: a tool for quality inspection of food and agricultural products. Journal of food science and technology, Oxford, v. 49, n.2, p. 123-141, 2012. DOI: https://dx.doi.org/10.1007/s13197-011-0321-4

PIMENTA, C. J.; ANGÉLICO, C. L.; CHALFOUN, S. M. Elaboration of an open source iterative bayesian classifier algorithm for quality assessment of processed coffee grains. Revista Virtual de Química, V. 9, N. 1, P. 54-66, 2018. DOI: https://dx.doi.org/10.5901/rvq2018424000118

SANO, E. E.; ASSAD, E. D.; CUNHA, S. A.; CORREA, T. B.; RODRIGUES, H. R. Quantifying adulteration in roast...
coffee powders by digital image processing. *Journal of food quality*, Waspport, v. 26, n. 2, p.123-134, 2003. DOI: https://dx.doi.org/10.1111/j.1745-4557.2003.tb00232.x

SMRKE, S.; KROSLAKOVA, I.; GLOESS, A. N.; YERETZIAN, C. Differentiation of degrees of ripeness of Catuai and Tipica green coffee by chromatographical and statistical techniques. *Food chemistry*, v. 174, p. 637-642, 2015. DOI: https://dx.doi.org/10.1016/j.foodchem.2014.11.060

VASSALLO-BARCO, M.; VIVES-GARNIQUE, L.; TUESTA-MONTEZA, V.; MEJÍA-CABRERAL, H. I. Automatic Detection of Nutritional Deficiencies in Coffee Tree Leaves Through Shape and Texture Descriptors. *Journal of Digital Information Management*, v. 15, n. 1, p.7-18, feb. 2017.

VENORA, G.; GRILLO, O.; SACCONE, R. Quality assessment of durum wheat storage centres in sicily: Evaluation of vitreous, starchy and shrunken kernels using an image analysis system. *Journal of Cereal Science*, v. 49, n. 3, p. 429-440, 2009. DOI: https://dx.doi.org/10.1016/j.jcs.2008.12.006

ZHANG, B.; ZHANG, L.; XIE, D.; YIN, X.; LIU, C.; LIU, G. Application of synthetic NDVI time series blended from Landsat and MODIS data for grassland biomass estimation. *Remote Sensing*, v. 8, n. 1, p. 1-21, 2016. DOI: https://dx.doi.org/10.3390/rs8010010