ISSN 1807-1929
Revista Brasileira de Engenharia Agrícola e Ambiental
v.23, n.10, p.782-786, 2019
Campina Grande, PB, UAEA/UFCG – http://www.agriambi.com.br

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v23n10p782-786

Proposal of automated computational method to support Virginia tobacco classification

Leonel P. C. Tedesco[1], Adriano da C. de Freitas[1], Rolf F. Molz[1] & Jacques N. C. Schreiber[1]

[1] Universidade de Santa Cruz do Sul/Programa de Pós-Graduação em Sistemas e Processos Industriais. Santa Cruz do Sul, RS, Brasil. E-mail: leoneltedesco@unisc.br (Corresponding author) - ORCID: 0000-0003-3010-8197; adrcostafreitas@gmail.com - ORCID: 0000-0002-6536-4775; rolf@unisc.br - ORCID: 0000-0002-4359-6751; jacques@unisc.br - ORCID: 0000-0002-2847-0130

ABSTRACT: This article proposes an automatic method for classification of cured tobacco leaves. Typically this process is performed manually, allowing the occurrence of human errors. In addition, the existence of an automated comparative procedure, helping to perform the classification, can make this process faster and more transparent. In order to implement the method, non-invasive to the agricultural product, 250 samples of Virginia tobacco digital images in the RGB and HSV color models were analyzed. The validation of the method was carried out using partial least squares (PLS) and artificial neural network (ANN), presenting a qualitative and quantitative analysis of both tools. It has been verified that the PLS can be applied to this method, as it has a shorter computational time, better suiting a real-time process. It can be verified that the ANN obtained better prediction results. Both methods employed had better results when adopting the RGB color model, reaching coefficient of determinations of 68 and 96% for the PLS and ANN methods, respectively.

Key words: image processing, partial least square, artificial neural network

Proposta de método computacional automatizado para apoio à classificação de tabaco Virgínia

RESUMO: Este artigo propõe um método automático para classificação de folhas de tabaco curado. Tipicamente este processo é realizado de modo manual, possibilitando erros humanos. Aliado a isso, a existência de um procedimento comparativo automatizado, auxiliando na realização da classificação, poderá tornar tal processo mais rápido e transparente. Para a implementação do método, não invasivo ao produto agrícola, analisou-se 250 amostras de imagens digitais de tabaco Virginia nos modelos de cores RGB e HSV. A validação do método foi desenvolvida empregando ferramentas de quadrados mínimos parciais (QMP) e rede neural artificial (RNA), apresentando uma análise qualitativa e quantitativa de ambos as ferramentas. Verificou-se que a técnica de QMP pode ser aplicada para este método, pelo fato de possuir um tempo computacional menor, adequando-se melhor a um processo em tempo real. Pode-se constatar que o método por RNA obteve melhores resultados de predição. Ambos os métodos empregados, tiveram melhores resultados adotando o modelo de cor RGB, atingindo coeficientes de determinação de 68 e 96% para o método de QMP e RNA, respectivamente.

Palavras-chave: processamento de imagem, quadrados mínimos parciais, rede neural artificial

Ref. 191847 – Received 20 Feb, 2018 • Accepted 23 Aug, 2019 • Published 28 Aug, 2019

Introduction

The use of data analysis methods coupled with image processing has been the subject of several studies focusing on the extraction of object characteristics. Carlinet & Géraud (2015) present a method of hierarchical construction of representations, applied in the processing of images referring to documents, meteorology, medicine, among others. Ji (2012) proposes a method that improves the robustness of the independent component analysis by adding an outlier rejection rule to solve problems in analysis of images generated by synthetic-aperture radar. Kucheryavskiy (2013) discusses the discrimination of objects in hyperspectral images, through the detection of patterns and performing quantitative evaluations. Zhao et al. (2009) used spectral imaging technology to determine the apple firmness.

In the processing of tobacco, its leaves are cured in an oven with controlled temperature and humidity, resulting in a color with the following characteristic shades (Matei et al., 2014): (i) L (lemon) - leaves with lemon shade, allowing brownish spots that occupy up to 50% of their surface; (ii) O (orange) - leaves of orange shade, allowing brownish spots that occupy up to 50% of their surface; (iii) R (reddish) - leaves in which the light brown to dark color occupies more than 50% of their surface. Evaluation at the time of the purchase of tobacco in the industry establishes the values to be paid for every 15 kg of the product, and color is an important characteristic to be assessed.

This article proposes a tool for classification of Virginia tobacco using techniques that integrate digital image processing (Gonzalez & Woods, 2008) with multivariate analysis (Hair Junior, 2009; Juneau et al., 2015) and data mining (Martínez-Álvarez et al., 2015). Hence, the objective is to support specific processes of classification that occur in tobacco processing industries during the purchasing process.

Material and Methods

Figure 1 illustrates the methodology proposed in this study to support and optimize tobacco classification. Samples of images of Virginia tobacco bales are captured by a digital camera (MAKO-G125C model, Allied Vision manufacturer) installed in the processing. This camera is fixed at a 1.5 m height from the conveyor belt, so as to ensure a constant distance between the camera lens and the product, avoiding disturbances in the capture. From this, a set of computational processes are applied in order to treat the captured images and allow automatic classification of the product. After the classification, stocking procedures begin.

The image-capturing structure remains located before the step of classification by the evaluator, which takes the photograph. The captured images are subsequently stored in bitmap format, with a resolution of 24-bit intensity and dimensions of 500 x 550 pixels, resulting in a size close to 600 kilobytes and approximate measurements of 17 x 19 cm. The environment in which the camera was inserted has lighting controlled at 1600 Lux, for uniform acquisition of digital images, causing the figures to have unchanged luminance levels, besides color temperature of 5000 K, keeping the environment as close as possible to natural light.

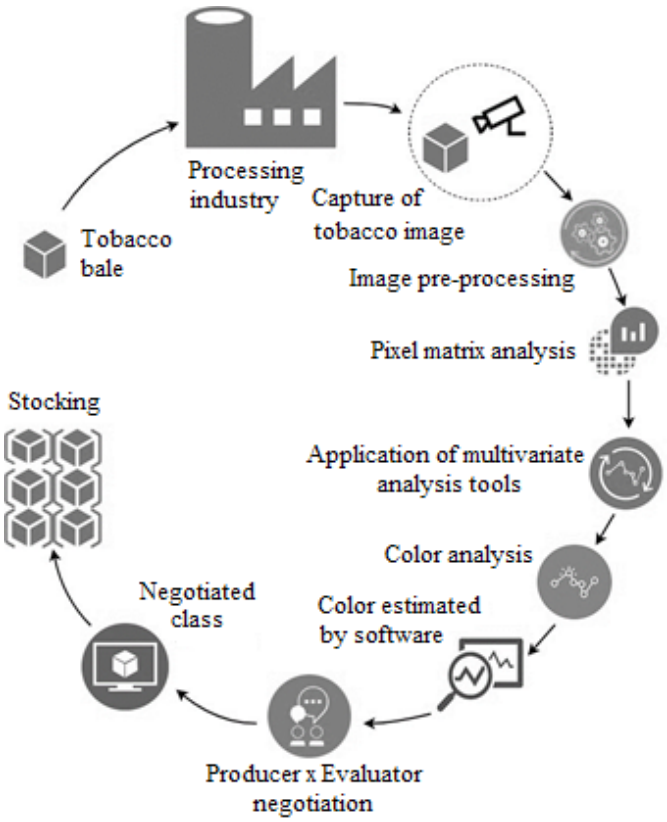


Figure 1. Proposed methodology

After capturing the images, they are subjected to the treatment of extraction of the characteristics according to the RGB (Red Green Blue) and HSV (Hue Saturation Value) standards. In this case, each image was subjected to a central ROI (Region of Interest) of 128 x 128 pixels, with dimensions of 3.4 x 3.8 cm. This size was used to make the processing more efficient, without losing information needed for analysis. Once this treatment was performed, software programs were employed in the classification and quantification of elements from the digital images acquired. A total of 250 images of tobacco were processed.

Subsequently, multivariate analysis and data mining tools were applied to verify possible samples considered as outliers. This verification used the multivariate calibration by the partial least squares (PLS) method (Prats-Montalbán et al., 2011). In this analysis, the samples of the images of tobacco bales were randomly divided into 2/3 for training (process in which samples are related to values indicated by human specialist) and 1/3 for validation (process in which the remaining samples have their values automatically assigned by the system). The division of samples was assisted by the algorithm Kennard-Stone (KS) (Kennard & Stone, 1969), which acts by selecting two samples with the longest Euclidean distance of the coordinates from each other, in a given space. The calculation of this distance for the RGB values extracted in the matrices of the images can be verified as follows (Wehrens, 2011).

$$d(P_1, P_2) = \sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2} \quad (1)$$

where:

d (P_1, P_2) - Euclidian RGB distance between point 1 and point 2;

R_1 - red value of point 1;
R_2 - red value of point 2;
G_1 - green value of point 1;
G_2 - green value of point 2;
B_1 - blue value of point 1; and,
B_2 - blue value of point 2.

For each remaining sample, the minimum Euclidean distance of the selected samples is calculated. After that, the samples with the longest minimum Euclidean distances are withdrawn, and the process is repeated until a certain number of samples are selected, according to the number of validations and training (Ji, 2012). For the HSV conversion process, the calculations are carried out according to the following equations (Helfer et al., 2015):

$$H = \begin{cases} 60 \times \frac{(G-B)}{(M-m)}, \text{ if } M = R \\ 60 \times \frac{(B-R)}{(M-m)} + 120, \text{ if } M = G \\ 60 \times \frac{(R-G)}{(M-m)} + 240, \text{ if } M = B \end{cases} \quad (2)$$

$$S = \begin{cases} \frac{(M-m)}{M}, \text{ if } M \neq 0 \\ 0, \text{ if } M = 0 \end{cases} \quad (3)$$

$$V = M \quad (4)$$

where:
H - (hue) color perception;
M - highest value of RGB color component;
m - lowest value of RGB color component;
S - (saturation) amount of gray in the color; and,
V - (value) color brightness.

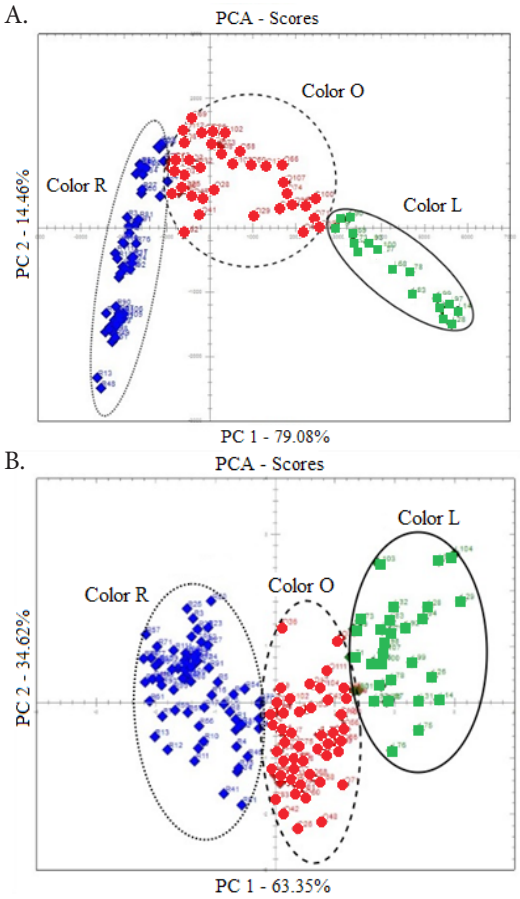
In the equations above, M and m respectively define the highest and lowest values of RGB color components. Subsequently, the results obtained by multivariate analysis were inserted, separated and discriminated. Initially, the digital images were classified according to human selection and then inserted into a database for multivariate analysis, using CHEMOSTAT® software (Helfer et al., 2015). In order to optimize the performance of this computational tool, the selection of Principal Components was performed to filter the samples by removing the outliers, using the Hotelling's T^2 method (Gonçalves et al., 2016). The Artificial Neural Network (ANN) algorithm (Martínez-Álvarez et al., 2015) was adopted for implementing such mining. The same 161 samples filtered initially were inserted in order to be interpreted by WEKA® software (Hall et al., 2009) and, then, the characteristics of the data matrices of the RGB and HSV color models were extracted.

The ANN employs the Multilayer Perceptron model (MLP), which uses the Backpropagation classifier algorithm for training. After defining the ANN and its algorithm, the set of information in the data matrices was processed, and these matrices were discriminated according to the color model. As done in the multivariate analysis, the training data were stipulated in 2/3, followed by 1/3 for validation.

In parallel, data mining was used to perform the analysis of predictions of the digital images of tobacco, which aims to generate values of correlation of the calibration coefficient and validation, root mean square error, among other performance indices, in order to obtain weighted optimization of the results of the proposed method (Mather & Tso, 2016). At the end of the analysis, it is possible to check the class estimated by the software program and compare it with samples classified according to a human expert, supporting the negotiation at the time of purchase.

Results and Discussion

The values to be interpreted by the analysis tools were obtained using the proposed color models. For the RGB model, the first two components (PC1 and PC2) reached 93.54% of the total variance of the samples, and thus the clusters defined by the colors of their classes were generated. In other words, red color was used for "O", green color was used for "L", and blue color was used for "R". For the HSV model, the principal components PC1 and PC2 explain 98.27% of the variance of the samples, and their clusters separated in a similar way to that applied to the RGB model. These results can be seen in Figure 2.



R - Reddish; O - Orange; L - Lemon

Figure 2. Clustering of samples based on first (PC1) and second (PC2) principal components for the RGB (A) and HSV (B) models

According to the discrimination of the data of the samples, it is possible to generate, after analysis of the outliers, well-defined clusters, favoring the data to create the model of multivariate calibration by PLS. Hence, 161 total samples were used in the software program for RGB and HSV analysis, with equivalences of classes around 20% for "L", 36% for "O" and 44% for "R"; these values were generated simultaneously, according to the tobacco purchase demand at the time of the procedure.

After analysis in the PCA model, the multivariate analysis software CHEMOSTAT® is applied to generate the PLS, in which the images of tobacco bales are imported from the database. Then, the matrices with the values of the RGB histogram of each image are extracted. With the original data matrix, the Kennard-Stone algorithm was applied, generating 108 samples for training and 53 samples for validation. Subsequently, the regression method was applied to obtain the 3 color classes.

In order to enable the interpretation of the analysis software, values '1' are assigned to the samples belonging to their respective classified color, and values '0' are assigned to the cases in which the sample does not match its class. As exemplified in Table 1, 6 examples of training samples are presented, from the 108 total.

After inserting all the data for the classification of the training samples, the trained data was analyzed, generating the results for each color model, according to Table 2.

For the RGB color model, it is possible to note the correlations of the colors of trained samples. Coefficients of determination higher than 98% for each color trained, root mean squared error of calibration (RMSEC) lower than 0.06, and root mean squared error of cross-validation (RMSECV) ranging from 0.3894 to 0.1881 were obtained. On the other hand, the training of samples in the HSV model resulted in coefficients of determination of 43, 16.05 and 77.06% for the sequence of colors "L", "O" and "R". The RMSEC showed a variation between 0.4424 and 0.2387, in addition to cross-validation of 0.2894. Hence, the discrepancy between the models trained for the calibrations of the samples is demonstrated. The evaluation of this step is followed by the loading of the prediction data, with the insertion of matrices with 53 samples for validation, with the same treatment of the data of the classes referring to the calibration step, as exemplified in Table 3.

According to the output data of the prediction, it can be seen that the software provides values of coefficients of determination (R^2) for the RGB model higher than 65% for samples of classes "L" and "R", and values of 26.77% for the color class "O". Likewise, the generated values of root mean squared error of prediction (RMSEP) were lower than 0.30 for the classes "L" and "R", along with 0.4612 for the class "O", after 5 s of processing.

For the HSV model, the generated values of prediction correlations were much lower than those of the previous model. For the color "O", the index was only 1.5%, followed by 47.45 and 41.2% for the colors "L" and "R", respectively. Finally, the root mean squared error of prediction varies between 0.4832 and 0.2601, being relatively superior to that of the RGB model. Therefore, comparing the results obtained by the multivariate calibration analysis, according to the indices of tobacco colors, the RGB model proposed better estimates of color trends, according to the samples tested. By analyzing these results, it can be observed that the greatest difficulty lies in the evaluation of the colors "L" and "O" because these are the colors that most influence the tobacco purchase evaluation.

Once the multivariate analysis was performed, the set of images previously used as input was subjected to the analysis based on data mining. This technique aims to serve as a reference for measuring performance for multivariate analysis, especially in terms of precision and execution time.

The configuration of the ANN applied to the RGB model had 2 layers of neurons, whereas 5 layers were required for the HSV model. A relevant fact during the training of the neural networks refers to the execution time. For the RGB color model, the network required processing time of approximately 303 s to elaborate a model, and 30 s for the HSV color model. The results of performance of this method and classes can be seen in Table 4, simultaneously comparing ANN and PLS. However, it should be highlighted that some metrics are not available in the CHEMOSTAT® software.

As observed in Table 4, the method of analysis of the results by neural network was more accurate, obtaining values of R^2 above 96% for the RGB model and above 91% for the HSV technique.

Table 1. Exemplification for training samples

Training samples	Color		
	L	O	R
L14	1	0	0
L15	1	0	0
O12	0	1	0
O13	0	1	0
R4	0	0	1
R5	0	0	1

R - Reddish; O - Orange; L - Lemon; L14, L15, O12, O13, R4 and R5 - Samples 14, 15, 12, 13, 4 and 5, respectively, of the colors lemon, orange and reddish

Table 2. Calibration of training samples

Model	Color	PLS		
		R^2 (%)	RMSEC	RMSECV
RGB	L	98.13	0.0575	0.2577
	O	98.80	0.0609	0.1881
	R	98.53	0.0569	0.3894
HSV	L	43.00	0.2814	0.2894
	O	16.05	0.4424	0.4560
	R	77.06	0.2387	0.2414

R - Reddish; O - Orange; L - Lemon; RMSEC - Root mean square error of calibration; RMSECV - Root mean square error of cross-validation; PLS - Partial least square; RGB - Red green blue model; HSV - Hue saturation value model

Table 3. Coefficients of determination (R^2) and root mean square error of prediction (RMSEP) for predicted samples based on partial least square (PLS)

Model	Color	PLS	
		R^2 (%)	RMSEP
RGB	L	65.22	0.2572
	O	26.77	0.4612
	R	68.10	0.3003
HSV	L	47.45	0.3195
	O	1.50	0.4832
	R	41.20	0.2601

R - Reddish; O - Orange; L - Lemon; RGB - Red green blue model; HSV - Hue saturation value model

Table 4. Performance of classification models

Method	Model	R^2 (%)	RMSE	MAE	Execution time (s)	Kappa
ANN	RGB	96.22	0.1635	0.033	303	0.942
	HSV	91.56	0.2348	0.166	30	0.915
PLS	RGB	68.10	0.2572	-	5	-
	HSV	47.45	0.2601	-	4	-

ANN - Artificial neural network; PLS - Partial least square; RGB - Red green blue model; HSV - Hue saturation value model; RMSE - Root mean squared error; MAE - Mean absolute error

Conclusions

1. Digital image processing coupled with partial least square (PLS) and artificial neural network (ANN) makes it possible to support the tobacco classification process.

2. The application of principal component analysis allows the classification of the images, based on color of the samples.

3. The ANN presents better performance compared to the PLS method, but required longer time for computational processing to generate the results.

4. The use of PLS tool makes it possible to observe the differentiation between the classes of Virginia tobacco color, and the obtained coefficients of determination are above 65% in some colors, as well as for the estimates of prediction errors lower than 0.28, but reaching 0.4612 for the color "O".

5. The ANN method provided more satisfactory results in terms of accuracy in the analysis of the RGB color model, reaching classificatory coefficients of determination around 96% against the maximum level of 68% by the proposed method, and both results are in line with the indications of the trends of the color patterns of the samples indicated by human classifiers.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil - Finance Code 001.

Literature Cited

Carlinet, E.; Géraud, T. MToS: A tree of shapes for multivariate images. IEEE Transactions on Image Processing, v.24, p.5330-5342, 2015. https://doi.org/10.1109/TIP.2015.2480599

Gonçalves, M. I. S.; Vilar, W. T. S.; Medeiros, E. P.; Pontes, M. J. C. An analytical method for determination of quality parameters in cotton plumes by digital image and chemometrics. Computers and Electronics in Agriculture, v.123, p.89-94, 2016. https://doi.org/10.1016/j.compag.2016.02.007

Gonzalez, R. C.; Woods, R. E. Digital image processing. 3.ed. Upper Saddle River: Prentice-Hall, 2008. 793p.

Hair Junior, J. F. Análise multivariada de dados. 6.ed. Porto Alegre: Bookman, 2009. 688p.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: An update. SIGKDD Explorations, v.11, p.10-18, 2009. https://doi.org/10.1145/1656274.1656278

Helfer, G. A.; Böck, F. C.; Marder, L.; Furtado, J. C.; Costa, A. B. da; Ferrão, M. F. Chemostat: Um software gratuito para análise exploratória de dados multivariados. Química Nova, v.38, p.575-579, 2015.

Ji, J. Robust approach to independent component analysis for SAR image analysis. IET Image Processing, v.6, p.284-291, 2012. https://doi.org/10.1049/iet-ipr.2009.0084

Juneau, P.-M.; Garnier, A.; Duchesne, C. The undecimated wavelet transform-multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information. Chemometrics and Intelligent Laboratory Systems, v.142, p.304-318, 2015. https://doi.org/10.1016/j.chemolab.2014.09.007

Kennard, R. W.; Stone, L. A. Computer aided design of experiments. Technometrics, v.11, p.137-148, 1969. https://doi.org/10.1080/00401706.1969.10490666

Kucheryavskiy, S. A new approach for discrimination of objects on hyperspectral images. Chemometrics and Intelligent Laboratory Systems, v.120, p.126-135, 2013. https://doi.org/10.1016/j.chemolab.2012.11.009

Martínez-Álvarez, F.; Troncoso, A.; Asencio-Cortés, G.; Riquelme, J. C. A survey on data mining techniques applied to electricity-related time series forecasting. Energies, v.8, p.13162-13193, 2015. https://doi.org/10.3390/en81112361

Matei, G.; Torcea, M.; Imbrea, F. Quality assessment of tobacco type Virginia F.I. using international rating system. Annals of the University of Craiova-Agriculture, Montanology, Cadastre Series, v.44, p.139-146, 2014.

Mather, P.; Tso, B. Classification methods for remotely sensed data. 2.ed. Boca Raton: CRC Press, 2016. 376p. https://doi.org/10.1201/9781420090741

Prats-Montalbán, J. M.; Juan, A. de; Ferrer, A. Multivariate image analysis: A review with applications. Chemometrics and Intelligent Laboratory Systems, v.107, p.1-23, 2011. https://doi.org/10.1016/j.chemolab.2011.03.002

Wehrens, R. Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences. Berlim: Springer Science & Business Media, 2011. 286p. https://doi.org/10.1007/978-3-642-17841-2

Zhao, J.; Chen, Q.; Vittayapadung, S.; Chaitep, S. Determination of apple firmness using hyperspectral imaging technique and multivariate calibrations. Transactions of the Chinese Society of Agricultural Engineering, v.25, p.226-231, 2009.