Data Article

Image datasets of cocoa beans for taxonomy nuances evaluation

F.A. Santos*, E.S. Palmeira, G.Q. Jesus

Universidade Estadual de Santa Cruz, Brazil

ARTICLE INFO

Article history:
Received 22 July 2019
Received in revised form 29 September 2019
Accepted 7 October 2019
Available online 15 October 2019

Keywords:
Cocoa beans
Cut Test
Object positioning
Taxonomy evaluation

ABSTRACT

There are some classification methods that generate nuances in the final accuracy caused by objects positioning, framing and damage. These occurrences may result in a drop of accuracy in computer vision systems that were trained with structured static datasets and are intended to be used in day-to-day applications in which the images are not always as organized as the trained dataset, like some biometric classification systems such as iris and fingerprint. In this regard, this paper presents six image datasets processed with different methods to help researchers analyze the impact of object positioning, framing and damage in their taxonomies.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Data

Six image datasets of cut-test-classified cocoa beans were created from the research presented in Ref. [3]. Each dataset contains 14 classes (namely: Compartmentalized Brown, Compartmentalized White, Compartmentalized Partially Purple, Compartmentalized Purple, Compartmentalized Slatty, Plated Brown, Plated White, Plated Partially Purple, Plated Purple, Plated Slatty, Moldered, Flattened,

![Fig. 1. A Compartmentalized Brown bean from the source dataset.](image-url)
Brittle and Agglutinated) with 100 images per class, totaling 1400 images per dataset. Fig. 1 presents an image of the source dataset and Fig. 2 presents the six results of the same beans accordingly with the preprocessing methods.

2. Materials and methods

This section brings the explanation about the processing applied to the source image to produce the resulting datasets. Each section of this chapter will present the methods used to reach each of the six preprocessed versions. Those methods were obtained through empirical tests with a fuzzy associative memories implementation sensible to objects positioning, those six datasets versions presented six different accuracies for the same classification method, with a standard deviation of approximately 10.07% between them.

2.1. A-Method: Background Removed

Presented under the name of “background_removed__version_1__method_a.rar” in the repository, five steps (see Fig. 3) were applied to generate the images of this dataset, being:

![Fig. 2. The six versions created from the bean from Fig. 1.](image-url)
Fig. 3. (a) Source; (b) Background removed; (c) Binarized; (d) Preservation of largest body; (e) Application of mask in source image.

Fig. 4. (a) and (b) are images from the "A-Method: Background Removed" and (c) and (d) are their respective cropped images to the minimum fitting rectangle.
i. The CIELAB color space were used to remove the background;
ii. The image was binarized, everything that was not pure RGB black were turned into pure RGB white;
iii. All connect white pixels were labeled as regions and only the largest region of the step (ii) were preserved;
iv. The (iii) result were applied as a mask in the source image.

2.2. A-Method: Only image

To create this dataset all images resulted from the “A-Method: Background Removed” were cropped to the minimum rectangle capable of fitting each bean, resulting in images with varying widths and heights, as shown in Fig. 4. This dataset is under the name “framed_and_centralized_-_version_1_-_method_a.rar” in the repository.

2.3. A-Method: Framed and centralized

All beans from “A-Method: Background Removed” were measured and then centralized in the smallest rectangle capable of fitting all beans, thus all images have the same dimensions: 3011x2851. Two samples of this process can be seen in Fig. 5. This dataset is under the name “framed_and_centralized_-_version_2_-_method_a.rar” in the repository.

Fig. 5. (a) and (b) are images from “A-Method: Background Removed” and (c) and (d) are their respective centralized beans in the created frame.
2.4. B-Method: Background Removed

This dataset (named “background_removed-version_2-method_b.rar” in the repository) was created with additional steps to restore the damage caused to some beans during the background removal, as shown in Fig. 6, being:

i. The CIELAB color space were used to remove the background;
ii. The image was binarized, everything that was not pure RGB black were turned into pure RGB white;
iii. All connect white pixels were labeled as regions and only the largest region of the step (ii) were preserved;
iv. The image was inverted;
v. All connect white pixels were labeled as regions and only the largest region of the step (iv) were preserved;
vii. The (vi) result were applied as a mask in the source image.

Important to perceive that the process to restore damages to the beans also restore parts of the backgrounds that were contained inside them, as shown in Fig. 7, caused by hollow areas (such as broken ones) in the beans.
Fig. 7. Sample of restored background.

Fig. 8. (a) and (b) are images from the “B-Method: Background Removed” and (c) and (d) are their respective cropped images to the minimum fitting rectangle.
2.5. **B-Method: Only beam**

The same process to create “A-Method: Only Image” were applied to the “B-Method: Background Removed” to create this dataset (see Fig. 8). This dataset is under the name “framed_and_centralized_-_version_3_-_method_b.rar” in the repository.

2.6. **B-Method: Framed and centralized**

The same process to create “A-Method: Framed and Centralized” were applied to the “B-Method: Background Removed” images to create this dataset. The rectangle obtained were of the same dimensions as the one from “A-Method: framed and centralized” (3011x2851). Two samples of this process can be seen in Fig. 9. This dataset in under the name “framed_and_centralized_-_version_4_-_method_b.rar” in the repository.

**Acknowledgments**

This research was made in partnership with the Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia (PPGMC) and with the Centro de Inovação do Cacau (CIC). The financial support was provided by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Fundação de Amparo à Pesquisa do Estado da Bahia.
Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Felipe A. Santos, Eduardo S. Palmeira, Gildson Q. Jesus, An image dataset of cut-test-classified cocoa beans, Data in Brief 24 (June 2019) 103916.
[2] Felipe A. Santos, Eduardo S. Palmeira, Gildson Q. Jesus, Color, structural and textural features for the classification of a cocoa beans image dataset using artificial neural network. In: XIV Workshop de Visão Computacional, 2018, Ilhéus, ANAIS WVC 2018 (2018) 80–84.
[3] Felipe A. Santos, Modelagem de Um Sistema de Visão Computacional para a Classificação de Amêndoas de Cacau na Prova de Corte (Master’s thesis), State University of Santa Cruz, Ilhéus, Brazil, 2019.
[4] Felipe A. Santos, “Image Datasets of Cocoa Beans for Taxonomy Nuances Evaluation”, Mendeley Data 2 (2019) https://doi.org/10.17632/pcx7mj68yn.4.