Automatic Grading of Green Intensity in Soybean Seeds

Rafael Namías[1], Carina Gallo[2], Roque M. Craviotto[2], Miriam R. Arango[2], and Pablo M. Granitto[1] *

[1] CIFASIS, French Argentine International Center for Information and Systems Sciences, UPCAM (France) / UNR-CONICET (Argentina)
Bv. 27 de Febrero 210 Bis, 2000, Rosario, Argentina
[2] Estación Experimental Oliveros, Instituto Nacional de Tecnología Agropecuaria
Ruta Nacional 11 km 353, 2206 Oliveros, Santa Fe, Argentina
granitto@cifasis-conicet.gov.ar

Abstract. In this work we introduce a low cost machine vision system for grading problems in agriculture. Instead of a careful evaluation of a given quantity over a reduced number of samples with a high cost dedicated equipment, we propose to measure the quantity with less precision but over a much bigger number of samples. The advantage of our procedure is that very low cost vision equipment can be used in this case. For example, we used a standard flatbed scanner as an integrated illumination plus acquisition hardware. Our system is aimed at the quantification of the amount of chlorophyll present in a production batch of soybean seeds. To this end we arbitrarily divided green seeds in four classes, with a decreasing amount of green pigment in each class. In particular, in this work we evaluate the possibility of an accurate discrimination among the four classes of green seeds using machine vision methods. We show that morphological features have low discrimination capabilities, and that a set of simple features measured over color distributions provides good separation among grades. Also, most errors are assignations to neighbor grades, which have a lower cost in grading. The good results are almost independent from the classifier being use, Random forest or Support Vector Machines with a Gaussian kernel in our case.

Key words: Soybean, Machine Vision, Green seeds, Grading

1 Introduction

Machine vision systems have been used increasingly in the food and agricultural industry over the last decade, mainly for inspection and evaluation purposes. They advantages over the traditional methods are multiple, including velocity, economy, repeatability and objectivity [4]. However, most systems are based on dedicated expensive hardware equipment, including controlled illumination systems, acquisition equipment and appropriate cameras. The relatively high cost of

* Author to whom all correspondence should be addressed.

the required hardware has prevented the extension of machine vision methods to many non-critical applications. Nevertheless, nowadays there is a growing access to simple and very cheap image equipment, like digital still cameras and flatbed scanners. Even if this equipment is not specifically designed for machine vision, its use in this context can provide a very low cost solution to some grading or quality control problems.

Soybean (Glycine max) is a highly relevant crop for agricultural countries like USA, Brazil or Argentina. A well-known problem with this crop, which has gained relevance in the last years, is the occurrence of green seeds. As explained by Cicero el at. [5], the problem is related to unfavorable climatic conditions, mainly drought, that occurs during the final stages of seed maturation. The green color is caused by the presence of chlorophyll, which is not properly degraded during maturation. Chlorophyll's degradation occurs naturally at the final stages of seed development, but this degradation is affected by maturity, drying conditions and climatic conditions.

The presence of green seeds degrades the quality of seed batches in two ways. First, green seeds have their vigor and viability significantly reduced, showing germination percentages below the minimum standard for commercialization in some countries [6]. Second, the presence of chlorophyll has a negative effect on the quality of the extracted oil. Chlorophyll can be extracted with the oil during the grinding, causing oxidation, reducing the shelf life and producing dark-colored oil. Also, high chlorophyll contents require additional bleaching steps in the hydrogenation process [14].

There is a high correlation between the content of chlorophyll and the quantity of green pigments that can be observed in a seed [14]. This suggest the possibility of using still images of small seed batches to estimate to total content of chlorophyll in a production batch, leading to an objective method to grade the quality of the batch in this aspect.

Previous work has been done on this and similar problems using machine vision techniques. Ahmad et al. characterized diverse symptoms of soybean deceases, including immaturity, using RGB images and a linear classifier. They found that immaturity was one of the most difficult deceases to diagnose [1]. Sinnecker et al. [14] evaluated the correlation between chlorophyll content and color in soybeans using a spectrophotometer over milled seeds. Cicero el at [5] used a chlorophyll fluorescence technique to identify seeds with shades of green. Less related, Melendez el al. [10] used direct color measurements, with a dedicated colorimeter, to estimate the correlation between carotenoids content and color in orange juice, and Vollmann et al. [16] compared two method to estimate the chlorophyll content in leaves, using a dedicated colorimeter and a standard digital still camera.

Typical asymptomatic soybean seeds are almost round, compact, and have a smooth seed coat and beige color. The size and shape of the seed may depend on the location of the seed in the pod, but the round compact shape is almost uniform. Opposite, green immature seeds vary not only in the quantity of green pigment (color) but usually also in the shape of the seed and the smoothness of

Table 1. Description and number of samples of the four categories in which the seeds were graded by a human expert.

Class	Number	Description
Type I	68	More than 50% of green pigmentation
Type II	155	Between 25% and 50% of green pigmentation
Type III	178	Less than 25% of green pigmentation
Type IV	202	No green pigmentation detected in the seed

the coat [1]. This fact suggests that, in addition to color measurements, information about the shape of the seed (morphological features) can be of help in the identification of green seeds.

This work is part of a research project aimed at the possibility of using a very low cost equipment to develop a machine vision system capable of estimating the content of chlorophyll in a soybean production batch. Instead of evaluating the content of the batch by careful measurements over a reduced number of seeds, we propose to estimate it by an average, over a much higher number of seed, of a discrete estimation of the content of each seed. To this end we arbitrarily divided green seeds in four classes, with a decreasing amount of green pigment in each class. In particular, in this work we evaluate the possibility of an accurate discrimination among the four classes of green seeds using a simple flatbed scanner and standard machine vision methods. The use of a flatbed scanner provides a low cost solution with controlled illumination and high portability and reproducibility.

The rest of this paper is organized as follows. In the next section we describe the methods used for image acquisition, segmentation, feature extraction and classification. In Section III we show and analyze our results and finally we close the work with some conclusions in Section IV.

2 Materials and methods

A sample with 603 soybean seeds with a variable amount of green pigmentation were provided by experts from the Estación Experimental Oliveros of the Instituto Nacional de Tecnología Agropecuaria (INTA). All seeds come from the same variety of soybean, collected in a uniform way at different locations near the named institution. A human expert from the same institution graded the seeds according to four classes depicted in Table 1. The graded was based only in visual information. This procedure, based on a human criteria, is prone to errors. A preliminary estimation, based on a single replication of the grading, indicated a potential error of around a 25%, uniformly distributed among neighbor grades. Figure 1 show some examples of each class.

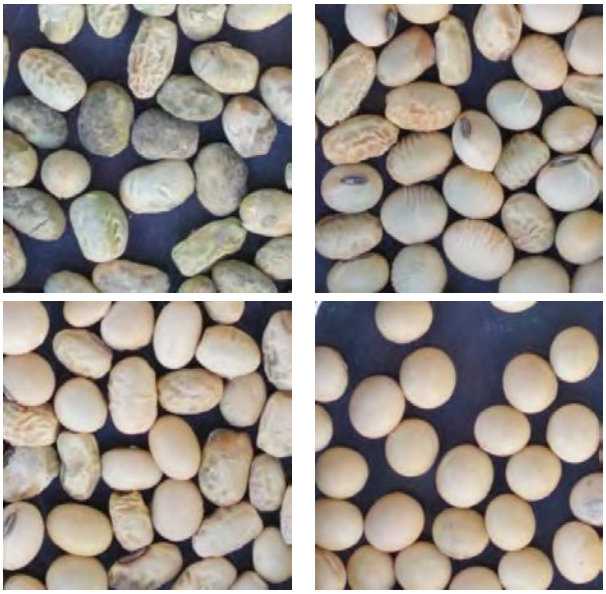

Fig. 1. Samples of the four classes in which the seeds were graded. Top row, on the left Type I and on the right Type II. Bottom row, on the left Type III and on the right Type IV.

2.1 Image Acquisition

Still images of small batches of seeds were taken using a flatbed scanner (Epson Stylus CX 5600). The standard image acquisition software provided with the scanner was used, without any automatic correction of color or bright. A fixed resolution of 600dpi was used. The white foreground of the scanner was replaced with a black one, in order to eliminate reflections and to make the segmentation process easier.

2.2 Segmentation

All image processing was made using open source software, in particular the openCV library [2]. Segmentation was performed as a three steps procedure, starting from the captured RGB image.

The first step is a thresholding of the color image, in order to separate foreground (the seeds) from the background. We used the two more relevant channels, Red and Green, to separate beige and green colors from the background. The optimal threshold for each channel was selected with the "Optimal threshold selection", Sonka et al. [15], page 181, which looks for the minimum density between two normal distributions corresponding to foreground and background pixels. All pixels above any of the thresholds is considered as foreground.

The second step is aimed to identify each individual seed. The binary mask produced at the first step usually has compact regions formed by more than one

seed, which we separate into its individual components using a simple erosion procedure. Finally, we apply a blob extraction procedure with the cvblob library [8].

The last step consists in detecting, as best as possible, the pixels corresponding to each seed. Starting from the erosioned mask and the number of blobs, we apply Meyer's WaterShed method [11] which returns all the regions of interest, i.e., the segmented individual seeds.

2.3 Feature Extraction

As explained in the Introduction, previous results suggest that morphological features can play a relevant role in the grading of green seeds. According to this, we extracted from each seed a set of 10 morphological features that characterize the size and shape of each seed:

1. Perimeter
2. Area
3. Elongation
4. Compactness
5. Roughness
6. Hull Area
7. Hull Perimeter
8. Major Axis
9. Minor Axis
10. Major/Minor axis ratio.

A definition of each feature can be found in the literature [13,9]. All features were converted from pixels to millimeters units using the corresponding resolution.

Color features are expected to be of high importance for the grading of soybean seeds. Even if flatbed scanners produce a highly controlled illumination, it is important to use color measures that are independent of the total illumination of each pixel. A simple method in this direction is to use the rgI color space instead of the original RGB. In rgI space, $I = R + G + B$, $r = R/(R + G + B)$ and $g = G/(R + G + B)$. Channels r and g are more related to the color of the pixel than to the intensity of light. Other standard procedure is to use the HSL color space, in which the hue (H) and saturation (S) are independent of the pixel level (L) [12].

In total we extracted 28 color features. For each seed and each of the four r, g, H and S channels, we extracted the corresponding distribution of observed values, from which we measured some basic statistics that are good descriptors of the distribution:

1. Mean
2. Variance
3. Covariance
4. Kurtosis
5. Skewness
6. Third quartile
7. Interquartile range

Table 2. Morphology features: Confusion matrices for the four types of soybean seeds using a Random Forest (RF) and a Support Vector Machine (SVM) Classifier. Rows are the true classes, columns are the classes assigned by each method.

RF	I	II	III	IV
I	0.16	0.32	0.37	0.15
II	0.05	0.34	0.52	0.09
III	0.07	0.31	0.55	0.07
IV	0.03	0.04	0.06	0.87

SVM	I	II	III	IV
I	0.00	0.24	0.53	0.24
II	0.04	0.17	0.66	0.12
III	0.00	0.09	0.83	0.07
IV	0.01	0.03	0.07	0.89

2.4 Classification

In this work we use two types of classifiers. On one side we utilize the well-known Support Vector Machine (SVM) [7], which nd the maximum-margin separating hyperplane between datapoints of different classes. We use a Gaussian kernel, which adds nonlinear capacity to the classifier. SVM is widely recognized as one of the best methods for classification. The SVM has two free parameters, one is the constant C that controls the margin and the second is the kernel constant γ that regulates the scale of the kernel. In both cases we set the constant selecting from a grid of values using an internal cross validation loop over training data.

On the other side we applied the Random Forest (RF) [3] classifier. RF is a recent ensemble algorithm where the individual classifiers are a set of de-correlated trees. They perform comparably well to other state of the art classifiers and are also very fast. The method has two parameters that control its performance, the total number of trees (that was set to the default value of 500) and the number of variables randomly sampled as candidates at each split, which was also set to its default value of the square root of the total number of variables.

In order to evaluate the classifiers in diverse situations we used a 10-fold cross validation procedure. As results we estimated mean accuracies (with the corresponding standard errors) and confusion tables [17] for both classification methods.

3 Results

After segmentation and feature extraction, we produced a dataset with 38 features (10 morphological, 28 color) measured over 603 soybean seeds.

We analyzed first the grading capability of morphological features. Table 2 show the corresponding results for both RF and SVM classifiers. Each row of the table shows how the classifier assigned the seeds of that class (in proportion, i.e., each row sums to one). As can be seen in the table, morphological features are only able to discriminate Type IV seeds (no green) from the other three classes. Types I, II and III are highly confused for both classification methods. The average accuracy is (0.56 ± 0.02) for RF and (0.59 ± 0.02) for SVM.

Table 3 show the corresponding results using only the 28 features with color information. Overall, these classifiers are clearly more accurate then previous

Table 3. Color features: Confusion matrices for the four types of soybean seeds using a Random Forest (RF) and a Support Vector Machine (SVM) Classifier. Rows are the true classes, columns are the classes assigned by each method.

RF	I	II	III	IV
I	0.71	0.25	0.04	0.00
II	0.11	0.66	0.22	0.01
III	0.01	0.22	0.72	0.05
IV	0.00	0.00	0.04	0.96

SVM	I	II	III	IV
I	0.62	0.32	0.06	0.00
II	0.10	0.70	0.19	0.01
III	0.01	0.24	0.70	0.06
IV	0.00	0.00	0.05	0.95

Table 4. Both morphology and color features: Confusion matrices for the four types of soybean seeds using a Random Forest (RF) and a Support Vector Machine (SVM) Classifier. Rows are the true classes, columns are the classes assigned by each method.

RF	I	II	III	IV
I	0.65	0.31	0.04	0.00
II	0.10	0.65	0.24	0.01
III	0.01	0.22	0.72	0.05
IV	0.00	0.00	0.05	0.95

SVM	I	II	III	IV
I	0.65	0.31	0.04	0.00
II	0.10	0.69	0.19	0.01
III	0.01	0.23	0.71	0.05
IV	0.00	0.01	0.05	0.94

ones, with an average accuracy of (0.78 ± 0.02) for RF and (0.77 ± 0.02) for SVM. Type IV seeds are again more easily identified than the other classes, with a high accuracy (over 0.95). There is more confusion among the other three classes, but correct identification is over 70% in most cases. A very important result is that using color features almost all errors take place between neighbor grades, which have a lower cost in grading problems. For example, for SVM only 1% of Type II seeds are assigned to Type IV, where 29% are assigned to Type I or Type III. As in the previous experiment, both RF and SVM classifiers showed the same accuracy, with only small differences among some of the classes.

Finally, in the last experiment we used both morphological and color features together. The results are shown in Table 4. The average accuracy is the same as when using color features only, (0.77 ± 0.02) for RF and (0.78 ± 0.02) for SVM, and overall the results are very similar to Table 3. This result suggest that morphological features are not relevant when color information is present.

4 Conclusions

In this work we proposed a new strategy for some grading problems in agriculture. Instead of a careful evaluation of a given quantity over a reduced number of samples with a dedicated equipment, we propose to measure the quantity with less precision but over a much bigger number of samples. The advantage of our procedure is that very low cost vision equipment can be used in this case. For example, we used a standard flatbed scanner as in integrated illumination plus acquisition hardware.

In particular we analyzed the problem of grading the content of green pigment in soybean seeds. We showed that morphological features have low discrimination capabilities, and that a set of simple features measured over color distributions

provides good separation among grades. Also, most errors are assignations to neighbor grades, which have a lower cost in grading. The results were almost independent from the classifier being use, Random Forest or Support Vector Machines with a Gaussian kernel.

These results show that our low cost system could grade the seeds with an error level similar to a human expert. This is a first step towards the development of a complete system able to quantify the content of chlorophyll in production batches. The same system could be used to estimate the quality of a sowing batch. In both cases, what is needed is an accurate quantification of the property of interest (chlorophyll content or germination power) over samples of each grade.

Further work is needed also to verify the portability of the method to other flatbed scanners and to estimate better the error level of human experts.

References

1. I.S. Ahmad, J.F. Reid, M.R. Paulsen, and J.B. Sinclair. Color classifier for symptomatic soybean seeds using image processing. *Plant disease*, 83(4):320–327, 1999.
2. Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, Cambridge, MA, 2008.
3. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
4. T. Brosnan and D.W. Sun. Inspection and grading of agricultural and food products by computer vision systems–a review. *Computers and Electronics in Agriculture*, 36(2-3):193–213, 2002.
5. S.M. Cicero, R. Van Der Schoor, and H. Jalink. Use of chlorophyll fluorescence sorting to improve soybean seed quality. *Revista Brasileira de Sementes*, 31(4):145–151, 2009.
6. N.P. Costa, J.B. Franca-Neto, J.E. Pereira, C.M. Mesquita, F.C. Krzyzanowsky, and A.A. Henning. Efeito da sementes verdes na qualidade fisiologica de sementes de soja. *Revista Brasileira de Sementes*, 23(2):102–107, 2001.
7. N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.
8. cvBlobsLib Visual Library. http://opencv.willowgarage.com/wiki/cvblobslib. accesed march 2012.
9. M.L. Hentschel and N.W. Page. Selection of descriptors for particle shape characterization. *Particle & Particle Systems Characterization*, 20(1):25–38, 2003.
10. A.J. Melendez-Martinez, I.M. Vicario, and F.J. Heredia. Application of tristimulus colorimetry to estimate the carotenoids content in ultra frozen orange juices. *Journal of agricultural and food chemistry*, 51(25):7266–7270, 2003.
11. F. Meyer. Color image segmentation. In *International Conference on Image Processing and its Applications*, pages 303–306, 1992.
12. S.E. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999.
13. P.M. Pieczywek and A. Zdunek. Automatic classification of cells and intercellular spaces of apple tissue. *Computers and Electronics in Agriculture*, 81:72–78, 2012.
14. P. Sinnecker, M.S.O. Gomes, AG José, and U.M. Lanfer-Marquez. Relationship between color (instrumental and visual) and chlorophyll contents in soybean seeds during ripening. *Journal of agricultural and food chemistry*, 50(14):3961–3966, 2002.

15. M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision, third edition*. PWS publishing Pacific Grove, CA, 2008.
16. J. Vollmann, H. Walter, T. Sato, and P. Schweiger. Digital image analysis and chlorophyll metering for phenotyping the effects of nodulation in soybean. *Computers and Electronics in Agriculture*, 75(1):190–195, 2011.
17. I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.