Automatic Identification of Weed Seeds

Pablo M Granitto, Pablo F Verdes and H A Ceccatto

Instituto de Física Rosario, CONICET and Universidad Nacional de Rosario,
Boulevard 27 de Febrero 210 Bis, 2000 Rosario, Argentina
{Granitto, Verdes, Ceccatto}@ifir.edu.ar

Abstract. We explore the feasibility of implementing fast and reliable computer-based systems for the automatic identification of weed seeds from color and black and white images. Seeds size, shape, color and texture characteristics are obtained by standard image-processing techniques, and their discriminating power as classification features is assessed. These investigations are performed on a database much larger than those used in previous studies, containing 10,310 images of 236 different weed species. We consider the implementation of a simple Bayesian approach (naïve Bayes classifier) and (single and bagged) artificial neural networks for seed identification. Our results indicate that the naïve Bayes classifier based on an adequately selected set of classification features has an excellent performance, competitive with that of the comparatively more sophisticated neural network approach. In addition, we discuss the possibility of using only morphological and textural characteristics as classification features, which would reduce the operational complexity and hardware cost of a commercial system since they can be obtained from black and white images. According to our results, under particular operational conditions this would result in a relatively small loss in performance when compared to the implementation based on color images.

1 Introduction

The process of manual identification of seeds by specialized technicians is slow, has low reproducibility, and possesses a degree of subjectivity hard to quantify, both in its commercial as well as in its technological implications. It is then of major technical and economical importance to implement computer-based methods for reliable and fast identification and classification of seeds. Automatic systems can be based on seed images, from which classification features associated to seed size, shape, color and texture (i.e., greytone variations on the surface) are readily obtained. For this task, numerous image-processing algorithms are available, which complemented with classification methods make the field of machine vision suitable for seed identification.

Most previous attempts to identify seeds by machine vision have concentrated on cultivated varieties[1-10], with image analysis essentially restricted to basic geometrical measurements (shape factor, aspect ratio, length/area, etc.). More recently, color images were also successfully used to establish seed quality and characterize damages and diseases[11,12].

Besides varietal identification and cereal grain grading, early identification of weeds from the analysis of strange seeds is also of major interest in the agricultural industry. This can be done for the purpose of chemically controlling weed growth or, as occurs in many countries, it can be routinely performed as part of official requirements before a seed batch can be made commercially available (purity analysis). Automatic identification of seeds of wild species is different from the identification of seeds of varieties of a single species. To be approved as a variety, the cultivated plants have to be homogeneous with respect to certain plant characters. Wild species, on the contrary, tend to have larger intra-species variations. Moreover, the variation between weed species will be in general larger, but seeds of some closely related species can be very similar. From the color point of view, most weed seeds are light to dark brownish or black. All these characteristics make the automatic identification of weed seeds *a priori* a difficult classification problem.

An early attempt to identify weed seeds[13] showed the importance of using color instead of black and white images to improve classification accuracy. More recently, Chtioui *et al.*[14] compared the capabilities of linear discriminant analysis and artificial neural networks (ANNs) to identify weed seeds from morphological and textural parameters. However, these investigations considered only four different species, which does not provide a good characterization of inter-species seed variations. In a previous work we have assessed the discriminating power of different seed characteristics for the unique identification of seeds of weed species[15]. We used a simple Bayesian approach (naïve Bayes classifier) to evaluate morphological, color and textural characteristics measured from video images, establishing their importance as classification features for weed seeds identification. In addition, we presented classification results obtained using the same feature set as input of a committee of ANNs. These studies were conducted on a much larger basis than previous ones[13,14], including seed images of 57 frequent weeds found in Argentina's commercial seed production industry. In particular, the species are those listed by Argentina's Secretary of Agriculture as prohibited and primary- and secondary-tolerated weeds.

Argentina's law regulations require the analysis by registered laboratories of a small sample before a seed batch can be made commercially available. In these analyses, commercial and strange seeds present are separated, and the latter ones identified one by one. The studies in [15] were part of a development to avoid the continuous training of new technicians to perform this task, providing an automatic classifier that could be used by less skilled operators. In spite of the good performances of the classifiers developed in this previous work, the number of species considered was still too small to draw definite conclusions on the viability of our approach. Here we present a more extensive investigation of this problem by considering a much larger database, consisting of 10,310 seed images of 236 common weed species. Although this number of species is not yet enough for a commercial system, which would probably require at least twice this capability, it is certainly large enough to reassess the feasibility and limitations of the proposed development.

This work is organized as follows. First, we give, for completeness, a brief description of the hardware used to capture the seed images and list the more relevant morphological, color and textural parameters for identification purposes. More details on image acquisition and a discussion on how these parameters were selected are

given in [15]. We present next the results obtained with the naïve Bayes and ANN classifiers and compare their performances. In addition, we investigate their capabilities for seed identification without using color features, a possibility of interest since black and white images are easier to process and the required hardware is much cheaper. Finally, in the last section we summarize our work and draw some conclusions.

2 Image acquisition and Classification Features

2.1 Hardware

The database contains 10,310 images of 236 different species (a list of these species is available on request). Images with 768×512 pixel resolution were obtained using a 2/3" CCD video camera (XC-711P, Sony Corp, Japan), connected to a color frame grabber (IC-PCI, Imaging Technology Inc, USA) with a 8-bit look-up table for each red (R), green (G) and blue (B) channel. Illumination was provided by a 150W light source (Fostec Inc, USA) with a standard 20V-150W halogen projector lamp (Ushio Inc, Japan)), through a quadruple fiber optic bundle of 12.7mm diameter. All images were taken to approximately fill the camera field of view by adjusting a 6.5X parfocal zoom (6000 System, Navitar Inc, USA) with 0.5X and 2X lens attachments. This is necessary given the large differences in seed sizes considered –from 0.2 to 15mm, approximately– since, otherwise, the images of the smallest seeds would have shown very little texture details.

2.2 Features

We initially measured 75 morphological, color and textural features from the raw seed images to be later used for classification purposes[15]. To choose the best features in each group (those with the largest discriminating power), we implemented standard sequential forward and backward selection algorithms[16], using the performance of a naïve Bayes classifier as the selection criterion. This selection reduced the parameters to nearly optimal sets of 10 morphological, 7 color and 7 textural features. The same procedure applied to these 24 remaining parameters selected 12 (6 morphological, 4 color and 2 textural) features, which were finally used to build the classifiers. A list of these final parameters is given below.

Morphology and size (see Fig. 1)
- Square root of seed area [SQRT(A)]
- Ratio of semi-axis lengths of the main principal axis [h_1 / h_2]
- Ratio of seed and enclosing box areas [$A / (h_1+h_2) \times (v_1+v_2)$]
- Moments of the planar mass distribution with respect to the principal axes [M_{20}, M_{21}, M_{22}]

Color

- Variance of the intensity histogram $[M_2(I)]$
- Skewness of the intensity histogram $[M_3(I) / M_2(I)^{3/2}]$
- Ratios of average pixel values in RGB channels $[E(R)/E(I), E(G)/E(I)]$

Texture

- Contrast[17] along the main principal axis direction
- Cluster Prominence[18] along the secondary principal axis direction
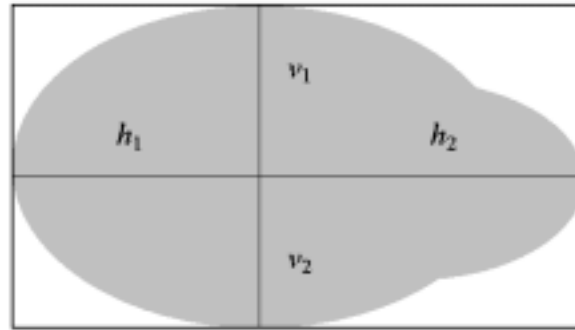
Fig. 1. Definition of quantities related to seed shape used to compute morphological features

The moments of the planar mass distribution with respect to the principal axes above defined are given by $M_{nm} = \sum_{xy} x^n y^m \delta_{xy} / \sum_{xy} \delta_{xy}$. Here x,y are the pixel coordinates with respect to the principal axes and with origin at the image center of mass, and δ_{xy} is equal to 1 if the pixel belongs to the image and 0 otherwise. Furthermore, all morphological quantities were made dimensionless by conveniently normalizing them by the required powers of the square root of the seed area (which was taken as the only dimensional quantity). Notice that since we use the principal axes as a reference frame for all measurements, the resulting values are independent of image orientation. For color features, we considered the intensity $I = (R+G+B)/3$, and E[.] in the above definitions means mean pixel value. Texture parameters were obtained from gray level co-occurrence and gray-level run-length matrices, whose precise definitions can be found in [17] and [18].

3 Classification Results

3.1 Color Images

Following the previous experience in [15], we implemented a simple Bayesian approach to the classification problem (naïve Bayes classifier) and compared its performance with that of the more sophisticated ANN technique (see [19] for an introduction to both methodologies). The naïve Bayes classifier fits the class conditional probabilities with a product of normal distributions of the individual features, considered as independent classification parameters. For the ANN approach we trained feedforward networks with 12 input, h hidden, and 236 output units. The numbers of input and output units correspond, respectively, to the number of parameters used and seed species to be identified. The number of hidden units was varied from $h=20$ to $h=100$, monitoring the performance on cross-validation samples set aside from the training data. The results presented below correspond to $h=80$ units, which lead to the smallest classification error on these samples. We employed output units with softmax (normalized exponential) activation functions to allow the interpretation of outputs as class probabilities. Furthermore, a cross-entropy error measure was used, which is the standard choice for classification problems.

Both for the Bayesian and ANN approaches we split the 10,310 images of the 236 species considered in training and test sets. For this, we randomly chose, for each species, 80% of the images to build the classifier and the remaining 20% to test it. This leaves a large database with 8,250 images for training and also a fairly large test set with 2,060 images. The ANNs were trained with the usual backpropagation rule until convergence, since only negligible overfitting problems were observed. This avoided the use of part of the training set for validation purposes (except for the initial selection of the optimal number of hidden units). In [15] we found that there is not much gain over the performance of a single ANN by structuring several of them in a simple committee. Here we explored the slightly more sophisticated "bagging" approach[20] to build a composite ANN classifier. All the results quoted below correspond to an average over 30 independent experiments.

Table 1 gives average performances and standard deviations for both the training and test sets. It also shows how performance increases when the system assigns a test image to any of the n most probable classes, starting from $n=1$ (standard classification) to $n=5$. That is, for $n > 1$ the classification is considered as correct if the test image corresponds to any of the n classes with the largest probabilities output by the classifier. This possibility is very useful in practice, since untrained operators can easily select the correct option by simple visual comparison with stored representative seed images of the n classes suggested by the classifier. We give the results of a single (generic) classifier and those obtained by ensembling 100 classifiers according to the bagging technique. In this last case, the classifiers were trained on bootstrap re-samples of the training set, which gives them some diversity. In particular, the class probabilities output by the 100 ensemble members were added and the image was assigned to the class with the largest sum value. Notice that, unlike

ANNs, the naïve Bayes (NB) classifier is a stable algorithm, so that its diversity comes only from the bootstrap re-samples of the learning set.

Table 1. Performances of different classifiers as percentage of correct seed identifications using the optimal set of 12 features. Mean values and standard deviations are estimated from 30 independent experiments, as described in the main text

Classifier		$n=1$		$n=2$	$n=3$	$n=4$	$n=5$
		Training	Test	Test	Test	Test	Test
N B	Single	95.0 ± 0.1	92.4 ± 0.4	97.0 ± 0.3	98.4 ± 0.2	98.9 ± 0.1	99.1 ± 0.1
	Bagging	95.2 ± 0.1	92.6 ± 0.4	97.3 ± 0.2	98.6 ± 0.1	99.0 ± 0.1	99.2 ± 0.1
A N N	Single	99.9 ± 0.1	92.5 ± 0.4	96.9 ± 0.3	98.2 ± 0.2	98.7 ± 0.2	99.1 ± 0.2
	Bagging	100	93.0 ± 0.4	97.4 ± 0.3	98.5 ± 0.1	99.0 ± 0.1	99.3 ± 0.1

We stress the excellent performance of the naïve Bayes classifier, which might be related to an effective near independence of the selected classification parameters. This point had already been remarked in [15]. However, there we (wrongly) speculated that for a much larger number of species the classification problem would be more demanding and a ANN ensemble might have an advantage over simpler methods. We also see that the performance of single classifiers leaves not much room for improvement by bagging them (the small differences between single and bagged algorithms are, however, statistically significant). Finally, notice that, for simplicity, feature selection was performed using the naïve Bayes classifier, which may not necessarily produce an optimal set for the ANN approach.

From Table 1 we see that, for the 236 different weed species considered, when the bagged classifiers are allowed to suggest four options for class membership they reach a performance of 99%. For comparison, in our previous work this performance was obtained with $n=3$ for only 57 different species in the database. Finally, we stress the fact that different realizations of training and test sets do not substantially change performances, as indicated by the low standard deviations observed in the 30 independent runs.

3.2 Black and White Images

In general the largest discriminating power corresponds to morphological features. This was established in [15], where it was also shown that, as expected, color features are not particularly good as classification parameters since many species are light to dark brownish or black. On the other hand, texture characteristics are even less reliable than color ones for classification purposes. Furthermore, if one combines any two sets of features, it was found that morphology plus color features have an edge over the combined use of morphology and texture characteristics. However, in this last case it would be enough to consider black and white images, which constitutes an

important simplification in system's operation and leads to a reduction in cost. In fact, color images require a much better control of illumination conditions than black and white ones, and the required acquisition hardware (RGB camera, frame grabber, etc.) is substantially more expensive.

Table 2. . Performances of different classifiers as percentage of correct seed identifications using the 10 features that can be obtained from black and white images. Mean values and standard deviations are estimated from 30 independent experiments, as described in the main text

	B&W 10 features	$n=1$ Training	Test	$n=2$ Test	$n=3$ Test	$n=4$ Test	$n=5$ Test
N B	Single	89.0 ± 0.3	85.2 ± 0.7	93.4 ± 0.6	96.0 ± 0.5	97.4 ± 0.4	98.1 ± 0.3
	Bagging	89.5 ± 0.2	85.5 ± 0.6	93.8 ± 0.7	96.5 ± 0.4	97.8 ± 0.3	98.4 ± 0.3
A N N	Single	98.9 ± 0.4	85.9 ± 0.8	93.3 ± 0.6	95.8 ± 0.5	97.1 ± 0.5	97.9 ± 0.4
	Bagging	99.6 ± 0.1	87.3 ± 0.6	94.1 ± 0.5	96.5 ± 0.4	97.6 ± 0.4	98.2 ± 0.3

To explore in more detail the possibility of working with black and white images to identify weed seeds, one can consider using only the 8 morphological and textural features listed in the previous section. However, since two of the color features there mentioned are related only to the intensity channel (variance and skewness of the intensity histogram), they can also be obtained from black and white images. Consequently, we have explored the classification capabilities of a system built in terms of the resulting set of 10 features (Table 2). For $n=5$ the best classifier has a performance above 98% of accuracy, only less than 1% below the accuracy obtained including color features. Unfortunately, for $n=1$ (standard classification) there is a larger loss (from 93% to 88%). Of course, this performance might still be acceptable depending on regulations tolerance in concrete applications.

4 Summary and Conclusions

We extended our previous work [15] on the feasibility of using machine vision algorithms for the identification of weed seeds. We considered a large database comprising 10,130 images of 236 species, and discussed the discriminating power of weed seed characteristic measured from color images. The features considered were the same used after a careful selection performed in [15], seeking for the best classification parameters using the performance of a naïve Bayes classifier as evaluation criterion. This lead to the nearly optimal set of 12 characteristics –6 morphological, 4 color and 2 textural properties– listed in Section 2.

For the large database considered in this work, the naïve Bayes classifier produced again excellent results, as shown in Table 1. This seems to confirm the conclusion,

already advanced in our previous work, that due to the careful parameter selection the set retained approximately fulfills the independence assumed by the naïve Bayes approach. Table 1 also shows that the performance of this simple Bayesian approach and those of classifiers based on ANNs are essentially the same (except perhaps for $n=1$, where the latter are slightly better). In particular, for the bagged versions these performances reach 99% for $n=4$ classification options.

We have also investigated the possibility of using only black and white images of weed seeds. This alternative is appealing since in this case illumination conditions are less critical and the required hardware much cheaper, which are important advantages for a potential commercial system. Our results in Table 2 show that, by using a set of 10 features which includes the 8 morphological and textural characteristics and 2 parameters associated to the intensity histogram, the ANN approach is able to reach approximately 98% of classification accuracy. This is only a 1% loss with respect to the use of color features, and might still be an acceptable performance depending on the application. Notice that even in this more difficult problem the ANN approach does not perform better than the simpler Bayesian approach.

The number of species considered in this study is large enough to draw some definite conclusions. First, for not too critical applications involving the identification of a few hundred species, the present approach is effective and might even be implemented using black and white seed images. This is the case of the intended application of our work, in commercial seed purity analysis. For more critical problems, where either the identification performances in Tables 1 and 2 are not enough, presenting several options to an operator is not feasible, and/or the number of species involved is extremely large –like, for instance, for botanical gardens use– the implemented classifiers have to be improved. There are several ways of doing it: As already mentioned, feature selection for the ANN classifiers should be performed using an ANN as the selection criterion, although this would not probably make them much more efficient; more sophisticated feature selection methods [16,21] that the one used in [15] could be implemented; a better control of illumination conditions than the one used in this study (for instance, keeping the background color constant by an electronic control of the light source) would enhance the discriminating power of color and texture parameters; etc. In addition, "boosting" techniques[22,23] can be applied to improve the capabilities of the basic learners studied here, other standard classification methods may be considered (boosted trees, for instance), and the whole classification strategy might be changed to a more *ad hoc* method better suited for this application . Some of these possibilities are currently under consideration.

Acknowledgements

References

1. Draper, S.R., Travis, A.J.: Preliminary observations with a computer based system for analysis of the shape of seeds and vegetative structures. Journal of National Institute of Agricultural Botany 16 (1984) 387-395
2. Keefe, P.D., Draper, S.R.: The measurement of new characters for cultivar identification in wheat using machine vision. Seed Sci. Technol. 14 (1986) 715-724
3. Zayas, I., Lai, F.S., Pomeranz, Y.: Discrimination between wheat classes and varieties by image analysis. Cereal Chem 63(1) (1986) 52-56
4. Sapirstein, H.D., Neuman, M., Wright, E.H., Shwedyk, E., Bushuk, W.: An instrumental system for cereal grain classification using digital image analysis. J. Cereal Sci. 6 (1987) 3-14.
5. Chen, C., Chiang, Y.P., Pomeranz, Y.: Image analysis and characterization of cereal grains with a laser range finder and camera contour extractor. Cereal Chem. 66(6) (1989) 466-470
6. Symons, S.J., Fulcher, R.G.: Determination of wheat kernel morphological variation by digital image analysis: I. Variation in Eastern Canadian Milling Quality Wheats. J. Cereal Sci. 8 (1988) 211-218
7. Zayas, I., Pomeranz, Y., Lai, F.S.: Discrimination of wheat and nonwheat components in grain samples by image analysis. Cereal Chem. 66(6) (1989) 233-237
8. Neuman, M.R., Sapirstein, H.D., Shwedyk, E., Bushuk, W.: Discrimination of wheat class and variety by digital image analysis of whole grain samples. J. Cereal Sci. 6 (1987) 125-132
9. Neuman, M.R., Sapirstein, H.D., Shwedyk, E., Bushuk, W.: Wheat grain color analysis by digital image processing I. Methodology. J. Cereal Sci. 10 (1989) 175-182
10. Neuman, M.R., Sapirstein, H.D., Shwedyk, E., Bushuk, W.: Wheat grain color analysis by digital image processing II. Wheat class discrimination. J. Cereal Sci. 10 (1989) 183-188
11. Jansen, P.I.: Seed production quality in *Trifolium balansae* and *T. resupinatum*: The role of seed color. Seed Sci. Technol. 23 (1995) 353-364
12. Ahmad, I.S., Reid, J.F., Paulsen, M.R., Sinclair, J.B.: Color classifier for symptomatic soybean seeds using image processing. Plant Disease 83 (1999) 320-327
13. Petersen, P.E.H., Krutz, G.W.: Automatic identification of weed seeds by color machine vision. Seed Sci. Technol. 20 (1992) 193-208
14. Chtioui, Y., Bertrand, D., Dattée, Y., Devaux, M.F. : Identification of seeds by color imaging: Comparison of discriminant analysis and artificial neural networks. J. Sci. Food Agric. 71 (1996) 433-441
15. Granitto, P.M., Navone, H.D., Verdes, P.F., Ceccatto, H.A.: Weed Seeds Identification by Machine Vision. Computers and Electronics in Agriculture 33 (2002) 91-103
16. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(2) (1997) 153-158
17. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3(6) (1973) 610-621
18. Conners, R.W., Trivedi, M.M., Harlow, C.A.: Segmentation of a high-resolution urban scene using texture operators. Computer Vision, Graphics and Image Processing 25 (1984) 273-310
19. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
20. Breiman L.: Bagging predictors. Machine Learning 24 (1996) 123-140
21. Chtioui, Y., Bertrand, D., Barba, D.: Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision. J. Sci. Food Agric. 76 (1998) 77-86
22. Freund Y., Schapire, R.E.: A decision theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1) (1997) 119-139
23. Sharkey, A.J.C. Ed.: Combining Artificial Neural Nets. Springer-Verlag, London (1999)