AUTOMATIC IDENTIFICATION OF WEED SEEDS BY COLOR IMAGE PROCESSING

P. M. Granitto, H. D. Navone, P. F. Verdes and H. A. Ceccatto

Instituto de Física Rosario (CONICET – Universidad Nacional de Rosario)

Boulevard 27 de Febrero 210 Bis, 2000 Rosario, Argentina

(granitto,navone,verdes,ceccatto)@ifir.edu.ar

ABSTRACT: The analysis and classification of seeds are essential activities contributing to the final added value in the crop production. Besides varietal identification and cereal grain grading, it is also of interest in the agricultural industry the early identification of weeds from the analysis of strange seeds, with the purpose of chemically controlling their growth. The implementation of new methods for reliable and fast identification and classification of seeds is thus of major technical and economical importance. Like the manual identification work, the automatic classification of seeds should be based on knowledge of seed size, shape, color and texture. In this work we present a study of the discriminating power of morphological, color and textural characteristics of weed seeds, which can be measured from video images. This study was conducted on a large basis, considering images of weed seeds found in Argentina's commercial seed production industry and listed by the Secretary of Agriculture as prohibited and primary- and secondary-tolerated weeds. We first describe the experimental setting and hardware used to capture the seed images. Then, we define the morphological, color and textural parameters measured from these images, and discuss the selection of the most relevant ones for identification purposes. Finally, we present results for the identification of test images obtained using a Naive Bayes classifier and a committee of Artificial Neural Networks.

KEYWORDS: machine vision, seed identification, pattern recognition, neural networks.

1. INTRODUCTION

The analysis and classification of seeds are essential activities contributing to the final added value in the crop production. These studies are performed at different stages of the global process, including the seed production, the cereal grading for industrialization or commercialization purposes, during scientific research for improvement of species, etc. For all these purposes, different procedures based on manual abilities and appreciation capabilities of specialized technicians are employed. In most cases these methods are slow, have low reproducibility, and possess a degree of subjectivity hard to quantify, both in their commercial as well as in their technological implications. It is then of major technical and economical importance to implement new methods for reliable and fast identification and classification of seeds. Like the manual identification work, the automatic classification should be based on knowledge of seed size, shape, color and texture (i.e., greytone variations on the surface). Numerous image analysis algorithms are available for such descriptions, which make machine vision a suitable candidate for such a task.

Most previous attempts to identify seeds by machine vision have concentrated on cultivated varieties. Initially it was assumed that varietal differences could be extracted from the structure of the kernel, so different geometrical measurements were used to describe a variety[1,2]. Other investigations have been conducted to separate different species of cereal grains[3,4], wheat from non-wheat components (weed seeds and stones)[5], different types of wheat[6-9] and special grading classes[9,10], etc. In these studies the image analysis was essentially restricted to basic geometrical measurements to obtain different parameters (shape factor, aspect ratio, length/area, etc.). In addition, color was successfully used to separate red-, amber- and white-colored wheat, but could not separate into grading classes. More recent studies have used color images to establish seed quality and hardseededness of some annual pasture legumes[11], to characterize fungal damage, viral diseases and immature soybean seeds[12], etc.

Besides varietal identification and cereal grain grading, it is also of major interest in the agricultural industry the early identification of weeds from the analysis of strange seeds, with the purpose of chemically controlling their growth. Weed seeds are also identified by seed testing stations and seed corporations to measure the purity of the harvest, and by research stations to detect changes in the seed banks in the soil. The 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 among 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. Consequently, a successful approach should include parameters associated to all the relevant characteristics of size, shape, color and texture above mentioned.

An early attempt to identify weed seeds[13] showed the importance of using color instead of black and white images to improve classification accuracy. However, this investigation was conducted considering only four different weed species, which does not provide a good characterization of seeds variations. In this work we present a study of the discriminating power of morphological, color and textural characteristics of weed seeds. This study was conducted on a much larger basis, including seed images of frequent weeds found in Argentina's commercial seed production industry. In order to avoid having a bias in the selection of species to be included in this study, we restricted ourselves to the 58 species listed by the Secretary of Agriculture as prohibited and

primary and secondary tolerated weeds. From this list we finally considered 57 species for which a good number (~ 40) of young exemplars were available in the seed bank of the Seed Analysis Laboratory at the Oliveros Experimental Station of the National Institute for Agricultural Technology (INTA).

This work is organized as follows. In Section 2 we describe the hardware and the experimental setting conditions used to capture the seed images. Then, in Section 3 we define the morphological, color and textural parameters measured from these images, and discuss the selection of the most relevant ones for identification purposes. In Section 4 we present the results obtained using two different classification methods (Naive Bayes and Artificial Neural Networks). Finally, in Section 5 we draw some conclusions.

2. EXPERIMENTAL SETTING FOR SEED IMAGES ACQUISITION

We have built a database containing 3163 images of the 57 species considered (a list of these species is available on request). To acquire the images we used a Sony XC-711P RGB video camera with a 2/3" CCD, connected to a AM-CLR frame grabber from Imaging Technology with 8 bits look-up tables per color channel. Illumination was provided by a 150W Fostec light source through a quadruple fiber optic bundle of 12.7mm diameter, with the four guides in a symmetric arrangement to produce an even illumination with good texture enhancement. Regardless of the seed sizes, all images were taken to approximately fill the camera field of view by adjusting a 6000 System 6.5X parfocal zoom from Navitar. Lens attachments of 0.5X and 2X allowed to cover the seeds size range of the species considered. Light intensity was regulated with an iris diaphragm in order to adjust the illumination to the changing field of view while keeping a constant color temperature (corresponding to a standard Ushio 20V-150W halogen projector lamp). A better control of illumination conditions would have enhanced the classification capabilities of color and texture parameters, which could be required for a commercial system. However, the experimental setting just described was considered enough for the purposes of the present work.

Images were taken with a 768×512 pixel resolution on a blue background, which can be easily subtracted by standard segmentation routines because of the difference in color with the seeds. The segmented images consist of arrays whose entries are 24-bit records, corresponding to the 256 pixel intensity levels (8 bits) for each of the red (R), green (G) and blue (B) channels. In order to reduce effects associated to illumination changes, we also considered the normalized red ($r=R/I$) and green ($g=G/I$) pixel values, where $I=(R+G+B)/3$ is the average intensity.

3. CLASSIFICATION PARAMETERS

We have measured a number of features from the raw seed images to be later used for classification purposes. As stated above, these features correspond to morphological, color and textural characteristics of the seeds. Below we briefly describe the different parameters considered.

MORPHOLOGY

Size and shape characteristics of the seeds can be easily obtained from the binarized images. In particular, we have measured the lengths of the principal axes and several moments of the planar mass distribution with respect to these axes, the size of the minimal rectangular box containing the seed and the ratio of its area to the seed area (compactness), etc. All these 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). Furthermore, since we used the principal axes as the reference frame for all the measurements, the resulting values are independent of the image orientation. In total, we have measured 21 morphological features.

COLOR

We have determined the gray level histograms in the *I,r,g* channels. From these histograms we considered standard features such as average, variance and skewness. In addition, we considered ratios of average histogram values in the RGB channels like, for instance, E[*R*]/E[*I*] and E[*G*]/E[*I*] (here E[.] means the average pixel value in the corresponding channel). We have measured 12 different color characteristics.

TEXTURE

Like in [13], two different textural analysis were used to describe the texture of the seed surface:

1. *Gray Level Co-Occurrence Matrix*: A two-dimensional matrix with entries A_{ij}, where *i,j* are gray levels and the entry value gives the number of nearest-neighbor pixels in the image having these gray levels along a given direction (we used alternatively both principal axis directions, which makes the textural features rotational invariant). In practice, we have considered a coarse-grained version of this two-dimensional histogram. First, we performed a dynamical equalization of the gray level histogram on each channel using 16 boxes in order to eliminate illumination intensity variations[14]. Then, the indices *i,j* were made to correspond to these box levels. From the resulting 16×16 matrix 17 textural features were obtained. The precise definition of these parameters and the interpretation of their discriminating properties can be found in [14,15].

2. *Gray Level Run Length Matrix*: The two dimensions in this matrix are the gray level and the so-called run length, *i.e.* the base 2 logarithm of the number of adjacent pixels in a given direction with the same gray level. We have considered both principal axis directions to compute the run lengths. In this case the matrix dimension was reduced by taking the same 16 gray level intervals used before. The resulting matrix allows to measure 4 new textural features. The precise definition of these parameters and the interpretation of their discriminating properties can be found in [16].

In total we have considered 42 textural characteristics. Then, from each color image we measured 75 parameters to be used for classification. By simple inspection we determined that several of them had erratic behaviors and could be discarded. Finally we retained 15 morphological, 8 color and 17 textural properties. Of course this large set of parameters still contained redundant, too noisy or even irrelevant information for classification purposes. In order to choose the best features in each group (those with the largest discriminating power), we implemented standard sequential forward and backward selection algorithms[17] using the performance of a Naive Bayes classifier as selection criterion. The Naive Bayes classifier fits the class conditional probabilities with a product of normal distributions of the individual features and, in spite of its simplicity, it has a very good performance for this problem (see next section). The selection algorithms reduced the parameters to nearly optimal sets of 10 morphological, 7 color and 7 textural features. The same procedure applied to the joint 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 parameters is given in the Appendix.

4. RESULTS

In order to compare the discriminating power of the different set of features, in Table I we present the generalization capabilities of Naive Bayes classifiers built solely in terms of the 10

morphological, 7 color or 7 textural features. For this we have split the 3163 images of the 57 species considered in training and test sets, using, for each species, 80% of the images to build the classifier and including the remaining 20% in the test set. This leaves 2527 images for training and 636 images for testing the system. Table I gives the performances on both the training and test sets, and also indicates how these performances increase when the system is given the chance to assign a given image to any of the n most probable classes, for n=1,2 and 3 (this possibility is very useful in practice, since untrained operators can easily select the correct option by simple visual inspection of stored representative seed images of the n classes suggested by the classifier).

FEATURES	FIRST OPTION		FIRST TWO OPTIONS		FIRST THREE OPTIONS	
	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
MORPHOLOGY	86.3	85.5	95.9	95.8	98.3	97.5
COLOR	62.1	49.2	74.4	64.5	82.1	73.0
TEXTURE	55.6	51.3	69.4	65.4	77.4	72.6

Table I: Naive Bayes classifier performances in % of correct seed identifications using only one particular set of features at a time.

A quick look at this table shows, as expected, the large discriminating power of morphological features. As anticipated, color is not particularly good because many species are light to dark brownish or black; its discriminating power is nearly equal to that of textural features. However, if we consider any two combined set of features (see Table II), morphology plus color features have and edge over the combined use of morphology and texture characteristics. Notice, however, that in this last case it would be enough to consider black and white images, which constitutes an important simplification and a reduction in hardware cost. Finally, the performances of the Naive Bayes classifier built in terms of the optimal set of 12 features listed in the Appendix are given in Table III.

FEATURES	FIRST OPTION		FIRST TWO OPTIONS		FIRST THREE OPTIONS	
	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
MORPHOLOGY + COLOR	96.7	95.4	99.3	98.4	99.5	99.4
MORPHOLOGY + TEXTURE	91.7	90.4	97.7	96.4	98.6	98.6
COLOR + TEXTURE	84.0	74.5	91.8	84.7	95.0	90.3

Table II: Naive Bayes classifier performances in % of correct seed identifications using different combination of two sets of features.

We have also developed a classifier based on Artificial Neural Networks (ANN)[18]. To this end we trained 10 feedforward networks with 12 input, h hidden, and 57 output units. The numbers of input and output units correspond to the number of parameters used and seed species to be identified respectively. The number of hidden units was varied from h=20 to h=80; the results presented below correspond to h=40 units, which was found to be nearly optimal. 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. We trained the ANN 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. The performance of a single (generic) ANN and the results obtained by structuring the 10 networks in a committee are shown in Table III. In the case of the ANN committee we have two options: i) each network votes for the class with the

largest probability according to its own outputs, and the image is finally assigned to the class with the majority of votes, and ii) the class probabilities output by the 10 networks are added and the image is assigned to the class with the largest sum value. These two options correspond respectively to the upper and lower results for the ANN committee in Table III.

CLASSIFIER	FIRST OPTION		FIRST TWO OPTIONS		FIRST THREE OPTIONS	
	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
NAIVE BAYES	97.3	96.2	99.4	98.7	99.8	99.4
SINGLE ANN	100	95.3	100	98.0	100	99.1
ANN COMMITTEE	100 100	96.7 96.4	100 100	98.1 98.6	100 100	98.1 99.2

Table III: Performances of different classifiers in % of correct seed identifications using the optimal set of 12 features listed in the Appendix.

Several comments are in order at this point. First, we stress the excellent performance of the Naive Bayes classifier, which for this problem competes with the more sophisticated ANN method. Second, since the performance of a single network is already very good, there is no much room left to improve this performance by adding several predictors in a committee. Notice that from the 636 images in the test set, finally only 5 images are misclassified when the system is allowed to suggest three options for class membership (the performance reaches 100% with five options). Of course, for a much larger number of species the classification problem would be more demanding and the ANN committee might have an edge over the other simpler methods. Also note that, for simplicity, the feature selection in Section 4 was performed using the Naive Bayes classifier, which not necessarily produces the optimal set for the ANN approach. Moreover, there are much more sophisticated feature selection method that can be implemented[17]. Finally, as an important remark, we mention that different realizations of training and test sets do not sensibly change the performances shown in Tables I to III.

5. CONCLUSIONS

We have performed a detailed and extensive analysis of the discriminating powers of different features associated to color weed seed images. First, we collected a fairly large database with images of seeds of the most frequent weed species present in the commercial seed production industry. This set of images was then suitably processed to extract a large number of morphological, color and textural properties, which were later considered for classification purposes. A careful selection of the best parameters, *i.e.* those having the largest discriminating power, reduced the measured seed characteristics to only 12 features (6 morphological, 4 color and 2 textural properties). The implementation of two different classifiers on the basis of these parameters produced excellent results and allowed us to establish the relative importance of the different kind of features in the identification process. As expected, morphology is the principal characteristic for seed identification, although color and texture are also contributing to the final classifier performance. These last properties have approximately the same discriminating power when considered independently of morphology.

For the number of species considered, the preprocessing of images and the careful selection of measured features reduced considerably the complexity of the classification problem. However, one might expect that this problem will become more demanding for databases containing several

hundreds of species, as required for a commercial system. In such a case, several important improvements on the classifier development can be implemented. For instance, when using ANN the feature selection must be performed using this method to evaluate the importance of different parameters, which would lead to the set best suited for this approach. In addition, the implementation of optimal ensemble techniques instead of a simple voting committee could take advantage of the ANN variance (nonidentifiability of the model) on complex problems[19]. Work in this direction previously requires the lengthy acquisition of the extended database, which is currently in progress.

ACKNOWLEDGEMENTS

We acknowledge the constant assistance of Ing. Roque Craviotto and technicians of the Seed Analysis Laboratory at EEA Oliveros of INTA. This project was partially financed through grant PICT 11-03834 from ANPCyT.

APPENDIX

The final 12 parameters selected for classification are the following:

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 axis [M_{20}, M_{21}, M_{22}]

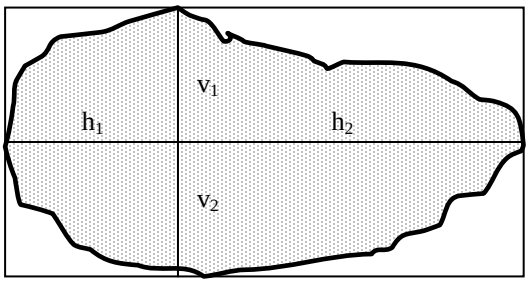

Figure 1

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[14] (main principal axis direction)
Cluster Prominence[15] (secondary principal axis direction)

REFERENCES

[1] Draper, S.R. and Travis, A.J.. "Preliminary observations with a computer based system for analysis of the shape of seeds and vegetative structures". *Journal of the National Institute of Agricultural Botany*, **16**, 387-395, 1984.

[2] Keefe, P.D. and Draper, S.R.. "The measurement of new characters for cultivar identification in wheat using machine vision". *Seed Science and Technology*, **14**, 715-724, 1986.

[3] Sapirstein, H.D., Neuman, M., Wright, E.H., Shwedyk, E. and Bushuk, W.. "An instrumental system for cereal grain classification using digital image analysis". *Journal of Cereal Science*, **6**, 3-14, 1987.

[4] Chen, C., Chiang, Y.P. and Pomeranz, Y.. "Image analysis and characterization of cereal grains with a laser range finder and camera contour extractor". *Cereal Chemistry*, **66**(6), 466-470, 1989.

[5] Zayas, I., Pomeranz, Y. and Lai, F.S.. "Discrimination of wheat and nonwheat components in grain samples by image analysis". *Cereal Chemistry*, **66**(6), 233-237 , 1989.

[6] Zayas, I, Lai, F.S. and Pomeranz, Y.. "Discrimination between wheat classes and varieties by image analysis". *Cereal Chemistry*, **63**(1), 52-56 , 1986.

[7] Symons, S.J. and Fulcher, R.G.. "Determination of wheat kernel morphological variation by digital image analysis: I. Variation in Eastern Canadian Milling Quality Wheats". *Journal of Cereal Science*, **8**, 211-218 , 1988.

[8] Neuman. M.R., Sapirstein, H.D., Shwedyk, E. and Bushuk, W.. "Wheat grain color analysis by digital image processing I. Methodology". *Journal of Cereal Science*, **10**, 175-182, 1989.

[9] Neuman. M.R., Sapirstein, H.D., Shwedyk, E. and Bushuk, W.. "Wheat grain color analysis by digital image processing II. Wheat class discrimination". *Journal of Cereal Science*, **10**, 183-188, 1989.

[10] Neuman. M.R., Sapirstein, H.D., Shwedyk, E. and Bushuk, W.. "Discrimination of wheat class and variety by digital image analysis of whole grain samples". *Journal of Cereal Science*, **6**, 125-132, 1987.

[11] Jansen, P.I.. "Seed production quality in *Trifolium balansae* and *T. resupinatum*: The role of seed colour". *Seed Science and Technology*, **23**, 353-364, 1995.

[12] Ahmad, I.S., Reid, J.F., Paulsen, M.R. and Sinclair, J.B.. "Color classifier for symptomatic 7soybean seeds using image processing". *Plant Disease*, **83**, 320-327, 1999.

[13] Petersen, P.E.H. and Krutz, G.W.. "Automatic identification of weed seeds by colour machine vision". *Seed Science and Technology*, **20**, 193-208, 1992.

[14] Haralick, R.M., Shanmugam, K. and Dinstein, I.. "Textural features for image classification". *IEEE Transactions on Systems, Man, and Cybernetics*, **3**(6), 610-621, 1973.

[15] Conners, R.W., Trivedi, M.M. and Harlow, C.A.. "Segmentation of a high-resolution urban scene using texture operators". *Computer Vision, Graphics and Image Processing*, **25**, 273-310, 1984.

[16] Galloway, M.M.. "Textural analysis using gray level run length". *Computer Graphics and Image Processing*, **4**, 172-179, 1975.

[17] Jain, A. and Zongker D.. "Feature selection: Evaluation, application, and small sample performance". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(2), 153-158, 1997.

[18] Bishop, C.M.. *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press. 1995

[19] Navone, H.D., Verdes, P.F., Granitto, P.M. and Ceccatto, H.A.. "A new algorithm for selecting diverse members of a neural network ensemble". In *Proceedings of the VI International Congress on Information Engineering*, Universidad de Buenos Aires, 2000.