Default

A Computer-based System for Seed Identification

T. Daoust,* K. Fujimura, M. B. McDonald and M. A. Bennett

ABSTRACT

An image processing computer application was developed to collect statistics of physical characteristics from seeds. A machine learning technique ensured that the software was applicable to a wide variety of species so that it could be used in purity analysis tests. The method presented requires an inexpensive scanner and a modern personal computer. The software operates by locating seeds within the digitized image of the purity sample and takes measurements on each seed (width, height, area, perimeter, average color, etc.). These measurements are inputted into a classification routine trained to recognize all potential seeds within the sample. The classification routine determines the closest matching species for each seed in the image and reports the results to the user. Because of the wide variation in seeds encountered in purity tests, the automated system of measurement and classification is highly configurable. The seed identification system is designed to be rapidly adapted to specific seed types and trained without knowledge of artificial intelligence techniques.

INTRODUCTION

Seed laboratories and companies routinely perform purity analyses since they are required by federal and state laws as information on a seed label describing the quality of the seed lot to customers. The traditional four-part purity analysis required by the AOSA (2002) is a process in which a seed analyst manually sorts and weighs pure seeds of the labeled species, other crop seeds, weed seeds, and inert matter within a sample (Effenberger, 2001). This is done by placing a representative sample on a clean, hard surface, drawing small portions of the sample across the surface and categorizing each seed or particle as it passes through the field of view. The final classification is often made using a magnifying lens or dissecting microscope. The speed of the test can vary widely based on the experience of the analyst and the quality and type of sample. An experienced analyst working with a clean sample may be able to conduct a purity analysis on 100 g of moderately sized seeds in approximately 15 min.

Seed identification systems using computers can detect seeds and count their number within a sample and may identify the species or indicate potential matches for seeds that the analyst does not immediately recognize. If they are combined with mechanical systems, they could even sort seeds that closely match the desired crop from the rest of the sample to perform automated purity analysis.

T. Daoust* and K. Fujimura, Department of Computer and Information Science, The Ohio State University, 2015 Neil Ave. Columbus, Ohio 43210-1210 USA (e-mail: daoustt@yahoo.com); M. B. McDonald and M. A. Bennett, Department of Horticulture and Crop Science, 2021 Coffey Rd., The Ohio State University, Columbus, Ohio 43210-1086 USA. *Corresponding author. Received July 27, 2004.

Studies have been reported on attempts to achieve this objective. Travis and Draper (1985) proposed the use of automated purity analysis methods, but concluded that the features and classification techniques used were insufficient. A successful early effort to distinguish between 'Arkan' and 'Arthur' wheat (*Triticum aestivum* L.) varieties using discriminant analysis (Zayas, et al., 1985) was followed by a report of a method that could distinguish between several other varieties with an overall classification rate of approximately 80% (Zayas, et al., 1986). Zayas, et al. (1989) later conducted a study using discriminant analysis to separate wheat, weeds, and stones. The method achieved accurate classification of wheat and non-wheat particles. Attempts were also made to classify 17 different species using both 3d volumetric data and seed contour information, but the system was only capable of analyzing an area of less than 2 cm^2 (Chen, et al., 1989). Using color and texture in seed classification has also been explored. Classification of wheat varieties based solely on color features showed accuracy rates between 34 and 90% (Neuman, et al., 1989). Color texture features were potentially valuable in weed seed identification (Petersen and Krutz, 1992). Using a complex set of features, Chtioui, et al. (1996) compared the performance of discriminant analysis and artificial neural networks at distinguishing four species. Their study showed accuracy rates that indicated automated purity analysis involving four seeds was feasible, but concluded that the high number of seeds and contaminants in routine purity testing would require a more complex system.

These studies relied on ideal circumstances including aligned orientation and clear separation among seeds. Seed identification systems may be required to operate in conditions with seeds that are in contact with each other at random orientations. The issue of seeds in contact with each other has been mentioned by several authors (Zayas et al., 1989; Chtioui et al, 1996), but a workable solution has yet to be presented. In addition, the underlying environment of a classification system must be adaptable to different crops and to a large number of other species to be of practical value for seed purity testing.

The purpose of this study was to determine the performance of a more general seed identification system that allows: 1) the use of 2d shape and color features to perform seed classification; and 2) a configuration of the software to permit the recognition of a wide range of species by rapid retraining.

MATERIALS AND METHODS

Image Acquisition

The specific computer imaging techniques used in this study have been reported elsewhere (Sako, et al., 2001a). The following procedure was used in purity sample image acquisition:

1. A solid white sheet of paper was placed in a custom-built black-painted box designed to eliminate outside light. The box had two inexpensive scanners (Umax Astra 4400, Freemont, CA) attached to the upper lid that were suspended approximately 1 cm above the sheet of paper when the lid was closed (Sako, et al., 2001a).

2. Seeds were separated on the paper and the lid was closed.
3. The scanner was set to 200 dpi, and a 32-bit color image in JPEG format was obtained.

Computer

A Sony Viao (Tokyo, Japan) laptop with an Intel (Santa Clara, CA) Pentium III 750 MHz CPU, 128 MB Ram, and a 10 GB hard drive was the principal development platform. The software was also developed and tested on a custom built system containing an AMD (Sunnyvale, CA) Athlon 750 MHz CPU, 256 MB Ram, and a 40 GB hard drive.

Software Processing of the Image

The software designed for this study was developed using Microsoft (Redmond, WA) Visual C++ 6.0, Silicon Graphics (Mountain View, CA) Image Format Library (IFL) 1.3.1, and Newmat10. Operation of the software requires an initial training phase for each type of seed the user wishes to identify. Once the software has been trained, the image of the sample under test is opened and analyzed by the program. The results of the classification are displayed to the user, who can also correct any mistakes made by the software.

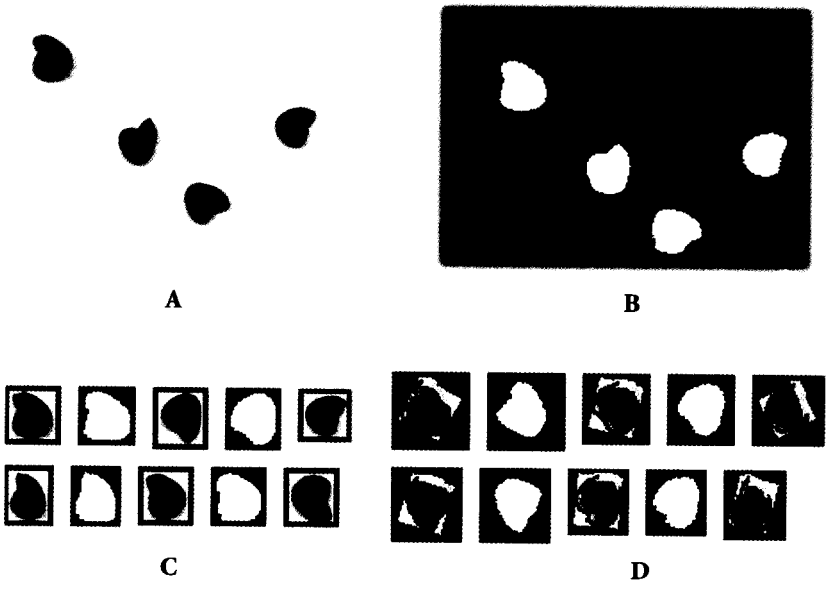
The operation of the software during the training and classification stages is essentially the same. An algorithm analyzes the full color image to separate foreground pixels from background pixels. Because a wide variety of backgrounds might be encountered in laboratories, the user can select the algorithm to best perform this operation. For the images in this study, a global thresholding routine with user defined values separated background and foreground pixels.

Once the foreground pixels (seeds) are identified, another algorithm groups contiguous pixels as a single unit and labels the region as a seed. This algorithm operates by scanning the image for unlabeled foreground pixels and recursively labeling all touching foreground pixels. A noise-removal algorithm was then applied to remove foreground regions smaller than a threshold size that would not represent seeds.

Because the seeds are randomly placed on the sheet of paper, the orientation of the seeds varies. To simplify the measuring algorithms, the orientation of the remaining foreground regions was determined by means of a moment calculation, which performed automatic alignment. A standard rotation algorithm was then run on the foreground regions (seeds) to align them vertically (Fig. 1).

Once seeds were rotated, a check was run to identify and separate contiguous seeds. This operation was performed by analyzing the concavity of a foreground region (seed) by computing the perpendicular distance from the convex hull to the seed perimeter (Fig. 2). The calculations were passed through a high pass filter before local maxima were computed. These maxima corresponded to potential locations where seeds touched. If two such locations were found within close proximity, the foreground region was split along a line that connected the two points. The separated regions were then rotated and the outline and convex hull recomputed.

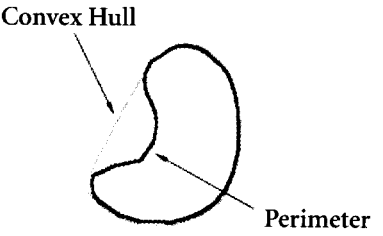
FIGURE 1. Phases of feature extraction process. (A) A portion of an input image before processing. (B) The image in part A after a thresholding routine has converted the image into a two color (binary) image. (C) Areas containing seeds isolated from the background. (D) Measurements taken after automated rotation is performed on the seeds.



The processing of the image in the previous steps allowed other algorithms to collect several statistics about each foreground region (seed). These included:

1. *Width* – Width was measured as the distance between the left and right most pixels within the rotated region.
2. *Height* – Height was measured as the distance between the top and bottom most pixels within the rotated region.
3. *Width to height ratio* – This ratio was computed as width/height.
4. *Perimeter* – Perimeter was the length of the outside border of the foreground region. Chain coding (Gonzalez and Woods, 1992) locates and stores the pixels comprising the border of the seed and determines the total border length around the object in pixels. Adjacent pixels along the perimeter were assigned a distance of 1 pixel and diagonally touching pixels were assigned a distance of 1.414 pixels (Fig. 2).
5. *Area* – Area was measured by counting the number of pixels in the

FIGURE 2. An example of a convex hull calculated from a shape perimeter.



foreground region by means of an algorithm based on chain coding of the perimeter.

6. *Average red, green, and blue color values* – Color was described by three values for the intensity of the red, green, and blue components between 0 and 255 resulting in approximately 16.7 million different colors. The average values were determined for each foreground region.
7. *Average hue, saturation, and intensity* – In addition, another system of color description was used that describes colors with three other values. The first value was the hue, which is a means of quantifying what humans describe as red, green, blue, etc. The two other values were intensity and saturation. Intensity is a measure of the brightness of the pixel. Saturation is a measure of how dominant the pure hue is in the color. The average values were determined for each foreground region.
8. *Convex hull area and perimeter* – The convex hull of a region is the smallest convex shape containing the entire region (Fig. 2). The convex hull area and perimeter were determined using the chain code of the pixels belonging to the perimeter of the seed.
9. *Area and perimeter ratios* – The area ratio was calculated as convex hull area divided by the actual area of the foreground region. The perimeter ratio was calculated as convex hull perimeter divided by the actual perimeter of the foreground region.
10. *Extent-fill* – Extent-fill was calculated as area/(width*height). It indicates how close the shape is to a rectangle. For elliptical and circular regions, the extent-fill area was designated as pi*(width/2)*(height/2) pixels resulting in an approximate calculation of 0.78 for most elliptical shapes.

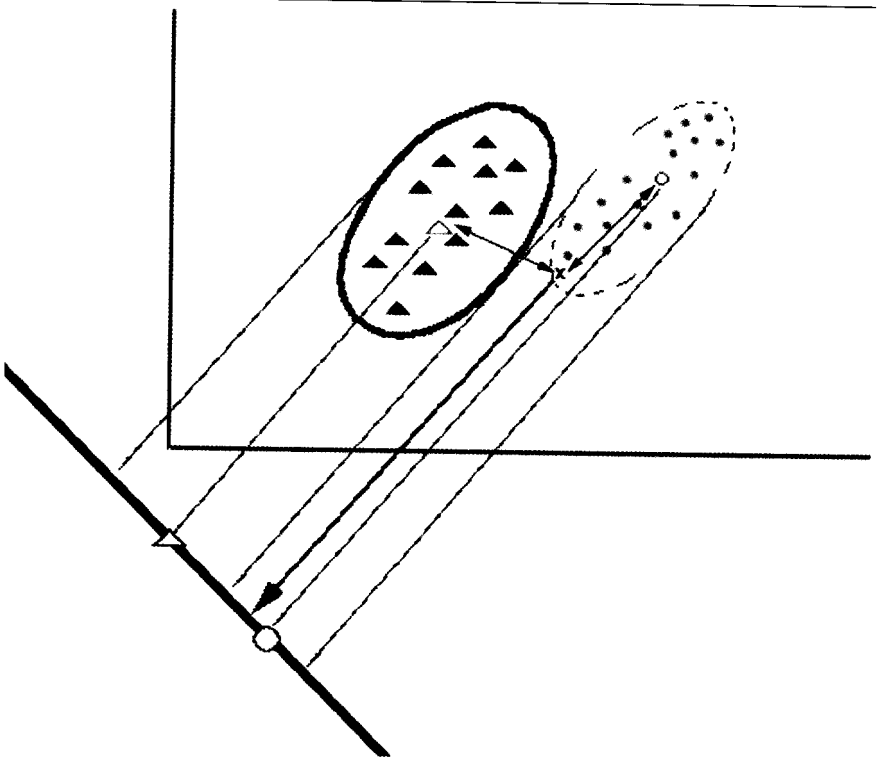
Seed Classification

The classification routine used in this study was previously described for giant ragweed (*Ambrosia trifida* L.) seeds (Sako, et al., 2001b). The software operates by taking a large number of measurements and recognizing similarities between foreground regions of the sample under test and foreground regions identified in the training period as seeds of known species.

Before classifying foreground regions in a test sample as seeds, a database containing statistics about the different seeds of the examined species must be developed during the training period. This is done either by clicking on the foreground region and typing the name of the species or assigning all foreground regions in an image to the same species. For a correct identification of seeds in a test sample, statistics of a large number of seeds for each species must be in the database.

The primary classification technique used in this study was the Fisher Linear Discriminant (FLD) (Duda and Hart, 1973). The FLD performed substantially better and faster than nearest neighbor classification and a neural network (Russel and Norvig, 1995). After calculation of the FLD (Fig. 3) using the training database, each seed was projected into the new feature space. The dimension of the data was reduced to a value specified by the user, which was usually three or four features often accounting for 99% of the information.

FIGURE 3. A simple schematic of Fisher's Linear Discriminant (FLD) in two dimensions. In the original feature space, the **x** is equidistant from both class centers, although it is clearly a member of the class represented by dots. The new feature space is represented by the black line, and the FLD is defined as the projection (the arrows) from the two dimensional feature space to the black line. Using the new feature space, the **x** is clearly a member of the circle class.



Classification was then performed in the transformed feature space using two methods.

The first method found the class of the closest projected seed in the new feature space. This method required the computation of the distance from the current seed to every training seed within the database. The seed was assigned to the class of its closest neighbor. This method was sensitive to seeds in the training set that deviated from the class average. The second method computed the class average of each class within the transformed feature space. The distance from the unknown seed to each class average was computed, and the closest class chosen as the unknown class. This method was used for classification in this study.

Procedure

To develop the databases for this study, 21 training images with seeds of 21 different species were scanned. Each training image had 100 to 500 individual

seeds (Fig. 4). The software analyzed each training image and saved the measurements of approximately 5,000 training seeds to a global database.

FIGURE 4. An example of a wheat training image.

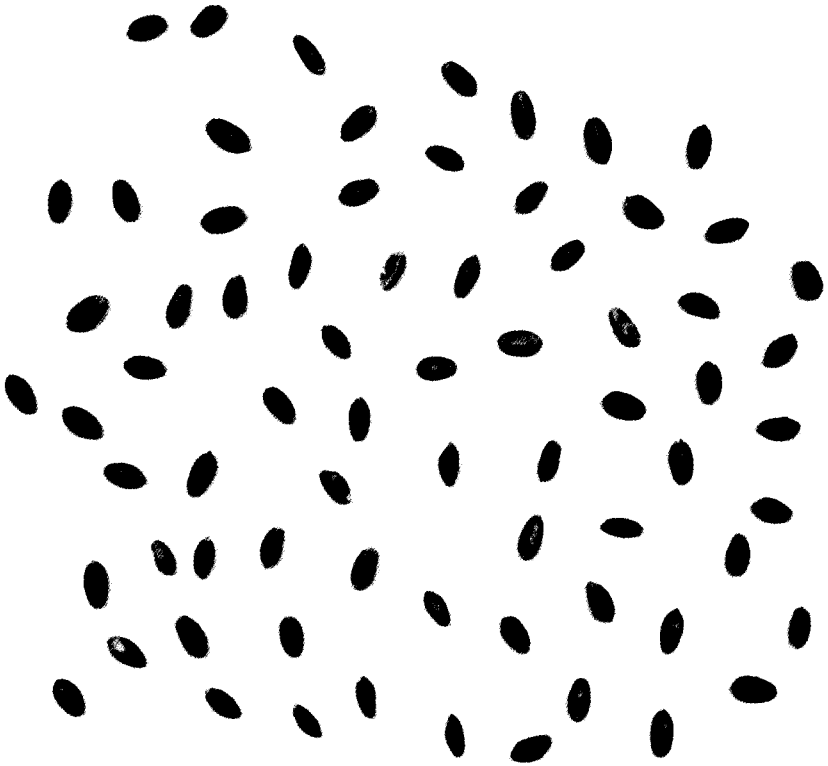


FIGURE 5. An example of a test image using ryegrass (A), rye (B), and spring triticale (C).

Analyzing a subset of the training database and saving the results created specific databases that generally contained three species. These specific databases were selected to discriminate species with extremely similar seeds, seeds with several features in common, and seeds with widely varying features. In these specific databases, the feature space was less crowded allowing finer distinctions.

A series of test images were created that contained the same seeds used to create the specific databases. For instance, the test image developed for a database that contained rye (*Secale cereale* L. subsp. *cereale)*, ryegrass (*Lolium perenne* L.), and spring triticale (×*Triticosecale* Wittm. ex A. Camus) only contained seeds of these three species. The test image was created by placing 20 seeds of each species in close proximity to each other on the sheet of paper (Fig. 5). Each test image was analyzed using the corresponding specific database containing the three species. A separate analysis was done using the global database. The results of the computer classification were then compared to the known identity of each seed.

The following 21 species were included in the global database:

Alfalfa (*Medicago sativa* L. subsp. *sativa*)
Broad Leaved Dock (*Rumex obtusifolius* L.)
Crabgrass (*Digitaria ischaemum* (Schreb. ex Schweigg.) Schreb. ex Muhl.)
Giant Foxtail (*Setaria faberi* R. A. Wilterrm.)
Ivyleaf Morning Glory (*Ipomoea hederacea* Jacq.)
Johnsongrass (*Sorghum halepense* (L.) Pers.)
Wild Carrot (*Daucus carota* L. subsp. *carota*)
Sorghum (*Sorghum bicolor* (L.) Moench.)
Turnip (*Brassica rapa* L. var. *rapa*)
Wheat (*Triticum aestivum* L.)
Red Clover (*Trifolium pratense* L.)
Kale (*Brassica oleracea* L. var. *viridis* L.)
Large Crabgrass (*Digitaria sanguinalis* (L.) Scop.)
Orchardgrass (*Dactylis glomerata* L.)
Poison Hemlock (*Conium maculatum* L.)
Rye (*Secale cereale* L. subsp. *cereale)*
Perennial Ryegrass (*Lolium perenne* L.)
Yellow Foxtail (*Setaria pumila* (Poir.) Roem. & Schult. subsp. *pumila*)
Spring Triticale (×*Triticosecale* Wittmack ex A. Camus)
Velvetleaf (*Abutilon theophrasti* Medik.)
White Clover (*Trifolium repens* L.)

RESULTS AND DISCUSSION

The overall rate of correct classification in all the tests was 76%. The rate of correct classification for tests with only three species in the database was 94%. The rate of correct classification for tests using the global database with 21

species was 67%. The lower rate of correct classification when using the global database was consistent with the expectation that within an increasing number of species, the distances between the species centers with the feature space decreased, making a correct decision more difficult. For example, in Fig. 3, the addition of a third class located on the horizontal axis between the listed classes makes the current projection calculated by FLD much less accurate. Finally, the feature spaces of the different classes have an increased likelihood of overlapping when multiple features are applied.

The results confirm that species with different seed features can be distinguished easily by FLD. For instance, the rate of correct classification for broadleaf dock, white clover and wheat was 100% when using a database containing only these species. Seeds of these species show considerable difference in size, shape and color. The other test with a 100% rate of correct classification involved sorghum, velvetleaf and ivyleaf morningglory. These seeds are similar in size, but have distinctive shape features that allow accurate classification even when using the global database containing the 21 species.

The ability of the system to distinguish seeds that are similar in appearance is evident when using the specific databases. For example, for distinguishing ryegrass, rye, and spring triticale, the system achieved a 90% rate of correct classification although the shape, size, and color of seeds of these three species are quite similar. The rates of correct classification of kale, alfalfa, and turnip were on average 93% using the specific database and on average 83% using the 21 species global database. The shape, size, and color of these seeds are also quite similar.

The misclassifications fell into two general groups. Misclassifications of the first group occurred when the global database was used to analyze an image that contained three species with similar seeds. For instance, under these circumstances, ryegrass was misclassified as either tall fescue or rye. The main reason was the high number of species and the inability to make a unique classification. Misclassifications of the second group occurred due to greater variability within a species resulting in considerable overlaps. For example, seeds were frequently mistaken as red clover because the shape, size, and color of some red clover seeds closely resemble alfalfa, white clover, and even turnip seeds. Under these conditions, it is difficult to find a projection that separates the seeds even when only the specific databases were used.

Although the system correctly distinguished between similar seeds, it made substantial errors in test cases that are quite simple for manual testing. In an image that contained broadleaf dock, yellow foxtail and tall fescue that was analyzed by the specific database with these three species, the dense, shiny black broadleaf dock seeds were frequently mistaken for the light yellow foxtail seeds. When using the global database, none of the broadleaf dock seeds were correctly identified. Although broadleaf dock and yellow foxtail seeds are substantially different in color, the shape of the two seeds can appear similar. Analysis of the standard deviation and class averages confirmed that overlap could occur for shape features, although the exact reason the FLD did not weigh the color component higher was not immediately evident.

The relative success of this system when used in purity analysis depends on how the technology is applied. To improve fully automated purity analysis, the system must routinely achieve rates of classification higher than 90% for a wide variety of seeds. The results of this study do not indicate that the classification method and features selected achieve this goal at this time. However, this study does document that this technology has the potential to assist seed analysts in purity tests. The majority of the seed analyst's time is spent examining crop seed within the purity sample. A number of seeds had high recognition rates in all tests, indicating that these seeds were good candidates for automatic identification. Ivyleaf morning-glory seeds had a 100% recognition rate (Tables 1 and 2) and white clover seeds had a 99% recognition rate (Tables 1–4). Several other seeds had overall recognition rates at or above 90%: spring triticale – 93% (Tables 1 and 2), kale – 90% (Tables 1–4), velvetleaf – 95% (Tables 1 and 2), and sorghum – 98% (Tables 1 and 2). If these seeds were consistently identified and other seeds in the sample were not mistaken for the crop seed, an automated

TABLE 1. The seeds grouped together in the table below were used to create a test specific database (seven databases in total). A test image that only contained the three listed seeds was then analyzed using the test specific database.

Seed	Classification		
	Correct	Incorrect	
	#		%
Ryegrass	18	2	90
Rye	17	3	85
Spring Triticale	19	1	95
Kale	19	1	95
Alfalfa	20	0	100
Turnip	17	3	85
Velvetleaf	20	0	100
Ivyleaf Morning-glory	20	0	100
Sorghum	20	0	100
Wild Carrot	18	2	90
Giant Foxtail	18	2	90
Poison Hemlock	20	0	100
Broad Leafed Dock	13	7	65
Yellow Foxtail	20	0	100
Tall Fescue	20	0	100
Johnsongrass	20	0	100
Large Crabgrass	17	3	85
Orchardgrass	19	1	95
Broadleaf Dock	20	0	100
White Clover	20	0	100
Wheat	20	0	100
Total	395	25	94

system could sort the crop seed from the contaminants in a purity sample. A seed analyst would then classify the seeds not automatically identified by the system in the traditional manner. Checks in the software could ensure that only unknown seeds that closely match the crop seed are sorted from the main sample to reduce the number of seeds being evaluated. These checks could be implemented by defining a maximum Euclidian distance from a class center where classification by computer is accepted. Additional checks would ensure that the regions where computer classification is accepted do not overlap. More advanced systems that model the acceptable areas of the feature space could also be developed. Such a system could reduce the workload of the purity analyst.

The results obtained in this study may be improved by adding other features to the classification routines. Texture varies widely among seeds, and has been demonstrated as a powerful feature in seed classification (Chtioui et al., 1996). A large variety of techniques already exist to express and classify texture (Haralick, 1983). More advanced methods for describing shape (Henderson, 1983) could also improve the rate of correct classification. In addition, the method to separate seeds that touch could be improved. Although the method used in this study is

TABLE 2. Results of classification of 20 seeds per species using the global database that contained all 21 species used in this study.

Seed	Classification		
	Correct	Incorrect	
	#		%
Ryegrass	3	17	15
Rye	16	4	80
Spring Triticale	18	2	90
Kale	18	2	90
Alfalfa	17	3	85
Turnip	15	5	75
Velvetleaf	18	2	90
Ivyleaf morning-glory	20	0	100
Sorghum	19	1	95
Wild Carrot	7	13	35
Giant Foxtail	8	12	40
Poison Hemlock	9	11	45
Broad Leaved dock (perianth)	0	20	0
Yellow Foxtail	11	9	55
Tall Fescue	17	3	85
Johnsongrass	17	3	85
Large Crabgrass	16	4	80
Orchardgrass	16	4	80
Broad Leaved Dock (achene)	8	12	40
White Clover	20	0	100
Wheat	8	12	40
Total	281	139	67

capable of separating two seeds, it may be difficult to adapt to more complex situations. Several other well known techniques exist to differentiate seeds that touch from each other (Martelli, 1976) and the capability of these methods to separate seeds poured on a scanner needs to be tested in future studies.

This study has demonstrated that classifiers that can distinguish between three species at an average rate of 94% can be rapidly developed and applied by seed analysts without any knowledge of artificial intelligence or image processing techniques. It has also demonstrated that the method is not powerful enough to distinguish between a large number of species at an acceptable rate. Future studies need to address how to develop classification methods of practical use to seed analysts that are capable of achieving high rates of accuracy in tests containing a wide variety of species and varieties within a species.

ACKNOWLEDGMENTS

Salaries and research support were provided by state and federal funds appropriated to the Ohio Agricultural Research and Development Center, The Ohio State University.

TABLE 3. Classification results using a database designed to only identify these six seeds.

Seed	Classification		
	Correct	Incorrect	
	#		%
Alfalfa	2	18	10
Turnip	2	18	10
Kale	17	3	85
Red Clover	17	3	85
White Clover	20	0	100
Wild Carrot	3	17	15
Total	61	59	51

TABLE 4. Classification results using the database that contained all 21 seeds in this study.

Seed	Classification		
	Correct	Incorrect	
	#		%
Alfalfa	11	9	55
Turnip	8	12	40
Kale	9	11	4
Red Clover	0	20	0
White Clover	19	1	95
Wild Carrot	2	18	10
Total	49	71	41

REFERENCES

AOSA. 2002. Rules of Testing Seeds. Assoc. of Offic. Seed Anal., Las Cruces, NM, USA.

Chen, C., Y. P. Chiang, and Y. Pomeranz. 1989. Image analysis and characterization of cereal grains with laser range finder and camera contour extractor. Cereal Chem. 66:466–470.

Chtioui, Y., D. Bertarand, Y. Dattee and M. F. Devaux. 1996. Identification of seeds by colour imaging: Comparison of discriminant analysis and artificial neural network. J. Sci. Food Agric. 71:431–441.

Duda, R. O. and P. E. Hart. 1973. Pattern Classification and Scene Analysis. J. Wiley. NY, USA.

Effenberger, J. 2001. Physical Purity Testing. p. 7.1–7.31. *In* M. B. McDonald et al. (ed.) Seed Technologist Training Manual. Soc. Commercial Seed Technol., Ithaca, NY, USA.

Gonzalez, R. and R. Woods. 1992. Digital Image Processing. 2nd ed. Addison-Wesley Publishing, Reading, MA.

Haralick, R. M. 1983. Image Texture Survey in Fundamentals. p. 145–172. *In* O.D. Faugeras (ed.) Computer Vision, Cambridge University Press, Cambridge, UK.

Henderson, T. C. 1983. Feature-Based 2-D shape Models. p. 263–265. *In* O.D. Faugeras (ed.) Computer Vision, Cambridge University Press, Cambridge, UK.

Martelli, A. 1976. An Application of heuristic search methods to edge and contour detection. Commun. Assoc. Computing Machinery 19:73–83.

Neuman, M. R., H. D. Sapirstein, E. Shwedyk, and W. Bushuk. 1989. Wheat grain colour analysis by digital image processing II. Wheat Class Discrimination. J. Cereal Sci. 10:183–188.

Petersen, P., and G. W. Krutz. 1992. Automatic identification of weed seeds by colour machine vision. Seed Sci. Technol. 20:193–208.

Russel, S. and P. Norvig. 1995. Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, NJ.

Sako, Y., M. B. McDonald, K. Fujimura, A. Evans, and M. Bennett. 2001a. A system for automated seed vigour assessment. Seed Sci. Technol. 27:625–636.

Sako, Y., E. Regnier, T. Daoust, K. Fujimura, K. Harrison, and M. B. McDonald. 2001b. Computer image analysis and classification of giant ragweed seeds. Weed Sci. 49:738–745.

Travis, A. J. and S. R. Draper. 1985. A computer based system for the recognition of seed shape. Seed Sci. Technol. 13:813–820.

Zayas, I., F. S. Lai, and Y. Pomeranz. 1986. Discrimination between wheat classes and varieties by image analysis. Cereal Chem. 63:52–56.

Zayas, I., Y. Pomeranz, and F. S. Lai. 1985. Discrimination between Arthur and Arkan wheat by image analysis. Cereal Chem. 62:478–480.

Zayas, I., Y. Pomeranz, and F. S. Lai. 1989. Determination of wheat and nonwheat components in grain samples by image analysis. Cereal Chem. 66:233–237.