Proceedings of The National Conference
On Undergraduate Research (NCUR) 2007
Dominican University of California
San Rafael, California
April 12 – 14, 2007

Characterization of Spatial Patterns in Microphotographies of Leaves Epidermis

Melissa López Serrano
Department of Mathematics
University of Puerto Rico at Humacao
Humacao, Puerto Rico

Faculty Advisors: Elio Ramos, Denny S. Fernández

Abstract

Microscopic images of leaves were analyzed. They were collected from Mona Island dry forest, which is located between Puerto Rico and the Dominican Republic. The images show a wide variety of textures, spatial patterns, cells, and stomata configurations. The main objective of this research is to characterize mathematically the observed textures in the images and to determine features that can be used to classify them. For this purpose, a gray level co-occurrence matrix (GLCM) method was used to analyze the spatial distribution of gray values. From the GLCM matrix several texture features were calculated namely: the angular second moment (ASM), the contrast, correlation, inverse difference moment (ISM), and entropy. Preliminary results indicate the formation of three groups of images based on the features derived from the GLCM analysis.

Keywords: Pattern recognition, image processing

1. Introduction

The characterization and classification of spatial patterns and textures is an important problem in areas like artificial vision and image processing. The natural world provides a wide variety of examples of textures and patterns that can be observed at different spatial scales. From the microscopic cells observed in a plant tissue, to the shape and patterns observed in the leaves, to the distribution of vegetation in a GIS image, different patterns can be observed. The characterization of patterns is extremely important in many areas of biology. This task is the basis for the discrimination between species in areas like taxonomy and fundamental issues like the relation between the observed structure and function. With the widespread availability of digital technology, like cameras, many tasks, like characterization, that used to be handled manually, can be performed in an automatic or semi-automatic way using several statistical and artificial intelligence methodologies. Besides the attractiveness of this possibility the actual implementation of the methods are complicated, mainly due to the heterogeneous quality of the images and noisiness.

In this paper we present the results of an analysis of a sample of images of microphographies of leaves epidermis. The images present a wide variety of textures, spatial patterns, cell structures, and stomata configurations. The main objective of the research is the development of a method to characterize mathematically and statistically the observed patterns in such a way that the original group can be divided in sub-samples with similar features. In the next section we present some basic biological background concerning the source and the basic structures observed in the images. Then we describe the Gray Level Co-occurrence Matrix (GLCM) method that was used to obtain the texture features that were used to characterize the images. Then we present a description of the Principal Component Analysis (PCA) and how it was used to construct sub-samples based on the features obtained from the GLCM analysis. Finally we present the results of both analyses.

2. Data Set

The data set consisted of 151 images of leaves epidermis of 1600x1200 pixels at 200x magnifications. The leaves were collected from the Mona Island dry forest, which is located between Puerto Rico and the Dominican Republic. The epidermis is the outermost cellular layer that covers the whole plant structure and typically can be observed as a set of closely packed cells without intercellular spaces[4]. Besides the epidermal cells a prominent structure known as the stomata can be observed. The stomata are basically a pore surrounded by two bean shaped cells known as the guard cells (Figure 1). The epidermis has many functions being the most important to allow the sunlight to pass through the chloroplasts which is crucial for the photosynthesis process and to avoid an excessive lost of water from the inner tissues. The stomata allow the gas exchange between the plant and the environment which again is necessary for photosynthesis and respiration. For different plant species a wide variety of patterns of cell epidermis and stomata configurations can be observed. In this sense the observed structures in the images can be used as a discriminator between species or group of species. Traditionally this type of task is carried out manually by visual inspection of the images and the corresponding classification. In this paper an automatic procedure is presented that is able to measure some features from the raw images, followed by a method that allows the creation of groups based on the features.

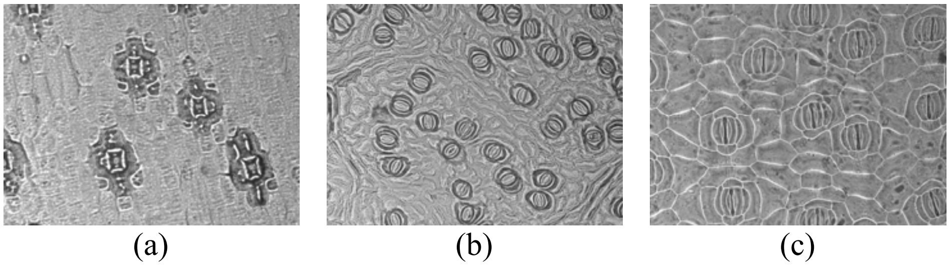

(a) (b) (c)

Figure 1: Some examples of the images that were analyzed. The observed bean shaped structures are the stomata.

3. Data Analysis

Starting with the raw images the data analysis procedure consists of several steps which are described in the following sections.

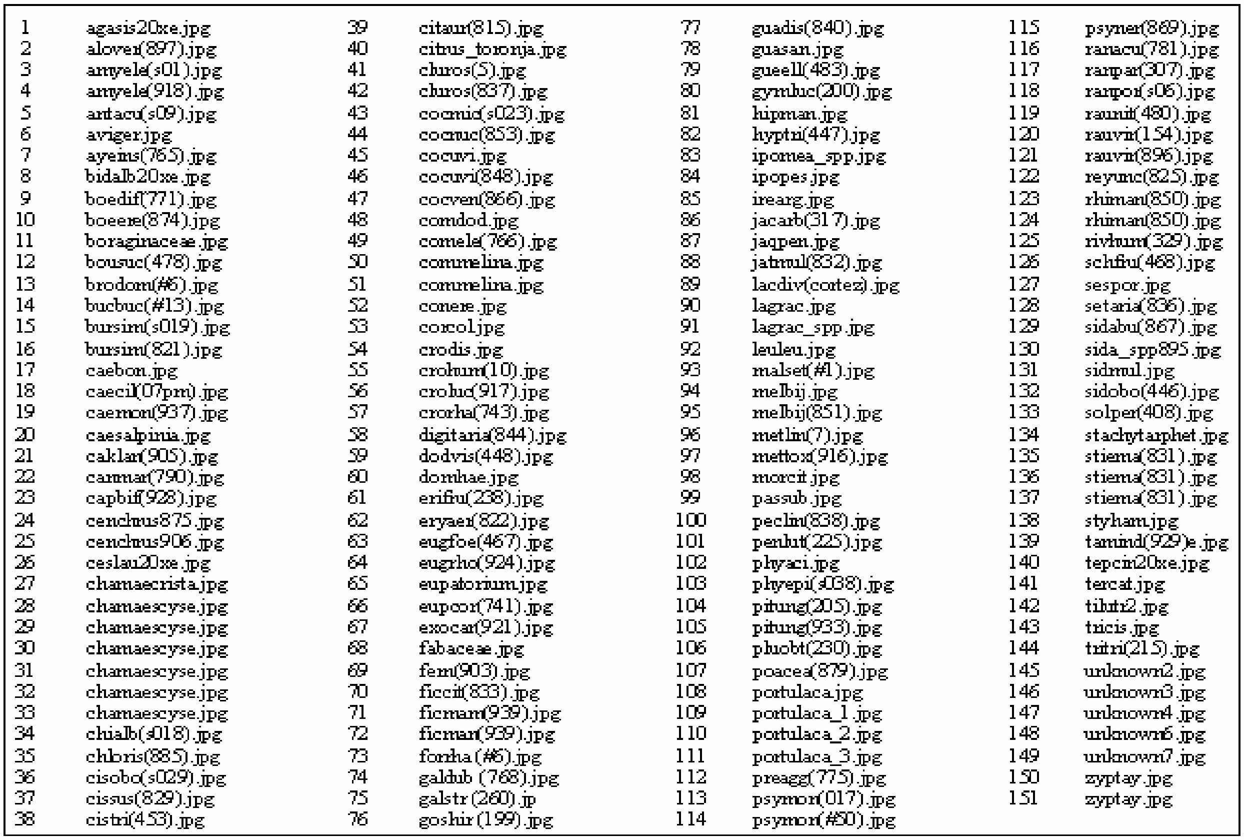
3.1 The Gray Scale Co-correspondence Matrix

Several statistical approaches can be considered to characterize the distribution of pixels in an image. First-order statistical approaches construct a histogram for the distribution of gray scales in the images. From this histogram several parameters can be derived e.g. mean, variance, skewness, and kurtosis. For a given image, and pixels distributions, this first order statistics gives information about the amount of symmetry (or asymmetry) in the image. In a second-order statistical approach an intermediate set of parameters are calculated and from this parameters the second-order statistics are derived. The Gray Scale Co-occurrence Matrix (GLCM) method is a second-order statistical method developed and improved by Haralick in 1979[2, 6]. The GLCM method considers the spatial relationship between pixels of different gray levels. The method calculates a GLCM by calculating how often a pixel with a certain intensity i occurs in relation with another pixel j at a certain distance d and orientation θ. For instance, if the value of a pixel is 1 the method looks, for instance, the number of times this pixel has a pixel 2 in the right side. Each element (i,j) in the GLCM is the sum of the number of times that the pixel with value i occurred in the specified relationship to a pixel with value j in the raw image. Once the GLCM is calculated several second-order texture statistics can be computed as is illustrated in Table 1 where $P_{d,\theta}(i,j)$ is the GLCM between pixels i and j.

Table 1: Texture features utilized in the GLCM analysis of the images with the corresponding formula[2]

Texture feature	Formula	Texture feature	Formula
Absolute value	$\sum_{i,j=0}^{G-1} \mid i-j \mid P_{d,\theta}(i,j)$	Entropy	$\sum_{i,j=0}^{G-1} P_{d,\theta}(i,j)\log_2[P_{d,\theta}(i,j)]$
Inverse difference	$\sum_{i,j=0}^{G-1} \frac{P_{d,\theta}(i,j)}{1+(i-j)^2}$	Correlation	$\frac{\sum_{i,j=0}^{G-1} ijP_{d,\theta}(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$
Homogeneity	$\sum_{i,j=0}^{G-1} \frac{P_{d,\theta}(i,j)}{1+\lvert i-j\rvert}$	Energy	$\sum_{i,j=0}^{G-1} (P_{d,\theta}(i,j))^2$

The GLCM and the corresponding features from Table 1 were calculated for each of the 151 images. The GLCM method was implemented as a plugin with the ImageJ[1] public domain Java image processing software. The plugin, `Headless_GLCM.java` was developed in Java and based on the texture analysis plugin developed by Julio E. Cabrera from NIH[4]. The original texture analysis plugin source was extended by adding some additional features and modified to handle a large amount of images by running ImageJ in a “headless” mode. The GLCM analysis of the images resulted in a *features table* of 151 rows corresponding to each image and 7 columns corresponding to the “features” derived from the GLCM. Each image in the features table was identified by an integer number between 1 and 151. The correspondence between the image number and the plant specie (from visual identification) are given in the Figure 2.



1	agasis20xe.jpg	39	citaur(815).jpg	77	guadis(840).jpg	115	psyner(869).jpg
2	alover(897).jpg	40	citrus_toronja.jpg	78	guasan.jpg	116	ranacu(781).jpg
3	amyele(s01).jpg	41	churos(5).jpg	79	gueell(483).jpg	117	ranpar(307).jpg
4	amyele(918).jpg	42	churos(837).jpg	80	gymluc(200).jpg	118	ranpor(s06).jpg
5	antacu(s09).jpg	43	cocmic(s023).jpg	81	hipman.jpg	119	raunit(480).jpg
6	aviger.jpg	44	cocnuc(853).jpg	82	hyptri(447).jpg	120	rauvir(154).jpg
7	ayeins(765).jpg	45	cocuvi.jpg	83	ipomea_spp.jpg	121	rauvir(896).jpg
8	bidalb20xe.jpg	46	cocuvi(848).jpg	84	ipopes.jpg	122	reyunc(825).jpg
9	boedif(771).jpg	47	cocven(866).jpg	85	irearg.jpg	123	rhiman(850).jpg
10	boeere(874).jpg	48	comdod.jpg	86	jacarb(317).jpg	124	rhiman(850).jpg
11	boraginaceae.jpg	49	comele(766).jpg	87	jaqpen.jpg	125	rivhum(329).jpg
12	bousuc(478).jpg	50	commelina.jpg	88	jatmul(832).jpg	126	schfiu(468).jpg
13	brodom(#6).jpg	51	commelina.jpg	89	lacdiv(cortez).jpg	127	sespor.jpg
14	bucbuc(#13).jpg	52	conere.jpg	90	lagrac.jpg	128	setaria(836).jpg
15	bursim(s019).jpg	53	corcol.jpg	91	lagrac_spp.jpg	129	sidabu(867).jpg
16	bursim(821).jpg	54	crodis.jpg	92	leuleu.jpg	130	sida_spp895.jpg
17	caebon.jpg	55	crohum(10).jpg	93	malset(#1).jpg	131	sidmul.jpg
18	caecil(07pm).jpg	56	croluc(917).jpg	94	melbij.jpg	132	sidobo(446).jpg
19	caemon(937).jpg	57	crorha(743).jpg	95	melbij(851).jpg	133	solper(408).jpg
20	caesalpinia.jpg	58	digitaria(844).jpg	96	metlin(7).jpg	134	stachytarphet.jpg
21	caklan(905).jpg	59	dodvis(448).jpg	97	mettox(916).jpg	135	stiema(831).jpg
22	canmar(790).jpg	60	domhae.jpg	98	morcit.jpg	136	stiema(831).jpg
23	capbif(928).jpg	61	erifiu(238).jpg	99	passub.jpg	137	stiema(831).jpg
24	cenchrus875.jpg	62	eryaer(822).jpg	100	peclin(838).jpg	138	styham.jpg
25	cenchrus906.jpg	63	eugfoe(467).jpg	101	penhut(225).jpg	139	tamind(929)e.jpg
26	ceslau20xe.jpg	64	eugrho(924).jpg	102	phyaci.jpg	140	tepcin20xe.jpg
27	chamaecrista.jpg	65	eupatorium.jpg	103	phyepi(s038).jpg	141	tercat.jpg
28	chamaescyse.jpg	66	eupcor(741).jpg	104	pitung(205).jpg	142	tihtr2.jpg
29	chamaescyse.jpg	67	exocar(921).jpg	105	pitung(933).jpg	143	tricis.jpg
30	chamaescyse.jpg	68	fabaceae.jpg	106	pluobt(230).jpg	144	tritri(215).jpg
31	chamaescyse.jpg	69	fern(903).jpg	107	poacea(879).jpg	145	unknown2.jpg
32	chamaescyse.jpg	70	ficcit(833).jpg	108	portulaca.jpg	146	unknown3.jpg
33	chamaescyse.jpg	71	ficmam(939).jpg	109	portulaca_1.jpg	147	unknown4.jpg
34	chialb(s018).jpg	72	ficmar(939).jpg	110	portulaca_2.jpg	148	unknown6.jpg
35	chloris(885).jpg	73	forrha (#6).jpg	111	portulaca_3.jpg	149	unknown7.jpg
36	cisobo(s029).jpg	74	galdub (768).jpg	112	preagg(775).jpg	150	zyptay.jpg
37	cissus(829).jpg	75	galstr (260).jp	113	psymon(017).jpg	151	zyptay.jpg
38	cistri(453).jpg	76	goshir (199).jpg	114	psymon(#50).jpg		

Figure 2: List of the 151 images and its corresponding numbers. The name of the files correspond to the abbreviated scientific name of the identified species.

3.2 Preliminary Analysis

Once the features table was constructed several elementary visualization techniques were considered in order to understand the relationships between the features. First, we constructed a scatter plot, better known as pairs plot[5], that relates all of the features against each other as is shown in Figure 3. In this plot we observed the strong correlation between some features like the homogeneity (hom) and the inverse difference (IND) , and between contrast and the absolute value (absval). Also, we noticed the formation of groups (clusters) by considering pairs of features like the correlation and the inverse difference and between the correlation and the homogeneity. In this sense the amount of dispersion between the variables was a good indicator of the discriminatory capability of different pairs of features. Furthermore, we constructed a stars plot (Figure 4) for each of the images. The stars plot is a multivariate visualization technique[5] where each observation is represented as a star-shape structure. In each observation the length of the ray is proportional to the size of the variables (features) associated with each observation (image). From the stars plot we can obtain interesting and valuable information about the differences and similarities between the images. From the stars plot in Figure 4 we can observe several families based on the shapes. For instance we can observe that several groups of images (e.g. 31, 32, 42,54, and 56) are characterized by the "airplane" pattern meanwhile other images (e.g. 18, 30, 101, and 102) are characterized by the "cobweb" pattern. Also, we can deduce a small amount of images (e.g. 12, 83, 110, and 11) belonging to the "butterfly" pattern. This type of qualitative characterization is important and will be contrasted with other quantitative techniques in the next sections.

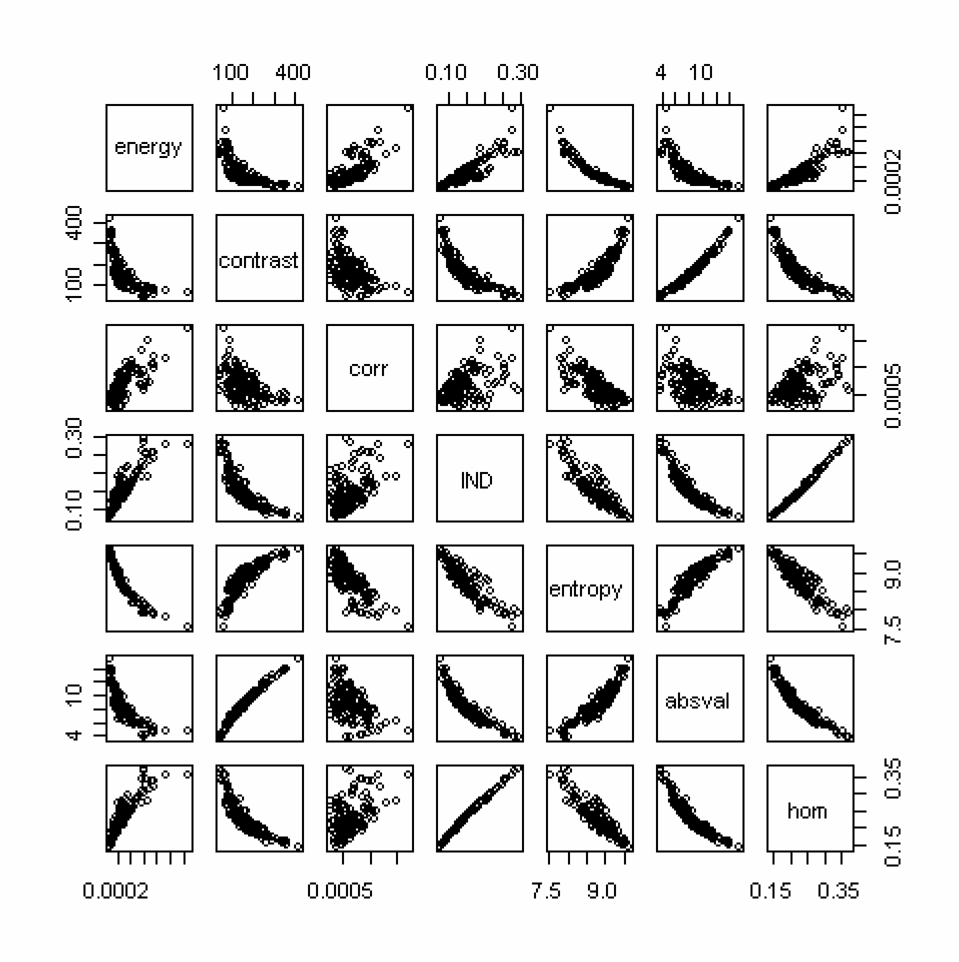


Figure 3: Pairs plot from the GLCM derived features from the images.

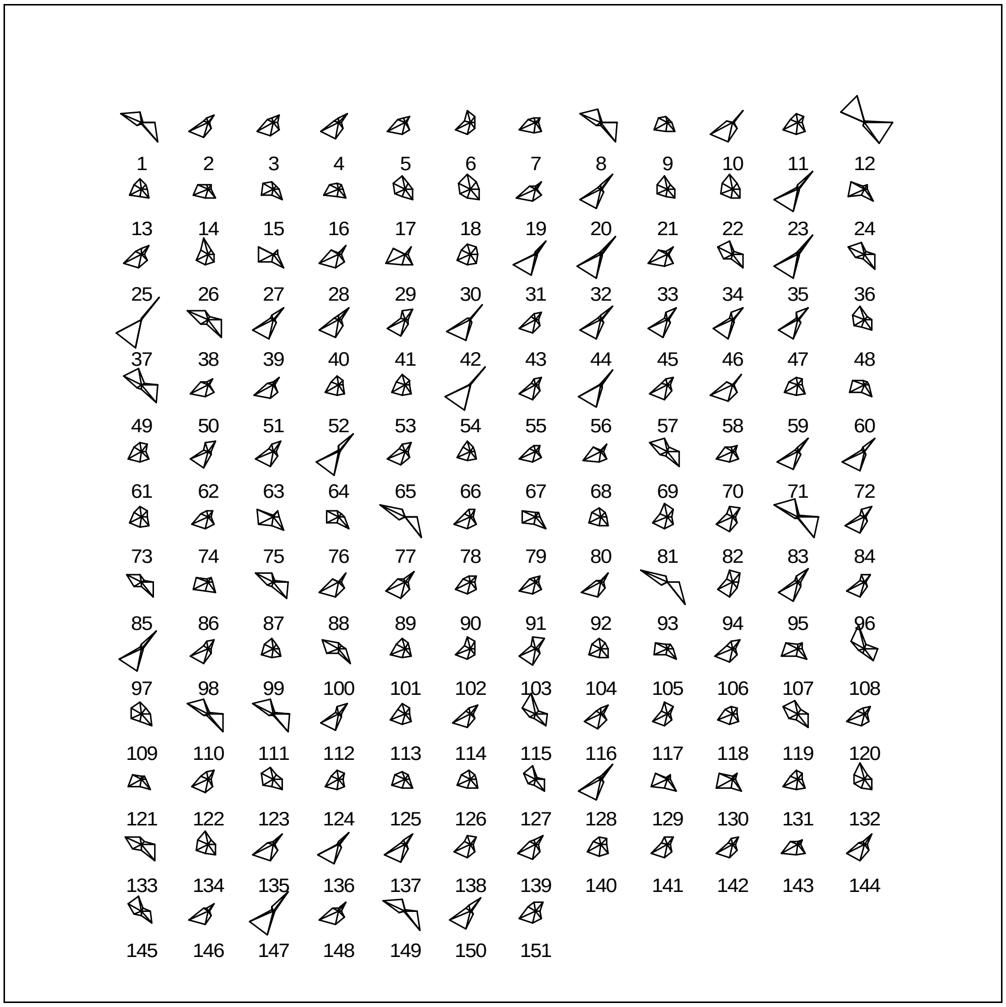


Figure 4: Stars plot from the GLCM derived features for each image.

3.3 Principal Component Analysis

The Principal Component Analysis (PCA) is one of the main tools of exploratory multivariate data analysis. A very common situation in multivariate data analysis, like in the features table, is the availability of several variables for a single observation. In this sense each observation is a point in a multidimensional space. However, in many cases multidimensional points are very difficult to visualize and consequently hard to identify patterns. The PCA method is a technique to reduce the dimension of a multidimensional space by constructing new variables (components) consisting of linear combinations of the original variables with the highest variance. The result of a PCA is a set of components where the first component has the highest variance, the second component the second highest variance, etc. In this sense the complex inter-relationships between many variables can be reduced to a much simpler two dimensional plot relating the first and second components. The PCA is one of the standard and most powerful techniques of multivariate data analysis and more detailed descriptions of the method can be found elsewhere[7, 8].

The PCA was implemented using the correlation matrix method and the calculation of the eigenvalues and eigenvectors. The results indicate that the first two component account for most of the variability observed in the data set. Examination of the plot in Figure 5 reveals the formation of several groups of images. The location of image 12 as an isolated singleton is not surprising by examination of the unusual "butterfly" shape in the star plot with share some of its features with image 83, both of them very close in the PCA plot. Examination of other groups in the PCA plot reveals a strong correspondence between the shapes in the star plot and the closeness in the PCA plot.

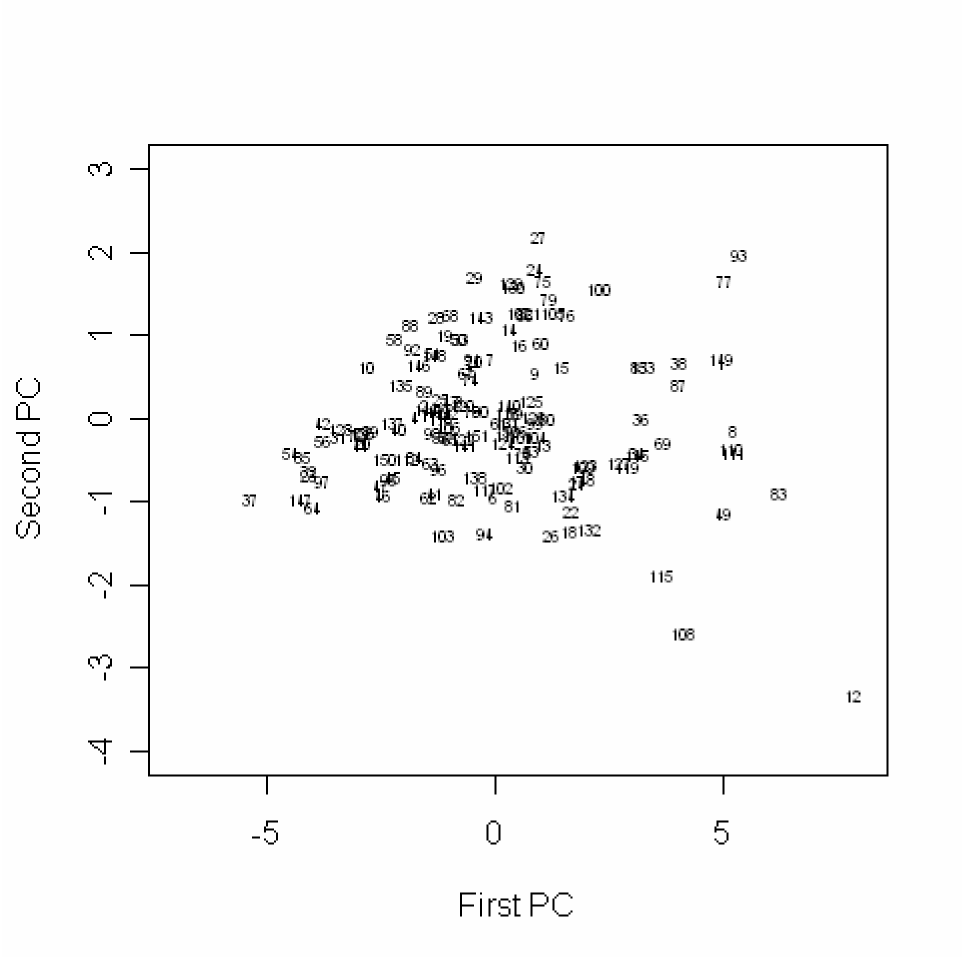
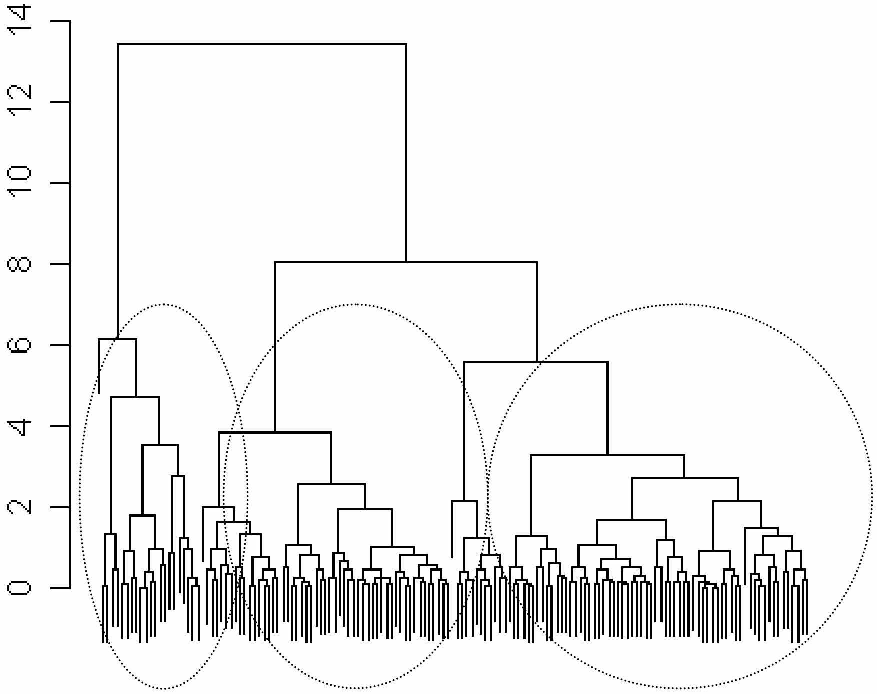


Figure 5: Principal Component Analysis (PCA) Plot.

3.4 Cluster Analysis

Cluster analysis is a general term for a set of algorithms and methods that construct groups based on similar features or categories[8]. From the PCA analysis in the previous section we were able to "qualitatively" identify groups of images by closeness of points in Figure 6. The cluster analysis was implemented by measuring the Euclidean distance between data points in Figure 6 and a hierarchical clustering method. From the distance measurements a *dendogram* plot can be constructed in which the similarity between two groups can be obtained by the height in which they join a single group. The dendogram for the images is shown in Figure 6. Besides the lack of resolution in the images numbers (in the lower part) three relatively large clusters can be identified below height 4 which are summarized in Table 2. The dendogram confirms the results from the stars plots and the PCA by locating image 12 as an isolated singleton near height 5. Furthermore, the dendogram was able to locate, in the same branch, several images sequences that effectively are from the same species (e.g. 111 and 111, 129 and 130). On the other hand, several sequences of images from the same species like 27, 28, 29, 30, 31, 32 , 33 reveals some of the complexities involved in this type of analysis. In this case images 27 and 29 were located by the dendogram in the left cluster, and images 28, 20, 31, 32, and 33 were located (correctly) in the right cluster. However, an examination of the raw images reveals dramatic differences in the visual appearance of the images mainly due to differences in the qualities of the photographed samples.

(a)



(b)

Cluster	Number of Images	Images
Left cluster	21	1 149 77 93 85 133 38 87 34 145 119 127 36 69 108 115 83 49 8 110 111
Central cluster	53	100 76 79 105 15 9 60 27 24 75 129 130 107 86 121 14 16 26 134 17 21 48 109 123 132 18 22 117 138 94 81 6 102 59 118 140 11 124 61 52 131 80 125 99 126 13 53 104 66 101 30 73 113
Right cluster	76	37 31 128 42 56 54 35 23 32 97 64 147 7 67 74 5 70 91 29 143 33 50 68 19 28 78 90 151 122 141 3 120 25 43 142 106 57 65 55 98 114 116 139 82 103 84 112 63 96 41 62 39 72 71 136 20 44 46 47 150 45 95 10 4 89 2 144 135 40 137 58 88 51 148 92 146

Figure 6: (a) Dendogram from the cluster analysis of the PCA plot and the main clusters (b) The clusters and the images contained in each of them.

4. Results and Conclusions

A multivariate data analysis was carried out on images of microphographies of leaves epidermis in search of patterns that allow an automatic or semi-automatic grouping or characterization of the images. In the first stage of the analysis seven textural features were estimated based on the Gray Scale Co-occurrence Matrix (GLCM) namely the absolute value, inverse difference, homogeneity, entropy, correlation, and energy. The pairs and stars plot visualization method were applied to the features data and preliminary groups of images were identified. A Principal Component Analysis (PCA) was carried out on the features data and several groups of images were identified. From the PCA data a cluster analysis revealed the formation of three large groups of images each one with its internal hierarchy of smaller clusters. Examination of the raw images for each cluster reveals a remarkable visual consistency between the groups and the images; however, in some cases the noisiness and bad quality of the images hamper or complicate the discrimination. In the future we will compare our results with an independent study carried out using manual and visual classification, and probably a hybrid procedure will be implemented for the recognition of the species using epidermal tissue.

5. Acknowledgments

This research has been sponsored by the National Security Agency (NSA) under the proposal #1498230-C-0486 and the NSF-CSEMS proposal of the Department of Mathematics at the UPRH. We would like to thank Dr. Elvia Meléndez-Ackerman from CREST-Center for Applied Tropical Ecology and Conservation (CATEC) for providing the images utilized in the analysis.

6. References

1. Abramoff, M.D., Magelhaes, P.J., Ram, S.J. "Image Processing with ImageJ", 2004.
2. Bharath Kumar, S.V. et al, Proc. of International Conference on Signal Processing and Communication, 2004
   Biophotonics International, volume 11, issue 7, pp. 36-42, 2004.
3. Cabrera J., Texture Analyzer, http//rbs.info.nih.gov/ij/plugins/texture.html 2005
4. Carpenter, K.J., *Am. J. Botany*, 92, 2005
5. Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. Graphical Methods for Data Analysis, Belmont, CA: Wadsworth. 1983.
6. Haralick, R.M., Proceedings of the IEEE, 67, 1979.
7. Smith, L.I., A tutorial on Principal Component Analysis, http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf 2002
8. Venaples, W.N., Ripley, B.D., Modern Applied Statistics with S, Springer-Verlag, New York, 2002.