UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 09-01
Statistics and Econometrics Series 01
January 2009

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

CLUSTERING AND CLASSIFYING IMAGES WITH LOCAL AND GLOBAL VARIABILITY

Andrea Giuliodori, Rosa Lillo and Daniel Peña

Abstract

A procedure for clustering and classifying images determined by three classification variables is presented. A measure of global variability based on the singular value decomposition of the image matrices, and two average measures of local variability based on spatial correlation and spatial changes. The performance of the procedure is compared using three different databases.

Keywords: images, cluster, classification

Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail addresses: mariaandrea.giuliodori@uc3m.es (Andrea Giuliodori), rosaelvira.lillo@uc3m.es (Rosa Lillo) and daniel.pena@uc3m.es (Daniel Peña). This research acknowledges the support of "Comunidad de Madrid" grant CCG07-UC3M/HUM-3260.

CLUSTERING AND CLASSIFYING IMAGES WITH LOCAL AND GLOBAL VARIABILITY

ANDREA GIULIODORI, ROSA LILLO AND DANIEL PEÑA

Abstract. A procedure for clustering and classifying images determined by three classification variables is presented. A measure of global variability based on the singular value decomposition of the image matrices, and two average measures of local variability based on spatial correlation and spatial changes. The performance of the procedure is compared using three different databases.

1. Introduction

One of the goals in exploratory image analysis is to classify images. We are interested in finding characteristic variables in an image which are fast to compute and can be used for classification and clustering. Some classification techniques used in this area include the linear discriminant functions proposed by Liu et al. (1993), the nearest neighbors method applied by Hastie and Simard (1998), the support vector machine proposed by Vapnik (1995) and improved by Marron and Todd (2002), and principal component analysis proposed by Turk and Pentland (1991). Additionally, other interesting studies in this field are those by Hastie and Tibshirani (1996), Vailaya et al. (1998), Le Cun Y. (1995), Vailaya et al. (1998) and Herwig et al. (1999). The purpose of this paper is to propose a new way to find distinctive features in the image to build clusters. We introduce three classification variables. The first one is a measure of the global variability, based on the singular value decomposition of the image as a matrix, the second and third are average measures of local variability, one based on spatial correlation and the other on spatial changes.

A color digital image can be represented by 3 matrices of N rows and M columns. These matrices represent the three primary colors produced by the light, which are red (R), green (G) and blue (B). Each matrix depicts one dimension of a picture and consists of elements x_{ij}, with $i = 1, ..., N$ and $j = 1, ..., M$, that represent color intensity of a pixel (picture element) extracted from digitized images. Therefore, the color of a pixel is the combination of the three elements that constitute the vector $[R(x, y), G(x, y), B(x, y)]$. Then, each pixel can be defined as a function $f(x, y) \epsilon R^3$. All the elements x_{ij} are in the range [0,255], where the value 0 represents the black color and the value 255 the white color. In order to facilitate data manipulation, this range is turned into the interval [0,1].

A preliminary visual inspection of an image provides an idea of its color variability. For instance, a picture of the blue sky has low variability; conversely an

Key words and phrases. **images, cluster, classification**.

image of a rainbow has higher variability in its structure. This variation in colors can be used as a criterion to classify images. The forgoing representation of an image allows to calculate different correlation and variability measures in a particular matrix.

The paper is organized as follows. In the next section we review briefly the statistical structure of an image. Section 3 introduces the variables we propose as summaries of the image, and comment on their properties and expected discriminant power. Section 4 presents the experimental results of the application of classification and grouping techniques through two illustrative examples. Section 5 extend these results to other data sets. Finally, Section 6 concludes and proposes some future research.

2. Classification variables

Classification represent one of the most important applications in Digital Images Processing which require the definition of discriminant variables. In this paper we propose to use three variables which describe the variability and spatial correlation.

The first variable is the local variability (δ), performed by Benito and Peña (2004). The $\delta(X)$ is a smoothing measure that represents the spatial dependence within pixels. The measure is obtained through the bi-dimensional derivative and is defined as follows. Given a pixel x_{ij}, the derivative of the intensity in this point is equal to

$$\nabla_{ij}(X) = x_{i+1,j+1} - x_{i+1,j} - x_{i,j+1} + x_{i,j}$$

and the local variability is given by

$$\delta(X) = \frac{1}{\tilde{d}} \sum_{n-1}^{m-1} \sum_{j=1}^{J-1} |\nabla_{ij}(X)| \tag{1}$$

where $\tilde{d} = (I-1) \times (J-1)$ and the notation $| \bullet |$ represents the absolute value. The coefficient δ assumes a value equal to 2 for the most complex image, since the maximum value of $\nabla_{ij}(X)$ is 2 when x_{ij} is standardized values of pixels. Images with variability in colors show high local variability.

The second variable is the effective variance denoted as $V_e(X)$. This concept is introduced by Peña and Rodriguez (2003) to compare the variability of sets of variables with different dimensions. The measure can be defined as follows. Let X be a matrix of $n \times m$, then V_e is given by

$$V_e(X) = (\phi_1 \phi_2 .... \phi_{r_e})^{1/r_e} \tag{2}$$

where ϕ_i are the singular values of the matrix $X^T X$ or $X X^T$. Besides, the value of r_e denote the effective range of X and represents the number of eigenvalues for which a relative error is less than ϵ; that is,

$$r_e = \{\sharp\phi / \frac{||X - \hat{X}_k||}{||X||} < \epsilon\}$$

where $\hat{X}_k$ is given by the singular decomposition of X, using only the k greatest singular values and theirs respective eigenvectors

$$\hat{X}_k = V_k D_k^{1/2} U_k^T$$

where V and U are orthogonal matrices and theirs columns are the eigenvectors of matrices XX^T and X^TX, respectively. D is a diagonal matrix and its elements are the singular values of XX^T or X^TX. Then,

$$||X - \hat{X}_k|| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{ij})^2} \tag{3}$$

Solving (3), the final value of k will be the *effective range* r_e. To compute $V_e(X)$ we include the greater singular values [1] for which the relative error are less than a sufficiently small value denoted as ϵ. As a result, of applying this measure to images, we observe that pictures with variability in colors show high effective variance.

The third variable to be considered is the correlation (ρ_h) between one pixel and its neighborhood located to a **h** distance, letting h=1,....15. The values of h were defined in accordance with the size of images. For greater values of h we would have a large proportion of zeros when calculating the correlation for edge pixels, in small-sized images. The calculation was done for each pixel, taking into account the ordination per row. In the simplest version, ρ is defined by:

$$\rho_h = \frac{\sum_i^n \sum_j^m \sum_{neig_h}(X_{ij} - \bar{X})(X_{neig_h} - \bar{X})}{\sum_i^n \sum_j^m \sum_{neig_h}(X_{ij} - \bar{X})^2} \tag{4}$$

where X_{ij} is the element ij of the correspond matrix in the selected image. Its value is between [0,1]. $\bar{X}$ is the total mean of all pixels in the matrix X. X_{neig_h} corresponds to the value of the pixel located to distance h of the element X_{ij} (neighbor h). The values n and m are the number of rows and column respectively, in the matrix X. And h is the order of the spatial correlation, that is the distance between X_{ij} and its neighbors.

In all cases the correlation decrease as we move away from the central pixel. Some images present a high correlation at the first order and decrease slowly along the h-order. Others show a smaller spatial correlation at the first order ($h = 1$) and even smaller value at greater order of h. As a generalization, each image has a similar spatial correlation structure for the three RGB matrices.

3. Methodology

In order to illustrate the classification power of the variables proposed in Section 2, we apply classification and grouping methods to three different datasets. The first one is a subset of 379 color images, already analyzed in Wang et al. (2001). To classify, we select images of same size (128×96) in order to avoid possible distortions in the selected measures. The set contains different types of pictures such as foods,

[1] The singular value is the square root of the eigenvalue. $\phi_i = \sqrt{\lambda_i}$.

buildings, monuments and landscapes (see Appendix 7.1). To simplify, this dataset is called as *WWL*.

The second dataset is collected from "google images" section. We collect 379 heterogeneous google color images (landscapes, ID cards, babies, animals, paintings, and the like) with approximately the same size than the previous one, i.e. (128×96). This dataset is denoted as *GOOGLE* (see Appendix 7.2).

The third database is a set of digits that were clipped from images of handwritten ZIPCODES. The original dataset was treated and improved by Bottou et al. (1994) and Le Cun Y. (1995), removing the unrecognizable digits from the set. Our subset is comprised for 7291 digits from 0 to 9 (10 classes), represented as a graycolor image of 16×16 dimension (see Appendix 7.3). We labelled this database as *ZIPCODES*.

Considering the measures proposed in Section 2, we obtain the following information for each image. First, we calculate the Local Variability (δ). Calculations are done for each RGB matrix obtaining one vector of three elements per variable for every image. In symbols, the vector is: $\delta = [\delta_R, \delta_G, \delta_B]$.

To calculate the Effective Variance (V_e) we use matrix $Xdev$, instead of the original matrix X, as the deviation of the mean per row (or column, the smaller) in order to remove the mean effect. That is, $Xdev_i = X_i - Xmean$ where the resulting values of $Xdev$ correspond to each value of the matrix X minus the mean of its row (or column). Besides we choose a norm equal to $\sqrt{\sum(diag(X'X))}$, $\epsilon = 0.01$ and the initial value of $k = 0.3 \times [\min(n, m)]$. Those values were chosen in order to include in the calculation the greater non-null eigenvalues. Calculations are done for each RGB matrix, obtaining one vector of three elements per variable for every image. In symbols, $V_e = [V_{eR}, V_{eG}, V_{eB}]$.

The Spatial Correlation (ρ) is obtained for each RGB matrix, taking the h-order from 1 to 15. As result, we have a vector of 15 elements ($h = 1, ..., 15$) for each RGB matrix of every image. That is,

$$\begin{aligned} \rho_R &= [\rho_{R1}, \rho_{R2}, ..., \rho_{R15}] \\ \rho_G &= [\rho_{G1}, \rho_{G2}, ..., \rho_{G15}] \\ \rho_B &= [\rho_{B1}, \rho_{B2}, ..., \rho_{B15}] \end{aligned}$$

We calculate and analyze the selected variables in all databases, observing the same characteristics in the three. However, the following results correspond to *WWL* one. There exists a high correlation among ρ_i of all orders given by the presence of spatial dependence. The application of principal component analysis technique (PCA) shows redundancy among ρ of different orders. The variability in the set is explained largely by ρ_1, with minimum contribution of spatial correlation of order 2 to 15. Consequently, superior orders of ρ are excluded from the analysis. Moreover,the analysis of variance to contrast the equality of means among ρ_{R1}, ρ_{G1} and ρ_{B1} indicates that at least one ρ_{i1} mean is significative different to the other (see Appendix Table A.2). Hence, we propose to use the Mean Spatial Correlation of order 1 $\bar{\rho_1}$ and $\Delta\rho_1$ as variables to be included in classification and grouping images. Where,

$$\begin{aligned} \bar{\rho_1} &= (\rho_{R1} + \rho_{G1} + \rho_{B1})/3 \\ \Delta\rho_1 &= \max(\rho_{R1}, \rho_{G1}, \rho_{B1}) - \min(\rho_{R1}, \rho_{G1}, \rho_{B1}) \end{aligned}$$

By the analysis of variance, V_e and δ evidence no meaningful difference among RGB matrices (see Appendix Table A.2). Therefore, we use the average of each variable to make the classification; that is, Average Local Variability ($\bar{\delta}$) and Average Effective Variance ($\bar{V_e}$). Where,

$$\bar{\delta} = (\delta_R + \delta_G + \delta_B)/3$$
$$\bar{V_e} = (V_{eR} + V_{eG} + V_{eB})/3$$

The previous variables will be used in cluster procedure to the whole image. However, in classification we also obtain the variables to pieces of images. This technique will be explained in Section 5.

4. Clustering images

The method consists on the classification of objects into different groups through the partitioning of a data set into subsets (clusters). The data in each subset share some common trait, often proximity, according to some defined distance measure. The exploratory analysis carried out in the previous section suggests a way to build groups, due to the discriminant power the analyzed variables exhibit in the set (see Fraiman et al. (200)). In cluster formation we use the *WWL* database of color images. Based on classification variables and because of the high dimension of data, we performed the "k-means" algorithm method to form the groups. The calculated measures suggest a way to split up groups of pictures with homogeneity in color (small variability and high correlation) from others. By visual inspection of *WWL* dataset we can deduce that there are approximately three groups of pictures clearly defined: 213 landscapes, 41 buildings and 125 foods. Then, in order to apply the k-means algorithm we initially choose an arbitrary number of groups, i.e. $k = 3$, trying to identify the exact number which shows differences between images.

In order to check the results obtained in Section 2, at the beginning we try to group, by considering all variables belonging to each RGB matrix. We confirm empirically the fact that there exists redundancy among V_{ei}, δ_i and ρ_i for each RGB matrices. Then, the variables used to perform cluster technique are the $\bar{V_e}$, the $\bar{\delta}$, the $\bar{\rho_1}$ and the $\Delta\rho_1$. The classification begins considering the fourth selected variables and forming 3 groups. The results suggest that we have 2 mains groups and another one comprised of only fifteen pictures which, according to the variables, do not constitute a cluster by itself. Then, we consider $k = 2$ and the resulting groups are well differentiated. One of them, identified as *cluster 1*, mainly constituted by landscapes. The other one, identified as *cluster 2* is constituted by the rest of pictures. Interpreting the outcomes we can note in the groups the following features: cluster 1: broadly speaking, landscapes have more homogeneity in color. Consequently, they show a high spatial correlation ($\bar{\rho_1}$ and $\Delta\rho_1$), a low average local variability ($\bar{\delta}$) and a low average effective variance ($\bar{V_e}$), and Cluster 2: high ($\bar{\delta}$) and ($\bar{V_e}$). Low $\bar{\rho_1}$ and $\Delta\rho_1$. However, according to visual inspection, there are some pictures that do not belong to the assigned cluster. Out of 214 landscape pictures, nine of them were grouped in the class mainly formed by foods and buildings (cluster 2). Moreover, out of 165 pictures of foods and buildings, eight were grouped with the landscapes (cluster 1). Images wrongly grouped with landscapes have similar characteristics than them, that are uniformity of colors and high correlations. In

TABLE 1. k-means algorithm- 2 and 3 groups

	number of cases for k=3	number of cases for k=2
cluster 1	200	214
cluster 2	164	165
cluster 3	15	
N	379	379

contrast, images wrongly grouped with foods and buildings present great variability in color intensities, may be not perceived by simple visual inspection (see Appendix 7.4).

We are also interested in analyzing the results in a gray color image to contrast them with RGB results. Then, the gray transformation is applied to the *WWL* color image database. The conversion from RGB to gray is established by the international norm for digital TV (CCIR-601). Each gray pixel can be obtained from RGB pixels through the equation below:

$$GRAYpixel_i = 0,299 \cdot R_i + 0,587 \cdot G_i + 0,114 \cdot B_i$$

where R_i, G_i and B_i are the values of the pixel i in red, green and blue matrices respectively. Given that our interest is to compare the grouping results between RGB and Gray images, the methodology applied is the same in both cases. However, a gray image has only one matrix instead of the three RGB. Then, the resulting variables are the Effective Variance (V_e), Local Variability (δ) and Spatial Correlation of order 1 (ρ_1). Following the methodology used with color images, we try the k-means algorithm for 2 groups. The clusters formed have similar characteristics than in RGB case. Cluster 1 mainly comprised by landscape has 216 pictures. The other one has 163 pictures. The solution includes ten landscape pictures wrongly grouped in cluster 2 and six foods and buildings images wrongly grouped in landscapes cluster (See Appendix 7.5).

TABLE 2. Percentage of success in Cluster Analysis: *WWL* database

	RGB images for k=3-	GRAY images for k=2-
cluster 1	95.79	95.37
cluster 2	95.15	96.31

The Table above shows the percentages of images well grouped in their belonging cluster.

5. CLASSIFYING IMAGES

Discriminant analysis is a technique for classifying a set of observations into predefined classes. The purpose is to determine the class of a new observation based on a set of variables known as predictors or input variables belongings to the

known groups. Lets $x_1, x_2, ..., x_n$ a set of elements (images), and $c_1, c_2, ...c_k$: "k" the classes or population in which the images are classified. Then, the probability to classify a new observation is given by:

$$p(c_k|x_{n+1})$$

We apply a linear discriminant technique to *WWL* database in RGB and GRAY colors. The procedure consists on dividing the database in two classes according to visual observation of pictures. *Group 1* is comprised by landscapes and *group 2* consisted of foods and buildings. The variables obtained in Section 3 are considered to make the classification in RGB images, i.e., $\bar{V}_e$, $\bar{\delta}$, $\bar{\rho_1}$ and $\Delta\rho_1$. Then, it is fitted a multivariate normal density to each group, with a pooled estimate of covariance. The classification procedure begins by splitting up one picture from the rest of the database. Then, ignoring its membership group, it is classified in one of the existing classes, according to the probability of belonging. The process continues until all observations are classified in some class. Defining the misclassification error as the probability of classifying wrongly a particular picture, in RGB images we the error is $4,22\%$. Applying exactly the same procedure to gray *WWL* set, taking the three corresponding variables, we obtain a misclassification error of $4,49\%$. It can be observed that previous results do not show clear differences between RGB and GRAY images neither in cluster, nor in discriminant analysis. As a consequence of this, we henceforth use the databases in gray color.

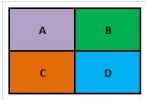
With the purpose of proving the effectiveness of the proposed variables we apply a nearest neighbor classification method (K-NN method). By means of k-nearest-neighbor algorithm, one set denoted as *training* set is used to classify each member of another set called *sample*. The structure of the data is given by a classification (categorical) variable of interest (e.g.,"landscape", or "non-landscape"), and a number of additional predictor variables (e.g., local variability, effective variance and spatial correlation). In general, the process is carried out as follows. Step 1: for each case (image) in the *sample* set, the k closest members are located (the k nearest neighbors) in the *training* set. The euclidean distance measure is used to calculate how close each member of the *training* set is to the each case in *sample*. Step 2: an object is classified by a majority vote of its neighbors, being assigned to the class most common amongst its k nearest neighbors (majority rule). The value of k is a positive integer, typically small and even, to avoid tie. Step 3: the procedure is repeated for the remaining cases in the *sample* set. Therefore, the classification is performed by using the *nearest neighbor method* considering the Euclidean distance and applying the majority rule. There is no consensus in defining the adequate number k of nearest neighbors. Hall et al. (2008) propose some properties to motivate new methods for choosing the value of k. However, in most previous work on nearest-neighbor classifiers, the value of k is held by cross-validation, which is the method that we used.

To illustrate this approach, we present two examples. First, we use *WWL* database as our *training* set, to classify the *GOOGLE* database, both in gray color. Second, we apply the nearest method to classify the *ZIPCODES* database. We use the discriminant variables $\bar{V}_e$, δ and ρ_1 to perform the K-NN method. Besides, we also we include ρ_2 and ρ_3, in order to analyze their discriminant power in the K-NN

method. Then, we consider five variables per image: $V_e, \delta, \rho_1, \rho_2$ and ρ_3.

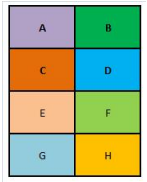
Specifically, the aforementioned *WWL* set is used to classify the *GOOGLE* group using the K-NN method, considering the classification variables and the number of neighbor equals to $k = 7$ *and* 9. Table 3 shows the results obtained to *GOOGLE* dataset. The lower misclassification error $(8,97\%)$ is obtained by considering 9 neighbors and only three variables (V_e, δ, and ρ_1).

In order to improve the classification, we consider appropriate dividing the images sections. We first divide the number of rows and columns by 2, resulting in four equally-sized sections. For instance, let X an image of dimensions 128×96, each section of the image will have a dimension 64×48. Graphically,



Afterwards, we recalculate the classification variables for each section. They are V_e, δ, ρ_1 for sections A, B, C and D. Due to the dimensionality reduction, we omit the calculation of ρ_2, ρ_3, because they become zero for high order of distance (high h).Consequently, there are 12 values of the variables for every image. We perform the K-NN method, obtaining a higher misclassification error, regardless the variables we choose. The lower error obtained for 4-sections classification is $12,14\%$.

Second, we divide the image in 8 sections, dividing the number of rows by 4 and the number of columns by 2. Graphically,



Then, we recalculate our three classification variables (V_e, δ, ρ_1) for all 8 sections of every image. As a result, we have 24 values of the variables per picture. The error is increased with respect to the previous analysis. Then, the technique of divide the image in sections seem no to be good enough for this dataset.

We also apply the same techniques to the *ZIPCODES* database. This database was already analyzed by Hastie et al. (2001). We begin the study of this dataset, by calculating the statistics measures for every classification variable (see Appendix Table A.3). Then, we analyze the means of the classification variables for every number class.

TABLE 3. Means by classes: *ZIPCODES* database

	true class									
variable	0	1	2	3	4	5	6	7	8	9
V_e	1.11	0.22	1.29	1.16	1.14	1.22	1.38	1.30	1.32	1.31
δ	0.34	0.07	0.33	0.34	0.24	0.33	0.34	0.24	0.37	0.29
ρ_1	0.54	0.60	0.51	0.50	0.50	0.51	0.52	0.54	0.50	0.51
ρ_2	0.19	0.24	0.17	0.19	0.17	0.16	0.16	0.22	0.18	0.19
ρ_3	-0.02	0.09	-0.01	0.05	0.01	0.02	-0.03	0.04	0.07	0.07

The Table above shows that the *number one* presents a differentiated behavior with respect to the rest of numbers, specially considering the V_e and δ.

The K-NN method is performed by dividing the *ZIPCODES* dataset (7291 images) in two halves. As a *training* set we use the first half which allow as to classify the second half (*sample* set). Then, we apply the K-NN method considering 13 and 15 nearest neighbor, numbers chosen by cross-validation. To classify we use the same procedure than in the previous example. The results are shown in the following table

TABLE 4. K Nearest Neighbor Method: Results

Image	Variables	GOOGLE database		ZIPCODE database	
		K-nearest Neighbors	misclassification error %	K-nearest Neighbors	misclassification error %
Whole	V_e, δ ρ_1	7	10.55	13	60.66
		9	**8.97**	15	60.08
	V_e, δ ρ_1 ,ρ_2 and ρ_3	7	10.29	13	52.46
		9	10.03	15	52.13
4-sections	V_e (4 values per image)	7	13.46	13	66.98
		9	12.66	15	63.79
	V_e and δ (8 values per image)	7	13.46	13	41.81
		9	13.19	15	42.44
	V_e, δ ρ_1 (12 values per image)	7	12.66	13	37.94
		9	12.14	15	38.60
8-sections	V_e (8 values per image)	7	21.64	13	36.08
		9	20.58	15	36.16
	V_e and δ (16 values per image)	7	22.16	13	27.85
		9	20.84	**15**	**27.43**
	V_e, δ ρ_1 (24 values per image)	7	21.90		
		9	19.26		

Considering the image as a whole, we obtain a misclassification error of $52,13\%$, taking into account five classification variables and 15 neighbors. However, when dividing the image in 4 sections, the error significantly decreases down to $37,94\%$, considering three variables per section (12 per image) and 13 neighbors. In order to improve the classification even more, we divide the whole image in 8 sections[2]. Now, the lower misclassification error is $27,43\%$ taking two variables (Ve and δ)

[2]Given that images in *ZIPCODES* database have dimension 16×16, the calculation of ρ_1 is omitted. This is because each section of image has dimension 8×2 and ρ_1 is in many cases undetermined.

and 15 neighbors. The division of image in 8 sections improves the classification considerably.

A final analysis is conducted by including a new variable. This variable is calculated over the image divided in four sections. Given sections A, B, C and D, we obtain the correlation pixel to pixel, among the four parts. This new correlation is obtain as follows,

$$\varrho_{pq} = \frac{\sum_i^n \sum_j^m (X_{ij_P} - \bar{X}_P)(X_{ij_Q} - \bar{X}_Q)}{\sigma_P * \sigma_Q} \tag{5}$$

Where, ϱ_{pq} is the correlation between section P and Q; X_{ij_P} and X_{ij_Q} are the pixels contended in sections P and Q. Finally, $\bar{X}_P$ and $\bar{X}_Q$ are the general means of sections P and Q. Then, we obtain six new variables per image. They represent all possible correlations among the four sections, that is ϱ_{AB}, ϱ_{BC}, ϱ_{CD} ϱ_{DA} ϱ_{AC} ϱ_{BD}. We add these new variables to those used in the 8-sections analysis, making a total of 22 variables. We obtain a new classification result shown in the following table. The

TABLE 5. K-NN method:*ZIPCODE* database

K-nearest Neighbors	misclassification error %
11	**20.63**
13	21.02
15	21.24

best classification rate is obtained considering 11 neighbors, obtaining an error of 20.63%. The final classification frequencies are shown in the following contingency table. The predicted classes are those estimated by the forgoing procedure, and the true classes correspond to the real belonging group of *ZIPCODES* images.

TABLE 6. Predicted vs True class: ZIPCODES database

	Predicted class										Total
True class	0	1	2	3	4	5	6	7	8	9	0
0	537	0	3	3	2	1	6	1	6	6	565
1	0	469	1	0	0	0	0	0	0	1	471
2	36	0	268	42	4	9	20	4	2	2	387
3	31	2	13	306	1	38	6	5	8	7	417
4	27	8	23	7	161	4	12	53	1	76	372
5	28	0	5	45	2	206	17	2	4	15	324
6	11	0	1	0	0	3	289	0	1	0	305
7	2	2	1	0	1	0	6	247	0	34	293
8	12	5	3	6	0	8	7	4	185	22	252
9	5	1	0	2	2	1	3	19	1	225	259
Total	689	487	318	411	173	270	366	335	208	388	3645

6. FINAL REMARKS

The goal of this research is to propose an alternative method to classify and group images. The performance of our proposal is given by the application of cluster and discriminant techniques using three discriminant variables.

In k-means method we obtain good grouping levels without significant differences between RGB and GRAY images. In nearest neighbor technique, we achieve the lower misclassification error in *GOOGLE* database, by considering only two of the proposed variables. However, in *ZIPCODES* database, the best results are obtained when dividing the images in 8 sections and including the correlations among 4 bigger sections.

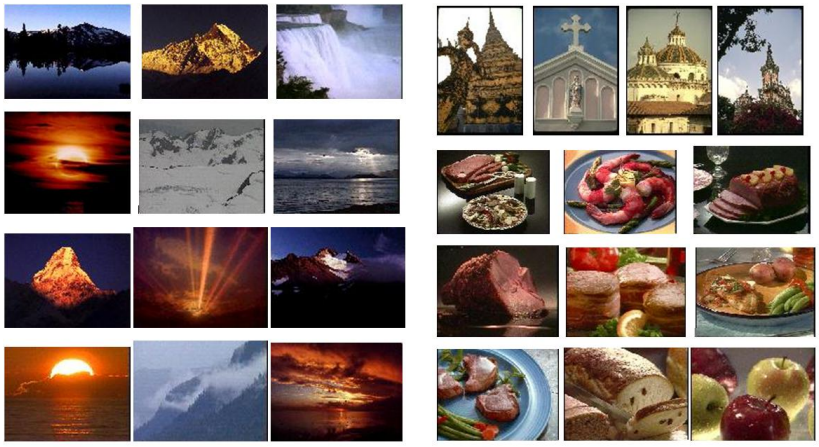
Cutting images in sections improves the classification error considerably in the *ZIPCODES* database. However, this procedure does not improve the results for the *GOOGLE* database. One potential reason that may explain these differences across databases may lay in the complexity of images. Images within *ZIPCODES* database are single traces on a solid background. Our measures seem to work better as these images are subdivided in smaller sections. Alternatively, the reasons may lay on the size of the images. *GOOGLE* database contained big-sized images that, even after divided them, remained large compared to those images within the *ZIPCODES* database. Future research should try disentangle the conditions under which dividing images offer a better classification rate.

7. Acknowledgements

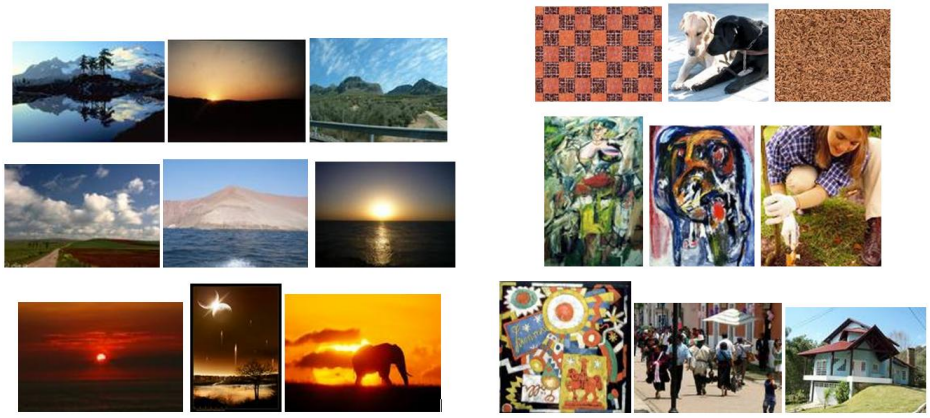
This research acknowledges the support of "Comunidad de Madrid" grant CCG07-UC3M/HUM-3260.

8. Appendix

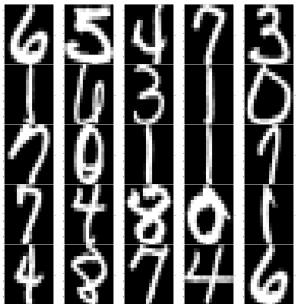
8.1. Typical images from the *WWL* set.



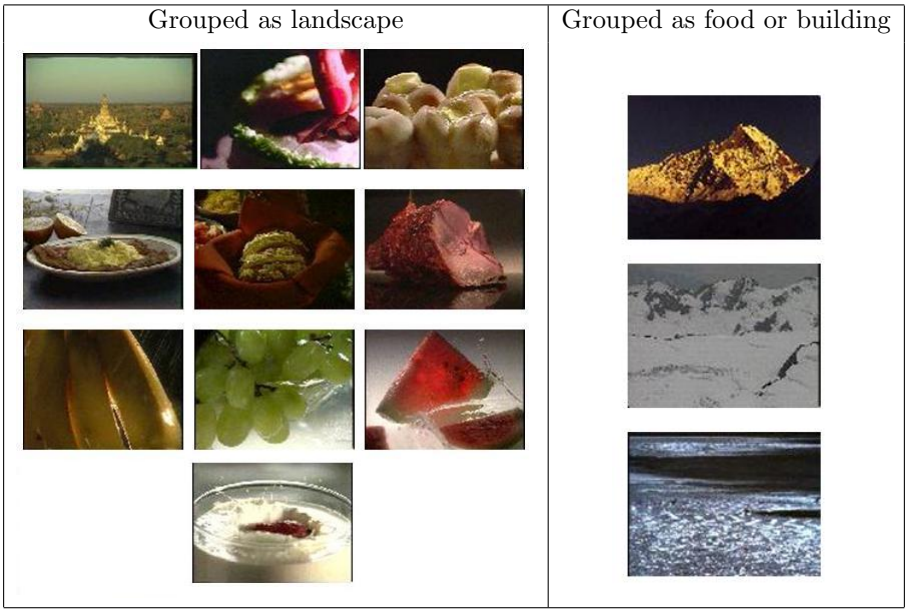
8.2. Typical images from the *GOOGLE* set.



8.3. Typical images from *ZIPCODES* set.



8.4. **Some misclassified RGB pictures.**



8.5. **Some misclassified gray pictures.**

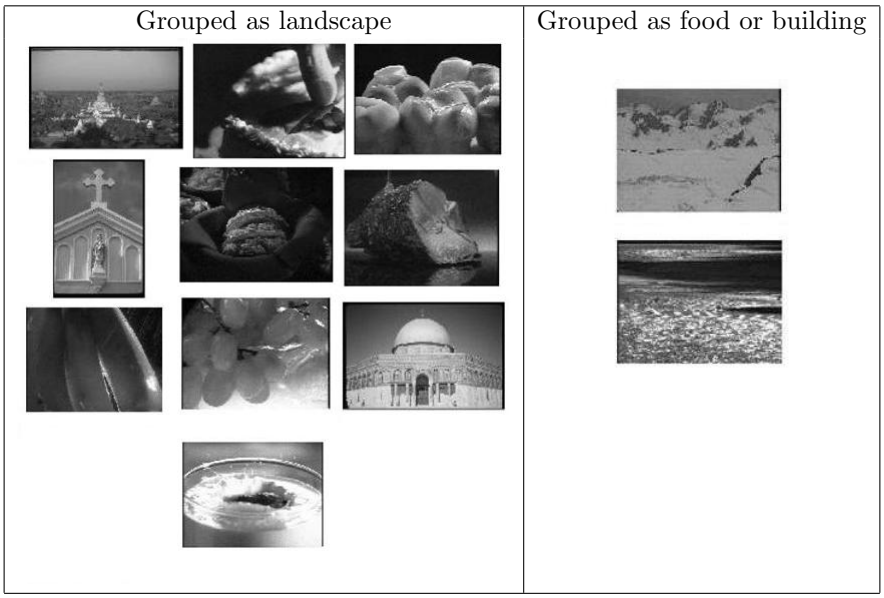


TABLE A.1. Descriptives Statistics- *WWL* dataset

Variable	matrix	Mean	St. Dev	St.Error
Effective Variance	R	0.343	0.2862	0.0147
	G	0.34	0.287	0.015
	B	0.33	0.278	0.014
	Total	0.39	0.284	0.008
Local Variability	R	0.04	0.0345	0.0017
	G	0.04	0.035	0.002
	B	0.04	0.034	0.002
	Total	0.04	0.034	0.001
Spatial Correlation 1	R	0.79	0.093	0.005
	G	0.77	0.103	0.005
	B	0.76	0.12	0.006
	Total	0.78	0.106	0.003

TABLE A.2. ANOVA - *WWL* dataset

Variable		Sum of squares	df	Mean square	F	Sig.
Effective Variance	within-groups	0.028	2	0.014	0.175	0.839
	betw-groups	91.357	1134	0.081		
	Total	91.385	1134			
Local Variability	within-groups	0.000	2	0.000	0.053	0.949
	betw-groups	1.339	1134	0.001		
	Total	1.339	1134			
Spatial Correlation	within-groups	0.191	2	0.095	8.523	0.000
	betw-groups	12.686	1134	0.011		
	Total	12.876	1134			

TABLE A.3. Descriptive Statistics: *ZIPCODES* database

Variable	N	Min	Max	Mean	Standard Deviation	Variance	Skewness	Standard Error
V_e	7291	0.00	2.28	1.10	0.42	0.18	-1.44	0.03
δ		0.02	0.56	0.28	0.11	0.01	-0.82	0.03
ρ_1		0.22	0.74	0.53	0.08	0.01	-0.22	0.03
ρ_2		-0.13	0.56	0.19	0.09	0.01	0.55	0.03
ρ_3		-0.25	0.41	0.03	0.09	0.01	0.57	0.03

References

Benito, M. and D. Peña (2004). Dimensionality reduction with image data. *Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science.Springer. Verlag*, 326–332.

Bottou, L., C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik (1994). Comparison of classifier method: A case study in handwritten digits recognition. *Pattern Recognition 2*.

Fraiman, R., A. Justel, and M. Svarc (200). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association 103*(483), 1294–1303.

Hall, P., B. Park, and R. Samworth (2008). Choice of neighbor order in nearest-neighbor classification. *Annals Statistics 36*(5), 2135–2152.

Hastie, T. and P. Simard (1998). Metrics and models for handwritten character recognition. *Statistical Science 13*(1), 54–65.

Hastie, T. and R. Tibshirani (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*(6), 607–616.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *Data Mining, Inference, and Prediction*. Springer.

Herwig, R., A. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien (1999). Large-scale clustering of cdna-fingerprinting data. *Genoma Research 9*, 1093–1105.

Le Cun Y., e. a. (1995). Handwritten digit recognition with a back propagation network. *International Conference on Artificial Neural Networks*, 53–60.

Liu, K., Y. Cheng, and J. Yang (1993). Algebraic feature extraction for image recognition based on an optimal discriminant criterion. *Pattern Recognition 26*, 903–911.

Marron, J. S. and M. Todd (2002). Distance weighted discrimination. *Technical Report*.

Peña, D. and J. Rodriguez (2003). Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis. 2*(58), 361–374.

Turk, M. and A. Pentland (1991). Face recognition using eigenfaces. *The Journal of Cognitive Neuroscience 3*.

Vailaya, A., A. Jain, and H. J. Zhang (1998). On image classification: City images vs. landscapes. *Pattern Recognition 31*, 1921–1935.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Wang, J., J. Li, and G. Wiederhold (2001). Semantics sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(9).