HUITIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 1er au 5 JUIN 1981

A SIGNAL DETECTION TECHNIQUE TO ACHIEVE DATA CLUSTERING FROM DATA HISTOGRAMS

GABRIEL E. LOWITZ

MATRA ESPACE, RUE DES COSMONAUTES, 31 400 TOULOUSE - FRANCE

RESUME

Cet article présente une stratégie nouvelle de classification non supervisée basée sur une analyse d'histogrammes. La philosophie générale a déjà été publiée (LOWITZ [1]). Cet article traite plus spécifiquement d'une solution de ce problème par filtrage adapté.

Le cas intéressant est celui d'une scène multispectrale en Télédétection. On la réduit, d'abord à ses deux composantes principales par transformation unitaire de K.L. L'histogramme bidimensionnel résultant est alors partitionné en segments connexes par détection directe des centres et des séparatrices des distributions gaussiennes sous-jacentes. Cette détection est effectuée par filtrage adapté dans le domaine de Fourier, le signal étant ici l'histogramme. Le filtrage adapté est optimisé à l'aide d'un développement en ondes de la Prolate Spheroïdale (Slepian[4]).

Mots Clés

Théorie du Signal - Classification - Reconnaissance des Formes - Traitement d'Images - Fonctions Orthogonales - Fonctions d'Ondes de la Prolate Spheroïdale - Imagerie Multispectrale.

SUMMARY

This paper presents a novel and computationally efficient strategy to achieve data clustering directly on histograms without iteration.

The general philosophy and motivations of clustering Data histograms by extraction of the self-information have been developped in a previously published paper (LOWITZ[1]). The contents of this present paper deals with an optimum signal theory implementation of these concepts.

The incoming raw multichannel data can be first reduced to its two principal (Karhunen-Loeve) components to augment the weight of the statistical evidence, the histogram of which is then partitionned in non overlopping radiometry domains after detection of the centers and separatrices of the underlying distributions. The novelty consists of the extraction of the self information, the detection of the underlying distribution and the required filtering by a Fourier transform of the histogram itself considered as an information carrying signal. The methodology is first explained in simple terms using the vocabulary of signal detection. This methodology is later refined using a model based on Slepian and AL'S Prolate Spheroidal Wave Fonctions.

KEY WORDS

Signal Theory - Clustering - Pattern - Recognition Image processing - Classification Orthogonal Functions - Prolate Spheroidal Wave Functions - Histograms - Multispectral Imagery.

A SIGNAL DETECTION TECHNIQUE TO ACHIEVE DATA CLUSTERING FROM DATA HISTOGRAMS

CLUSTERING ON HISTOGRAMS

In a previously published paper (LOWITZ[1]) the motivations and the basic principles of clustering on histograms were given. These motivations and principles are briefly reviewed here.

WATANABE[2] has demonstrated mathematically (The Ugly Duckling Theorem) that classification is possible only if a complete set of predicates is altered by an extra logical distribution of weights on these various predicates either in the form of a "Professor" or in the form of a global statistical distribution, a frame of imagery for instance. In the first case the classification is said to be "supervised" in the second case, the classification is (wrongly) said to be "unsupervised" and is commonly refered as "clustering". It is quite clear that clustering is a form of classification internally supervised by the own statistics of the data : The WATANABE conditions are satisfied.

Given an image for instance and the radiometric values as the set of predicates, the histogram of the radiometries is the best possible statistics to supervise the clustering : A histogram is the best estimator of the superposition of the underlying distributions and the a priori probabilities thereof.

The clustering information, however, is not directly derived from the histogram, it is derived from the information law attached to the histogram. To start with take a gaussian shaped histogram, i.e. a gaussian distribution with a given a-priori probability. The self information attached to the radiometry x in the "message" :

$$x_0 \leq x \leq x_0 + \Delta x$$

Where x is a realization of a random variable governed by a probability of occurrence such that : $h(x) = K_1 \exp\left\{-x^2/2\sigma^2\right\}$

within a constant factor, is given by the self-information law $I(x)$ obtained by taking the logarithms of the counts :

$$I(x,\sigma) = -\mathrm{Log}_2\{h\} = k_2 x^2/2\sigma^2$$

In this simple case the information law is a parabola whose invariant is its curvature : $\frac{d^2I}{dx^2} = 1/\sigma^2$

A gaussian shaped histogram is therefore characterized by a parabolic information law which has a constant curvature equal to This fact was recognized by FISHER[3] who called this quantity the information content of the distribution. In our context $I(x)$ is called the information content of the (gaussian) histogram.

The general problem of clustering on histograms consists of finding its "primitives" i.e. the various distributions whose superposition, within an additive noise contribution, will adequately represent the given histogram. If one is interested in gaussian primitives, then the clustering problem is reduced to the search of the centers and the curvatures of the underlying parabolic information laws.

The intended objective of the present paper is to reduce this general problem of clustering on histograms to a signal detection problem whose optimal solution is known and can be derived in the FOURIER domain by efficient digital means.

CLUSTERING ON HISTOGRAMS, A SIGNAL DETECTION PROBLEM

Under the assumptions of gaussians primitives the information law attached to each underlying distribution is parabolic. In order to find the underlying parabolas, buried in lot of noise, the appropriate pre-processing of the histogram consists of taking the logarithm of the counts. Within a small interval (x_i', x_i''), the parabolas should be detectable by appropriate means :

$$y_i = \frac{(x - x_i)^2}{2\sigma^2}, \quad x' < x \leq x'', \quad x'' - x' = T.$$

Their essential characteristic is a constant curvature within a small interval of measure T. Another essential characteristic is the fact that, in pratical cases, the radiometric values are discrete and sampled at Shannon frequency w_0. In order to enhance the signal from the noise, the detection should be done with a cut off frequency Ω smaller than w_0.

It seems therefore logical to attempt the detection of a signal of known shape, but unknown time of arrival, by matched filtering in the FOURIER domain to bank on the translation invariance property and the ease of zonal filtering offered by the FOURIER Transform.

A SIGNAL DETECTION TECHNIQUE TO ACHIEVE DATA CLUSTERING FROM DATA HISTOGRAMS

Because the discrete fourier amplitude spectrum of a centered parabola has $1/w^2$ (w integer) as an enveloppe, the proper filter is known :

$$F(\omega) = \omega^2, \; |\omega| < \Omega \leq \omega_0 .$$

Limiting the bandwidth to Ω (a threshold to be determined) provides a re-sampling in signal space that permits to control both the noise and the maximum number of detected centers. After filtering and inverse transformations, the detected centers will correspond to a signal exhibiting zero curvature within the small interval $\mathcal{T}$, i.e. points corresponding to the maxima or minima of the reconstructed wave. The value of the reconstructed wave at these extrema will be proportional to the inverse of the variance of the detected underlying gaussian distribution when this variance is positive. When this "variance" is negative one has detected a parabola with inverse curvature i.e. the separation between two gaussians distributions.

The significance of this "LORENTZ metric" in terms of apparent variance has been explained in LOWITZ[1].

Using the discrete Fourier transform scheme detailed in Figure 1 has one drawback : The classification relies on the detection of the extremal points of a wave. One would prefer to rely on the detection of zero-crossings because such points are readily observable and easily detectable by digital means.

To achieve this goal, the true Fourier transform has to be replaced by the sine and cosine transforms.

These transforms are implemented from the conventional (FFT) algorithms in reconstructing on positive frequencies only : the phasing information contained in the complex Fourier component is then reconstructed as the inverse sine transform and its zero-crossings occur at the extrema of the inverse cosine transform, which, within a constant, is identical to the inverse Fourier transform and carries only the variance information of the underlying gaussian distribution. The only information not readily obtained is the a-priori probability of each distribution. An estimate of these a-priori probabilites can be obtained by summing the counts of the original histogram over each of the detected classes.

In one dimension, when the domain of the histogram is on a line, the classification problem has been completely solved.

In several dimensions the method permits to positively detect Kernels of classes with a measure of variance for each point of these Kernels. The remaining points can easily be classified by a K-neighbourgh method with a Mahalanobis metric to take advantage of the availibility of a measure of variance. Separatrices also forms Kernels that help the classification by forbiding to belong to a class whose Kernel is on the other side of such a barrier.

It should be stressed that the classification is done in the domain of definition of the histogram (feature space) and not in image space. For each point of this domain a number of image points are dealt with at the same time. As a result the classification is done with a great speed. Once the domain of the histogram is classified, a look up table and a single reading back of the image radiometries permits the production of the thematic map.

ADDITIONAL SIGNAL FILTERING

Up to now nothing has been said about the implementation of the detection of the parabolic signal within the small interval $\mathcal{T}$ in an effort to optimize the detection in presence of the inevitable statistical noise. The required filtering is a convolution of the $I(x)$ function by a gaussian filter whose sigma is expressed in units of $\mathcal{T}$. In Fourier space this filter component will also be gaussian with variance inverse of the signal space variance. The following section outlines the required optimization model in terms of the so called Prolate Spheroidal wave functions. Though this model does not add any new concept to the classical signal detection procedure outlined above it permits an elegant optimization perfectly adapted to the problem, easily implemented by digital means and discrete arithmetic.

THE PROLATE SPHEROIDAL WAVE FUNCTION MODEL

An elegant theoretical model relevant to a large class of problems including discrete filtering has been published in 1961 by SLEPIAN and Al[4] and has been recently used by SHANMUGAM and Al[5] for the optimum detection of edges in digital images. Here the same model is proposed to implement clustering on histograms.

The prolate spheroidal functions are in fact the eigenfunctions of the discrete FOURIER transform. Their basic properties are briefly reviewed in this section, compiled from reference 4.

Given any $T>0$ and any $\Omega>0$ there exists a countably infinite set of real functions :

$\varphi_0(t), \varphi_1(t), \ldots$

and a set of positive numbers $\lambda_0 > \lambda_1 > \lambda_2 > \ldots$

with the following properties :

- The $\varphi's$ are band limited, orthonormal on the real line and complete for the representation of band limited functions of $\mathcal{L}^2_\infty$:

$$\int_{-\infty}^{+\infty} \varphi_i(t)\,\varphi_j(t)\,dt = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \qquad i, j = 0, 1, 2, \ldots$$

- In the interval $-T/2 < t \leq +T/2$ the $\varphi's$ are orthogonal and complete in $\mathcal{L}^2_{T/2}$:

$$\int_{-T/2}^{+T/2} \varphi_i\,\varphi_j\,dt = \begin{cases} 0, & i \neq j \\ \lambda_i, & i = j \end{cases} \qquad i, j = 0, 1, 2 \ldots$$

- For all values of t real or complex :

$$\lambda_i\,\varphi_i(t) = \int_{-T/2}^{T/2} \frac{\sin \Omega(t-s)}{\pi(t-s)}\,\varphi_i(s)\,ds, \quad i = 0, 1, 2.$$

In fact both the $\varphi's$ and the $\lambda's$ are functions of a single parameter c given by SLEPIAN's relation : $c = \Omega T/2$

In can be demonstrated that the FOURIER transform of $\varphi_i(c,t)$ is given by the relation :

$$\mathcal{F}\{\varphi_i(c,t)\} = K(j)^i\,\varphi_i(c, \omega T/2\Omega)$$

Where K is a constant and $j = \sqrt{-1}$

The table 2 below gives some values of and illustrates general shape of $\varphi_0(c,t)$, $\varphi_1(c,t)$ and the influence of the parameter c on their shape.

From the graphs in table 2 it is quite clear that c is a linear expansion factor for t. As a result, for computational convenience, φ_0 and φ_1 can be closely approximated, for $c \leq 2$ and $|x| = \frac{2t}{T} \leq 1$ by the following HERMITE functions :

$$\varphi_0(c, 2t/T) \simeq \sqrt{\frac{c}{\pi}}\,\exp\left\{-\frac{c^2 t^2}{2T^2}\right\}$$

$$\varphi_1(c, 2t/T) \simeq \sqrt{\frac{c}{\pi}}\,\frac{ct}{T}\,\exp\left\{-\frac{c^2 t^2}{2T^2}\right\}$$

These approximations have been arrived at by fitting the amplitude of φ_0 and the slope of φ_1 for t = 0. These approximations hold quite closely for values of c up to 2.

If f (x) and g (x) denote respectively the input signal within T and the output signal after filtering, if h (x) is the spread function of the filter, F (w), G (w), H (w) their FOURIER transforms, the optimum detection problem can be stated as follows : T is the symetrical interval inside which the signal (the parabola) is to detected. Ω is a frequency cut off. One wants to maximize the following energy ratio

$$\gamma = \frac{\int_{-T/2}^{T/2} |g(x)|^2}{\int_{-\infty}^{+\infty} |g(x)|^2}$$

with the constraints :

$$H(\omega) = G(\omega)/F(\omega), \quad |\omega| < \Omega \leq \omega_0$$
$$H(\omega) = 0 \qquad |\omega| \geq \Omega$$
$$F(\omega) = K/\omega^2$$

the Fourier amplitude spectrum of the parabola.

Looking for a representation of the filtered signal g (x) in terms of the prolate spheroidal wave functions such as :

$$g(x) = \sum_{n=0}^{\infty} a_n\,\varphi_n(c,x)$$

the scalar γ to maximize becomes :

$$\gamma = \frac{\hat{\sum} |a_n|^2 \lambda_n}{\hat{\sum}_n |a_n|^2}$$

But the eigen values λ_n are positive and strictly decreasing.

Therefore :

$$0 < \gamma \leq \frac{\lambda_0 \hat{\sum} |a_n|^2}{\hat{\sum} |a_n|^2} = \lambda_0$$

The optimum output corresponds to the maximum value of which is reached when the series representation is limited to its first term :

$g(x) = a_0\,\varphi_0$

The Fourier transform is obtained using the relation :

$$\mathcal{F}\{\varphi_0(c,x)\} = K\,\varphi_0\left(c, \frac{\omega T}{2\Omega}\right)$$

and the optimum filter becomes, for $|\omega| \leq \Omega$

$$H(\omega) = G(\omega)/F(\omega) = K\omega^2\,\varphi_0\left(c, \frac{\omega T}{2\Omega}\right)$$

A SIGNAL DETECTION TECHNIQUE TO ACHIEVE DATA CLUSTERING FROM DATA HISTOGRAMS

Now approximating φ_0 by the appropriate Hermite function yields :

$$H(\omega) = K_1 \omega^2 \exp\left\{-\frac{c^2}{2\pi^2}\left(\frac{\omega T}{2\Omega^2}\right)^2\right\} = K_1 \omega^2 \exp\left\{-\frac{c^2\omega^2}{8\Omega^2}\right\}$$

Using Slepian's relation $c = \Omega T/2$ permits a further simplification :

$$H(\omega) = K_1 \omega^2 \exp\left\{-\frac{\omega^2}{2}\left(\frac{T}{4}\right)^2\right\} \qquad |\omega| \leq \Omega$$
$$H(\omega) = 0 \qquad |\omega| > \Omega$$

As the inverse transform of a Gaussian shape with variance σ^2 is a gaussian with variance $1/\sigma^2$, it follows that the exponential component of the filter corresponds in signal space to a convolution by a gaussian shape with $\sigma = T/4$

On the other hand the hard limiting after at Ω corresponds in image space to a convolution by $\sin(\Omega x)/x$ which, in discrete topology, resamples the spatial domain according to a new Shanon frequency Ω . Finally the component ω^2 corresponds, within a sign, to the second derivative of the signal, i.e. the extraction of the curvature of the information function $I(x)$ attached to the histogram under processing.

The use of the Prolate Spheroidal Wave functions model has permited to optimize the detection of the relative extrema of $I(x)$ which are normally buried in large amounts of statistical and quatification noise.

On a pratical point of view varying Ω permits to control the average number of classes to be detected, while increasing T permits to merge those classes that are very close to each other. The overall detection performances will only be function of $c = \Omega T/2$, the "frequency-space" compression parameter.

The translation invariance property of the Fourier transform simplifies the multiclass detection problem, while the FFT algorithm computing speed increases the overall efficiency of the proposed methodology.

The extension to higher dimensions does not present any difficulty due to the even symetry of the optimum filter. In two dimensions for instance the filter becomes :

$$H(u,v) = K_1 (u^2+v^2) \exp\left\{-\frac{u^2+v^2}{2}\left(\frac{T}{4}\right)^2\right\}, \quad u^2+v^2 \leq \Omega^2$$
$$H(u,v) = 0 \qquad u^2+v^2 > \Omega^2$$

EXPERIMENTAL RESULTS

The clustering methodology outlined in the preceding sections has been extensively experimented on histograms of single black and white pictures (one dimension) and on composite histograms cartesian products of the first two eigenimages of multichannel imagery. Such a dimensionality reduction preserves 97 to 98% of the total scene variance.

The problem encountered in dealing with composite histograms cartesian products of N channels, $N > 2$, has nothing to do with the efficiency of the proposed strategy. Such higher dimension histograms have extremely poor weight of statistical evidence simply because the average count per histogram cell becomes zero almost everywhere resulting in an intolerable statistical noise.

In the reduction to practice the only difficulty encountered resulted from an inconsiderate choise of the integral number implementing the cut off frequency Ω in the discrete implementation of the optimal filter. For certain even values of this parameter the reconstructed waves presented parasitic bumps (false classes). The problem was easily traced to a beat phenomenon due to the reverse Gibbs effect of the sharp cut off Ω accentuated by the spectrum overlap between the frequencies Ω an ω_0 the original shanon frequency. This parasitic effect was eliminated by choosing for Ω an odd integer : This choice decreases the beat frequency in signal space and rejects the parasitic bumps of the reconstructed wave outside the domain of definition for the histogram.

Adequate noise filtering and a proper number of detected classes (16 to 20) where obtained for values of Ω roughly equal to one third of Shannon frequency w_0. The optimum value of T was found between 1 and 8, and, as expected, permited a "fine tuning" on the number of classes by effecting a merge when desired.

The photographic reproductions at the end of the paper illustrate typical clustering results.

It has been stated that the histogram is a good statistics because it contains the estimation of the underlying distributions and those of their higher moments. This property has been thoroughly used. It has been shown that, under the gaussian hypothesis, the information law attached to the histogram is a superposition of parabolas whose centers can be optimally detected, within small intervals of measure τ with a matched filter in the Fourier domain.

The Prolate Spheroidal Wave function model has been used to optimize the filter and an approximation with Hermite functions has permited to implement the values of the filter parameters without extensive computations.

Some results of digital simulation have also be presented to validate the theoretical concepts.

REFERENCES

1 - LOWITZ - LEBOUCHER "What an Histogram can Really Tell the Classifier". Pattern Recognition, Vol 10, PP 351 - 357, 1978

2 - WATANBE "Knowing and Guessing". The Ugly Duckling Theorem. John WILEY, 1969 pp 376 - 379

3 - FISHER "On the Mathematical Foundations of Theoretical Statistics" Reading at the Royal Academy of Sciences Pp 309 - 368 - Nov. 17th, 192

4 - SLEPIANS, AL "Prolate Spheroidal Wave Functions, Analysis and Uncertainty I" Bell Syst. Tech. J. Vol 40, pp 43 - 46,61

5 - SHANMUGAM, AL "An Optimal Frequency Domain Filter for Edge Detection in Digital Pictures". IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. PAMI 1, n° 1, Janv. 1979.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the contribution of Mr. B. POUCHIN a cooperative student from ECOLE SUPERIEURE DE PHYSIQUE DE MARSEILLE who managed to program the clustering by Fourier Transform strategy in PDP - 11 machine language on the interactive image processing system. A very special acknowledgement goes to Mr. V. DUBE a cooperative student from the UNIVERSITY OF WATERLOO, ONTARIO, CANADA who developped the algorithm in FORTRAN on IBM-370-158 and who was responsible for many improvements.

Many thanks also to MM. J.M. VIVIER and G. BOURRY from the permanent staff of the Space Division Image Processing Laboratory for their help and many usefull discussions. Many thanks also to Mr G. LEBOUCHER for his constant interest and suggestions.

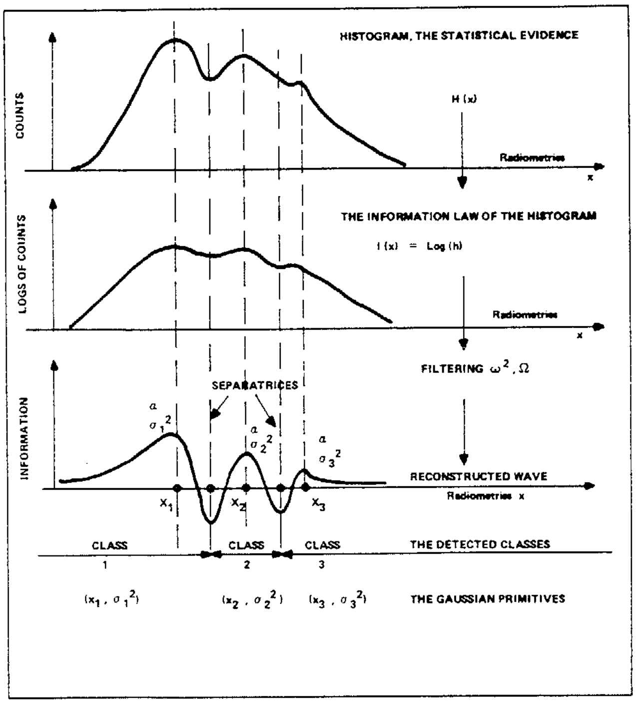


FIGURE 1 – PRINCIPLE OF CLUSTERING BY FOURIER FILTERING

i	C = .5	C = 1	C = 2
0	.30969	.57280	.88056
1	.00858	.06279	.35564
2	.00004	.00124	.03587

EIGENVALUES λ_i

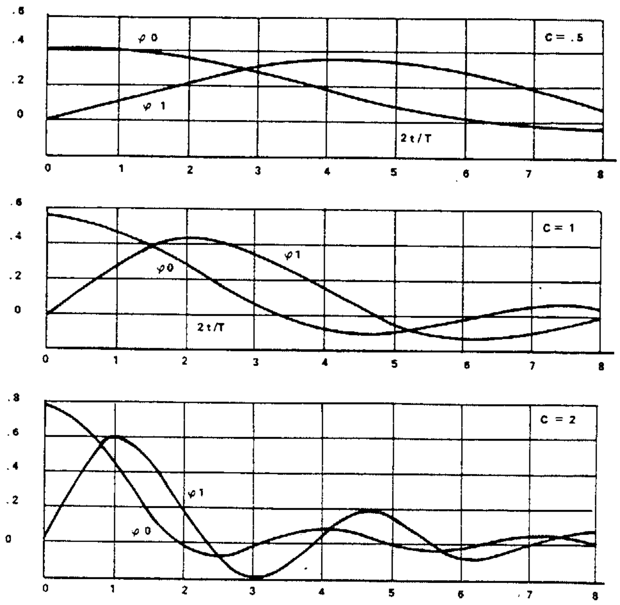
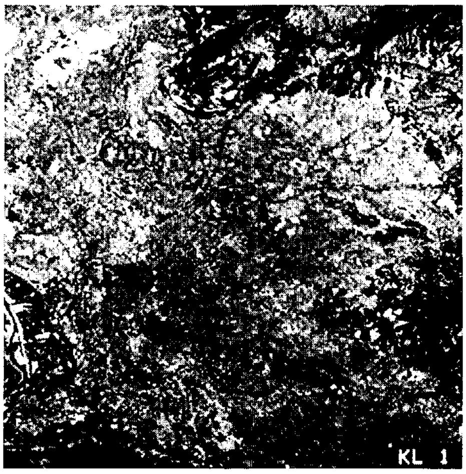


TABLE 2 – EIGENVALUES AND GRAPHS OF φ_0 AND φ_1 THE FIRST TWO PROLATE SPHERICAL WAVE FUNCTIONS

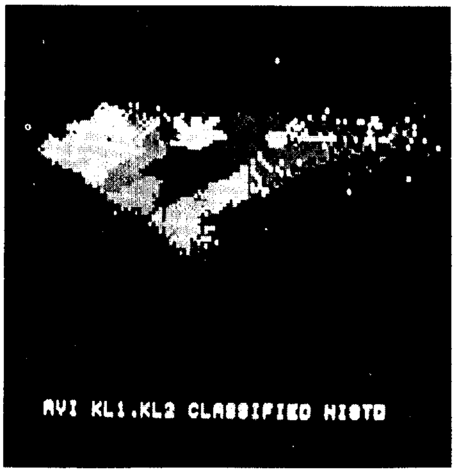
A SIGNAL DETECTION TECHNIQUE TO ACHIEVE DATA CLUSTERING FROM DATA HISTOGRAMS



KL1 EIGENIMAGE

KL2 EIGENIMAGE

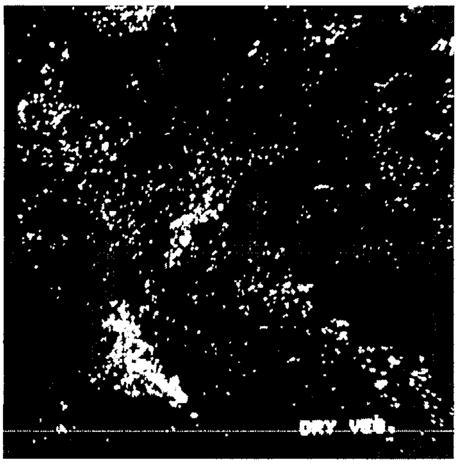
(KL1 AND KL2 CONTAIN 18% OF THE SCENE VARIANCE)



KL1 X KL2 CLASSIFIED HISTOGRAM

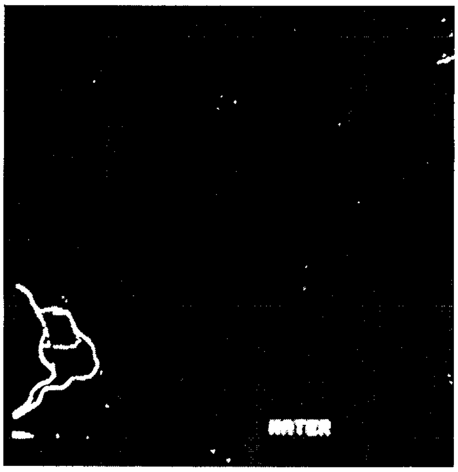
EXAMPLE OF CLUSTERING IN MULTIDIMENSION SPACE :
PRINCIPAL COMPONENT PROCESSING OF THE AVIGNON SCENE
RESULTS OF FEATURE SPACE CLUSTERING (10 CLASSES)

MATRA-ESPACE (LABORATOIRE DE TRAITEMENT DES IMAGES)

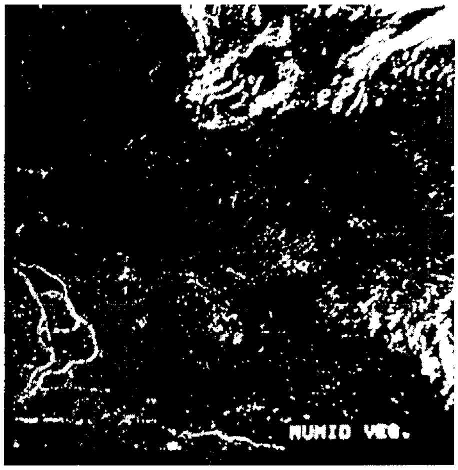


OPEN COUNTRY CULTURES

THE CITY, CIMENT AND ROCK CLASS



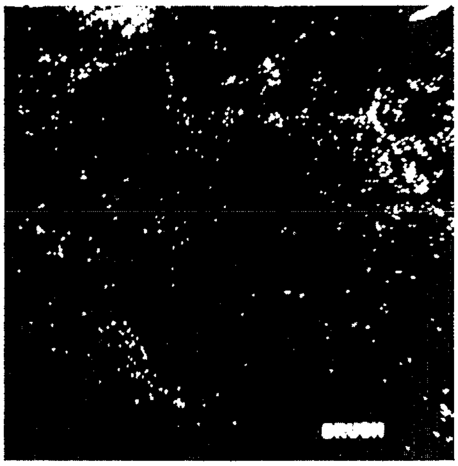
THE WATER CLASS (RHONE RIVER)



HUMID VEGETATION

MATRA-ESPACE (LABORATOIRE DE TRAITEMENT DES IMAGES)

MOUNTAIN FORESTS



BRUSH

RESULTS OF CLUSTERING IN THE IMAGE SPACE
A FEW CLASSES IN ISOLATION FROM THE AVIGNON SCENE