PERGAMON

Pattern Recognition 36 (2003) 1799–1811

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding

I. Epifanio[a,*], J. Gutiérrez[b], J. Malo[c]

[a] *Departament de Matemàtiques, Universitat Jaume I, Campus del Riu Sec, 12071 Castelló, Spain*
[b] *Departament d'Informàtica, Universitat de València, Dr. Moliner 50, 46100 Burjassot, Spain*
[c] *Departament d'Òptica, Universitat de València, Dr. Moliner 50, 46100 Burjassot, Spain*

Received 20 August 2001; received in revised form 5 June 2002; accepted 9 September 2002

Abstract

Two types of redundancies are contained in images: statistical redundancy and psychovisual redundancy. Image representation techniques for image coding should remove both redundancies in order to obtain good results. In order to establish an appropriate representation, the standard approach to transform coding only considers the statistical redundancy, whereas the psychovisual factors are introduced *after* the selection of the representation as a simple scalar weighting in the transform domain.

In this work, we take into account the psychovisual factors in the definition of the representation together with the statistical factors, by means of the perceptual metric and the covariance matrix, respectively. In general the ellipsoids described by these matrices are not aligned. Therefore, the optimal basis for image representation should simultaneously diagonalize both matrices. This approach to the basis selection problem has several advantages in the particular application of image coding. As the transform domain is Euclidean (by definition), the quantizer design is highly simplified and at the same time, the use of scalar quantizers is truly justified. The proposed representation is compared to covariance-based representations such as the DCT and the KLT or PCA using standard JPEG-like and Max-Lloyd quantizers.

Keywords: Image compression; Transform coding; Statistical redundancy; Psychovisual Redundancy; Perceptual metric

1. Introduction

The basis selection problem is ubiquitous in many fields of science. In physics the dynamics of a system is described by the eigenfunctions of its hamiltonian [1]. In image science the problem is finding the appropriate basis that optimally represents the features of the image [2].

In many pattern recognition and image processing problems taking into account the probability density function (PDF) of the signal is enough given the requirements of the application [2,3]. In this case, the basis selection problem consists of looking for the set of functions that factorize the PDF [2–4]. If the process is Gaussian the problem is reduced to an eigenvalue problem on the covariance matrix (the second-order term in the expansion of the PDF). Then the solution is just the classical principal component analysis (PCA) or Karhunen–Loève transform (KLT) [2–6]. In the physics analogy, this Gaussian (covariance-only) problem is similar to linear systems of two bodies (second order correlations) such as the harmonic oscillator, where the eigenfunctions are plane waves [1]. Recently in image analysis, some higher order moments of the PDF

* Corresponding author. Tel.: +34-964728390; fax: +34-964728429.
E-mail address: epifanio@uji.es (I. Epifanio).

0031-3203/03/$30.00
PII: S0031-3203(02)00325-4

are being taken into account in the so-called independent component analysis (ICA) [4,7,8]. In most practical cases (specially those dealing with natural images) fixed transforms such as the DCT or wavelets are used because of their similarity to the optimal PCA or ICA solutions [7–11].

However, in some problems taking into account the distribution of the data is not enough. For instance, in problems involving natural images perceived by humans, such as in multimedia applications, the different dimensions of the image space are not equally significant. In this case the space is highly non-Euclidean due to the particular properties of human perception [12–16].

This is the case in the transform coding approach for image coding. In transform coding the image has to be represented in a meaningful space before its components are scalarly quantized [5]. The aim of the transform is removing the redundancies between the coefficients of the image in order to allow a fair scalar quantization.

In the natural images there are two kinds of redundancies: statistical and psychovisual. As it is widely known, the statistical redundancy is related to the fact that not every sample or feature in a signal is equally important because of the statistical relations between the samples of the signal. In the same way, from the perceptual point of view not every sample or feature is equally important: not every scale, texture or color component has the same relevance for the human visual system (HVS).

In the image coding problem the appropriate representation is the one that removes both redundancies.

However, the standard approach to transform coding looks for the appropriate representation just by taking into account the statistical redundancy through the diagonalization of the covariance matrix [5,9]. In the standard approach the psychovisual factors are empirically introduced *after* the selection of the representation as a simple scalar weighting in the transform domain [17–19].

The main idea of this work is taking into account the psychovisual factors in the definition of the representation together with the statistical factors. To this end, we use the perceptual metric matrix, which describes the non-uniformity of the image space from the perceptual point of view [14,20], together with the covariance matrix used in the standard approach. Of course, this is just a second order approach from both perceptual and statistical points of view. The idea could be extended to higher order interactions taking into account higher order moments in the expansion of the PDF and higher order terms in the expansion of the non-linear response of the HVS. In this work we take this second order approach just as an example to illustrate the benefits of considering the HVS in the definition of the image representation.

Geometrically, the ellipsoids described by the covariance and the perceptual metric are not aligned. Therefore, the optimal basis for image representation should simultaneously diagonalize both matrices. The consideration of the perceptual metric at this level can be seen as a particular transform of the input data to meet the linear model assumptions taken by the standard approach [21]. The proposed approach to the basis selection problem has several advantages in the particular application of image coding. On the one hand, the distortion metric in the transform domain is diagonal by definition so the use of scalar quantizers is truly justified (which is not in the standard approach). On the other hand, the quantizer design is highly simplified because the final transform domain is Euclidean.

According to the second order approach selected here, the proposed representation is compared to standard covariance-based representations such as the DCT and the KLT or PCA using standard JPEG-like and Max-Lloyd quantizers. As stated above, a comparison with other linear transforms such as ICA solutions or wavelet transforms would imply taking into account higher order terms in the expansion of the non-linear behavior of the HVS. We consider this is beyond the scope of this work.

The outline of this paper is as follows: in the next section covariance and perceptual matrix are presented. These matrices describe the statistical and perceptual relations among coefficients and are jointly used to decorrelate coefficients statistically and perceptually. Section 3 elaborates on the assumed visual model that permits the perceptual matrix definition. Section 4 explains the technique used for the simultaneous diagonalization of the perceptual metric and the covariance matrix. In Section 5 several experiments are conducted and their results discussed. The suggested transformation is compared with a completely statistical approach (KLT) and the standard JPEG. Experiments illustrate the importance of including the anisotropies of the space together with the shape of the PDF. In the last section some conclusions are drawn.

2. Matricial expressions for the statistical and perceptual relations among coefficients and their interpretation

The purpose of the transform is to remove both statistical and perceptual relations among coefficients. Firstly, the aforementioned relations have to be quantified and formally formulated. Two matrices will define these interactions: the covariance matrix for the statistical ones and the perceptual metric matrix for the psychovisual ones.

2.1. Covariance matrix

The luminances of an image in the spatial domain can be represented by an array, A. This array can be seen as an ensemble of random variables. The *statistical deviations* from a point (or image), A_0, can be described in that domain

by the matrix $\Sigma_A(A_0)$,

$$\Sigma_A(A_0) = E[(A - A_0)(A - A_0)^{\mathrm{T}}], \quad (1)$$

where E stands for the expected value and T stands for transposition. When A_0 is the mean, this matrix will be simply referred as the covariance matrix Σ_A in the A (spatial) domain.

The diagonal element of the covariance matrix, σ_{ii}^2, is the variance of the ith coefficient. The element σ_{ij} represents the covariance between the ith and jth coefficients, a second-order relation.

2.2. Perceptual metric matrix

If a L^2 norm is assumed [12], the perceptual deviation from A_0 due to a distortion ΔA is determined by [14]

$$d(A_0, A_0 + \Delta A)^2 = \Delta A^{\mathrm{T}} W_A(A_0) \Delta A = \sum_i W_{ii} \Delta A_i^2 + \sum_{i \neq j} W_{ij} \Delta A_i \Delta A_j, \quad (2)$$

where $W_A(A_0)$ is the perceptual metric of the domain A at the point A_0.

The diagonal components of the perceptual metric represent the contribution of each coefficient to the global distortion. Non-zero off-diagonal elements induce additional contributions to the distortion due to combinations of deviations in different dimensions, i.e. they represent perceptual interactions between features that modify the distortion perception. This is a convenient way to represent what is commonly referred to as *masking* [22,23]: a distortion in one coefficient could mask the subjective distortion in another one. This kind of interaction is not usually considered in the Transform Coding overall distortion.

2.3. Geometrical interpretation of these matrices

As covariance and perceptual metric matrices are positive definite matrices, two ellipsoids can be associated respectively to each one (see Fig. 1). On the one hand, Σ describes the shape of the distribution of image samples around A_0, it gives information about the data dispersion. On the other hand, W describes the shape of the (ellipsoidal) locus of perceptually equidistant patterns from A_0 (constant distortion in Eq. (2)). W describes the underlying geometry of the feature space: off-diagonal elements in W represent second-order perceptual interactions. It is important to stress two facts:

- The ellipsoids described by the covariance and the metric are not aligned with the axes of the spatial domain representation. This is so in natural images because, on the one hand, the luminance values, A_i, in each spatial location, i, have strong statistical correlations with the values, A_j, in neighboring locations, j [6,9]. On the other hand, there is a strong perceptual correlation between neighboring pixels as well [24], giving rise to strongly non-diagonal metrics in the spatial domain [14,20].

  These interactions between coefficients mean that the spatial domain representation is quite inadequate for a scalar quantization. The core of transform coding is that some transform to a new domain, a, is needed in order to remove these correlations prior to the scalar quantization.
- In general these ellipsoids are not aligned between them. This means that the standard KLT, PCA or DCT approach based on the diagonalization of Σ does not imply a diagonalization of W.

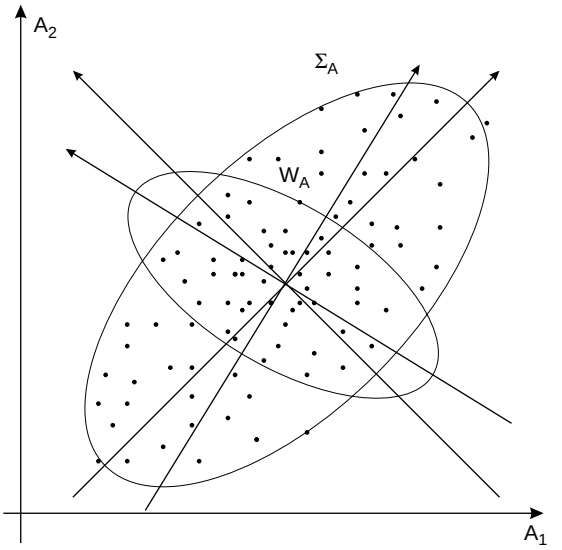


Fig. 1. Ellipsoids describing the data distribution and the space geometry. Note the different meaning of the ellipsoids defined by Σ and W. While the most important features from the statistical point of view are given by the directions of highest variance (i.e. the major axis of the ellipsoid defined by Σ), the most important features from the perceptual point of view are given by the directions in which the discrimination is highest (i.e. the minor axis of the ellipsoid defined by W). This is why in pattern recognition applications where the distance is completely given by the data distribution (when there is no additional W), the metric is defined by Σ^{-1} (the Mahalanobis metric [2,3]).

As shown in Section 5.1, the classical quantizer design methods assume a diagonal distortion metric. If W is not diagonal in the selected representation these results cannot be strictly applied.

This is why, in principle, the standard approach may be improved taking into account W in the selection of the representation.

Therefore, the appropriate representation is the one that not only diagonalizes Σ, but also W. In this simultaneous diagonalization case, the scalar quantization will be effective (because the statistical and the perceptual relations will have

been removed), and the classical quantizer design results will be strictly applicable (because the distortion metric will be diagonal).

2.4. Matricial changes under linear mappings

In order to look for the appropriate set of transforms that diagonalize Σ and W, it is worth knowing how these matrices change when a linear transformation is applied.

Let A be the vector mapped into a by a linear transformation L, $a = L \cdot A$. Then, the covariance matrix in the domain, a, is [2]

$$\Sigma_a = L \cdot \Sigma_A \cdot L^{\mathrm{T}} \tag{3}$$

and the metric in $a_0 = L \cdot A_0$ is (see Ref. [25] and the appendix):

$$W_a(a_0) = (L^{-1})^{\mathrm{T}} \cdot W_A(A_0) \cdot L^{-1}. \tag{4}$$

3. Vision model and perceptual metric matrix

The previous step, before defining the perceptual metric matrix, is to introduce the perceptual model which has been assumed. The standard model of human low-level image analysis has two basic stages [22,26,27],

$$A \xrightarrow{T} a \xrightarrow{R} r \tag{5}$$

in which the input image, A (array of luminances in the *spatial domain*), is first transformed into a vector, $a = T \cdot A$ (with components a_f, $f = 1 \ldots M$), in a local frequency domain (the *transform domain*) using a linear filter bank, T, and then a set of mechanisms responds to each coefficient of the transformed signal giving an output, $r = R(a)$, which is the image representation in the *response domain*.

The first linear perceptual transform T is similar to the class of transforms employed in image coding. The local DCT has been used here as a model of the perceptual transform T, followed by an amplitude normalization of the coefficients. The transform coefficients are expressed in contrast (amplitude over mean luminance of the block, the DC coefficient) [15]. As not all the basis functions of the transform T are equally perceived, additional processing (the transform R) is included to explain these heterogeneities. The HVS models assume that all the components of the r vector are equally important and there is no perceptual interaction between them [12,22,23], therefore the response domain is Euclidean. The response model that has been used here is basically the energy-normalization model of Refs. [12,22,23,27] where the energy of each transform coefficient (in contrast) is normalized by a weighted sum of the energy of its neighbors. The dependence with the neighbor coefficients is given by the convolution with an interaction kernel h,

$$r_i = \frac{\alpha_i}{100}|a_i| + \alpha_i \frac{|a_i|^2}{\beta_i + (h * |a|^2)_i}, \tag{6}$$

where the index i corresponds to spatial frequency. Fig. 2 shows the parameters of this non-linear energy normalization model and an example of the response for some basis functions of different frequencies.

The values of α and β have been fitted to reproduce amplitude discrimination thresholds without inter-coefficient masking measured at our lab (Legge-like experimental data [28,29]). A frequency-dependent Gaussian kernel has been heuristically introduced according to the results of Refs. [12,22,23,30],

$$h_{ij} = k_i \cdot \mathrm{e}^{-(f_i - f_j)^2/\sigma(f_i)^2}, \tag{7}$$

where f_i is the spatial frequency meaning of the coefficient a_i, $\sigma(f_i)$ is the variable width of the kernel, $\sigma(f_i) = \frac{1}{3}|f_i| + 0.05$, with $|f|$ in cycl/deg, and k_i is a constant to obtain a unit-volume kernel.

Assuming the above T and R transforms and a Euclidean (identity) perceptual metric in the response domain, the perceptual metric in the local frequency domain can be obtained by using the properties of a Riemannian metric when a change of co-ordinate systems is considered [14,25].

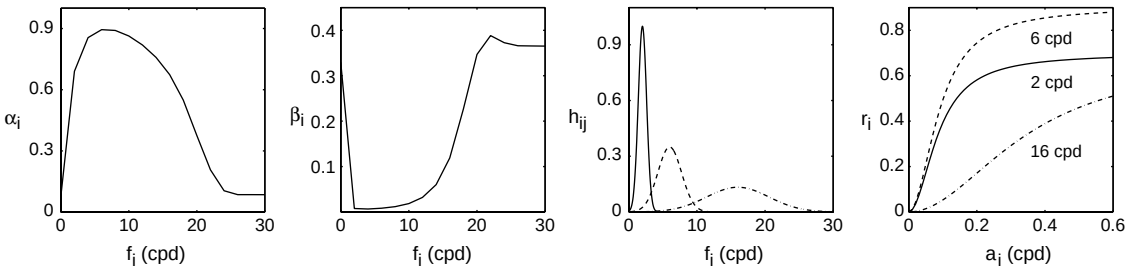


Fig. 2. Parameters of the vision model and non-linear response functions. The values in these figures assume that the amplitude of the coefficients is expressed in contrast (amplitude over mean luminance). The response examples of the last figure show the basic (sigmoid) behavior of Eq. (6), but they are not general because the response to one coefficient depends on the background (it depends on the value of the neighbor coefficients). These particular curves were computed for the particular case of no additional masking pattern (zero background).

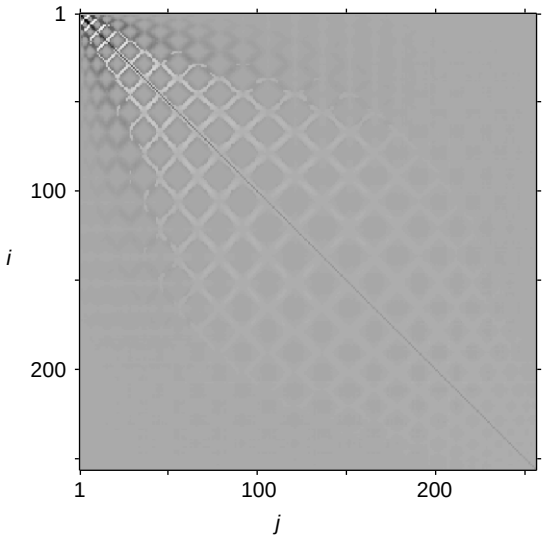


Fig. 3. Perceptual metric in the transform (DCT) domain using $\bar{J}$. Dark and light pixels indicate positive and negative values, respectively. Notice the non-diagonal nature of W. An exponent has been applied to enhance the visibility of the off-diagonal elements.

If J denotes the Jacobian matrix of the transformation R and

$$J_{ij} = \frac{\partial R_i}{\partial |a_j|} = \frac{\alpha_i}{100}\delta_{ij} + 2\alpha_i \left(\frac{|a_i|}{\beta_i + (h * |a|^2)_i}\delta_{ij} - \frac{|a_i^2 \cdot a_j|}{(\beta_i + (h * |a|^2)_i)^2} h_{ij} \right), \quad (8)$$

the metric in the corresponding transform domain a is

$$W_a = J^{\mathrm{T}} \cdot J. \quad (9)$$

Fig. 3 shows the perceptual metric that comes from expression (9), replacing J with the average of the Jacobian matrices of a set of images, $\bar{J}$.

The qualitative meaning of the metric elements, which give the relations between different coefficients of the feature vectors, depends on how the 2D DCTs are scanned to construct the 1D feature vectors. Fig. 4 explains the zigzag scanning that has been applied to the 2D DCTs. According with the zigzag scanning, the frequency meaning of the diagonal elements of h, J, W and Σ progressively increases from zero to the Nyquist frequency.

From Fig. 3 it is clear that the relative perceptual relevance of the transform coefficients highly depends on frequency (the diagonal of W has a low-pass shape), i.e. the frequency domain is perceptually anisotropic. It is also clear that transform coefficients are not perceptually independent because W is not diagonal, i.e. the perceptually privileged directions of the frequency domain are not aligned with the axes of the space. This implies that an additional transform is needed to remove the perceptual (as well as statistical) correlation between the transform coefficients and process them individually afterwards.

A point worth noting is that J (and then the metric) is input-dependent and to be rigorous, the decorrelation transform should be local. However, only one metric matrix can be established for representing the rest since the metric W_a does not vary greatly at this domain. Results shown in Section 5 corroborate this supposition. This assumption is analogue to the stationarity assumption in order to consider a single Σ. The average of the Jacobian matrices of a set of images, $\bar{J}$, has been chosen to calculate the model metric which has been used in the experiments. Other options such as the Jacobian matrix of the mean of transformed images were explored for summarizing the metrics in sole one matrix, but the final results did not change substantially. The

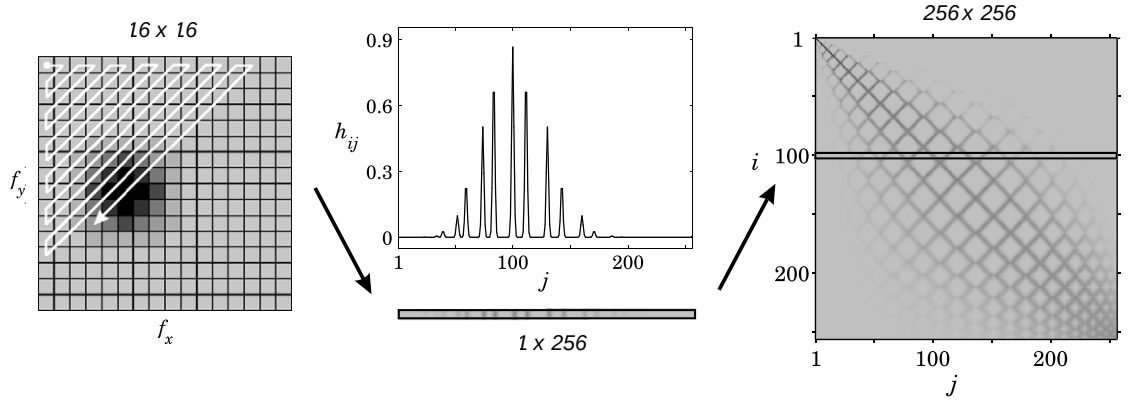


Fig. 4. DCT zigzag scanning. A convenient way to scan the 2D DCT is the zigzag scheme used in the JPEG standard [18] because it groups together coefficients with similar frequency. In the final vector the frequency progressively increases from the DC component in the first coefficient, up to the diagonal Nyquist frequency for the last coefficient, i.e. (32,32) *cpd* if the sampling frequency in each direction is 64 *cpd*. In Fig. 4 the kernel for the coefficient $f_0 = (18, 12)$ *cpd* is zigzag scanned and introduced in its corresponding position in h. In accordance with this scanning scheme, the coefficient of frequency f_0 is in the position $i = 100$, so the values that give the relations of a_{100} with a_j for $j = 1, \ldots, 256$ form a row vector that goes in the row $i = 100$ to be applied on the input column vectors to give $(h * |a|^2)_{100}$.

condensation into one metric allows us to design a global (non-local), linear transform which diagonalizes both covariance and perceptual metric matrices. In this way, a domain where features are statistically and perceptually uncorrelated is attained.

4. Simultaneous diagonalization of the perceptual metric W and the covariance matrix Σ

The idea of a simultaneous diagonalization of both matrices follows the principles behind transformations used for satisfying the hypotheses which support a well-established theory, for instance the Box–Cox transforms to normality [21]: if the data do not meet the assumptions required by the model (scalar processing) you should transform the data instead of devising a more complex model (vector processing). In this case the data are first transformed to a perceptually Euclidean domain before the standard KLT is applied to get the final statistically and perceptually decorrelated domain.

Let A be a vector in the spatial domain. The diagonalization process is as follows:

(1) Firstly, W_A is *whitened* by a perceptual transform called T_P defined by

$$a^p = T_P \cdot A = J \cdot T \cdot (A - \bar{A}), \tag{10}$$

where $\bar{A}$ is the mean of a set of images, T is the filter bank of the vision model of Section 3 (a DCT in our implementation) and J is the jacobian of the perceptual non-linearity (Eq. (8)). In this way, Σ_{a^p} is $J \cdot T \cdot \Sigma_A \cdot T^{\mathrm{T}} \cdot J^{\mathrm{T}}$ (Eq. (3) is applied) and W_{a^p} is the identity (Eqs. (4) and (9) are applied). With this transform the image, A, is mapped into a perceptually decorrelated domain a^p.

(2) Secondly, a statistical orthonormal transformation to diagonalize Σ_{a^p} is applied: the KLT. That is,

$$a^{sp} = T_K \cdot a^p, \tag{11}$$

where T_K^{T} and Λ are the eigenvector and eigenvalue matrices of Σ_{a^p} as

$$\Sigma_{a^p} \cdot T_K^{\mathrm{T}} = T_K^{\mathrm{T}} \cdot \Lambda \quad \text{and} \quad T_K \cdot T_K^{\mathrm{T}} = I. \tag{12}$$

With this second transform the perceptually decorrelated vector, a^p, is mapped into a statistically and perceptually decorrelated domain, a^{sp}. Thus,

$$\Sigma_{a^{sp}} = T_K \cdot \Sigma_{a^p} \cdot T_K^{\mathrm{T}} = \Lambda, \tag{13}$$

$$W_{a^{sp}} = (T_K^{\mathrm{T}})^{-1} \cdot W_{a^p} \cdot T_K^{-1} = (T_K^{\mathrm{T}})^{-1} \cdot I \cdot T_K^{-1} = I. \tag{14}$$

Hence, both matrices are diagonalized. The combination of the two (perceptual and statistical) steps gives the overall transformation,

$$a^{sp} = T_{KP} \cdot A = T_K \cdot T_P \cdot A = T_K \cdot J \cdot T \cdot (A - \bar{A}). \tag{15}$$

Fig. 5 shows a two-dimensional (two coefficients) example of this process.

5. Compression results

In this section we compare the proposed representation, T_{KP}, with the standard covariance-based representations such as the KLT or PCA, T_K, and the DCT, T, in image coding applications.

Transform coding for image compression has two associated problems [5]. First you have to select a certain transform for image representation, and then you have to design the quantizer in this representation domain.

For a fair comparison of the representations, we use several standard designs for the quantizer. We use rate-distortion based quantizers with two design criteria: (1) minimizing the *mean square error* (MSE) (either Euclidean or Perceptual) and, (2) restricting the *maximum perceptual*

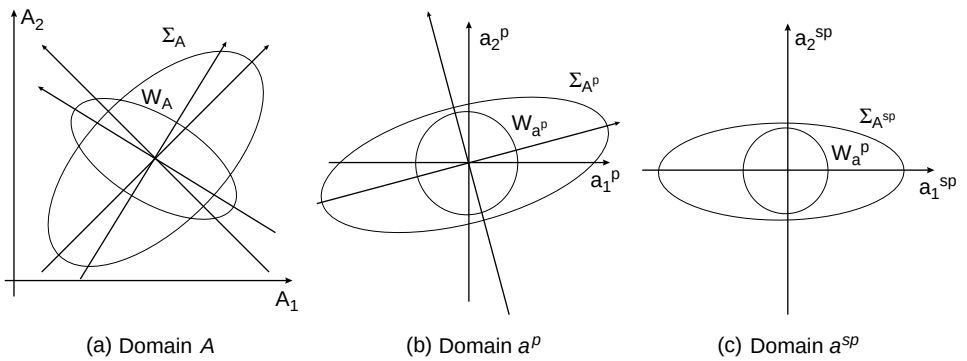


Fig. 5. Simultaneous diagonalization of W_A and Σ_A. Note that initially A_1 and A_2 are highly correlated from both, perceptual and statistical, points of view. The first transform, T_p, implies the complete decorrelation from the perceptual point of view. As T_P involves a DCT (J is formulated in a DCT domain) it also implies some reduction of the statistical correlation because the KLT basis for natural images is similar to the DCT basis. Strictly speaking the complete decorrelation is only obtained when the second transform, T_K, is applied on the a^p domain. (a) Domain A, (b) Domain a^p, and (c) Domain a^{sp}.

error (MPE). The rate-distortion theory used in transform quantizer design assume that the global distortion in the transform domain is given by a weighted sum of the distortions in each coefficient with no interaction between coefficients (i.e. Eq. (2) in the transform domain with a diagonal metric).

The results for a DCT-based JPEG are included as a useful reference. In this case, a bit allocation based on the human *contrast sensitivity function* (CSF) is used as recommended by the JPEG standard [17].

Section 5.1 reviews the optimal results for MSE and MPE criteria. Implementation details are analyzed in Section 5.2. Finally, examples of the compression results on some standard images are presented in Section 5.3.

5.1. Quantizer design

The scalar quantizers applied to the transform vector, $X = (x_1, \ldots, x_M)$, are defined by the *bit allocation*, N_i, the number of quantization levels to encode the coefficient, x_i, and by the *point densities*, $\lambda_i(x_i)$, the densities of quantization levels to encode each coefficient, x_i.

The rate-distortion theory used in transform quantizer design assumes that the global distortion in the transform domain is given by a weighted sum of the distortions in each coefficient with no interaction between coefficients,

$$D^2 = \sum_{i=1}^{M} D_i^2. \tag{16}$$

The standard quantizer design procedure consists of obtaining the optimal λ_i to minimize each D_i and then choosing N_i to obtain the same error contribution per coefficient. This standard procedure may consider amplitude dependent weights in the distortion for each coefficient [31], but it cannot deal with interactions between coefficients (Eq. (16) must hold). This is why, the distortion metric should be diagonal to allow a straightforward application of the formalism.

In this section we review the results for bit allocation and point densities using two standard design criteria for D_i [31]: (1) minimizing the MSE (either Euclidean or Perceptual) and, (2) restricting the MPE. The details may be found in Refs. [5,31–33].

Any input coefficient, x_i, pertaining to the quantization region, R_{ij}, is represented by the quantization level y_{ij}, so the MSE in each coefficient is

$$D_{i,MSE}^2 = \sum_{j=1}^{N_i} \int_{R_{ij}} (x_i - y_{ij})^2 f_{x_i}(x_i)\,\mathrm{d}x_i, \tag{17}$$

where N_i is the number of quantization levels, y_{ij} is the jth quantization level for the coefficient x_i, R_{ij} is the quantization region corresponding to the jth level and $f_{x_i}(x_i)$ is the PDF of x_i. This widely used distortion can be modified in order to consider the physical meaning of the coefficient x_i and hence a perceptual MSE can be defined,

$$D_{i,PMSE}^2 = \sum_{j=1}^{N_i} \int_{R_{ij}} (x_i - y_{ij})^2 W_i(y_{ij}) f_{x_i}(x_i)\,\mathrm{d}x_i, \tag{18}$$

where $W_i(x)$ is a weight that depends on the input x.

From the asymptotic quantization approach (high resolution regular quantizers), the point density function that minimizes the distortion, $D_{i,PMSE}^2$, is

$$\lambda_{i,PMSE,opt}(x_i) = \frac{(W_i(x_i) f_{x_i}(x_i))^{1/3}}{\int (W_i(x) f_{x_i}(x))^{1/3}\,\mathrm{d}x} \tag{19}$$

and the asymptotic expression for the average distortion with the optimal point density function is

$$D_{i,PMSE,opt}^2 = \frac{\sigma_{x_i}^2}{12N_i^2} \left(\int (W_i(\sigma_{x_i} x_i) \tilde{f}_{x_i}(x_i))^{1/3}\,\mathrm{d}x_i \right)^3 = \frac{\sigma_{x_i}^2}{N_i^2} H_i, \tag{20}$$

where $\tilde{f}_{x_i}(x_i) = \sigma_{x_i} f_{x_i}(\sigma_{x_i} x_i)$ is the normalized unit-variance pdf. Analogous expressions for MSE can be obtained by taking $W_i = 1$.

The optimal bit allocation per coefficient, b_i, is obtained solving N_i from Eq. (20) and assuming constant distortion per coefficient:

$$b_i = log_2(N_i) = \frac{B}{M} + \frac{1}{2} log_2(\sigma_{x_i}^2 H_i) - \frac{1}{2M} \sum_{i=1}^{M} log_2(\sigma_{x_i}^2 H_i), \tag{21}$$

where B is the total number of available bits.

The aforementioned average design criterion cannot guarantee a satisfactory subjective performance on a particular image. In order to prevent high perceptual errors on individual images arising from outlier coefficient values, the overall performance could be assessed by a *worst-case* value distortion [5]. In order to do so, it has been proposed to restrict the MPE in each coefficient [31–33]. The MPE criterion implies a perceptually uniform distribution of the available quantization levels. A key factor of a worst-case measure is that the values depend only on the support of the pdf but not on the actual distribution [5]. Let us examine the analogous expressions adopted in the MPE-based approach by the equations presented above.

If a given coefficient is represented by N_i quantization levels distributed according to a density, $\lambda_i(x_i)$, the maximum Euclidean quantization error at an amplitude, x_i, will be bounded by half the Euclidean distance between two levels:

$$(x_i - y_{ij}) \leqslant \frac{1}{2N_i\,\lambda_i(x_i)}. \tag{22}$$

Assuming a generic diagonal frequency and amplitude-dependent metric, the MPE at that amplitude will be related to the metric and the density of levels:

$$D^2_{i,MPE}(x_i) = \frac{W_i(x_i)}{4N_i^2\lambda_i^2(x_i)}. \tag{23}$$

According to this, the maximum perceptual distortion bound is constant over the amplitude range only if the point density varies as the square root of the metric, so the optimal quantizers under the MPE criterion are given by

$$\lambda_{i,MPE,opt}(x_i) = \frac{W_i(x_i)^{1/2}}{\int W_i(x)^{1/2}\,\mathrm{d}x} \tag{24}$$

and the MPE with the optimal point density function is

$$D^2_{i,MPE,opt} = \frac{1}{4N_i^2}\left(\int W_i(x_i)^{1/2}\,\mathrm{d}x_i\right)^2. \tag{25}$$

Fixing the same maximum distortion for each coefficient, $D^2_{i,MPE,opt} = k^2$, the optimal bit allocation is given by

$$b_i = log_2 N_i = log_2\left(\frac{1}{2k}\int W_i(x_i)^{1/2}\,\mathrm{d}x_i\right). \tag{26}$$

Although all these expressions only hold in the high resolution case, they often turn out to be a reasonable approximation even in the medium to low resolution cases.

5.2. Experiments and implementation details

The proposed transform with two quantizers was compared with JPEG and a KLT (only statistical) approach. A set of well-known images (most of them can be found in http://sipi.usc.edu/services/database/Database.html), 8 bits/pixel has been used as a sample. These images have been partitioned into 16×16 blocks and vectors obtained by ordering of the pixels within the block have been used to estimate the covariance matrix. Although for natural images the DCT is a very close approximation of KLT [9,10], the KLT has been calculated. The images in Section 5.3 are also segmented into blocks to which the computed KLT is applied. The transform coefficients are quantized with a simple MSE quantizer (Eqs. (19) and (21) with $W_i = 1$). This procedure will be referred to as T_K-MSE.

Two options for the quantizer design are contemplated with our transform, T_{KP}. The first one is PMSE quantizer. The weight $W_i(x)$ that appears in Eqs. (19) and (21) represents the diagonal element of the perceptual metric matrix which corresponds with the coefficient that is being quantized. As the perceptual metric matrix in the a^{sp} domain is the Identity, PMSE is simply reduced to MSE with our transform. This option will be referred to as T_{KP}-MSE. On the other hand, MPE quantizer is also greatly simplified. As $W = I$, it is converted into a simple uniform quantizer with uniform bit allocation. Regarding to bit allocation, the distribution of the number of quantization levels depends exclusively on the range of each coefficient. This later option will be referred to as T_{KP}-MPE.

Therefore, the comparison between the transforms T, T_K and T_{KP} is made comparing the performance of the schemes JPEG, T_K-MSE, T_{KP}-MSE and T_{KP}-MPE.

The aforementioned results (Section 5.1) are strictly applicable under the high rate approach [5]. The actual low resolution MSE quantizers (for T_K-MSE and T_{KP}-MSE), have been obtained using the LBG method [5,34] initialized with the asymptotic results. The final quantizers were quite consistent with the asymptotic assumptions. As said above, for T_{KP}-MPE we just have a uniform quantizer of the a^{sp} domain.

The actual bit allocation for all alternatives was determined by means of a greedy integer-constrained allocation algorithm [5] based on the sequential allocation of one bit to the coefficient with the largest distortion in each iteration. Distortions are given by Eqs. (20) and (25), according to the corresponding case. In all schemes, the DC coefficient is separately encoded using DPCM.

As was mentioned in Section 3, the average of the Jacobian matrices (J) of the sample images ($\bar{J}$) has been used to compute the perceptual metric matrix. Table 1 shows how these Jacobian matrices vary. The 2-*norm* (a matrix norm) [35] of the difference between $\bar{J}$ and the Jacobian matrix of a set of images (different from those used to compute $\bar{J}$) was calculated. This norm is normalized by the 2-*norm* of $\bar{J}$. A summary of these quantities are displayed in Table 1. These values are quite small. In addition, the acceptable experimental results in Section 5.3 confirm that the assumption is satisfactory.

Table 1
Summary of the variation of J

$\|J-\bar{J}\|/\|\bar{J}\|$	Mean	Median	Std. deviation
	1.58	0.98	1.83

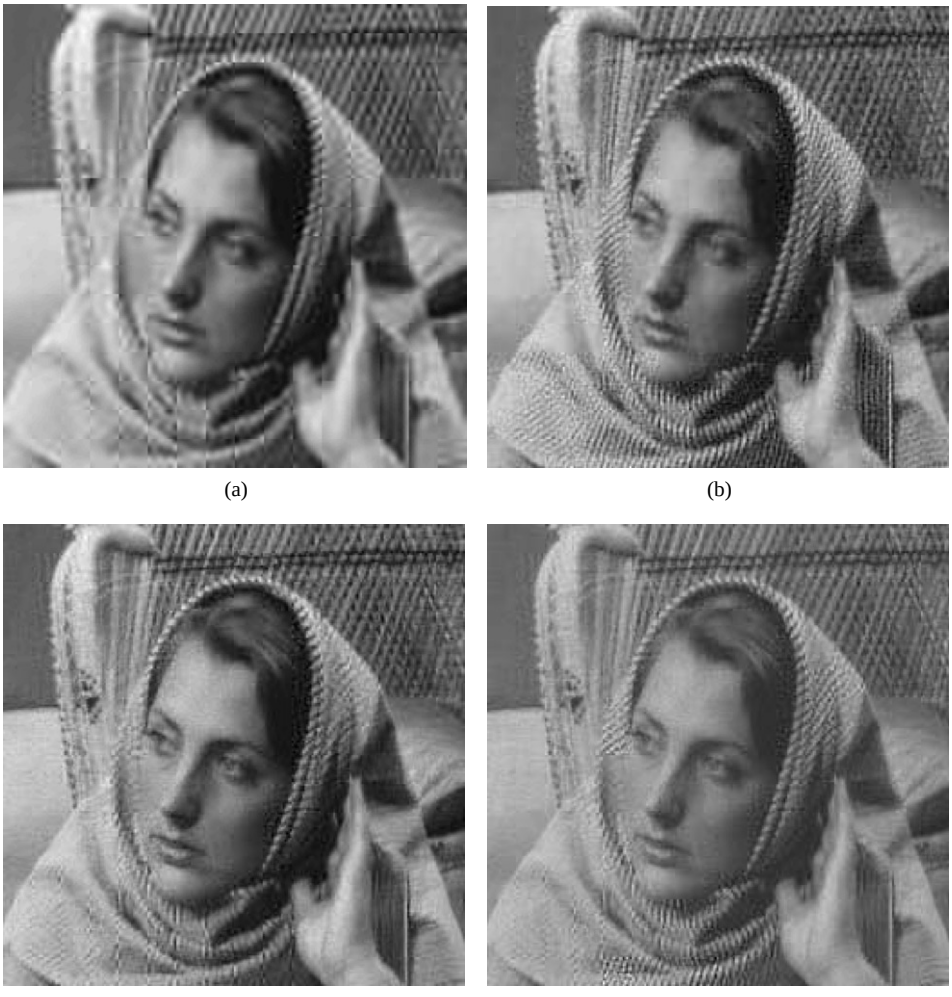
5.3. Decoded images

In order to evaluate the performance of the schemes, compression results on three standard images (not included in the training set) are displayed. These images are: Barbara, Peppers and Baboon. All schemes were used at the same compression ratio (0.5 bpp).

Goodness of the different schemes was evaluated both qualitatively and quantitatively. Fig. 6 displays the original images. Figs. 7 and 8 show the decoded images, using the different schemes.

(a) (b)

Fig. 6. Grayscale images used to assess the performance: (a) (a close up of) Barbara and (b) peppers.



(a) (b)

(c) (d)

Fig. 7. Barbara at 0.5 bits/pixel. (a) JPEG, (b) T_K-MSE,(c) T_{KP}-MSE and (d) T_{KP}-MPE.

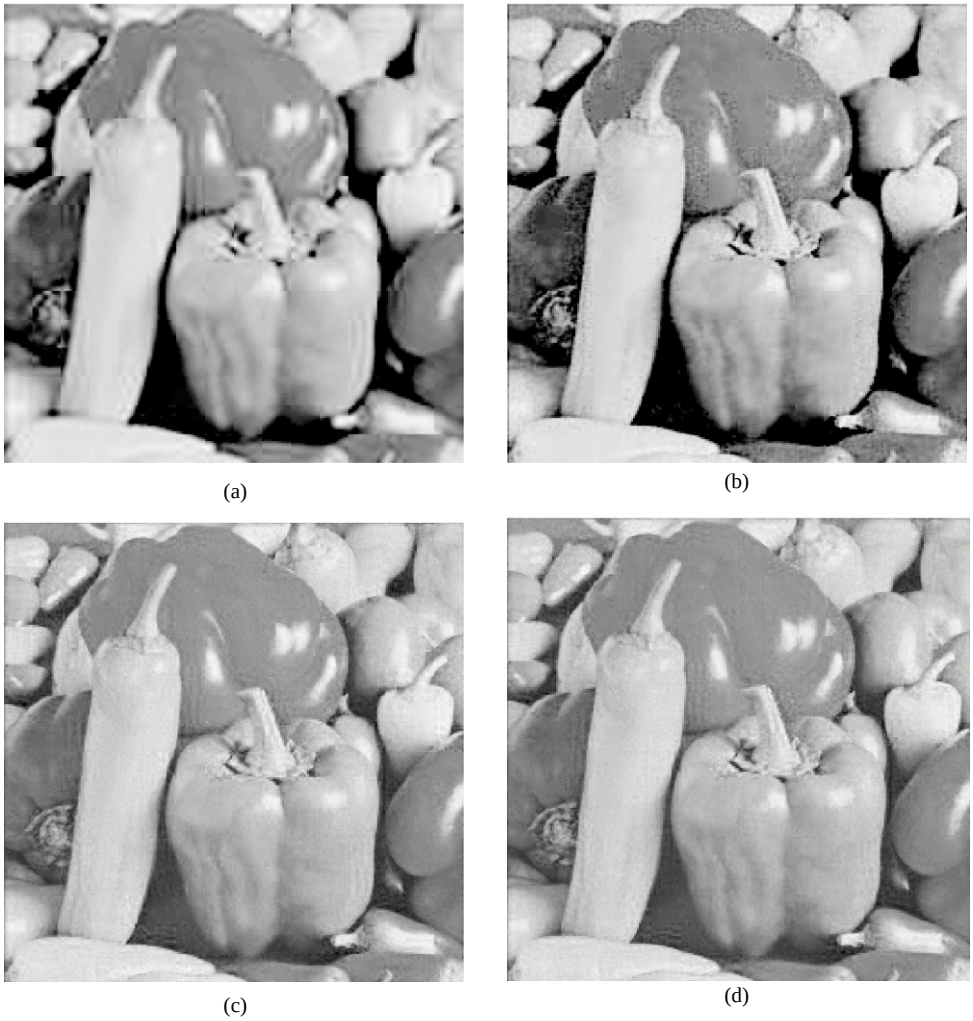


Fig. 8. Peppers at 0.5 bits/pixel. (a) JPEG, (b) T_K-MSE,(c) T_{KP}-MSE and (d) T_{KP}-MPE.

These images allow us to get a feeling for the corresponding subjective quality. For comparison purposes, we have also employed some quality measures. A widely used objective quality measure is the peak-to-noise ratio or PSNR. The interpretation of this measure is that the larger the PSNR, the better the quality of the processed image. However, this easy computing measure is not always an adequate measure of the subjective quality [12–14,19]. As seen in Section 3, the visual perception of the HVS is not as simple. Therefore, two objective quality measures based on human visual perception have been also considered (Eq. (2) and [14]). Here, the values $\beta = 2$ and ∞ have been adopted by the so-called summation index. These distances will be denoted as D_2 and D_∞. The closer they are to zero, the better the quality is. Results are displayed in Table 2.

At this point we will analyze the results. Figs. 7 and 8 show several behaviors. JPEG-images are blurred due to truncation of high-frequency coefficients. High-frequency and high-contrast details (Barbara's clothes or Barbara's armchair are clear examples) are removed. Moreover, the blocking effect is quite noticeable (see Fig. 7). On the other hand, graininess (smooth zones in peppers or the Barbara's hand) and somewhat strong blocking effect can be observed in KLT-images. The first artifact may be due to coarse quantization of some coefficients. With regard to the transform presented here, blocking artifacts are not so easily perceived and main details are preserved. Slight graininess appears, but it is not as noticeable as that of the T_K-MSE scheme. In general, both T_{KP}-MSE and T_{KP}-MPE give rise to decoded images with superior quality to the ones obtained by JPEG or T_K-MSE. There is not a significant difference

Table 2
Quality measurements for the different approaches at 0.5 bits/pixel

Image	Measure	JPEG	T_K-MSE	T_{KP}-MSE	T_{KP}-MPE
Barbara	PSNR	23.48	25.87	25.07	25.48
	D_2	29.85	64.17	13.91	13.78
	D_∞	6.42	24.18	3.87	3.76
Baboon	PSNR	23.46	25.34	26.93	27.16
	D_2	59.12	54.77	14.27	12.49
	D_∞	6.55	19.89	3.74	3.71
Peppers	PSNR	29.37	32.45	33.41	33.54
	D_2	22.2	35.07	9.22	8.04
	D_∞	5.43	13.42	3.3	2.47

between the use of the MSE-quantizer or the MPE-quantizer, although the MPE criterion seems to give a slight improvement on the subjective quality (Barbara's clothes). These subjective interpretations (based on the observation of the decoded images) can be also corroborated by the objective measures displayed in Table 2. The highest values (best quality) for PSNR are reached by T_{KP}-MPE approach (except with Barbara). These values are not greatly different from the ones with T_{KP}-MSE approach. Perceptual distances (D_2 and D_∞) again validate the previous remarks. For all images and PSNR measure, JPEG gives the poorest results while if the perceptual distances are considered T_K-MSE is the worst for the most part. T_{KP}-MPE is the best with the perceptual distances.

6. Conclusions

This paper introduces a linear transform for removing second-order statistical and perceptual relations between coefficients in the transform coding context. The perceptual correlation between features of an image representation was formalized through the perceptual metric matrix in the same way as the statistical correlation is represented by the covariance matrix. The proposed transform facilitates the application of scalar quantization and the overall performance of the coding system can be accurately determined from the sum of the mean distortions of each coefficient.

In view of the pictures and the distortion results, a statistical decorrelating transform (KLT) on its own is not sufficient. However, the joint statistical and perceptual decorrelation gives rise to a better overall performance on natural imagery. According to the results presented here, a transform eliminating higher-order relations may be interesting, although its computation could be too intensive. It is certain that a non-linear and local (adaptive) transform could obtain better results at the cost of increase in processing.

Acknowledgements

The authors are grateful to Guillermo Ayala and Amelia Simó for their support. This work has been partially supported by grants CICYT BSA2001-0803-C02-02 and Fundació Caixa Castelló P1-1B2001-10.

Appendix

In this appendix we review the properties of the Riemannian metrics that are significant to our problem. A more extensive coverage of this concept is given (for example) by Dubrovin et al. [25].

Definition 1. A Riemannian metric in a region of the space $\mathbb{R}^n$ is a positive definite quadratic form defined on vectors originating at each point P of the region and depending smoothly on P.

This definition can be stated more explicitly:

Definition 2. A Riemannian metric in a region of a space, relative to arbitrary coordinates $(z_1,\dots,z_n)$ is a family of smooth functions $g_{ij}=g_{ij}(z_1,\dots,z_n)$, $i,j=1,\dots,n$, with the following two properties:

(1) the matrix (g_{ij}) is positive definite,
(2) if $(y_1,\dots,y_n)$ are new co-ordinates for the region, and $z_i=f_i(y_1,\dots,y_n)$, $i=1,\dots,n$, then relative to these new co-ordinates the Riemannian metric is represented by the family of functions $g'_{ij}=g'_{ij}(y_1,\dots,y_n)$, $i,j=1,\dots,n$, given by

$$g'_{ij}=\sum_{k=1}^{n}\frac{\partial f_k}{\partial y_i}\sum_{l=1}^{n}g_{kl}\frac{\partial f_l}{\partial y_j}. \tag{A.1}$$

The second property can be rewritten in a matricial form:

$$G_y = J^{\mathrm{T}} G_z J, \tag{A.2}$$

where G_y is the metric matrix in the domain y with coefficients g'_{ij}, G_z is the metric matrix in the domain z with coefficients g_{ij}, and J stands for the Jacobian matrix of the (inverse) function $z = f(y)$:

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \cdots & \frac{\partial f_1}{\partial y_n} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \cdots & \frac{\partial f_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial y_1} & \frac{\partial f_n}{\partial y_2} & \cdots & \frac{\partial f_n}{\partial y_n} \end{pmatrix}. \tag{A.3}$$

Positive definiteness of the matrix (g_{ij}) means simply that $\zeta^{\mathrm{T}} G \zeta > 0$ for non-zero vectors ζ, i.e. that the quadratic form is positive definite.

References

[1] J. Sakurai, Modern Quantum Mechanics, Addison-Wesley, Menlo Park, CA, 1985.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Boston, 1990.

[3] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[4] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[5] A. Gersho, R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Press, Boston, 1992.

[6] A. Akansu, R. Haddad, Multiresolution Signal Decomposition, Academic Press, Boston, 1992.

[7] B. Olshausen, D. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.

[8] A. Bell, T. Sejnowski, The independent components of natural scenes are edge filters, Vision Res. 37 (23) (1997) 3327–3338.

[9] R. Clarke, Relation between the Karhunen-Loève and cosine transforms, Proc. IEE, Part F 128 (6) (1981) 359–360.

[10] W. Niehsen, M. Brunig, Covariance analysis of motion-compensated frame differences, IEEE Trans. Circuit Systems Video Technol. 9 (4) (1999) 536–539.

[11] E. Simoncelli, B. Olshausen, Natural image statistics and neural representation, Annu. Rev. Neurosci. 24 (2001) 1193–1216.

[12] P. Teo, D. Heeger, Perceptual image distortion, Proc. First IEEE Int. Conf. Image Process 2 (1994) 982–986.

[13] J. Malo, A. Pons, J. Artigas, Subjective image fidelity metric based on bit allocation of the HVS in the DCT domain, Image Vision Comput. 15 (7) (1997) 535–548.

[14] A. Pons, J. Malo, J. Artigas, P. Capilla, Image quality metric based on multidimensional contrast perception models, Displays 20 (1999) 93–110.

[15] S. Winkler, Issues on vision modeling for video quality assessment, Signal Process. 4 (12) (1999) 2401–2417.

[16] A. Watson, J. Hu, J. McGowan, Digital video quality metric based on human vision, J. Electr. Imaging 10 (1) (2001) 20–29.

[17] G. Wallace, The JPEG still picture compression standard, Commun. ACM 34 (4) (1991) 31–43.

[18] A. Tekalp, Digital Video Processing, Prentice-Hall, Upper Saddle River, NJ, 1995.

[19] Y. Shi, H. Sun, Image Compression for Multimedia Engineering: Fundamentals, algorithms and standards, CRC Press, Boca Raton, 2000.

[20] J. Malo, R. Navarro, I. Epifanio, F. Ferri, J. Artigas, Non-linear Invertible Representation for Joint Statistical and Perceptual Feature Decorrelation, Lecture Notes in Computer Science, Vol. 1876, Berlin, Springer, 2000, pp. 658–667.

[21] G. Box, G. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, MA, 1973.

[22] A. Watson, J. Solomon, A model of visual contrast gain control and pattern masking, J. Opt. Soc. Am. A 14 (1997) 2379–2391.

[23] E. Simoncelli, O. Schwartz, Modeling surround suppression in V1 neurons with a statistically derived normalization model, in: M. Kearns (Ed.), Adv. in Neural Inf. Proc. Syst., Vol. 11, MIT Press, Cambridge, MA, 1999.

[24] D. Kersten, Predictability and redundancy of natural images, J. Opt. Soc. Am. A 4 (12) (1987) 2395–2400.

[25] B. Dubrovin, S. Novikov, A. Fomenko, Modern Geometry: Methods and Applications, Springer, New York, 1982.

[26] H. Wilson, Pattern discrimination, visual filters and spatial sampling irregularities, in: M. Landy, J. Movshon (Eds.), Computational Models of Visual Processing, MIT Press, Cambridge, MA, 1991, pp. 153–168.

[27] M. Carandini, D. Heeger, Summation and division by neurons in visual cortex, Science 264 (1994) 1333–1336.

[28] G. Legge, J. Foley, Contrast masking in human vision, J. Opt. Soc. Am. 70 (1980) 1458–1471.

[29] A. Pons, Human visual system contrast response functions, Ph.D. Thesis, Dpt. d'Òptica, Facultat de Física, Universitat de València, July, 1997.

[30] J. Solomon, A. Watson, A. Ahumada, Visibility of DCT basis functions: effects of contrast masking, in: Proceedings of Data Compression Conference, Snowbird, Utah, IEEE Computer Society Press, Silver Spring, MD, 1994, pp. 361–370.

[31] J. Malo, F. Ferri, J. Albert, J. Soret, J. Artigas, The role of perceptual contrast non-linearities in image transform coding, Image Vision Comput. 18 (3) (2000) 233–246.

[32] J. Malo, F. Ferri, J. Albert, J. Soret, Comparison of perceptually uniform quantization with average error minimization in image transform coding, Electron. Lett. 35 (13) (1999) 1067–1068.

[33] J. Malo, J. Gutiérrez, I. Epifanio, F. Ferri, J. Artigas, Exploiting perceptual feed-back in multigrid motion estimation using an improved DCT quantization, IEEE Trans. Image Process. 10 (10) (2001) 1411–1427.

[34] Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design, IEEE Trans. Commun. 28 (1) (1980) 84–95.

[35] G. Golub, C.V. Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, 1996.

About the Author—IRENE EPIFANIO was born in València, Spain, in 1975. She graduated in Mathematics in 1997 and received the Ph.D. degree in Statistics in 2002, both from the Universitat de València, València, Spain. In 1999 she joined the Computer Science Department, Universitat de València. In October 2000, she joined the Department of Mathematics, Universitat Jaume I, Castelló, Spain, where she is an Assistant Professor. Currently her research interests are focused on texture analysis and image compression.

About the Author—JUAN GUTIÉRREZ received the Licenciado degree in Physics (Electricity, Electronics, and Computer Science) in 1995 from the Universtitat de Valencia, Valencia, Spain, where he is currently pursuing the Ph.D. degree in Motion Estimation and Segmentation.

Since 1997, he has been with the Computer Science Department, Universtitat de Valencia, Valencia, Spain. He has performed two research stays, one at the Digital Imaging Research Centre at Kingston University (UK) for 7 months working on multi-object tracking, and other at the Department of Informatics and Mathematical Modeling, Technical University of Denmark for 2 months, working on optical flow regularization. His current research interests include image analysis, motion understanding and regularization theory.

About the Author—JESÚS MALO (1970) received the M.Sc. degree in Physics in 1995 and the Ph.D. degree in Physics in 1999 both from the Universitat de València.

Since 1994 he has been with the Vision Group of the Universitat de València. Dr. Malo was the recipient of the Vistakon European Research Award in 1994. In the fall of 1999 he worked with the Image and Vision Group at the Institute of Optics (CSIC). In 2000 and 2001 he worked as Fulbright Postdoc at the Vision Group of the NASA Ames Research Center, and at the Lab of Computational Vision of the Center for Neural Science (NYU).

He is interested in models of low-level human vision, their relations with information theory, and their applications to computer vision, image processing and vision science experimentation. He also likes Matlab, modern art, European movies, chamber music, comics and more. http://taz.uv.es/∼jmalo