Non-linear Invertible Representation for Joint Statistical and Perceptual Feature Decorrelation

J. Malo[1], R. Navarro[3], I. Epifanio[2], F. Ferri[2], and J.M. Artigas[1]

[1] Dpt. d'Òptica, Universitat de València
[2] Dpt. d'Informàtica, Universitat de València
Av. Vicent Andrés Estellés S/N, 46100 Burjassot, València, Spain
[3] Instituto de Óptica (CSIC)
C/ Serrano 122, 28006 Madrid, Spain

Abstract. The aim of many image mappings is representing the signal in a basis of decorrelated features. Two fundamental aspects must be taken into account in the basis selection problem: data distribution and the qualitative meaning of the underlying space. The classical PCA techniques reduce the *statistical* correlation using the data distribution. However, in applications where human vision has to be taken into account, there are *perceptual* factors that make the feature space uneven, and additional interaction among the dimensions may arise.
In this work a common framework is presented to analyse the perceptual and statistical interactions among the coefficients of any representation. Using a recent non-linear perception model a set of input-dependent features is obtained which simultaneously remove the statistical and perceptual correlations between coefficients. A fast method to invert this representation is also presented, so no input-dependent transform has to be stored. The decorrelating power of the proposed representation suggests that it is a promising alternative to the linear transforms used in image coding, fusion or retrieval applications[1].

1 Introduction

Independence among the features is recognized as an intrinsic advantage of a given signal representation because it allows simple scalar data processing and a better qualitative interpretation of the feature vector [1,2]. This is why the aim of most feature extraction transforms is to find out a complete set (a basis) of independent features. Two main factors should determine the basis selection problem: the data distribution and the qualitative (geometric) properties of the underlying space. The basis functions should not only reflect the principal axis of the training set but also the eventual anisotropies of the space.

This is particularly important in applications involving natural imagery or texture description, such as indexing and retrieval, fusion, or transform coding.

[1] The authors wish to thank A.B. Watson and E.P. Simoncelli for their fruitful comments. This work has been partially supported by the CICYT-FEDER (TIC) projects 1FD97-0279 and 1FD97-1910 and CICYT (TIC) 98-677-C02

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 658–667, 2000.

In these cases, in addition to the data distribution, it is usually necessary to take into account the properties of Human Visual System (HVS): not every scale, texture or colour component has the same relevance for the HVS, and undesired perceptual interactions among the coefficients may arise if they are scalarly processed. Therefore, in many applications the concept of independence of image coefficients has not only a statistical meaning, but it may also be related to the intrinsic (perceptual) geometry of the space. On the other hand, the HVS has developed efficient representations to deal with natural imagery [3,4,5,6], so the knowledge of the geometry of the low-level representation of a general-purpose biological vision system is of theoretical interest for image processing.

Wavelet and local DCT transforms, are widely used in many applications due to both statistical and perceptual reasons. On one hand, they are used as an approximate fixed-basis *Principal Component Analysis* (PCA). On the other hand, these transforms are similar to the first linear stage in HVS processing. However, the statistical and perceptual decorrelation obtained with these transforms is not complete.

Recently developed perception models with non-linear interactions between the coefficients of wavelet-like representations [7,8,9], can show interesting statistical decorrelation properties [6], but they cannot be used in image processing applications because they are not analytically invertible.

In this work the basis selection problem is analysed from both the statistical and the perceptual points of view. Here the covariance and the perceptual metric matrices are used together to evaluate the statistical and perceptual interactions between the features under a common framework. Also, a fast method to invert the most recent perceptual representation [7,8,9,6] is developed and tested. It is shown that excellent decorrelation results are obtained from both statistical and perceptual points of view just taking into account the perceptual geometry of the wavelet-like feature space. In this context, the decorrelating power of this representation is compared with fixed linear transforms and (unpractical) PCA-like methods that require the storage of ad-hoc basis functions.

2 Aim of the Feature Extraction Transform

Matrices of Second Order Relations. The *statistical deviations* from an image a_0 in a certain feature space are described by the covariance matrix, $\mathit{\Gamma}$:

$$\mathit{\Gamma}(a_0) = \mathcal{E}\left[(a-a_0)\cdot(a-a_0)^T\right] = \mathcal{E}\left[\mathit{\Delta} a \cdot \mathit{\Delta} a^T\right] \tag{1}$$

Assuming a L^2 norm [8], the *perceptual deviation* from a_0 due to a distortion $\mathit{\Delta} a$ is determined by the perceptual metric, W, of the domain at that point,

$$d(a_0, a_0+\mathit{\Delta} a)^2 = \mathit{\Delta} a^T \cdot W(a_0) \cdot \mathit{\Delta} a = \sum_i W_{ii} \mathit{\Delta} a_i^2 + 2\sum_{i\neq j} W_{ij} \mathit{\Delta} a_i \mathit{\Delta} a_j \tag{2}$$

Associated non-aligned ellipsoids. The covariance and the perceptual metric matrices are quadratic forms that describe two different interesting ellipsoids.

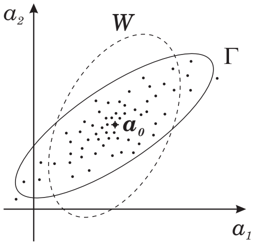

Fig. 1. Ellipsoids describing the data distribution and the space geometry around a_0.

On one hand, Γ describes the shape of the distribution of image samples around a_0. Non-zero off-diagonal elements in Γ indicate a deviation between the data ellipsoid and the axis of the space. This deviation implies a *statistical correlation* between features in the training set. On the other hand, W describes the shape of the (ellipsoidal) locus of perceptually equidistant patterns from a_0. The diagonal elements of W represent the contribution of each coefficient to the perceived distortion (eq. 2). Non-zero off-diagonal elements induce additional contributions to the distortion due to deviations in different dimensions, i.e. they represent *perceptual interactions* between features that modify the perceived distortion. This is a convenient way to represent what is commonly referred to as *masking*: a distortion in a_i may mask the subjective distortion in a_j.

In the most general case these two ellipsoids are not aligned, so their eigenaxis, and the corresponding PCA-like basis functions, are not the same.

Measuring the Statistical and Perceptual Relations Among Features. The decorrelating efficiency of a feature extraction transform has been traditionally referred to the diagonal nature of the resulting covariance matrix. As the non-diagonal elements in W represent the 2nd-order perceptual interactions between the dimensions of the feature space, here we propose to evaluate the transforms from the perceptual point of view applying to W the same measures that have been used for Γ in the context of transform coding [2].

In this way, given a matrix, M, that describe the (statistical or perceptual) relations between the features, a scalar measure (the statistical interaction, η_s, or the perceptual interaction, η_p), can be defined comparing the magnitude of the off-diagonal coefficients with the magnitude of the diagonal coefficients,

$$\eta = \frac{\sum_{i \neq j} |M_{ij}|}{\sum_i |M_{ii}|} \tag{3}$$

Aim of the Feature Extraction Transform. In order to minimise the final correlations from both statistical and perceptual points of view, the transform should find out the eigenaxis of both ellipsoids, i.e., it should simultaneously diagonalise, Γ and W, or simultaneously minimise η_s, and η_p.

Given a perceptual matrix, $W(a_0)$, and using simple linear algebra it can be obtained the linear transform that simultaneously removes both correlations [10]. However, due to the highly point-dependent nature of W, a different linear transform would be necessary for each possible input, which is not a practical solution.

In this work we take a different approach: we use the current non-linear perceptual model [7,8,9,6] to map a local DCT into a perceptually Euclidean space. Beyond the obvious perceptual decorrelation, we show that this non-linear transform has also statistical interest: due to its structure, it also removes the residual statistical correlations that remain in the DCT, strongly reducing μ_s. In this way, both measures, μ_s and μ_p, are simultaneously minimised by a single adaptive transform which can be inverted without the storage of ad-hoc basis functions.

3 Visual Models and Associated Perceptual Geometry

Metric and Visual Response. The standard model of human low-level image analysis has two basic stages. First the image, A, (in the *spatial domain*) is transformed into a vector, a, in a local frequency domain (the *transform domain*) using a linear filter bank, T. Then a set of mechanisms respond to each coefficient of the transformed signal giving an array, r (the *response representation*):

$$A \xrightarrow{T} a \xrightarrow{R} r \tag{4}$$

It is well established that the first linear perceptual transform, T, is similar to the class of wavelet-like transforms employed in many image analysis applications. This is not a casual result, because the low-level algorithms used by the HVS should be mainly determined by the statistics of natural images [3,4,5,6], and, as a result, linear PCA-like solutions have been developed in the low-level HVS.

Not all the basis functions of the transform T are equally perceived so additional processing, R, is included to explain these non-homogeneities. The HVS models assume that all the components of the r vector are equally important and there is no perceptual interaction between them [8,9,6] (i.e. the response domain is Euclidean), so the (perceptual) geometry of the transform domain (and also of the spatial domain) must depend on the nature of the response.

Given a response model, R, an explicit expression for the perceptual metric in any representation space can be obtained. The change of the elements of a tensor under a coordinate mapping depend on the jacobian of the transform [11]. Applying the expressions for tensor transformation to our case, eq. 4, we have,

$$W(a) = \nabla R(a)^T \cdot W'(r) \cdot \nabla R(a) \tag{5}$$

where ∇R is the gradient (or jacobian matrix) of the non-linear response and, $W' = I$, is the metric in the response domain.

Given a particular perception model, i.e. a (T, R) pair, eq. 5 can also be used to compute the metric, and μ_p, in any other representation domain.

Non-Linear Energy Normalisation Model. The current models for R assume that after the application of the linear filter bank, T, the energy of each transform coefficient is normalised by a weighted sum of the energy of its neighbours. The dependence with the neighbour coefficients is given by the convolution with an interaction kernel h [8,6,9],

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

Figure 2 shows the parameters of this *non-linear energy normalisation model* and an example of the response for some basis functions of different frequencies.

The parameters α_i, fig. 2.a, define a band-pass function that modulates the strength of the response for each coefficient i. The parameters β_i, fig. 2.b, determine the point of maximum slope in each response. The values of α and β have been fitted to reproduce amplitude discrimination thresholds without inter-coefficient masking [12]. A frequency-dependent (one octave width) Gaussian kernel, fig. 2.c, has been heuristically introduced according to the refs. [7,8, 9,6].

For mathematical convenience (see section 5) a small linear term (proportional to $|a_i|$) has been included in the response model. This linear band-pass term (fig. 2.a) dominates for very low amplitudes. It is consistent with the fact that for low amplitude patterns the HVS response is roughly linear and it is well described by a band-pass function, the *Contrast Sensitivity Function* (CSF) [13].

In our implementation, the linear transform T is a block DCT, and h includes no spatial interactions between neighbour blocks, but the analytical results can also be applied to any wavelet-like transform with spatial interactions.

Perceptual Metric using the Non-Linear Normalisation. Taking partial derivatives in eq. 6 we have the following gradient matrix:

$$\nabla R(a)_{ij} = \frac{\alpha_i}{100}\delta_{ij} + 2\alpha_i \left(\frac{|a_i|}{\beta_i + (h * |a|^2)_i}\delta_{ij} - \frac{|a_i^2 \cdot a_j|}{(\beta_i + (h * |a|^2)_i)^2} h_{ij} \right) \tag{7}$$

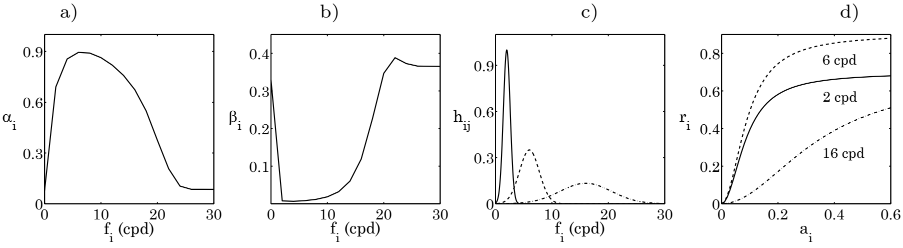

Fig. 2. Parameters of the vision model and non-linear response functions. Here, the amplitude of the coefficients is expressed in contrast (amplitude over mean luminance) which ranges between 0 and 1. The response examples of fig. 2.d show the basic (sigmoid) behaviour of eq. 6, but they are not general because the response to one coefficient depends on the value of the neighbour coefficients. These particular curves were computed for the particular case of no additional masking pattern (zero background).

The slope of the response has three contributions: two diagonal contributions and one off-diagonal contribution given by the interaction kernel. Note that from medium to high amplitudes the slope decreases with amplitude, i.e. the increase in the response is inhibited for high energy coefficients. Also note that the off-diagonal contribution is always negative, i.e. the increase in the response to one coefficient is also inhibited by high energy neighbour coefficients.

The non-diagonal contributions in ∇R give non-diagonal elements in W (see fig. 6). It is clear that the relative perceptual relevance of the DCT features highly depend on frequency (the diagonal of W has a low-pass shape), i.e. the DCT feature space is perceptually anisotropic. It is also clear that DCT features are not perceptually independent because W is not diagonal, i.e. the perceptually privileged directions of the DCT feature space are not aligned with the axis of the space. This implies that an additional transform is needed to remove the perceptual correlation between the DCT features.

As the metric is input-dependent there are no global privileged directions in the space. This implies that the decorrelation transform must be local.

4 Joint Statistical and Perceptual Decorrelation through the Non-linear Normalisation Model

In this work the *non-linear normalisation model* is proposed as a feature decorrelation mapping from both statistical and perceptual points of view. First because it transforms the DCT domain in a perceptually Euclidean space, and second, because, its structure makes it a special form of predictive coder, therefore the output, r, should show less statistical correlation than the input DCT.

The basic idea of predictive coding (or DPCM) [14] is to remove from each coefficient the part that can be predicted from its neighbours. If a prediction of each coefficient is discounted from the original signal in some way, the cross-correlation between neighbour coefficients of the result will be highly reduced. In the commonly used DPCM the discount is linear: the prediction is substracted from the input giving a decorrelated error signal [14].

The normalisation by a weighted sum of the neighbour coefficients can be interpreted as a (non-linear) divisive DPCM (see fig. 3): if the central point of the kernel is set to zero (i.e. if the coefficient a_i is not taken into account in $(h * |a|^2)_i$ as is done in [6]), the convolution in the denominator can be seen as a prediction of the energy of each coefficient from the energies of its neighbours. The division will be a different way of discounting this prediction from the input.

In fact, the prediction stage in the *non-linear normalisation model* is similar to the prediction scheme that has been successfully used in [15] to exploit the conditional probabilities of the transform coefficients to encode them in a more efficient way. This suggest that the normalisation could certainly remove the statistical correlation in r. It has been shown that the parameters in eq. 6 can be optimised to maximise the decorrelation of the output [6]. However, it is important to remark that the parameters used in this work are empirical, i.e. not optimised to improve the decorrelation in a given training set.

5 Quasi-Analytical Inversion of the Normalisation

Problem Statement. The prediction kernel which makes the model useful for statistical decorrelation also makes it non-invertible. As h is not diagonal, each response r_i is coupled with every transform coefficient a_j. Therefore, the inversion of R gives rise to a set of non-linear equations which have no analytical solution. There are, of course, a number of numerical methods based on the iterative search of a solution, $\widehat{a}$, which minimises some distance, $|r - R(\widehat{a})|$, but their convergence is not guaranteed and may be very sensitive to the initialisation.

Quasi-analytical Inversion. In spite of the non-invertible nature of R, around a point a_a, the inverse function can be locally written as,

$$a = R^{-1}(r_a + dr) = a_a + \nabla R^{-1}(r_a) \cdot dr \tag{8}$$

where the unknown gradient of the inverse function can be be related to the (known) gradient of the response (see fig. 4). This differential equation represents the local evolution of the inverse response. If it is integrable, it can be used to propagate the solution from any *initial conditions*, (r_a, a_a), up to the desired point r_b. The computation of the inverse can be analytically formulated as a definite integral. As this integral must be numerically solved we have called this method *quasi-analytical* in contrast to the numerical search-based methods.

Convergence of the Solution. The existence and uniqueness of the solution of an *initial value problem* is guaranteed if the gradient to be integrated is bounded [16]. In our case, $\nabla R(a)$ should not vanish anywhere. The small linear term

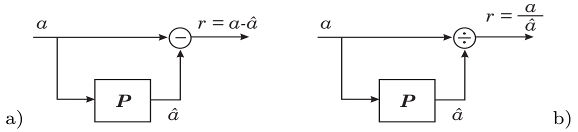

Fig. 3. Alternative DPCM schemes. Fig. 3.a shows the classical *substractive* DPCM. Fig. 3.b shows the *divisive* DPCM interpretation of the non-linear normalisation.

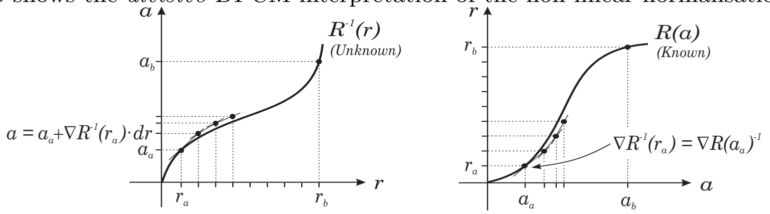

Fig. 4. Inverse computation integrating the increments of the inverse function. In each iteration, the unknown gradient is computed from the known response at that point.

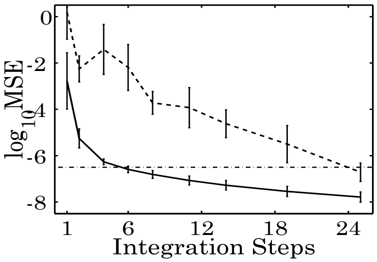

Fig. 5. Reconstruction errors of a *tipical* block (solid line) and a *difficult* block (dashed line). The curves are the average over several initial conditions: the mean DCT, a $1/f$ DCT and a flat DCT. The bars show the dispersion in the distortion due to the different initial conditions. The differences below the dashdot line are visually negligible.

in the response avoids ensures a non-zero slope for every a. This guarantees that the integration of eq. 8 is possible and always gives the appropriate solution.

To test the speed and robustness of the inverse computation, the 16×16 DCT blocks of a set of 256×256 natural images were transformed according to the *non-linear normalisation* and then inverted back integrating the eq. 8 with a 4^{th} order Runge-Kutta algorithm. The effect of the initialisation and the number of integration steps was explored. Figure 5 shows the DCT reconstruction error as a function of the number of integration steps for two different blocks and different initial conditions. The inversion experiments show the following trends:

- **The solution is always found**. The experiments confirm the theoretical existence and uniqueness result: the proposed method achieves the appropriate inverse (with negligible distortion), for every response block, no matter the initial conditions with a reasonably small number of integration steps.
- **Speed**. Most of the responses (~90% in the explored images) appropriately converge to its corresponding DCT in 3-6 integration steps from very different initial conditions. The solid line in fig. 5 is an example of this kind of blocks. However, we found that ~10% of the DCT blocks, usually corresponding to sharp spectrum regions, require a more accurate integration. The dashed line represents the worst-behaved block of the training set.
- **Robustness**. The inverse does not substantially depend on the initial conditions, but on the nature of the block (see fig. 5), so the algorithm is insensitive to the initialisation. Generic $1/f$ or flat spectra give quite good results.

6 Decorrelation Experiments

The decorrelation properties of the proposed representation were compared with the standard PCA representation (i.e. the domain of eigenfunctions of Γ) and with the domain of eigenfunctions of W, which will be referred to as *Perceptual Principal Component Analysis* (PPCA). The local DCT which is the best fixed-basis approximation to PCA analysis for natural images [2] was also explored. The local DCT is also interesting because it is the first linear stage, T, in the

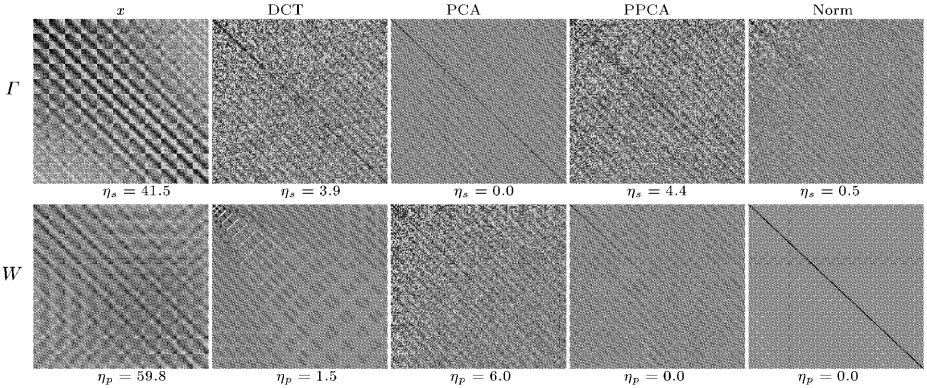

Fig. 6. Covariance (upper row) and perceptual metric (lower row) in different domains. The qualitative meaning of the elements of these matrices depends on how the 2D domains are scanned to construct the 1D feature vectors. The matrices in the spatial domain are the result of a raster scanning. A JPEG-like zigzag scanning has been used in the DCT and the other transform domains because the coefficients of similar frequency are grouped together. According to this, the frequency meaning of the diagonal elements of W, and Γ in these domains progressively increases from zero to the Nyquist frequency. For the sake of clarity only the upper-left 176×176 submatrix is shown. The frequency values of the displayed elements in the DCT domain range from 0 to 26 *cpd.*

proposed representation, (T, R), so it is useful to assess the benefits of the non-linear normalisation R. The spatial representation has been included as a useful example of highly correlated domain.

The PCA representation was computed from the covariance around the average of a set of natural images. The PPCA was computed from the average perceptual metric, originally defined over the DCT blocks of the training set. The values of Γ, W, η_s and η_p, in the different domains are shown in figure 6.

The highly non-diagonal nature of W in the spatial domain is an additional argument against the spatial domain representation that complements the classical reasonings exclusively based on the non-diagonal nature of the covariance matrix [1,2]. The DCT domain certainly reduces the statistical and perceptual interactions by an order of magnitude with regard to the spatial domain but it still doesn't completely remove none of them. The linear approaches that only take into account one of the relations, PCA or PPCA, are not acceptable because, in these cases, the other relation is increased in the resulting representation.

The proposed representation, DCT plus non-linear normalisation transform, gives the best results. On one hand, it achieves a complete perceptual decorrelation for every input because it works with local (not average) metrics. In this sense the perceptual decorrelation is better than in the PPCA or any other PCA-like approach such as [10]. On the other hand, the statistical interaction is also highly reduced, almost an order of magnitude with regard to the DCT.

7 Concluding Remarks

In this paper the perceptual correlation between the features of an image representation has been formalised through the perceptual metric matrix in the same way as the statistical correlation is described by the covariance matrix.

We have presented a perceptually inspired image representation that simultaneously reduces the statistical and perceptual correlation between the features. It first uses a linear local frequency transform and after a non-linear energy normalisation is applied to the coefficients. The good statistical behaviour of this perceptual model relies on its divisive-DPCM structure. The proposed representation improves the decorrelation properties of a fixed basis representation such as the DCT without the basis storage problem of linear input-dependent PCA-like transforms because an efficient method to invert it has been presented.

According to the results presented here, the non-linear mapping R may be a very interesting second stage after the linear DCT or wavelet-like transforms used in many image analysis applications [17].

References

1. K. Fukunaga. *Introd. to Statistical Pattern Recognition.* Acad. Press, MA, 1990.
2. R.J. Clarke. *Transform Coding of Images.* Academic Press, NY, 1985.
3. J.G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Tran. Biomed. Eng.*, 36:107–114, 1989.
4. B.A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by a learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
5. A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338, 1997.
6. E.P. Simoncelli and O. Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In M.S. Kearns, editor, *Adv. in Neural Inf. Proc. Syst.*, volume 11. MIT Press, 1999.
7. J.A. Solomon, A.B. Watson, and A. Ahumada. Visibility of DCT basis functions: Effects of contrast masking. *IEEE Proc. Data Compress. Conf.*, (1):361–370, 1994.
8. P.C. Teo and D.J. Heeger. Perceptual image distortion. *Proc. of the First IEEE Intl. Conf. Im. Proc.*, 2:982–986, 1994.
9. A.B. Watson and J.A. Solomon. A model of visual contrast gain control and pattern masking. *JOSA A*, 14:2379–2391, 1997.
10. I. Epifanio and J. Malo. Linear transform for simultaneous diaginalization of covariance and perceptual metric in image coding. Technical Report 3, Grup de Visió. Dpt. d'Informàtica, Universitat de València, 2000.
11. B. Dubrovin and S. Novikov. *Modern Geometry.* Springer Verlag, NY, 1982.
12. G.Legge and J.Foley. Contrast masking in human vision. *JOSA*, 70:1458–71, 1980.
13. B.A. Wandell. *Foundations of Vision.* Sinauer Assoc. Publish., MA, 1995.
14. A.M. Tekalp. *Digital Video Processing.* Prentice Hall, NJ, 1995.
15. R. Buccigrossi and E. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Im. Proc.*, 8(12):1688–1701, 1999.
16. D.J. Logan. *Non-linear partial differential equations.* Wiley&Sons, NY, 1994.
17. J. Malo, F. Ferri, R. Navarro, and R. Valerio. Perceptually and statistically decorrelated features for image representation: Application to transform coding. *Proc. XV Intl. Conf. Patt. Recog.*, 2000.