Chapter XIII

Perceptual Image Representations for Support Vector Machine Image Coding

Juan Gutiérrez, Universitat de València, Spain
Gabriel Gómez-Perez, Universitat de València, Spain
Jesús Malo, Universitat de València, Spain
Gustavo Camps-Valls, Universitat de València, Spain

Abstract

Support-vector-machine image coding relies on the ability of SVMs for function approximation. The size and the profile of the ε-insensitivity zone of the support vector regressor (SVR) at some specific image representation determines (a) the amount of selected support vectors (the compression ratio), and (b) the nature of the introduced error (the compression distortion). However, the selection of an appropriate image representation is a key issue for a meaningful design of the ε-insensitivity profile. For example, in image-coding applications, taking human perception into account is of paramount relevance to obtain a good rate-distortion performance. However, depending on the accuracy of the considered perception model, certain image representations are not suitable for SVR training. In this

chapter, we analyze the general procedure to take human vision models into account in SVR-based image coding. Specifically, we derive the condition for image representation selection and the associated ε-insensitivity profiles.

Introduction

Nowadays, the volume of imaging data increases exponentially in a very wide variety of applications, such as remote sensing, digital and video camera design, medical imaging, digital libraries and documents, movies, and videoconferences. This poses several problems and needs for transmitting, storing, and retrieving images. As a consequence, digital image compression is becoming a crucial technology. However, compressing an image is significantly different than compressing raw binary data given their particular statistical properties, and thus the application of general-purpose compression methods would be far from optimal. Therefore, statistical knowledge about the problem becomes extremely important to develop efficient coding schemes. Another critical issue in visual communications to be judged by human observers is introducing perception models in the algorithm design procedure.

The efficient encoding of images relies on understanding two fundamental quantities, commonly known as *rate* and *distortion*. The rate expresses the cost of the encoding (typically in bits) and the distortion expresses how closely the decoded signal approximates the original image. Extensive literature has shown that the problem can be made much more tractable by transforming the image from an array of pixels into a new representation in which rate or distortion are more easily quantified and controlled. In this framework, the goal of the transform is removing the statistical (Gersho & Gray, 1992) and perceptual (Epifanio, Gutiérrez, & Malo, 2003; Malo, Epifanio, Navarro, & Simoncelli, 2006) dependence between the coefficients of the new representation in order to allow an efficient scalar quantization and zero-order entropy coding of the samples. To this end, the current transform coding standards, JPEG and JPEG2000 (Taubman & Marcellin, 2001; Wallace, 1991), use fixed-basis linear transforms (2-D block discrete cosine transform [DCT] or wavelets), which are similar to adaptive linear transforms that remove second-order or higher order statistical relations of natural image samples (principal components analysis, PCA, and ICA; Hyvarinen, Karhunen, & Oja, 2001), and resemble the first linear stage in human perception models (A. Watson & Solomon, 1997). However, natural images are not that simple as a Gaussian process (fully described by its PCA components) or a linear superposition of independent patterns (the basic assumption in ICA). In fact, significant relations between the energy of the transform coefficients remain in the linear domains that are widely used for transform coding or denoising (Buccigrossi & Simoncelli, 1999; Gutiérrez, Ferri, & Malo, 2006; Malo et al.). Besides this, masking experiments reveal that linear local frequency basis functions are not perceptually independent either (A. Watson & Solomon). Recent results confirm the link between human perception and statistics of natural images in this context: The statistical effect of the current cortical vision mechanisms suggests that the use of these biological image representations is highly convenient to reduce the statistical and the perceptual dependence of the image samples at the same time (Epifanio et al.; Malo et al.). These results are consistent with the literature that seeks for statistical explanations of the cortical sensors' organization (Barlow, 2001; Malo & Gutiérrez, in press; Simoncelli, 2003).

These statistical and perceptual results suggest that achieving the desired independence necessarily requires the introduction of nonlinearities after the commonly used linear image representations *prior* to their scalar quantization. Recently, two different nonlinear approaches have been used to improve the results of image-coding schemes based on linear transforms and linear perception models. On one hand, more accurate nonlinear perception models have been applied after the linear transform for image representation (Malo et al., 2006). On the other hand, support-vector-machine (SVM) learning has been used for nonlinear feature selection in the linear local DCT representation domain (Gomez, Camps-Valls, Gutiérrez, & Malo, 2005; Robinson & Kecman, 2003). Both methodologies will be jointly exploited and analyzed in this chapter.

The rationale to apply the support vector regressor (SVR) in image-coding applications is taking advantage of the sparsity property of this function approximation tools. This is carried out by using tunable ε-insensitivities to select relevant training samples, thus representing the signal with a small number of support vectors while restricting the error below the ε bounds.

An appropriate ε-insensitivity profile is useful to (a) discard statistically redundant samples, and (b) restrict the perceptual error introduced in the approximated signal. Therefore, the choice of the domain for ε-insensitivity design is a key issue in this application.

The use of SVMs for image compression was originally presented in Robinson and Kecman (2000), where the authors used the standard ε-insensitive SVR (Smola & Schölkopf, 2004) to learn the image gray levels in the spatial domain. A constant insensitivity zone per sample is reasonable in the spatial domain because of the approximate stationary behavior of the luminance samples of natural images. However, these samples are strongly coupled both from the statistical and the perceptual points of view. First, there is a strong *statistical correlation* between neighboring luminance values in the spatial domain, and second, coding errors independently introduced in this domain are *quite visible* on top of a highly correlated background. These are the basic reasons to make the promising SVR approach inefficient in this domain.

The formulation of SVRs in the local DCT domain was fundamental to achieve the first competitive results (Robinson & Kecman, 2003). In this case, Robinson and Kecman also used a constant ε-insensitivity, but according to a qualitative human-vision-based reasoning, they a priori discarded the high-frequency coefficients in the SVR training. This is equivalent to using a variable ε-insensitivity profile: a finite value for the low-frequency samples and an infinite value for the high-frequency samples. This heuristic makes statistical and perceptual sense since the variance of the local frequency samples is concentrated in the low-frequency coefficients and the visibility of these patterns is larger. Therefore, it is appropriate to ensure a limited error in the low-frequency region while allowing more distortion in the high-frequency region.

However, the qualitative ideal low-pass ε-insensitivity profile in Robinson and Kecman (2003) can be improved by using a rigorous formulation for the ε-insensitivity design. Specifically, the ε-insensitivity in a given image-representation domain has to be constructed to restrict the maximum perceptual error (MPE) in a perceptually Euclidean domain. This MPE restriction idea is the key issue for the good subjective performance of the quantizers used in the JPEG and JPEG2000 standards (Wallace, 1991; Zeng, Daly, & Lei, 2002), as pointed out in Malo et al. (2006) and Navarro, Gutiérrez, and Malo (2005).

In Camps-Valls et al. (2006) and Gomez et al. (2005), the MPE restriction procedure was applied using different perception models to obtain the appropriate ε-insensitivity, which may be constant or variable (Camps-Valls, Soria-Olivas, Pérez-Ruixo, Artés-Rodriguez, Pérez-Cruz, & Figueiras-Vidal, 2001). However, it is worth noting that depending on the accuracy of the considered perception model, certain image representations are not suitable for SVR training.

In this chapter, we analyze the general procedure to take human vision models into account in SVR-based image-coding schemes. Specifically, we derive the condition for image representation selection and the associated ε-insensitivity profiles.

The structure of the chapter is as follows. The next section motivates the need for considering human perception in dealing with the coding noise, which poses the problem of a proper (perceptually meaningful) transformation. Then we review the computational models that account for the effects shown in the previous section. Next, the chapter formulates the problem of the design and application of suitable insensitivity profiles for SVM training that give rise to perceptually acceptable distortions. Finally, we show some experimental results on benchmark images, and then end this chapter with some conclusions and further work.

Visibility of Noise

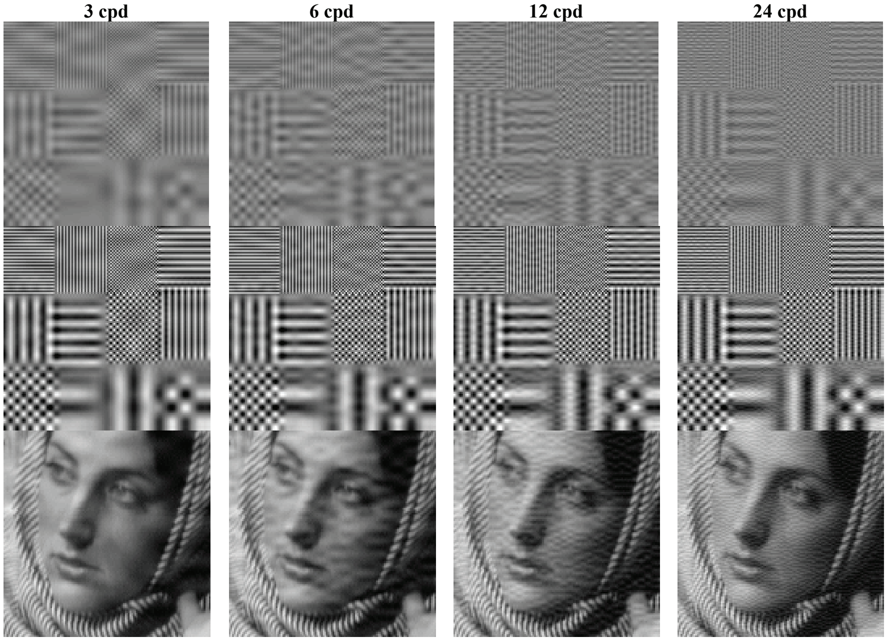
The distortion introduced in SVM-based image coding is due to the approximation error given by the use of a limited number of support vectors. However, the qualitative nature of this distortion strongly depends on the image-representation domain and on the insensitivity profile used to select support vectors.

In this section, we show that the visibility of some distortion not only depends on the energy of the noise (mean squared error, MSE, and PSNR), but also on its frequency nature and on the background signal. Figure 1 summarizes the effects reported in the perception literature (Campbell and Robson, 1968; Heeger, 1992; A. Watson & Solomon, 1997), namely, *frequency sensitivity and masking*, that will determine the appropriate image-representation domain and insensitivity profiles for SVM training. The issue of selecting an appropriate domain of image representation is an important concern in other domains of vision computing and image processing (see Chapters XII and XIV in this book).

In this example, equal-energy random noise of different frequency bands and horizontal orientation has been added on top of three different background images: two synthetic images and one natural image. The synthetic images consist of patches of periodic functions of different frequency bands (3 cpd, 6 cpd, 12 cpd, and 24 cpd) and different orientations (horizontal, vertical, and diagonal). The two synthetic background images (first and second rows) differ in their energy (contrast or amplitude of the periodic functions). Several conclusions can be extracted from Figure 1.

1. **Mean squared error is not perceptually meaningful.** A remarkable fact is that the energy of the noise (or the MSE) does not correlate with the visibility of the noise at

Figure 1. Equal-energy noise of different-frequency content—3 cpd, 6 cpd, 12 cpd, and 24 cpd—shown on top of different backgrounds. All images have the same MSE distance with regard to the corresponding original image, but the visibility of the noise is quite different.



all (Girod, 1993; Malo, Pons, & Artigas, 1997; Pons, Malo, Artigas, & Capilla, 1999; Teo & Heeger, 1994).

2. **Frequency selectivity.** Noise visibility strongly depends on its frequency nature: low-frequency and high-frequency noise are less visible in all cases. This is because human perception is mediated by frequency analyzers that have different sensitivity. Therefore, more noise (larger SVM insensitivity) will be perceptually acceptable in different frequency bands.
3. **Automasking.** For a given noise frequency (for a given column in Figure 1), noise visibility decreases with the energy of the background signal: The same distortion is less visible in high-contrast backgrounds. This is because the perceptual frequency analyzers are nonlinear; their slope (sensitivity) is bigger for low-contrast signals while it is smaller for high-contrast signals. This phenomenon is usually referred to as masking since a high-contrast signal masks the distortion because it saturates the response of the frequency analyzers. Therefore, more noise (larger SVM insensitivity) will be perceptually acceptable in high-contrast regions.
4. **Cross-masking.** Note however, that for the same noise frequency and background signal contrast (within every specific image in Figure 1), noise visibility depends on the similarity between signal and distortion. Low-frequency noise is more visible in

high-frequency backgrounds than in low-frequency backgrounds (e.g., left figure of second row). In the same way, high-frequency noise is more visible in low-frequency backgrounds than in high-frequency backgrounds (e.g., right image of the second row). That is, some signal of a specific frequency strongly masks the corresponding frequency analyzer, but it induces a smaller sensitivity reduction in the analyzers tuned to different frequencies. Besides that, the reduction in sensitivity of a specific analyzer is larger as the distance between the background frequency and the frequency of the analyzer is smaller. For instance, in the left image of the second row, the visibility of the low-frequency noise (sensitivity of the perceptual low-frequency analyzer) is small in the low-frequency regions (that mask this sensor) but progressively increases as the frequency of the background increases. This is because the response of each frequency analyzer not only depends on the energy of the signal for that frequency band, but also on the energy of the signal in other frequency bands (cross-masking). This implies that a different amount of noise (different SVM insensitivity) in each frequency band may be acceptable depending on the energy of that frequency band and on the energy of neighboring bands.

According to these perception properties, an input dependent ε-insensitivity in a local frequency domain is required.

Linear and Nonlinear Perception Models

The general model of early visual (cortical) processing that accounts for the previously described effects includes a linear local frequency transform, T, followed by a response transform, R (A. Watson & Solomon, 1997):

$$A \xrightarrow{T} y \xrightarrow{R} r, \tag{1}$$

where in the first stage, each spatial region A of size $N \times N$ is analyzed by a filter bank, T, for example, a set of local (block) DCT basis functions. This filter bank gives rise to a vector, $y \in R^{N^2}$, whose elements y_f represent the local frequency content of the signal in that spatial region (or block). The second stage, R, accounts for the different linear (or eventually nonlinear) responses of the local frequency analyzers of the first stage.

The last image representation domain, $y \in R^{N^2}$, is assumed to be perceptually Euclidean (Legge, 1981; Pons et al., 1999; Teo & Heeger, 1994; A. Watson & Solomon, 1997); that is, the distortion in any component, Δr_f, is equally relevant from the perceptual point of view. This implies that the perceptual geometry of other image representations is not Euclidean, but depends on the Jacobian of the response model ∇R (Epifanio et al., 2003; Malo et al., 2006). As it will be shown in the next section, this geometric fact is the key issue to select the representation domain and the ε-insensitivity profile for perceptually efficient SVM training.

Given the first linear filter-bank stage T, different response models have been proposed for the second stage R to account for the perception effects illustrated in Figure 2, either linear or nonlinear. In the following sections, we will first review the functional form of the model and its Jacobian. Afterward, we will show that given the shape of the responses, different amounts of noise, Δy_f, are needed to obtain the same subjective distortion, thus explaining the effects described in the previous section.

Linear, Frequency-Dependent Response

In the linear model approach, each mechanism of the filter bank has a different (but constant) gain depending on its frequency:

$$r_f = \alpha_f \cdot y_f \qquad (2)$$

This linear, frequency-dependent gain is given by the contrast sensitivity function (CSF; see Figure 2; Campbell & Robson, 1968).

In this case, the Jacobian of the response is a constant diagonal matrix with the CSF values on the diagonal:

$$\nabla R_{ff'} = \alpha_f \; \delta_{ff'}. \qquad (3)$$

Figure 2. Frequency-dependent linear gain, α_f, of the CSF model

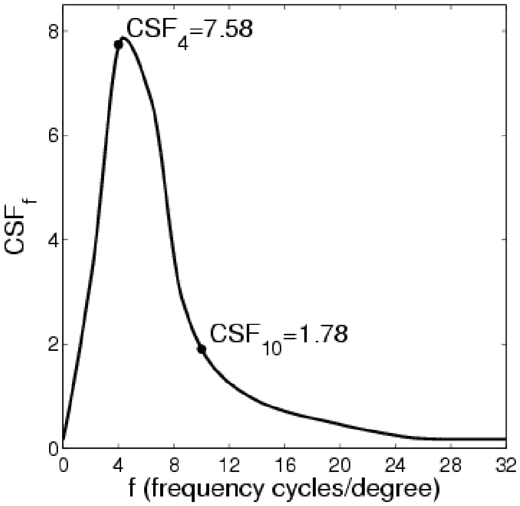


Figure 3 shows the response of two linear mechanisms tuned to different frequencies as a function of the energy (or contrast) of the stimuli of these frequencies (Peli, 1990). This linear response model accounts for the general frequency-dependent visibility of the noise shown in Figure 1: The larger the slope of the response of a mechanism tuned to a specific frequency, *f*, the smaller the distortion Δy_f needed to give rise to the same perceptual distortion $\Delta r_f = \tau$. According to the CSF, the visibility of medium frequencies is larger than the visibility of very low and high frequencies.

However, this linear model is too simple to account for masking: Note that the slope of the responses (the sensitivity of the mechanisms) is constant, so the same signal distortion on top of a signal of larger energy (or contrast) generates the same perceptual distortion. This problem can be alleviated by introducing more sophisticated (nonlinear) response models.

Nonlinear Response: Adaptive Gain Control or Divisive Normalization

The current response model for the cortical frequency analyzers is nonlinear (Heeger, 1992; A. Watson & Solomon, 1997). The outputs of the filters of the first linear stage undergo a nonlinear transform in which the energy of each linear coefficient (already weighted by a CSF-like function) is normalized by a combination of the energies of the neighboring coefficients in frequency:

$$r_f = \frac{\mathrm{sgn}(y_f)\cdot |\alpha_f \cdot y_f|^\gamma}{\beta_f + \sum_{f'=1}^{N^2} h_{ff'} |\alpha_{f'} \cdot y_{f'}|^\gamma}, \tag{4}$$

Figure 3. Responses and associated visibility thresholds of the two sensors whose slopes have been highlighted in Figure 2. The Euclidean nature of the response domain implies that two distortions, Δy_f and $\Delta y_{f'}$, induce perceptually equivalent effects if the corresponding variations in the response are the same: $\Delta r_f = \Delta r_{f'} = \tau$.

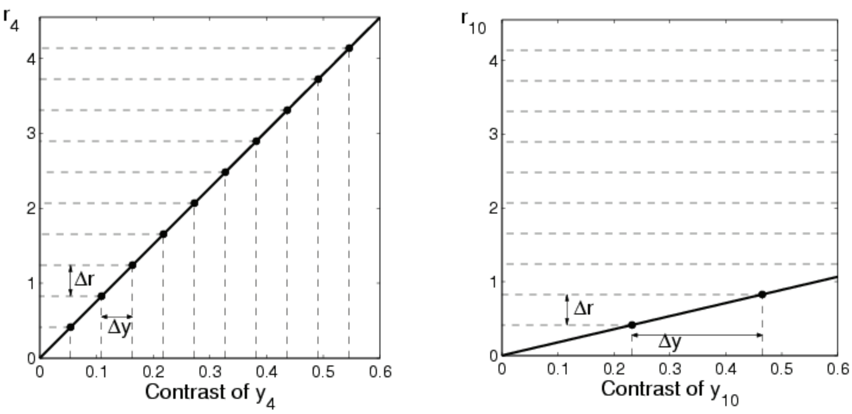
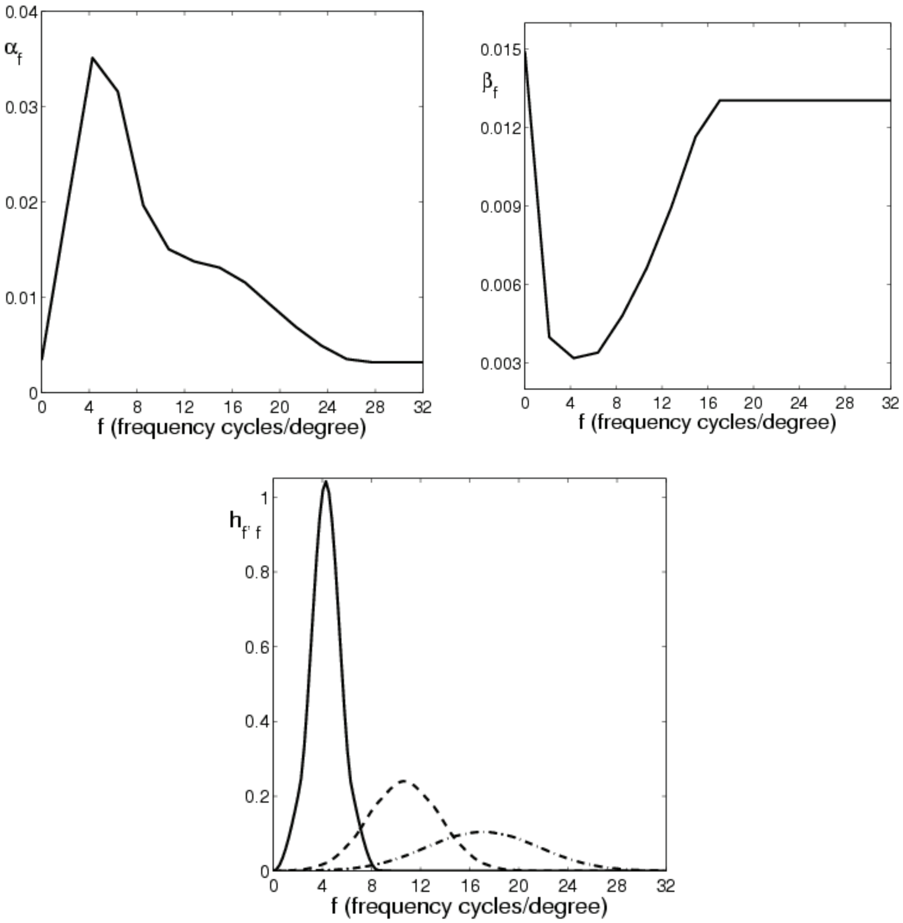


Figure 4. Parameters α *and* β *and three frequency interaction neighborhoods (rows of h) in equation 4. The different line styles represent different frequencies: 4 cpd (solid), 10 cpd (dashed), and 18 cpd (dash-dot).*



where $h_{ff'}$ determines the interaction neighborhood in the nonlinear normalization of the energy, which is assumed to be Gaussian (A. Watson & Solomon, 1997), and

$$h_{ff'} = K_f \cdot \exp(-\|f - f'\|^2 / \sigma_{|f|}^2), \qquad (5)$$

where $\sigma_{|f|} = \frac{1}{6}|f| + 0.05$ and $|f|$ is given in cpd. See Figure 4 for the parameters in equations 4 and 5.

In this case, the Jacobian of the response is input dependent and nondiagonal:

$$\nabla R(y)_{ff'} = \mathrm{sgn}(y_f)\gamma \left(\frac{|\alpha_f \cdot y_f|^{\gamma-1}}{\beta_f + \sum_{f'=1}^{N^2} h_{ff'} |\alpha_{f'} \cdot y_{f'}|^{\gamma}} \cdot \delta_{ff'} - \frac{|\alpha_f \cdot y_f|^{\gamma} |\alpha_{f'} \cdot y_{f'}|^{\gamma-1}}{(\beta_f + \sum_{j=1}^{N^2} h_{ff'} |\alpha_{f'} \cdot y_{f'}|^{\gamma})^2} \cdot h_{ff'} \right). \tag{6}$$

Figures 5 and 6 show examples of the response of two nonlinear mechanisms tuned to different frequencies as a function of the energy (or contrast) of stimuli of these frequencies in different masking conditions. In the first case (Figure 5), an automasking situation is considered; that is, this figure shows the response r_4 (or r_{10}) as a function of y_4(or y_{10}) when all the other mechanisms are not stimulated, that is, $y_{f'} = 0$, $\forall f' \neq 4$ (or $\forall f' \neq 10$). In the second case (Figure 6), this automasking response is compared with the (cross-masking) response obtained when showing the optimal stimulus (y_4 or y_{10}) on top of another pattern that generates $y_6 \neq 0$.

This more elaborated response model also accounts for the frequency-dependent visibility of distortions: Note that the slope is larger for 4 cpd than for 10 cpd, thus a larger amount of distortion is required in 10 cpd to obtain the same perceived distortion. Equivalently, 4 cpd of noise is more visible than 10 cpd of noise of the same energy. This general behavior is given by the band-pass function α_f.

Moreover, it also accounts for automasking since the amount of distortion needed to obtain a constant perceptual distortion increases with the contrast of the input (see Figure 5). This is due to the fact that the response is attenuated when increasing the contrast because of the normalization term in the denominator of equation 4.

It also accounts for cross-masking since this attenuation (and the corresponding response saturation and sensitivity decrease) also occurs when other patterns $y_{f'}$ with $f' \neq f$ are present. Note how in Figure 5 the required amount of distortion increases as the contrast

Figure 5. Responses and associated visibility thresholds of the two sensors tuned to frequencies 4 and 10 cpd in automasking (zero background) conditions. The required amount of distortion Δy_f to obtain some specific distortion in the response domain τ is shown for different contrasts of the input pattern.

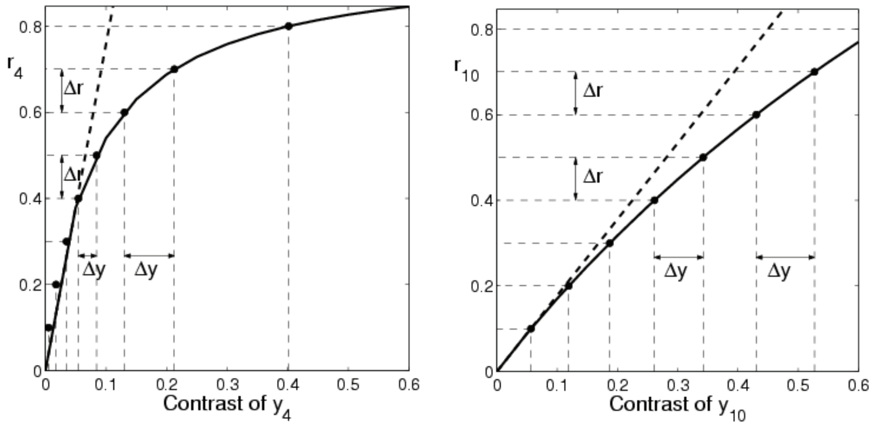
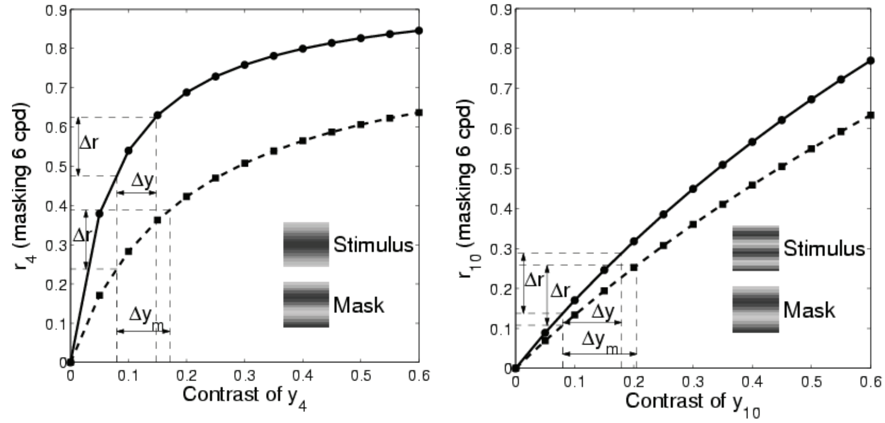


Figure 6. Responses and associated visibility thresholds of the two sensors tuned to frequencies 4 and 10 cpd when masked by a pattern of different frequency (6 cpd) at different contrast: 0 (automasking, solid line) and 0.5 (dashed line). In this case, the required amount of distortion Δy_f to obtain some specific distortion in the response domain τ at a given contrast of the stimulus increases when the contrast of the mask increases.



of the mask of different frequency is increased. Moreover, given the Gaussian shape of the interaction neighborhood, patterns of closer frequencies mask the distortion more effectively than background patterns of very different frequency. This is why the 6 cpd mask induces a larger variation of the acceptable noise in 4 cpd than in 10 cpd.

SVM ε-Insensitivity from Perceptually Acceptable Distortions

In order to obtain image approximations with a small number of support vectors, while keeping the perceptual appearance of the image good, it is necessary to derive the kind of distortions that guarantee that all the distortion coefficients in the response domain will be below a certain threshold. The appropriate domain for SVM training will be the one that makes the ε-insensitivity design feasible.

A small distortion in a spatial region of the signal ΔA induces a distortion in the perceptually meaningful response representation Δr. This distortion can be approximated by using the Jacobian of the response function:

$$\Delta r \cong \nabla R(y) \cdot \Delta y = \nabla R(T \cdot \Delta A) \cdot T \cdot \Delta A. \quad (7)$$

Then, the MPE for that spatial region, *s*, is given by:

$$\mathrm{MPE}_s = \| \Delta r \|_\infty = \max(\nabla R(y) \cdot \Delta y) = \max(\nabla R(T \cdot \Delta A) \cdot T \cdot \Delta A). \quad (8)$$

The global perceived distortion in an image with *n* spatial regions will be a particular spatial pooling (*q*-norm) of these *n* local distortions from each local (block) response representation:

$$\mathrm{MPE} = \| (\mathrm{MPE}_1, \cdots, \mathrm{MPE}_n) \|_q = \left(\sum_s \mathrm{MPE}_s^q \right)^{1/q}, \quad (9)$$

where *q* is the summation exponent in this spatial pooling.

Some distortion in a block ΔA, or in the local frequency domain, Δy, is perceptually acceptable if it generates a distortion in the perceptual response domain in which the distortion in every response coefficient is below a threshold τ. As stated in the introduction, the role of SVM regression in the context of image coding is reducing the size of the signal description (reducing the number of support vectors) while keeping the MPE below a threshold in every spatial region, that is, $\mathrm{MPE}_s < \tau, \forall\, s$. This can be incorporated in the SVM training by using an appropriate ε-insensitivity profile, that is, given a different ε for each training sample.

The MPE restriction determines a geometric condition to derive the ε profile. In the response representation domain, this condition is quite simple, taking:

$$\varepsilon_r(f) = \tau, \quad (10)$$

it is obviously guaranteed that the distortions Δr_f will be bounded by τ, $\forall f$. This set of scalar restrictions $\varepsilon_r(f)$ is equivalent to constrain the vector distortion Δr into an *n*-dimensional cube of side τ.

The corresponding *n*-dimensional boundary region in other representation domains can be obtained operating backward from the cube in the response representation. However, depending on the complexity of the transformation from the response representation to the desired representation, the shape of the *n*-dimensional boundary region may be quite complicated for ε design. Let us analyze the difficulty for ε design in previously reported image-representation domains, namely, the spatial domain (Robinson & Kecman, 2000) and the block DCT domain (Gomez et al., 2005; Robinson & Kecman, 2003).

In the block DCT domain, the profile $\varepsilon_y(f)$, for a given input signal y_o, is determined by the boundaries of the *n*-dimensional region that fulfills the condition:

$$[\nabla R(y_o) \cdot \Delta y]_f \le \tau \quad (11)$$

for all possible distortions Δy with $|\Delta y_f| \le \varepsilon_y(f)$.

Analogously, in the spatial domain, the profile $\varepsilon_A(x)$ for a given input signal A_o is determined by the boundaries of the n-dimensional region that fulfills the condition:

$$[\nabla R(A_o) \cdot T \cdot \Delta A]_f \leq \tau \tag{12}$$

for all possible distortions ΔA with $|\Delta A_x| \leq \varepsilon_A(x)$.

If the matrix,∇R or $\nabla R \cdot T$, that acts on the distortion vectors, Δy or ΔA, is not diagonal, the above conditions determine n-dimensional boxes not aligned with the axes of the representation. This implies that the distortions in different coefficients should be coupled, which is not guaranteed by the SVM regression.

This is an intrinsic problem of the spatial domain representation because T is an orthogonal filter bank that is highly nondiagonal. Therefore, the spatial domain is very unsuitable for perceptually efficient SVM regression.

The situation may be different in the DCT domain when using a simplified (linear) perception model. In this particular case, the Jacobian ∇R is diagonal (equation 3), so the condition in equation 11 reduces to:

$$|\Delta y_f| \leq \frac{\tau}{\alpha_f}, \tag{13}$$

or equivalently,

$$\varepsilon_y(f) = \frac{\tau}{\alpha_f}, \tag{14}$$

which is the frequency-dependent ε-insensitivity proposed in Gomez et al. (2005). This implies a more accurate image reproduction (smaller $\varepsilon_y(f)$) for low and medium frequencies, and it allows the introduction of substantially more distortion (larger $\varepsilon_y(f)$) in the high-frequency region. This reasoning is the justification of the ad hoc coefficient selection made in the original formulation of SVMs in the DCT domain for image coding (Robinson & Kecman, 2003). This ideal low-pass approximation of the CSF may be a too-crude approximation in some situations, leading to blocky images.

Despite the relatively good performance of the SVM coding approaches based on either rigorous (Gomez et al., 2005) or oversimplified (Robinson & Kecman, 2003) linear models, remember that linear models cannot account for automasking and cross-masking effects, and thus the introduction of nonlinear response models is highly convenient.

The current nonlinear response model (equation 4) implies a nondiagonal Jacobian (equation 6). Therefore, when using this perception model, it is not possible to define the ε-insensitivity in the DCT domain because of the perceptual coupling between coefficients in this domain. According to this, when using the more accurate perception models, the appropriate domain for SVM training is the perceptual response domain, hence using a constant insensitivity (Camps-Valls et al., 2006).

Experimental Results

In this section, we show the performance of several SVM-based image-coding schemes. First, some guidelines for model development are given, where explicit examples of the different behavior of the SVM sample selection are given depending on the image-representation domain. Then we analyze the developed coding strategies in terms of (a) the distribution of support vectors, and (b) the effect that these distributions have in the compression performance.

Model Development

In the SVR image-coding framework presented here, the whole image is first divided in blocks, and then a particular SVR is trained to learn some image representation of each domain. Afterward, the signal description (the weights) obtained by the SVR are subsequently quantized, giving rise to the encoded block. In this section, we analyze examples of this procedure using four image representations (or ε-insensitivity design procedures): (a) the spatial representation using constant insensitivity profile (Robinson & Kecman, 2003), (b) the DCT representation using constant insensitivity in the low-frequency coefficients and discarding (infinite insensitivity) the coefficients with frequency bigger than 20 cpd, RK-i as reported in Robinson and Kecman (2005), (c) the DCT representation using CSF-based insensitivity (equation 14), CSF-SVR as reported in Gomez et al. (2005), and (d) the proposed nonlinearly transformed DCT representation using equations 4 and 5 (NL-SVR). In this latter case, and taking into account the perceptual MPE concept, constant ε-insensitivity was used.

After training, the signal is described by the Lagrange multipliers of the support vectors needed to keep the regression error below the thresholds ε_i. Increasing the thresholds reduces the number of required support vectors, thus reducing the entropy of the encoded image and increasing the distortion. In all experiments, we used the RBF kernel, trained the SVR models without the bias term *b*, and modeled the absolute value of the DCT (or response) coefficients. For the sake of a fair comparison, all the free parameters (ε-insensitivity, penalization parameter *C*, Gaussian width of the RBF kernel σ, and weight quantization coarseness) were optimized for all the considered models. In the NL-SVM case, the parameters of the divisive normalization used in the experiments are shown in Figure 4. In every case, rather than adjusting directly the ε, we tuned iteratively the modulation parameter τ to produce a given compression ratio (target entropy). Note that high values of τ increase the width of the ε tube, which in turn produce lower numbers of support vectors and consequently yield higher compression ratios.

Figure 7. SVR performance (selected samples) in different image-representation domains (at 0.5 bits/pix). Each panel represents the image vector (and the selected samples by the SVR learning) corresponding to the highlighted block as well as a zoom of the resulting reconstructed image. The top-left figure represents the result of the encoding and reconstruction of the image in the spatial domain. In the DCT-based cases, after transforming the 16x16 blocks using a 2-D DCT, and zigzagging its 256 coefficients, we perform support vector regression through (top-right) the RKi-1 algorithm, (bottom-left) CSF-SVR, and (bottom-

Figure 7. SVR performance (selected samples) in different image-representation domains (at 0.5 bits/pix)

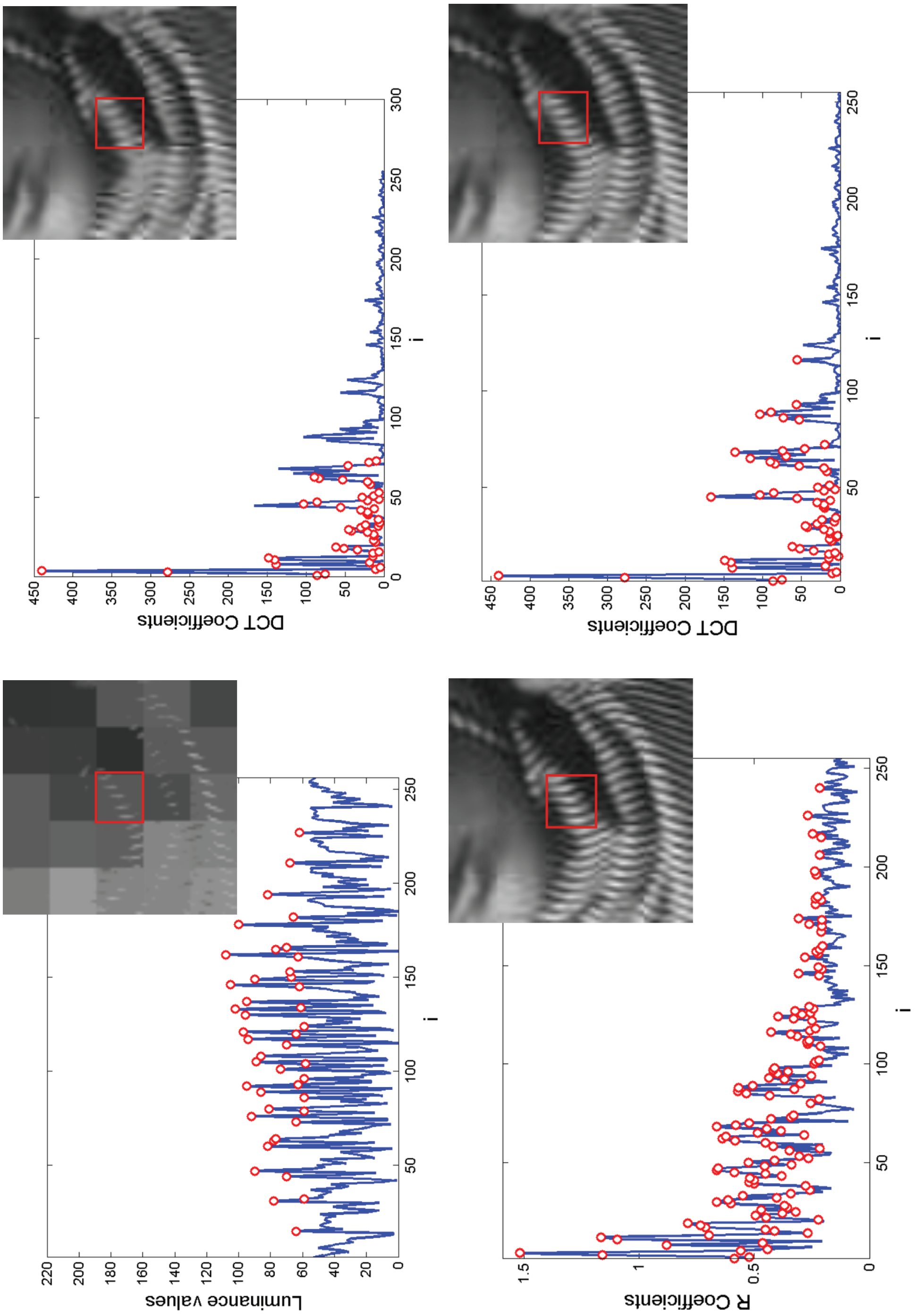
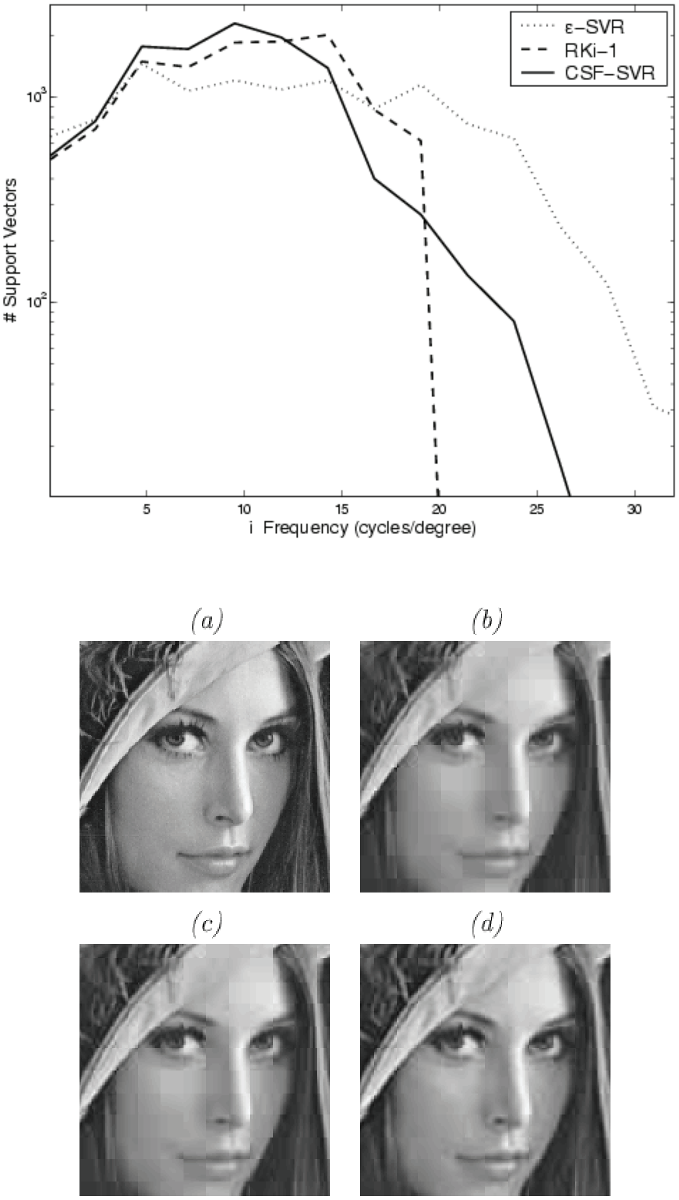


Figure 8. Analysis of the influence of the profile on SVM-based image-coding schemes in the DCT domain. The left panel shows the distribution of support vectors for each ε profile as a function of the frequency in the Lena image. The right panel shows (a) a zoom of the original Lena image at 8 bits/pixel (the bit rate for this example is 0.3 bpp; 27:1), (b) the ε-SVR (constant insensitivity in the DCT domain), (c) RKi-1, and (d) CSF-SVR.



right) NL-SVR. Then, the resulting support vectors (circles) are used to reconstruct the signal. These illustrative examples show that using the same SVR abilities, (a) the spatial domain representation is highly unsuitable for SVR-based image coding, and (b) SVR-CSF and the NL-SVR take into account higher frequency details by selecting support vectors in the high-frequency range when necessary, something that is not accomplished with the RKi-1 method. This second effect enhances the subjective behavior of these methods. The advantage of the nonlinear method is that it gives rise to better contrast reproduction.

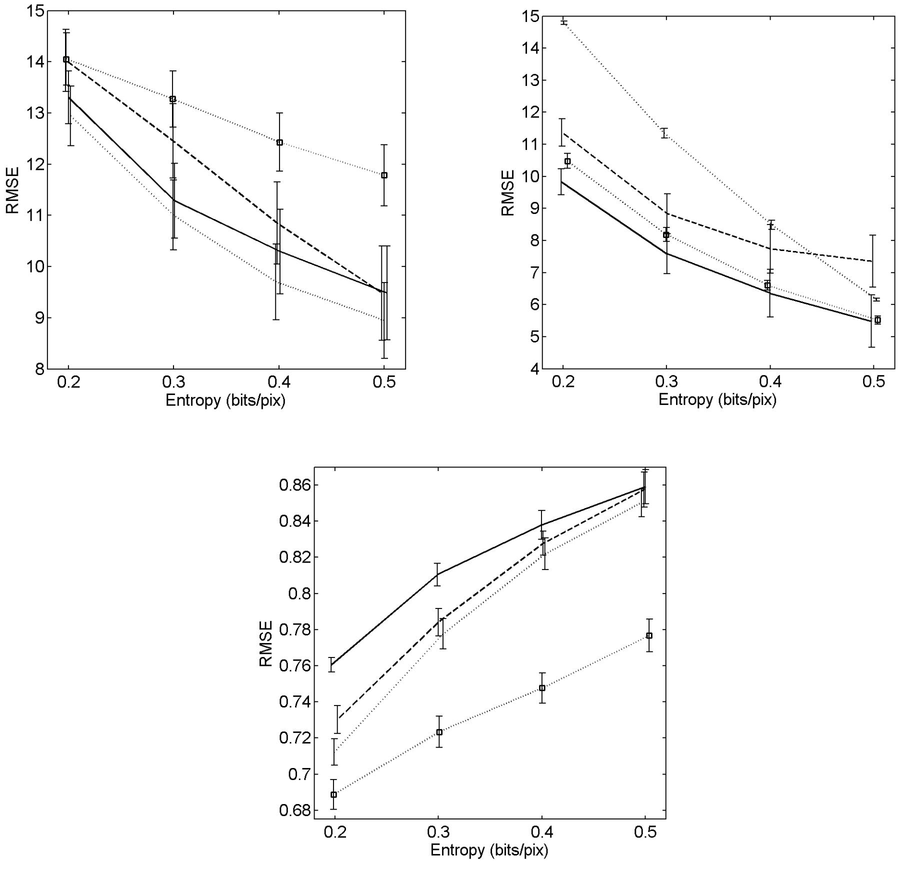
Distribution of Support Vectors in the DCT Domain

Tailoring different ε profiles produces critically different support vector distributions in the frequency domain and hence different error distributions in this domain. Therefore, different ε profiles lead to results of quite different perceptual quality. Figure 8 (left) shows a representative example of the distribution of the selected SVs by the ε-SVR (constant insensitivity in the DCT domain), the RKi-1, and the CSF-SVR models. These distributions reflect how the selection of a particular insensitivity profile modifies the learning behavior of the SVMs.

In Figure 8 (right), we illustrate the effect that the considered ε*i* profile has on the encoded images. Using a straightforward constant ε for all coefficients (ε-SVR approach) concentrates more support vectors in the low-frequency region because the variance of these DCT coefficients in natural images is higher (Clarke, 1981; Malo, Ferri, Albert, Soret, & Artigas, 2000). However, it still yields a relatively high number of support vectors in the high-frequency region. This is inefficient because of the low subjective relevance of that region (see Figure 2). Considering these vectors will not significantly reduce the (perceptual) reconstruction error while it increases the entropy of the encoded signal. The RKi-1 approach (Robinson & Kecman, 2003) uses a constant ε, but the authors solve the above problem by neglecting the high-frequency coefficients in training the SVM for each block. This is equivalent to the use of an arbitrarily large insensitivity for the high-frequency region. As a result, this approach relatively allocates more support vectors in the low- and medium-frequency regions. This modification of the straightforward uniform approach is qualitatively based on the basic low-pass behavior of human vision. However, such a crude approximation (that implies no control of the distortion in the high-frequency region) can introduce annoying errors in blocks with sharp edges. The CSF-SVR approach uses a variable ε according to the CSF, and thus takes into account the perception facts reviewed previously, giving rise to a (natural) concentration of support vectors in the low- and medium-frequency region. Note that this concentration is even bigger than in the RKi-1 approach. However, the proposed algorithm does not neglect any coefficient in the learning process. This strategy naturally reduces the number of allocated support vectors in the high-frequency region with regard to the straightforward uniform approach, but it does not prevent selecting some of them when it is necessary to keep the error below the selected threshold, which may be relevant in edge blocks.

The visual effect of the different distribution of the support vectors due to the different insensitivity profiles is clear in Figure 8 (right). First, it is obvious that the perceptually based training leads to better overall subjective results: The annoying blocking artifacts of the ε-SVR and RKi-1 algorithms are highly reduced in the CSF-SVR, giving rise to smoother and perceptually more acceptable images. Second, the blocking artifacts in the ε-SVR and RKi-1 approaches may come from different reasons. On the one hand, the uniform ε-SVR wastes (relatively) too many support vectors (and bits) in the high-frequency region in such a way that noticeable errors in the low-frequency components (related to the average luminance in each block) are produced. However, note that due to the allocation of more vectors in the high-frequency region, it is the method that better reproduces details. On the other hand, neglecting the high-frequency coefficients in the training (RKi-1 approach) does reduce the blocking a little bit, but it cannot cope with high-contrast edges that also produce a lot of energy in the high-frequency region (for instance, Lena's cheek on the dark hair background).

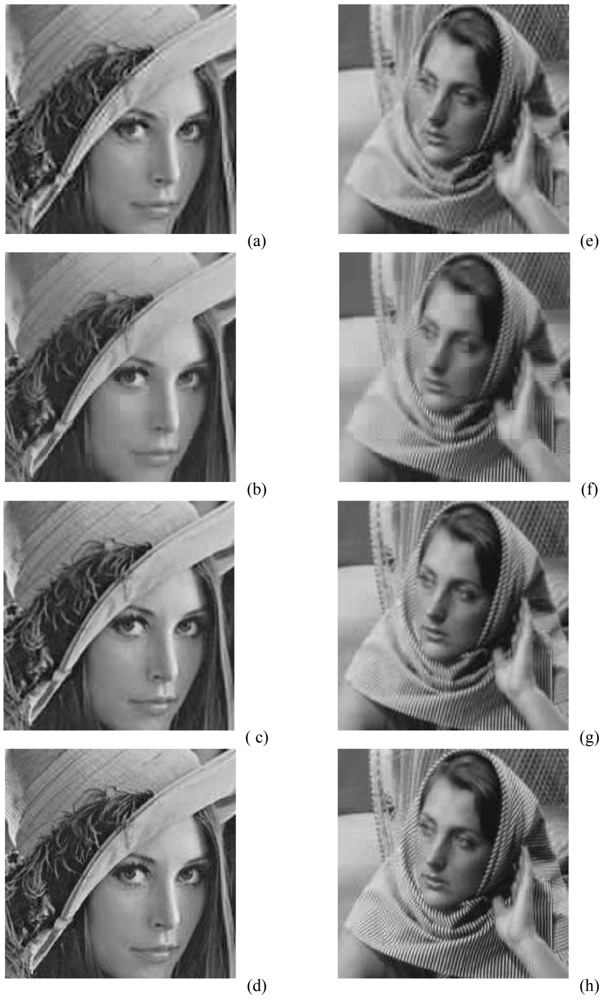
Figure 9. Average rate-distortion curves over four standard images (Lena, Barbara, Boats, Einstein) using objective and subjective measures for the considered SVM approaches (RKi-1 is dotted, CSF-SVR is dashed, and NL-SVR is solid). JPEG (dotted squares) has also been included for reference purposes. The following are also given: (a) RMSE distortion (left), (b) MPE (Gomez et al., 2005; Malo et al., 2006; Malo et al., 2000) (center), and (c) SSIM (Wang et al., 2004) (right).



Compression Performance

In this section, we analyze the performance of the algorithms through rate-distortion curves (Figure 9) and explicit examples for visual comparison (Figure 10). In order to assess the quality of the coded images, three different measures were used: the standard (Euclidean) RMSE, the MPE (Gomez et al., 2005; Malo et al., 2006; Malo et al., 2000), and the (also perceptually meaningful) structural similarity (SSIM) index (Wang, Bovik, Sheikh, & Simoncelli, 2004).

Figure 10. Examples of decoded Lena (a-d) and Barbara (e-h) images at 0.3 bits/pix, encoded by using JPEG (a, e), RKi-1 (b, f), CSF-SVR (c, g), and NL-SVR (d, h)



According to the standard Euclidean MSE point of view, the performance of the SVM algorithms is basically the same (note the overlapped big deviations in Figure 9a). However, it is widely known that the MSE results are not useful to represent the subjective quality of images, as extensively reported elsewhere (Girod, 1993; Teo & Heeger, 1994; A. B. Watson & Malo, 2002). When using more appropriate (perceptually meaningful) quality measures (Figures 9b-9c), the NL-SVR outperforms previously reported SVM methods.

Figure 10 shows representative results of the considered SVM strategies on standard images (Lena and Barbara) at the same bit rate (0.3 bits/pix). The visual inspection confirms that the numerical gain in MPE and SSIM shown in Figure 9 is also perceptually significant. Some conclusions can be extracted from this figure. First, as previously reported in Gomez et al. (2005), RKi-1 leads to poorer (blocky) results because of the too-crude approximation of the CSF (as an ideal low-pass filter) and the equal relevance applied to the low-frequency DCT coefficients. Second, despite the good performance yielded by the CSF-SVR approach to avoid blocking effects, it is worth noting that high-frequency details are smoothed (e.g., see Barbara's scarf). These effects are highly alleviated by introducing SVR in the nonlinear domain. See, for instance, Lena's eyes, her hat's feathers, or the better reproduction of the high-frequency pattern in Barbara's clothes.

Conclusion, Discussion, and Further Work

In this chapter, we have analyzed the use and performance of SVR models in image compression. A thorough revision of perception and statistics facts became strictly necessary to improve the standard application of SVR in this application field. First, the selection of a suitable (and perceptually meaningful) working domain requires one to map the input data first with a linear transformation and then with an (optional) nonlinear second transformation. Second, working in a given domain also imposes certain constraints on the definition of the most suitable insensitivity zone for training the SVR. As a consequence, the joint consideration of domain and insensitivity in the proposed scheme widely facilitates the task of function approximation, and enables a proper selection of support vectors rather than a reduced amount of them.

Several examples have illustrated the capabilities of using SVR in transform coding schemes, and revealed it to be very efficient in terms of perceptually meaningful rate-distortion measurements and visual inspection. However, further studies should be developed in the immediate future, which could lead to some refinements and improved compression, for example, by performing sparse learning regression in more sophisticated nonlinear perceptual domains, or by replacing the DCT with wavelet transforms. Finally, further research should be addressed to analyze the statistical effects of SVM feature extraction and its relation to more general nonlinear ICA techniques (Hyvarinen et al., 2001; Murray & Kreutz-Delgado, 2004).

References

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Comp. Neur. Sys., 12*, 241-253.

Buccigrossi, R., & Simoncelli, E. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing, 8*(12), 1688-1701.

Campbell, F., & Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology, 197*(3), 551-566.

Camps-Valls, G., Soria-Olivas, E., Pérez-Ruixo, J., Artés-Rodriguez, A., Pérez-Cruz, F., & Figueiras-Vidal, A. (2001). A profile-dependent kernel-based regression for cyclosporine concentration prediction. *Neural Information Processing Systems (NIPS): Workshop on New Directions in Kernel-Based Learning Methods.* Retrieved from http://www.uv.es/~gcamps

Clarke, R. (1981). Relation between the Karhunen-Loeve transform and cosine transforms. In *Proceedings IEE, Pt. F, 128*(6), 359-360.

Epifanio, I., Gutiérrez, J., & Malo, J. (2003). Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding. *Pattern Recognition, 36*(8), 1799-1811.

Gersho, A., & Gray, R. (1992). *Vector quantization and signal compression.* Boston: Kluwer Academic Press.

Girod, B. (1993). What's wrong with mean-squared error. In A. B. Watson (Ed.), *Digital images and human vision* (pp. 207-220). MIT Press.

Gomez, G., Camps-Valls, G., Gutierrez, J., & Malo, J. (2005). Perceptual adaptive insensitivity for support vector machine image coding. *IEEE Transactions on Neural Networks, 16*(6), 1574-1581.

Gutierrez, J., Ferri, F., & Malo, J. (2006). Regularization operators for natural images based on non-linear perception models. *IEEE Trans. Im. Proc., 15*(1), 189-200.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*, 181-198.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* New York: John Wiley & Sons.

Legge, G. (1981). A power law for contrast discrimination. *Vision Research, 18*, 68-91.

Malo, J., Epifanio, I., Navarro, R., & Simoncelli, E. (2006). Non-linear image representation for efficient perceptual coding. *IEEE Trans. Im. Proc., 15*(1), 68-80.

Malo, J., Ferri, F., Albert, J., Soret, J., & Artigas, J. M. (2000). The role of perceptual contrast non-linearities in image transform coding. *Image & Vision Computing, 18*(3), 233-246.

Malo, J., & Gutiérrez, J. (in press). V1 non-linearities emerge from local-to-global non-linear ICA. *Network: Comp. Neural Syst., 17*(1).

Malo, J., Pons, A., & Artigas, J. (1997). Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. *Image & Vision Computing, 15*(7), 535-548.

Murray, J. F., & Kreutz-Delgado, K. (2004). Sparse image coding using learned overcomplete dictionaries. *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2004)*.

Navarro, Y., Gutiérrez, J., & Malo, J. (2005). Gain control for the chromatic channels in JPEG2000. In *Proceedings of 10th International Conference of AIC, 1*(1), 539-542.

Peli, E. (1990). Contrast in complex images. *JOSA A, 7*, 2032-2040.

Pons, A. M., Malo, J., Artigas, J. M., & Capilla, P. (1999). Image quality metric based on multidimensional contrast perception models. *Displays, 20*, 93-110.

Robinson, J., & Kecman, V. (2000). The use of support vector machines in image compression. In *Proceedings of the International Conference on Engineering Intelligence Systems, EIS2000, 36* (pp. 93-96).

Robinson, J., & Kecman, V. (2003). Combining support vector machine learning with the discrete cosine transform in image compression. *IEEE Transactions on Neural Networks, 14*(4), 950-958.

Simoncelli, E. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology, 13*, 144-149.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*, 199-222.

Taubman, D., & Marcellin, M. (2001). *JPEG2000: Image compression fundamentals, standards and practice.* Boston: Kluwer Academic Publishers.

Teo, P., & Heeger, D. (1994). Perceptual image distortion. *Proceedings of the SPIE Conference: Human Vision, Visual Processing, and Digital Display V, 2179* (pp. 127-141).

Wallace, G. (1991). The JPEG still picture compression standard. *Communications of the ACM, 34*(4), 31-43.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Im. Proc., 13*(4), 600-612.

Watson, A., & Solomon, J. (1997). A model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A, 14*(9), 2379-2391.

Watson, A. B., & Malo, J. (2002). Video quality measures based on the standard spatial observer. In *Proceedings of the IEEE International Conference on Image Proceedings* (Vol. 3, pp. 41-44).

Zeng, W., Daly, S., & Lei, S. (2002). An overview of the visual optimization tools in JPEG2000. *Sig.Proc.Im.Comm., 17*(1), 85-104.