CCIT Report #571

January 2006

On Subband Transform Coding for Color Image Compression

Evgeny Gershikov and Moshe Porat

Department of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel.

Abstract

Although subband transform coding is a useful approach to image compression, the performance of this method has not been analyzed so far for color images, especially when the selection of color components is considered. Obviously, the RGB components are not suitable for such a compression method due to their high inter-color correlation. On the other hand, the common selection of YUV or YIQ is rather arbitrary and in most cases not optimal.

In this work we introduce a rate-distortion model for color image compression and employ it to find the optimal color components and optimal bit allocation (optimal rates) for the compression. We show that the DCT (Discrete Cosine Transform) can be used to transform the RGB components into an efficient set of color components suitable for subband coding. The optimal rates can be also used to design adaptive quantization tables in the coding stage with results superior to fixed quantization tables. Based on the presented results, our conclusion is that the new approach can improve presently available methods for color compression.

Corresponding author:
Moshe Porat,
Dept. of Electrical Engineering,
Technion, Haifa 32000, Israel.

Email: mp@ee.technion.ac.il
Fax: +972-4-829-5757
Tel.: +972-4-829-4684

1 Introduction

It is well known that natural images are characterized by high correlation between their RGB components [1], [2], [5], [8], [20]. This data redundancy has to be considered in order to reduce the volume of information that has to be stored or transmitted for a given image. Most of the techniques for color image compression reduce the redundancies between the colors components by transforming the colors primaries into a decorrelated color space such as YIQ or YUV [4], [10], [18] or by performing the Karhunen-Loeve Transform on the color components in some color space [3], [6], [19]. Another approach to exploiting the high correlations of the RGB primaries is by approximating subordinate colors as a function of a base color in the RGB domain [2], [12]. When using decorrelation methods, the choice of the YUV or YIQ color space is the most common, but it is usually not optimal as demonstrated in this work in the context of subband transform coding systems.

1.1 Subband transforms

Subband transforms include the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT) and Karhunen-Loeve Transform (KLT), as well as wavelet tree decompositions, wavelet packets and filter banks. The most familiar systems based on subband transform coding are the JPEG [18] and JPEG2000 [4], [10] standards for image coding. Other examples include the EZW wavelet based algorithm [14] or the CEB algorithm [16] for images and for example Uniform DFT Filter Banks [13] for speech coding.

Subband transforms are a generalization of block transforms. If we consider a non-expansive transform (which transforms an input signal into an output signal of the same size), the input sequence x is divided into m dimensional vectors $\mathbf{x}[k]$ and each is transformed to form the (m dimensional) output vector $\mathbf{y}[k]$. However, in contrast to a simple block transform, multiple input vectors are used to form one output vector according to:

$$\mathbf{y}[k] = \sum_{i \in z} \mathbf{A}^H[i]\mathbf{x}[k-i], \tag{1}$$

where $\mathbf{A}^H[i]$ is a series of matrices. If $\mathbf{A}[i] \neq 0$ only for $i = 0$, (1) describes a block transform. The inverse transform, if exists, is of the form:

$$\mathbf{x}[k] = \sum_{i \in z} \mathbf{S}[i]\mathbf{y}[k-i]. \tag{2}$$

1.2 Filter bank interpretation of subband transforms

The subband transforms can be considered as a filter bank. It can be shown that (1) can be rewritten in the form:

$$y_q[n] = y[nm+q] = \sum_i \mathbf{a}_q^H[i]\mathbf{x}[n-i], \tag{3}$$

where $y_q[n]$ is the sequence of the q^{th} components of each output vector $\mathbf{y}[k]$, that we will refer to as the q^{th} subband. Also $\mathbf{a}_q[i]$ denotes the q^{th} column of $\mathbf{A}[i]$. Writing the inner product in (3) explicitly, we get:

$$y_q[n] = \sum_i \sum_{j=0}^{m-1} (\mathbf{a}_q[i])_j^* \; x[m(n-i)+j] \tag{4}$$

with $(\mathbf{a}_q[i])_j^*$ denoting the complex conjugate of the j^{th} element of $\mathbf{a}_q[i]$. The last equation can be rewritten as:

$$y_q[n] = \sum_l h_q[l] \; x[mn-l] = (x * h_q)[mn], \tag{5}$$

where the scalar filter responses $h_q[l]$ are related to the matrix series $\mathbf{A}[i]$ (to the element at the (q,j) position) according to:

$$(\mathbf{A}^H[i])_{q,j} = h_q[mi-j]. \tag{6}$$

It is clear that the meaning of (5) is that the input signal x passes through a bank of (analysis)

filters h_q and then is down-sampled by a factor of m to obtain the subband $\mathbf{y}_q[n]$. In a similar way the synthesis equation (2) can be rewritten in the form:

$$\begin{aligned} x_p[n] &= x[nm+p] = \sum_{i\in z}\sum_{q=0}^{m-1}(s_q[i])_p y_q[n-i] = \sum_{q=0}^{m-1}\sum_{i\in z} s_q[n-i])_p y_q[i] \\ &= \sum_{q=0}^{m-1}\sum_{i\in z} g_q[nm+p-mi]y_q[i] = \sum_{q=0}^{m-1}(\tilde{y}_q * g_q)[nm+p]. \end{aligned} \tag{7}$$

Here $\tilde{y}_q[i]$ is defined as:

$$\tilde{y}_q[i] = \begin{cases} y_q\left[\frac{i}{m}\right] & \text{if} \quad m \quad \text{divides} \quad i \\ 0 \quad \text{otherwise} \end{cases},$$

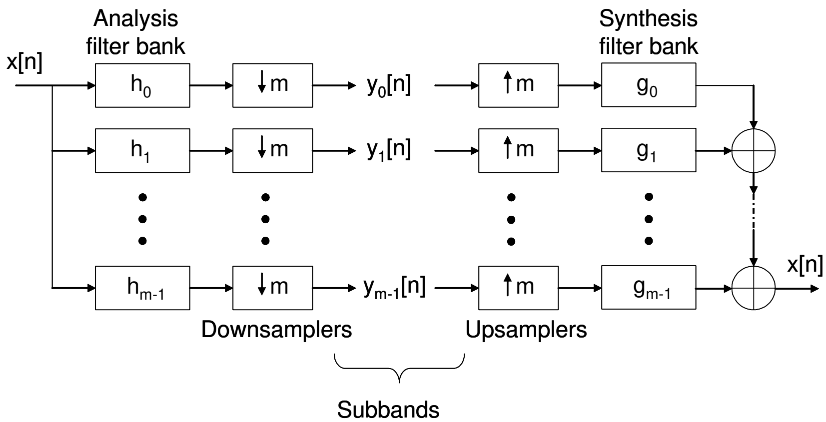


Figure 1: **Subband transforms filter bank representation: analysis and synthesis.**

$\mathbf{s_q}[i]$ is the q^{th} column of the matrix $\mathbf{S}[i]$, $(\mathbf{s_q}[i])_p$ is its p^{th} element and the synthesis scalar filters g_q are related to the matrix series $\mathbf{S}[i]$ by:

$$g_q[mi+j] = (\mathbf{s}_q[i])_j = (\mathbf{S}[i])_{j,q}. \tag{9}$$

Here $\mathbf{x}_p[n]$ denotes the p^{th} polyphase component of the input sequence x similarly to the notation $\mathbf{y}_q[n]$.

The meaning of (7) is that the subbands $\mathbf{y}_q[n]$ are up-sampled by a factor of m and when filtered by the synthesis filter bank g_q and summed up, the result is the reconstructed input sequence x. Both the analysis and synthesis operations in the form of a filter bank are summed up in Fig. 1.

1.3 Vector space interpretation of subband transforms

Equation (5) for the filter bank analysis can be rewritten in the form of the vector product of $\mathcal{C}^n$ in the following way:

$$\begin{aligned} y_q[n] &= \sum_l h_q[l]\; x[mn-l] = \sum_l h_q[-l]\; x[mn+l] \\ &= \sum_l h_q[mn-l]\; x[l] = \sum_l \alpha_q^*[l-mn]\; x[l] = <\mathbf{x}, \boldsymbol{\alpha}_q^{(n)}>, \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha}_q[k] = h_q^*[-k]$ are sequences derived from the subband analysis filters and $\boldsymbol{\alpha}_q^{(n)}$ it the sequence $\boldsymbol{\alpha}_q[k]$ delayed by mn samples in vector form. Therefore, (10) suggests that the transform coefficients $\mathbf{y}_q[n]$ can be regarded as the result of the inner product of the input vector $\mathbf{x}$ and the analysis vectors $\boldsymbol{\alpha}_q^{(n)}$. In a similar way (7) describing the filter bank synthesis operation can be rewritten as:

$$\begin{aligned} x[k] &= \sum_{q=0}^{m-1} \sum_{i \in z} g_q[k-mi] y_q[i] \\ &= \sum_{q=0}^{m-1} \sum_{i \in z} y_q[i] s_q^{(i)}[k] \end{aligned} \tag{11}$$

with $s_q^{(i)}[k]$ denoting the sequences $s_q[k] = g_q[k]$ (that are identical to the synthesis filters) delayed by mi samples ($s_q^{(i)}[k] = g_q[k-mi]$). Therefore, in vector form:

$$\mathbf{x} = \sum_{q=0}^{m-1} \sum_{i \in z} y_q[i] \mathbf{s}_q^{(i)}, \tag{12}$$

where $\mathbf{s}_q^{(i)}$ are the synthesis vectors. Here the outer sum (on q) is on the subbands and the inner one (on i) is on the transform coefficients in each subband.

Note that if the synthesis vectors $\mathbf{s}_{q1}^{(i)}$ are mutually orthogonal, the subband transform is called orthogonal and it is called orthonormal if they also have unit norm. Also note that the subband transform can be applied iteratively to the subbands resulting in a multi-level or tree decomposition. Then the subbands will no longer be uniform as in Fig. 1 in their sample rate (the number of coefficients in each subband) and passband bandwidth.

This paper is organized as follows. In Section 2 we introduce the rate-distortion model for subband transforms and extend it to color images. In section 3 we discuss the derivation of the optimal CCT and subband rates based on the model. Section 4 is devoted to applications of the model to image compression and discussion of the results. Finally, conclusions are summarized in Section 5.

2 Rate-Distortion theory of subband coders

2.1 Rate Distortion of the PCM scheme

The rate-distortion performance of a scalar quantizer with independently coded samples for a stochastic source x with variance σ_x^2 can be modeled as [15]:

$$d(R) = g(R) \sigma_x^2 2^{-2R}, \tag{13}$$

where $d()$ is the MSE (Mean Square Error) distortion, R is the rate in bits per sample and $g()$ is a weak function of the source. For large enough R, $g(R) \cong \varepsilon^2$, where ε^2 is a constant dependent upon the distribution of x and therefore:

$$d(R) = \varepsilon^2 \sigma_x^2 2^{-2R}. \tag{14}$$

The scheme that performs scalar quantization with independent coding of the source samples is called PCM (Pulse Code Modulation). An example of such a system is a uniform scalar quantizer with entropy coded output.

2.2 Rate-Distortion of an orthonormal subband transform

Consider an encoder that first transforms an N samples source signal x into a set of subbands by an orthonormal (hence energy preserving) subband transform and then each subband coefficients y_b are coded independently by the PCM scheme, while the decoder reconstructs the signal $\mathbf{x}_{rec}$ from the dequantized transform coefficients $\hat{y}_b$. We get [15]:

$$d_x = E\left[\frac{1}{N}\sum_k (\mathbf{x}[k] - \mathbf{x}_{rec}[k])^2\right] = \frac{1}{N}E\left[\|\mathbf{x} - \mathbf{x}_{rec}\|^2\right] \quad (N \text{ - length of the signals}). \tag{15}$$

Substituting

$$\mathbf{x} = \sum_{q=0}^{B-1}\sum_{i\in z} y_q[i]\mathbf{s}_q^{(i)}, \qquad \mathbf{x_{rec}} = \sum_{q=0}^{B-1}\sum_{i\in z} \hat{y}_q[i]\mathbf{s}_q^{(i)}, \tag{16}$$

according to (12) where B is the number of subbands, d_x becomes:

$$d_x = \frac{1}{N}E\left[\left\|\sum_{b=0}^{B-1}\sum_i (y_b[i] - \hat{y}_b[i])\,\mathbf{s}_b^{(i)}\right\|^2\right] = \frac{1}{N}E\left[\left\|\sum_{b=0}^{B-1}\sum_i \delta y_b[i]\mathbf{s}_b^{(i)}\right\|^2\right], \tag{17}$$

where

$$\delta y_b[i] \triangleq y_b[i] - \hat{y}_b[i] \tag{18}$$

is the (quantization) error in the coefficient i of subband b. Therefore:

$$d_x = \frac{1}{N}\sum_{b=0}^{B-1}\sum_{p=0}^{B-1}\sum_i\sum_j E\left[\delta y_b[i]\delta y_p[j]\right]\left\langle \mathbf{s}_b^{(i)}, \mathbf{s}_p^{(j)}\right\rangle \tag{19}$$

Using the orthonormality assumption:

$$\left\langle \mathbf{s}_b^{(i)}, \mathbf{s}_p^{(j)} \right\rangle = \delta[b-p]\delta[i-j], \tag{20}$$

(19) is reduced into the form:

$$d_x = \frac{1}{N}\sum_{b=0}^{B-1}\sum_i E\left[(\delta y_b[i])^2\right] = \frac{1}{N}\sum_{b=0}^{B-1}\sum_i d_b, \tag{21}$$

where d_b is the MSE distortion of subband b:

$$d_b \triangleq E\left[(\delta y_b[i])^2\right]. \tag{22}$$

(21) can be rewritten in the form:

$$d_x = \sum_{b=0}^{B-1} \eta_b d_b \tag{23}$$

with η_b denoting the ratio between the number of coefficients in subband b and the total number of samples in the source N. Since the transform is also non-expansive, the following equation holds:

$$\sum_{b=0}^{B-1} \eta_b = 1 \tag{24}$$

Substituting $d(R)$ of (14) as the MSE distortion expression of each subband d_b into (23) we get:

$$d_x = d(R) = \sum_{b=0}^{B-1} \eta_b \sigma_b^2 \varepsilon^2 2^{-2R_b}. \tag{25}$$

where σ_b^2 is the variance of the subband indexed b ($b \in [0, B-1]$) and R_b is the rate allocated to it.

2.3 Rate-Distortion of a non-orthogonal subband transform

When examining a non-orthogonal subband transform, the derivation of the expression for the MSE distortion for such a transform is the same as for orthogonal transforms in section 2.2 up to (19), i.e., for a signal $\mathbf{x}$ encoded by a subband transform coder and decoded into the signal $\mathbf{x}_{rec}$:

$$d_x = E\left[\frac{1}{N}\sum_k (\mathbf{x}[k] - \mathbf{x}_{rec}[k])^2\right] = \frac{1}{N}\sum_{b=0}^{B-1}\sum_{p=0}^{B-1}\sum_i\sum_j E\left[\delta y_b[i]\delta y_p[j]\right]\left\langle \mathbf{s}_b^{(i)}, \mathbf{s}_p^{(j)} \right\rangle. \tag{26}$$

However, we cannot assume now that the vectors $\mathbf{s}_b^{(i)}$ and $\mathbf{s}_p^{(j)}$ are orthogonal for $i \neq j$ or $b \neq p$. The assumption here is that the quantization errors $\delta y_b[i]$ in each subband coefficient $y_b[i]$ of each subband are uncorrelated with zero mean [15]. Therefore:

$$E\left[\delta y_b[i]\delta y_p[j]\right] = \delta[b-p]\delta[i-j] \tag{27}$$

and

$$d_x = \frac{1}{N}\sum_{b=0}^{B-1}\sum_i E\left[(\delta y_b[i])^2\right]\left\|\mathbf{s}_b^{(i)}\right\|^2 = \frac{1}{N}\sum_{b=0}^{B-1}\sum_i G_b d_b. \tag{28}$$

Here

$$G_b \triangleq \left\|\mathbf{s}_b^{(i)}\right\|^2 \tag{29}$$

is independent of i and is the energy gain of subband b. Similar to (23) we rewrite (28) as:

$$d_x = \sum_{b=0}^{B-1} \eta_b G_b d_b. \tag{30}$$

Substituting the PCM MSE of (14) for d_b, we get:

$$d_x = \sum_{b=0}^{B-1} \eta_b G_b \sigma_b^2 \varepsilon^2 2^{-2R_b}, \tag{31}$$

which is a generalization of (25), since for an orthogonal transform we can use (31) with $G_b = 1$ for all $b \in [0, ..., B-1]$.

2.4 Extension to a color image

Denote each pixel in a color image in the RGB domain by a 3x1 vector $\mathbf{x} = [R\ G\ B]^T$. We first apply a color component transform to the image, denoted by a matrix $\mathbf{M}$ to obtain at each pixel a new vector of 3 components $C1, C2, C3$, denoted $\widetilde{\mathbf{x}} = [C1\ C2\ C3]^T$ and related to $\mathbf{x}$ by:

$$\widetilde{\mathbf{x}} = \mathbf{M}\mathbf{x}. \tag{32}$$

Then each component in the C1,C2,C3 color space is subband transformed, quantized and its samples are independently encoded (e.g. entropy coded). This description corresponds to such image compression algorithms as JPEG [18] and JPEG 2000 [4], when applied to a color image up to and including the quantization stage (note that after the quantization stage, JPEG or JPEG2000 encode the transform coefficients not independently, but using the correlation between them, so the above description does not apply).

We denote by $\widetilde{\mathbf{x}}_{\mathbf{rec}}$ the reconstructed image in the C1C2C3 domain after inverse quantization, and by $\mathbf{x_{rec}}$ the reconstructed image in the RGB domain, when

$$\widetilde{\mathbf{x}}_{\mathbf{rec}} = \mathbf{M}\mathbf{x_{rec}} \tag{33}$$

similarly to (32). Now we can define the error covariance matrix in the RGB domain $\mathbf{Er}$, given by

$$\mathbf{Er} = E\left[(\mathbf{x} - \mathbf{x_{rec}})(\mathbf{x} - \mathbf{x_{rec}})^T\right] \tag{34}$$

and the error covariance matrix in C1C2C3 domain $\widetilde{\mathbf{Er}}$, given by

$$\widetilde{\mathbf{Er}} = E\left[(\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_{\mathbf{rec}})(\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_{\mathbf{rec}})^T\right] \tag{35}$$

where $E[\cdot]$ stands for statistic mean.
It is easy to show that:

$$\begin{aligned}\widetilde{\mathbf{Er}} &= E\left[(\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_{\mathbf{rec}})(\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_{\mathbf{rec}})^T\right] \\ &= E\left[(\mathbf{Mx} - \mathbf{Mx_{rec}})(\mathbf{Mx} - \mathbf{Mx_{rec}})^T\right] \\ &= \mathbf{M}E\left[(\mathbf{x} - \mathbf{x_{rec}})(\mathbf{x} - \mathbf{x_{rec}})^T\right]\mathbf{M}^T \\ &= \mathbf{MErM^T}.\end{aligned} \tag{36}$$

Equivalently:

$$\mathbf{Er} = \mathbf{M}^{-1}\widetilde{\mathbf{Er}}\mathbf{M}^{-T} \tag{37}$$

The average MSE between the original and reconstructed images in the RGB domain is then simply:

$$\begin{aligned}MSE = \frac{1}{3}trace(\mathbf{Er}) &= \frac{1}{3}trace\left(\mathbf{M}^{-1}\widetilde{\mathbf{Er}}\mathbf{M}^{-T}\right) \\ &= \frac{1}{3}trace\left(\widetilde{\mathbf{Er}}\mathbf{M}^{-T}\mathbf{M}^{-1}\right) \\ &= \frac{1}{3}trace\left(\widetilde{\mathbf{Er}}(\mathbf{MM}^T)^{-1}\right).\end{aligned} \tag{38}$$

Assuming that the errors in the three color components C1, C2, C3 (that occur due to the quantization inherent in the compression process) are uncorrelated, i.e.

$$E\left[(Ci - Ci_{rec})(Cj - Cj_{rec})\right] = 0, \quad i,j \in \{1,2,3\}, \quad i \neq j \tag{39}$$

(where Ci_{rec} is the reconstructed Ci component), it is clear that $\widetilde{\mathbf{Er}}$ is diagonal and therefore the expression for the average MSE of (38) simplifies to:

$$
\begin{aligned}
MSE &= \frac{1}{3}\sum_{i=1}^{3}\widetilde{\mathbf{Er}}_{ii}\left((\mathbf{M}\mathbf{M}^{T})^{-1}\right)_{ii} \\
&= \frac{1}{3}\sum_{i=1}^{3}MSE_{Ci}\left((\mathbf{M}\mathbf{M}^{T})^{-1}\right)_{ii},
\end{aligned}
\tag{40}
$$

where MSE_{Ci} denotes the MSE of the component Ci.
Using (31) for this MSE, one can easily derive the following expression for the average MSE:

$$
MSE = \frac{1}{3}\sum_{i=1}^{3}\sum_{b=0}^{B-1}\eta_b G_b \sigma_{bi}^2 \varepsilon_i^2 2^{-2R_{bi}}\left((\mathbf{M}\mathbf{M}^{T})^{-1}\right)_{ii}. \tag{41}
$$

R_{bi} stands for the rate allocated for the subband b of component i and σ_{bi}^2 is this subband's variance. The expression can also be rewritten in the following equivalent form:

$$
MSE = \frac{1}{3}\sum_{i=1}^{3}\sum_{b=0}^{B-1}\eta_b G_b \sigma_{bi}^2 \varepsilon_i^2 e^{-aR_{bi}}\left((\mathbf{M}\mathbf{M}^{T})^{-1}\right)_{ii}, \tag{42}
$$

where $a = 2ln2$.

The expression obtained can be used to find the optimal subband rates allocation for minimal MSE for a given color components transform, to find the optimal color components transform for a given rates allocation or to find both optimal rates allocation and color transform. In the next section we minimize the MSE function.

2.5 Finding the optimal rates and color components transform

We would like to minimize the MSE expression of (42) subject to the constraint of some total rate allocation R for the image (in bits per pixel). We note that the rates of the 3 color components, denoted R_i, simply sum up, i.e.,

$$
R = \sum_{i=1}^{3} R_i, \tag{43}
$$

while the subband rates of each component sum up as a weighted sum :

$$
R_i = \sum_{b=0}^{B-1}\eta_b R_{bi}. \tag{44}
$$

Using Lagrange multipliers method, we therefore have to minimize the function

$$L\left(\{R_{bi}\}, \mathbf{M}, \lambda\right) = \frac{1}{3}\sum_{i=1}^{3}\sum_{b=0}^{B-1}\eta_b G_b \sigma_{bi}^2 \varepsilon_i^2 e^{-aR_{bi}} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} + \lambda\left(\sum_{i=1}^{3}\sum_{b=0}^{B-1}\eta_b R_{bi} - R\right), \tag{45}$$

where λ denotes the Lagrange multiplier.
It can be shown (Appendix A), that minimizing for R_{bi} and $\mathbf{M}$ yields the following equations:

$$R_{bi} = \frac{R}{3} + \frac{1}{a} ln\left(\frac{\varepsilon_i^2 G_b \sigma_{bi}^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii}}{\left(\prod_{k=1}^{3} GM_k\right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} \varepsilon_k^2\right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}\right)^{\frac{1}{3}}}\right). \tag{46}$$

GM_k here is the weighted geometric mean of the subband variances of the component k (corrected by the energy gains G_b), i.e.,

$$GM_k = \prod_{b=0}^{B-1} (G_b \sigma_{bk}^2)^{\eta_b}. \tag{47}$$

We can see that the equation for optimal rates allocation depends on the color components transform $\mathbf{M}$. The optimal color component transform is the one minimizing the following target function:

$$\tilde{f}(\mathbf{M}) = \prod_{k=1}^{3} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk} GM_k. \tag{48}$$

Taking the energy gains in GM_k out of the target function, since not dependent on $\mathbf{M}$ the target function simplifies to:

$$f(\mathbf{M}) = \prod_{k=1}^{3} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk} \prod_{b=0}^{B-1} (\sigma_{bk}^2)^{\eta_b} \tag{49}$$

We will now concentrate on the $\tilde{f}(\mathbf{M})$ target function minimization. As can be seen, no constraints are needed here for the matrix minimizing this function to be invertible because of the $\prod_{k=1}^{3}\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}$ part. Also, the solution of this problem is a transform which is adaptive to the image, similar to the result of Pinhasov et al. in [9] in the context of disparity estimation in stereo vision of color images.

2.5.1 Optimal color components transform - approximated target function

The target function of (48) is not easy to minimize. In order to simplify it a bit, we introduce the concept of the subband transform coding gain. This coding gain is defined as the ratio of the MSE distortion of simple PCM coding (i.e., simple scalar quantization with independent coding of the sources samples) and the MSE of the subband transform coder (subband transform followed by PCM coding) for the same rate R. In [15], the coding gain expression G_T for a signal x with variance σ_x^2 coded by a subband coder (SC) is found to be:

$$G_T = \frac{d^{PCM}(R)}{d^{SC}(R)} = \frac{\sigma_x^2}{\prod_{b=0}^{B-1}\left(G_b \sigma_b^2\right)^{\eta_b}} = \frac{\sigma_x^2}{GM}. \tag{50}$$

Therefore, the coding gain is:

$$G_T = \frac{\sigma_x^2}{GM}. \tag{51}$$

and the geometric means GM_k of the color components can be expressed via their variances in the image domain $\widetilde{\sigma}_k^2$ and the coding gains G_{T_k}:

$$GM_k = \frac{\widetilde{\sigma}_k^2}{G_{T_k}}. \tag{52}$$

The variances $\widetilde{\sigma}_k^2$ are the diagonal elements of the covariance matrix in the C1C2C3 image domain:

$$\widetilde{\mathbf{\Lambda}} \triangleq E\left[\left(\widetilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)\left(\widetilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)^T\right], \qquad \tilde{\boldsymbol{\mu}}_{\mathbf{x}} \triangleq E\left[\tilde{\mathbf{x}}\right] \tag{53}$$

and can also be expressed using the $\mathbf{M}$ matrix and the covariance matrix in the RGB image domain:

$$\mathbf{\Lambda} \triangleq E\left[\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\right)\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\right)^T\right], \qquad \boldsymbol{\mu}_{\mathbf{x}} \triangleq E\left[\mathbf{x}\right] \tag{54}$$

as:

$$\widetilde{\sigma}_k^2 = \mathbf{m_k}^T \mathbf{\Lambda} \mathbf{m_k}, \tag{55}$$

where $\mathbf{m_1}, \mathbf{m_2}, \mathbf{m_3}$ denote the $\mathbf{M}$ matrix rows in column vector form. Substituting (52) and (55) into (48), we get:

$$\tilde{f}(\mathbf{M}) = \frac{\prod_{k=1}^{3}\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk} \mathbf{m_k}^T \mathbf{\Lambda} \mathbf{m_k}}{\prod_{k=1}^{3} G_{T_k}}. \tag{56}$$

If we attempt to minimize the numerator of the expression only, then we seek to minimize:

$$g(\mathbf{M}) \triangleq \prod_{k=1}^{3}\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk} \mathbf{m_k}^T \mathbf{\Lambda} \mathbf{m_k}, \tag{57}$$

or equivalently $log(g(\mathbf{M}))$, given by:

$$log(g(\mathbf{M})) = \sum_{k=1}^{3} log\left[\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}\right] + \sum_{k=1}^{3} log\left(\mathbf{m_k}^T \mathbf{\Lambda} \mathbf{m_k}\right), \tag{58}$$

it can be shown (Appendix B), that the gradients of this function according to the rows of $\mathbf{M}$, when examining a unitary matrix (so that $\mathbf{M}\mathbf{M}^T = \mathbf{I}$), are:

$$\nabla_{\mathbf{m_i}} log(g(\mathbf{M})) = -2\mathbf{m_i} + \frac{2\Lambda\mathbf{m_i}}{\mathbf{m_i}^T\Lambda\mathbf{m_i}}. \tag{59}$$

These equations, when equaled to zero are equivalent to:

$$\Lambda\mathbf{m_i} = \left(\mathbf{m_i}^T\Lambda\mathbf{m_i}\right)\mathbf{m_i}, \tag{60}$$

and have an interesting solution, namely the eigen vectors of $\mathbf{\Lambda}$ or the KLT. The KLT matrix is indeed orthogonal (unitary) and therefore suggests that the target function $g(\mathbf{M})$ can be minimized by a matrix with orthogonal rows.

However, even though the KLT minimizes the $g(\mathbf{M})$ function, it is only the numerator of the target function of (56). The denominator of this target function $\prod_{k=1}^{3} G_{T_k}$ is not maximal for the KLT and in fact the Karhunen-Loeve transform is not the solution of the full target function. In the next section we discuss this problem's solution.

2.5.2 Optimal color components transform - approximated solution

In order to minimize (48), we first define the subband covariance matrices $\widetilde{\mathbf{\Lambda}}_\mathbf{b}$ as:

$$\widetilde{\mathbf{\Lambda}}_\mathbf{b} \triangleq E\left[\left(\widetilde{\mathbf{Y}}_\mathbf{b} - \tilde{\boldsymbol{\mu}}_{\mathbf{Y_b}}\right)\left(\widetilde{\mathbf{Y}}_\mathbf{b} - \tilde{\boldsymbol{\mu}}_{\mathbf{Y_b}}\right)^T\right] \qquad \tilde{\boldsymbol{\mu}}_{\mathbf{Y_b}} \triangleq E\left[\widetilde{\mathbf{Y}}_\mathbf{b}\right], \tag{61}$$

where $\widetilde{\mathbf{Y}}_b = [\widetilde{Y}_{b1} \quad \widetilde{Y}_{b2} \quad \widetilde{Y}_{b3}]^T$ is the vector of the subband b coefficients of the three color components C1, C2, C3 at each index in the subband. Due to the linearity of the transform, it is easy to show that (32), suggests that each subband vector $\mathbf{Y_b} = [Y_{bR} \quad Y_{bG} \quad Y_{bB}]^T$ of the R, G, B transform coefficients goes through the same color components transform to the C1C2C3 domain:

$$\widetilde{\mathbf{Y}}_\mathbf{b} = \mathbf{M}\mathbf{Y_b}. \tag{62}$$

Therefore, $\widetilde{\mathbf{\Lambda}}_b$ can be expressed by $\mathbf{\Lambda_b}$ - the RGB subband b covariance matrix defined by:

$$\mathbf{\Lambda_b} \triangleq E\left[\left(\mathbf{Y_b} - \boldsymbol{\mu}_{\mathbf{Y_b}}\right)\left(\mathbf{Y_b} - \boldsymbol{\mu}_{\mathbf{Y_b}}\right)^T\right] \qquad \boldsymbol{\mu}_{\mathbf{Y_b}} \triangleq E\left[\mathbf{Y_b}\right] \tag{63}$$

according to:

$$\widetilde{\mathbf{\Lambda}}_\mathbf{b} = \mathbf{M}\mathbf{\Lambda_b}\mathbf{M}^T. \tag{64}$$

Then we can express the variance σ_{bk}^2 as

$$\sigma_{bk}^2 = \mathbf{m_k}^T\mathbf{\Lambda_b}\mathbf{m_k}. \tag{65}$$

Substituting in (49) and taking the logarithm of the resulting target function $f(\mathbf{M})$, we get:

$$log f(\mathbf{M}) = \sum_{k=1}^{3} \left(log\left((\mathbf{M}\mathbf{M}^T)^{-1} \right)_{kk} + \sum_{b=0}^{B-1} \eta_b log\left(\mathbf{m_k}^T \mathbf{\Lambda_b} \mathbf{m_k} \right) \right) \tag{66}$$

The exact analytic solution of this problem is difficult to find, however, based on the solution of the approximated target function $g(\mathbf{M})$ in section 2.5.1, we define by $\mathbf{M_{GKLT}}$ (GKLT - Generalized KLT) the matrix satisfying the equations:

$$\nabla_{\mathbf{m_i}} log(f(\mathbf{M})) = -2\mathbf{m_i} + 2\sum_{b=0}^{B-1} \eta_b \frac{\Lambda_b \mathbf{m_i}}{\mathbf{m_i}^T \Lambda_b \mathbf{m_i}} = 0, \tag{67}$$

which are the gradient equations for the $log(f(M))$ target function with the assumption of orthogonality for $\mathbf{M}$ ($\mathbf{M}\mathbf{M}^T = \mathbf{I}$) and are somewhat similar to (59) (see Appendix C). They can be rewritten in the form:

$$\sum_{b=0}^{B-1} \eta_b \frac{\mathbf{\Lambda_b} \mathbf{m_i}}{\mathbf{m_i}^T \mathbf{\Lambda_b} \mathbf{m_i}} = \mathbf{m_i}, \tag{68}$$

similar to (60). We further propose an iterative approach to find the $\mathbf{M_{GKLT}}$ matrix, however beforehand we are obliged to note that the vectors solving (68) are not orthogonal (so that these equations' solution does not exactly minimize the target function), but they are usually close to orthogonality, i.e., defining the cosine of the angle between $\mathbf{m_i}$ and $\mathbf{m_j}$ as

$$\cos(\theta_{ij}) = \frac{\mathbf{m_i}^T \mathbf{m_j}}{\sqrt{(\mathbf{m_i}^T \mathbf{m_i})(\mathbf{m_j}^T \mathbf{m_j})}}, \quad i, j \in \{1, 2, 3\}, \quad i \neq j, \tag{69}$$

and we can state that $\cos(\theta_{ij})$ are usually close to 0.
Also note that (68) means that $\mathbf{m_i}$ is an eigen vector of the matrix $\sum_{b=0}^{B-1} \eta_b \frac{\mathbf{\Lambda_b}}{\mathbf{m_i}^T \mathbf{\Lambda_b} \mathbf{m_i}}$ that belongs to the eigen value 1.

An iterative algorithm for finding the GKLT We propose the following algorithm:

1. Take a random 3x1 vector $\mathbf{v}$. Given the B subband covariance matrices $\mathbf{\Lambda_b}$, calculate the $\sum_{b=0}^{B-1} \eta_b \frac{\mathbf{\Lambda_b}}{\mathbf{v}^T \mathbf{\Lambda_b} \mathbf{v}}$ matrix as in (68).

2. Find the eigen values λ_i and eigen vectors of the matrix calculated in the previous step.

3. For the i^{th} eigen vector $\mathbf{v_i}$, $i \in \{1, 2, 3\}$ do:
   while($|\lambda_i - 1| \geq \varepsilon$)

   (a) Calculate the $\sum_{b=0}^{B-1} \eta_b \frac{\mathbf{\Lambda_b}}{\mathbf{v_i}^T \mathbf{\Lambda_b} \mathbf{v_i}}$ matrix.

(b) Find its eigen values and eigen vectors decomposition.

(c) Take the i^{th} eigen vector as the new $\mathbf{v_i}$ and the i^{th} eigen value as the new λ_i.

Here the ε parameter is the threshold, indicating how close will $\mathbf{v_i}$ get to solving (68). When the algorithm converges, the GKLT is the matrix with $\mathbf{v_i}, (i = 1, 2, 3)$ as its rows.

2.6 Optimal rates with down-sampling

Some coding systems (e.g. JPEG [18]) perform down-sampling on some of the color components prior to coding. For such systems the down-sampling can be taken into account by introducing down-sampling factors α_i, so that the constraint on the global rate for the image is:

$$\sum_{i=1}^{3} \alpha_i \sum_{b=0}^{B-1} \eta_b R_{bi} = R, \tag{70}$$

where for example, if the down-sampling is by a factor of 2 horizontally and vertically, then

$$\alpha_i = \begin{cases} 1 & \text{full component} \\ 0.25 & \text{down-sampled component} \end{cases}$$

Therefore, the Lagrangian of (45) changes to:

$$L\left(\{R_{bi}\}, \mathbf{M}, \lambda\right) = \frac{1}{3} \sum_{i=1}^{3} \sum_{b=0}^{B-1} \eta_b G_b \sigma_{bi}^2 \varepsilon_i^2 e^{-aR_{bi}} \left(\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\right)_{ii} + \lambda \left(\sum_{i=1}^{3} \alpha_i \sum_{b=0}^{B-1} \eta_b R_{bi} - R \right), \tag{71}$$

and the solution for the optimal rates is (Appendix D):

$$R_{bi} = \frac{R}{\sum_{j=1}^{3} \alpha_j} + \frac{1}{a} ln \left(\frac{\frac{\varepsilon_i^2 G_b \sigma_{bi}^2 \left(\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\right)_{ii}}{\alpha_i}}{\prod_{k=1}^{3} \left(\frac{\left(\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\right)_{kk} \varepsilon_k^2 GM_k}{\alpha_k} \right)^{\frac{\alpha_k}{\sum_{j=1}^{3} \alpha_j}}} \right), \quad GM_k = \prod_{b=0}^{B-1} \left(G_b \sigma_{bk}^2\right)^{\eta_b}. \tag{72}$$

The solution of (72) actually covers also the case when no down-sampling is employed of (46) if we substitute 1 for all α_i.

Note that in the derivation of the Lagrangian, we have taken the quantization errors into account for the MSE and have neglected the down-sampling errors. The reason is that taking the down-sampling errors into account complicates the theoretical analysis and makes it dependent upon the down-sampling and up-sampling schemes employed.

2.7 Optimal rates with $R_{bi} \geq 0$ constraints

Observing (72), one can note that the $ln(\cdot)$ part and as a result the whole R_{bi} expression can be negative especially when considering low rates R and small energy subbands (small σ_{bi}^2). To avoid this, we have to add the constraints $Rbi \geq 0$ to our minimization problem and therefore, the Lagrangian of (71) becomes:

$$\begin{aligned} L\left(\{R_{bi}\}, \mathbf{M}, \lambda, \{\mu_{bi}\}\right) \quad &= \quad \frac{1}{3}\sum_{i=1}^{3}\sum_{b=0}^{B-1} \eta_b G_b \sigma_{bi}^2 \varepsilon_i^2 e^{-aR_{bi}} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} \qquad (73) \\ &+\lambda\left(\sum_{i=1}^{3}\alpha_i \sum_{b=0}^{B-1}\eta_b R_{bi} - R\right) - \sum_{i=1}^{3}\sum_{b=0}^{B-1}\mu_{bi}R_{bi}, \qquad (74) \end{aligned}$$

where μ_{bi} are the Lagrange multipliers for the new constraints and the expression of (73) is relevant

to subband coders with or without down-sampling of some of the color components. Minimizing the Lagrangian for the rates R_{bi} requires finding the rates that are positive and those that are zero. We define by Act_i the set of all the active subbands in color component i, that is those subbands with positive rates:

$$Act_i \triangleq \{b \in [0, B-1] \mid R_{bi} > 0\}.$$

We also define the following:

$$\xi_i \triangleq \sum_{b \in Act_i} \eta_b, \quad GMA_i \triangleq \prod_{b \in Act_i} (G_b \sigma_{bi}^2)^{\frac{\eta_b}{\xi_i}}\ , \qquad (75)$$

i.e., the relative part of the coefficients in the active subbands from the total signal length (ξ_i) and

the weighted geometric mean of their variances (corrected by the energy gains G_b) GMA_i. Then the solution becomes (Appendix E):

$$R_{bi} = \frac{R}{\sum_{j=1}^{3} \alpha_j \xi_j} + \frac{1}{a} ln \left(\frac{\frac{\varepsilon_i^2 G_b \sigma_{bi}^2 \left((\mathbf{MM}^T)^{-1}\right)_{ii}}{\alpha_i}}{\prod_{k=1}^{3} \left(\frac{\left((\mathbf{MM}^T)^{-1}\right)_{kk} \varepsilon_k^2 GMA_k}{\alpha_k} \right)^{\frac{\alpha_k \xi_k}{\sum_{j=1}^{3} \alpha_j \xi_i}}} \right) \quad (b \in Act_i). \tag{76}$$

2.7.1 How do we know which subbands are active ($R_{bi} > 0$) ?

A reasonable question that rises considering (76) is how can we know which subbands are the active ones. The following algorithm can be used to find the active subbands iteratively:

Proposed algorithm

1. Assume all the subbands are active and calculate the rates.
2. While some $R_{bi} < 0$
   - Set $Act_i = \{b \in [0, B-1] \mid R_{bi} > 0\}$
   - Calculate new rates.
3. Check that the Lagrange multipliers $\mu_{bi} \geq 0$, where (Appendix E):

$$\mu_{bi} = \begin{cases} \frac{a}{3}\eta_b \left((\mathbf{MM}^T)^{-1}\right)_{ii} \varepsilon_i^2 \left(\sigma_{0i}^2 e^{-aR_{0i}} - \sigma_{bi}^2\right) & b \notin Act_i \\ 0 & b \in Act_i \end{cases}.$$

   Here σ_{0i}^2 and R_{0i} are the variance and rate of the subband 0 of component i, respectively, when subband 0 is assumed to have the maximal variance for that component (e.g. the DC subband for the DCT) and therefore is assumed to be active ($0 \in Act_i$).

Next we present the DCT (Discrete Cosine Transform) color components transform and its connection to the $f(M)$ target function.

2.8 DCT color components transform performance

The one dimensional 3x1 DCT transform matrix

$$\mathbf{M_{DCT}} = \begin{pmatrix} 0.5774 & 0.5774 & 0.5774 \\ 0.7071 & 0.0000 & -0.7071 \\ 0.4082 & -0.8165 & 0.4082 \end{pmatrix} \tag{77}$$

or (if normalized to L_1 norm of 1 for each row, so that it transforms for example the intensity values range of [0,255] to a new range of the same width):

$$\widetilde{\mathbf{M}}_{\mathbf{DCT}} = \begin{pmatrix} 0.3333 & 0.3333 & 0.3333 \\ 0.5000 & 0.0000 & -0.5000 \\ 0.2500 & -0.5000 & 0.2500 \end{pmatrix} \tag{78}$$

can be used on the $[R\ G\ B]^T$ vector as the color components transform. It has been shown [17], [7] that the DCT can be used as an approximation for the KLT transform. However, we have stated earlier that the KLT is not optimal from the minimal MSE point of view and it does not minimize the $f(\mathbf{M})$ target function of (48). Therefore, the DCT is not of interest as an approximation of the KLT, but because of the following property:
although the DCT is not a solution to (48), it usually achieves very close values of the target function to those of the solution and the optimal transform, i.e. the one minimizing $f(\mathbf{M})$. The performance of the DCT vs. other transforms is demonstrated in the next section.

2.9 Performance of different color component transforms for a DCT coder

In this section we present results of images compression using a DCT based coder with different color components transforms (CCT) and optimal rates. The coding system consists of the following stages:

* Apply the CCT as preprocessing.
* Transform each color components by two-dimensional DCT.
* Uniform scalar quantization is then applied to each subband separately. The quantization step sizes are chosen so that optimal subband rates are achieved as will be described in subsection 3.1.
* Apply lossless coding of the transform coefficients, similar to the baseline JPEG algorithm [18]:
  - Use differential coding to the DC coefficients
  - Use zigzag scan, run length coding and Huffman variable length coding (combined with variable-length integer codes) for the AC coefficients.

We use (76) for the optimal rates. The results for different CCTs and the same image rate or same compression ratio (CR) are summarized in Table 1. The 'Opt Trans' CCT is the one minimizing the target function of (49) found by numeric methods of optimization. The performance criterion here is the PSNR (Peak Signal to Noise Ratio) defined as:

$$PSNR = 10log_{10}\left(\frac{255^2}{MSE}\right) \tag{79}$$

for a 8 bit per pixel (each color component) image.

It can be concluded from Table 1, that the optimal CCT indeed yields the best results on average, but the DCT is close behind. The GKLT can sometimes serves as a good approximation of the optimal transform (for example for the Lena, Girl, Landscape images), however, sometimes it fails (e.g., for Peppers and Baboon). All these transforms are on average better than the RGB to YUV transform and the KLT, which fails for low inter-color correlation images in the RGB domain, such as Peppers and Baboon, but gives good performance for such images as Lena, where the RGB components inter-color correlations are high.

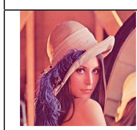
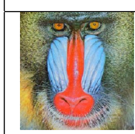
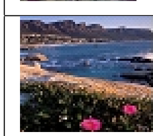
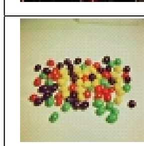
	Image	**YUV**	**DCT**	**KLT**	**GKLT**	**Opt Trans**	**CR**
	Lena	30.019	30.372	30.355	30.243	30.285	37.75
	Peppers	30.013	30.144	29.475	29.965	30.148	30.05
	Baboon	30.010	30.468	28.595	30.204	30.540	13.33
	Girl	30.015	30.359	30.343	30.417	30.450	43.87
	Tree	30.018	30.295	30.601	30.508	30.649	13.76
	Landscape	30.019	30.382	30.195	30.140	30.145	13.26
	Jelly Beans	30.019	30.294	30.294	30.225	30.314	49.04
	Mean PSNR	**30.016**	**30.330**	**29.980**	**30.243**	**30.362**	

Table 1: **PSNR for DCT compression with optimal rates and different CCT at the same CR (compression ratio).**

We, therefore, propose the DCT as CCT - additional results are given in the next section.

3 Application: DCT Based Color Compression

In this section we propose a new algorithm for color image compression based on the 2D DCT block transform. Features of this algorithm include:

* It employs the DCT as color component transform
* It uses the optimal rates equation to design its quantization tables.

The stages of the algorithm are:

1. Apply the CCT (DCT) to the RGB color components of a given image to obtain new color components $C1, C2, C3$.
2. Apply the 2D block DCT to each color component Ci.
3. Quantize each subband of each color component independently using uniform scalar quantizers to achieve the optimal subband rates. This is the lossy stage of the compression.
4. Code the quantized DCT coefficients similarly to JPEG [18]: differential coding for the DC coefficients and zigzag scan, run-length coding and Huffman coding (combined with variable-length integer codes) for the AC coefficients. This stage is lossless coding.

Our rate-distortion model provides the expression for the optimal rates as derived in (76) taking into account both down-sampling of some of the components and the non-negativity of the rates.

However, several questions are still unanswered:

1. How do we determine the quantization step Δ_{bi} required to achieve the rate R_{bi} for subband b of color component i?
2. How do we measure the rates of the quantized subbands?

3.1 Determining the quantization steps

Consider a stochastic source X with distribution $f_X(x)$, uniformly quantized to $\hat{X}$ with (small) step size of Δ and then entropy coded. To calculate the entropy of $\hat{X}$, we divide the axis of the real numbers into equal intervals of size Δ, so that the q^{th} interval is displayed in Fig. 2. Note that q is an integer that can be both positive, negative or zero.

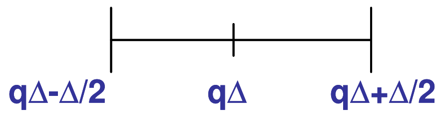


Figure 2: **Quantization intervals for uniform scalar quantization.**

Assuming that $f_X(x)$ is approximately constant inside the intervals, the entropy of $\hat{X}$ is:

$$\begin{aligned} H(\hat{X}) &= -\sum_{q=-\infty}^{\infty} P_q \log_2 P_q \cong -\sum_{q=-\infty}^{\infty} f_X(q\Delta)\Delta \log_2 \left(f_X(q\Delta)\Delta\right) \\ &\cong -\int_{-\infty}^{\infty} f_X(x) \log_2 \left(f_X(x)\right) dx - \int_{-\infty}^{\infty} f_X(x) \log_2 \Delta dx \\ &= h(X) - \log_2 \Delta. \end{aligned} \tag{80}$$

Here P_q is the probability that $\hat{X} = q$, that is, the probability that X falls into the q^{th} quantization interval. Also $h(X)$ is the entropy of the continuous variable X. Since we assume that $\hat{X}$ is efficiently entropy coded (e.g. by Huffman or arithmetic coding), its bit rate R is given by:

$$R \cong H(\hat{X}). \tag{81}$$

Note that even if $\hat{X}$ is not entropy coded and we wish to know the average error in the MSE sense caused by its quantization and reconstruction, we can use our rate-distortion model when the rate is measured by its entropy. In such a case we can write $R = H(\hat{X})$ without approximation.

Using (80) and (81)

$$\Delta = 2^{h(X)-R} \tag{82}$$

and when 2 quantization steps $\Delta 1$ and $\Delta 2$ are considered:

$$\frac{\Delta 1}{\Delta 2} = 2^{-(R1-R2)}. \tag{83}$$

Using (83) the following algorithm is proposed:

Optimal quantization steps algorithm:

1. Calculate the optimal rates R^*_{bi} (by (76)).

2. Set some initial quantization steps Δ_{bi} and calculate the resulting rates R_{bi}.

3. Update the quantization steps according to:

$$\Delta_{bi}^{new} = \Delta_{bi} 2^{-(R_{bi}^{*} - R_{bi})}$$

until the optimal rates R_{bi}^{*} are sufficiently close, i.e., $E\left(|R_{bi}^{*} - R_{bi}|\right) < \varepsilon$ for some small constant ε.

Based on practical considerations the following should be noted:

- The algorithm can be run for a limited number of iterations (typically 5-10).
- The rates R_{bi} are estimated using subband entropies.
- The initialization of Δ_{bi} and estimated entropies can be computed based on the Laplacian distribution assumption for the DCT coefficients.

3.2 Results of the algorithm

The algorithm was implemented and compared to the baseline JPEG [18]. First we present quantitative measures based on the MSE (PSNR) and WMSE and then some visual results. Similar to the PSNR (Peak Signal to Noise Ratio) definition:

$$PSNR = 10 \log_{10} \frac{255^2}{MSE},$$

we define the PSPNR (Peak Signal to Perceptible Noise Ratio):

$$PSPNR = 10 \log_{10} \frac{255^2}{WMSE}, \tag{84}$$

where $WMSE$ for each color component is calculated as:

$$WMSE = \sum_{b=0}^{B-1} \eta_b W_b G_b d_{b.} \tag{85}$$

Here d_b denotes the MSE of subband b, G_b it energy gain, η_b its sampling rate and W_b is its visual perception weight. For convenience we have taken the WMSE suggested in the JPEG2000 algorithm [15]. Therefore, the subbands in (85) are of the DWT (Discrete Wavelet Transform). We consider 256x256 or similar size images displayed on a screen as 12cm x 12cm size images. The visual weights suggested for a 5-levels DWT decomposition and a viewing distance of 46.875 cm (for our image size) are given in Table 2 based on the CSF (Contrast Sensitivity Function) of the human visual system for the YCbCr color space (the weights here squared are the W_b). HL stands for horizontal high-pass, vertical low-pass subbands, LH for horizontal low-pass, vertical high-pass subbands and HH for subbands high-pass filtered in both directions. The LL subband does not appear in the table, but should be assigned the highest weight of $W_b = 1$.

		Level				
		5	4	3	2	1
	HL	1.000000	1.000000	1.000000	0.998276	0.756353
Y	LH	1.000000	1.000000	1.000000	0.998276	0.756353
	HH	1.000000	1.000000	1.000000	0.996555	0.573057
	HL	0.883196	0.793487	0.650482	0.450739	0.230503
Cb	LH	0.883196	0.793487	0.650482	0.450739	0.230503
	HH	0.833582	0.712295	0.531700	0.309177	0.113786
	HL	0.910877	0.841032	0.725657	0.552901	0.336166
Cr	LH	0.910877	0.841032	0.725657	0.552901	0.336166
	HH	0.872378	0.776180	0.625103	0.418938	0.200507

Table 2: **Visual (CSF) weights for YCbCr.**

For convenience we use a single measure for the PSPNR: the average PSPNR for the three color components.

3.2.1 Quantitative results

The quantitative results are given in Table 3 for the same rate or compression ratio for both algorithms: the new algorithm and JPEG. It can be seen that the new algorithm (New Alg) improves JPEG performance by 1.24dB PSNR and 1.69dB PSPNR on average. It is clear that the PSNR is not a good measure when the human observer is considered, for example, when the difference in the PSNR is small. Then the PSPNR difference may be large as, for instance, in the Lena image, meaning that the small difference in the images is in details important to the human observer (such as low spatial frequencies of the image). On the other hand, larger differences in the PSNR, do not necessarily correspond to larger differences in the PSPNR when most of the error is concentrated in the high spatial frequencies, for example, to which the human eye is less sensitive.

3.2.2 Visual results

1. Lena (256x256) at 0.469 bpp (bits per pixel):

The results in terms of PSNR and PSPNR of JPEG and the new algorithm are given below.

		PSNR			**Average PSPNR**				
	Image	**New Alg**	**JPEG**	**Diff**	**New Alg**	**JPEG**	**Diff**	**Rate**	**CR**
	Lena	29.015	28.447	0.566	39.846	37.679	2.167	0.469	51.20

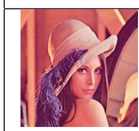
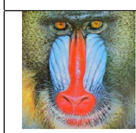
		PSNR			Average PSPNR			
	Image	**New Alg**	**JPEG**	**Diff**	**New Alg**	**JPEG**	**Diff**	**CR**
	Lena	30.019	29.785	0.234	39.073	37.559	1.514	41.87
	Peppers	30.015	28.640	1.375	37.848	35.787	2.061	31.59
	Baboon	30.010	26.370	3.639	38.957	36.061	2.896	14.43
	Fruit	29.977	29.302	0.675	39.185	37.042	2.143	38.72
	Girl	29.960	29.072	0.888	38.557	37.413	1.143	47.72
	House	29.975	29.213	0.762	38.962	37.970	0.991	45.20
	Tree	29.963	28.837	1.126	39.236	38.124	1.112	14.03
	Mean	**29.988**	**28.746**	**1.243**	**38.831**	**37.137**	**1.694**	

Table 3: **PSNR and PSPNR for the new algorithm and JPEG at the same compression ratio.**

The original image and the compressed versions for both algorithms: the new algorithm and JPEG are displayed in Fig. 3. Notice the color artifacts that JPEG introduces in the face area, which are much less visible in the image of the new algorithm.

2 Peppers (256x256) at 0.731 bpp:

		PSNR			**Average PSPNR**				
	Image	**New Alg**	**JPEG**	**Diff**	**New Alg**	**JPEG**	**Diff**	**Rate**	**CR**
	Peppers	29.995	28.273	1.722	37.844	35.429	2.416	0.731	32.82

For the Peppers image, the results are displayed in Fig. 4. Note the artifacts introduced by JPEG in the marked area. Fig. 5 zooms in on the area of interest and presents the mentioned artifacts more clearly. Those artifacts are almost absent form the new algorithm's image.

3 Baboon (256x256) at 0.287 bpp:

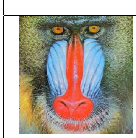
		PSNR			**Average PSPNR**				
	Image	**New Alg**	**JPEG**	**Diff**	**New Alg**	**JPEG**	**Diff**	**Rate(bpp)**	**CR**
	Baboon	22.028	21.469	0.558	32.065	29.693	2.372	0.287	83.67

Fig. 6 displays the result of the compression of the baboon image at the rate of 0.287bpp. Note the red and green color artifacts for the JPEG algorithm in the marked area which are absent in the new algorithm's image.

4 Summary

We have introduced a Rate-Distortion model for color image compression using subband transform coders. This tool provides prediction of the distortion (MSE) of such coders for given subband rates and a color components transform. Based on the model, a target function for an optimal CCT and optimal rates allocation are derived. The Generalized KLT has been introduced as an approximated solution to the target function, and an iterative algorithm for its calculation has been proposed. The performance of various color components transforms for image compression has been studied and it has been shown that the DCT can be used as a sub-optimal CCT, close to the optimal adaptive CCT and superior to the commonly used RGB to YUV transform or KLT. This solution also has an advantage over adaptive CCTs, since it is a fixed straightforward transform. Referring to optimal subbands rates, we have also considered and analyzed the case of down-sampling color components, as well as additional constraints for non-negativity of the rates. An algorithm for designing optimal quantization tables of subband coefficients has been introduced and implemented in the context of a new compression algorithm for color images. Both quantitative (MSE and WMSE) and visual results have been presented, showing that the proposed compression algorithm outperforms baseline JPEG. Our conclusion is that in addition to the theoretical aspects of the new Rate-Distortion model it can also serve as a tool for improving color image compression systems compared to presently available algorithms.

Acknowledgement

This research was supported in part by the HASSIP Research Program HPRN-CT-2002-00285 of the European Commission, and by the Ollendorff Minerva Center. Minerva is funded through the BMBF.

Appendices

A Minimization of the Lagrangian of (45)

Taking the gradient of $L\left(\{R_{bi}\}, \mathbf{M}, \lambda\right)$ according to R_{bi} equal to 0, we get:

$$\lambda = \frac{a}{3} G_b \sigma_{bi}^2 \varepsilon_i^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} e^{-aR_{bi}} \tag{86}$$

Therefore, the rate of subband b of component i can be expressed via its 0 subband rate, for example as:

$$R_{bi} = R_{0i} + \frac{1}{a} \ln\left(\frac{G_b \sigma_{bi}^2}{G_0 \sigma_{0i}^2}\right). \tag{87}$$

using the equality:

$$\lambda = \frac{a}{3} G_b \sigma_{bi}^2 \varepsilon_i^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} e^{-aR_{bi}} = \frac{a}{3} G_0 \sigma_{0i}^2 \varepsilon_i^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} e^{-aR_{0i}}. \tag{88}$$

Similarly by (86) the rates of the 0 subbands of different color components i and j are connected according to:

$$R_{0i} = R_{0j} + \frac{1}{a} \ln \left(\frac{\varepsilon_i^2 \sigma_{oi}^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii}}{\varepsilon_j^2 \sigma_{0j}^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{jj}} \right) \tag{89}$$

Expressing all the rates R_{bi} via R_{01} and solving the constraint $\sum_{i=1}^{3} \sum_{b=0}^{B-1} \eta_b R_{bi} = R$ gives us the solution for R_{01} and therefore by (87) and (89) for all the rates R_{bi}:

$$R_{bi} = \frac{R}{3} + \frac{1}{a} ln \left(\frac{\varepsilon_i^2 G_b \sigma_{bi}^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii}}{\left(\prod_{k=1}^{3} GM_k\right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} \varepsilon_k^2\right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}\right)^{\frac{1}{3}}} \right), \tag{90}$$

where GM_k is defined in (47). Substituting the rates in (42) results in:

$$MSE = e^{-\frac{aR}{3}} \left(\prod_{k=1}^{3} \varepsilon_k^2 \right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} GM_k \right)^{\frac{1}{3}} \left(\prod_{k=1}^{3} \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk} \right)^{\frac{1}{3}} \tag{91}$$

and minimizing this expression for $\mathbf{M}$ is the same as minimizing (48).

B Minimization of $log\left(g(\mathbf{M})\right)$

Starting from $log\left(g(\mathbf{M})\right)$ of (58) we get:

$$\nabla_{\mathbf{m}_1} \log(g(\mathbf{M})) = \frac{-6\mathbf{x}_1}{\det(\mathbf{M}\mathbf{M}^T)} + \frac{2\left[\left(\mathbf{m}_3^T\mathbf{m}_3\right)\mathbf{m}_1 - \left(\mathbf{m}_1^T\mathbf{m}_3\right)\mathbf{m}_3\right]}{\left(\mathbf{m}_1^T\mathbf{m}_1\right)\left(\mathbf{m}_3^T\mathbf{m}_3\right) - \left(\mathbf{m}_1^T\mathbf{m}_3\right)^2} + \frac{2\left[\left(\mathbf{m}_2^T\mathbf{m}_2\right)\mathbf{m}_1 - \left(\mathbf{m}_1^T\mathbf{m}_2\right)\mathbf{m}_2\right]}{\left(\mathbf{m}_1^T\mathbf{m}_1\right)\left(\mathbf{m}_2^T\mathbf{m}_2\right) - \left(\mathbf{m}_1^T\mathbf{m}_2\right)^2} + \frac{2\Lambda\mathbf{m_1}}{\mathbf{m_1}^T\Lambda\mathbf{m_1}}, \tag{92}$$

where

$$\begin{aligned}\mathbf{x}_1 \quad = \quad & \left[\left(\mathbf{m}_2^T\mathbf{m}_2\right)\left(\mathbf{m}_3^T\mathbf{m}_3\right) - \left(\mathbf{m}_2^T\mathbf{m}_3\right)^2\right]\mathbf{m}_1 + \left[\left(\mathbf{m}_1^T\mathbf{m}_3\right)\left(\mathbf{m}_2^T\mathbf{m}_3\right) - \left(\mathbf{m}_3^T\mathbf{m}_3\right)\left(\mathbf{m}_1^T\mathbf{m}_2\right)\right]\mathbf{m}_2 \\ & + \left[\left(\mathbf{m}_1^T\mathbf{m}_2\right)\left(\mathbf{m}_2^T\mathbf{m}_3\right) - \left(\mathbf{m}_2^T\mathbf{m}_2\right)\left(\mathbf{m}_1^T\mathbf{m}_3\right)\right]\mathbf{m}_3.\end{aligned} \tag{93}$$

The sum of first 2 terms in (92) is the gradient of $\sum_{k=1}^{3} log\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}$, while the 3^{rd} term is the gradient of $\sum_{k=1}^{3} log\left(\mathbf{m_k}^T\Lambda\mathbf{m_k}\right)$ according to $\mathbf{m}_1$. If $\mathbf{M}\mathbf{M}^T = \mathbf{I}$, then:

$$\nabla_{\mathbf{m}_1} \log(g(\mathbf{M})) = -2\mathbf{m}_1 + \frac{2\Lambda\mathbf{m_1}}{\mathbf{m_1}^T\Lambda\mathbf{m_1}}. \tag{94}$$

Similar equations can be derived for $\mathbf{m}_2$ and $\mathbf{m}_3$.

C Minimization of $log\left(f(\mathbf{M})\right)$

It is clear that the $\sum_{k=1}^{3} log\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{kk}$ part of $logf(\mathbf{M})$ in (66) is the same as in (58) which minimization is discussed in Appendix B. Therefore we will receive the same expression for its gradient according to $\mathbf{m}_1$ as in (92) and under the assumption $\mathbf{M}\mathbf{M}^T = \mathbf{I}$ it will reduce to $-2\mathbf{m}_1$ as in (94). The gradient of the second term in (66): $\sum_{k=1}^{3}\sum_{b=0}^{B-1}\eta_b log\left(\mathbf{m_k}^T\mathbf{\Lambda_b}\mathbf{m_k}\right)$ is:

$$\nabla_{\mathbf{m}_1}\left(\sum_{k=1}^{3}\sum_{b=0}^{B-1}\eta_b log\left(\mathbf{m_k}^T\mathbf{\Lambda_b}\mathbf{m_k}\right)\right) = 2\sum_{b=0}^{B-1}\eta_b\frac{\Lambda_b\mathbf{m_1}}{\mathbf{m_1}^T\Lambda_b\mathbf{m_1}}, \tag{95}$$

and therefore the total gradient equation becomes:

$$-2\mathbf{m}_1 + 2\sum_{b=0}^{B-1}\eta_b\frac{\Lambda_b\mathbf{m_1}}{\mathbf{m_1}^T\Lambda_b\mathbf{m_1}} = 0. \tag{96}$$

Similar equations are received for $\mathbf{m}_2$ and $\mathbf{m}_3$.

D Deriving the optimal rates with down-sampling

Similarly to the mathematical analysis of Appendix A we take the gradient of $L\left(\{R_{bi}\},\mathbf{M},\lambda\right)$ of (71) equalled to 0 to receive the equations:

$$\lambda = \frac{a}{3\alpha_i}G_b\sigma_{bi}^2\varepsilon_i^2\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii}e^{-aR_{bi}}. \tag{97}$$

We then repeat the same process of expressing the rates R_{bi} via R_{01}as in (87) and (89). While the first equation remains the same, the second one changes to:

$$R_{0i} = R_{0j} + \frac{1}{a}\ln\left(\frac{\frac{\varepsilon_i^2\sigma_{oi}^2\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii}}{\alpha_i}}{\frac{\varepsilon_j^2\sigma_{0j}^2\left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{jj}}{\alpha_j}}\right). \tag{98}$$

where j is taken as 1. Solving the constraint equation

$$\sum_{i=1}^{3}\alpha_i\sum_{b=0}^{B-1}\eta_b R_{bi} = R \tag{99}$$

for R_{01} allows us to express all the rates R_{bi} as in (72).

E Deriving the optimal rates with $R_{bi} \geq 0$ constraints

We repeat here the operations in Appendix A, i.e. we take the gradient of $L\left(\{R_{bi}\}, \mathbf{M}, \lambda, \{\mu_{bi}\}\right)$ of (73) equal to 0 to receive:

$$\lambda = \frac{a}{3\alpha_i} G_b \sigma_{bi}^2 \varepsilon_i^2 \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} e^{-aR_{bi}} + \frac{\mu_{bi}}{\alpha_i \eta_b}. \tag{100}$$

However, now we have unknown variables μ_{bi} in the equations. The solution we propose is to assume that the 0 (DC) subbands are active, i.e. $R_{0i} > 0$ (which is a logical assumption since a large part of the energy is concentrated in these subbands). Therefore, $\mu_{0i} = 0$ and the equation for the 0 subbands (and any other active subband) is the same as (97). For some other subband b of color component i: if it is active too, then its rate R_{bi} can be expressed through R_{0i} as in (87). If it is not active, then $R_{bi} = 0$ by definition and (100) can be used to express its Lagrange multiplier through R_{0i}:

$$\mu_{bi} = \frac{a}{3} \eta_b \left((\mathbf{M}\mathbf{M}^T)^{-1}\right)_{ii} \varepsilon_i^2 \left(G_0 \sigma_{0i}^2 e^{-aR_{0i}} - G_b \sigma_{bi}^2\right). \tag{101}$$

The 0 subbands rates are expressed through R_{01}, for example, according to (98) and then all the rates are substituted in the total rate constraint, which is slightly rewritten:

$$\sum_{i=1}^{3} \alpha_i \sum_{b=0}^{B-1} \eta_b R_{bi} = \sum_{i=1}^{3} \alpha_i \sum_{b \in Act_i} \eta_b R_{bi} = R,$$

i.e. we sum up only the non zero rates of the active subbands. The solution for R_{01} and therefore for all the other non zero rates is given by (76).

References

[1] Gershikov E. and Porat M., Does Decorrelation Really Improve Color Image Compression?, *the International Conference on Systems Theory and Scientific Computation (ISTASC05)*, Malta, 2005.

[2] Goffman-Vinopal L. and Porat M., Color image compression using inter-color correlation, *IEEE International Conference on Image Processing*, Vol.2, pp. II-353 - II-356, 2002.

[3] Han S.E., Tao B., Cooper T. and Tastle I., Comparison between Different Color Transformations for the JPEG 2000, *PICS 2000*, pp. 259-263, 2000.

[4] JPEG 2000 Part I: Final Draft International Standard (ISO/IECFDIS15444-1), ISO/IEC JTC1/SC29/WG1 N1855, Aug. 2000.

[5] Kotera H. and Kanamori K., A Novel Coding Algorithm for Representing Full Color Images by a Single Color Image, *J. Imaging Technology*, Vol. 16, pp. 142-152, Aug. 1990.

[6] Kouassi R.K., Devaux J.C., Gouton P. and Paindavoine M., Application of the Karhunen-Loeve transform for natural color images analysis, *IEEE Conference on Signals, Systems & Computers*, Vol.2, pp. 1740-1744, 1997.

[7] Kouassi R.K., Gouton P. and Paindavoine M., Approximation of the Karhunen-Loève transformation and its application to colour images, *Signal Processing: Image Communication*, Vol. 16, pp. 541-551, 2001.

[8] Limb J. O. and Rubinstein C.B., Statistical Dependence Between Components of A Differentially Quantized Color Signal, *IEEE Trans. on Communications,* Vol. Com-20, pp. 890-899, Oct. 1971.

[9] Pinhasov E. D., Shimkin N. and Zeevi Y. Y., Optimal Usage of Color for Disparity Estimation in Stereo Vision, *The European Signal Processing Conference (USIPCO2005)*, Sept. 2005.

[10] Rabbani M. and Joshi R., An overview of the JPEG2000 still image compression standard, *Signal Processing: Image Communication*, Vol. 17, No. 1, pp 3-48, 2002.

[11] Rao K. R. and Yip P., Discrete cosine transform: algorithms, advantages, applications, Academic Press, 1990.

[12] Roterman Y. and Porat M., Progressive image coding using regional color correlation, *the 4th EURASIP Conference,* Vol. 1, pp. 65-70, 2003.

[13] Satt A. and Malah D., Design of Uniform DFT Filter Banks Optimized for Subband Coding of Speech, *IEEE Trans. on Acoustics, Speech and signal Processing*, Vol. 37, No. 11, pp. 1672-1679, Nov. 1989.

[14] Shapiro J. M., Embedded Image Coding Using Zerotrees of Wavelet Coefficients, *IEEE Trans. on Signal Processing*, Vol. 41, No. 12, pp 3345-3462, 1991.

[15] Taubman D. S. and Marcellin M. W., JPEG2000: image compression, fundamentals, standards and practice, Kluwer Academic Publishers, 2002.

[16] Tu C. and Tran T. D., Context-Based Entropy Coding of Block Transform Coefficients for Image Compression, *IEEE Trans. on Image Processing*, Vol. 11, No. 11, pp 1271-1283, 2002.

[17] Unser M., On the Approximation of the Discrete Karhunen-Loeve Transform for Stationary Processes, *Signal Processing*, Vol. 7, 231-249, 1984.

[18] Wallace G. K., The JPEG Still Picture Compression Standard, *IEEE Transactions on Consumer Electronics* , Vol. 38, pp. xviii-xxxiv, 1992.

[19] Wolf S. G., Ginosar R. and Zeevi Y. Y., Spatio-chromatic Model for Colour Image Processing, *IEEE Proc. of the IAPR*, Vol. 1, pp. 599-601, 1994.

[20] Yamaguchi H., Efficient Encoding of Colored Pictures in R, G, B Components, *Trans. on Communications,* Vol. 32, pp. 1201-1209, Nov. 1984.

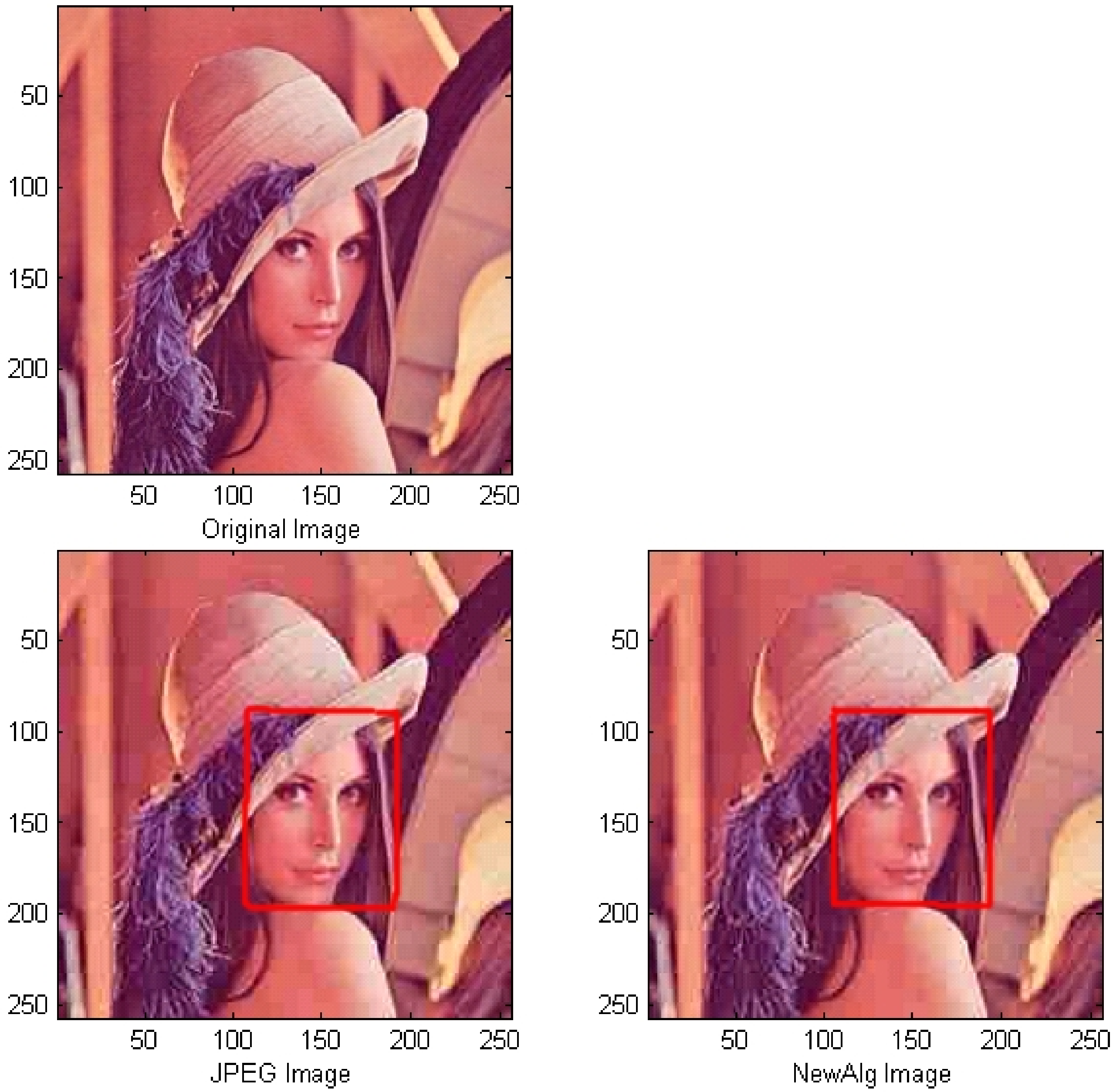


Figure 3: **Lena at 0.469 bpp: original (top), compressed by JPEG (bottom left) and compressed by the new algorithm (bottom right).**

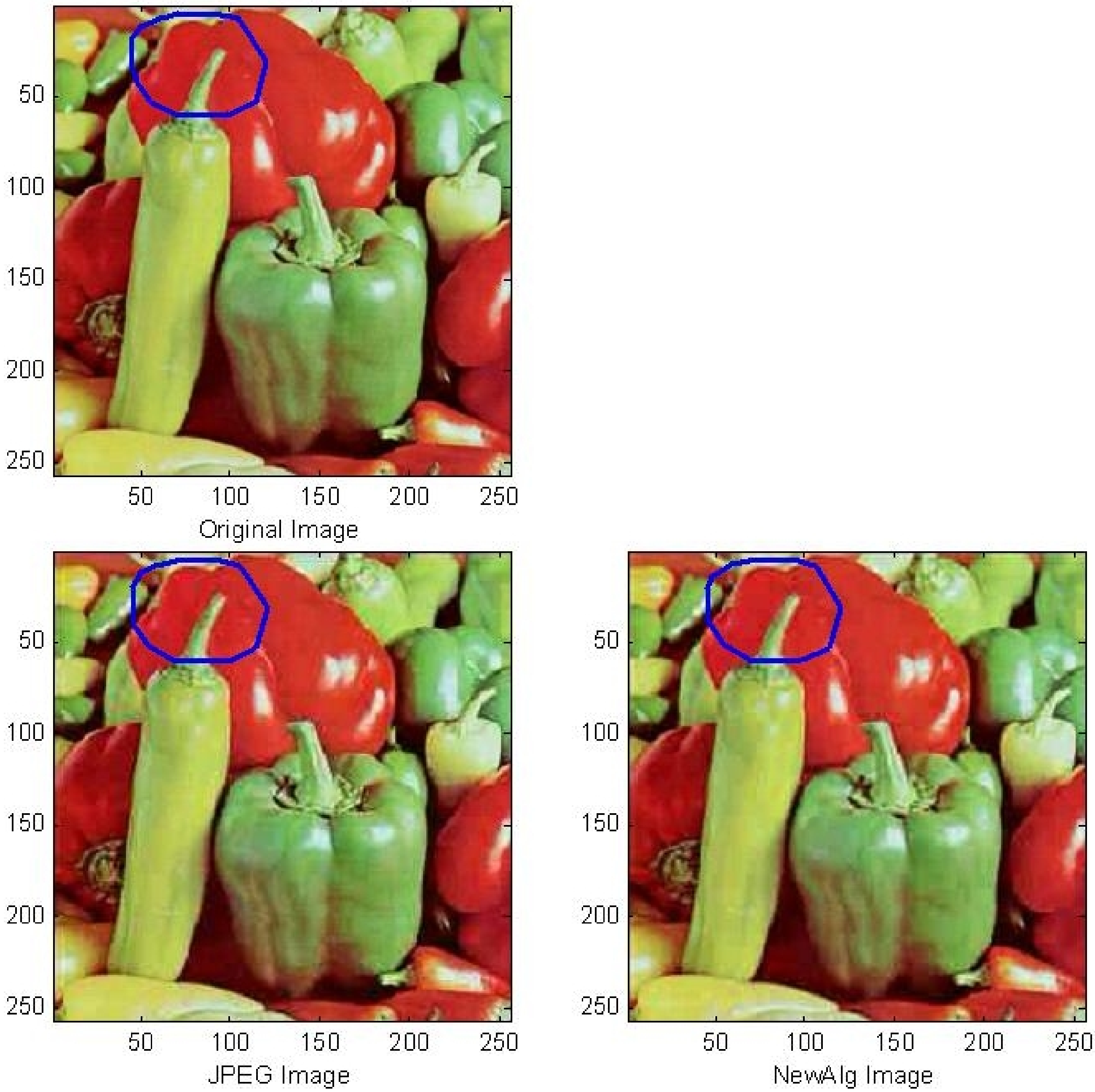


Figure 4: **Peppers at 0.731 bpp: original (top), compressed by JPEG (bottom left) and compressed by the new algorithm (bottom right).**

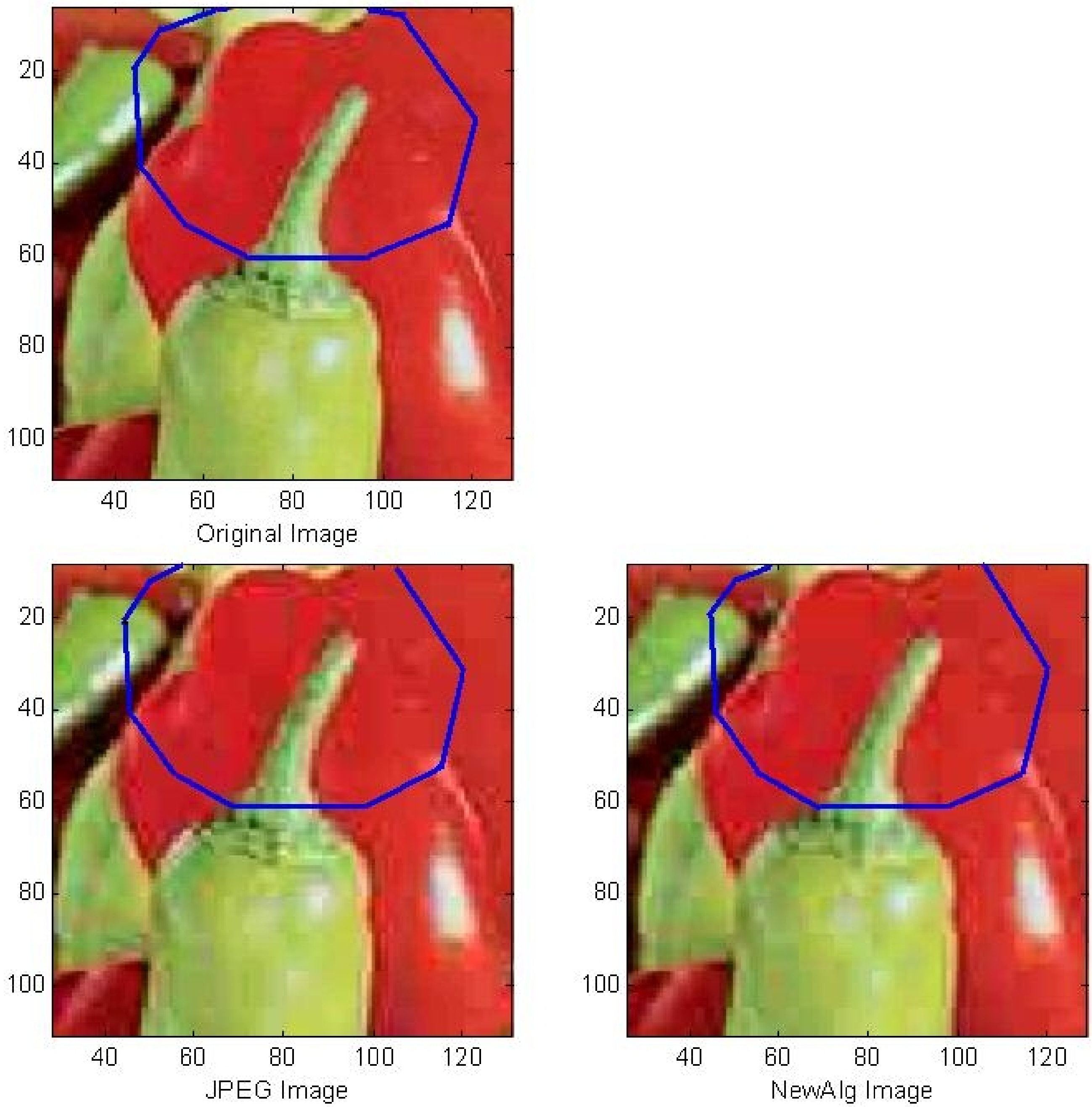


Figure 5: **Peppers at 0.731 bpp zoomed: original (top), compressed by JPEG (bottom left) and compressed by the new algorithm (bottom right).**

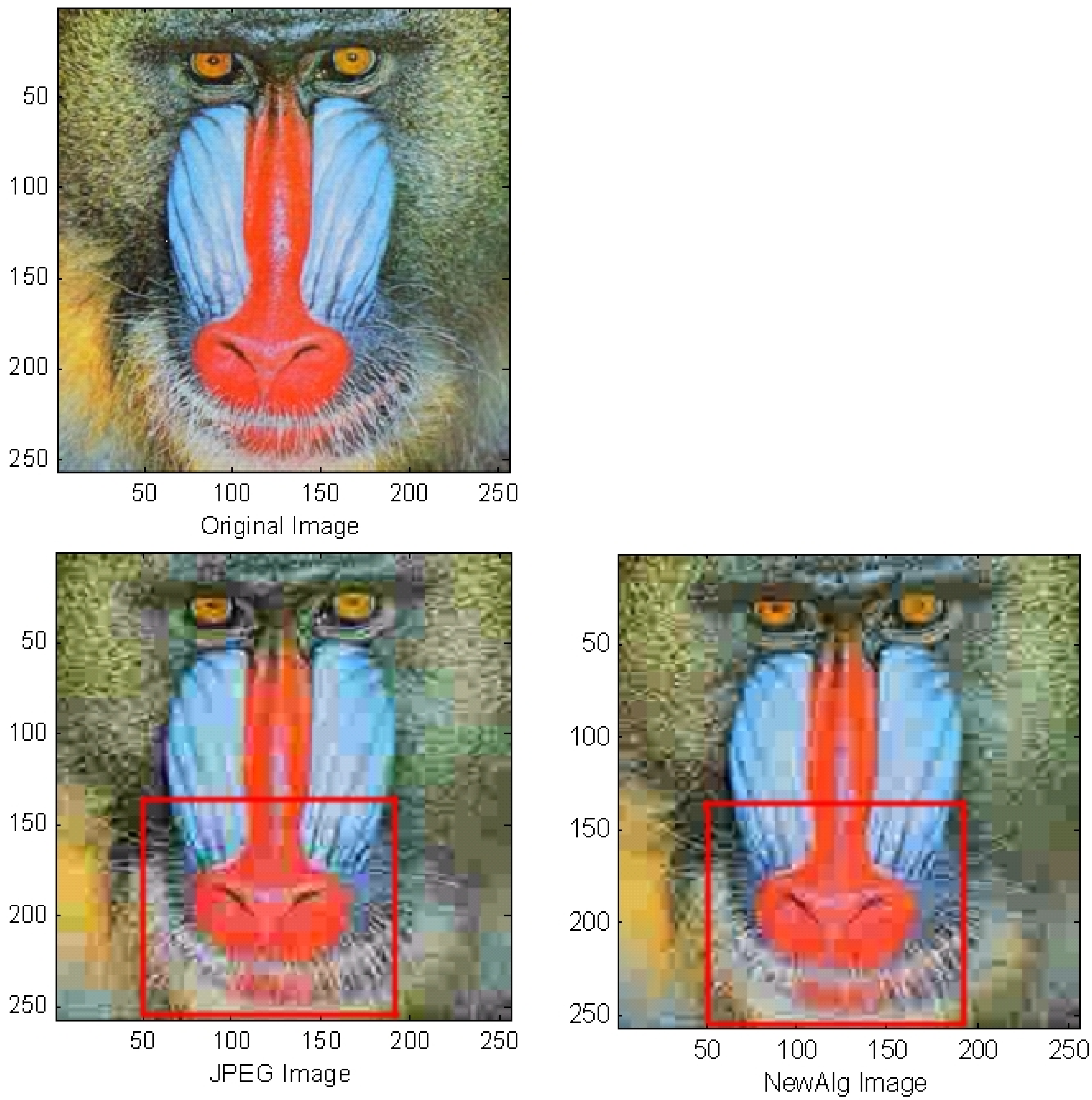


Figure 6: **Baboon at 0.287 bpp: original (top), compressed by JPEG (bottom left) and compressed by the new algorithm (bottom right).**