



TITLE: Report on CE V1 (Exploitation of Visual Masking through Control of Individual Block Contributions)

SOURCE: Wenjun Zeng, (zengw@sharplabs.com)
Scott Daly, (daly@sharplabs.com)
Shawmin Lei, (shawmin@sharplabs.com)
Sharp Labs. of America, Inc., 5750 NW Pacific Rim Blvd, Camas, WA 98607

PROJECT: JPEG 2000

STATUS: Information

REQUESTED ACTION: Voted to be accepted as an option in WD

DISTRIBUTION: WG 1 Mailing List

Contact:

ISO/IEC JTC 1/SC 29/WG 1 Convener - Dr. Daniel T. Lee
Hewlett-Packard Company, 11000 Wolfe Road, MS42U0, Cupertino, California 95014, USA
Tel: +1 408 447 4160, Fax: +1 408 447 2842, E-mail: Daniel_Lee@hp.com

Core Experiment Description/Results Summary
Experiment Name: Exploitation of Visual Masking through Control of
Individual Block Contributions
Sub-Group: Visual Number: V1

Description:

Core experiment partners	UNSW, EPFL, Sharp Laboratories of America, HP
Core experiment objective	<p>Modify the distortion measure used for post-compression R-D optimization in the VM, so as to take into account the visual masking effect. This represents a way of exploiting masking without any impact upon the decoder or the bit-stream syntax, since distortion estimates and control of the contribution of different blocks to layers in the final bit-stream are the responsibility of the encoder alone.</p> <p>The experiment will compare this approach with that proposed by Sharp and Kodak and incorporated into VM4, which is based on modifications to both the encoder and the decoder, as well as the bit-stream syntax. A third approach which combines the strength of both approaches will also be described and tested</p>
JPEG 2000 Requirement Focus	Visual quality
What will change from Verification Model 4.0	Formation of distortion estimates in the encoder
Key Benefit of change	Improved visual quality, particularly on large images with diverse content
Related Experiments	Visual Weighting, Visual Masking
Expected Memory Decrease/increase	No change
Expected Complexity Decrease/increase	Negligible change at encoder; none at decoder
Other expected results	None

Core experiment detail description	The distortion metric used for Post-Compression R-D Optimization will be modified to reflect masking properties of the Human Visual System. Specifically, within each code-block, the distortion will be computed by summing the weighted MSE contribution from the distortion at each sample in the block, where the weight itself is a function, which varies from sample to sample in accordance with local activity surrounding the relevant sample. Experiments will be performed to assess the visual quality improvement available under monitor display conditions and also with prints. In the case of printing, visual CSF and printer MTF effects will be included in the testing. The benefit of combining this approach with that proposed by Sharp will also be investigated.
------------------------------------	---

Results:

Web Document Number: ___ WG1N1303 ___

VM Mode Used in Experiment	7x9 filter
-----------------------------------	------------

What has changed from Verification Model 4 (provide level of integration)	Formation of distortion estimates in the encoder
--	--

Was this experiment performed on the VM or in a testbed	VM4.1
--	-------

Key Benefit of change	Improved image visual quality at given bit rate
Key Cost of change	Negligible complexity increase

Other performance results	
----------------------------------	--

1. Introduction

Visual masking is a perceptual phenomenon where artifacts are locally masked by the image acting as a background signal. For example, in the wavelet transform domain, a larger coefficient can tolerate a larger distortion since the larger coefficient results in a large background signal that masks the visual distortion. One way to exploit this masking effect in image compression is to let the coefficients go through a nonlinear transducer function such as a power function, before a uniform quantization is applied [2]. At the decoder, the inverse process is applied between dequantization and inverse wavelet transform. This way of making use of visual masking effect of the human visual systems (HVS) has been implemented in VM4.0 and VM4.1 software [1]. This approach will be referred to as Approach 1 in the sequel.

Another way of exploiting visual masking in the context of JPEG 2000 VM is through control of individual code-block contribution as proposed in [3]. In this approach, the embedded coding of each code-block is performed without considering visual masking effect. However, in the post-compression rate-distortion optimization process, the distortion metric takes into account the visual masking effect. More specifically, the distortion of each coefficient is weighted by a visual masking factor that is in general a function of the neighboring coefficients, i.e., it treats each coefficient value, V_i , as though it were equal to V'_i (from the perspective of distortion estimates), where

$$V'_i = V_i / M_i \quad (1)$$

and the masking strength function is

$$M_i = A * \sum_{\{k \text{ near } i\}} \text{sqrt}(|V_k|) \quad (2)$$

with A being a normalization factor. We will refer to this approach as Approach 2.

The key difference between the above mentioned two approaches is that the former one is a point-wise compensation, whereas the latter one uses spatially extensive masking to adjust the distortion measure associated with all features within a code-block. They have their strength and weakness: the latter one adjusts only the distortion metric and can be spatially extensive which are pros from an implementation and theoretical point of view; on the other hand, it can only adjust the truncation points of each code-block, which is a spatially much coarser adjustment than the sample-by-sample compensation offered by the former approach. Within a code-block which is usually no less than 32x32, the bit stream is ordered in a way that does not take into account any visual masking effect.

It is worth pointing out that there are several problems in Approach 1. The first is that it assumes the wavelet band structure and filters are a good match to the visual system's underlying channels. Although, the wavelet structure is a much better model of the visual system than the DCT, it has a problem with the diagonal band due to the Cartesian separable approach. In the visual system, frequencies at 45 degrees orientation have very little masking effect on those at -45 degrees, but the diagonal band in the wavelet has no way of distinguishing the two. This

gives rise to artifacts perpendicular to the diagonal edge. Another diagonal related problem is that the horizontal and vertical bands encroach on the diagonal signals at multiples of the Nyquist/ 2^L , i.e., 0.5, 0.25, 0.125 cy/pix. This means that diagonal edges cause high values in the H and V bands, which causes high quantization for the bands at the edges, giving rise to horizontal vertical artifacts along slanted edges. This problem can't be fixed by merely turning off masking in the diagonal band, but may be helped by a different choice of filters whose Cartesian product has low energy near 45 degrees (for the H and V bands). Another issue with the wavelet filter shape is that the visual system's spatial frequency receptive fields (of which the wavelets are attempting to model) appear to be much broader than those of the current implementation default. To model them more accurately would unfortunately require larger kernels. These overall problems lead to overmasking at diagonal edges.

The second problem with Approach 1 relates to phase. The visual system has a phase uncertainty of about 90 degrees phase [4], meaning that the masking effect will spread over two neighboring coefficients (in 1D), for the highest frequency of a band, and over 4 coefficients (1D) for the lowest frequency of the band. As a result, the masking that occurs underestimates the masking at the zero crossings in a band.

In Section 2, we describe a third approach that combines the strength of both approaches, and discuss two ways to implement this approach. We conducted some testing to compare the visual performance of these different approaches. Some observations based on the experiments are summarized in Section 3.

2. An approach that exploits both self contrast-masking and neighborhood masking

Approach 1 described above applies a non-linear transducer function to the coefficients prior to uniform quantization. It essentially protects low amplitude coefficients, whereas the distortion introduced by more coarsely quantizing high amplitude coefficients is well masked by the coefficients themselves. Approach 2 is essentially a region/block based adaptive bit allocation scheme. It exploits the masking capability of a complex region, therefore allocating more bits to smooth regions or regions with simple edge structures. The masking weighting factor for each coefficient in this approach is based on a neighborhood average. An advantage of this strategy is its ability to distinguish between large amplitude coefficients that lie in a region of simple edge structure and those in a complex region. This feature will assure the good visual quality of simple edges in a smooth background, which is critical to the overall perceived visual quality. This feature is not exploited in Approach 1 because this method assumes the wavelet transform is more like the visual system (in terms of band filter shapes and oriented structure) than it really is. In the following, we describe an improved approach that combines the strength of both Approach 1 and Approach 2.

We treat visual masking as a combination of two separate processes, i.e., self contrast masking and neighborhood masking. The visual masking effect is therefore exploited for compression purpose in two steps. The first step is to apply a coefficient-wise non-linear transducer function $f(\cdot)$ such as a power function to the original coefficient x_i , i.e., $x_i \rightarrow y_i = f(x_i)$. This step assumes each signal with which a coefficient is associated is lying on a common flat background. Under

this assumption, $\{y_i\}$ are perceptually uniform. In a real image, however, this is usually not the case. Each signal is superimposed on other neighboring signals. There is some masking effect contributed from neighboring signals due to the phase uncertainty, receptive field sizes, as well as possible longer range effects. To further exploit this neighborhood masking effect, the second step normalizes y_i by a masking weighting factor w_i which is a function of the amplitudes of the neighboring signals, i.e., $y_i \rightarrow z_i = y_i / w_i$, where w_i is a function $g(\cdot)$ of the neighboring signals denoted in a vector form as $N_i(\{y_k\})$, i.e., $w_i = g(N_i(\{y_k\}))$.

From an implementation point of view, the first step that is a point-wise operation has to be implemented at the encoder, and the inverse process has to be implemented at the decoder. The second step can be implemented in two ways under the context of JPEG2000 VM. The first one is to use a similar approach as Approach 2 where the masking weighting is incorporated into the distortion metric for each code-block. Note the difference between this implementation and Approach 2 is that the uniform quantization is applied to y_i , as opposed to x_i in Approach 2. Therefore both self-contrast-masking and neighborhood masking are exploited. A particular implementation is to simply concatenate Approach 1 and Approach 2. For example, if a power function $y_i = |x_i|^\alpha$ is used in Step 1, then a direct concatenation of Approaches 1 and 2 results in

$$z_i = y_i / (\sum_{\{k \text{ near } i\}} |y_k|^\rho / \|\phi_i\|) = |x_i|^\alpha / (\sum_{\{k \text{ near } i\}} |x_k|^{\alpha\rho} / \|\phi_i\|) \quad (3)$$

where $\|\phi_i\|$ denotes the size of the neighborhood. For comparison purpose, note that Approach 2 implements $y_i = x_i / (\sum_{\{k \text{ near } i\}} |x_k|^\rho / \|\phi_i\|)$.

In the above implementation, the second step is used to adjust the truncation points of each code-block, which is again a coarse adjustment. Another way of implementation is to do

$$z_i = y_i / w_i = f(x_i) / g(N_i(\{f(x_k)\})) \quad (4)$$

at the encoder before uniform quantization, and do the inverse process at the decoder. To make sure the inverse process is feasible, the neighborhood has to be causal in the sense that each coefficient x_k in this neighborhood has to be recovered before the current one x_i . An example is to use the non-linear transform

$$z_i = |x_i|^\alpha / (1 + a \sum_{\{k \text{ near } i\}} |x_k|^\beta / \|\phi_i\|) \quad (5)$$

where a is a normalization factor and the neighborhood contains coefficients that lie within an $N \times N$ window centered at the current coefficient and also appear earlier than the current coefficient in the raster scan order. The neighborhood does not include the current coefficient itself in order to have an explicit solution for the inverse process. Note that a potential problem with this approach is that the decoder only has the quantized version of the coefficients x_k . Therefore the inverse process may only be an approximation of the ideal inverse process. This may cause some problem especially at low bit rates where coefficients are coarsely quantized. Note that, unfortunately, the encoder can not do the non-linear transformation based on the exact final compressed/quantized version of the coefficient x_k because the nonlinear transform is performed prior to scalable compression and the decoder can have any bitstream that has a lower rate than the final rate. Nevertheless, the discrepancy can be reduced or eliminated by using only

the *same* (potentially very coarsely quantized) coefficients that are relatively large to calculate the masking weighting factor w_i at both the encoder and the decoder, with the compromise of a coarser granularity of w_i .

3. Observations based on visual testing of different approaches

We use VM4.1 based extended code that has the options of -Cvis which implements Approach 2, and -Qmask which implements Approach 1. In the original implementation of -Qmask in VM4.0 and VM4.1, the coefficients are first normalized by the csf weights, then are subject to the non-linear transducer function prior to uniform quantization. The csf normalization is applied prior to non-linear transducer function to ensure calibration of the transducer function so that the facilitation region of non-uniform quantization corresponds to the facilitation region of the observer, as previous study suggests. The non-linear function is realized using a table, and the table is encoded into the bitstream. This is to reduce the complexity and to accommodate a large variety of different transducer functions. As a result, the csf weights are also encoded into the bitstream so that the decoder can do the inverse csf normalization. If the non-linear transducer function is a power function, then it is possible to interchange the non-linear transducer function step and the csf normalization step so that the csf weights need not be encoded into the bitstream.

To compare the performance of different visual masking implementations, the entries of the csf weighting table are set to all one, i.e., no csf weighting is taken into account. This is reasonable for monitor display condition where the viewers are allowed to zoom in to see the details. The code is also modified so that the non-linear power function is implemented by calling the power function of the system library. This is to facilitate the adjustment of the power factor.

The test images include lenna (512x512), woman, café and bike. For the third approach, we test two implementations. One is a direct concatenation of Approach 1 and Approach 2 (i.e., both -Qmask option and -Cvis option are on). The other one is to implement Eq. (5). At the time of this writing, only very preliminary results are available for the second implementation, and we are able to achieve at least similar visual quality as Approach 2. However, in order to make a conclusion about the feasibility and effectiveness of this implementation, fine tuning and further investigation is needed. Therefore, the following observations about Approach 3 are made based on the results of direct concatenation of Approach 1 and Approach 2. The α of the power function is chosen as 0.8 0.7 or 0.6. These three values in general work fine, with 0.6 sometimes being a little bit too strong on sharp edges.

Image: Lenna

At higher bit rates (0.75-1 bpp), Approach 1 appears to have better visual quality than Approach 2, especially on the face where more fine texture of the face skin is observed. Approach 3 gives similar quality as Approach 1 at higher bit rates, but has better visual quality than Approach 1 at lower bit rates (0.25-0.5 bpp). Using α of 0.6 gives a little bit better quality for face and hat at higher bit rates than 0.8, but shows some artifact on sharp edges, especially at 0.25 bpp.

Explanation: The image size is small. Approach 2 does not work well because of large-block-based bit allocation. No masking is exploited within blocks.

Image: Woman

Approach 2 significantly outperforms Approach 1 for most bit rates, especially on sharp edges, sometimes even on the face texture at lower bit rate (0.25 bpp). However, Approach 3 outperforms Approach 2 at lower bit rates (0.25-0.75 bpp). There are more skin textures on the palms and face. The visual quality is also better at boundary areas that lie between smooth and complex areas. The main weakness with Approach 2 occurs when a code-block straddles a smooth area next to a textured area such that the block is predominantly complex. A high degree of masking is then assumed for the whole block, resulting in high quantization for coefficients near the edge, causing ringing artifacts into the smooth area, where they become visible because the visual system's masking effect is more localized than the block segments.

Explanation: The image is large. There are a lot of highly complex areas that can be exploited by adaptive bit allocation among different code-blocks. Also, there are substantially more blocks that are entirely consisting of complex areas (textures), so as more masking is invoked over a larger percentage of the image.

Image: Café and Bike

Approach 2 significantly outperforms Approach 1 for most bit rates, especially on background texture and sharp edges. Approach 3 with α of 0.8 gives similar visual quality as Approach 2. A value of 0.6 for α is a little bit too strong to be used in Approach 3.

Explanation: The image is highly complex. Low amplitude texture areas are small and are not the focus of visual interest. Therefore, preserving the low amplitude texture does not give an overall good perception. On the other hand, the edges of the objects are perceptually important and are well preserved by the block-based adaptive bit allocation scheme of Approach 2.

Another observation made based on the above testing is that Approach 2 in fact takes into account some of the frequency weighting effect. Since blocks that consist entirely of textures (which invoke the higher quantization of masking) are much more predominant in the high frequency bands, more errors are introduced in these bands. It penalizes the very complex area that is composed of a lot of high frequency components. As a result, it consistently provides good quality over different bit rates.

4. A note on viewing distances and frequency weighting

In these experiments we removed any consequences of frequency weighting by setting those table to 1.0 for all frequencies. Since the masking approaches in general allocate more quantization steps to the low amplitude coefficients and away from the high, they can become problematic if the CSF is not considered. This is because the minimum quantization level of the highest frequency band is higher than other bands (even at fairly close viewing distances). By

not taking into account the CSF and using masking approaches, one can allocate bits to the low amplitude coefficients in the highest frequency band which are below visible threshold due to the CSF. Depending on the power spectra of the image, this can consume a lot of bits. Finally, when using a viewing distance for an application or image study, it is important to use a distance set for the closest distance expected. Therefore for studies using free-range viewing, a value of around 5 inches is suggested, since that is about as close as the average viewer can optically accommodate. However, for an application, it is much better to use a viewing distance corresponding the actual used. For example, in consumer imaging, distances closer to 14 inches are more appropriate and less than 11 Inches are unrealistic.

5. Summary

Approach 1 that performs a non-linear transducer function of each individual coefficient tends to preserve low-to-mid amplitude texture well. This is especially suitable for high quality photographic images that contain human faces. The exploitation of visual masking through the control of individual code-block contribution is very effective for large images with diverse content. In many cases, Approach 2 outperforms Approach 1. Approach 3 combines the strength of both approaches and therefore provides an overall best visual quality over different bit rates and for a large variety of images.

References

1. "JPEG 2000 Verification Model 4.1", ISO/IEC JTC1/SC29/WG1 N1286, June 8, 1999
2. Scott Daly, "Report on CE V1 (Quality Improvement Using Visual Masking)", ISO/IEC JTC1/SC29/WG1 N1202, March 15, 1999.
3. David Taubman, "High performance Scalable Image Compression with EBCOT", submitted to IEEE Transactions on Image Processing, March, 1999
4. T. Caelli et al. (1985) "The detection of phase shifts in two-dimensional images". Perception and Psychophysics, V. 37, pp. 536-542.