Journal of King Saud University – Computer and Information Sciences (2014) **26**, 202–217

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

A General scheme for dithering multidimensional signals, and a visual instance of encoding images with limited palettes

Mohamed Attia [a,b,c,*,1,2], **Waleed Nazih** [d,3], **Mohamed Al-Badrashiny** [e,4], **Hamed Elsimary** [d,3]

[a] *The Engineering Company for the Development of Computer Systems, RDI, Giza, Egypt*
[b] *Luxor Technology Inc., Oakville, Ontario L6L6V2, Canada*
[c] *Arab Academy for Science & Technology (AAST), Heliopolis Campus, Cairo, Egypt*
[d] *College of Computer Engineering and Sciences, Salman bin Abdulaziz University, AlKharj, Saudi Arabia*
[e] *King Abdul-Aziz City for Science and Technology (KACST), Riyadh, Saudi Arabia*

Received 12 March 2013; revised 30 August 2013; accepted 5 December 2013
Available online 12 December 2013

KEYWORDS

Digital signal processing;
Digital image processing;
Dithering;
Multidimensional signals;
Quantization noise;
Soft vector clustering

Abstract The core contribution of this paper is to introduce a general neat scheme based on soft vector clustering for the dithering of multidimensional signals that works in any space of arbitrary dimensionality, on arbitrary number and distribution of quantization centroids, and with a computable and controllable quantization noise. Dithering upon the digitization of one-dimensional and multi-dimensional signals disperses the quantization noise over the frequency domain which renders it less perceptible by signal processing systems including the human cognitive ones, so it has a very beneficial impact on vital domains such as communications, control, machine-learning, etc. Our extensive surveys have concluded that the published literature is missing such a neat dithering scheme. It is very desirable and insightful to visualize the behavior of our multidimensional dithering scheme; especially the dispersion of quantization noise over the frequency domain. In general, such visualization would be quite hard to achieve and perceive by the reader unless the

* Corresponding author at: Luxor Technology Inc., Oakville, Ontario, Canada. Tel.: +1 6475349166.
E-mail addresses: m_atteya2004@yahoo.com, m_atteya@rdi-eg.com (M. Attia).

[1] www.RDI-eg.com.
[2] www.AAST.edu.
[3] www.sau.edu.sa.
[4] www.KACST.edu.sa.

Peer review under responsibility of King Saud University.

ELSEVIER **Production and hosting by Elsevier**

http://dx.doi.org/10.1016/j.jksuci.2013.12.002

target multidimensional signal itself is directly perceivable by humans. So, we chose to apply our multidimensional dithering scheme upon encoding true-color images – that are 3D signals – with palettes of limited sets of colors to show how it minimizes the visual distortions – esp. contouring effect – in the encoded images.

1. Introduction

The main contribution of this paper is to introduce a general neat scheme for the *dithering* of multidimensional signals that is able to deal with arbitrary dimensionality, arbitrary number and distribution of quantization centroids, and with computable and controllable noise power. In order to proceed with presenting this novel multidimensional dithering scheme, it is necessary first to formally review one-dimensional signal digitization, quantization noise, and dithering.

The digitization of an analog one dimensional signal – known as *Analog-to-Digital* ("A-to-D" or "A2D") conversion – aims at mapping any given sample of the signal within its dynamic range $q_{\min} \leqslant q \leqslant q_{\max}$ to one element of a pre-defined set of quantum levels $\{c_1, c_2, \ldots, c_i, \ldots, c_L\}$; $q_{\min} \leqslant c_i \leqslant q_{\max}$, $L \geqslant 2$. In order to minimize the digitization error, this mapping is typically done through the *minimum-distance* criterion; i.e. the signal sample is mapped to the nearest quantum level, which can be formulated as follows:

$$q \xrightarrow{A\text{-}to\text{-}D} i_0 : i_0 = \arg \min_{\forall k; 1 \leq k \leq L} \{d(q, c_k)\}, \tag{1}$$

where $d(q_1, q_2)$ is any legitimate distance criterion between $q_1, q_2 \in \Re^1$. The digitization of a given signal sample in the 1D space is reduced into a simple selection of one of – at most – the two quantum levels enclosing that signal sample (Roberts, 2007; Widrow and Kollár, 2008) as illustrated by Fig. 1 below.

The sum of the squared digitization errors of all the emerging signal samples make the quantization noise which is formulated as follows (Roberts, 2007; Widrow and Kollár, 2008):

$$E_q^2 = \sum\nolimits_{\forall q} e_q^2(q) = \sum\nolimits_{\forall q} (q - c_{i_0})^2. \tag{2}$$

The distribution of the set of quantum levels over the dynamic range of the signal may be regular that $c_i = q_{\min} + (i-1) \cdot \frac{q_{\max} - q_{\min}}{L}$ and is then called *regular quantization*. When the distribution of emerging signal samples to be digitized is significantly irregular, the distribution of the quantum levels may be designed to track that irregular one of emerging samples, and is then called *adaptive quantization*.[5] Adaptive quantization aims at minimizing the quantization noise for any given number L of quantum levels (Roberts, 2007; Widrow and Kollár, 2008).

Increasing L obviously decreases both the digitization errors and quantization noise; however, there are hardware and/or computational cost limitations on the size of L to be deployed in a given digitization scheme. When L is not large enough to adequately capture the resolution of the analog signal, the digitized signal suffers from obtrusive artifacts that render its information content into a significantly distorted version from that carried by the original analog signal. This may turn into a serious drop of quality if the digitized signal is destined for human perception; e.g. digital audio, or may turn into a serious source of error if the digitized signal is forwarded to some further processing; e.g., machine learning, control systems, etc.

For example, consider an audio signal of a single tone – i.e. a purely sinusoidal wave – at 500 Hz. In the frequency domain, this analog signal shows a single impulse at 500 Hz and nothing elsewhere. When, this audio signal is digitized via 16-bit quantization; i.e. $2^{16} = 65{,}536$ quantum levels, the resulting digital signal in the frequency domain seems (almost) the same as the original analog one as illustrated by the blue curve at the top of Fig. 2. On the other hand, when the same audio signal is digitized via 6-bit quantization; i.e. $2^6 = 64$ quantum levels, the resulting digital signal in the frequency domain shows a major peak at 500 Hz but also other considerable harmonics like the one around 4500 Hz as illustrated by the red curve at the middle of Fig. 2. These obtrusive harmonics mean that the digitized signal is not corresponding any more to a pure single tone, but is corresponding to a composite one where irritating false whistles are superimposed on the original pure tone (Pohlmann, 2005).

Researchers and engineers had realized since decades that this problem is caused by the concentration of the digitization errors within narrow bands of the signal, and has accordingly realized that dispersing the digitization errors over wider bands in the frequency domain would produce a better digitized signal where obtrusive artifacts are less conspicuous. With signals digitized this way, humans would perceive a better quality, and digital signal processing systems would perform more robustly. Dispersing the digitization errors over wider bands is typically achieved through adding controlled noise to the analog signal just before the A-to-D conversion (Petri, 1996; Schuchman, 1964). This process is popularly known as *"dithering"* whose simplest – and also most commonly used – variant adds to each analog signal sample q some $\pm$ random value whose amplitude is half the distance between the two enclosing quantum levels c_i and c_{i-1}. Digitization with this kind of dithering may be formulated as follows:

$$i_0^* = \underset{\forall k: i-1 \leq k \leq i}{\operatorname{argmin}} \left\{ d\left(q + rand\left(\frac{c_{i-1} - c_i}{2}, \frac{c_i - c_{i-1}}{2}\right), c_k\right)\right\}. \tag{3}$$

Digitization with dithering of a given signal sample as described by Eq. (3) is still a selection of one of the two quantum levels enclosing that signal sample; however, unlike Eq. (1) this selection is a stochastic process where the chances of attributing the sample to each of the two quantum levels are given by:

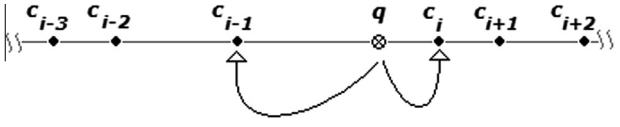

Figure 1 Dithering in 1D space; *only the two* enclosing quantum levels compete for the given point.

[5] The distribution of the quantum levels in Fig. 1 is assumed to belong to this second kind of quantization.

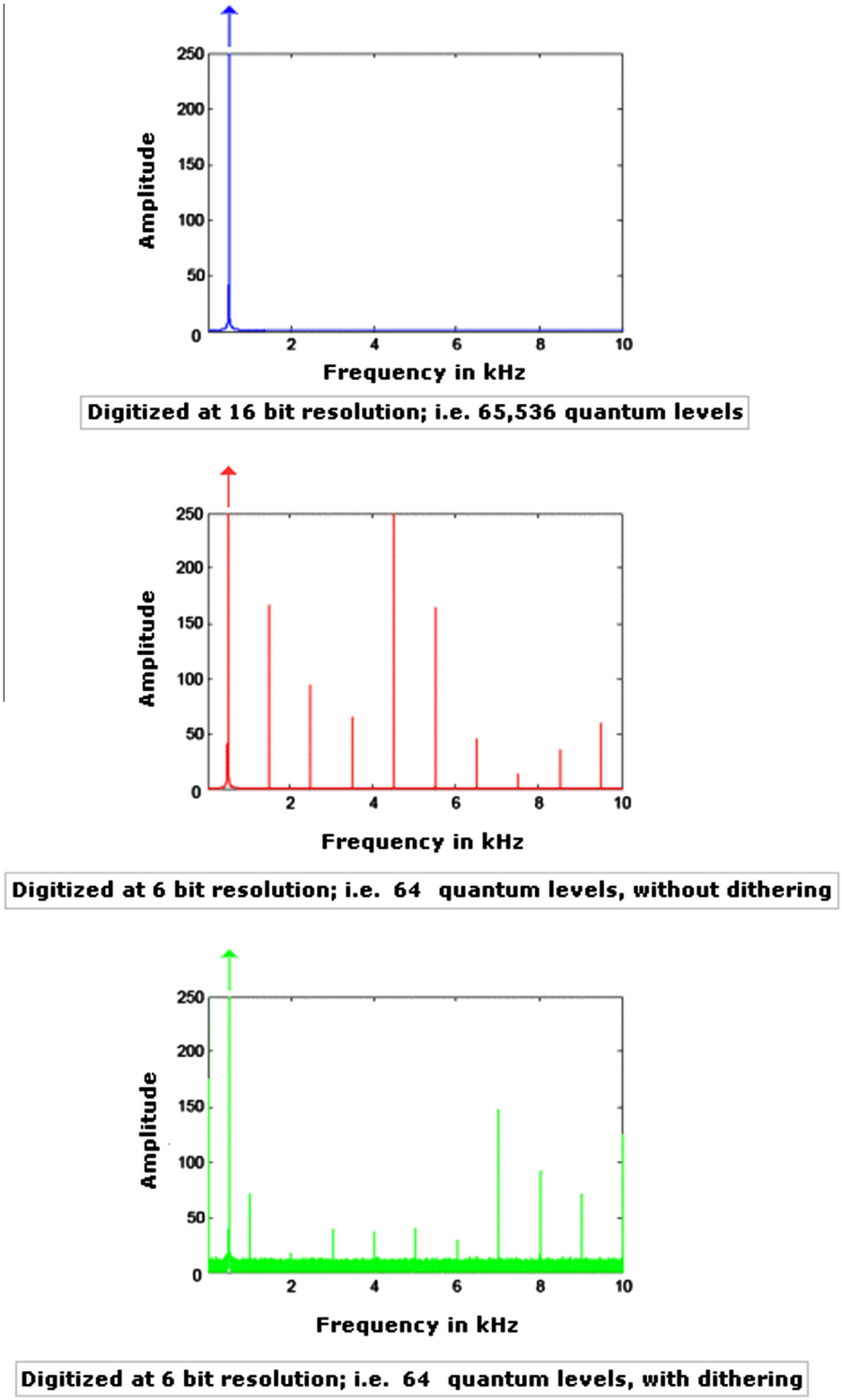

Figure 2 Dithering for the dispersion of quantization noise over the frequency domain.

$$P\left(q \xrightarrow{\text{Dithering}} i-1\right) = \frac{c_i - q}{c_i - c_{i-1}}, P\left(q \xrightarrow{\text{Dithering}} i\right) = 1 - P\left(q \xrightarrow{\text{Dithering}} i-1\right) = \frac{q - c_{i-1}}{c_i - c_{i-1}}. \quad (4)$$

The quantization noise of digitizing some population of signal samples with dithering is obviously larger than that of the quantization noise of the same population without dithering. This increase of quantization noise is the price paid to disperse

the quantization noise over wider frequency bands (Petri, 1996; Schuchman, 1964).

Applying this dithering model to the digitization of our exemplar single-tone audio signal via 6-bit quantization, the resulting digital signal in the frequency domain – illustrated at the bottom of Fig. 2 – shows white noise all over the signal spectrum and shows also some obtrusive harmonics but now with significantly lower amplitudes than those resulting from the digitization via 6-bit quantization without dithering. Apparently, the total quantization noise power in the white noise plus the obtrusive harmonics has increased; however, this noise has also been much more dispersed over the frequency spectrum. With the whiz of the white noise plus some much fainter obtrusive whistles, the original tone is much clearly identified in the digitized audio via 6-bit quantization with dithering than in the same audio digitized via 6-bit quantization without dithering where higher obtrusive whistles are irritatingly obscuring the original tone (Vanderkooy and Lipshitz, 1987).

Multidimensional signals – which arise in countless advanced modern applications of vital fields like control, communications, electronics, machine learning, image processing... etc. – need also be digitized before being digitally processed, and dithering is then an indispensible operation for alleviating the vagarities of digitization errors especially when multidimensional signals are digitized via limited sets of "quantum points"[6] that are not large enough for capturing the multidimensional resolution of these signals.

So, the next section discusses why the dithering of multidimensional signals is qualitatively more challenging than that of one-dimensional signals. Then, Section 2 proceeds to present our general neat scheme of dithering signals in any space of arbitrary dimensionality, and with arbitrary number and distribution of quantization centroids. Section 3 provides a quantitative analysis of the quantization noise due to our dithering scheme compared to that resulting from quantization without dithering.

It is very desirable and insightful to visualize the behavior of our multidimensional dithering scheme; especially the dispersion of quantization noise over the frequency domain. In general, such visualization would be quite hard to achieve and perceive by the reader unless the target multidimensional signal itself is directly perceivable by humans. While sound signals are one-dimensional, images are three-dimensional ones and make an ideal – and actually rare – instance for the sought visualization. As a means of visualizing the dynamic behavior of our novel multidimensional dithering scheme, we apply this scheme upon encoding true-color images with palettes of limited sets of colors to show how it minimizes the visual distortions – esp. contouring effect – in the encoded images. This application is discussed in detail in Section 4. Finally, Section 5 presents and discusses the comparative results of our experimentation with this application to a catalogue of miscellaneous true-color images.

2. The dithering of multidimensional signals

Consider Fig. 3 where the set of centroids in 2D space – or in multidimensional space in general – are regularly distributed over the dynamic range of the signal in each dimension.

[6] A quantization point in a multidimensional space is called *centroid*.

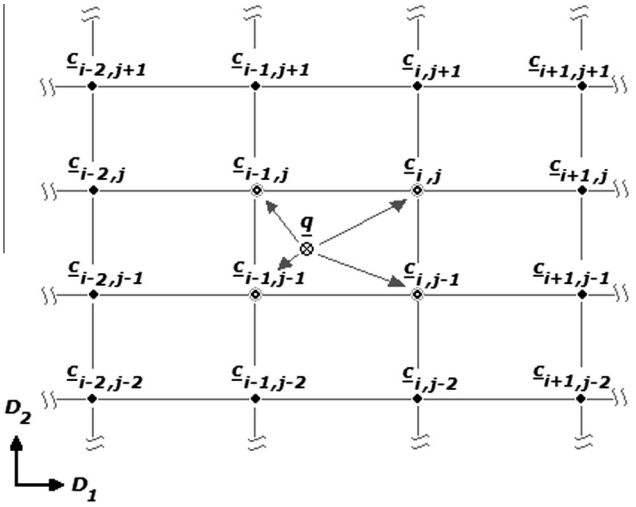

Figure 3 Dithering in two dimensional (or multi dimensional) space with regular quantization; only the quantum levels on the corners of the enclosing rectangle (or hyper rectangular prism) compete for the given point.

Given an emerging signal sample $\underline{q} = (q_1, q_2) \in \Re^2$ – located inside the rectangle whose four corners are $\underline{c}_{i-1,j-1}$, $\underline{c}_{i,j-1}$, $\underline{c}_{i-1,j}$, $\underline{c}_{i,j}$ that all belong to $\Re^2$ – to be digitized within this regular quantization setup, it is straightforward to augment Eq. (3) to two-dimensional space for the digitization with dithering of $\underline{q}$ that stochastically attributes it to one of the corners of its enclosing rectangle $\underline{c}_{i_0^* j_0^*}$ as follows:

$$\begin{aligned} i_0^* &= \underset{\forall k: i-1 \le k \le i}{\operatorname{argmin}} \left\{ d\left(q_1 + rand\left(\tfrac{c1_{i-1}-c1_i}{2}, \tfrac{c1_i-c1_{i-1}}{2}\right), c1_k\right)\right\} \\ j_0^* &= \underset{\forall k: j-1 \le k \le j}{\operatorname{argmin}} \left\{ d\left(q_2 + rand\left(\tfrac{c2_{j-1}-c2_j}{2}, \tfrac{c2_j-c2_{j-1}}{2}\right), c2_k\right)\right\} \end{aligned} \quad (5)$$

This procedure is extensible and applicable to any regular quantization setup in a space of any number of dimensions to attribute $\underline{q} = (q_1, q_2 \ldots q_D) \in \Re^D; D \geqslant 1$ to one of the 2^D corners of its enclosing hyper rectangular prism.

In such a regular quantization setup, dividing the dynamic range of the signal over each dimension into $(\ell_d - 1)$ intervals requires an overall number of centroids that is equal to:

$$L = \prod_{d=1}^{D} \ell_d; \ell_{d\in\{1,2\ldots D\}} \geq 2 \Rightarrow L = \ell^D; \ell_{d\in\{1,2\ldots D\}} = \ell. \quad (6)$$

This formula of exponential nature is prohibitively expensive with high dimensionality which is very common in real-life applications that need to digitize vectors of up to several tens of components; e.g. machine-learning feature-vectors. If, for example, we have to digitize a vectorial signal in a 20-dimensional space via a regular quantization setup where each dimension is to be modestly divided into 3 intervals, an immense number of $4^{20} \approx 1.1 \times 10^{12}$ centroids would be needed!

Due to their hyper-linear computational complexity patterns, real-life signal processing systems tend to deal with small sets of centroids and can at most deal with a couple-of-thousand centroids. Therefore, regular quantization schemes are seldom deployed and adaptive quantization schemes are instead resorted to. Vector clustering algorithms;

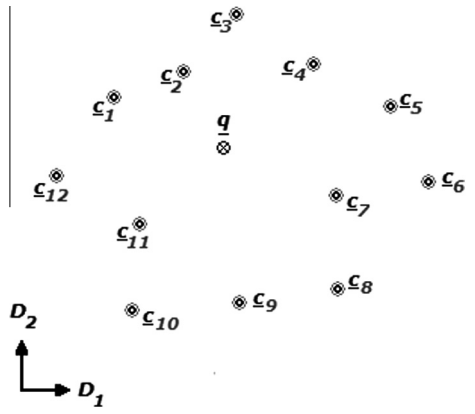

Figure 4 Dithering in two dimensional (or multi dimensional) space with irregular/adaptive quantization: all the quantum levels compete for the given point.

e.g. *k*-means and LBG,[7] are used to obtain an optimum set of centroids that are adaptive to the multidimensional signal to be digitized (Linde et al., 1980; Gray, 1984; Gray and Neuhoff, 1998; Jain et al. 2008). Fig. 4 illustrates an exemplar adaptive quantization setup:

The hard-deciding (i.e. deterministic) minimum-distance criterion for the digitization of any given signal sample in a multidimensional space with any distribution of centroids; regular or irregular, can be simply formulated as the vectorial form of Eq. (1) as follows[8]:

$$\underline{q} \overset{\text{Hard\_VQ}}{\longrightarrow} i_0 : i_0 = \underset{\forall k;1\le k\le L}{\operatorname{argmin}}\{d(\underline{q},\underline{c}_k)\}. \tag{7}$$

On the other hand, the dithering formula of Eq. (5) in the case of multidimensional regular quantization cannot be applied to the general case of multidimensional adaptive quantization because no enclosing rectangles (or "hyper rectangular prisms") can be decisively identified within irregular adaptive quantization setups like the exemplar one illustrated in Fig. 4.

In order to go around this hurdle, and to preserve the maximum compatibility with the basic dithering scheme dissected previously, our new dithering scheme for the general case of vector quantization in multidimensional spaces has to comply with the following three principal constraints:

1. As hyper rectangular prisms enclosing a given $\underline{q} \in \Re^D$ are absent in the general case, we allow *all* the centroids $\underline{c}_{j\in\{1,2,\ldots,L\}} \in \Re^D$ to compete for replacing $\underline{q}$ in the digitized signal representation.
2. This competition is arbitrated stochastically through what we call *Soft VQ* so that:
   $\forall i \in \{1,2,\ldots,L\} : P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i\right) > 0$ and $\sum_{i=1}^{L} P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i\right) = 1$.
3. With respect to a given $\underline{q}$, a more distant centroid $\underline{c}_{i_1}$ has a lower chance of replacing $\underline{q}$ than a less distant centroid $\underline{c}_{i_2}$; i.e. $d(\underline{q},\underline{c}_{i_1}) > d(\underline{q},\underline{c}_{i_2}) \Rightarrow P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i_1\right) < P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i_2\right)$.

[7] LBG is the acronym denoting the famous vector clustering algorithm developed by Linde, Buzo, and Gray (Linde et al., 1980).

[8] The digitization of vectorial signal samples is known also as Vector Quantization or VQ for shorthand.

Putting together all these constraints, our general dithering scheme based on Soft VQ can be formulated as:

$$P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i\right) = \frac{f(d(\underline{q},\underline{c}_i))}{\sum_{k=1}^{L} f\left(d(\underline{q},\underline{c}_k)\right)} = \frac{f(d_i)}{\sum_{k=1}^{L} f(d_k)}. \tag{8}$$

The function $f(d_i)$ must obey the following conditions:

1. $f(d_i) \ge 0 \quad \forall d_i \ge 0$,
2. $f(d_i)$ is continuous $\forall d_i \geqslant 0$,
3. $f(d_i)$ is a monotonically decreasing function $\forall d_i \geqslant 0$,
4. $d_i = 0 \Rightarrow P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} i\right) = 1 \wedge P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} k\neq i\right) = 0$.

In addition to satisfying the four conditions mentioned above, it is much desirable for the design of the function $f(x)$ to have the following properties:

1. Simplicity.
2. Tuning parameters to control the probability vanishing speed with increasing distance.
3. Analytic computability of the quantization noise energy of the resulting Soft VQ with respect to that of Hard VQ.

While the third of these desirable properties is subject to a detailed discussion over the next section, we select for our dithering scheme the inverse power-function that realizes all the necessary conditions and the first two desirable properties above. Our $f(x)$ is then defined as:

$$f(d_k) = d_k^{-m}; m > 0. \tag{9}$$

3. Quantization noise of Soft VQ based dithering

Dithering disperses the quantization noise over the frequency spectrum of the digitized signal, which is a great gain for the perceived signal quality and for the robustness of any subsequent digital signal processing as well. However, the price of this gain is the increase of total quantization noise, which might ruin the perceived signal quality or volatilize the stability of subsequent digital signal processing. Therefore, this section is devoted to a quantitative analysis of the additional quantization noise due to our general dithering scheme of multidimensional signals in order to see how safe its application to such signals is.

Digitized through Hard VQ formulated by Eq. (7) above, the local participation of each signal sample to quantization noise energy is given by:

$$e_{\text{Hard\_VQ}} = \min_{\forall k;1\le k\le L} (d(\underline{q},\underline{c}_k))^2 = d_{\min}^2. \tag{10}$$

The total quantization noise energy due to the digitization of a given population of signal samples of size N via Hard VQ is then given by Petri (1996) and Schuchman (1964):

$$E_{\text{Hard\_VQ}} = \sum\nolimits_{n=1}^{N} \min_{\forall k;1\le k\le L} (d(\underline{q}_n,\underline{c}_k))^2. \tag{11}$$

Digitized through Soft VQ formulated by Eq. (8) above, the local participation of each signal sample to quantization noise energy is given by:

$$e_{\text{Soft\_VQ}} = \sum_{k=1}^{L} \left(d_k^2 \cdot P\left(\underline{q} \overset{\text{Soft\_VQ}}{\longrightarrow} k\right)\right) = \frac{\sum_{k=1}^{L} (d_k^2 \cdot f(d_k))}{\sum_{k=1}^{L} f(d_k)}. \tag{12}$$

In Eq. (12): $d_k^2 \geq \left(d_{\min}^2 = (d(\underline{q}, \underline{c}_{i_0}))^2\right) \forall k; 1 \leq k \leq L$ is weighted by probabilities $\geqslant 0$, which results together with Eqs. (10) and (11) into:

$$e_{\text{Hard\_VQ}} \leq e_{\text{Soft\_VQ}} \Rightarrow 1 \leq r \equiv \frac{e_{\text{Soft\_VQ}}}{e_{\text{Hard\_VQ}}} \leq r_{\max} \Rightarrow 1 \leq \frac{E_{\text{Soft\_VQ}}}{E_{\text{Hard\_VQ}}} \leq r_{\max}. \tag{13}$$

The question is then to study how big the ratio r is, or at least how big its upper-bound $r_{\max}$ could be. Our selection of inverse power-function probability distributions proves to be quite useful at simplifying this task by substituting Eq. (9) in Eq. (12) to get:

$$e_{\text{Soft\_VQ}} = \frac{\sum_{k=1}^{L}(d_k^{2-m})}{\sum_{k=1}^{L}(d_k^{-m})}, \tag{14}$$

Then by substituting Eqs. (14) and (10) in Eq. (13), we get:

$$r \equiv \frac{e_{\text{Soft\_VQ}}}{e_{\text{Hard\_VQ}}} = \frac{\sum_{k=1}^{L}(d_k^{2-m}) / \sum_{k}^{L}(d_k^{-m})}{d_{\min}^2} = \frac{\sum_{k=1}^{L}(d_{\min}/d_k)^{m-2}}{\sum_{k=1}^{L}(d_{\min}/d_k)^{m}}. \tag{15}$$

Eq. (15) can be re-written more conveniently as:

$$r = \frac{1 + \sum_{\substack{k=1 \\ k \neq i_0}}^{L} \alpha_k^{m-2}}{1 + \sum_{\substack{k=1 \\ k \neq i_0}}^{L} \alpha_k^{m}}; \alpha_k \equiv \frac{d_{\min} = d_{i_0}}{d_k} \leqslant 1, \alpha_k \in [0, 1]. \tag{16}$$

Attia et al. (2010, 2011) investigate in detail the maximization of Eq. (16) – whose full derivation is unfolded in Appendix I at the end of this paper – and we present their findings as follows:

For $m < 2$ r can grow infinitely huge; i.e. $r_{\max} = \infty$ regardless to the value of L. So, this dithering scheme should not be used in this range in order to avoid the risk of producing devastating noise in the digitized signal.

For $m = 2$ $r_{\max} = L$ which may be a proper operating point for a small number of centroids; e.g. $L = 2$ or $L = 3$, but turns to be risky for a large number of centroids; e.g. $L = 256$ or $L = 1,024$.

For $m > 2$ there is one and only one maximum of $r = r_{\max}$ that occurs at $\underline{\alpha} = [\widehat{\alpha}_1, \widehat{\alpha}_2 \ldots \widehat{\alpha}_k \ldots \widehat{\alpha}_L]$ which – according to (Attia et al., 2012, 2011) – are both obtained via:

$$\widehat{\alpha}_k|_{\forall k \neq i_0} = \widehat{\alpha}, \tag{17}$$

$$r_{\max} = \frac{m-2}{m \cdot \widehat{\alpha}^2}. \tag{18}$$

…where $\widehat{\alpha}$ is the solution of the following polynomial:

$$\widehat{\alpha}^m + \left(\frac{m}{2} \cdot \frac{1}{L-1}\right) \cdot \widehat{\alpha}^2 - \frac{m-2}{2} \cdot \frac{1}{L-1} = 0; \quad m > 2, L \geq 2, \widehat{\alpha} \in [0, 1]. \tag{19}$$

Only for $m \in \{4, 6, 8, 10\}$ this polynomial has closed-form exact solutions (Jacobson, 2009); for example: for $m = 4$ the solution is:

$$\widehat{\alpha}^2\big|_{m=4} = \frac{1}{\sqrt{L}+1}, r_{\max}|_{m=4} = \frac{\sqrt{L}+1}{2}. \tag{20}$$

For $m \to \infty$ regardless to the value of L Soft VQ approaches the behaviour of Hard VQ, and we get:

$$\lim_{m \to \infty} r == 1. \tag{21}$$

For $L \to \infty$ – which is interpreted in practice as $L > > 1$; e.g. $L = 256$ or $L = 512$ ... etc. – we have the following excellent approximate solution that:

$$\lim_{L \to \infty} \widehat{\alpha} = \left(\frac{2 \cdot L}{m-2}\right)^{-1/m}, \lim_{L \to \infty} r_{\max} = \frac{m-2}{m} \cdot \left(\frac{2 \cdot L}{m-2}\right)^{\frac{2}{m}}. \tag{22}$$

Otherwise $\widehat{\alpha}$ is obtained numerically by finding the peak of Eq. (16) constrained by Eq. (17) like the exemplar charts illustrated in Fig. 5, Fig. 6, Fig. 7, and Fig. 8 below.

4. Encoding true-color images with a limited palette via Soft VQ

As mentioned in Section 1 previously, dithering is a common useful operation to accompany the fundamentally vital A-to-D conversion of one-dimensional signals. Dithering upon digitizing multidimensional signals is also useful for the same reason it is useful upon the digitization of one-dimensional signals; as it disperses quantization noise over the frequency domain which renders it less perceptible by signal processing systems – including the human cognitive ones – embedded in countless number of applications in vital domains such as communications, control, machine-learning… etc.

It is very desirable and insightful to visualize the behavior of our multidimensional dithering scheme; especially the dispersion of quantization noise over the frequency domain. In general, such visualization would be quite hard to achieve and perceive by the reader unless the target multidimensional signal itself is directly perceivable by humans. While sound signals are one-dimensional, images are three-dimensional ones and make an ideal – and actually rare – instance for the sought visualization. As a means of visualizing the dynamic behavior of our novel multidimensional dithering scheme, we apply this scheme upon encoding true-color images with palettes of limited sets of colors to show how it minimizes the visual distortions – esp. contouring effect – in the encoded images.

Color images are typically modeled as a three-dimensional signal in some *color space* like RGB, CIE-XYZ, or CIE-Lab (Hunt, 1998). The term "true-color image" denotes a digitized image where each color dimension is independently divided into 255 intervals or more. Therefore, each pixel in such an image is represented by $3 \times \log_2(255 + 1) = 24$ bits or more. The classic problem in this regard is to represent such images with an arbitrarily limited set of colors ($L \geqslant 2$) – called *palette* – with the minimum loss of quality (Dixit, 1991).

Virtually all the up-to-date high-end digital displays can manipulate true-color images; however, this has never been the case 15 or more years ago. However, there is still a need to deal with devices and setups with a limited – sometimes very limited – color display capabilities that each pixel can only be switched to one of a small set of colors. Here are some examples of such devices and setups that may still be in use on a wide scale:

1. Displays of low-end of digital gadgets; e.g. watches, calculators, wireless/cell phones…, especially when such gadgets are connected to the World Wide Web.
2. Printing devices with limited color capabilities such as monochrome printers, and fax machines.
3. Transferring images over slow/very slow Internet connections.

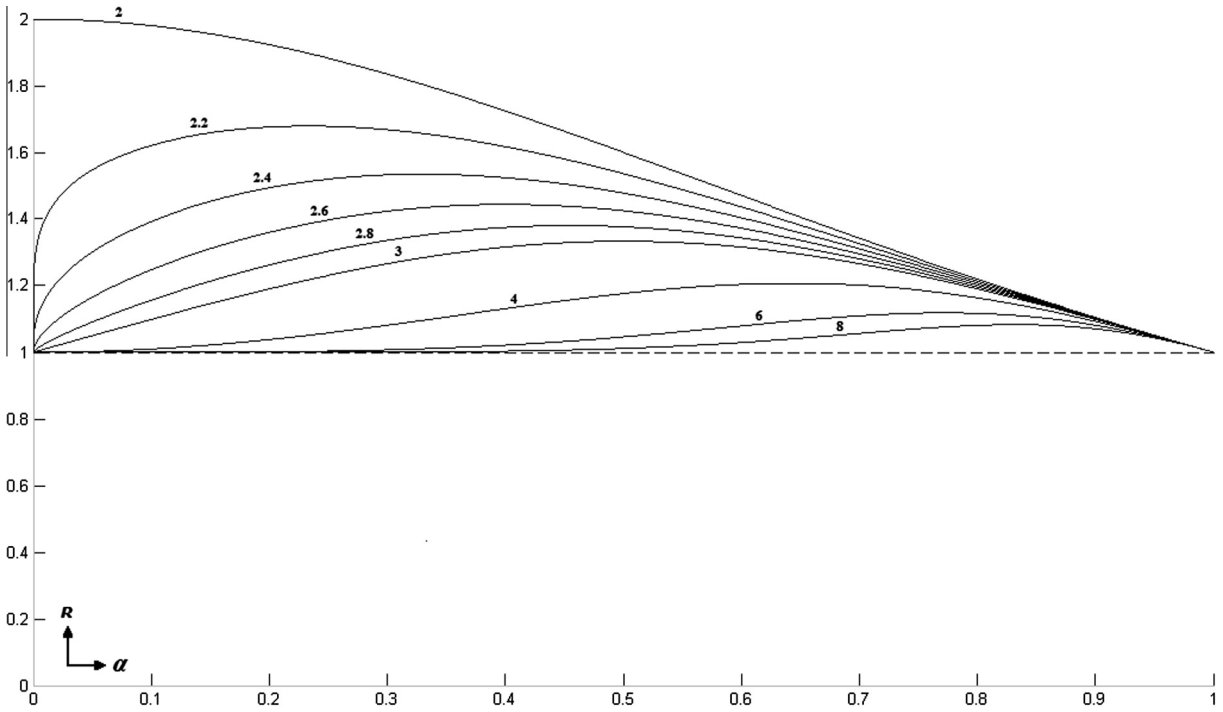

Figure 5 Quantization noise energy of our soft vector clustering relative to that of the hard clustering at two centroids ($L = 2$), and at various values of the power m (written in bold over each curve).

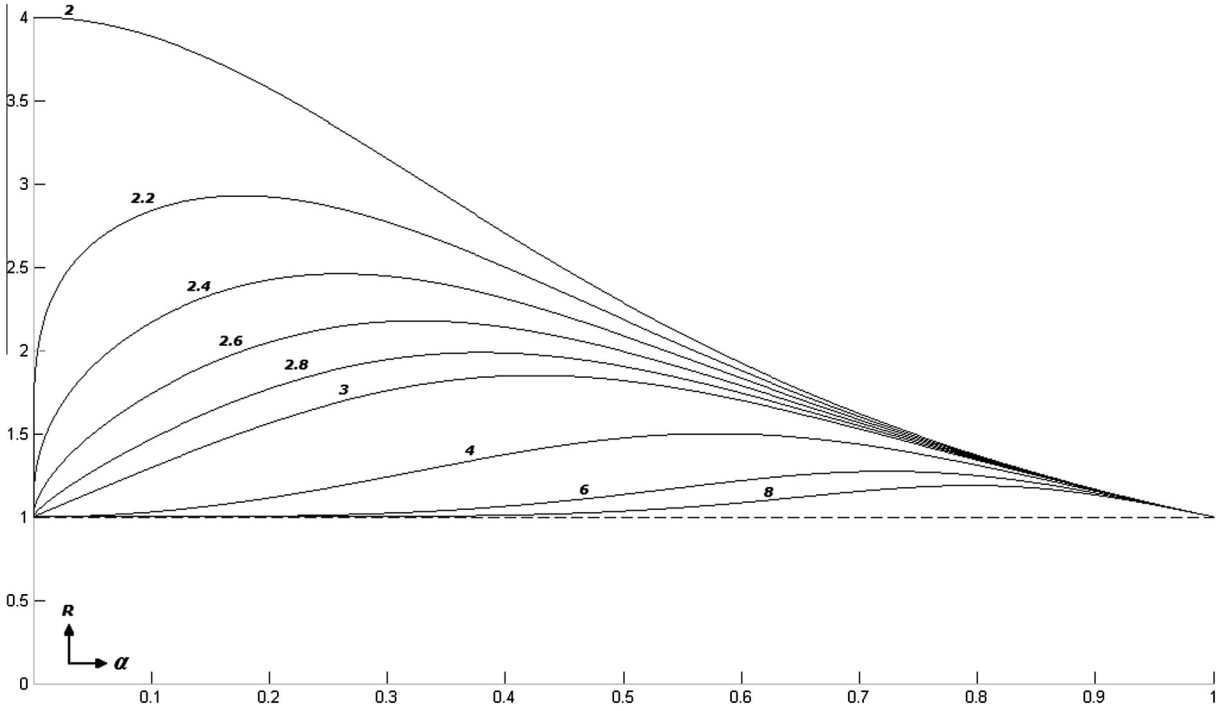

Figure 6 Quantization noise energy of our soft vector clustering relative to that of the hard clustering at four centroids ($L = 4$), and at various values of the power m (written in bold over each curve).

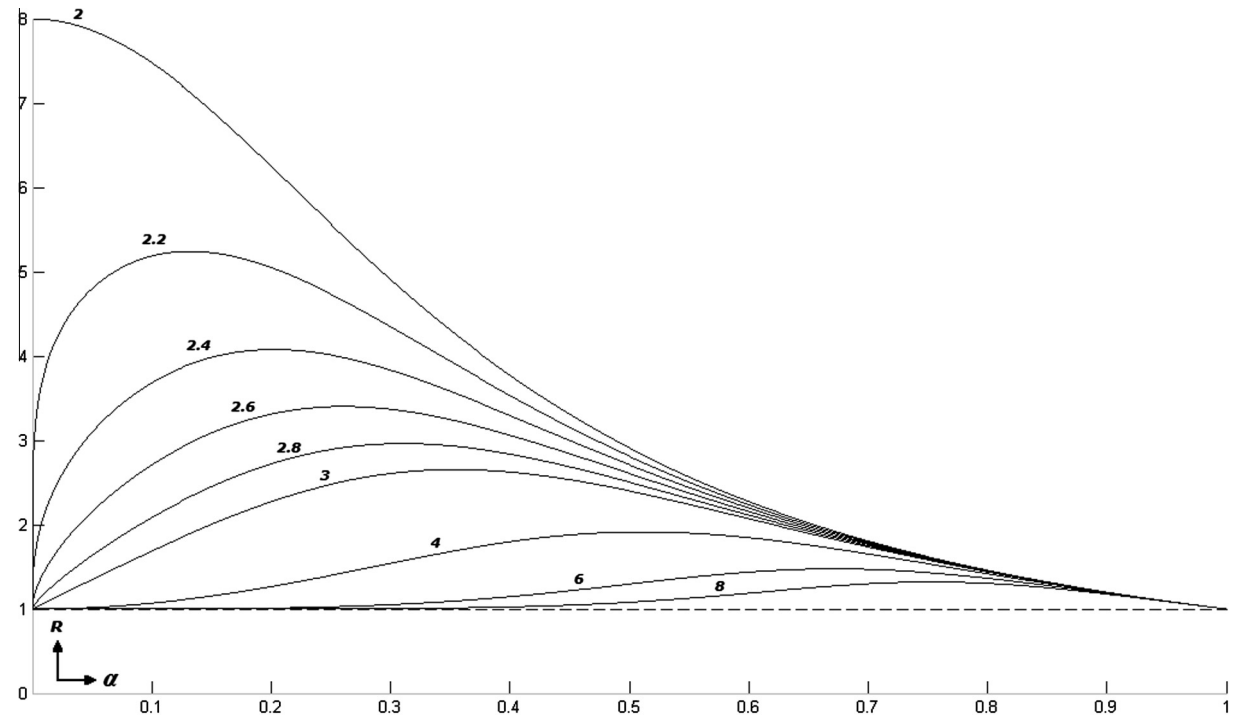

Figure 7 Quantization noise energy of our soft vector clustering relative to that of the hard clustering at eight centroids ($L = 8$), and at various values of the power m (written in bold over each curve).

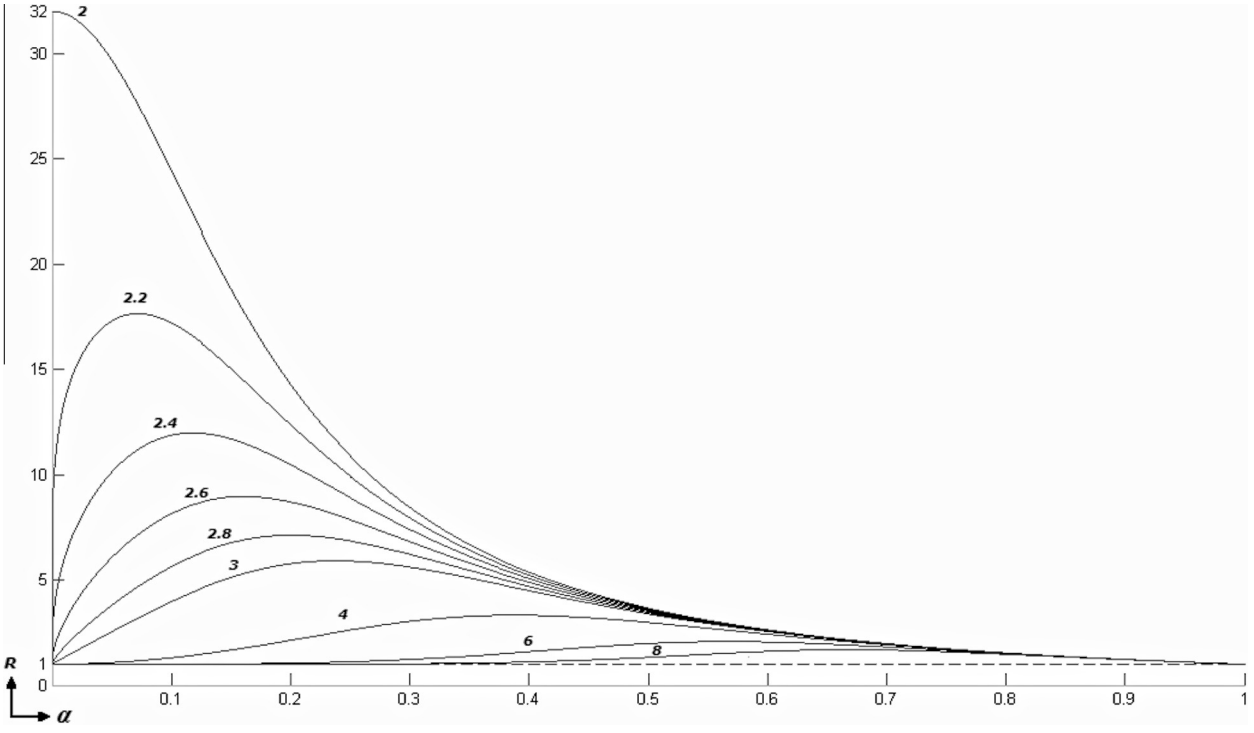

Figure 8 Quantization noise energy of our soft vector clustering relative to that of the hard clustering at thirty two centroids ($L = 32$), and at various values of the power m (written in bold over each curve).

4. Certain display modes of operating systems e.g. MS-Windows' safe mode, certain image formats; e.g. GIF, fast previewing of archived images; e.g. while browsing... etc. Such situations allow only a 16-color or 256-color palette-oriented image display.
5. Sizeable mosaic and mosaic-like image compositions.
6. Old fashioned sizeable electronic ad boards serving in places where it might be too expensive to replace them.

True-color images suffer from loss of image quality when encoded via Hard VQ with a limited palette as the original colors of some image details are replaced by different ones in the palette. This may not only confuse the image colors, but may also obscure the very characteristic features of the image. Such an encoding with a predefined palette produces more severe distortion than the case of an adaptive palette of the same size. Moreover, for the same palette acquisition and encoding methods, smaller palettes produce more severe image distortion than larger ones.

The oldest method to address this issue is *half-toning* that has been invented and deployed in the pre-digital era to display gray-scale analog photographs on old fashioned printed press produced via monochrome (i.e. black and white) printing. Half-toning prints a dot with a black area proportional to the darkness of the gray level of each point in the image; i.e. spatial resolution is traded for color resolution.

With the emergence of the digital age, more involved variants of half-toning have been developed to utilize the capabilities of monitors and digital graphic cards that can display multiple gray levels beyond the binary ones (Floyd and Steinberg, 1976). New variants of image dithering for colored images then emerged with the prevalence of color monitors and graphics cards with wider storage. Earlier such variants worked on each color dimension independently (Gentile et al., 1990), and then more sophisticated algorithms have been devised to simultaneously acquire the palette and perform the image digitization so that the loss of image quality in the digitized image is minimal (Kollias and Anastassion, 1991; Flohr and T., 1993; Ketterer et al., 1998; Cheng et al., 2009). The algorithm that provides the best results in this regard is commonly known as *scolorq* (for Spatial Color Quantization) (Ketterer et al., 1998); however, it is also the most mathematically sophisticated and computationally demanding one, because it is not a mere dithering algorithm but is actually an intricate solution of a combinatorial cost minimization problem.

On the other hand, Fig. 9 depicts our much simpler solution to this problem that is based on the direct application of our general dithering scheme presented in Section 3 of this paper.

The input digital image is assumed to be a true-color image in the RGB space which is the most common color space used for the representation of digital images on electronic displays. However, the Euclidean distance between two different points (i.e. colors) in the RGB space does not correspond to the difference in the human perception of these two colors (Hunt, 1998). So, the first module M_1 in our system converts the input image from the RGB color space to the CIE-Lab color space where the Euclidean distance between two different color vectors does correspond to the perceived visual difference of these two colors (Hunt, 1998).

Module M_2 then applies the LBG vector clustering algorithm (Linde et al., 1980; Gray, 1984; Gray and Neuhoff, 1998) to the population of the CIE-Lab color vectors of the pixels in the output image from M_1 in order to infer the optimal palette of a given size $L \geqslant 2$.

The CIE-Lab color vector of each pixel in the image is then dithered via our Soft VQ with inverse power-function distributions detailed in Section 3 above. So, given a power $m \geqslant 2$, the module M_3 stochastically attributes each CIE-Lab color vector $\underline{q}$ to one of the palette colors $\underline{c}_i$; $1 \leqslant i \leqslant L$ with probabilities calculated according to Eqs. (8) and (9).

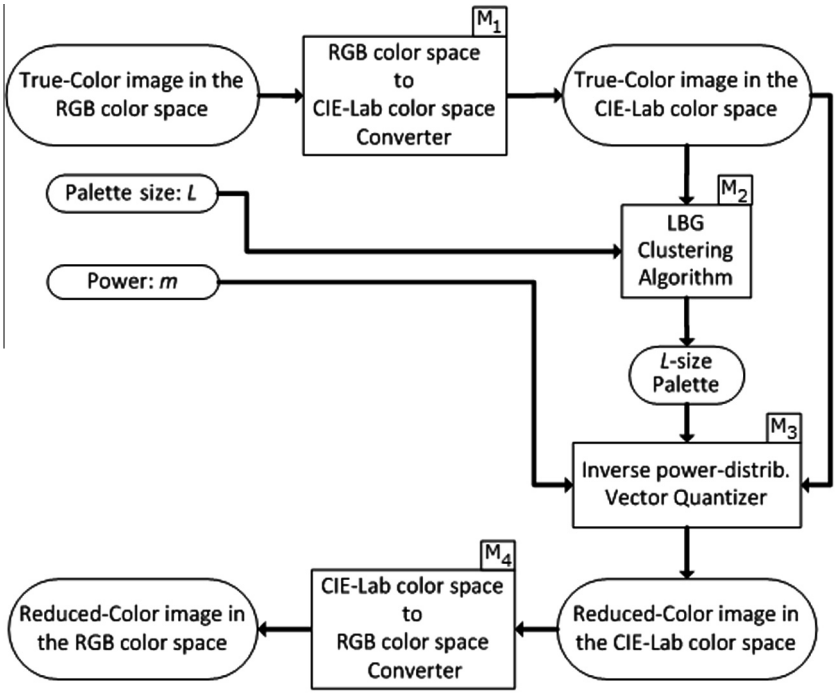

Figure 9 The overall structure of our system for optimally encoding true-color images with a limited palette.

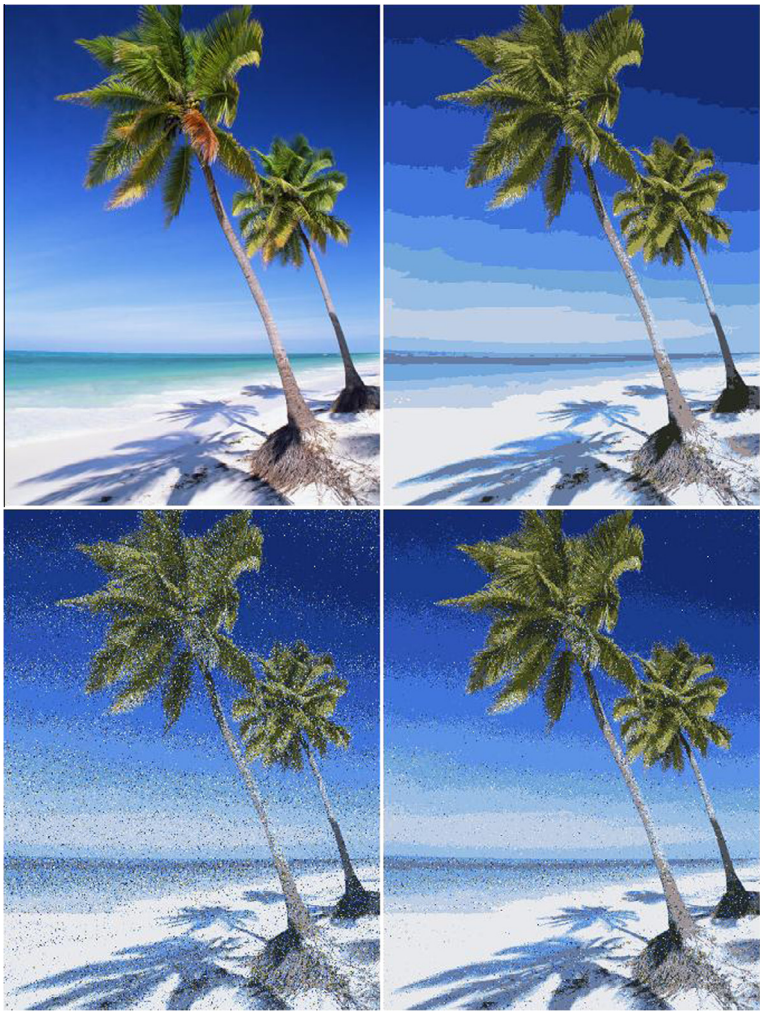

Figure 10 Original true-color image (upper left corner), image encoded with a 16-color palette (upper right corner) via LBG clustering and hard-deciding nearest-distance VQ, image encoded via LBG and Soft VQ with $m = 2$ (lower left corner), and finally the image encoded via LB and Soft VQ with $m = 3$ (lower right corner).

Figure 11 Original true-color image (upper left corner), image encoded with a 16-color palette (upper right corner) via LBG clustering and hard-deciding nearest-distance VQ, image encoded via LBG and Soft VQ with $m = 2$ (lower left corner), and finally the image encoded via LB and Soft VQ with $m = 3$ (lower right corner).

Figure 12 Original true-color image (upper left corner), image encoded with a 16-color palette (upper right corner) via LBG clustering and hard-deciding nearest-distance VQ, image encoded via LBG and Soft VQ with $m = 2$ (lower left corner), and finally the image encoded via LB and Soft VQ with $m = 3$ (lower right corner).

Module M_4 converts the output of M_3 from the CIE-Lab color space back into the RGB color space as the output of our system. The output of M_4 is then a digital image encoded with a limited palette that are compliant with digital image displays with any *color depth* $\geqslant \log_2(L)$.

Fig. 10 shows an illustrative example on the performance of our solution for encoding true-color images with a limited palette through Soft VQ with inverse power-function distributions. The upper left corner of the figure shows the input original true-color photographic image.

The upper right corner shows the result of encoding the original image via Hard VQ with an optimal (in the sense of minimum quantization noise) 16-color palette obtained through the LBG vector clustering algorithm. A severe obtrusive contouring effect[9] is quite apparent as a manifestation of the parasitic harmonics due to the concentration of quantization noise within tight frequency bands as discussed in Section 1 of this paper. Such a poor distribution does not only obscure the original content from the human viewers, but also creates false details that were never present in the original image.

The lower left and lower right corners of Fig. 10 show the results of applying our dithering scheme based on Soft VQ with $(L = 16, m = 2)$ and $(L = 16, m = 3)$ respectively. Obviously, the contouring effect has been alleviated at the price of higher quantization noise in both of these two images than that of the image obtained via Hard VQ in accordance with Eq. (13). It is also apparent that the quantization noise with $m = 3$ is lower than that with $m = 2$ in accordance with Eq. (16). The visual quality with $m = 3$ is a more balanced compromise of both alleviating the obtrusive effects and limiting the quantization noise compared with the two other cases in Fig. 10; the one at the upper right corner with too much

[9] Those false edges arising from abrupt transitions between disjoint color shades inadequately representing an area of a color gradient.

Figure 13 Original standard test-image of "Peppers".

Figure 14 Original standard test-image of "Mandrill" (also known as "Baboon").

contouring effect, and the one at the lower left corner with too much quantization noise.

Two more such examples like that of Fig. 10 are shown in Fig. 11 and Fig. 12. This illustrates how our dithering scheme is working in action on this kind of 3D signals by trading signal to quantization noise ratio for smoother distribution of this noise over the frequency spectrum of the signal. This behavior of algorithm is not specific to the self-visualizing 3D signals of true-color digital images but is also consistent with multidimensional signals in general.

5. Experiments and assessment

In order to learn in depth about its actual behavior, we have applied our dithering scheme depicted in Fig. 9 above to a

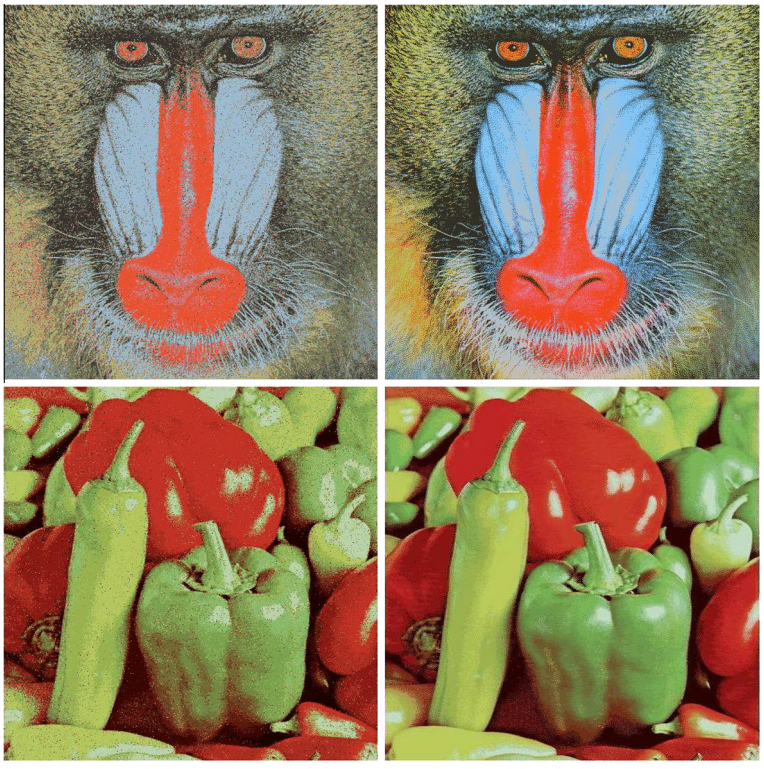

Figure 15 Mandrill image represented by 8-color palette via LBG and our Soft-VQ based dithering with $m = 2.5$ (upper left corner), then via scolorq (upper right corner). Peppers image represented by 16-color palette via LBG and our Soft-VQ based dithering with $m = 2.5$ (lower left corner), then via scolorq (lower right corner).

Table 1 Mean Opinion Score (MOS) of the image quality of our color encoding compared to *scolorq*.

Average MOS of *original* standard test images	Average MOS of standard test images encoded via *scolorq*	Average MOS of standard test images encoded via *LBG vector clustering and Soft VQ*
4.89 from 5	**4.47 from 5**	**3.98 from 5**

catalog containing tens of miscellaneous photos. For each photo; we archived the true-color photo, the one encoded with 16-color palette via Hard VQ, and three ones encoded with 16-color palette via Soft VQ with $m \in \{2, 3, 4\}$.[10] This archive is downloadable at http://www.rdi-e.g.com/downloads/Color-Reduction-Project_2012.rar. Each processed image confirms the behavior of our dithering algorithm as described in the end of the previous section. We have also empirically found that for $8 \leqslant L \leqslant 16$ the most visually accepted image encoding via Soft VQ are obtained at $m \in [2.5, 2.8]$.[11]

The experimental results presented in what follows are meant only to measure how successful our dithering scheme is with dispersing the quantization noise of the 3D signals of true-color images which is perceived by human viewers as alleviating the contouring effects. To put this measurement on a significant scale we compare the human judgments on the dithered images via our scheme with both the original true-color images and with the same images processed by *scolorq* (Ketterer et al., 1998) which is the ultimate algorithm for the representation of true-color images with a limited color palette.

With color images, *scolorq* goes much beyond than signal dithering as it produces the optimal color-reduced images via formally solving a combinatorial cost minimization problem in both the selection of the palette's colors and simultaneously the assignment of each pixel to one of these colors. Obviously, this intricate optimization is computationally expensive.

So, even if the images processed our way realize comparable Mean-Opinion-Score (MOS), this would be a considerable visual indication that our multidimensional signal dithering works so successfully at alleviating the contouring effect; i.e. dispersing concentrated noise around narrow bands in the 3D frequency domain, that the human viewers do not find a great quality difference between its output and the output of a quite elaborate technique as *scolorq* that takes care of many other aspects of color images.

Therefore, we picked five of the most famous bench-marking images used to assess image processing algorithms to compare our true-color image dithering with *scolorq*. While Fig. 13 and Fig. 14 show the true-color version of two of these five images, Fig. 15 shows the reduced color outputs of our image dithering and of *scolorq* for both images at $m = 2.5$.

We asked 13 volunteers independently to judge the visual quality of the output of both algorithms for each of the 5 input bench-marking images on a scale of five marks $\{Very\_Poor = 1,\ Poor = 2,\ Passable = 3,\ Good = 4,\ Excellent = 5\}$. The average results are presented in Table 1.

The results shown in the table suggest that the visual quality produced by *scolorq* are better – but not by far – than that produced by our method; however, we should also consider that ours is much mathematically simpler to perceive and implement, and seems to be less computationally complex.[12] It has to be restated that the scope of the presented multidimensional dithering scheme is wider and more profound than the color-reduction problem that we used as an easy-to-visualize instance of application, and the benefits of this dithering scheme to other vital domains as communications, control, machine-learning... etc. are great.

6. Conclusion

The paper started with a quick formal review of dithering upon analog-to-digital (A-to-D) conversion of one-dimensional signals along with its benefits to the perceived quality of the digitized signals. We then identified the difficulties of generalizing the simple dithering schemes applied to the one-dimensional signals to multidimensional signals. The paper proceeded with its main contribution of introducing a general neat scheme for the dithering of multidimensional signals that is able to deal with arbitrary dimensionality, arbitrary number and distribution of quantization centroids, and with computable and controllable noise power.

In order to visualize the dynamic behavior of our presented multidimensional dithering scheme, the paper then introduced a visually perceivable instance of applying this scheme via projecting it on the problem of encoding the 3D signals of true-color images for optimal viewing – in the sense of minimum visual distortion – on displays with limited sets of colors. We formally reviewed and chronologically surveyed the history of the significant approaches to tackle this problem along with the cons and pros of each. Then, the dithering scheme is instantiated for this specific 3D signal processing problem and incorporated into the architecture of our new solution of this problem. This solution is then run on a variety of images so that the optimal operating settings are empirically determined.

The experimental results are meant only to measure how successful our dithering scheme is with dispersing the quantization noise of the 3D signals of true-color images which is perceived by human viewers as alleviating the contouring effect. To put this measurement on a significant scale we compare the human judgments on the dithered images via our scheme with both the original true-color images and with the same images processed by *scolorq* which is the ultimate algorithm

[10] Very few photos in this archive are encoded with $L = 8$.

[11] For large palettes (e.g.$L = 256$; higher powers should be used to preserve a reasonable r_{max} according to eq. No. (22).

[12] The inventors of *scolorq* (Ketterer et al., 1998) did not provide a formula of the order of complexity of their sophisticated algorithm; however, we estimate their full-fledged optimization approach produces some hyper-linear (yet sub-exponential) order of complexity in the product of the image size and the size of reduced ···colors set. On the other hand, our dithering scheme tends to be linear in this product and might be reduced further through several optimizations in the implementation.

for the representation of true-color images with a limited color palette.

So, even if the images processed our way realize less yet comparable mean-opinion-scores (MOS), this is a considerable visual indication that our multidimensional signal dithering works so successfully at alleviating the contouring effect – that is concentrated noise around narrow bands in the frequency domain – that the human viewers does not find a great quality difference between its output and the output of a quite elaborate technique as scolorq that takes care of many other aspects of color images.

Acknowledgement

The work presented in this paper has been achieved through the 12-month research project (27/A.H./1432) funded by the Deanship of Scientific Research at Salman bin Abdulaziz University www.sau.edu.sa, AlKharj, Kingdom of Saudi Arabia (KSA) from October 2011 up to September 2012.

Appendix I.

The whole range of the positive power m in Eq. (16) – presented in Section 3 above – splits into three intervals with each defining one of the following three cases:

For $0 < m < 2$, it is obvious that the numerator of Eq. (16) grows indefinitely faster than the denominator for arbitrarily infinitesimal values of some $\alpha_k; k \in \Omega \subset \{1, 2, \ldots, L\}$ so that:

$$r_{\max} = \lim_{\alpha_k \to 0 \forall k \in \Omega} r\Big|_{0<m<2} = \frac{1 + \xi_N + \sum_{\forall k \in \Omega} \left(\lim_{\alpha_k \to \delta \to 0} \alpha_k^{m-2}|_{0<m<2} \right)}{1 + \xi_D + \sum_{\forall k \in \Omega} \left(\lim_{\alpha_k \to \delta \to 0} \alpha_k^{m}|_{0<m<2} \right)}$$
$$= \frac{1 + \xi_N + \omega \cdot \lim_{\delta \to 0} \left(\frac{1}{\delta}\right)^{2-m}}{1 + \xi_D + \omega \cdot 0} = \frac{1 + \xi_N + \infty}{1 + \xi_D + 0} = \infty, \tag{23}$$

where $\omega = SizeOf(\Omega)$ and $0 < \xi_N, \xi_D < L$ - ω - 1 .

This result necessitates the avoidance of the interval of $0 < m < 2$ as the possible unlimited growth of the soft quantization noise energy with respect to that of hard quantization would be devastating to the stability of whatever machine-learning procedure.

For $m = 2$, Eq. (16) reduces into:

$$r|_{m=2} = \frac{L}{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \alpha_j^2}; 1 \leqslant j \leqslant L, \tag{24}$$

...and one can easily guess that:

$$r_{\max} = \max(r|_{m=2}) = \lim_{\alpha_j \to 0 \forall j \neq i_0; 1 \leq j \leq L} (r|_{m=2}) = L. \tag{25}$$

For $m > 2$, the following three special cases of Eq. (16) can easily be noticed:

$$\lim_{m \to \infty} r = \frac{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{m \to \infty} \alpha_j^{m-2} \right)}{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{m \to \infty} \alpha_j^{m} \right)} = \frac{1 + \tau + 0}{1 + \tau + 0} = 1, \tag{26}$$

...where τ is the number of α_j's that are exactly equal to one. This shows that the quantization noise energy of our proposed Soft VQ with the power m growing larger is approaching the one of the hard-deciding VQ; however its distributions are also turning less smooth and more similar to those of the hard-deciding VQ.

$$\lim_{\forall \alpha_j \to 0; j \neq i_0} r = \frac{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{\alpha_j \to 0} \alpha_j^{m-2} \right)}{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{\alpha_j \to 0} \alpha_j^{m} \right)} = \frac{1 + (L-1) \cdot 0}{1 + (L-1) \cdot 0} = 1, \tag{27}$$

...which occurs only when $\underline{q} = \underline{c}_{i_0}$.

$$\lim_{\forall \alpha_j \to 1} r = \frac{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{\alpha_j \to 1} \alpha_j^{m-2} \right)}{1 + \sum_{\substack{j=1 \\ j \neq i_0}}^{L} \left(\lim_{\alpha_j \to 1} \alpha_j^{m} \right)} = \frac{1 + (L-1) \cdot 1}{1 + (L-1) \cdot 1} = 1, \tag{28}$$

...which occurs when $d(\underline{q}, \underline{c}_i)$ is exactly the same $\forall i; 1 \leqslant i \leqslant L$.

Only for these three special cases $r = 1$, otherwise $r > 1$. It is crucial to calculate the maximum value of $r = r_{\max}$; i.e. the worst case, which – according to Eq. (13) presented in Section 3 above – is an upper bound of the ratio between the total quantization noise energy of the proposed Soft VQ to that of the conventional hard VQ.

To obtain $r_{\max}$, the $(L-1)$ dimensional region within $\alpha_{j \neq i_0} \in [0, 1] \forall j$; $1 \leq j \leq L$ has to be searched for those $\widehat{\alpha}_{j \neq i_0}$ where that maximum is produced. This is analytically achievable via solving the following set of (L - 1) equations:

$$\partial r / \partial \alpha_k|_{\forall k \neq i_0} = 0; 1 \leq k \leq L. \tag{29}$$

For the sake of convenience, let us re-write Eq. (16) as:

$$r = \frac{A_k + \alpha_k^{m-2}}{B_k + \alpha_k^m}; A_k \equiv 1 + \sum_{\forall j \neq i_0, j \neq k} \alpha_j^{m-2}, B_k \equiv 1 + \sum_{\forall j \neq i_0, j \neq k} \alpha_j^{m}. \tag{30}$$

Then:

$$\partial r / \partial \alpha_k|_{\forall k \neq i_0} = 0 \Rightarrow \frac{(m-2) \cdot \widehat{\alpha}_k^{m-3}}{A_k + \widehat{\alpha}_k^{m-2}}\Big|_{\forall k \neq i_0} = \frac{m \cdot \widehat{\alpha}_k^{m-1}}{B_k + \widehat{\alpha}_k^{m}}\Big|_{\forall k \neq i_0}, \tag{31}$$

...that reduces into:

$$\frac{A_k + \widehat{\alpha}_k^{m-2}}{B_k + \widehat{\alpha}_k^{m}}\Big|_{\forall k \neq i_0} = r_{\max} = \left(\frac{m-2}{m} \right) \cdot \widehat{\alpha}_k^{-2}\Big|_{\forall k \neq i_0}. \tag{32}$$

In order for Eq. (31) to hold true, all $\widehat{\alpha}_{k \neq i_0}$ must be equal so that:

$$\widehat{\alpha}_k|_{\forall k \neq i_0} = \widehat{\alpha}, \tag{33}$$

...which reduces formula No. (30) into:

$$A = 1 + (L-2) \cdot \widehat{\alpha}^{m-2}, B = 1 + (L-2) \cdot \widehat{\alpha}^{m} \Rightarrow$$
$$r_{\max} = \frac{A + \widehat{\alpha}^{m-2}}{B + \widehat{\alpha}^{m}} = \frac{1 + (L-1) \cdot \widehat{\alpha}^{m-2}}{1 + (L-1) \cdot \widehat{\alpha}^{m}} = \left(\frac{m-2}{m} \right) \cdot \widehat{\alpha}^{-2} \tag{34}$$

Re-arranging the terms of (34), we get the polynomial equation:

$$\widehat{\alpha}^m + \left(\frac{m}{2} \cdot \frac{1}{L-1} \right) \cdot \widehat{\alpha}^2 - \frac{m-2}{2} \cdot \frac{1}{L-1} = 0;$$
$$m > 2, L \geq 2, \widehat{\alpha} \in [0, 1]. \tag{35}$$

For any $m > 2$, Eq. (35) can be shown to have one and only one real solution in the interval $\widehat{\alpha} \in [0, 1]$ through the following three-step proof:

1- Put $\widehat{\beta} = \widehat{\alpha}^2, c = \frac{1}{L-1}$, and re-write Eq. (35) as: $g(\widehat{\beta}) = \widehat{\beta}^{m/2} + \frac{m}{2} \cdot c \cdot \widehat{\beta} - \frac{m-2}{2} \cdot c = 0$.

2- $g(\widehat{\beta} = 0) = -\frac{m-2}{2} \cdot \frac{1}{L-1} < 0$

$g(\widehat{\beta} = 1) = \frac{2 \cdot L - 2 + m - m + 2}{2 \cdot (L-1)} = \frac{L}{L-1} > 0$

$\therefore g(\widehat{\beta})$ has roots ϵ [0, 1]

3- $\because dg(\widehat{\beta})/d(\widehat{\beta}) = \frac{m}{2} \cdot \widehat{\beta}^{m/2-1} + \frac{m}{2 \cdot (L-1)} > 0$

$\therefore g(\widehat{\beta})$ is a monotonically increasing function.

4- From steps 2 & 3, $g(\widehat{\beta})$ has only one root ϵ [0, 1].

A closed-form solution of Eq. (35) is algebraically extractable only for $(m/2) \epsilon \{2, 3, 4, 5\}$. (Jacobson, 2009)

When $m = 4$, for example; Eq. (35) turns into essentially a quadratic equation of the form $\widehat{\beta}^2 + 2 \cdot c \cdot \widehat{\beta} - c = 0$ whose well known closed-form solution is:

$$\widehat{\beta} = \widehat{\alpha}^2 = \frac{-2 \cdot c \pm \sqrt{4 \cdot c^2 + 4 \cdot c}}{2} = \sqrt{c^2 + c} - c,$$

Producing:

$$\begin{aligned} &\widehat{\alpha}^2|_{m=4} = \tfrac{\sqrt{L}-1}{L-1}, r_{\max}|_{m=4} = \tfrac{1}{2} \cdot \tfrac{L-1}{\sqrt{L}-1} \\ &\lim_{L\to\infty} \widehat{\alpha}^2|_{m=4} = \tfrac{1}{\sqrt{L}}, \lim_{L\to\infty} r_{\max}|_{m=4} = \tfrac{1}{2} \cdot \sqrt{L} \end{aligned} \tag{36}$$

As another example, when $m = 6$, Eq. (35) turns into $\widehat{\alpha}^6 + 3 \cdot c \cdot \widehat{\alpha}^2 - 2 \cdot c = 0$ which is a 3rd order equation of the form $\widehat{\beta}^3 + \eta_1 \cdot \widehat{\beta} + \eta_0 = 0$ whose closed-form solution is given (according to Jacobson, 2009) by:

$$\widehat{\beta} = -\frac{1}{3} \cdot \sqrt[3]{\left(\frac{1}{2}\right) \cdot \left(27 \cdot \eta_0 - \sqrt{27 \cdot \eta_0^2 + 4 \times 27 \cdot \eta_1^3}\right)} - \frac{1}{3} \cdot \sqrt[3]{\left(\frac{1}{2}\right) \cdot \left(27 \cdot \eta_0 + \sqrt{27 \cdot \eta_0^2 + 4 \times 27 \cdot \eta_1^3}\right)};$$

$\widehat{\alpha}^2|_{m=6}$ and $r_{\max}|_{m=6}$ are then given by:

$$\widehat{\alpha}^2|_{m=6} = \frac{\sqrt[3]{\left(1 + \sqrt{1 + \frac{1}{L-1}}\right)} - \sqrt[3]{\left(1 - \sqrt{1 + \frac{1}{L-1}}\right)}}{\sqrt[3]{L-1}},$$

$$r_{\max}|_{m=6} \frac{(2/3) \cdot \sqrt[3]{L-1}}{\sqrt[3]{\left(1 + \sqrt{1 + \frac{1}{L-1}}\right)} - \sqrt[3]{\left(1 - \sqrt{1 + \frac{1}{L-1}}\right)}},$$

$$\lim_{L\to\infty} \widehat{\alpha}^2|_{m=6} = \sqrt[3]{\frac{2}{L}}, \lim_{L\to\infty} r_{\max}|_{m=6} = \frac{2}{3} \cdot \sqrt[3]{\frac{L}{2}}. \tag{37}$$

For $(m/2) \notin \{2, 3, 4, 5\}$: one can only derive an expression for $\widehat{\alpha}^2$ and $r_{\max}$ at $L \to \infty$; i.e. with a large codebook. From Eqs. (36) and (37), one can speculate the generalization that:

$$\lim_{L\to\infty} \widehat{\alpha} = \left(\frac{2 \cdot L}{m-2}\right)^{-1/m}, \lim_{L\to\infty} r_{\max} = \frac{m-2}{m} \cdot \left(\frac{2 \cdot L}{m-2}\right)^{\frac{2}{m}} \tag{38}$$

Substituting that guess in the terms of Eq. (35) gives:

$$\lim_{L\to\infty} \frac{T_3}{T_1} = \lim_{L\to\infty} \left(-\left(\frac{2 \cdot L}{m-2}\right)^{-1} / \left(\frac{2 \cdot (L-1)}{m-2}\right)^{-1}\right) = -1,$$

$$\lim_{L\to\infty} \frac{T_2}{T_1} = \left(\frac{m}{m-2}\right) / \lim_{L\to\infty} \left(\frac{2 \cdot L}{m-2}\right)^{2/m} = 0,$$

$$\lim_{L\to\infty} \frac{T_2}{T_3} = \left(\frac{m}{m-2}\right) / \lim_{L\to\infty} \left(\frac{2 \cdot L}{m-2}\right)^{2/m} = 0,$$

. . .which confirms the validity of Eq. (38) as an approximation at $L \gg 1$.

References

Roberts, M.J., 2007. Fundamentals of Signal Processing. McGraw Hill.

Widrow B., Kollár, I., 2008. Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications, Cambridge University Press, pp. 485-528: Chapter 19 Dither, http://oldweb.mit.bme.hu/books/quantization/dither.pdf Cambridge, UK, 2008.

Pohlmann, K.C., 2005. Principles of Digital Audio. McGraw-Hill, ISBN 0071441565.

Petri, D., 1996. Dither signals and quantization. Elsevier Measur. 19 (3–4), 147–157.

Schuchman, L., 1964. Dither Signals and Their Effect on Quantization Noise. IEEE Trans. Commun. Technol. 12 (4), 162–165, ISSN: 0018-9332.

Vanderkooy, J., Lipshitz, S.P., 1987. Dither in digital audio. J. Audio Eng. Soc. 35 (12), 966–975.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for Vector Quantization Design. IEEE Trans. Commun. COM-28, 4–95.

Gray, R.M., 1984. Vector quantization. IEEE Signal Process. Magaz., 4–29.

Gray, R.M., Neuhoff, D.L., 1998. Quantization. IEEE Trans. Inf. Theory 44 (6), 2325–2383.

Jain, A.K., 2008. Data Clustering: 50 Years Beyond K-means http://biometrics.cse.msu.edu/Presentations/FuLectureDec5.pdf. Plenary Talk at The IAPR's 19th International Conference on Pattern Recognition http://www.icpr2008.org/, Tampa, FL USA, December 2008.

Attia, M., Almazyad, A., El-Mahallawy, M., Al-Badrashiny, M., Nazih, W., 2011. Soft vector quantization with inverse power-function distributions for machine learning applications, first ed.. In: Lecture Notes on Electrical Engineering: Intelligent Automation and Systems Engineering (Chapter 26), vol. 103 Springer-Verlag, Berlin Heidelberg, pp. 339–351, www.SpringerOnline.com, ISBN 978-1-4614-0372-2, http://www.Springer.com/engineering/circuits + %26 + systems/book/978-1-4614-0372-2.

Attia, M., Al-Mazyad, A., El-Mahallawy, M., Al-Badrashiny, M., Nazih, W., 2010. Post-clustering soft vector quantization with inverse power-function distribution, and application on discrete hmm-based machine learning. In: Proceedings of the International Conference on Signal Processing and Imaging Engineering (ICSPIE'10) ISBN: 978-988-17012-0-6/World Congress on Engineering and Computer Science 2010 (WCECS 2010). San Francisco, USA, October 2010, pp. 574–580 http://www.iaeng.org/publication/WCECS2010/WCECS2010_pp574-580.pdf.

Jacobson, N., 2009, second ed.. In: Basic Algebra, vol. 1 Dover, ISBN 978-0-486-47189-1.

Hunt, R.W., 1998. Measuring Colour, third ed. Fountain Press, UK, ISBN 0-86343-387-1.

Dixit, S., 1991. Quantization of color images for display/printing on limited color output devices. Comput. Graph. 15 (4), 561–567.

Floyd, R.W., Steinberg, L., 1976. An adaptive algorithm for spatial grey scale. Proc. Soc. Inf. Display 17, 75–77.

Gentile, R., Walowit, E., Allebach, J., 1990. Quantization and multilevel halftoning of color images for near original image quality. In: SPIE Human Vision and Electronic Models, Methods, and Applications, vol. 1249, pp. 249–259.

Kollias, S., Anastassion, D., 1991. A unified neural network approach to digital image halftoning. IEEE Trans. Signal Process. 39 (4), 980–984.

Flohr, T., Kolpatzik, B., Balasubramanian, R., Carrara, D., Bouman, C., Allebach, J., 1993. Model based color image quantization. In: Proceedings of the SPIE: Human Vision, Visual Processing, and Digital Display IV, vol. 1913, pp. 265–270.

Ketterer, J., Puzicha, J., Held, M., Fischer, M., Buhmann, J.M., Fellner, D., 1998. Computer vision – ECCV'98. In: On Spatial Quantization of Color Images, Lecture Notes on Computer Science (LNCS) Book Series, vol. 1406. Springer-Verlag, Berlin, Heidelberg, pp. 563–577, www.SpringerOnline.com.

Cheng, Cheuk-Hong, Au O.C., et al., 2009. Low color bit-depth image enhancement by contour region dithering. In: Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim 2009), Victoria, British Columbia, Canada, 23–26 August 2009.