Continued Fractions, Diophantine Approximations, and Design of Color Transforms

Yuriy A. Reznik*

Qualcomm Inc., 5775 Morehouse Dr., San Diego, CA 92121; USA

ABSTRACT

We study a problem of approximate computation of color transforms (with real and possibly irrational factors) using integer arithmetics. We show that precision of such computations can be significantly improved if we allow input or output variables to be scaled by some constant. The problem of finding such a constant turns out to be related to the classic Diophantine approximation problem. We use this relation to explain how best scaled approximations can be derived, and provide several examples of using this technique for design of color transforms.

1. INTRODUCTION

Color transforms are fundamental operations used in image/video acquisition, processing, encoding/decoding, and reproduction. They typically map colors presented in a reference color space (such as CIE's RGB or XYZ color spaces[1]) into a space, which is more suitable for a particular system or a device. For example, color television systems (NTSC, PAL, SECAM) have defined YUV-type color spaces (YIQ, YUV, YDbDr), which allowed to communicate chrominance information separately (which was needed for compatibility with black-and-white systems[1]) and in a way that it consumes much smaller bandwidth. Cameras, displays, and printers use their own types of color spaces specific to sensor, illuminant, or ink characteristics.

In most cases, color transforms are linear operators specified by matrices of transform factors. In some transforms, such as ones between YUV and YIQ color spaces,[1] transform factors are irrational, but more often they are specified as decimal fractions with finite (e.g. $3 \ldots 7$) number of digits after the period.

Nevertheless, most practical implementations of color transforms in today's digital systems use *approximations of transform factors* (denoted here as $\theta_1, \ldots, \theta_m$, $m \geqslant 2$), by *dyadic fractions*:

$$\theta_1 \approx p_1/2^k\,,\ \ \ldots\ ,\ \theta_m \approx p_m/2^k\,, \tag{1}$$

where $p_1, \ldots, p_m$, and k are integers. This way, multiplications or dot products with $\theta_1, \ldots, \theta_m$ can be efficiently implemented in integer arithmetics as follows:

$$\begin{array}{ccccc} x\theta_i & \approx & xp_i/2^k & \rightsquigarrow & (x * p_i) \gg k\,, \\ \sum_i x_i\theta_i & \approx & \sum_i x_i p_i/2^k & \rightsquigarrow & (\sum_i x_i * p_i) \gg k\,, \end{array}$$

where $*$ and $\gg$ denote integer multiplication and bit-wise right shift operations correspondingly.

The key parameter that influences the complexity of transforms using dyadic approximations (1) is the number of "precision bits" k. In software implementations, this parameter is often constrained by the width of registers (e.g. 8, 16 or 32), and failure to meet such a constraint can possibly double (or even quadruple) the execution time. In hardware designs, the parameter k directly affects the number of gates needed to implement adders and multipliers.

The precision of approximations (1) also depends on the parameter k. Thus, given k and θ_i, the best choice of p_i produces

$$\left|\theta_i - p_i/2^k\right| = 2^{-k}\left|2^k\theta_i - p_i\right| = 2^{-k}\min_{z\in\mathbb{Z}}\left|2^k\theta_i - z\right| \leqslant 2^{-k-1},$$

*Significant part of this paper was written while author was at Information Systems Laboratory, Stanford University, CA. Contact information: yreznik@ieee.org, phone: +1 (858) 658-1866.

which means, that minimum worst case magnitude of error

$$\Delta(k) = \min_{p_1,\ldots,p_m} \max_i \left\{ \left| \theta_i - p_i/2^k \right| \right\} . \tag{2}$$

is also bounded by

$$\Delta(k) \leqslant 2^{-k-1} . \tag{3}$$

In simple terms, this means, that on average, each bit of precision in dyadic approximations (1) reduces their worst case error at least by half.

In this paper, we develop a technique for improving precision of such approximations, based on the assumption[†] that *input or output variables of color transforms can be uniformly scaled by some constant*. In other words, instead of approximating constants $\theta_1, \ldots, \theta_m$ that are specified by a transform matrix, we suggest to approximate their scaled values:

$$\theta_1 \xi \approx p_1/2^k , \ \ldots \ , \ \theta_m \xi \approx p_m/2^k . \tag{4}$$

where $p_1, \ldots, p_m$, and k are integers, and ξ is a new "common factor" parameter that we introduce. We note, that in many applications, such uniform scaling can be perfectly acceptable (or even desirable, e.g. to provide headroom preventing overflows), or it can be easily "neutralized" by applying the inverse factor $1/\xi$ in an adjacent stage in image processing stack.

We show, that for infinitely many k, by carefully choosing the value of a common factor ξ the equivalent (scaled by $1/\xi$) worst case error of approximations (4):

$$\Delta_\xi(k) = \tfrac{1}{\xi} \min_{p_1,\ldots,p_m} \max_i \left\{ \left| \theta_i \xi - p_i/2^k \right| \right\} \tag{5}$$

can be made as small as

$$\Delta_\xi(k) \lesssim 2^{-k\left(1+\frac{1}{m-1}\right)} . \tag{6}$$

In other words, we show that common-factor-based approximations can be significantly more precise than direct ones. In applications to color transforms, where, typically $m = 3$, this means that we should be able to find approximations with $2^{-3/2k}$ error rate, and which means that compared to non-scaled designs we should be able to achieve same precision using 50% less bits! We confirm this prediction by providing several examples of transforms designs that demonstrate such complexity savings.

This paper is organized as follows. In the next section, we survey several known facts about rational approximations of real numbers. In Section 3, we study the problem of finding common factors ξ minimizing worst case errors of scaled dyadic approximations. We show that this problem is connected to one of finding best rational (Diophantine) approximations, and then we use this connection to derive our main results. Section 4 provides several practical examples showing how to apply our technique for design of color transforms. Finally, in Section 5 we provide our concluding remarks.

2. SOME FACTS FROM DIOPHANTINE APPROXIMATION THEORY

We start by recalling few properties of continued fractions and Diophantine approximations.[2]

2.1 Continued fractions and convergents

A finite *continued fraction* is a rational number presented in the form:

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots \cfrac{1}{a_n}}} , \tag{7}$$

[†]This idea is similar to one that we have previously suggested for implementations of Discrete Cosine Transforms.[4,5]

where a_0 is an integer, and a_k are positive integers for all $k \geqslant 1$. The usual compact notation for a continued fraction is:

$$[a_0, a_1, \ldots, a_n]\,.$$

Using, for example, Euclidean algorithm, it is easy to show that every rational number has a finite continued fraction expansion. Furthermore, if we consider an infinite sequence of integers $a_0, a_1, \ldots$, such that $a_k > 0, k \geqslant 1$, then a limit

$$\theta = \lim_{n\to\infty} [a_0, a_1, \ldots, a_n]$$

exists and is denoted by the infinite continued fraction expression $\theta = [a_0, a_1, \ldots]$. Conversely, if $\theta = \theta_0$ is an irrational number and if we recursively set

$$a_n = \lfloor \theta_n \rfloor\,, \quad \theta_{n+1} = \frac{1}{\theta_n - a_n}\,, \quad n = 1, 2, \ldots$$

then $\theta = [a_0, a_1, \ldots]$.

Two irrational numbers θ and θ' are said to be *equivalent* if:

$$\begin{aligned} \theta &= [a_0, a_1, \ldots, a_l, c_1, c_2, \ldots] \\ \theta' &= [b_0, b_1, \ldots, b_m, c_1, c_2, \ldots] \end{aligned}$$

for some suitable l, m and $a_0, a_1, \ldots, a_l, b_0, b_1, \ldots, b_m, c_1, c_2, \ldots$.

The *convergents* of an irrational number θ with infinite continued fraction expansion $\theta = [a_0, a_1, \ldots]$ are defined as

$$\frac{p_n}{q_n} = [a_0, a_1, \ldots, a_n]$$

where integers p_n, q_n are coprime. By setting $p_{-1} = 1, q_{-1} = 0, p_0 = a_0$ and $q_0 = 1$, these integers can be recursively obtained by

$$p_n = a_n p_{n-1} + p_{n-2}\,, \quad q_n = a_n q_{n-1} + q_{n-2}\,, \quad n = 1, 2, \ldots.$$

It can be noted that p_n, q_n are growing exponentially fast, and that convergents $\frac{p_n}{q_n}$ $(n = 1, 2, \ldots)$ produce rational approximations of θ such that:

$$|\theta - p_n/q_n| < 1/q_n^2\,. \tag{8}$$

Furthermore, obtained in such a manner approximations turn out to be *best*[2] in a sense that: $\big|\theta - p_n/q_n\big| < \big|\theta - p/q\big|$ for any integers p, $0 < q < q_n$.

2.2 Precision of rational approximations

The following fact (cf. [2, p. 11,Theorem V]) is an important consequence and refinement of the precision bound (8) that holds for convergents.

FACT 2.1. *Let θ be irrational. Then there exist infinitely many integers q and p such that*

$$|\theta - p/q| < \kappa(\theta) q^{-2}, \tag{9}$$

where:

$$\kappa(\theta) = \begin{cases} \frac{1}{\sqrt{5}}\,, & \text{if } \theta \text{ equivalent to } \frac{\sqrt{5}-1}{2} \quad (\text{root of } \theta^2 + \theta - 1 = 0)\,, \\ \frac{1}{2\sqrt{2}}\,, & \text{if } \theta \text{ equivalent to } \sqrt{2} - 1 \quad (\text{root of } \theta^2 + 2\theta - 1 = 0)\,, \\ \frac{5}{\sqrt{221}}\,, & \text{if } \theta \text{ equivalent to } \frac{\sqrt{221}-11}{10} \quad (\text{root of } 5\theta^2 + 11\theta - 5 = 0)\,, \\ \frac{13}{\sqrt{1517}}\,, & \text{if } \theta \text{ equivalent to } \frac{\sqrt{1517}-29}{26} \quad (\text{root of } 13\theta^2 + 29\theta - 13 = 0)\,, \\ \ldots \end{cases} \tag{10}$$

is a chain producing a sequence $\frac{1}{\sqrt{5}}$, $\frac{1}{2\sqrt{2}}$, $\frac{5}{\sqrt{221}}$, $\frac{13}{\sqrt{1517}}$, ... that tends to $\frac{1}{3}$.

In a case when we have multiple irrational constants that need to be approximated by rational numbers:

$$\theta_1 \approx p_1/q\,,\ \ \ldots\ ,\ \theta_m \approx p_m/q\,, \tag{11}$$

the following result holds (cf. [2, p. 14, Theorem III], [3, p.138]):

FACT 2.2. *Let $\theta_1, \ldots, \theta_m$, $(m \geqslant 2)$ be irrationals. Then, there are infinitely many integers q and $p_1, \ldots, p_m$, such that*

$$\max_i \{|\theta_i - p_i/q|\} < \tfrac{m}{m+1}\, q^{-1-1/m}\,. \tag{12}$$

We note, that in case when $m = 2$, formula (12) produces a somewhat weaker bound than (9), but it is certainly more general (valid for any number of constants $m \geqslant 2$).

3. PRECISION OF SCALED DYADIC APPROXIMATIONS

Given a set of real (and possibly irrational) constants $\theta_1, \ldots, \theta_m$, $(m \geqslant 2)$ we are now tasked with studying precision of scaled dyadic approximations

$$\theta_1\xi \approx p_1/2^k\,,\ \ \ldots\ ,\ \theta_m\xi \approx p_m/2^k\,. \tag{13}$$

where $p_1, \ldots, p_m$, and k are integers, and ξ is a new "common factor" parameter that we can adjust.

We immediately notice, that by picking some integer q, setting $\xi := q/2^k$, and then then multiplying both sides in (13) by $1/\xi$ we arrive at:

$$\tfrac{1}{\xi}\left|\xi\,\theta_i - p_i/2^k\right| = |\theta_i - p_i/q|\ ,\quad i = 1, \ldots, n\,,$$

which maps our problem into one of finding m simultaneous Diophantine approximations. The relevant result for this case is already provided by Fact 2.2.

Nevertheless, as we will show in this Section, there exists an even better value for ξ, which not only maps the problem to one of simultaneous Diophantine approximations, but also reduces the dimensionality of that problem.

We start by considering a special case when $m = 2$.

3.1 Minimizing errors of pairs of approximations.

By $\delta_1(\xi)$ and $\delta_2(\xi)$ we denote individual errors of approximations (13) as:

$$\delta_1(\xi) = \theta_1\xi - p_1/2^k\,,\qquad \delta_2(\xi) = \theta_2\xi - p_2/2^k\,, \tag{14}$$

and we are trying to find minimum of $\max\{|\delta_1(\xi)|\,, |\delta_2(\xi)|\}$ by adjusting ξ.

We make the following observation.

LEMMA 3.1. *Let θ_1, θ_2 be real numbers, such that $\theta_1\theta_2 > 0$, and let k, p_1, and p_2 be integers. Then, there exist values ξ^* and δ^*, such that*

$$\delta^* = \max\{|\delta_1(\xi^*)|\,, |\delta_2(\xi^*)|\} = \min_\xi \max\{|\delta_1(\xi)|\,, |\delta_2(\xi)|\}\ .$$

These values are:

$$\xi^* = \tfrac{1}{2^k}\tfrac{p_1+p_2}{\theta_1+\theta_2}\,, \tag{15}$$

and

$$\delta^* = \tfrac{1}{2^k}\left|\theta_1\tfrac{p_1+p_2}{\theta_1+\theta_2} - p_1\right| = \tfrac{1}{2^k}\left|\theta_2\tfrac{p_1+p_2}{\theta_1+\theta_2} - p_2\right|\ . \tag{16}$$

Proof. Condition $\theta_1\theta_2 > 0$ implies that both $\delta_1(\xi)$ and $\delta_2(\xi)$ are non-constant and have the same direction of growth with ξ.

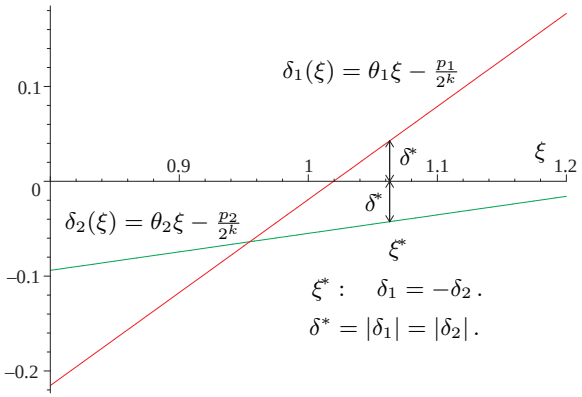

Figure 1. Finding $\min\limits_{\xi} \max\left\{\left|\,\theta_1\xi - \frac{p_1}{2^k}\,\right|, \left|\,\theta_2\xi - \frac{p_2}{2^k}\,\right|\right\}$.

If $\delta_1(\xi)$ and $\delta_2(\xi)$ intersect 0 at the same location, then there exists point ξ^* such that $\delta_1(\xi^*) = \delta_2(\xi^*) = 0$. This implies that

$$\xi^* = \tfrac{1}{2^k}\tfrac{p_1}{\theta_1} = \tfrac{1}{2^k}\tfrac{p_2}{\theta_2}\,,$$

which is a special case of (15).

If $\delta_1(\xi)$ and $\delta_2(\xi)$ intersect 0 at different locations, then there exists ξ^* such that (see Figure 1):

$$\delta_1(\xi^*) = -\delta_2(\xi^*)\,. \tag{17}$$

Moreover, since both $\delta_1(\xi)$ and $\delta_2(\xi)$ have same direction of growth, moving ξ away from ξ^* will lead to asymmetric changes in absolute values of $\delta_1(\xi)$ or $\delta_2(\xi)$. That is, one of them will increase. Therefore, ξ^* is the point of minimum of $\max\{|\delta_1(\xi)|\,, |\delta_2(\xi)|\}$.

By solving (17) with respect to ξ^* we arrive at formula (15), and by plugging (15) in (14), and using (17) we arrive at (16). □

3.2 Associated Diophantine approximation

Let us now further assume that p_1, p_2 have same signs as θ_1, and θ_2. Then, by denoting $p = p_1$, $q = p_1 + p_2$, and

$$\theta^* = \tfrac{\theta_1}{\theta_1+\theta_2}\,. \tag{18}$$

we observe that both parts of (16) turn into

$$\delta^* = \tfrac{|q|}{2^k}\,|\theta^* - p/q|\,.$$

By further de-scaling this quantity by ξ^* we arrive at

$$\delta^*/\xi^* = |\theta_1 + \theta_2|\;|\theta^* - p/q|\,, \tag{19}$$

which means, that by plugging $\xi = \xi^*$, the problem of finding minimum of the worst case error of a pair of scaled dyadic rational approximations

$$\Delta_\xi(k) = \tfrac{1}{\xi}\min_{p_1,p_2}\max\{|\delta_1(\xi)|\,, |\delta_2(\xi)|\}\,.$$

becomes equivalent to the problem of finding rational approximations of a single number θ^*

$$\theta^* \approx p/q\,. \tag{20}$$

Furthermore, if θ^* is irrational, then (20) turns into a classic Diophantine approximation problem (cf. Fact 2.1).

3.3 Main result for approximations of pairs of constants

We claim the following.

THEOREM 3.2. *Let θ_1, θ_2 be irrational numbers of the same sign. Then, there exist infinitely many integers k and real numbers ξ, such that*

$$\begin{aligned} \Delta_\xi(k) &= \tfrac{1}{\xi} \min_{p_1,p_2} \max \left\{ \left|\theta_1\xi - p_1/2^k\right|, \left|\theta_2\xi - p_2/2^k\right| \right\} \\ &< \kappa\left(\tfrac{\theta_1}{\theta_1+\theta_2}\right) \tfrac{4}{|\theta_1+\theta_2|}\, 2^{-2k} = O\left(2^{-2k}\right) . \end{aligned} \tag{21}$$

Proof. We use the following construction.

By assuming that $\xi = \xi^*$, and solving the associated Diophantine approximation problem (20), we find integers p, q satisfying precision constraint (9) of Fact 1. This also gives us integer factors $p_1 = p$ and $p_2 = q - p$ for our dyadic approximations. In order to select k, we can use some additional constraints. For example, we can require

$$1/2 < \xi^* \leqslant 1 , \tag{22}$$

which is satisfied by choosing $k = \lceil \log_2 (q/(\theta_1 + \theta_2)) \rceil$.

Then, by plugging precision bound (9) in (19), using lower bound for ξ^* from (22), and some simple algebra, we arrive at expression (21) claimed by the theorem. □

3.4 Extension of analysis to m-ary case

We now turn our attention to a problem of finding dyadic rational approximations for larger ($m > 2$) sets of numbers:

$$\theta_1\xi \approx p_1/2^k , \ \ldots \ , \ \theta_m\xi \approx p_m/2^k . \tag{23}$$

For simplicity, we assume that all numbers $\theta_1, \ldots, \theta_m$ and $p_1, \ldots, p_m$ are either positive or negative.

From Lemma 1, we know that for any pair of numbers θ_i, θ_j, $i \neq j$, we can compute factor

$$\xi^*_{ij} = \tfrac{1}{2^k} \tfrac{p_i+p_j}{\theta_i+\theta_j} , \tag{24}$$

which will "symmetrize" errors of approximations:

$$\delta^*_{ij} = \tfrac{1}{2^k} \left| \theta_i \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_i \right| = \tfrac{1}{2^k} \left| \theta_j \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_j \right| . \tag{25}$$

and which will turn them into a Diophantine approximation:

$$\delta^*_{ij} = \tfrac{|q_{ij}|}{2^k} \left| \theta^*_{ij} - p_{ij}/q_{ij} \right| . \tag{26}$$

where $p_{ij} = p_i$, $q_{ij} = p_i + p_j$, and

$$\theta^*_{ij} = \tfrac{\theta_i}{\theta_i+\theta_j}. \tag{27}$$

By applying ξ^*_{ij} to the remaining constants $\{\theta_k, k \neq i, j\}$, we note that their scaled approximations also turn into standard Diophantine forms:

$$\left|\theta_k\xi^*_{ij} - p_k/2^k\right| = \tfrac{1}{2^k}\left|\theta_k \tfrac{p_i+p_j}{\theta_i+\theta_j} - p_k\right| = \tfrac{|q_{ij}|}{2^k} \left| \theta^*_{k|ij} - p_k/q_{ij} \right| ,$$

where, however, the resulting constants

$$\theta^*_{k|ij} = \tfrac{\theta_k}{\theta_i+\theta_j} , \tag{28}$$

and errors of their approximations are different.

This means that by using factor ξ^*_{ij} we can reduce the problem of finding m dyadic rational approximations (23) to one of finding $m-1$ simultaneous Diophantine approximations:

$$\theta^*_{ij} \approx p_{ij}/q_{ij}, \left\{\theta^*_{k|ij} \approx p_k/q_{ij}, k \neq i,j\right\}. \tag{29}$$

This leads to the following result.

THEOREM 3.3. *Let $\theta_1, \ldots, \theta_2$ be $m > 2$ irrational numbers of the same sign. Then, there exist infinitely many integers k and real values ξ, such that*

$$\begin{aligned}\Delta_\xi(k) &= \tfrac{1}{\xi} \min_{p_1,\ldots,p_m} \max_i \left\{\left|\theta_i\xi - p_i/2^k\right|\right\} \\ &< \tfrac{m-1}{m}\left(\min_{ij}\{|\theta_i+\theta_j|\}\right)^{-\frac{1}{m-1}} 2^{-(k-1)\left(1+\frac{1}{m-1}\right)} \\ &= O\left(2^{-k\left(1+\frac{1}{m-1}\right)}\right). \end{aligned} \tag{30}$$

Proof. We use the following construction.

We scan all $\binom{m}{2}$ pairs of indices i,j, and find a pair, for which the normalized (by $1/\xi^*_{ij}$) worst case error:

$$\begin{aligned}&\tfrac{1}{\xi^*_{ij}} \min_{p_{ij},p_k} \tfrac{|q_{ij}|}{2^k} \max\left\{\left|\theta^*_{ij} - \tfrac{p_{ij}}{q_{ij}}\right|, \left|\theta^*_{k|ij} - \tfrac{p_k}{q_{ij}}\right|, k \neq i,j\right\} \\ &\quad = |\theta_i+\theta_j| \min_{p_{ij},p_k} \max\left\{\left|\theta^*_{ij} - \tfrac{p_{ij}}{q_{ij}}\right|, \left|\theta^*_{k|ij} - \tfrac{p_k}{q_{ij}}\right|, k \neq i,j\right\}\end{aligned}$$

is the smallest one.

Then, by applying Fact 2, using (24) to replace q_{ij} with 2^k and ξ^*_{ij}, and subsequently, bounds $1/2 < \xi^*_{ij} \leqslant 1$ (which is attainable by choice of k), and $|\theta_i+\theta_j| \geqslant \min_{ij}\{|\theta_i+\theta_j|\}$, we arrive at estimate (30) claimed by the theorem. □

4. APPLICATIONS TO THE DESIGN OF COLOR TRANSFORMS

In Table 1 we summarize specifications of several color transforms that we will use as examples. Most of these transforms belong to a class of RGB-to-YUV-type transforms, which are well known, and widely used in existing analog and digital television systems.[1] Color space YSbSr (see last column of Table 1), is a more recent proposal, and it seems to be of interest for future image and video coding applications.[6]

The transform parameters α, β, γ are used to compute luminance components

$$Y = \alpha R + \beta G + \gamma B,$$

whereas δ and ϵ, or alternatively κ, λ, μ, ν are the factors involved in computation of chrominance components according to flow-graphs presented in Figure 2. We note that flow-graph on the right in Figure 2 is a more general one, and it allows specification of transforms with rotations of chroma coordinates, such as transforms to YIQ or YSbSr spaces.

As a first example for using our approximation technique, consider computations of products by factors $\theta_1 := \delta$, and $\theta_2 := \epsilon$ involved in computation of chrominance (U,V) coordinates. There are only 2 factors that need to be simultaneously approximated in this case, and so we need to find convergents for the associated factor:

$$\theta^* = \tfrac{\theta_1}{\theta_1+\theta_2} \approx p/q$$

and then use them for deriving scale factor ξ^*. We illustrate all steps in this process in Table 2.

We notice, that the first scaled approximation:

$$\delta\xi \approx 1/2, \quad \epsilon\xi \approx 1/2, \tag{31}$$

Table 1. Parameters of several existing RGB-to-YUV-type color transforms.

Factors	**Color spaces / Standards**					
	YUV PAL	**YDbDr** SECAM	**YPbPr/YCbCr** ITU-R BT.601	**YPbPr/YCbCr** ITU-R BT.709	**YIQ** NTSC	**YSrSb** [6]
α	0.299	0.299	0.299	0.2125	0.299	0.3227
β	0.587	0.587	0.587	0.7154	0.587	0.3447
γ	0.114	0.114	0.114	0.0721	0.114	0.3326
δ	$0.615\,\frac{1}{1-\alpha}$	$1.333\,\frac{1}{1-\alpha}$	$\frac{1}{2}\frac{1}{1-\alpha}$	$\frac{1}{2}\frac{1}{1-\alpha}$		
ϵ	$0.436\,\frac{1}{1-\gamma}$	$-1.333\,\frac{1}{1-\gamma}$	$\frac{1}{2}\frac{1}{1-\gamma}$	$\frac{1}{2}\frac{1}{1-\gamma}$		
κ	0.615	1.333	$\frac{1}{2}$	$\frac{1}{2}$	$0.877\,(1-\alpha)\cos(33)$ $+0.492\,\alpha\,\sin(33)$	-0.1643
λ	$0.436\,\frac{\alpha}{1-\gamma}$	$-1.333\,\frac{\alpha}{1-\gamma}$	$\frac{1}{2}\frac{\alpha}{1-\gamma}$	$\frac{1}{2}\frac{\alpha}{1-\gamma}$	$-0.877\,(1-\alpha)\sin(33)$ $+0.492\,\alpha\,\cos(33)$	0.5095
μ	$0.615\,\frac{\gamma}{1-\alpha}$	$1.333\,\frac{\gamma}{1-\alpha}$	$\frac{1}{2}\frac{\gamma}{1-\alpha}$	$\frac{1}{2}\frac{\gamma}{1-\alpha}$	$0.877\,\gamma\,\cos(33)$ $+0.492\,(1-\gamma)\sin(33)$	0.3470
ν	0.436	-1.333	$\frac{1}{2}$	$\frac{1}{2}$	$-0.877\,\gamma\,\sin(33)$ $+0.492\,(1-\gamma)\cos(33)$	0.3870

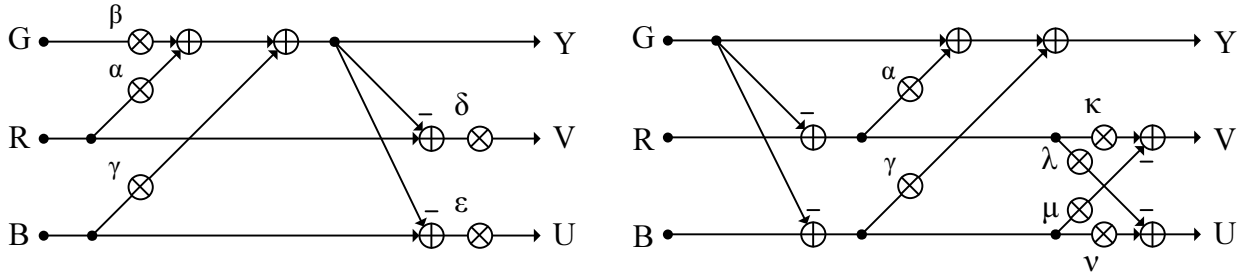

Figure 2. Flow-graphs and factors in RGB to YUV-type transforms.

leads to exactly the same dyadic fractions as non-scaled one, but the worst case error of such approximations turns our to be almost 3 times smaller (0.07.. vs 0.2..)!

The next pair of scaled approximation (using 3-bit denominator):

$$\delta\xi \approx 4/8 = 1/2\,, \quad \epsilon\xi \approx 5/8\,, \tag{32}$$

produces error that is almost 20 times smaller than one of non-scaled approximation (0.003 vs. 0.06). The error of these appoximations is so small that they may become useful in practice. They may be suitable, for example, in cases when input and output values are quantized to 8-bit resolution, in which case the maximum error resulting from using these approximations is less than a quantization step size.

Indeed, if further precision is consider necessary, Table 2 lists a couple of additional options. In all cases it can be seen that scaled approximations are remarkably more precise than non-scaled ones.

Next, we apply our technique for approximation of 3 constants α, β, γ involved in computation of luminance. This a bit more complex process, where we need to enumerate pairs of indices i, j with the goal of finding associated factor θ^*_{ij} leading to best set of rational approximations:

$$\theta^*_{ij} = \tfrac{\theta_i}{\theta_i+\theta_j} \approx p/q\,, \quad \text{and} \quad \left\{\theta^*_{k|ij} = \tfrac{\theta_k}{\theta_i+\theta_j} \approx p_k/q, k \neq i, j\right\}\,,$$

and then use their denominator q for deriving scale factor ξ^*. Details of this process are summarized in Table 3.

Table 2. Approximations of a pair of constants $\theta_1 = \delta(\text{YCbCr}) \approx 0.5643340858$, and $\theta_2 = \epsilon(\text{YCbCr}) \approx 0.7132667618$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k, \quad \theta_2 \approx p_2/2^k$				Associated rational appr-s: $\theta^* = \theta_1/(\theta_1+\theta_2) \approx p/q$			Scaled dyadic approximations: $\theta_1\xi^* \approx p_1/2^k, \quad \theta_2\xi^* \approx p_2/2^k$			
k	p_1	p_2	$\max_i \left\lvert\theta_i - \frac{p_i}{2^k}\right\rvert$	q	p	$\left\lvert\theta^* - \frac{p}{q}\right\rvert$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_1+\theta_2}$	p_1	p_2	$\frac{1}{\xi^*}\max_i \left\lvert\theta_i\xi^* - \frac{p_i}{2^k}\right\rvert$
1	1	1	0.2132667618	2	1	0.0582860744	0.7827170762	1	1	0.0744663380
2	2	3	0.0643340858							
3	5	6	0.0606659142	9	4	0.0027305188	0.8805567108	4	5	0.0030718336
4	9	11	0.0257667618							
5	18	23	0.0054832382	43	19	0.0001465395	1.0517760712	19	24	0.0001872190
6	36	46	0.0054832382							
7	72	91	0.0023292618	163	72	0.0000038658	0.9967412768	72	91	0.0000049389
8	144	183	0.0018340858							
9	289	365	0.0003761368							
10	578	730	0.0003761368							
11	1156	1461	0.0001190392							

Table 3. Approximations of constants: $\theta_1 = \alpha = 0.299$, $\theta_2 = \beta = 0.587$, and $\theta_3 = \gamma = 0.114$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k$, $\theta_2 \approx p_2/2^k$, $\theta_3 \approx p_3/2^k$					Associated rational appr-s: $\theta^*_{ij} = \theta_i/(\theta_i+\theta_j) \approx p/q$					Scaled dyadic approximations: $\theta_1\xi^* \approx p_1/2^k, \quad \theta_2\xi^* \approx p_2/2^k, \quad \theta_3\xi^* \approx p_3/2^k$				
k	p_1	p_2	p_3	$\max_i \left\lvert\theta_i - \frac{p_i}{2^k}\right\rvert$	i	j	q	p	$\left\lvert\theta^*_{ij} - \frac{p}{q}\right\rvert$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_i+\theta_j}$	p_1	p_2	p_3	$\frac{1}{\xi^*}\max_i \left\lvert\theta_i\xi^* - \frac{p_i}{2^k}\right\rvert$
1	1	1	0	0.2010000000										
2	1	2	0	0.1140000000										
3	2	5	1	0.0490000000										
4	5	9	2	0.0245000000	1	3	7	5	0.00968523	1.0593220339	5	10	2	0.0040000000
5	10	19	4	0.0135000000	2	3	19	8	0.00548089	0.8470042796	8	16	3	0.0038421053
6	19	38	7	0.0067500000										
7	38	75	15	0.0031875000										
8	77	150	29	0.0017812500	1	3	76	55	0.00028673	0.7188256659	55	108	21	0.0001184211
9	153	301	58	0.0008906250										
10	306	601	117	0.0002578125	1	3	413	299	0	0.9765625000	299	587	114	0
11	612	1202	233	0.0002304688										

Here again, we see that scaled approximations turn out to be significantly more precise. For example, scaled approximation (with $k = 4$):

$$\alpha\xi \approx 5/16\,, \quad \beta\xi \approx 10/16\,, \quad \gamma\xi \approx 2/16\,,$$

produces an error which is more than 6 times smaller (0.004 vs 0.0245) than one of the non-scaled approximation. Such error is already small enough, such that it could be useful for some applications. Moreover, the factors produced by such scaled approximations are remarkably easy to use for multiplier-less computations. Thus the computation of luminance can be accomplished by using just 3 additions, as follows:

$$Y = \alpha R + \beta G + \gamma B \quad \rightsquigarrow \quad \left[\begin{array}{rcl} x & = & G + (R \gg 1); \\ y & = & (x + B) \gg 3; \\ Y' & = & y + (x \gg 1); \end{array}\right.$$

where $Y' \approx Y\xi$, x, y are temporary variables, and $*$, $\gg$ are multiplication and binary right shift operators correspondingly.

Table 3 also offers several higher precision solutions, and one that is absolutely precise:

$$\alpha\xi = 299/1024\,, \quad \beta\xi = 587/1024\,, \quad \gamma\xi = 114/1024\,.$$

This exact solution is here due to the fact that our transform factors are actually rational numbers (decimal fractions converted to dyadic by scale factor $\xi = 1000/1024$).

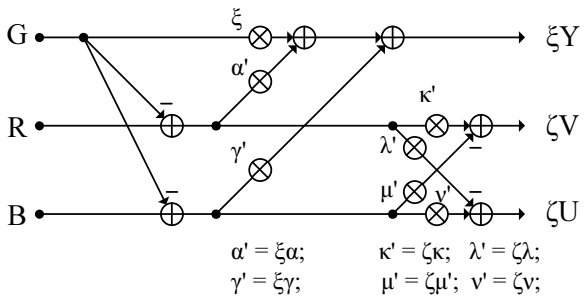

Figure 3. Color transforms with added scale-factors (parameters ξ, ζ).

4.1 Framework for design of scaled RGB↔YUV transforms

In Figure 4 we show a generalized flow-graph of an RGB-to-YUV-type transform, where we have introduced two scale factors:

ξ – affecting the luminance (Y),

ζ – affecting the chrominance components $(U, V), (Cb, Cr)$, etc.

Inside the flow-graph, the scale factor ξ is applied to a triple of factors $(1, \alpha, \gamma)$, and the scale factor ζ is applied to 4 factors $\kappa, \lambda, \mu, \nu$. We note, that in some transforms, such as YCbCr and YPbPr, chrominance constants κ and ν are identical, so the number of distinct constants affected by ζ in these cases is also 3.

We show the approximations of luminance and chrominance-related groups of factors for RGB→YCbCr/YPbPr transforms in Tables 4 and 5 correspondingly. It can be seen that both luminance and chrominance groups can be approximated with approximately $2e^{-3}$ error by using just 4 or 5 fixed-point mantissa bits (parameter k)! Same level of precision with non-scaled dyadic approximations is reachable by using 8 or more bits.

Furthermore, it can be seen that the reduction in bit-width of factors also translates in simpler factorizations of multiplications. For example, by using $\alpha' \approx 1/4, \gamma' \approx 3/32, \xi \approx 27/32$ approximations for luminance factors, and $\lambda' \approx 1/8, \mu' \approx 1/16, \kappa' = \nu' \approx 3/8$ for chrominance, we arrive at the algorithm for computing full scaled transform:

$$\begin{aligned} r1 &= R - G; \\ b1 &= B - G; \\ g1 &= G + (G \ll 3); \\ y1 &= g1 + b1; \\ Y' &= (y1 - (y1 \gg 2) + r1) \gg 3; \\ V' &= (r1 + (r1 \ll 1) - (b1 \gg 1)) \gg 3; \\ U' &= (b1 + (b1 \ll 1) - r1) \gg 3; \end{aligned}$$

that uses only 10 additions. Here $r1, b1, g1, y1$ are intermediate quantities, and $Y' \approx Y\xi$, $V' \approx V\zeta$, and $U' = U\zeta$.

This described scaled factorization and approximation technique can be easily applied to derive many other useful transforms (such as ones for YUV, YIQ, YDbDr, YSbSr, and other color spaces).

5. CONCLUSION

We have proposed and studied a technique for improving implementations of color transforms by introduction of scale factors. We have shown that this technique is related to one of finding simultaneous Diophantine approximations, and have shown how it can be solved and used to produce efficient implementations of color transforms.

Table 4. Approximations of luminance group of constants: $\theta_1 = \alpha = 0.299$, $\theta_2 = \gamma = 0.114$, and $\theta_3 = 1.0$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k$, $\theta_2 \approx p_2/2^k$, $\theta_3 \approx p_3/2^k$					Associated rational appr-s: $\theta^*_{ij} = \theta_i/(\theta_i+\theta_j) \approx p/q$					Scaled dyadic approximations: $\theta_1\xi^* \approx p_1/2^k$, $\theta_2\xi^* \approx p_2/2^k$, $\theta_3\xi^* \approx p_3/2^k$				
k	p_1	p_2	p_3	$\max_i \left\lvert\theta_i - \frac{p_i}{2^k}\right\rvert$	i	j	q	p	$\left\lvert\theta^*_{ij} - \frac{p}{q}\right\rvert$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_i+\theta_j}$	p_1	p_2	p_3	$\frac{1}{\xi^*}\max_i \left\lvert\theta_i\xi^* - \frac{p_i}{2^k}\right\rvert$
1	1	0	2	0.2010000000										
2	1	0	4	0.1140000000										
3	2	1	8	0.0490000000										
4	5	2	16	0.0135000000	0	1	7	5	0.0096852300	1.0593220339	5	2	17	0.0040000000
5	10	4	32	0.0135000000	1	2	30	3	0.0023339318	0.8415619390	8	3	27	0.0026000000
6	19	7	64	0.0046250000										
7	38	15	128	0.0031875000										
8	77	29	256	0.0017812500	0	1	76	55	0.0002867338	0.7188256659	55	21	184	0.0001184211
9	153	58	512	0.0007187500										
10	306	117	1024	0.0002578125	0	1	413	299	0	0.9765625000	299	114	1000	0
11	612	233	2048	0.0002304688										

Table 5. Approximations of chrominance group of constants (in RGB→YPbPr/YCbCr transforms): $\theta_1 = \lambda \approx 0.1687358916$, $\theta_2 = \mu \approx 0.0813124108$, and $\theta_3 = \mu = 0.5$.

Direct dyadic approximations: $\theta_1 \approx p_1/2^k$, $\theta_2 \approx p_2/2^k$, $\theta_3 \approx p_3/2^k$					Associated rational appr-s: $\theta^*_{ij} = \theta_i/(\theta_i+\theta_j) \approx p/q$					Scaled dyadic approximations: $\theta_1\xi^* \approx p_1/2^k$, $\theta_2\xi^* \approx p_2/2^k$, $\theta_3\xi^* \approx p_3/2^k$				
k	p_1	p_2	p_3	$\max_i \left\lvert\theta_i - \frac{p_i}{2^k}\right\rvert$	i	j	q	p	$\left\lvert\theta^*_{ij} - \frac{p}{q}\right\rvert$	$\xi^* = \frac{1}{2^k}\frac{q}{\theta_i+\theta_j}$	p_1	p_2	p_3	$\frac{1}{\xi^*}\max_i \left\lvert\theta_i\xi^* - \frac{p_i}{2^k}\right\rvert$
1	0	0	1	0.1687358916										
2	1	0	2	0.0813124108										
3	1	1	4	0.0437358916										
4	3	1	8	0.0188124108	0	1	3	2	0.0081465193	0.7498551205	2	1	6	0.0020370233
5	5	3	16	0.0124858916										
6	11	5	32	0.0031874108										
7	22	10	64	0.0031874108	1	2	93	13	0.0000923544	1.2498657975	27	13	80	0.0000536867
8	43	21	128	0.0007671416										
9	86	42	256	0.0007671416										
10	173	83	512	0.0002577233										
11	346	167	1024	0.0002305579										

REFERENCES

1. A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
2. J. W. S. Cassels, *An Introduction to Diophantine Approximations*, Cambridge University Press, 1957.
3. M. Gröetschel, L. Lovácz, and A. Schrijver, *Geometric algorithms and combinatorial optimization*, (Springer, Berlin 1988)
4. Y. A. Reznik, A. T. Hinds, L. Yu, Z. Ni, and C-X. Zhang, Efficient fixed-point approximations of the 8x8 inverse discrete cosine transform, Proceedings of SPIE – Volume 6696, Applications of Digital Image Processing, Sep. 24, 2007.
5. Y. A. Reznik, A. T. Hinds, and J. L. Mitchell, Improved Precision of Fixed-Point Algorithms by Means of Common Factors, IEEE Int. Conf. Image Processing (ICIP'08), – to appear.
6. H.M. Kim, W-S.Kim, and D-S. Cho, A new color transform for RGB coding, IEEE Int. Conf. Image Processing (ICIP'04), vol.1, pp. 107-110, 24-27 Oct. 2004.