# AN ADAPTIVE COLOR TRANSFORM APPROACH AND ITS APPLICATION IN 4:4:4 VIDEO CODING

*Detlev Marpe, Heiner Kirchhoffer, Valeri George, Peter Kauff, and Thomas Wiegand*

Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany
phone: +493031002-0, fax: +49303927200, email: [marpe|kirchhoffer|george|kauff|wiegand]@hhi.de
web: ip.hhi.de

## ABSTRACT

*This paper deals with an approach that extends block-based video-coding techniques by an adaptive color space transform. The presented technique allows the encoder to switch between several given color space representations with the objective to maximize the overall rate-distortion gain. Simulations based on the current draft of the H.264/MPEG4-AVC 4:4:4 extensions demonstrate that our technique guarantees a rate-distortion performance equal or better than that obtained when using any of the individual color spaces only.*

## 1. INTRODUCTION

Currently, research in the area of video coding specifically related to the 4:4:4 color sampling format is attracting a lot of interest both from academia and industry. In contrast to typical consumer applications, many high-quality video applications such as professional digital video recording or digital cinema / large-screen digital imagery are requiring all three color components to be represented with identical spatial resolution. Moreover, for this kind of applications sample values in each color component of a video signal are expected to be captured and displayed with a precision of more than 8 bits. These specific characteristics are posing new questions and new challenges, especially regarding the choice of an optimal color space representation.

Typically, both for video capture and display purposes the *RGB* (red, green, and blue) color space representation can be considered as the natural choice. From a coding point-of-view, however, the *RGB* domain is often not the optimum color space representation, mainly because for natural source material usually a quite significant amount of statistical dependencies between the *RGB* components can be observed. Thus, in order to take advantage of these statistical properties, the usage of a decorrelating transformation from the original *RGB* domain to some appropriate color space is often recommended. Several such color transforms are specified by standardization bodies like, e.g., ITU or SMPTE with the most relevant of them referenced in Annex E of the H.264/MPEG4-AVC video coding standard [1]. The corresponding color spaces are usually denoted by *YCbCr* to indicate conversion to a representation consisting of one luma (*Y*) and two color difference (*Cb,Cr*) components.

Apart from the decorrelating properties, *RGB*-to-*YCbCr* color transforms are also beneficial in distinguishing between what the human visual system perceives as brightness and as color attributes. Coding in the 4:2:0 color format, as it is mostly used in consumer applications, typically exploits these effects by first transforming to *YCbCr* and then subsampling the resulting color difference (i.e., chroma) components both horizontally and vertically by a factor of 2. In addition, existing video coding standards such as H.264/MPEG4-AVC often provide only restricted tool sets for predictive coding of chroma components relative to those available for the luma component [1].

In high-fidelity video applications, such as, for instance, currently addressed by a new standardization activity within H.264/MPEG4-AVC towards the specification of so-called Advanced 4:4:4 profiles [2], more care has to be taken to ensure a minimum level of distortion across all color components at a given bit rate. As one consequence of these increased requirements, the current draft of the Advanced 4:4:4 coding architecture [3] specifies all three input components to be treated exactly in the same way as the luma component in the existing High profiles of the H.264/MPEG4-AVC standard. Due to the lack of sufficient understanding of the potential merits of color transforms with respect to the currently addressed application space [4], it was decided to not include any specific color space transform in the initial draft [3] of the Advanced 4:4:4 profiles for the time being.

Interestingly, the same coding results that were presented in [4] for motivating the new H.264/MPEG4-AVC 4:4:4 profiles have also clearly shown that the effectiveness of a decorrelating color transform in terms of rate-distortion (R-D) performance highly depends on the specific coding conditions as well as the given source characteristics. In many cases, however, both the characteristics of the source and the coding conditions are not known in advance and may also change with respect to the spatial and temporal domain. Consequently, any fixed, a priori chosen color representation may result in a suboptimal R-D behavior.

As a solution to this problem, we propose a technique which allows to adapt the color space representation of a 4:4:4 formatted video signal on a block-by-block basis. We further propose to apply this adaptive color space transform to the spatially transformed prediction residual only which saves computational cost as well as memory bandwidth and which, in addition, enables us to apply the color space transform to selected low frequency components only. In this way, it is possible to exclude high frequency components

from the application of the color transform in case the given video signal contains a substantial amount of uncorrelated noise across the *RGB* components (as it is, e.g., often the case for film-scanned video sources).

The next section highlights the background of our work and describes the underlying problem with a particular focus on the current standardization of H.264/MPEG4-AVC Advanced 4:4:4 profiles. Section 3 then presents the concept of our new adaptive color space transform. In Section 4, finally some representative simulation results are given for demonstrating the R-D gain of our proposed approach.

## 2. BACKGROUND AND PROBLEM STATEMENT

From a theoretical point-of-view, the Karhunen-Loève (KL) transform can be considered as the R-D optimal transform. This has been proven at least for the case of Gaussian sources within the settings of high-resolution quantization theory [5]. The corresponding KL basis functions are given as the eigenvectors of the covariance matrix. Applied to the problem of finding an R-D optimal color space representation for a given set of image or video signals, the task would then consist in first estimating the 3×3 covariance matrix of the three color random variables *R*, *G*, and *B* drawn from a given representative set. By diagonalizing the covariance matrix, the resulting KL basis functions would then represent the optimum color representation – at least in principle and without further considering the interdependencies between the color transform and the subsequent coding process.

In practice, however, the KL transform is rarely used due to its excessive computational costs. Also, the assumption of dealing with a Gaussian process is seldom fulfilled. Real measured probability distributions are often highly non-uniform and non-stationary, as will be discussed further below. But even if all conditions for guaranteeing optimality of the KL representation are met, the KL transform often just serves as a benchmark for the performance of other custom-built or application-specific transform types.

Recently, a low-complexity, reversible color transform has been proposed in the context of the H.264/MPEG4-AVC standardization [6]. This color transform maps *RGB* to the so-called *YCoCg* color space, and it has some remarkable properties. First, this color transform has been shown to be capable of achieving a decorrelation that is much better than that obtained by various *RGB*-to-*YCbCr* transforms and which, in fact, is very close to that of the KL transform (at least, when measured for a representative set of high-quality *RGB* test images) [6]. Secondly, the transform is reversible in the sense that each original *RGB* triple can be exactly recovered from the corresponding *YCoCg* triple if the color difference components *Co* and *Cg* are represented with one additional bit accuracy relative to the bit depth used for representing *RGB,* and if furthermore, no information loss in any subsequent coding step is assumed. Thirdly and finally, both the forward and inverse *RGB*-to-*YCoCg* transform require only a few shift and add operations per triple which, in addition, can be performed inline, i.e., without the need of some extra memory apart from one single auxiliary register:

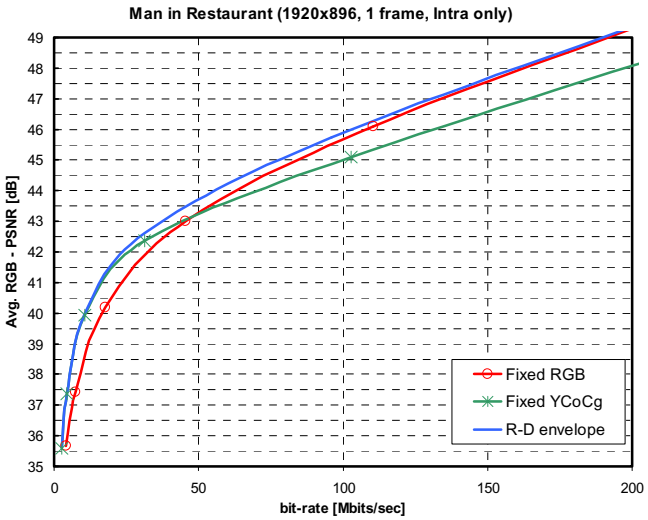


**Fig. 1 -** R-D curves for encoding a single picture of a film-scanned test sequence in two fixed color space representations together with the corresponding R-D envelope (in blue, partly covering the R-D curves for *RGB* and *YCoCg*). Note that the distortion D is measured as the mean of *R*, *G*, and *B* PSNR values.

$$\begin{aligned} Co &= R - B \\ t &= B + (Co \gg 1) \\ Cg &= G - t \\ Y &= t + (Cg \gg 1) \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} t &= Y - (Cg \gg 1) \\ G &= Cg + t \\ B &= t - (Co \gg 1) \\ R &= B + Co \end{aligned} \qquad (1)$$

The "≫"-operator in (1) denotes the bitwise right shift operator. Note that the transform steps have to be performed in the order from top to bottom in (1) to guarantee that the memory locations for *G*, *B*, and *R* can be re-used for *Y*, *Cg*, and *Co*, respectively, and vice versa.

In order to evaluate the R-D performance of the *RGB*-to-*YCoCg* color transform for the intended high-quality 4:4:4 video coding applications, we tested a variety of natural video material ranging from film-scanned sequences to film material that was directly captured from a high-quality 3-CCD camera [7]. For these coding simulations, we have used an implementation of the current draft Advanced 4:4:4 profiles [3], driven in intra-only coding mode. Fig. 1 exemplifies the characteristic output of our experiments, where the red and green curves represent the R-D performance for encoding the same source in the *RGB* and *YCoCg* domain, respectively. As can be seen from Fig. 1, at relatively low bit rates the *YCoCg* representation performs significantly better than the corresponding *RGB* representation, whereas *RGB*-based encoding leads to an increasingly better performance when moving towards higher bit rates.

This phenomenon can be mainly attributed to the fact that most natural video sources contain a relatively high amount of signal-independent, uncorrelated noise with varying noise power over all three primary channels. Typically, the blue (*B*) channel exhibits the most dominant noise signal which, when transformed to the *YCoCg* domain, gets spread over all three resulting components. As a result, the overall noise power in the *YCoCg* representation is increased relative to that in the *RGB* domain. This, in turn, leads to the observed degradation of coding efficiency in the medium to

high bit-rate range where more and more noise components typically are supposed to survive the quantization process.

As a further consequence, a cross-over region can be observed, indicating a suboptimal R-D performance of both fixed alternative color space representations relative to the R-D envelope, as illustrated by the blue curve in Fig. 1. Note that in either case, for encoding this sample picture in a single color space representation, one can only move along one or the other R-D curve.

In general, the effectiveness of a color transform in terms of decorrelation will also depend on other source characteristics such as, e.g., given by the degree of color saturation or the amount of texture information and its alignment across the *R*, *G*, and *B* components. Typically, this kind of statistical properties are often subject to changes both within a given picture and from picture to picture.

Based on these observations, the fundamental problem to be addressed in this paper is how to generate a color space representation (out of a given variety) that always guarantees an optimal R-D performance in 4:4:4 video coding scenarios, independently from the chosen coding conditions and the specific source characteristics.

## 3. ADAPTIVE COLOR SPACE TRANSFORM

### 3.1 Basic idea

Our present approach is based on the finding that a picture or a series of pictures can be encoded more efficiently when each picture is partitioned into smaller regions such that each region carries the picture information either in a primary color space representation (e.g., *RGB*) or in one of several secondary color space representations (e.g., *YCoCg*). In this way, the color transform such as, e.g., *RGB*-to-*YCoCg*, can be applied in a spatially adaptive way by taking into account the varying statistical properties of the source. The decision which color space representation to use is made by the encoder and signaled to the decoder as side information.

### 3.2 Simple in-loop decorrelating transform

Additionally to the *YCoCg* color space we further utilize a very simple transform from *RGB* to another "color" space that we have abbreviated by $Gr_Br_R$ and that is defined by the following relation:

$$\begin{bmatrix} G \\ r_B \\ r_R \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} G \\ B \\ R \end{bmatrix} \qquad (2)$$

According to this transform, *G* is not modified and acts as a predictor for *B* and *R*. Because of the vanishing off-diagonal elements in the second and third column of the transform matrix in (2), potential noise components cannot be propagated from the blue or red channel to any of the other channels. Furthermore, since the matrix on the right hand side of (2) is a (lower) triangular matrix, it can be applied to a triplet of primaries containing the quantized *G* component instead of the original *G* signal. This "closed-loop" processing enables the encoder to use exactly the same quantized *G* signal for prediction (of *B* and *R*) as the decoder which, in turn, has the advantage that the quantization noise of *G* will not be propagated to the transformed $r_B$ and $r_R$ components.

### 3.3 Frequency-selective color transform

Typically, for high-frequency spatial transform coefficients the power of the noise signal as resulting, e.g., from film grain or from a CCD, relative to the power of the wanted signal can expected to be quite dominant. Thus, to prevent the propagation of those noisy coefficients from the G component to the corresponding coefficients of the $r_B$ and $r_R$ components, we introduce a frequency-selective application of the corresponding color transform by restricting its application to spatial low frequency coefficients in a way as further described below.

### 3.4 Integration into H.264/MPEG4-AVC

Our proposed approach has been integrated into the current draft design of H.264/MPEG4-AVC Advanced 4:4:4 profiles. By doing so, several design decisions with respect to a specific incarnation of the rather general idea of an adaptive color space transform, as described above, have been adjusted to the given environment. First, the partition size has been chosen to be aligned with the fixed macroblock (MB) partitioning (of 16×16 samples for each color component) in H.264/MPEG4-AVC. This choice was also motivated by the desire to achieve a reasonable trade-off between the additional side information needed to indicate the choice of the color space and the coding gain obtained by spatially adapting the color space representation in a most accurate way.

The second important design decision is related to the specific application of the color space transform itself. Mathematically, there is no difference in applying a color transform before forming and applying the prediction, or vice versa, provided the same (linear) prediction operator is applied to all three color components[1], as it is the case in the current draft H.264/MPEG4-AVC 4:4:4 design [3]. However, operating on the original domain would imply that the color transform has to be applied to the reference signal as well, whenever required by the encoder's choice. This, together with the need for an increased bit depth in the *Co* and *Cg* or $r_B$ and $r_R$ representation, would have led to a considerable overhead in computational cost as well as memory bandwidth when compared to the relatively resource-friendly way of applying the color space transform to the prediction residual.

Consequently, we decided to operate on the residual signal and dubbed our approach the *macroblock-adaptive residual color transform* (MB-ARCT). Furthermore, in our proposed MB-ARCT approach we have reversed the order of application of spatial transforms and the proposed adaptive color space transform. This commutation is necessary to accomplish the frequency-selective application of the *RGB*-to-$Gr_Br_R$ color space transform.

[1] In a mathematically strict sense this is, however, only true if any rounding or clipping operations are neglected.

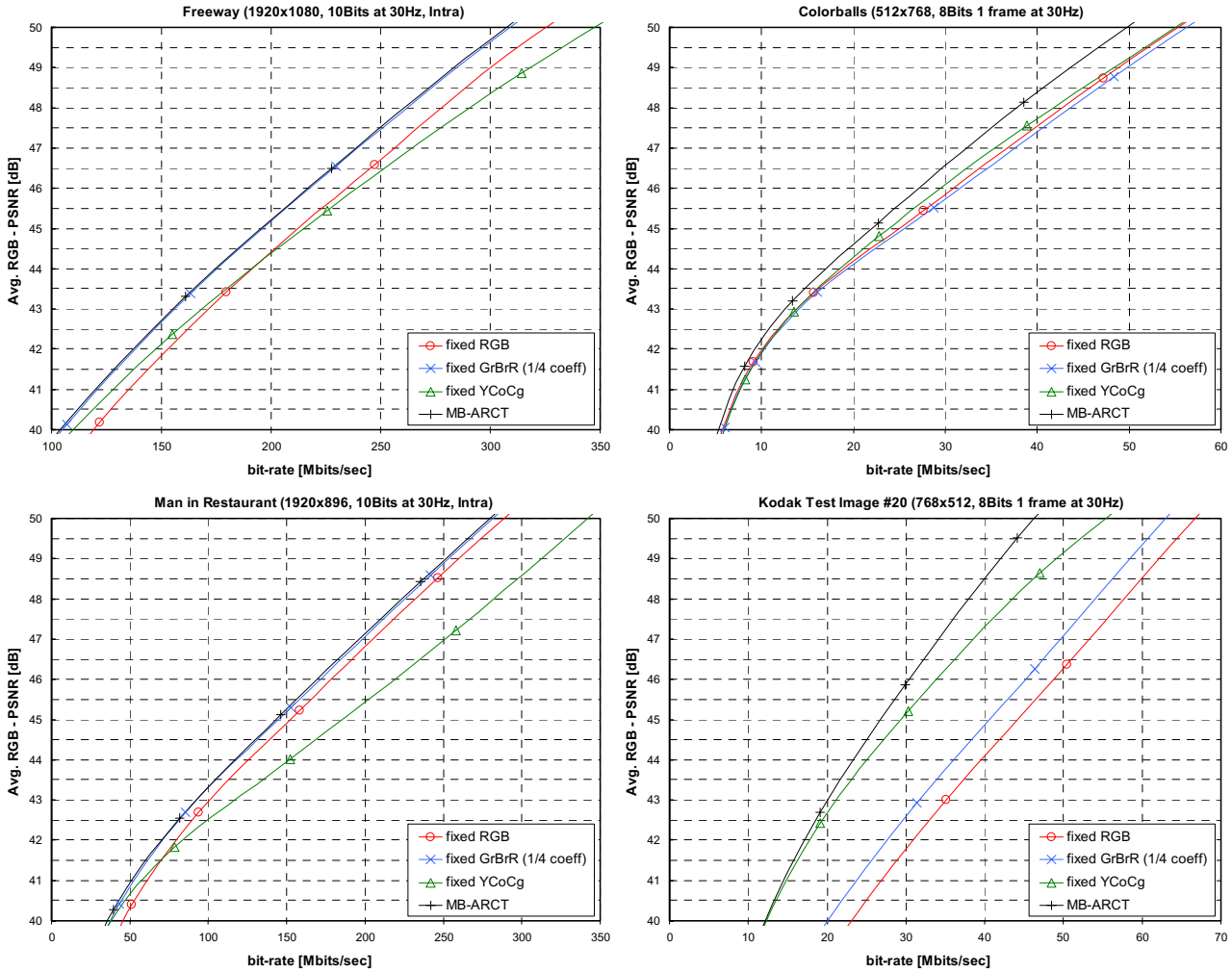


**Fig. 2** - R-D curves comparing fixed *RGB*, $Gr_Br_R$, and *YCoCg* encoding with the proposed MB-ARCT method. "1/4 coeff" denotes the proposed frequency selective application of the $Gr_Br_R$ transform.

For achieving the desired amount of adaptivity, our simulations have shown that it is sufficient to choose between the following three color space representations per macroblock:

- *RGB*: no coefficient of the corresponding macroblock is color transformed.
- *YCoCg*: all coefficients of the corresponding macroblock are transformed to *YCoCg* according to (1).
- $Gr_Br_R$: only one quarter of coefficients along the zig-zag scan, starting from the DC in a spatially transformed block are transformed to $Gr_Br_R$ along (2).

The choice of the encoder is signaled by means of two separate flags at the macroblock layer. This implies a maximum overhead of 2 bits/MB with a considerably lower effective rate, especially in case of using the CABAC entropy coding mode [1]. The modeling part of CABAC has been extended to include two additional adaptive probability models for coding those flags.

### 3.5 R-D optimized color space selection process

In our specific encoder implementation, the MB-adaptive choice of the residual color transform is performed as a straightforward extension of the usual mode decision process based on a Lagrangian cost function $J = D + \lambda R$ [8]. For each given macroblock, the color space selection process proceeds as follows. First, for each of the three color space representations of the prediction residual, a candidate prediction mode is selected along the same strategy as described in [8]. Then, the underlying color space of the candidate with the lowest overall Lagrangian cost $J$ is selected. Note that for the distortion measurement $D$ of the corresponding Lagrangian $J$, the averaged sum of squared differences over all three components in the primary color representation (*RGB*) has always been used.

## 4. EXPERIMENTAL RESULTS

We implemented our proposed approach as described in the previous section on top of an implementation of the current draft Advanced 4:4:4 profiles [3]. In Fig. 2, coding results are shown comparing the R-D performance of our proposed MB-ARCT method including the frequency-selective $Gr_Br_R$ transform with that of using fixed color space representations. The upper and lower left charts of Fig. 2 depict the R-D graphs for intra-only encoding the Viper test sequence "Freeway" and the film-scanned sequence "Man in Restau-

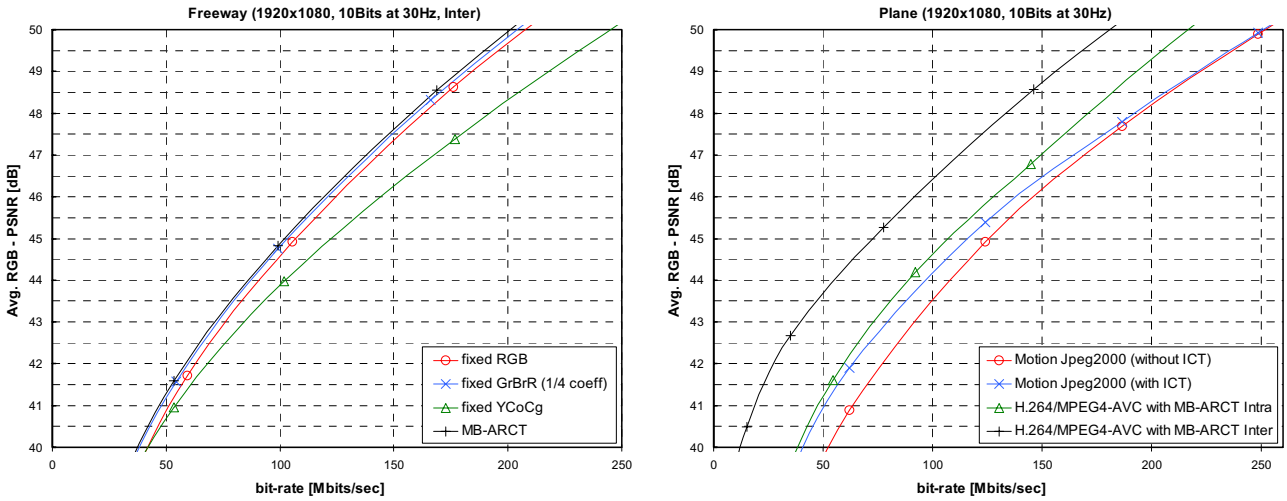


**Fig. 3** – Left: R-D curves comparing fixed *RGB*, $Gr_Br_R$, and *YCoCg* encoding with the MB-ARCT method for the inter-case. Right: R-D comparison of H.264/MPEG4-AVC using MB-ARCT and Motion JPEG2000.

rant", respectively. Typically, for encoding this kind of video material in view of the prospective target applications, objective reconstruction qualities of around 45 dB average *RGB* PSNR will be required to achieve a visually transparent quality. At this target range, however, we often observe the critical cross-over of the R-D curves for encoding in the fixed *RGB* and *YCoCg* domain, as can be seen from Fig. 1 and Fig. 2 (left). By using our proposed adaptive color space transform, it is guaranteed to always achieve the optimum R-D performance relative to the alternative of using one of the fixed color spaces. In fact, for that kind of video material the frequency-selective transform to $Gr_Br_R$ has contributed most to the overall gain in the range of 0.5-1.0 dB avg. *RGB* PSNR relative to the fixed *YCoCg* or *RGB* representation.

The R-D graphs on the right hand side of Fig. 2 demonstrate intra-coding results for two particular still images. For those images, MB-ARCT actually achieved an overall R-D gain compared to the best performing fixed color space representations due to its ability of spatially adapting the residual color representation to the varying color characteristics of the source. As can be seen from the corresponding graphs, substantial R-D gains of up to 1.5 dB average *RGB* PSNR have been observed in favor of MB-ARCT.

Fig. 3 show some representative simulation results for the *inter* case, *i.e.*, the case of using motion-compensated prediction with both *P* and *B* pictures. We observed a similar R-D behavior as in the intra-only case with the general trend of smaller overall R-D gains, as exemplified in Fig. 3 (left) for the "Freeway" sequence. An R-D performance comparison of our MB-ARCT enhanced H.264/MPEG4-AVC implementation with Motion-JPEG2000 [9] was made as well by using an JPEG2000 software implementation [10]. Fig. 3 (right) illustrates a sample of the corresponding coding results. As can be seen from that R-D graph, H.264/MPEG4-AVC intra-only coding with MB-ARCT performs better than Motion-JPEG2000, even if the latter is using the JPEG2000-specific irreversible color transform (ICT). In most cases, H.264/MPEG4-AVC inter coding shows a substantially improved R-D performance compared to H.264/MPEG4-AVC intra-only coding both with and without using MB-ARCT, as illustrated for the former case in Fig. 3 (right).

## 5. CONCLUSIONS

We presented a relatively simple but rather efficient approach to resolve the critical issue of selecting an appropriate color space representation in compression applications considered to be addressed by the current standardization of H.264/MPEG4-AVC Advanced 4:4:4 profiles. For that purpose, we introduced an adaptive approach for selecting between a number of given color space representations for the prediction residual on a macroblock-by-macroblock basis. Our coding results have shown that by using the proposed macroblock-adaptive residual color transform approach, it is ensured to always achieve the same or better coding efficiency than by using any of the fixed color space representations. In particular, we have demonstrated that by utilizing the frequency-selective *RGB*-to-$Gr_Br_R$ transform in addition to the *RGB* and *YCoCg* representations, remarkable coding gains can be achieved for some representative video sources.

## REFERENCES


[1] *ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC*, "Advanced Video Coding for Generic Audiovisual Services," Version 3, 2005.

[2] T. Suzuki *et al.*, "Justification for new 4:4:4 video coding profile(s)", *ISO/IEC JTC 1/SC 29/WG 11,* N7313, July 2005.

[3] H. Yu, "Draft Text of H.264/AVC Amendment 2", *ISO/IEC JTC 1/SC 29/WG 11, ITU-T Q6/SG16,* JVT-Q205r1, Oct. 2005.

[4] H. Yu and L. Liu, "Advanced 4:4:4 Profile for MPEG4-Part10/H.264," *ISO/IEC JTC 1/SC 29/WG 11, ITU-T Q6/SG16,* JVT-P017, July 2005.

[5] T. Berger, *Rate Distortion Theory*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[6] H. Malvar and G. J. Sullivan, "YCoCg-R: A Color Space with RGB Reversibility and Low Dynamic Range", *ISO/IEC JTC1/SC 29/WG 11, ITU-T Q6/SG16,* JVT-I014, July 2003.

[7] Thomson Grass Valley Viper FilmStream Camera, online: http://www.thomsongrassvalley.com/products/cameras/viper/

[8] K.-P. Lim, G. Sullivan, and T. Wiegand, "Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods", *ISO/IEC JTC 1/SC 29/WG 11, ITU-T Q6/SG16*, Document JVT-N046r1, Jan. 2005.

[9] *ITU-T Rec. T.802 & ISO/IEC 15444-3 Motion JPEG2000*, 2002.

[10] Kakadu software online: http://www.kakadusoftware.com/