|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  12th Meeting: Geneva, CH, 14–23 January 2013 | Document: JCTVC-L0316-v2 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Updated proposal for frame packing arrangement SEI for 4:4:4 content in 4:2:0 bitstreams** | | |
| *Status:* | Input Document to JCT-VC | | |
| *Purpose:* | Proposal | | |
| *Author(s) or Contact(s):* | Yifu Zhang  Yongjun Wu  Sandeep Kanumuri  Shyam Sadhwani  Gary J. Sullivan  Henrique S. Malvar  One Microsoft Way Redmond WA 98052-6399  USA | Tel: Email: | |  | | --- | | +1-425-703-5308 [yifuzh@microsoft.com](mailto:yifuzh@microsoft.com)  [yongjunw@microsoft.com](mailto:yongjunw@microsoft.com)  [skanumu@microsoft.com](mailto:skanumu@microsoft.com)  [shyams@microsoft.com](mailto:shyams@microsoft.com)  [garysull@microsoft.com](mailto:garysull@microsoft.com)  [malvar@microsoft.com](mailto:malvar@microsoft.com) | |
| *Source:* | Microsoft Corporation | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This contribution proposes a method to extend the use of the frame packing arrangement SEI message to represent 4:4:4 content in nominally 4:2:0 bitstreams. The contribution is an update of the prior contribution JCTVC-K0240 that includes the proposal of an additional content interpretation type (indicating band separation filtering) and provides additional experiment results. With the proposed method, it is reported that one constituent frame (e.g. in a top-bottom packing or alternating-frame coding scheme) can be decoded compatibly as an ordinary 4:2:0 image, or can be supplemented with the data from another constituent frame to form a complete 4:4:4 image representation. It is proposed to include support for the additional scheme into the frame packing arrangement SEI message in both AVC (Rec. ITU-T H.264 | ISO/IEC 14496-10) and HEVC, to facilitate deployment of systems using this method. Since 4:2:0 is the most widely supported format in products, it is asserted that having an effective way of conveying 4:4:4 content through such decoders can provide the substantial benefit of enabling widespread near-term deployment of 4:4:4 capabilities (especially for screen content coding). The proposed method operates by packing the samples of a 4:4:4 frame into two 4:2:0 frames and encoding the two 4:2:0 frames as the constituent frames of a frame packing arrangement. The content\_interpretation\_type is extended to signal this packing arrangement. The proposed scheme is asserted to be of high practical value for applications involving screen content. Relative to native 4:4:4 encoding, the proposed scheme can provide the advantage of compatibility with the ordinary 4:2:0 decoding process that is expected to be more widely supported in decoding products.

# Introduction

Most video codecs that are commercially available today support only the 4:2:0 format, which sub-samples the chroma resolution, as opposed to using a 4:4:4 format, in which the chroma information is represented at the same resolution used for the luma. The YUV 4:2:0 format is considered sufficient and efficient for "mainstream" content (i.e. most natural and animated video content) since users do not ordinarily see a perceptible difference between the two formats for such content. However, there are a variety of existing and emerging applications (e.g., as discussed in [1] and [2]) that operate with screen content, and for such content, the difference between these two formats can be easily perceived (see Appendix). In such applications, a 4:4:4 format is strongly preferred over the YUV 4:2:0 format. However, the lack of wide-spread support for video codecs supporting 4:4:4 formats is a hindrance for these applications. Moreover, for certain critical uses such as scrolling titles and hard-edged graphics, there may sometimes be a significant benefit for 4:4:4 use in other scenarios.

In this contribution, we propose an approach to use codecs designed for YUV 4:2:0 content to compress and represent 4:4:4 content through the use of frame packing. The contribution is an update of the prior contribution JCTVC-K0240 [2] that includes the proposal of an additional content interpretation type (indicating band separation filtering) and provides additional experiment results. This proposal is similar to the frame packing of stereo (3D) content into 2D images, and builds on the framework established for that by extending the semantics of the frame packing arrangement SEI message. Unlike the frame packing of stereo content wherein there is a left and right view, the frame packing of the 4:4:4 content is done using a main view and an auxiliary view. Both the main and auxiliary views are in an equivalent of a 4:2:0 format. The main view may be independently useful, while the auxiliary view is useful when interpreted appropriately together with the main view.

# Proposed modifications

## Packing a YUV444 frame into main and auxiliary views

A YUV (YCbCr, YCoCg, GBR, etc.) 4:4:4 frame can be represented as follows, wherein Y444, U444, and V444 are the Y, U, and V planes comprising the YUV444 frame. Let the resolution of these planes be represented by width and height .

Y444

U444

V444

The YUV444 frame represented above can be packed into two YUV420 frames (as main and auxiliary frames) as follows:

Y420 U420 V420

Main view

B1

B2

B3

(YUV420 frame)

B6

B7

B8

B9

B4

B5

Auxiliary view

(YUV420 frame)

The areas marked as B1 to B9 make up the Y, U and V planes of the two YUV420 frames representing the main and auxiliary views. These areas are related to Y444, U444, and V444 as follows:

**Main view**

Area B1, , where the range of is .

Area B2, , where the range of is .

Area B3, , where the range of is .

**Auxiliary view**

Area B4, , where the range of is .

Area B5, , where the range of is .

Area B6, , where the range of is .

Area B7, , where the range of is .

Area B8, , where the range of is .

Area B9, , where the range of is .

In the above equations, and are either the same as or represent filtered versions of and respectively, where the range of is . This choice is explained in more detail in section 2.4.

### Advantages

The proposed packing method is designed such that:

* The main view is a YUV420 equivalent of the original YUV444 frame
  + Systems can optionally just display just the main view if YUV420 output is needed
* The auxiliary view fits the content model of a YUV420 frame and is well suited for compression in this manner, in terms of
  + Geometric consistency across its Y, U and V components
  + Motion is highly correlated across its Y, U and V components

The packing method is illustrated by the following example wherein a YUV444 frame contains a circle represented using gray color (checkerboard pattern) for the Y plane, blue color (horizontal lines) for the U plane and red color (vertical lines) for the V plane and how the resultant main and auxiliary views are formed in YUV420 format.

Y444 U444 V444

YUV444 frame

Y420 U420 V420

Main view

(YUV420 frame)

Y420 U420 V420

Auxiliary view

(YUV420 frame)

## Extension to frame packing arrangement SEI message

The proposed method to signal the frame packing of YUV444 content involves extending the frame packing arrangement SEI message as defined in the AVC [3] and HEVC [4] specifications.

In particular, the table representing the semantics of the syntax element '**content\_interpretation\_type**' is proposed to be extended as shown in green below:

Table D‑9 – Definition of content\_interpretation\_type

|  |  |
| --- | --- |
| **Value** | **Interpretation** |
| 0 | Unspecified relationship between the frame packed constituent frames |
| 1 | Indicates that the two constituent frames form the left and right views of a stereo view scene, with frame 0 being associated with the left view and frame 1 being associated with the right view |
| 2 | Indicates that the two constituent frames form the right and left views of a stereo view scene, with frame 0 being associated with the right view and frame 1 being associated with the left view |
| 3 | Indicates that the two constituent frames form the main and auxiliary views (4:2:0 frames) representing a 4:4:4 frame, with frame 0 being associated with the main view and frame 1 being associated with the auxiliary view. Indicates that the chroma samples of frame 0 should be interpreted as unfiltered samples of the 4:4:4 frame (without anti-alias filtering). |
| 4 | Indicates that the two constituent frames form the main and auxiliary views (4:2:0 frames) representing a 4:4:4 frame, with frame 0 being associated with the main view and frame 1 being associated with the auxiliary view. Indicates that the chroma samples of frame 0 should be interpreted as having been anti-alias filtered prior to frame packing. |
| 5 | Indicates that the two constituent frames form the main and auxiliary views (4:2:0 frames) representing a 4:4:4 frame, with frame 0 being associated with the main view and frame 1 being associated with the auxiliary view. Indicates that the chroma samples of frame 0 should be interpreted as having been band-separation filtered prior to frame packing. |

It is also proposed to establish the following constraints:

1. When content\_interpretation\_type is equal to 3, 4, or 5, the following syntax elements are proposed to be required to be set equal to 0:
   * quincunx\_sampling\_flag
   * spatial\_flipping\_flag
   * frame0\_grid\_position\_x
   * frame0\_grid\_position\_y
   * frame1\_grid\_position\_x
   * frame1\_grid\_position\_y
2. When content\_interpretation\_type is equal to 3, the following syntax elements should be required to be set as follows since these values represent the correct location type for chroma in the main view in this case:
   * chroma\_loc\_info\_present\_flag shall be equal to 1.
   * chroma\_sample\_loc\_type\_top\_field and chroma\_sample\_loc\_type\_bottom\_field shall be equal to 2.
3. When content\_interpretation\_type is equal to 5, the following syntax elements should be required to be set as follows since these values represent the correct location type for chroma in the main view in this case:
   * chroma\_loc\_info\_present\_flag shall be equal to 1.
   * chroma\_sample\_loc\_type\_top\_field and chroma\_sample\_loc\_type\_bottom\_field shall be equal to 1.

## System operation at the decoding end

The proposed extension to frame packing arrangement SEI message informs the decoder that the decoded pictures contain main and auxiliary views of a 4:4:4 frame as the constituent frames of the frame packing arrangement. This information can be used to process the main and auxiliary views appropriately for display or other purposes.

When the system at the decoding end desires the video in 4:4:4 format and is capable of reconstructing the 4:4:4 frames from the main and auxiliary views, it should do so and the output format should be 4:4:4. Otherwise, only the main view should be given as output and the output format will then be 4:2:0.

## Pre-processing and post-processing (informative)

When content\_interpretation\_type is set to 3, the indication is that none of the chroma samples underwent an anti-alias filtering operation during the process of frame packing i.e. and =. In such a case, the chroma samples comprising the main view are a result of a direct sub-sampling of the chroma planes representing the 4:4:4 frame. As shown in the appendix, direct sub-sampling without filtering can create aliasing artifacts for certain types of screen content when only the main view is used to generate a 4:2:0 output.

In order to reduce the aliasing artifacts and improve the visual quality for the case where only the main view is used, the content\_interpretation\_type can be set to 4 and the main view can be generated using filtered/pre-processed versions of the 4:4:4 chroma planes. In such a case, it is recommended that the filter choice be made based on the chroma sample grid alignment with luma sample grid (inferred from chroma\_sample\_loc\_type\_top\_field and chroma\_sample\_loc\_type\_bottom\_field). For simplicity, in the case when the chroma sample grid aligns with the luma sample grid for each particular direction (horizontal/vertical), it is sugggested that that the 3-tap filter [0.25 0.5 0.25] be used in that direction. In the chroma sample grid positions are centered between the luma sample positions for a particular direction (horizontal/vertical), then it is suggested that the 2-tap filter [0.5 0.5] be used in that direction. Another possible filter choice for the latter case is [0.125 0.375 0.375 0.125].

For example, if we consider the case where the chroma sample grid is not aligned with the luma sample grid in both horizontal and vertical directions (i.e. when chroma\_sample\_loc\_type\_top\_field and chroma\_sample\_loc\_type\_bottom\_field are equal to 1), the 2-tap filter [0.5 0.5] would be applied in both directions, such that and are obtained as follows:

When pre-processing is used (content\_interpretation\_type set to 4), the main view does not contain samples and but contains their filtered counterparts and . The auxiliary view contains the other chroma samples.

If the decoding system decides to output a 4:4:4 frame, a post-processing step should be applied to estimate the samples , as , from the encoded packed frame. For example, a simple suggested estimation of and would be as follows:

In the proposed form, with content\_interpretation\_type equal to 4 and chroma\_sample\_loc\_type\_top\_field and chroma\_sample\_loc\_type\_bottom\_field equal to 1, with the suggested anti-alias filter of [0.5 0.5], the value would perfectly reconstruct the input values in the absence of quantization error and rounding error. When considering quantization error, using somewhat different values would be advised (e.g., as determined by quantization step-size-dependent cross-correlation analysis).

## Band separation filtering for the auxiliary frame

In the frame packing scheme illustrated in section 2.1, pixel values of and frames are placed directly into (and are directly unpacked from) the auxiliary frames. We thus refer to these schemes as “direct” packing approaches. Alternatively, we can consider the auxiliary frame samples as an enhancement layer signal to be combined with the main frame (or base layer frame) data. The main and auxiliary frame data can formed using low-pass and high-pass band separation filtering, instead of direct sample packing. With this variation, the primary signal energy can be concentrated into the main frame, and arbitrarily low bit rates can be allocated to the supplemental auxiliary frame data that forms the enhancement signal.

Instead of encoding auxiliary frame pixels directly, a two-dimensional, three-band wavelet decomposition can first be applied to and before the actual encoding process. Mathematically, for an array , where *=* or , define the following:

, for .

, for .

, for .

, for .

, for

, for

, for

, for

, for

, for

, for

, for

, for

By the above approach, a typical four-band wavelet decomposition breaks the frame into “LL”, “LH”, “HL” and “HH” subbands (“LL” = low-pass in both vertical and horizontal directions, “LH” = low-pass vertical, high-pass horizontal, and so forth). In our wavelet packing scheme, though, the “HL” and “HH” bands are not created; instead, the vertical high-pass signal is kept at full horizontal resolution, i.e., B2 and B3 are the “LL” bands of and respectively, B4 and B5 are vertical high-pass signals, i.e. a vertical “H” band of and , respectively, B6 and B8 consist of even-numbered rows of the “LH” band of , and B7 and B9 consist of odd-numbered rows the “LH” band of . That way, the decoder would apply the corresponding inverse wavelet operations after decoding the main and auxiliary frames to obtain and pixels. Moreover, an additional vertical band separation can be performed, such that B6 and B8 are an “LHL” and “LHH” band of , and B7 and B9 are an “LHL” and “LHH” band of .

When the auxiliary frames are transmitted at lower bit rates (lower quality relative to the main frame), the chroma information from the main frame ( and ) sets the minimum level of quality for the and reconstruction, and any information from the auxiliary frame is used to improve beyond that minimum quality level. In the case of the “direct” frame packing method, wherein pixels from the auxiliary frame are directly unpacked into and frames, such an approach would cause the chroma pixels obtained from the main frame (3 out of 4) to have a lower quality compared to the chroma pixels obtained from the main frame. However, the band-separation frame packing approach incurs a larger rounding error in the pre-processing steps than the direct frame packing approach because of the additional filtering operations involved (in the absence of bit-depth expansion).

We have experimented with the use of lifting-based band-separation filtering to mitigate the rounding error effects, with clipping to eliminate the bit-depth expansion. However, this introduces a different type of distortion. This topic is further discussed in an upcoming publication [5].

# Experiment results

We tested an end-to-end system for packing a 4:4:4 frame into two 4:2:0 frames, based on Microsoft’s implementation of an AVC software encoder and decoder with a simple “IPPP” (forward-predictive) coding structure. We also conducted similar tests using the HEVC HM 9.0 encoder [6]. Each encoder starts with a 4:4:4 input frame, constructs a 4:2:0 frame with twice the height of the 4:4:4 frame, places the main view in the top half and the auxiliary view in the bottom half of the 4:2:0 frame, and encodes the 4:2:0 frame. This corresponds to the use of the SEI message as described in Section 2.2 in the top-bottom frame packing mode (frame\_packing\_arrangement\_type equal to 4). The decoder decodes the 4:2:0 frame, extracts the main and auxiliary views and reassembles the 4:4:4 frame for output (using to simplify the initial testing).

We tested both the “direct” frame packing approach and one variation of the band-separation frame packing approach both on AVC and HEVC platforms. To show the effectiveness of the scheme on screen content encoding, the test sequences we used were captured computer screen recordings (Windows OS, contents are desktop and file folders, Windows applications, etc.). The tested band separation approach was to use a Haar wavelet (i.e., [1 1] / 2 and [1 −1] / 2 filtering with rounding).

Figure 1 and Figure 2 are the resulting rate-distortion performance comparison between these approaches at different bit rates for the auxiliary frame, using the Microsoft AVC and HEVC HM 9.0 encoders [6], respectively. Each frame is divided into two slices, the first one for the main frame and the second one for the auxiliary. We see that the band-separation frame packing approach performs well at low bit rates for the auxiliary frame but suffers at high bit rates due to rounding error, while the direct frame packing approach works better at high bit rates since it introduces no rounding error.

Figure : Rate-distortion performance comparison between the direct frame packing and band-separation approaches for a fixed main frame bit rate using a Microsoft AVC encoder with a screen content sequence of resolution 1920×1200 and length 57 frames, at 30 fps. Auxiliary frame QP varies from −12 to +4 relative to main frame QP, which in this case is set to 39. The bit rate for the main view is 445 kbps, with a PSNR of 31.3 dB.

Figure : Rate-distortion performance comparison between the direct frame packing and band-separation approaches for a fixed main frame bit rate using the HEVC test model HM 9.0 encoder [6], with two screen content sequences of resolution 1024×768 and length 10 frames, at 30 fps. The auxiliary frame QP varies from −12 to +4 relative to the main frame QP, which in this case is set to 26. The left sequence bit rate for the main view is 45 kbps, with a PSNR of 46.2 dB. The right sequence bit rate for the main view is 127 kbps, with a PSNR of 42.9 dB.

# Conclusion

This proposal enables the creation of a system in which the existing 4:2:0 decoding process becomes the core component of a 4:4:4 decoder. Moreover, a subset of the decoded output can provide compatibility with existing 4:2:0 decoding systems. Since 4:2:0 is the most widely supported format in products, having an effective way of conveying 4:4:4 content through such decoders can provide the substantial benefit of enabling widespread near-term deployment of 4:4:4 capabilities.

# References

1. T. Lin *et. al,* “Syntax and semantics of Dual-coder Mixed Chroma-sampling-rate (DMC) coding for 4:4:4 screen content”, JCTVC-J0233, 10th Meeting: Stockholm, SE, 11–20 July 2012.
2. Y. Wu*,* S. Kanumuri, S. Sadhwani, L. Zhu, S. Sankuratri, G. J. Sullivan, and B. A. Kumar, “Frame packing arrangement SEI for 4:4:4 content in 4:2:0 bitstreams”, JCTVC-K0240, 11th Meeting: Shanghai, CN, 10-19, October, 2012.
3. Rec. ITU-T H.264 | ISO/IEC 14496-10, *Advanced video coding*.
4. B. Bross *et. al*, “High efficiency video coding (HEVC) Text Specification Draft 9”, JCTVC-L1003, 11th meeting, Shanghai, October 2012.
5. Y. Wu, S. Kanumuri, Y. Zhang, S. Sadhwani, G. J. Sullivan, and H. S. Malvar, “Tunneling High-Resolution Color Content through 4:2:0 HEVC and AVC Video Coding Systems”, *Proc. IEEE Data Compression Conf.* (DCC 2013), Snowbird, Utah, March 2013 (*to appear*).
6. “HEVC software repository”, <https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/>.

# Patent rights declaration(s)

**Microsoft Corporation may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, is prepared to grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ITU-T Recommendation | ISO/IEC International Standard (per box 2 of the ITU-T/ITU-R/ISO/IEC patent statement and licensing declaration form).**

# Appendix

We have attached a set of BMP pictures for comparison. Note that in order to view the true differences, the BMP files have to be opened and displayed in original resolution of without any scaling.

OriginalText.bmp is the original screen capture in the RGB color space.

444Text.bmp is the screen capture with the color conversion from RGB to YUV444, and from YUV444 back to RGB.

420AverageSubsamplingText.bmp is the screen capture with the color conversion from RGB to YUV444, applying the [0.5 0.5] filter on chroma while downsampling from YUV444 to YUV420, using replication to upsample from YUV420 to YUV444, and from YUV444 back to RGB.

420DirectSubsamplingText.bmp is the screen capture with the color conversion from RGB to YUV444, no chroma filtering while downsampling from YUV444 to YUV420, using replication to upsample from YUV420 to YUV444, and from YUV444 back to RGB.

For example, in areas with red text on green background, the difference in quality between the different variants can be easily seen. 444Text.bmp is almost identical to OriginalText.bmp and is expected. 420DirectSubsamplingText.bmp has the worst quality. While 420AveragetSubsamplingText.bmp is better than 420DirectSubsamplingText.bmp, the quality is significantly lower than that of 444Text.bmp.

Similar chroma artifacts are observed with AVC-compressed screen content without using the proposed approach and were avoided by using the proposed approach.