October 2023

General Multi-channel Input Video Support for Video Encoding

Danny Hong

Deb Mukherjee

General Multi-channel Input Video Support for Video Encoding

ABSTRACT

Traditional video coding standards process video frames in the YCbCr color format, with a single luminance channel (Y) and two chroma channels (Cb, Cr). There is a strong correlation between the three channels that can enable efficient encoding and good compression, since decisions such as motion estimation and block partitioning can be made once for the Y plane and applied to the chroma channels. Also, subsampling can be used to reduce the resolution of the chroma planes. However, these standards are unable to efficiently encode additional data such as depth, alpha, velocity, etc. that is needed for immersive applications such as virtual reality (VR), augmented reality (AR), or extended reality (XR) to provide a high quality photorealistic experience. This disclosure describes configuring the additional data as a single input that can optionally be combined with the YCbCr input. Regardless of the number of channels, the implementation of the encoder can be designed such that the complexity of the encoder is a function of one dominant plane/channel. Current video coding standards can be extended, and future video coding standards can be implemented to incorporate such input by including a provision to signal the number of input channels and their resolution. A dominant plane can also be explicitly signaled, or it can be implied to be the first plane.

KEYWORDS

- Video encoder
- Video streaming
- Photorealistic video
- Immersive video
- Augmented reality
- Extended reality (XR)
- Virtual reality
- Video coding
- Chroma subsampling
- Multi-channel video

BACKGROUND

Current video coding standards such as H.264 (AVC)/H.265 (HEVC) and VP9/AV1 define input sequences of video frames in the YCbCr color format. Y (luminance), Cb (blue difference chroma) and Cr (red difference chroma) are separate planes of values. However, since the details in these planes are highly correlated, these are simply coded together for better performance and compression efficiency.

For example, objects in the Y plane have similar characteristics as the corresponding ones in both the Cb and Cr planes such that there is no need to go through separate encoding decision processes like motion estimation or block partitioning, e.g., quad-tree arrangement of coding units (CUs) or prediction units (PUs) within a coding tree unit (CTU) for the three planes. As a result, there is also no need to send separate motion vectors or block partitioning information. Additionally, as human perception is less sensitive to chroma details, basic tools of the standards support smaller resolution Cb and Cr planes, e.g., 4:2:0 is a subsampling where the width and height of the chroma planes are both halved). Many hardware implementations of video coding standards, particularly encoder implementations, result in the same performance whether 4:4:4 (no chroma subsampling) or 4:2:0 is used. This is because most computationally expensive decisions for encoding, including motion estimation, are done only once at the Y plane resolution (the largest resolution) as the three planes are highly correlated.

Traditional 2D video encoding for video playback only requires the three planes (Y, Cb, and Cr). However, immersive applications such as virtual reality (VR), augmented reality (AR), extended reality (XR), etc. require far more information than what the YCbCr channels include. In cloud gaming, the complex rendering of video content is offloaded, encoded, and streamed to low-end devices. A similar mechanism is necessary to provide high quality photorealistic

experience to head-mounted displays (HMDs), augmented reality/virtual reality (AR/VR) glasses, etc. that may have limited compute and/or battery capacity.

For a high quality experience with such devices, additional channels or planes of data are required to allow late stage reprojection. Some of the additional planes of data encoded and transmitted include a depth map plane and the planes for the X, Y, and Z components of velocity. Also, if alpha blending is needed between the transmitted and locally generated video frames, an additional plane/channel for opacity values is also needed. With reprojection and alpha blending, in addition to the traditional three channel/plane color video, additional channels/planes of data are needed to be encoded/transmitted. In the particular example described above, 5 additional planes of data are needed.

However, including additional channels such as depth, alpha, velocity X, Y, and Z components, requires additional encoding resources with the additional channels packed as one of the supported YCbCr color formats. This can be wasteful of encoding resources. Further, since not all the channels are not correlated with each other, a lack of compression efficiency results, requiring individual encoding decisions for different channels. Existing video coding standards and their hardware implementations accommodate additional channels by creating extra encoding sessions in which these channels are incorporated into one of the supported YCbCr color formats.

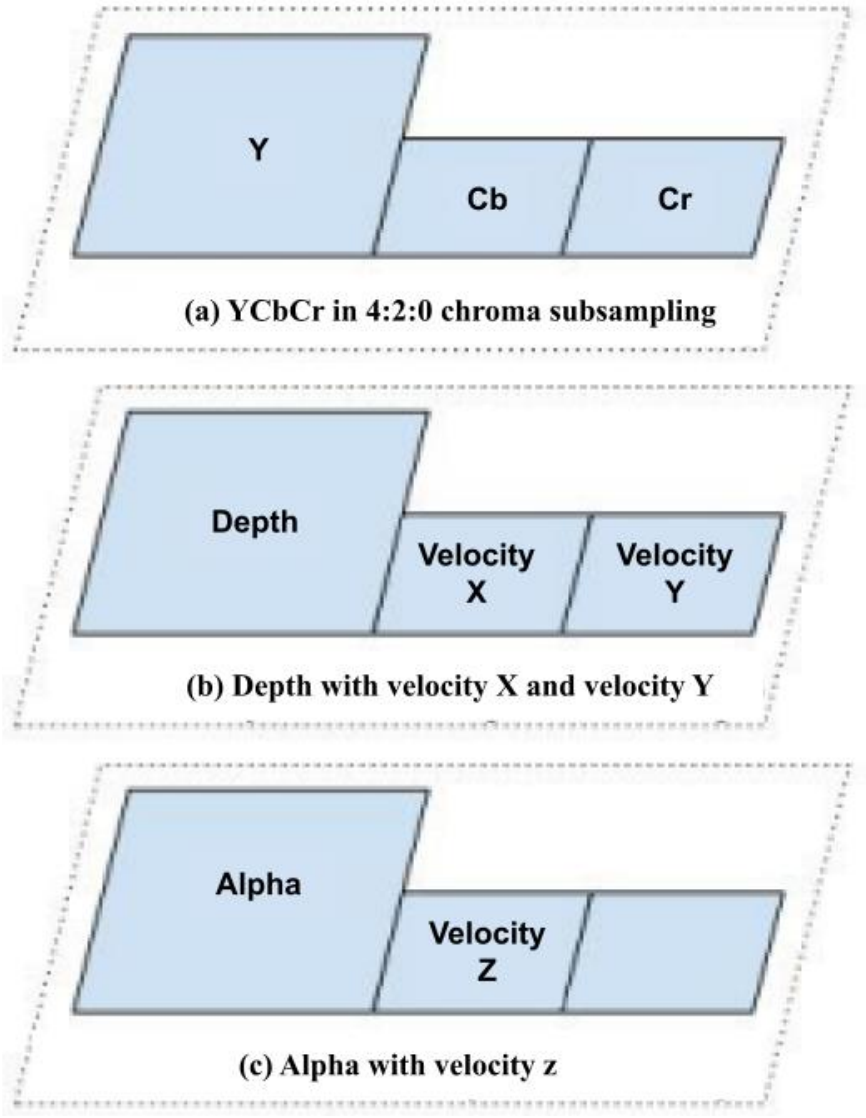

Fig. 1 : Example scenario with 5 data channels and 3 separate inputs: (a) YCbCr in 4:2:0 chroma subsampling; (b) Depth with velocity X and velocity Y; (c) Alpha with velocity Z.

Fig. 1 illustrates an example scenario involving 5 additional data channels (depth, alpha, velocity X, velocity Y and velocity Z). The data includes the standard YCbCr input for color information, as seen in Fig. 1(a). A separate input is generated for the depth map and the X and Y components of velocity, as seen in Fig. 1(b). For the remaining channels, which include alpha and the Z component of velocity, form another input, as shown in Fig. 1(c). The Cr plane is not

utilized for the third input in this setup. In this example, the velocity components are packed into smaller chroma planes since they do not require the same high resolution as other channels.

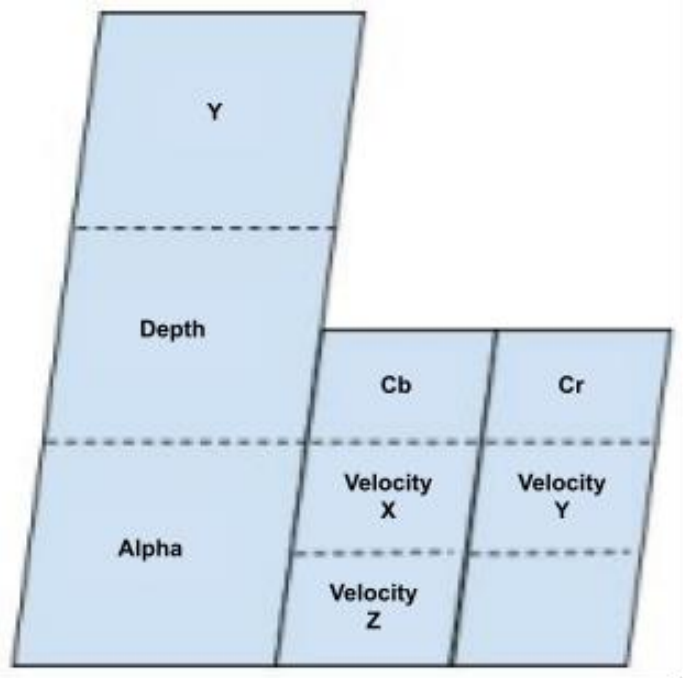

Fig. 2: Example scenario involving 5 data channels and 3 inputs stitched together to create a single larger resolution input.

To enable a single encoding session, the planes of three separate sessions, as shown in Fig. 1, can be stitched together to create a single larger resolution input. As seen in Fig. 2, all the corresponding planes are put together top to bottom. However, this does not address the issues related to encoder resource usage or compression efficiency (since correlation between some planes of data is not taken into account).

DESCRIPTION

If the input planes are correlated, there is no need to go through separate encoding decision processes. Hardware implementations can be designed to support not just three, but any number of planes of data without incurring noticeably more complexity. This is possible since

most encoding decisions are done based on one dominant plane (traditionally, the Y plane), and the remaining planes (traditionally, the Cb and Cr planes) can reuse the decisions. Based on this approach, whether there are two or more additional plans, the increase in complexity is limited. This disclosure describes general multi-channel video where a single input includes more than the three traditional Y, Cb and Cr channels. For example, one or more of the additional channels as described above with reference to Figs. 1 and 2 can be packed as a single input as shown in Fig. 3, instead of two inputs.

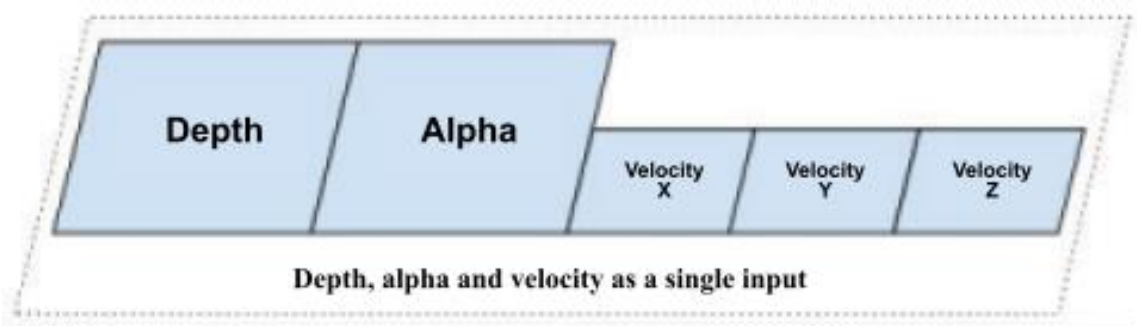

Fig. 3: Depth, alpha and velocity as a single input

Furthermore, all video channels that include the traditional color channels along with the five additional channels can also be packed as illustrated in Fig. 4 below.

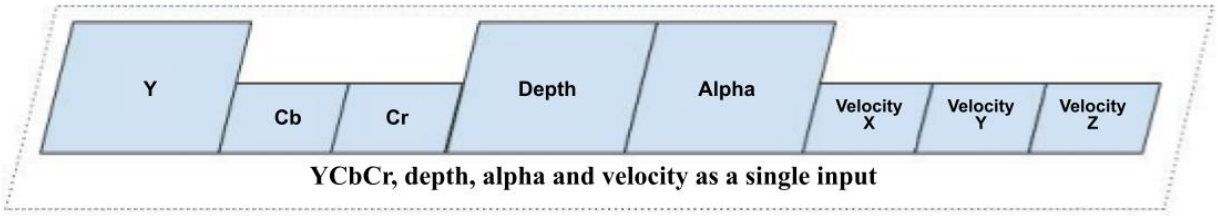

Fig. 4: YCbCr, Depth, alpha and velocity as a single input

Regardless of the number of channels, the implementation of the encoder can be designed such that the complexity of the encoder is a function of one dominant plane/channel. Hence, whether the input is the traditional 3 channel YCbCr or as shown in Fig. 3 or Fig. 4, since the

dominant channel resolution is the same, encoding decision processes can be implemented to be similar in complexity.

Current video coding standards can be extended, and future video coding standards can be implemented to incorporate such input by including a provision to signal the number of input channels and their resolution. Current standards such as H.265 (HEVC) and AV1 define ways to specify the input resolution and chroma subsampling via parameter sets or sequence header. New standards can be implemented to enable setting of number of input channels and their respective resolutions in such sequence level syntax. A dominant plane can also be explicitly signaled, or it can be implied to be the first plane.

Existing standards can support an extension via introduction of new profiles. With multiple-channel support as described herein, what each plane represents can be application specific. That information is not relevant for the actual encoding/decoding processes. The details can be provided as part user-defined metadata.

Any video encoder (or other encoder) that has a requirement to compress multiple channels of correlated data can benefit from the concepts described herein.

CONCLUSION

This disclosure describes configuring the additional data as a single input that can optionally be combined with the YCbCr input. Regardless of the number of channels, the implementation of the encoder can be designed such that the complexity of the encoder is a function of one dominant plane/channel. Current video coding standards can be extended, and future video coding standards can be implemented to incorporate such input by including a provision to signal the number of input channels and their resolution. A dominant plane can also be explicitly signaled, or it can be implied to be the first plane.

REFERENCES

1. "Web video codec guide - Web media technologies | MDN" available online at https://developer.mozilla.org/en-US/docs/Web/Media/Formats/Video_codecs accessed September 25, 2023.