



Title: Resolution switching for coding efficiency and resilience
Status: Input Document
Purpose: Proposal
Author(s) or Contact(s): Thomas Davies Email: thdavies@cisco.com
Source: Cisco Systems (UK)

Abstract

A method for changing the resolution of frames within a sequence without causing an IDR or new SPS to be sent is proposed. Frames may be predicted across resolutions by re-scaling reference pictures in a similar manner to H.263 Annex P. The purpose is to allow video communications to use re-scaling to adapt seamlessly to adverse network conditions. It is observed that in HEVC Intra frames are relatively more expensive than in AVC, and forcing the use of Intra frames can worsen losses and increase delays. It is reported that predicting instead of inserting an IDR frame when down-scaling gives average gains of 5.4-5.6% in Low-Delay common conditions, with 15.6-16.4% gain in Class E. When the size of intra frames is reduced to emulate realistic video conferencing, average gains of 6.9-7.4% are observed, with 20.5-21.6% gain in Class E.

1 Introduction

This document proposes allowing spatial resolution to vary within a video stream without triggering a new Sequence Parameter Set (SPS) and IDR frame. It is proposed instead that an encoder can signal a change of resolution and predict frames across resolutions. The method used for prediction is very close to H.263-style reference-frame resampling [1]: reference frames are simply re-scaled as needed.

The motivation for this proposal is the requirements of video communication. Here video is frequently re-scaled for a number of reasons, but especially to maximise perceptual quality in adverse or varying network conditions: video conferencing equipment often contains advanced scaling algorithms to achieve this. However, in AVC prediction across resolutions is not permitted and an IDR frame must be sent to re-initialise the stream.

Resolution is therefore often changed at times of network stress. Although IDR frames break dependencies, this is not necessarily a good thing for error resilience. The sheer size of Intra frames make them difficult to transmit in a low-delay environment. If part of an IDR frame is then lost, another IDR frame must be generated after a delay, sometimes leading to an “Intra storm” as bandwidth is consumed by repeated resynchronisation frames which cannot get through. Often it is the case that known good long-term references exist that could be used at the point of resolution change for a lower-cost recovery.

The current development of HEVC suggests that much less gain over AVC has been found for Intra frame coding than for Inter coding. This increases the disparity between Intra and Inter frames, and therefore the impact of changing resolution compared to AVC.

Although it is proposed that a stream may contain different resolutions, this is not a scalable coding proposal. However, the proposal is certainly highly compatible with future scalable extensions, and the syntax changes will provide hooks for scalable functionality.

The proposed algorithm has been implemented in HM3.1 and results are reported showing the impact of an inserted IDR frame versus cross-resolution prediction from an old reference frame.

2 Requirements for dynamic resolution

2.1 *Seamless network adaption and error resilience*

Applications such as video conferencing and streaming over packets networks frequently require that the encoded stream adapt to changing network conditions, especially when bit rates are too high and data is being lost. Such applications typically have a return channel allowing the encoder to detect the errors and perform adjustments. The encoder has two main tools at its disposal: bit rate reduction and changing the resolution, either temporal or spatial. Temporal resolution changes can be effectively achieved by coding using hierarchical prediction structures. However, for best quality spatial resolution changes are needed as well as part of a well-designed encoder for video communications.

Changing spatial resolution within AVC requires an IDR frame is sent and the stream is reset. This causes significant problems. An IDR frame at reasonable quality will be much larger than an Inter picture, and will be correspondingly more complex to decode: this costs time and resource. This is a problem if the resolution change is requested by the decoder for loading reasons. It can also break low-latency buffer conditions, forcing an audio re-sync, and the end-to-end delay of the stream will increase, at least temporarily. This gives a poor user experience.

To minimize these problems, the IDR is typically sent at low quality, using a similar number of bits to a P frame, and it takes a significant time to return to full quality for the given resolution. To get low enough delay, the quality can be very low indeed and there is often a visible blurring before the image is “refocused”. In effect, the Intra frame is doing very little useful work in compression terms: it is just a method of re-starting the stream.

So there is a requirement for methods in HEVC which allow resolution to be changed, especially in challenging network conditions, with minimal impact on subjective experience.

2.2 *Fast start*

It would be useful to have a “fast start” mode where the first frame is sent at reduced resolution and resolution is increased over the next few frames, in order to reduce delay and get to normal quality more quickly without unacceptable image blurring at the beginning.

2.3 *Conference “compose”*

Video conferences also often have a feature whereby the person speaking is shown full-screen and other participants are shown in smaller resolution windows. To support this efficiently, often the smaller pictures are sent at lower resolution. This resolution is then increased when the participant becomes the speaker and is full-screened. Sending an intra frame at this point causes an unpleasant hiccup in the video stream. This effect can be quite noticeable and unpleasant if speakers alternate rapidly.

2.4 *Relationship with scalable coding*

It is not proposed to add scalable coding functionality into the core standard. Likewise it is not desirable to require a scalable coding solution in order to switch resolution efficiently, since in many video conference scenarios not all participants may support scalable technology but will still need to change resolution. Instead, since resolution adaption is a core video coding technique across many applications to maximize subjective quality, there is a requirement to support dynamic changes of resolution in HEVC.

3 Tool description

There are a number of different ways in which changing the resolution mid-stream could be achieved. The focus in this proposal is simplicity and so the following constraints have been applied:

- the resolution may differ from the nominal resolution by a factor 0.5, applied to both dimensions at once; resolution may increase or decrease, yielding scaling ratios of 0.5 and 2.0 (but see below).
- aspect ratios, chroma formats or other parts of the video format are not changed;
- cropping areas are scaled in proportion to the picture;
- reference pictures are simply re-scaled as needed and picture prediction is as normal.

(More different resolutions could be supported, subject to limiting the number of different scalings that would need to be implemented to a manageable level.)

A Picture Resolution Index (PRI) is signaled in the PPS. This will indicate slices using this PPS will be from pictures having the resolution indicated by the index. This will set values for the width and height. This index will form part of the meta-data associated with each stored reference picture.

When a subsequent picture is decoded, reference picture lists need to be formed. If the current reference picture has index different from the reference picture, it is rescaled. Both resolutions are available for future reference – that is, if the resolution changes back, no further rescaling is needed.

If a reference frame has a different size to the current frame, temporal motion vector prediction is disabled.

It is possible that complexity could be controlled by limiting the number of scalings per frame period, to prevent several reference frames from being re-scaled at once.

3.1 Compatibility with SVC

In case the PRI is never changed, the stream will correspond to a single resolution which may, in case $PRI > 0$, be lower than the nominal resolution. An enhancement layer could therefore be added with a different PRI, with the potential for cross-layer prediction. H.263 Annex 0-style spatial scalability could be added very simply as a base-level scalability system if desired.

3.2 Scaling

It is proposed to use simple, zero-phase separable down- and up-scaling filters. Note these filters are for prediction only: a decoder may also use more sophisticated scaling for display purposes.

The factor 2 down-scaling filter has zero phase and taps:

$$(-1, 9, 16, 9, -1)/32$$

Down-sampling points are at even sample positions and are co-sited. The same filter is used for luma and chroma.

2:1 up-sampling is performed by inserting zeroes at odd sample positions, and filtering with the filter

$$(-3 \ 0 \ 19 \ 32 \ 19 \ 0 \ -3)/32$$

Values are computed using unbiased rounding, e.g.

$$y(2n+1) = (19*(x(2n)+x(2n+2)) - 3(x(2n-2)+x(2n+4))+16) \gg 5$$

Edge extension is used at image boundaries.

Combined up- and down-sampling will not change phase or the position of chroma sampling points.

4 Modifications to bitstream syntax and semantics

Various modifications to the current Working Draft (WD) [3] are needed.

4.1 Sequence parameter set (SPS)

The SPS is modified as follows:

| seq_parameter_set_rbsp() { | Descriptor |
|--|------------|
| profile_idc | u(8) |
| reserved_zero_8bits /* equal to 0 */ | u(8) |
| level_idc | u(8) |
| seq_parameter_set_id | ue(v) |
| max_temporal_layers_minus1 | u(3) |
| sequence_pic_width_in_luma_samples | u(16) |
| sequence_pic_height_in_luma_samples | u(16) |
| bit_depth_luma_minus8 | ue(v) |
| bit_depth_chroma_minus8 | ue(v) |
| pcm_bit_depth_luma_minus1 | u(4) |
| pcm_bit_depth_chroma_minus1 | u(4) |
| log2_max_frame_num_minus4 | ue(v) |
| pic_order_cnt_type | ue(v) |
| if(pic_order_cnt_type == 0) | |
| log2_max_pic_order_cnt_lsb_minus4 | ue(v) |
| else if(pic_order_cnt_type == 1) { | |
| delta_pic_order_always_zero_flag | u(1) |
| offset_for_non_ref_pic | se(v) |
| num_ref_frames_in_pic_order_cnt_cycle | ue(v) |
| for(i = 0; i < num_ref_frames_in_pic_order_cnt_cycle; i++) | |
| offset_for_ref_frame[i] | se(v) |
| } | |
| max_num_ref_frames | ue(v) |
| gaps_in_frame_num_value_allowed_flag | u(1) |
| log2_min_coding_block_size_minus3 | ue(v) |
| log2_diff_max_min_coding_block_size | ue(v) |
| log2_min_transform_block_size_minus2 | ue(v) |
| log2_diff_max_min_transform_block_size | ue(v) |
| log2_min_pcm_coding_block_size_minus3 | ue(v) |
| max_transform_hierarchy_depth_inter | ue(v) |
| max_transform_hierarchy_depth_intra | ue(v) |
| chroma_pred_from_luma_enabled_flag | u(1) |
| loop_filter_across_slice_flag | u(1) |
| sample_adaptive_offset_enabled_flag | u(1) |
| adaptive_loop_filter_enabled_flag | u(1) |
| pcm_loop_filter_disable_flag | u(1) |
| cu_qp_delta_enabled_flag | u(1) |
| temporal_id_nesting_flag | u(1) |
| rbsp_trailing_bits() | |
| } | |

4.2 Picture parameter set (PPS)

The PPS is modified as follows:

| | Descriptor |
|--|------------|
| pic_parameter_set_rbsp() { | |
| pic_parameter_set_id | ue(v) |
| seq_parameter_set_id | ue(v) |
| picture_resolution_idx | ue(v) |
| if (picture_resolution_idx == 0){ | |
| pic_width_in_luma_samples = sequence_pic_width_in_luma_samples | |
| pic_height_in_luma_samples = sequence_pic_height_in_luma_samples | |
| } else if (picture_resolution_idx == 1){ | |
| pic_width_in_luma_samples = (sequence_pic_width_in_luma_samples+1)/2 | |
| pic_height_in_luma_samples = (sequence_pic_height_in_luma_samples+1)/2 | |
| } | |
| entropy_coding_mode_flag | u(1) |
| num_temporal_layer_switching_point_flags | ue(v) |
| for(i = 0; i < num_temporal_layer_switching_point_flags; i++) | |
| temporal_layer_switching_point_flag[i] | u(1) |
| num_ref_idx_l0_default_active_minus1 | ue(v) |
| num_ref_idx_l1_default_active_minus1 | ue(v) |
| pic_init_qp_minus26 /* relative to 26 */ | se(v) |
| constrained_intra_pred_flag | u(1) |
| slice_granularity | u(2) |
| shared_pps_info_enabled_flag | u(1) |
| if(shared_pps_info_enabled_flag) | |
| if(adaptive_loop_filter_enabled_flag) | |
| alf_param() | |
| if(cu_qp_delta_enabled_flag) | |
| max_cu_qp_delta_depth | u(4) |
| rbsp_trailing_bits() | |
| } | |

4.2.1 PPS semantics

picture_resolution_idx shall have a value between 0 and 1.

4.3 Slice Header

There are no changes to the slice header syntax

4.4 Prediction

The value of `picture_resolution_idx` present in the PPS shall be assigned to each frame whose slices reference that PPS, and made available via a function `PicResolutionIdx`. If a reference frame in `RefPicList0` or `RefPicList1` or `RefPicListComb` has a different resolution to the current frame, it is first rescaled to have resolution corresponding to that of the current picture.

4.4.1 Prediction semantics

4.4.1.1 Co-located predictors

Section 8.4.2.1.6 “Derivation process for temporal luma motion vector prediction” of the WD [3] is modified as follows:

...

Outputs of this process are

- the motion vector prediction `mvLXCcol`,
- the availability flag `availableFlagLXCcol`.

The function `RefPicOrderCnt(pic, refidx, LX)` is specified by the value of `PicOrderCnt` of the picture that is the reference picture `RefPicListX[refidx]` of `pic` with `X` being 0 or 1. `PicOrderCnt` of the reference picture shall be maintained until the picture is marked as “non-existing.”

The function `RefPicResolutionIdx(pic, refidx, LX)` is specified by the value of `PicResolutionIdx()` of the picture that is the reference picture `RefPicListX[refidx]` of `pic` with `X` being 0 or 1. If `RefPicResolutionIdx(pic, refidx, LX)` is not equal to `PicResolutionIdx(pic)` then `availableFlagLXCcol` shall be false.

...

4.4.1.2 Scaling

Section 8.4.2.2.1 “Reference picture selection process” in the WD [3] will be modified to require scaling when `RefPicResolutionIdx` is not equal to `PicResolutionIdx`.

5 Simulations

HM3.1 was modified to incorporate the changes to the PPS and SPS, and the scaling functions. In single-resolution mode, the code gave almost identical results to the HM3.1 anchors. Differences were due to the different PPS and SPS (the HM3.1 SPS does not include `seq_parameter_set_id`), and to padding the input pictures to a multiple of 16 by default to support the lower resolution mode.

Simulations were performed using `LDB_HE`, `LDB_LC`, `LDP_HE` and `LDP_LC` configurations [2]. A resolution change scenario was simulated at frame 20. It was assumed that frames 16-19 were lost or too badly damaged to be used as references, and hence were marked as non-reference. From frame 20 pictures were coded at 1/4 resolution. All pictures were padded to a multiple of 16 to support an 8x8 CU size at the lowest resolution.

Six configurations were simulated:

Configuration 1 :

- Flags: `--ResSwitchFrameNum=20 --ResSwitchType=3`
- IDR frame at frame 0 and frame 20
- pictures 16-19 inclusive unavailable for reference

- pictures with POC \geq 20 coded at 1/4 resolution:
- prediction structure as per common conditions

Configuration 2:

- Flags: --ResSwitchFrameNum=20 --ResSwitchType=1
- IDR frame at frame 0 and predicted frame at frame 20
- pictures 16-19 inclusive unavailable for reference
- pictures with POC \geq 20 coded at 1/4 resolution
- prediction structure as per common conditions

Configuration 3:

- Flags: --ResSwitchFrameNum=20 --ResSwitchType=3 --IntraQPOffset=6
- IDR frame at frame 0 and frame 20 with QP=nominal QP+6
- pictures 16-19 inclusive unavailable for reference
- pictures with POC \geq 20 coded at 1/4 resolution
- --RateGOPSize=2

Configuration 4:

- Flags: --ResSwitchFrameNum=20 --ResSwitchType=1 --IntraQPOffset=6
- IDR frame at frame 0 with QP=nominal QP+6 and predicted frame at frame 20
- pictures 16-19 inclusive unavailable for reference
- pictures with POC \geq 20 coded at 1/4 resolution
- --RateGOPSize=2

Configuration 5:

- Flags: --ResSwitchFrameNum=20 --ResSwitchType=3 --IntraQPOffset=12
- As Configuration 3 but with Intra QP=nominal QP+12

Configuration 6

- Flags: --ResSwitchFrameNum=20 --ResSwitchType=1 --IntraQPOffset=12
- As Configuration 4 but with Intra QP=nominal QP+12

Configurations 3-6 represent more realistic video conference configurations. Configurations 5 and 6 perhaps represent the most acceptable Intra frame size in delay terms. Configuration 1 was compared against configuration 2, configuration 3 against configuration 4 and configuration 5 against configuration 6.

Simulations were run on a homogenous cluster using 64 bit Intel Xeon X7560 8-core processors at 2.26GHz, running 64-bit Debian Linux. A single process was run per physical core (no hyper-threading).

5.1 Common conditions comparison (configs 1 and 2)

Summary results are as follows:

| | | |
|--|----------------|----------------|
| | Low delay B HE | Low delay B LC |
|--|----------------|----------------|

| | Y | U | V | Y | U | V |
|----------------|-------------|--------------|--------------|-------------|--------------|--------------|
| Class A | | | | | | |
| Class B | -5.2 | -10.2 | -11.2 | -5.1 | -9.9 | -10.5 |
| Class C | -2.2 | -4.2 | -4.1 | -2.1 | -4.3 | -3.7 |
| Class D | -1.7 | -5.7 | -4.3 | -1.6 | -2.0 | -2.8 |
| Class E | -16.4 | -27.7 | -33.3 | -15.6 | -31.1 | -36.6 |
| Overall | -5.7 | -10.8 | -11.8 | -5.4 | -10.5 | -11.8 |
| Enc Time[%] | | 100% | | | 100% | |
| Dec Time[%] | | 92% | | | 92% | |

| | Low delay P HE | | | Low delay P LC | | |
|----------------|----------------|--------------|--------------|----------------|--------------|--------------|
| | Y | U | V | Y | U | V |
| Class A | | | | | | |
| Class B | -4.9 | -9.8 | -11.4 | -5.1 | -10.2 | -11.3 |
| Class C | -2.1 | -4.4 | -3.9 | -2.2 | -5.1 | -4.6 |
| Class D | -1.6 | -7.2 | -5.4 | -1.7 | -3.0 | -3.1 |
| Class E | -16.5 | -29.0 | -35.3 | -16.3 | -32.8 | -38.2 |
| Overall | -5.6 | -11.4 | -12.5 | -5.6 | -11.4 | -12.6 |
| Enc Time[%] | | 100% | | | 100% | |
| Dec Time[%] | | 92% | | | 93% | |

For class E, the configuration 1 lower resolution IDR frame size as a multiple of mean frame size is:

| | Vidyo1 | Vidyo3 | Vidyo4 | Mean (frame periods) | Milliseconds |
|-------|--------|--------|--------|-------------------------|--------------|
| QP=22 | 14.9 | 13.1 | 13.9 | 14.0 | 248 |
| QP=37 | 26.9 | 24.7 | 25.1 | 25.7 | 428 |

5.2 Reduced intra frame quality comparison

5.2.1 Intra QP + 6 (configs 3 and 4)

Summary results are as follows:

| | Low delay B HE | | | Low delay B LC | | |
|----------------|----------------|--------------|--------------|----------------|--------------|--------------|
| | Y | U | V | Y | U | V |
| Class A | | | | | | |
| Class B | -6.4 | -11.9 | -12.5 | -6.3 | -11.5 | -11.7 |
| Class C | -2.7 | -5.6 | -5.0 | -2.5 | -5.6 | -5.1 |
| Class D | -2.1 | -6.6 | -5.4 | -1.9 | -2.7 | -3.9 |
| Class E | -21.3 | -30.0 | -36.6 | -20.5 | -31.9 | -38.6 |
| Overall | -7.2 | -12.4 | -13.4 | -6.9 | -11.7 | -13.1 |
| Enc Time[%] | | 100% | | | 100% | |
| Dec Time[%] | | 98% | | | 95% | |

| | Low delay P HE | | | Low delay P LC | | |
|--|----------------|---|---|----------------|---|---|
| | Y | U | V | Y | U | V |

| | | | | | | |
|----------------|-------------|--------------|--------------|-------------|--------------|--------------|
| Class A | | | | | | |
| Class B | -6.0 | -11.9 | -13.0 | -6.2 | -11.4 | -12.1 |
| Class C | -2.6 | -5.4 | -4.9 | -2.5 | -5.8 | -4.9 |
| Class D | -2.0 | -5.9 | -5.6 | -1.9 | -2.8 | -3.9 |
| Class E | -21.4 | -32.5 | -38.5 | -21.5 | -32.7 | -40.7 |
| Overall | -7.0 | -12.6 | -13.9 | -7.1 | -11.8 | -13.6 |
| Enc Time[%] | 100% | | | 100% | | |
| Dec Time[%] | 93% | | | 96% | | |

For class E, the configuration 1 lower resolution IDR frame sizes as a multiple of mean frame size are:

| | Vidyo1 | Vidyo3 | Vidyo4 | Mean (frame periods) | Milliseconds |
|-------|--------|--------|--------|-------------------------|--------------|
| QP=22 | 7.6 | 6.8 | 6.6 | 7.0 | 117 |
| QP=37 | 11.7 | 8.5 | 11.0 | 10.4 | 173 |

5.2.2 Intra QP + 12 (configs 5 and 6)

Summary results are as follows:

| | Low delay B HE | | | Low delay B LC | | |
|----------------|----------------|--------------|--------------|----------------|--------------|--------------|
| | Y | U | V | Y | U | V |
| Class A | | | | | | |
| Class B | -6.6 | -12.3 | -12.8 | -6.7 | -12.0 | -12.0 |
| Class C | -2.9 | -6.0 | -5.6 | -2.8 | -5.8 | -5.2 |
| Class D | -2.2 | -7.0 | -6.3 | -2.3 | -4.4 | -5.0 |
| Class E | -21.5 | -30.5 | -37.3 | -20.9 | -33.6 | -39.9 |
| Overall | -7.4 | -12.8 | -14.0 | -7.3 | -12.6 | -13.8 |
| Enc Time[%] | 100% | | | 100% | | |
| Dec Time[%] | 98% | | | 98% | | |

| | Low delay P HE | | | Low delay P LC | | |
|----------------|----------------|--------------|--------------|----------------|--------------|--------------|
| | Y | U | V | Y | U | V |
| Class A | | | | | | |
| Class B | -6.4 | -12.3 | -13.0 | -6.6 | -12.2 | -12.4 |
| Class C | -2.7 | -5.7 | -5.4 | -2.8 | -6.4 | -5.4 |
| Class D | -2.3 | -5.8 | -7.4 | -2.3 | -4.1 | -4.8 |
| Class E | -21.5 | -32.4 | -40.2 | -21.6 | -35.1 | -42.0 |
| Overall | -7.3 | -12.8 | -14.8 | -7.4 | -13.0 | -14.3 |
| Enc Time[%] | 100% | | | 100% | | |
| Dec Time[%] | 93% | | | 96% | | |

For class E, the configuration 1 lower resolution IDR frame sizes as a multiple of mean frame size are:

| | Vidyo1 | Vidyo3 | Vidyo4 | Mean (frame periods) | Milliseconds |
|-------|--------|--------|--------|-------------------------|--------------|
| QP=22 | 4.1 | 3.5 | 3.5 | 3.7 | 62 |

| | | | | | |
|-------|-----|-----|-----|-----|----|
| QP=37 | 5.2 | 4.2 | 5.0 | 4.8 | 80 |
|-------|-----|-----|-----|-----|----|

6 Recommendations

It is recommended that support for dynamic resolution be incorporated into the HEVC Working Draft and HM, and that an AHG be started to study the technology, covering issues such as filter design and the resolutions to be used.

Patent rights declaration(s)

Cisco Systems may have IPR 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).

References

- [1] ITU-T H.263 (01/2005): Video coding for low bit rate communication, Annex O “Temporal, SNR and Spatial Scalability mode” and Annex P “Reference Picture Resampling”
- [2] JCTVC-E700, “Common test conditions and software reference configuration”, F. Bossen (DoCoMo), 5th JCT-VC Meeting, Geneva, CH, 16-23 March, 2011
- [3] JCTVC-E603_d8 “WD3: Working Draft 3 of High-Efficiency Video Coding”, T. Wiegand, W.-J. Han, B. Bross, J.-R. Ohm, G. J. Sullivan (Editors), 5th JCT-VC Meeting, Geneva, CH, 16-23 March 2011