|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  32nd Meeting: Ljubljana, SI, 12–18 July 2018 | Document: JCTVC-AF0024-v2 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Mask information for Annotated region SEI message** | | |
| *Status:* | Input document to JCT-VC | | |
| *Purpose:* | Proposal | | |
| *Author(s) or Contact(s):* | Xiaozhen Zheng, Xuyang Feng, Liliang Zhang, Cong Zhao | Email: | xiaozhen.zheng@dji.com |
| *Source:* | DJI | | |

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

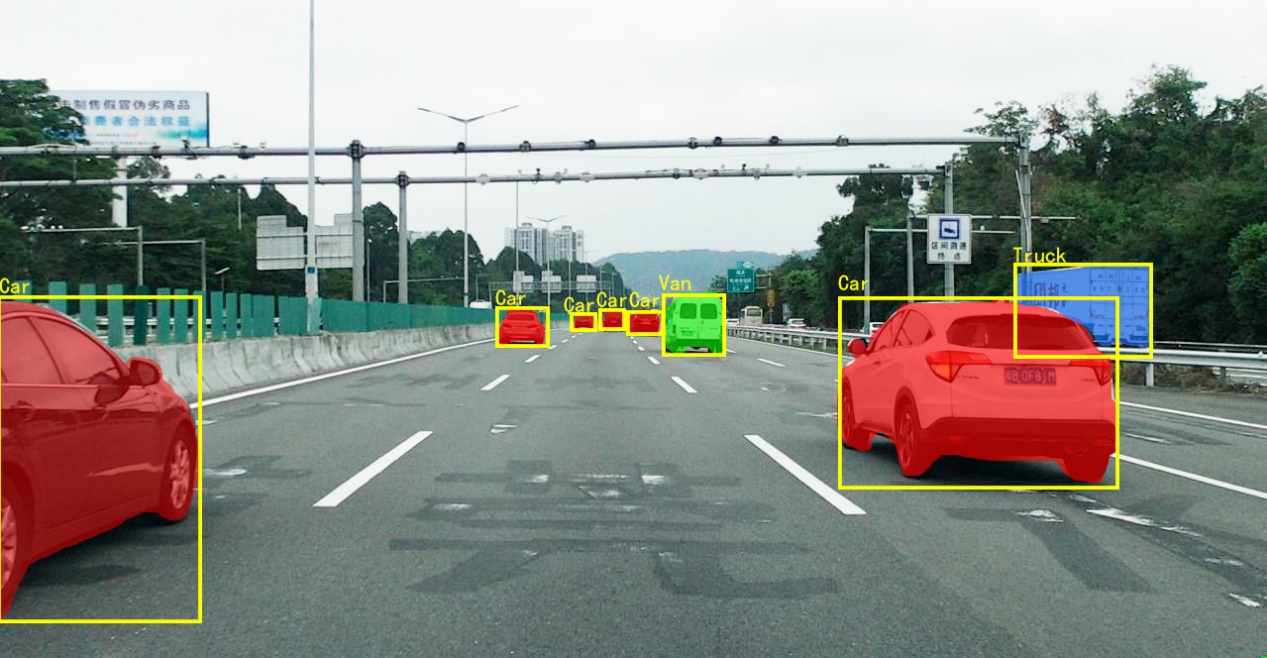
# Abstract

This contribution proposes to add mask information for annotated regions to annotated regions SEI message. With the help of mask information, a decoder can detect actual sharp of a tracking object without additional object recognition method which can bring benefits to decoder-side application. Base on concept of annotated region SEI, a signaling method for mask information is also proposed at the contribution.

# Motivation

Annotated region SEI message was proposed at SanDiego meeting [1] that uses syntax elements to describe the location and size of tracked objects. This kind of SEI message show great benefits for some particular applications. For an example, video surveillance scenario always perform object tracking at encoder side with the purpose that video camera can track specified object real time. In order to observe track the objects and perform online scene analysis, a decoder may perform object tracking again which cause the waste of computing resource. Since annotated region SEI can transmit the tracked objects’ location and size to decoder-side, decoder will not perform same analysis action as encoder side that can save power of decoder side. Meanwhile, annotated region SEI can also improve the quality of object tracking because an encoder can do analysis on original video which has better quality than decoder-side’s compressed video, and get more actual tracking results.

Besides a tracked object’s location and size, some applications might also need the object’s shape can help a decoder-side device performs more intelligent activity. For an example, a video surveillance decoder-side device may need to highlight the tracked objects and dim background and other untracked objects with the purpose to help observer to track interested objects more easily. Moreover, some entertainment apps have requirement to change a person’s background real time that user can get funs from background switch. Fig.1 shows an example of those scenario. At this scenario, tracked cars, van and truck are marked by the objects’ shape with different color.

Fig. 1: real-time tracking and object’s shape detection for a road

Perform objects tracking, recognition and get objects’ shape real time is doable at some encoder-side devices, e.g. drone, because those devices always have powerful computing module. For an example, drone usually contains high performance computing module with the purpose to achieve automatic avoid-obstacle and auto-driving. However, same analysis activities may cause trouble at decoder-side device since those are always mobile devices that are sensitive to power.

Therefore, based on above analysis, this contribution suggests add mask information of objects as optional syntax element to annotated region SEI message. It can help encoder-side device to save object tracking and recognition results, and decoder-side can use the results to perform intelligent activity directly.

# Proposal

A single bit mask for a tracked object is proposed to added to annotated region SEI. One bit for one luma sample is used to signal whether the sample belongs to the object or background.

To align with concept of annotated region SEI and improve signaling efficiency for object mask, a syntax element is proposed to indicate whether a object’s mask can be inferred from the same object in previous frame.

A high level flag that indicates whether current frame contains object mask is also proposed for the cases that object mask information is not necessary.

The modification to annotated region SEI syntax and the added semantics are show as below:

**Modification to Annotated regions SEI message syntax**

|  |  |
| --- | --- |
| annotated\_regions( payloadSize ) { | **Descriptor** |
| **ar\_cancel\_flag** | u(1) |
| **ar\_not\_optimized\_for\_viewing\_flag** | u(1) |
| **ar\_true\_motion\_flag** | u(1) |
| **ar\_occluded\_object\_flag** | u(1) |
| **ar\_partial\_object\_flag\_present\_flag** | u(1) |
| **ar\_object\_label\_present\_flag** | u(1) |
| **ar\_object\_confidence\_present\_flag** | u(1) |
| **ar\_object\_mask\_present\_flag** | u(1) |
| if( ar\_object\_confidence\_present\_flag ) |  |
| **ar\_object\_confidence\_length\_minus1** | u(4) |
| if( ar\_object\_label\_present\_flag ) { |  |
| **ar\_object\_label\_language\_present\_flag** | u(1) |
| if( ar\_object\_label\_language\_present\_flag ) { |  |
| while( !byte\_aligned( ) ) |  |
| **ar\_bit\_equal\_to\_zero** /\* equal to 0 \*/ | f(1) |
| **ar\_object\_label\_language** | st(v) |
| } |  |
| **ar\_num\_cancelled\_labels** | ue(v) |
| for( i = 0; i < ar\_num\_cancelled\_labels; i++ ) |  |
| **ar\_cancelled\_label\_idx**[ i ] | ue(v) |
| **ar\_num\_new\_labels** | ue(v) |
| for( i = 0; i < ar\_num\_new\_labels; i++ ) |  |
| **ar\_label\_idx**[ i ] | ue(v) |
| while( !byte\_aligned( ) ) |  |
| **ar\_bit\_equal\_to\_zero** /\* equal to 0 \*/ | f(1) |
| **ar\_label**[ ar\_label\_idx[ i ] ] | st(v) |
| } |  |
| } |  |
| **ar\_num\_cancelled\_objects** | ue(v) |
| for( i = 0; i < ar\_num\_cancelled\_objects; i++ ) |  |
| **ar\_cancelled\_object\_idx**[ i ] | ue(v) |
| **ar\_num\_objects\_minus1** | ue(v) |
| for( i = 0; i  <=  ar\_num\_objects\_minus1;i++ ) { |  |
| **ar\_object\_idx**[ i ] | ue(v) |
| **ar\_new\_object\_flag[** ar\_object\_idx[ i ] ] | u(1) |
| if( !ar\_new\_object\_flag[ ar\_object\_idx[ i ] ] ) |  |
| **ar\_bounding\_box\_update\_flag[** ar\_object\_idx[ i ] ] | u(1) |
| if ( ar\_object\_mask\_present\_flag ) { |  |
| **ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ]** | u(1) |
| if ( ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ] && (!ar\_new\_object\_flag[ ar\_object\_idx[ i ] ] || !ar\_bounding\_box\_update\_flag[ ar\_object\_idx[ i ] ] )) { |  |
| **ar\_bounding\_box\_mask\_infer\_flag[ ar\_object\_idx[ i ] ]** | u(1) |
| } |  |
| } |  |
| if( ar\_new\_object\_flag[ ar\_object\_idx[ i ]  &&  ar\_object\_label\_present\_flag ) |  |
| **ar\_object\_label\_idc**[ ar\_object\_idx[ i ] ] | ue(v) |
| if( ar\_partial\_object\_flag\_present\_flag ) |  |
| **ar\_partial\_object\_flag**[ ar\_object\_idx[ i ] ] | u(1) |
| if( ar\_object\_bounding\_box\_update\_flag[ ar\_object\_idx[ i ] ]  | | ar\_new\_object\_flag[ ar\_object\_idx[ i ] ] ) { |  |
| **ar\_object\_top[** ar\_object\_idx[ i ] ] | u(16) |
| **ar\_object\_left**[ ar\_object\_idx[ i ] ] | u(16) |
| **ar\_object\_width**[ ar\_object\_idx[ i ] ] | u(16) |
| **ar\_object\_height**[ ar\_object\_idx[ i ] ] | u(16) |
| if( ar\_object\_confidence\_present\_flag ) |  |
| **ar\_object\_confidence**[ ar\_object\_idx[ i ] ] | u(v) |
| } |  |
| if ( ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ] && !ar\_bounding\_box\_mask\_infer\_flag[ ar\_object\_idx[ i ] ] || ar\_new\_object\_flag[ ar\_object\_idx[ i ] ] || ar\_bounding\_box\_update\_flag[ ar\_object\_idx[ i ] ] ) { |  |
| for (m = 0; m < ar\_object\_height[i ]; m++) |  |
| for (n = 0; n < ar\_object\_width[i ]; n++) |  |
| **mask[ ar\_object\_idx[ i ] ][m][n]** | u(1) |
| } |  |
| } |  |
| } |  |

**Added annotated region SEI message semantics**

**ar\_object\_mask\_present\_flag** equal to 1 indicates mask information syntax elements of annotated regions are present. ar\_object\_mask\_present\_flag equal to 0 indicates mask information syntax elements of annotated regions are not present.

**ar\_bounding\_box\_mask\_present\_flag[** ar\_object\_idx[ i ] **]** equal to 1 indicates that mask information syntax elements of the annotated region with the value of ar\_object\_idx[ i ] is present. ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ] equal to 0 indicates that mask information syntaxs element of the annotated region with the value of ar\_object\_idx[ i ] is not present. If ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ] syntax element is not present, the value of ar\_bounding\_box\_mask\_present\_flag[ ar\_object\_idx[ i ] ] is infered to 0.

**ar\_bounding\_box\_mask\_infer\_flag** [ ar\_object\_idx[ i ] ] equal to 1 indicates that mask information syntax elements of the corresponding object is inferred from mask information syntax elements specified in the previous annotated regions SEI message in output order in the CLVS with the same value of ar\_object\_idx[ i ]. ar\_bounding\_box\_mask\_infer\_flag[ ar\_object\_idx[ i ] ] equal to 0 indicates that mask information syntax elements of the corresponding object is not inferred from the previous annotated regions SEI message in output order in the CLVS that contains the same value of ar\_object\_idx[ i ]. If ar\_bounding\_box\_mask\_infer\_flag[ ar\_object\_idx[ i ] ] syntax element is not represent, the value of ar\_bounding\_box\_mask\_infer\_flag [ ar\_object\_idx[ i ] ] is infered to 0.

**mask[ ar\_object\_idx[ i ] ][m][n]** equal to 1 indicates that a luma sample with right offsets *n* and bottom offset *m* realted to ar\_object\_top[ ar\_object\_idx[ i ] ] and ar\_object\_left[ ar\_object\_idx[ i ] ] belongs to the annotated region with the value of ar\_object\_idx[ i ]. mask[ ar\_object\_idx[ i ] ][m][n] equal to 0 indicates that the luma sample does not belong to the annotated region with the value of ar\_object\_idx[ i ].

# Example Scenarios

## Picture 0

At picture 0, 4 objects are present in the image, 2 car, 1 van and 1 truck.

Pic 0 key syntax:

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_object\_mask\_present\_flag | 1 |
| ar\_num\_new\_labels | 4 |
| ar\_label\_idx[ 0 ] | 0 |
| ar\_label\_idx[ 1 ] | 1 |
| ar\_label\_idx[ 2 ] | 2 |
| ar\_label\_idx[ 3 ] | 3 |
| ar\_label[ 0 ] | car |
| ar\_label[ 1 ] | car |
| ar\_label[ 2 ] | van |
| ar\_label[ 3 ] | truck |
| ar\_num\_objects\_minus1 | 3 |
| ar\_object\_idx[ 0 ] | 0 |
| ar\_new\_object\_flag[ 0 ] | 1 |
| ar\_bounding\_box\_mask\_present\_flag[0] | 0 |
| ar\_object\_idx[ 1 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 1 |
| ar\_bounding\_box\_mask\_present\_flag[1] | 1 |
| ar\_object\_idx[ 2 ] | 2 |
| ar\_new\_object\_flag[ 2 ] | 1 |
| ar\_bounding\_box\_mask\_present\_flag[2] | 1 |
| ar\_object\_idx[ 3 ] | 3 |
| ar\_new\_object\_flag[ 3 ] | 1 |
| ar\_bounding\_box\_mask\_present\_flag[3] | 0 |
| ar\_object\_label\_idc[ 0 ] | 0 |
| ar\_object\_label\_idc[ 1 ] | 1 |
| ar\_object\_label\_idc[ 2 ] | 2 |
| ar\_object\_label\_idc[ 3 ] | 3 |
| ar\_object\_top, left, width, height[ 0 ] | BB\_A |
| ar\_object\_top, left, width, height[ 1 ] | BB\_B |
| mask[1] | car1’s mask |
| ar\_object\_top, left, width, height[ 2 ] | BB\_C |
| mask[2] | van’s mask |
| ar\_object\_top, left, width, height[ 3 ] | BB\_D |

## Picture 1

At picture 1, the car0 (object 0), car1(object 1) and truck (object 3) stayed in the same position and van (object 2) moved to a new position. Car1’s mask is inferred from previous frame.

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_object\_mask\_present\_flag | 1 |
| ar\_num\_new\_labels | 0 |
| ar\_num\_objects\_minus1 | 3 |
| ar\_object\_idx[ 0 ] | 0 |
| ar\_new\_object\_flag[ 0 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 0 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[0] | 0 |
| ar\_object\_idx[ 1 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 1 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[1] | 1 |
| ar\_bounding\_box\_mask\_infer\_flag[ 1 ] | 1 |
| ar\_object\_idx[ 2 ] | 2 |
| ar\_new\_object\_flag[ 2 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 2 ] | 1 |
| ar\_bounding\_box\_mask\_present\_flag[2] | 1 |
| ar\_object\_idx[ 3 ] | 3 |
| ar\_new\_object\_flag[ 3 ] | 1 |
| ar\_object\_bounding\_box\_update\_flag[ 3 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[3] | 0 |
| ar\_object\_label\_idc[ 0 ] | 0 |
| ar\_object\_label\_idc[ 1 ] | 1 |
| ar\_object\_label\_idc[ 2 ] | 2 |
| ar\_object\_label\_idc[ 3 ] | 3 |
| ar\_object\_top, left, width, height[ 0 ] | BB\_A |
| ar\_object\_top, left, width, height[ 1 ] | BB\_B |
| ar\_object\_top, left, width, height[ 2 ] | BB\_E |
| mask[2] | van’s mask |
| ar\_object\_top, left, width, height[ 3 ] | BB\_D |

## Picture 2

At picture 2, all of objects stayed in the same position. Car1’s mask is updated and van’s mask is inferred from previous frame.

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_object\_mask\_present\_flag | 1 |
| ar\_num\_new\_labels | 0 |
| ar\_num\_objects\_minus1 | 3 |
| ar\_object\_idx[ 0 ] | 0 |
| ar\_new\_object\_flag[ 0 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 0 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[0] | 0 |
| ar\_object\_idx[ 1 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 1 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[1] | 1 |
| ar\_bounding\_box\_mask\_infer\_flag[ 1 ] | 0 |
| ar\_object\_idx[ 2 ] | 2 |
| ar\_new\_object\_flag[ 2 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 2 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[2] | 1 |
| ar\_bounding\_box\_mask\_infer\_flag[ 2 ] | 1 |
| ar\_object\_idx[ 3 ] | 3 |
| ar\_new\_object\_flag[ 3 ] | 1 |
| ar\_object\_bounding\_box\_update\_flag[ 3 ] | 0 |
| ar\_bounding\_box\_mask\_present\_flag[3] | 0 |
| ar\_object\_label\_idc[ 0 ] | 0 |
| ar\_object\_label\_idc[ 1 ] | 1 |
| ar\_object\_label\_idc[ 2 ] | 2 |
| ar\_object\_label\_idc[ 3 ] | 3 |
| ar\_object\_top, left, width, height[ 0 ] | BB\_A |
| ar\_object\_top, left, width, height[ 1 ] | BB\_B |
| mask[1] | car1’s mask |
| ar\_object\_top, left, width, height[ 2 ] | BB\_E |
| ar\_object\_top, left, width, height[ 3 ] | BB\_D |

# References

[1] J. Boyce, P. Guruva reddiar, JCTVC-AE0027, “Object tracking SEI message”, April, 2018, San Diego, US.

[2] J. Boyce, Y.-K. Wang, G. J. Sullivan, JCTVC-AE1012, “Annotated Regions SEI message for HEVC (Draft 1)”, April, 2018, San Diego, US.

# Patent rights declaration(s)

**SZ DJI Technology Co., Ltd 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).**