|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  31st Meeting: San Diego, US, 13–20 Apr. 2018 | Document: JCTVC-AE0027-v4 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Object tracking SEI message (now Annotated region SEI message)** | | |
| *Status:* | Input document to JCT-VC | | |
| *Purpose:* | Proposal | | |
| *Author(s) or Contact(s):* | Jill Boyce Palanivel Guruva reddiar | Email: | jill.boyce@intel.com palanivel.guruvareddiar@intel.com |
| *Source:* | Intel | | |

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

# Abstract

This contribution proposes to define an annotated region SEI message that carries parameters to describe annotated regions using bounding boxes representing the size and location of tracked objects within the compressed video bit stream. Attributes about the annotated objects can also be optionally signaled, including a label for the object and the confidence level associated with the tracking and detection.

The SEI message is also proposed for inclusion in AVC.

The v4 version of this documents updates the proposal based upon the discussion during the initial presentation to the JCT-VC, which includes renaming the proposed SEI message Annotated region SEI, rather than Object tracking SEI.

# Motivation

Object detection and tracking is common workload for end-to-end video systems, such as depicted in Figure 1, which includes smart cameras, video gateway, and datacenter/cloud.

For an intermediate processing stage, the following steps are typical:

1. Decode compressed video
2. Perform analytics on the video to identify object locations using a rectangular bounding box for each picture
3. Composite video with a graphical outline of the bounding box(es)
4. Re-encode the video containing bounding box(es)



**Smart Cameras Video Gateway Datacenter / Cloud**

Figure 1. End-to-End intelligent video solution

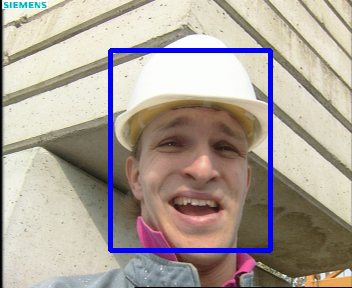


Figure 2. Video composited with graphical bounding box outline

Compositing the graphical bounding box outline onto the video has clear disadvantages. The unmodified video is no longer available. Refinement of the bounding box in later processing stages is not supported. Significant computation complexity is involved in the re-encoding process.

Using an SEI message to send the bounding box parameters as metadata addresses those disadvantages.

# Proposal

The proposed SEI message carries parameters to describe the bounding box of tracked objects within the compressed video bitstream. Optionally, object labels and confidence levels associated with the detected and tracked objects may be provided. Object detection and tracking is computationally expensive, so it is advantageous in some systems to allow separation of objection detection/tracking functionality from video decoding.

The object label syntax contains a text field. Multiple attributes for a given object can be signalled using a single label, e.g. separated by commas or another delimiter. For example, “Person, Brad Pitt” can be sent as a single label.

The syntax uses persistence of parameters to avoid the need to re-signal information already available in previous SEI messages within the same persistence scope, e.g. within the same coded video sequence. For example, if one tracked object remains stationary for the current picture relative to a previously coded picture while another tracked object moves during that interval, the bounding box parameters are signaled only for the moving object.

A syntax flag is included to indicate if a coded video sequence is not optimized for user viewing, but rather is intended for machine learning applications. For example, this flag could be used when the tracked objects are represented at high quality but areas outside of the tracked objects are represented at very low video quality.

Another syntax flag is included to indicate if the motion information (motion vectors, modes) were selected in order to accurately track objection motion, rather than to optimize coding efficiency, i.e. with RDO.

A syntax flag indicate if bounding boxes may represent the estimated position of occluded or partially occluded objects (vs. representing only the visible portion).

A flag per tracked object may optionally indicate if the bounding box represents the size and location of an object that is only partially visible within the coded picture.

The number of bits of granularity of the confidence level is explicitly signaled, up to 16 bits.

## Annotated region SEI message syntax

|  |  |
| --- | --- |
| annotated\_region ( payloadSize ) { | **Descriptor** |
| **ar\_cancel\_flag** | u(1) |
| **ar\_seq\_parameter\_set\_id** | ue(v) |
| **ar\_not\_optimized\_for\_viewing\_flag** | u(1) |
| **ar\_true\_motion\_flag** | u(1) |
| **ar\_occluded\_objects\_flag** | u(1) |
| **ar\_partial\_object\_flag\_present\_flag** | u(1) |
| **ar\_object\_label\_present\_flag** | u(1) |
| **ar\_object\_detection\_confidence\_info\_present\_flag** | u(1) |
| if (ar\_object\_detection\_confidence\_info\_present\_flag ) |  |
| **ar\_object\_detection\_confidence\_precision\_num\_bits** | 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\_zero\_bit** /\* equal to 0 \*/ | f(1) |
| **ar\_object\_label\_language** | st(v) |
| } |  |
| **ar\_num\_cancel\_labels** | ue(v) |
| for( i = 0; i  <  ar\_num\_cancel\_labels; i++ ) { |  |
| **ar\_cancel\_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\_zero\_bit** /\* equal to 0 \*/ | f(1) |
| **ar\_label[** ar\_label\_idx [ i ] ] | st(v) |
| } |  |
| } |  |
| **ar\_num\_cancel\_objects** | ue(v) |
| for( i = 0; i  <  ar\_num\_cancel\_objects; i++ ) { |  |
| **ar\_cancel\_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\_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\_detection\_confidence\_info\_present\_flag ) |  |
| **ar\_object\_detection\_confidence** [ ar\_object\_idx[ i ] ] | u(v) |
| } |  |
| } |  |
| } |  |

## Annotated region SEI message semantics

The annotated region SEI message carries parameters to describe annotated regions using bounding boxes representing the size and location of tracked objects within the compressed video bitstream and also describe optional elements such as object labels and object detection confidence levels.

**ar\_cancel\_flag** equal to 1 indicates that the SEI message cancels the persistence of any previous annotated region SEI message in output order that is associated with one or more primary picture layers to which this SEI applies. ar\_cancel\_flag equal to 0 indicates that annotated region information follows.

**ar\_seq\_parameter\_set\_id** indicates and shall be equal to the sps\_seq\_parameter\_set\_id value of the active SPS. The value of ar\_seq\_parameter\_set\_id shall be in the range of 0 to 15, inclusive.

**ar\_not\_optimized\_for\_viewing\_flag** equal to 1 indicates that decoded picture is not optimized for user viewing, but for other purposes. ar\_not\_optimized\_for\_viewing\_flagequal to 0 indicates that the decoded picture is optimized for user viewing.

**ar\_true\_motion\_flag** equal to 1 indicates that the motion information in the coded picture was selected with a goal of accurately representing object motion for annotated objects. ar\_true\_motion\_flag equal to 0 makes no indication about motion vector accuracy of annotated objects.

**ar\_occluded\_objects\_flag** equal to 1 indicates that the ar\_object\_top, ar\_object\_left, ar\_object\_width, and ar\_object\_height[ ar\_object\_idx[ i ] ] syntax elements represent the size and location of an object that may not visible or may be only partially visible in the coded picture. ar\_occluded\_objects\_flagequal to 0 indicates that the ar\_object\_top, ar\_object\_left, ar\_object\_width, and ar\_object\_height[ ar\_object\_idx[ i ] ] syntax elements represent the size and location of the visible portion of an object within the coded picture.

**ar\_partial\_object\_flag\_present\_flag** equal to 1 indicates that ar\_partial\_object\_flag[ ar\_object\_idx[ i ] ] syntax elements are present in the coded bit stream. ar\_partial\_object\_flag\_present\_flag equal to 0 indicates that that ar\_partial\_object\_flag[ ar\_object\_idx[ i ] ] syntax elements are not present in the coded bit stream.

**ar\_object\_label\_present\_flag** equal to 1 indicates that the label information corresponding to the annotated objects is present in the coded bit stream. ar\_object\_label\_present\_flag equal to 0 indicates that the label information corresponding to the annotated objects is not present in the coded bit stream.

**ar\_object\_detection\_confidence\_info\_present\_flag** equal to 1 indicates that ar\_object\_detection\_confidence[ ] is present in the bitstream. ar\_object\_detection\_confidence\_info\_present\_flag equal to 0 indicates that ar\_object\_detection\_confidence[ ] is not present in the bitstream.

**ar\_object\_detection\_confidence\_precision\_num\_bits** indicates the number of bits used to represent ar\_object\_detection\_confidence[ ].

**ar\_object\_label\_language\_present\_flag** equal to 1 indicates that the ar\_object\_label\_language is present in the bit stream. ar\_object\_label\_language\_present\_flagequal to 0 indicates that the ar\_object\_label\_language is not present and that the language of the label is unspecified.

**ar\_zero\_bit** shall be equal to zero.

**ar\_object\_label\_language** contains a language tag as specified by IETF RFC 5646 followed by a null termination byte equal to 0x00. The length of the ar\_object\_label\_language syntax element shall be less than or equal to 255 bytes, not including the null termination byte.

**ar\_num\_cancel\_labels** indicates the number of canceled labels associated with the annotated objects. ar\_num\_cancel\_labels shall be in the range of 0 to 255, inclusive.

**ar\_cancel\_label\_idx**[ i ]cancels the persistence of the ar\_cancel\_label\_idx[ i ]-th label. The value of ar\_cancel\_label\_idx[ i ] shall be in the range of 0 to 255, inclusive.

**ar\_num\_new\_labels** indicates the total number of new labels associated with the annotated objects that will be signaled. The value of ar\_num\_new\_labels shall be in the range of 0 to 255, inclusive.

**ar\_label\_idx**[ i ]indicates the index to the label associated with the corresponding annotated object. The value of ar\_label\_idx[ i ]shall be in the range of 0 to 255, inclusive.

**ar\_label**[ ar\_label\_idx[ i ] ] contains the label of the bounding box. The length of the ar\_label[  ar\_label\_idx[ i ] ] syntax element shall be less than or equal to 255 bytes, not including the null termination byte.

**ar\_num\_cancel\_objects** indicates the number of canceled annotated objects. ar\_num\_cancel\_objects shall be in the range of 0 to 255.

**ar\_cancel\_object\_idx**[ i ]cancels the persistence of the ar\_cancel\_object\_idx[ i ]-th annotated object. The value of ar\_cancel\_object\_idx[ i ]shall be in the range of 0 to 255, inclusive.

**ar\_num\_objects\_minus1** plus 1 indicates the total number of annotated objects being tracked in the current decoded picture. ar\_num\_objects\_minus1 shall be in the range of 0 to 255.

**ar\_object\_idx**[ i ]specifies the index of the object present in the list of objects present in the current coded picture.ar\_object\_idx[ i ]shall be in the range of 0 to 255.

**ar\_new\_object\_flag**[ ar\_object\_idx[ i ]]equal to 1 indicates that the corresponding object was not represented in earlier annotated region SEI messages within the persistance scope. ar\_new\_object\_flag[ ar\_object\_idx[ i ]]equal to 0 indicates that the corresponding object was represented in earlier annotated region SEI messages within the persistance scope.

**ar\_object\_bounding\_box\_update\_flag**[ ar\_object\_idx[ i ]]equal to 1 indicates that the bounding box of the corresponding object has been changed from the values represented in earlier annotated region SEI messages within the persistance scope. ar\_object\_bounding\_box\_update\_flag[ ar\_object\_idx[ i ]]] equal to 0 indicates that the bounding box of the corresponding object persists from earlier annotated region SEI messages within the persistance scope.

**ar\_object\_label\_idc**[ ar\_object\_idx[ i ] ] specifies the index of the label corresponding to the object. The value of ar\_object\_label\_idc[ ar\_object\_idx[ i] ] persists from earlier annotated region SEI messages within the persistance scope. If ar\_object\_label\_idc[ ar\_object\_idx[ i] ] was not present in earlier annotated region SEI messages within the persistance scope, its value is undefined.

**ar\_partial\_object\_flag**[ ar\_object\_idx[ i ] ] equal to 1 indicates that ar\_object\_top, ar\_object\_left, ar\_object\_width and ar\_object\_height[ ar\_object\_idx[ i ] ] syntax elements represent the size and location of an object that is only partially visible within the coded picture. ar\_partial\_object\_flag[ ar\_object\_idx [i] ] equal to 0 indicates that ar\_object\_top, ar\_object\_left, ar\_object\_width and ar\_object\_height[ ar\_object\_idx[ i ] ] syntax elements represent the size and location of an object that is fully visible within the coded picture.

**ar\_object\_top**[ ar\_object\_idx[ i ] ] and **ar\_object\_left**[ ar\_object\_idx[ i] ]specify, as luma samples, the top and left location, respectively, of the ar\_object\_idx[ i ]–th object in the decoded picture.The values of ar\_object\_left[ ar\_object\_idx[ i ] ] shall be in the range of 0 to pic\_width\_in\_luma\_samples inclusive and the value of ar\_object\_top[ ar\_object\_idx[ i] ] will be in the range of 0 to pic\_height\_in\_luma\_samples inclusive. The values of ar\_object\_top[ ar\_object\_idx[ i ] ] ] and ar\_object\_left[ ar\_object\_idx[ i ] ] persist from earlier annotated region SEI messages within the persistance scope. If ar\_object\_top[ ar\_object\_idx[ i ] ] or ar\_object\_left[ ar\_object\_idx[ i] ] were not present in earlier annotated region SEI messages within the persistance scope, their values are undefined.

**ar\_object\_width** [ ar\_object\_idx[ i ] ] and **ar\_object\_height**[ ar\_object\_idx[ i ] ] specify as luma samples the width and height, respectively, of the ar\_object\_idx[ i ]–th object in the decoded picture. When the ar\_partial\_object\_flag\_present\_flag is 0, the value of ar\_object\_left[ ar\_object\_idx[ i ] ]+ar\_object\_width[ ar\_object\_idx[ i ] ] shall be in the range of 0 to pic\_width\_in\_luma\_samples inclusive and the value of ar\_object\_top[ ar\_object\_idx[ i ] ]+ar\_object\_height [ ar\_object\_idx[ i ] ] shall be in the range of 0 to pic\_height\_in\_luma\_samples inclusive. The values of ar\_object\_width[ ar\_object\_idx[ i ] ] and ar\_object\_height[ ar\_object\_idx[ i ] ] persist from earlier annotated region SEI messages within the persistance scope. If ar\_object\_width[ ar\_object\_idx[ i ] ] or ar\_object\_height[ ar\_object\_idx[ i ] ] was not present in earlier annotated region SEI messages within the persistance scope, their values are undefined.

**ar\_object\_class\_detection\_confidence**[ ar\_object\_idx[ i ] ] specifies the confidence associated with the ar\_object\_idx[ i ]-th object, in units of 2**ar\_object\_detection\_confidence\_precision\_num\_bits**. The length of the ar\_object\_class\_detection\_confidence[ ar\_object\_idx[ i ] ] syntax element is ar\_object\_detection\_confidence\_precision\_num\_bitsbits. The value of ar\_object\_class\_detection\_confidence[ ar\_object\_idx[ i ] ] persists from earlier annotated region SEI messages within the persistance scope. If ar\_object\_class\_detection\_confidence[ ar\_object\_idx[ i ] ] was not present in earlier annotated region SEI messages within the persistance scope, its value is undefined.

# Example Scenarios

## Picture 0

At picture 0, two objects are present in the image, a car and a person.

Pic 0 key syntax:

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_num\_new\_labels | 2 |
| ar\_label\_idx[ 0 ] | 0 |
| ar\_label\_idx[ 1 ] | 1 |
| ar\_label[ 0 ] | car |
| ar\_label[ 1 ] | person |
| ar\_num\_objects\_minus1 | 1 |
| ar\_object\_idx[ 0 ] | 0 |
| ar\_new\_object\_flag[ 0 ] | 1 |
| ar\_object\_idx[ 1 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 1 |
| ar\_object\_label\_idc[ 0 ] | 0 |
| ar\_object\_label\_idc[ 1 ] | 1 |
| ar\_object\_top, left, width, height[ 0 ] | BB\_A |
| ar\_object\_top, left, width, height[ 1 ] | BB\_B |

## Picture 1

At picture 1, the car (object 0) stayed in the same position and the person (object 1) moved to a new position

Pic 1 key syntax:

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_num\_new\_labels | 0 |
| ar\_num\_objects\_minus1 | 1 |
| ar\_object\_idx[ 0 ] | 0 |
| ar\_new\_object\_flag[ 0 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 0 ] | 0 |
| ar\_object\_idx[ 1 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 1 ] | 1 |
| ar\_object\_top, left, width, height[ 1 ] | BB\_C |

The position of the car (object 0) persists from Pic 0, as BB\_A.

## Picture 2

At picture 2, the first car (object 0) is no longer in the picture, the person (object 1) moved within the picture, a different car (object 2) entered the picture, and a dog (object 3) entered the picture.

Pic 2 key syntax:

|  |  |
| --- | --- |
| ar\_object\_label\_present\_flag | 1 |
| ar\_num\_new\_labels | 1 |
| ar\_label\_idx[ 0 ] | 2 |
| ar\_label[ 2 ] | dog |
| ar\_num\_objects\_minus1 | 2 |
| ar\_object\_idx[ 0 ] | 1 |
| ar\_new\_object\_flag[ 1 ] | 0 |
| ar\_object\_bounding\_box\_update\_flag[ 1 ] | 1 |
| ar\_object\_idx[ 1 ] | 2 |
| ar\_new\_object\_flag[ 2 ] | 1 |
| ar\_object\_label\_idc[ 2 ] | 0 |
| ar\_object\_idx[ 2 ] | 3 |
| ar\_new\_object\_flag[ 3 ] | 1 |
| ar\_object\_label\_idc[ 3 ] | 2 |
| ar\_object\_top, left, width, height[ 1 ] | BB\_D |
| ar\_object\_top, left, width, height[ 2 ] | BB\_E |
| ar\_object\_top, left, width, height[ 3 ] | BB\_F |

# Patent rights declaration(s)

**Intel 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).**