INTERNATIONAL STANDARD

ISO/IEC 23001-10

Second edition
2020-04

Information technology — MPEG systems technologies —

Part 10: Carriage of timed metadata metrics of media in ISO base media file format

Technologies de l'information — Technologies des systèmes MPEG — Partie 10: Transport de métriques de métadonnées de temporisation de supports du format de fichier de support en base ISO

Reference number
ISO/IEC 23001-10:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

This second edition cancels and replaces the first edition (ISO/IEC 23001-10:2015), which has been technically revised.

The main changes compared to the previous edition are as follows:

— addition of carriage of special information in new Clause 6 and Annex A with support for encoded regions of interest;

— ISO/IEC 14496-12 and ISO/IEC 23008-2 moved from Bibliography to Clause 2 and other minor editorial changes to align fully with ISO/IEC Directives Part 2.

A list of all parts in the ISO/IEC 23001 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html..

Introduction

This document specifies the carriage of timed metadata in files belonging to the family based on ISO/IEC 14496-12. The families of metadata are 'green' metadata (related to energy conservation), quality measurements of the associated media data (related to video quality metrics) and coordinates describing relationship between media data.

Information technology — MPEG systems technologies —

Part 10:
Carriage of timed metadata metrics of media in ISO base media file format

1 Scope

This document defines a storage format for timed metadata. The timed metadata can be associated with other tracks in the ISO base media file format. Timed metadata such as quality and power consumption information and their metrics are defined in this part for carriage in files based on the ISO base media file format (ISO/IEC 14496-12). The timed metadata can be used for multiple purposes including supporting dynamic adaptive streaming.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14496-10, *Information technology — Coding of audio-visual objects — Part 10: Advanced video coding*

ISO/IEC 14496-12, *Information technology — Coding of audio-visual objects — Part 12: ISO base media file format*

ISO/IEC 23001-11, *Information technology — MPEG Systems Technologies — Part 11: Energy-Efficient Media Consumption (Green Metadata)*

ISO/IEC 23008-2, *Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 2: High efficiency video coding*

ITU-T Recommendation J.144, *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*

ITU-T Recommendation J.247, *Objective perceptual multimedia video quality measurement in the presence of a full reference*

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 14496-10 and ISO/IEC 23008 apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.2 Abbreviated terms

FSIG	frame significance
MOS	mean opinion score
MSE	mean signal error
MS-SSIM	multi-scale structural similarity index
ROI	region of interest
PEVQ	perceptual evaluation of video quality
PSNR	peak signal to noise ratio
SSIM	structural similarity index
VQM	video quality metric

4 Carriage of quality metadata

4.1 General

If quality metrics are carried in an ISO base media file format, they shall be carried in the metadata tracks within the ISO base media file format in accordance with ISO/IEC 14496-12. Different metric types and corresponding storage formats are identified by their unique code names. This clause defines those quality metrics.

The metadata track is linked to the track it describes by means of a `'cdsc'` (content describes) track reference.

Codes not defined in this document are reserved and files shall use only codes defined here.

4.2 Quality metadata

4.2.1 Definition

Sample Entry Type: `'vqme'`

Container: Sample Description Box (`'stsd'`)

Mandatory: No

Quantity: 0 or 1

The sample entry for video quality metrics is defined by the `QualityMetricsSampleEntry`.

The quality metrics sample entry shall contain a `QualityMetricsConfigurationBox`, describing metrics that are present in each sample, and the constant field size that is used for the values. The quality metrics are defined in subclause 4.3.

Each sample is an array of quality values, corresponding one for one to the declared metrics. Each value is padded by preceding zero bytes, as needed, to the number of bytes indicated by field_size_bytes.

The codecs parameter value for this track as defined in RFC 6381[6] shall be set to `'vqme'`. The sub-parameter for the `'vqme'` codec is a list of the metrics present in the track as indicated by the metrics code names, joined by "+", e.g., `'vqme.psnr+mssm'`.

4.2.2 Syntax

```
aligned(8) class QualityMetricsSampleEntry()
   extends MetadataSampleEntry ('vqme') {
   QualityMetricsConfigurationBox();
}

aligned(8) class QualityMetricsConfigurationBox
   extends FullBox('vqmC', version=0, 0){
   unsigned int(8) field_size_bytes;
   unsigned int(8) metric_count;
   for (i = 1 ; i <= metric_count ; i++){
      unsigned int(32) metric_code;
   }
}
```

4.2.3 Semantics

`field_size_bytes` indicates the constant size in byte of the value for a metric in each sample.

`metric_count` the number of metrics for quality values in each sample.

`metric_code` is the code name of the metrics in the sample.

4.3 Quality metrics

4.3.1 Peak signal to noise ratio (PSNR)

4.3.1.1 Definition

PSNR for encoded video sequence is defined based on per-picture mean square error (MSE) differences:

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j)-K(i,j)]^2$$

where

I is the luma plane of the reference $m \times n$ picture;

K is the luma plane of the reconstructed picture;

i,j are indices enumerating all pixel locations.

The picture-level PSNR is defined as:

$$PSNR = 10 \times \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$PSNR = 20 \times \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

where $MAX_I = 2^B - 1$ where B is the number of bits per sample in pictures.

PSNR for a given video sequence is computed as an average of all picture-level PSNR values obtained for all pictures in the sequence, i.e., for a sequence with N pictures:

$$PSNR_{sequence} = \frac{1}{N}\sum_{n=0}^{N-1} PSNR_{picture(n)}$$

Only luma component of the video signal is used for PSNR computation.

NOTE 1 This is the traditional metric referred to as PSNR in academic literature and in the context of video compression research.

NOTE 2 In cases when the spatial resolution of the reference pictures and the reconstructed ones do not match, reconstructed pictures are up-sampled to match the spatial resolution of the reference.

NOTE 3 In cases when the pictures of reconstructed video represent only a subset of pictures in the reference video sequence, reconstructed pictures are replicated to produce time-aligned reconstructed pictures for all pictures in the reference sequence.

4.3.1.2 Metric code name

PSNR quality metric values shall be provided as ones under the `'psnr'` metric code name.

4.3.1.3 Sample storage format

Each PSNR metric value shall be stored as an unsigned 16-bit integer value.

4.3.1.4 Decoding operation

Given stored 16-bit integer value x, the corresponding PSNR value (in dB) is derived as follows (expressed in floating point):

PSNR = (real) x / 100; with the exception of PSNR = infinity for x=0

4.3.2 SSIM

4.3.2.1 Definition

SSIM for encoded video sequence is defined based on SSIM index map obtained for each picture. Per-picture SSIM index map is computed as follows:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where

x is the 8×8 window in the reference picture;

y is the 8×8 window in the reconstructed picture;

μ_x is the average sample value for pixels in x;

μ_y is the average sample value for pixels in y;

σ_x^2 is the variance computed for pixel values in x;

σ_y^2 is the variance computed for pixel values in y;

σ_{xy} is the covariance computed for pixel values in x and y.

and where

$$c_1 = (k_1 L)^2,\ c_2 = (k_2 L)^2$$

are constants computed using

$$k_1 = 0.01,\ k_2 = 0.03,\ \text{and}\ L = 2^B - 1$$

where B is the number of bits per sample in reference video.

This formula is applied using an 8×8 sliding window and producing a map of SSIM index values for all pixel positions within a picture. The overall SSIM index is then computed as the average of index values in the SSIM map.

This formula is applied only on luma components in each picture.

SSIM for video sequence is computed as an average of all picture-level SSIM values obtained for all pictures in the sequence, i.e., for a sequence with N pictures:

$$SSIM_{sequence} = \frac{1}{N}\sum_{n=0}^{N-1} SSIM_{picture}(n)$$

NOTE 1 This is the traditional metric referred to as SSIM in academic literature and in the context of video compression research[1].

NOTE 2 The nominal range of SSIM index values is [−1..1].

NOTE 3 In cases when the resolution of the reference pictures and the reconstructed ones do not match, reconstructed pictures are up-sampled to match the resolution of the reference.

NOTE 4 In cases when the pictures of reconstructed video represent only a subset of pictures in the reference video sequence, reconstructed pictures are replicated to produce time-aligned reconstructed pictures for all pictures in the reference sequence.

4.3.2.2 Metric code name

SSIM quality metric values shall be provided under the `'ssim'` metric code name.

4.3.2.3 Sample storage format

Each SSIM metric value shall be stored as an unsigned 8-bit integer value.

4.3.2.4 Decoding operation

Given stored 8-bit integer value x, the corresponding SSIM value is derived as follows (expressed in floating point):

SSIM = (real) (x – 127) / 128.

4.3.3 MS-SSIM

4.3.3.1 Definition

The MS-SSIM calculation procedure is described in Figure 1. Taking the reference and distorted image signals as the input, the system iteratively applies a low-pass filter and downsamples the filtered image by a factor of 2. The original scale is indexed by j = 1 and the highest scale is indexed by j = M, for M-1 levels of iteration. Further details can be found in Reference [2].