INTERNATIONAL STANDARD

ISO/IEC 23001-10

First edition
2015-09-01

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 au format de fichier de support en base ISO

Reference number
ISO/IEC 23001-10:2015(E)

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

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. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

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).

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

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 23001 consists of the following parts, under the general title *Information technology — MPEG systems technologies*:

— *Part 1: Binary MPEG format for XML*

— *Part 2: Fragment request UNITS*

— *Part 3: XML IPMP messages*

— *Part 4: Codec configuration representation*

— *Part 5: Bitstream Syntax Description Language (BSDL)*

— *Part 7: Common encryption in ISO base media file format files*

— *Part 8: Coding-independent code-points*

— *Part 9: Common encryption for MPEG-2 Transport Streams*

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

— *Part 11: Energy-efficient media consumption (green metadata)*

Introduction

This part of ISO/IEC 23001 specifies the carriage of timed metadata related to two fields, in files belonging to the family based on ISO/IEC 14496-12 the ISO base media file format. The two families of metadata are “green” metadata (related to energy conservation) and quality measurements of the associated media data (related to video quality metrics).

Information technology — MPEG systems technologies —

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

1 Scope

This part of ISO/IEC 23001 defines a storage format for timed metadata metrics. The timed metadata metrics can be associated with other tracks in the ISO Base Media File Format. Typical timed metadata, quality and power consumption information and their metrics, are defined in the specification for carriage in files based on the ISO Base Media File Format (ISO/IEC 14496-12 and ISO/IEC 15444-12). The timed metadata can be used for multiple purposes including supporting dynamic adaptive streaming.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. 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 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 23008-2 apply.

3.2 Abbreviated terms

FSIG	Frame SIGnificance
MOS	Mean Opinion Score
MSE	Mean Signal Error
MS-SSIM	Multi-Scale Structural SIMilarity index
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 Introduction

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. Different metric types and corresponding storage formats are identified by their unique code names. This section 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 specification are reserved and files must 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 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 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:

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

where I denotes luma plane of the reference $m \times n$ picture, K denotes luma plane of the reconstructed picture, and i,j denote indices enumerating all pixel locations.

The picture-level PSNR is defined as:

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

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

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 we have

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

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

Note 1 This is the traditional metric referred to as PSNR in the 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 must be 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 must be 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:

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

where

x denotes the 8x8 window in the reference picture;

y denotes the 8x8 window in the reconstructed picture;

μ_x denotes the average sample value for pixels in x;

μ_y denotes the average sample value for pixels in y;

σ_x^2 denotes the average sample value for pixels in x;

σ_y^2 denotes the average sample value for pixels in y

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

and where

$$c_1=\left(k_1L\right)^2, c_2=\left(k_2L\right)^2$$

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 8x8 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 we have

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

Note 1 This is the traditional metric referred to as SSIM in the 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 must be 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 must be 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):

$$\mathrm{SSIM} = (\mathrm{real})(x - 127)/128.$$

4.3.3 MS-SSIM

4.3.3.1 Definition

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].

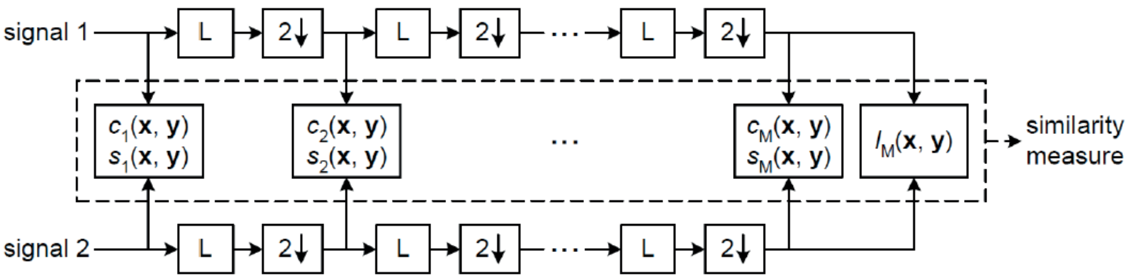

Figure 1 — MS-SSIM calculation procedure