 ISO/IEC JTC 1/SC 29/WG 3 N1563

**ISO/IEC JTC 1/SC 29/WG 3**

**MPEG Systems   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** Preliminary Draft of ISO/IEC 23001-10 AMD 3 Additional quality metrics

**Status:** Approved

**Date of document:** 2020-10-06

**Source:** ISO/IEC JTC 1/SC 29/WG 3

# Expected action: ACT

**Action due date:** 2020-10-05

**No. of pages:** 7 (with cover page)

**Email of Convenor:** young.L@samsung.com

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg3>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 3**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 3 N** **01563**

**Jun-Jul 2025, Daejeon, KR**

**ISO/IEC 23001-10:2020/AMD3**

ISO/IEC TC JTC 1/SC 29/WG3

Secretariat: JISC

**Information technology — MPEG systems technologies — Part 10: Carriage of timed metadata metrics of media in ISO base media file format — Amendment 3: Support for VMAF quality metric**

WD stage

**Warning for WDs and CDs**

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

*A model manuscript of a draft International Standard (known as “The Rice Model”) is available at* [*https://www.iso.org/iso/model\_document-rice\_model.pdf*](https://www.iso.org/iso/model_document-rice_model.pdf)

© ISO 2025

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO’s member body in the country of the requester.

ISO copyright office

CP 401 • Ch. de Blandonnet 8

CH-1214 Vernier, Geneva

Phone: +41 22 749 01 11

Email: copyright@iso.org

Website: www.iso.org

Published in Switzerland

Information technology — MPEG systems technologies — Part 10: Carriage of timed metadata metrics of media in ISO base media file format — Amendment 2: Support for VMAF quality metric

*Clause 4.3*

*Add the following subclauses at the end of the clause.*

### 4.3.x VMAF

#### 

#### Definition

VMAF metric is a full reference quality metric using machine learning to fuse the scores from several elementary quality metrics to produce the quality score for the video. The metric is trained to simulate the quality assessment obtained as the result of a subjective test. The full implementation of the metric is available in [6].

#### 

VMAF objective quality is obtained by merging values obtained from several elementary features into a single score using a Support Vector Regression (SVR). The metrics is trained to approximate the subjective quality test scores obtained on a variety of video material with compression distortions. The elementary features used in VMAF are: Mean Co-Located Pixel Difference (MCPD), visual information fidelity (VIF), and detail loss measure (DLM). VMAF is only determined based on the luma component. The calculation of the VMAF features is performed as follows.

## Mean Co-Located Pixel Difference (MCPD)

This feature is calculated as the mean absolute difference between the co-located samples in the current and previous frames.

*MCPD* = 1 / (M \* N) \* ∑i,j | *x*c( *i, j* ) - *x*c-1( *i, j* ) |

where, *M* is picture width, *N* is picture height, *c* is the current frame, *x*( *i*, *j* ) is the Gaussian filtered sample value at location *i* , *j*. The Gaussian filtering uses a 5-tap Gaussian kernel with = 0.5.

## Video Information Fidelity (VIF)

VIF models the image quality assessment problem as a communication channel, and the final VIF score can be considered as the ratio between two mutual information measures *I*(*C*; *F* | *s*) and *I*(*C*; *E* | *s*). This is illustrated in the system diagram below:

A diagram of a network

Description automatically generated

Here *C* represents the original source signal, and *E* presents the source signal as perceived by the human vision system (HVS). The source follows a model of *C* = *sU*, where *s* is estimated through low-pass filtering, and *U* is a Gaussian random variable characterizing the local variability. The HVS is modeled as white additive noise channel with noise *N*. *I*(*C*; *E* | *s*) represents the maximum information that can be conveyed through this channel. Similarly, *D* represents the distorted signal and *F* represents the distorted signal as perceived by the HVS. The distortion is modeled by two terms: 1) a gain term *g*, and 2) an additive noise term *V* ~ N(0, σ*V*2). Given the input *C* and the output *D*, the parameters *g* and σ*V* can be solved by maximum likelihood estimation (MLE) with the following solution:

,

where is the covariance between *C* and *D*.

VMAF implements the pixel-domain version of VIF, where the coefficients are calculated in the pixel domain after successive Gaussian filtering. In total, coefficients are calculated in 4 different scales. The VIF value at scale () is calculated as:

The VIF scores calculated in the 4 different scales are submitted into the SVR as 4 distinct features.

## Detail loss measure (DLM)

DLM uses 4-scale Daubechies 2 (db2) wavelets. After the wavelet decomposition, the target image *T* is decomposed into a restored image *R* and an additive impairment image *A*, guided by the original image *O*. Specifically: , where the restored image is calculated as:

Here *T*, *R*, *A* and *O* each represents a coefficient of a location (*x*, *y*) in a sub-band of scale .

The restored signal *R* and the original signal *O* go through a contrast sensitivity function (CSF) and a contrast masking (CM) function. The formula of DLM is summarized as:

The coefficients in *R* after CSF and CM (and *O* after CSF) are then Minkowski-pooled with power 3 and summed within the center region of each subband with a border factor 0.1. The results are then summed over the vertical (), horizontal () and diagonal () subbands, and then over the 4 scales ().

*Temporal pooling*

The sequence VMAF value is obtained as an average of the corresponding frame VMAF values.

#### Metric code name

VMAF quality metric values shall be provided under the ‘vmaf’ metric code name.

#### 

#### Sample storage format

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

#### Decoding operation

Given stored 16-bit integer value x, the corresponding VMAF value shall be derived as follows (expressed in floating point):

VMAF = (real) (x) / 512;

*Bibliography*

*Add the following subclauses at the end of the clause.*

[ 6 ] VMAF metric implementation: <https://github.com/Netflix/vmaf>