INTERNATIONAL STANDARD

ISO/IEC 15938-15

First edition
2019-07

Information technology — Multimedia content description interface —

Part 15: Compact descriptors for video analysis

Technologies de l'information — Interface de description du contenu multimédia —
Partie 15: Descripteurs compacts pour analyse de vidéo

Reference number
ISO/IEC 15938-15:2019(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 specified 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 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 Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

A list of all parts in the ISO/IEC 15938 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 http://www.iso.org/members.html.

Introduction

ISO/IEC 15938 (all parts), also known as "Multimedia content description interface", provides a standardized set of technologies for describing multimedia content. It addresses a broad spectrum of multimedia applications and requirements by providing a metadata system for describing the features of multimedia content.

The following are specified in this ISO/IEC 15938 (all parts):

Description schemes (DS) describe entities or relationships pertaining to multimedia content. Description schemes specify the structure and semantics of their components, which may be description schemes, descriptors or datatypes.

Descriptors (D) describe features, attributes or groups of attributes of multimedia content.

Datatypes are the basic reusable datatypes employed by description schemes and descriptors.

Description definition language (DDL) defines description schemes, descriptors and datatypes by specifying their syntax, and allows their extension.

Systems tools support delivery of descriptions, multiplexing of descriptions with multimedia content, synchronization, file format, etc.

The ISO/IEC 15938 series is subdivided into 15 published parts with further parts in development:

— **Part 1: Systems**: specifies the tools for preparing descriptions for efficient transport and storage, compressing descriptions, and allowing synchronization between content and descriptions.

— **Part 2: Description definition language**: specifies the language for defining the series set of description tools (DSs, Ds and datatypes) and for defining new description tools.

— **Part 3: Visual**: specifies the description tools pertaining to visual content.

— **Part 4: Audio**: specifies the description tools pertaining to audio content.

— **Part 5: Multimedia description schemes**: specifies the generic description tools pertaining to multimedia including audio and visual content.

— **Part 6: Reference software**: provides a software implementation of the series.

— **Part 7: Conformance testing**: specifies the guidelines and procedures for testing conformance of implementations of the series.

— **Part 8: Extraction and use of MPEG-7 descriptions**: provides guidelines and examples of the extraction and use of descriptions.

— **Part 9: Profiles and levels**: provides guidelines and standard profiles.

— **Part 10: Schema definition**: specifies the schema using description definition language.

— **Part 11: MPEG-7 profile schemas**: listing of profile schemas using description definition language.

— **Part 12: Query format**: contains the tools of the MPEG query format (MPQF).

— **Part 13: Compact descriptors for visual search**: specifies an image description tool for visual search applications.

— **Part 14: Reference software, conformance and usage guidelines for compact descriptors for visual search**: provides the reference software and guidelines, specifies the conformance testing.

— **Part 15: Compact descriptors for video analysis (this document)**: specifies a video description tool designed to enable efficient and interoperable video analysis applications, allowing visual content matching in videos.

The structure of this document is as follows:

— Clause 5 specifies the binary representation syntax and descriptor component semantics for a CDVA descriptor.

— Clause 6 specifies the extraction and encoding process for a CDVA descriptor.

— Annex A specifies recommended values for the parameters of the encoding process of Clause 6.

— Annex B specifies parameters and a neural network model of the deep feature extraction process.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO and IEC take no position concerning the evidence, validity and scope of this patent right. The holder of this patent right has assured ISO and IEC that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO and IEC. Information may be obtained from:

Joanneum Research Forschungagesellshaft mbH

Leonhardstrasse 59

8010 Graz, Austria

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified above. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — Multimedia content description interface —

Part 15: Compact descriptors for video analysis

1 Scope

This document addresses descriptor technology for search and retrieval applications, i.e. for visual content matching in video. Visual content matching includes matching of views of large and small objects and scenes, with robustness to partial occlusions as well as changes in vantage point, camera parameters and lighting conditions. The objects of interest comprise planar or non-planar, rigid or partially rigid, textured or partially textured objects, but exclude the identification of people and faces. The databases can be large, for example broadcast archives or videos available on the internet. Such applications thus require video descriptors that enable matching with smaller descriptor sizes and shorter runtimes as compared to application enabled by single-frame (still image) descriptors (e.g. CVDS, ISO/IEC 15938-13) in the video domain.

Compact descriptors for video analysis for search and retrieval applications:

— enable design of interoperable object instance search applications;

— minimize the size of video descriptors;

— ensure high matching performances of objects (in terms of accuracy and complexity);

— enable efficient implementation of those functionalities on professional or embedded systems.

This document provides a complementary tool to the suite of existing standards, such as ISO/IEC 15938-13.

2 Normative references

The following referenced documents are indispensable for the application 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 15938-13:2015, *Information technology — Multimedia content description interface — Part 13: Compact descriptors for visual search*

Neural Network Exchange Format, The Khronos Group, Version 1.0, Revision 3, 2018-06-13.

RFC 3986, *Uniform Resource Identifier (URI): Generic Syntax,* Jan. 2005.

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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/

3.1
image descriptor
descriptor extracted from a single *key frame* (3.6) sampled from the *input video* (3.8), which contains *global descriptor* (3.2), *local feature descriptor* (3.3) and *deep feature descriptor* (3.4)

Note 1 to entry: Image descriptors are encoded as described in Clause 6.

3.2
global descriptor
aggregation of local feature descriptors into a compact representation of the *image* (3.5)

Note 1 to entry: The aggregation is as described in subclause 6.1.2.

3.3
local feature descriptor
descriptor of a local region, extracted around an interest point (a point in an *image* (3.5) showing detection stability under local and global perturbations in the image domain, including perspective transformations, changes in image scale, and illumination variations)

Note 1 to entry: The extraction is as described in subclause 6.1.3.

3.4
deep feature descriptor
feature descriptor extracted from a layer of a trained convolutional neural network

Note 1 to entry: The extraction is as described in subclause 6.1.4.

3.5
image
input *key frame* (3.6) to the *image descriptor* (3.1) encoder

Note 1 to entry: The image is as described in Clause 6.

3.6
key frame
frame extracted from the *input video segment* (3.7) by the frame difference process of colour histogram

Note 1 to entry: The extraction is as described in subclause 6.2.

3.7
input video segment
time range (temporal segment) of a video and from which a descriptor is extracted

3.8
input video
image sequence to be processed by the system containing a number of *input video segment(s)* (3.7) to CDVA extraction process

Note 1 to entry: Input video is as described in Clause 6.

3.9
segment descriptor
descriptor extracted from the sampled *key frames* (3.6) of an *input video segment* (3.7)

Note 1 to entry: Segment descriptors are encoded as described in Clause 6. They are contructed from the *image descriptors* (3.1) of the sampled key frames of the input video segment.

3.10
representative frame
frame of an *input video segment* (3.7) for which an uncompressed descriptor is represented and which is used as the basis for differential encoding

3.11
pixel
indexable element on an integer grid of the original image or the converted image, comprising spatial coordinates, a luminance value and (optional) chrominance values

4 Abbreviated terms, operators, mnemonics, functions and symbols

4.1 General

The mathematical symbols used in this document are similar to those used in the C programming language. However, integer divisions with truncation and rounding are specifically defined. Numbering and counting loops generally begin with zero.

4.2 Abbreviated terms

ABAC	adaptive binary arithmetic coding
CDVA	compact descriptors for visual analysis as defined by this document
CDVS	compact descriptors for visual search as defined by ISO/IEC 15938-13
CNN	convolutional neural network
MPEG-7	ISO/IEC 15938 (all parts)
NIP	nested invariance pooling
NN	neural network
NNEF	neural network exchange format as defined by the Khronos specification referenced in Clause 2
PCA	principal component analysis
RGB	red-green-blue colour space
ROI	region of interest
SCFV	scalable compressed fisher vector
URI	uniform resource identifier as defined by RFC 3986
XOR	binary exclusive OR operation

4.3 Arithmetic operators

+	addition
-	subtraction (as a binary operator) or negation (as a unary operator)
++	increment, i.e. x++ is equivalent to x = x+1
--	decrement, i.e. x-- is equivalent to x = x–1
*	multiplication
×	multiplication

^	power
/	integer division with truncation of the result towards zero For example, 7/4 and –7/–4 are truncated to 1, and –7/4 and 7/–4 are truncated to –1
//	integer division with rounding to the nearest integer; half-integer values are rounded away from zero unless otherwise specified For example, 3//2 is rounded to 2, and –3//2 is rounded to –2.
÷	indicates division in mathematical equations where no rounding is intended
%	modulus operator, defined only for positive numbers
ceil	minimum integer number greater than or equal to the given floating point number
sqrt	square root

4.4 Logical operators

||	logical OR
&&	logical AND
!	logical NOT
⊕	bit-wise difference (XOR) operator

4.5 Relational operators

>	greater than
>=	greater than or equal to
≥	greater than or equal to
<	less than
<=	less than or equal to
≤	less than or equal to
==	equal to
!=	not equal to

4.6 Bitwise operators

|	OR
&	AND

4.7 Interval specification

[a;b]	inclusive range from a to b

4.8 Mnemonics

The following mnemonics are defined to describe the different data types used in the coded bitstream.

bslbf — bit string, left bit first, where "left" is the order in which bits are written in this document

Bit strings are generally written as a string of 1s and 0s within single quote marks, e.g. '1000 0001'. Blanks within a bit string are for ease of reading and have no significance. For convenience, large strings are occasionally written in hexadecimal, in which case conversion to a binary in the conventional manner will yield the value of the bit string. Thus, the left-most hexadecimal digit is first and in each hexadecimal digit the most significant of the four digits is first.

uimsbf — unsigned integer, most significant bit first

vlclbf — variable length code, left bit first, where "left" refers to the order in which the VLC codes are written in this document

The byte order of multibyte words is most significant byte first.

4.9 Functions

$\text{argmax}_i()$ — maximum value in argument list

$\text{argmin}_i()$ — minimum value in argument list

δ_g — distance function for global descriptors

δ_l — distance function for local descriptors

$\sum_{i=a}^{i<b} f(i)$ — summation of $f(i)$ with i taking integer values from a up to, but not including b

L0 norm — $\mathrm{L0}(\mathrm{x},\mathrm{y}) = \|x-y\|_0 = \sum_i \delta(x_i - y_i)$, where $\delta(\mathrm{a}) = \begin{cases} 1(a!=0) \\ 0(a=0) \end{cases}$

L1 norm — $\mathrm{L1}(\mathrm{x},\mathrm{y}) = \|x-y\|_1 = \sum_i |x_i - y_i|$

L2 norm — $\mathrm{L2}(\mathrm{x},\mathrm{y}) = \|x-y\|_2 = \mathrm{sqrt}\left(\sum_i (x_i - y_i)^2\right)$

Euclidean distance — $\mathrm{D}(x,y) = \mathrm{sqrt}\left(\sum_i (x_i - y_i)^2\right)$

hist(I,C) — histogram of image I for colour channel C

4.10 Symbols

β — selection priority of local feature

c — number of channels of feature map (dimension of descriptor extracted from CNN)

Δ_k — deep feature descriptor for frame k

D_k — binarized deep feature descriptor for frame k

f — feature vector of local descriptor

$\boldsymbol{G}$	set of global descriptors
G_k	global descriptor of frame k
γ_k	result of pooling operation for feature map for frame k
h	feature map height
I_k	RGB image of frame k
k	key frame index
m	feature map index
n_k	number of key frames in a video segment
n_l^k	number of local descriptors of frame k
n_r	number of rotation transformations
n_s	number of scale transformations
ρ	representative frame of video segment
p_r, p_s, p_t	statistical moment for pooling operation
P_s	scale invariance pooling
P_r	rotation invariance pooling
P_t	translation invariance pooling
q	quantization function
$\boldsymbol{L}^k$	set of local descriptors of frame k
L	list of local feature descriptors of a video segment
l_i^k	i^{th} local descriptor of frame k
θ_l	threshold for local descriptor distance
w	feature map width
x	horizontal image coordinate
y	vertical image coordinate

5 CDVA bitstream syntax

5.1 CDVA descriptor

5.1.1 Binary representation syntax

CDVADescriptor {	Number of bits	Mnemonics
CDVAHeader	≥80	vlclbf
for (i=0; i<NrSegments; i++) {		