TECHNICAL REPORT

ISO/IEC TR 15938-8

First edition
2002-12-15

AMENDMENT 6
2011-11-01

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

Information technology — Multimedia content description interface —

Part 8: Extraction and use of MPEG-7 descriptions

AMENDMENT 6: Extraction and matching of video signature tools

Technologies de l'information — Interface de description du contenu multimédia —

Partie 8: Extraction et utilisation des descriptions MPEG-7

AMENDEMENT 6: Extraction et correspondance des outils de signature vidéo

Reference number
ISO/IEC TR 15938-8:2002/Amd.6:2011(E)

© ISO/IEC 2011

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2011

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide to publish a Technical Report. A Technical Report is entirely informative in nature and shall be subject to review every five years in the same manner as an International Standard.

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.

Amendment 6 to ISO/IEC TR 15938-8:2002 was prepared jointly by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

© ISO/IEC 2011 – All rights reserved

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

Information technology — Multimedia content description interface —

Part 8:
Extraction and use of MPEG-7 descriptions

AMENDMENT 6: Extraction and matching of video signature tools

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

After 4.9.1, add:

4.9.2 Video Signature

The visual content descriptors in Clauses 6-9 of ISO/IEC 15938-3:2002 are very useful when trying to find videos with similar content. These descriptors are intended to be general and were found to be unsuitable for the task of finding duplicate content. The video signature descriptor is designed to identify duplicate video content. This descriptor is robust to a wide range of common video editing operations, but is sufficiently different for every original content to identify it reliably.

The video signature is composed of three main elements: a frame signature, a set of compact summary frame signatures, referred to as words, and a group-of-frames representation for a temporal segment, referred to as a bag of words.

4.9.2.1 Extraction

11.4.5 to 11.4.8 of ISO/IEC 15938-3:2002 describe the extraction of the video signature.

4.9.2.2 Matching

A Video Signature is composed of multiple temporal segments, each represented by a BagOfWords element, and multiple frames, each represented by a FrameSignature element and a FrameConfidence element.

The matching between two Video Signatures v^1 and v^2 is carried out in three stages, designed to maximize matching speed and true positives and to minimize false positives. The first stage uses the BagOfWords element to identify candidate matching segments. The second stage uses the FrameSignature element to identify candidates of frame rate ratio and temporal offset between the candidate matching segments. The third stage performs frame-by-frame matching to determine candidate matching intervals using the FrameSignature and FrameConfidence elements, and then determines the best match between the Video Signatures v^1 and v^2. These matching stages are explained in more detail below.

Stage 1 (Segment matching with BagOfWords)

All of the temporal segments of Video Signature v^1 are compared with all of the temporal segments of Video Signature v^2.

For two segments f^1 and f^2, their similarity is assessed by comparing the bag-of-words representation for each vocabulary *j* and merging the results to reach a decision. More specifically, for $BagOfWords^1[j]$ and $BagOfWords^2[j]$, their distance is measured by the Jaccard distance metric given by

$$\underset{J}{D_j}\left(BagOfWords^1[j], BagOfWords^2[j]\right) = 1 - \frac{\#(BagOfWords^1[j] \cap BagOfWords^2[j])}{\#(BagOfWords^1[j] \cup BagOfWords^2[j])}$$

where # denotes the number of elements in a set. This measures the distance of the segments f^1 and f^2 in a given vocabulary as a function of the distinct words they have in common and all the distinct words that they contain jointly.

For $Q = 5$ vocabularies, we have Jaccard distances $\underset{J}{D_0}$, $\underset{J}{D_1}$, …,. $\underset{J}{D_{Q-1}}$. These distances are fused to give the composite distance $\underset{J}{D}$ as

$$\underset{J}{D} = \sum_{k=0}^{Q-1} \underset{J}{D_k}$$

Then a decision on the similarity of the segments is reached by thresholding each of Jaccard distances $\underset{J}{D_0}$, $\underset{J}{D_1}$, …,. $\underset{J}{D_{Q-1}}$, and the composite distance $\underset{J}{D}$. That is, the segments f_i^1 and f_j^2 are passed to stage 2 of matching if more than half of the Q Jaccard distances $\underset{J}{D_0}$, $\underset{J}{D_1}$, …,. $\underset{J}{D_{Q-1}}$ are less than a threshold T_1 and the composite distance $\underset{J}{D}$ is less than another threshold T_2, otherwise they are declared not matching.

Stage 2 (Frame rate ratio & time shift estimation using Hough transform)

For the segment pairs passed to this stage, a Hough transform is used to estimate the temporal parameter differences, i.e. frame rate ratio and time shift, between the segments. These are linear properties and can therefore be estimated using two strongly corresponding frame pairs.

First, the L1 distance between the FrameSignature elements of the frame pairs between the segments are calculated and the pairs whose distance is smaller than a threshold are selected as strongly corresponding frame pairs. Then, two strongly corresponding frame pairs are selected to calculate the time shift and frame rate ratio, and the bin corresponding to the calculated parameters in the Hough space is incremented. This is done for all possible combinations of two strongly corresponding frame pairs. Finally, multiple temporal parameters with high response in the Hough space are selected as candidate parameters, and are passed to stage 3 of matching. If the highest response in the Hough space is below a certain threshold, the segment pairs are declared not matching.

Stage 3 (Frame-by-frame matching on frame signature)

The matching interval (the start and end position of the match) is determined by temporal interval growing based on a frame-by-frame matching on the full frame signature. The candidate temporal parameters between two sequences are used in this frame-by-frame matching.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

© ISO/IEC 2011 – All rights reserved

First, the estimated time shift is used to determine the initial temporal matching position. Then, using the estimated frame rate ratio, the temporal interval is extended frame-by-frame towards both directions by calculating the L1 distance between the FrameSignature elements of corresponding frames. The temporal extension stops when the distance exceeds a certain threshold in order to determine the matching interval. If the length of the matching interval is shorter than a given minimum duration, the matching interval is eliminated as a non-match. Otherwise, the FrameConfidence element associated with each frame in the matching interval is checked to verify the match. The overall confidence of the matching interval is calculated as the ratio of the number of frames which has a FrameConfidence that is higher than a certain threshold. If the overall confidence is below a certain level, the matching interval is eliminated as a false match caused by frames with low information content.

This process is carried out for all of the candidate temporal parameters, thus generating multiple candidates of matching intervals. Then, one candidate interval is selected as the best matching result, based on the mean L1 distance of FrameSignatures and the length of the interval. The best interval is selected by first selecting multiple intervals with mean L1 distances below a threshold and then selecting amongst them the one which has the longest length.

Clause 7 of ISO/IEC 15938-6:2003 contains an exemplary implementation and source code for this matching technique, including default threshold values.

4.9.2.3 Fast Matching using Index

The process of fast matching using index tables may be used as a pre-filtering step to quickly determine whether two Video Signatures v^1 (a query video) and v^2 (a reference video) have possible matching intervals, in which case the matching process described in 4.9.2.2 follows.

Index tables using the word elements of the Video Signature are utilized in the fast matching process. For each video, index tables for each word ($Q = 5$) are built, which maps the values of the word (0-242) to the frame numbers of which they appear. By using the index tables, we can quickly locate frame pairs between two videos which have the same word values.

The index tables are built before matching is carried out. For the reference video, all of the frames are used to build the index tables, referred to as all-frame index tables. For the query video, only selected keyframes are used to build the index tables, referred to as keyframe index tables. The keyframe selection of a query video is carried out using a keyframe detection algorithm. The proposed algorithm uses the FrameSignature element, and proceeds as follows.

Keyframe selection for a query video

1. Calculate L1 distance between FrameSignature elements of each frame and its previous frame.

2. Set a sliding window and find the maximum L1 distance within the window. The frame which the distance with its previous frame gives the maximum value is a keyframe candidate. A recommended size of the sliding window is 4 seconds. In case the expected matching duration is less than 4 seconds, the recommended window size is half of the expected matching duration. The keyframe candidates for the whole video should be selected by sliding the window by one frame.

3. If consecutive frames are selected as keyframe candidates, remove all such consecutive keyframe candidates except for the last one.

4. If the FrameConfidence element of the keyframe candidate is smaller than a predefined threshold, discard that keyframe candidate and select as a keyframe candidate the first posterior frame of which the FrameConfidence element is larger than or equal to the threshold.

5. The remaining keyframe candidates are selected as keyframes.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

The matching process between a query video and a reference video uses the keyframe index tables of the query video and all-frame index tables of the reference video to identify matching frame pairs between the query and the reference videos. The detailed matching procedure is as follows.

Matching procedure using index tables

1. For each word ($Q = 5$), keyframe index table of the query video and all-frame index table of the reference video are compared to identify frame pairs with the same word value between the query and the reference video. Frame pairs having the same value for multiple of the $Q = 5$ words are selected as candidate of matching frame pairs to be passed to the next step.

2. Each candidate frame pair is further verified by calculating the L1 distance between the FrameSignature elements. This can be done in two steps, 1) first by calculating the L1 distance using only the 25 out of 380 dimensions used to construct the words, 2) then by calculating the L1 distance for the whole 380 FrameSignature dimensions. If the L1 distance is smaller than a predefined threshold, the frame pairs are identified as matching frame pairs.

3. If the number of the matching frame pairs between the query and reference video identified in the previous step is larger than a predefined threshold, the query and the reference videos are passed to the matching process described in 4.9.2.2 for identifying the matching intervals.

Clause 7 of ISO/IEC 15938-6:2003 contains an exemplary implementation and source code for this matching technique, including default threshold values.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 15938-8:2002/AMD6:2011

© ISO/IEC 2011 – All rights reserved