**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N15338**

**July 2015, Warsaw, Poland**

|  |  |
| --- | --- |
| **Source** | **Requirements** |
| **Status** | **Approved** |
| **Title** | **Evaluation Framework for Compact Descriptors for Video Analysis - Search and Retrieval** |

# Introduction

This document outlines an evaluation framework (version 1.0) for assessing the responses to the Call for Proposals on Compact Descriptors for Video Analysis addressing Search and Retrieval applications [1]. It focuses on scenarios of visual search in video, primarily in the media and entertainment domain. Please note that further updates to the CDVA Evaluation Framework may be issued in the future, so users are advised to check the latest MPEG output documents to ensure the use the latest version.

# Objectives and definitions of terms

The evaluation framework is designed to assess the performance of descriptors in the task of identifying scenes and views of objects in a database of videos. The objects of interest comprise planar or non-planar, rigid or partially rigid, textured or partially textured objects. The objective excludes identification of people and faces as there are existing methods and standards specifically addressing these functionalities. The databases can be large, for example broadcast archives or videos available on the Internet.

The following definitions and terms are used throughout this document.

**Query images or videos / Query set:** images or videos used to initiate search.

**Reference videos / Reference set:** videos containing views of objects or scenes intended to be identified.

**Distractor videos / Distractor set:** videos containing unrelated content w.r.t. query and reference (i.e. no frame contains views of objects or scenes intended to be identified).

**Dataset:** the query set and reference set.

**Database:** the dataset and distractor set or their descriptors and associated information.

**Database index:** file or set of files, enabling fast retrieval of all database items (including distractors).

**Relevant videos or video segments:** reference videos or segments in video that contain view(s) of object(s) or the scene present in the query.

**Identified video segments:** list of all video segments returned in response to a query request. This will include video material ID, and start/end frame for the segment.

**Precision:**  the ratio betweenthe number of relevant items among all identified items and the number of identified items, where items are video segments.

**Recall:** the ratio between the number of relevant items among the identified items and the total number of relevant items, where items are video segments.

**Rank:** position of an item in the ranked list of identified items, where items are video segments.

**Precision at a given cut-off rank r for a single query:**

P(r) = (number of relevant items of rank r or less) / r

**Recall at a given cut-off rank r for a single query:**

R(r) = (number of relevant items of rank r or less) / (number of relevant items)

**Average precision:** defined as follows

****

where N is the number of identified items, R is the number of relevant items, and rel(r) = 1 if item at rank r is relevant, 0 otherwise.

**Mean average precision:** average precision for a set of queries is defined as follows:



where Q is the number of queries.

**Success rate for top match:** (number of times the top identified item is relevant) / (number of queries).

**Spatio-temporal Localization information:** description of temporal and spatial location of objects within the reference and query videos. Please note that version 1 of the CVDA Evaluation Framework does not specify spatial locations of objects, but spatial location may be specified in subsequent revisions of the evaluation frameworks.

# Datasets and annotations

## Original videos and images

The evaluation will be conducted on 5,289 videos and images depicting 796 items of interest across three different categories, namely large objects, small objects and scenes. The videos have durations of between 1 sec to 1 min+. Table 1 summarises the numbers of items of interest and their instances, i.e. their videos and images, for each category. This summery is for information purposes only, as this categorisation is not used in the evaluation framework.

Table 1. Number of Items and Instances across Categories.

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Items of interest | Instances | |
| Videos | Images |
| Large objects \*1 | 489 | 2922 | 50 |
| Small objects \*2 | 236 | 1707 | 180 |
| Scenes \*3 | 71 | 400 | 30 |
| Total | 796 | 5029 | 260 |

\*1 Stationary large objects e.g. buildings, landmarks (most likely background objects, possibly partially occluded or a close-up).

\*2 Generally smaller items (e.g. paintings, books, CD covers, products) which typically appear in front of background scenes, possibly occluded.

\*3 Scenes (e.g. interior scenes, natural scenes, multi-camera shots, etc.). General camera recordings showing the same scenes or actions from different view-points.

Table 2 shows the compositions of the original videos and images in terms of Broadcast and User-Generated Content (UGC).

Table 2. Number of Items and Instances for Broadcast and UGC.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Items of interest | Instances | |
| Videos | Images |
| Broadcast | 181 | 1283 | 0 |
| UGC | 615 | 3746 | 260 |
| Total | 796 | 5029 | 260 |

## Query set

All images for the items of interest are used as query images.

For the videos, the original videos of ~80% of the items of interest were embedded in unrelated content of duration 5-30 seconds before the original video and 5-30 seconds after the original video to create partial queries, and the original videos of the remaining ~20% of the items of interest were not embedded in unrelated content, giving direct queries. It should be noted that the CDVA evaluation framework does not distinguish between direct and partial queries and that proposed methods should be able to cope with both types without prior knowledge of whether a query is direct or partial.

Then, the query videos of ~20% of the items of interest were modified according to Table 3, to produce modified queries. It should be noted that the CDVA evaluation framework does not distinguish between unmodified and modified queries and that proposed methods should be able to cope with both types without prior knowledge of whether a query is modified or not.

Table 4 summarises the query set.

Table 3. Query video modifications.

|  |  |
| --- | --- |
| Modification | Strength |
| Text/Logo overlay | ~20% of screen area |
| Frame rate | 7 fps |
| Interlaced/Progressive Comversion | P🡪I🡪P or I🡪P |
| Transcoding | MPEG-2 SD @ 512kbps |
| Color to monochrome and contrast change | 20% contract increase |
| Add grain | fine motion picture grain (modern 35mm material), monochrome, signal dependent (stronger on mid tones) |
| Captured displayed content | "Zoom out": Clips should be at least 20% of area  "Zoom in": At least 50% of clip should be visible |

Table 4. Query set.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Items of interest | Instances | |
| Videos | Images |
| Unmodified  (Direct & Partial) | 796 | 5029 | 260 |
| Modified  (Direct & Partial) | 123 | 4686 | 0 |
| Total | 796 | 9715 | 260 |

## Reference set

No images for the items of interest are used as references.

For the videos, the original videos of ~80% of the items of interest were embedded in unrelated content (different to that used in the queries) of duration 5-30 seconds before the original video and 5-30 seconds after the original video to create partial references, and the original videos of the remaining ~20% of the items of interest were not embedded in unrelated content, giving direct references. It should be noted that the CDVA evaluation framework does not distinguish between direct and partial references and that proposed methods should be able to cope with both types without prior knowledge of whether a reference is direct or partial.

No modifications were applied to the reference videos.

Table 5 summarises the reference set.

Table 5. Reference set.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Items of interest | Instances | |
| Videos | Images |
| All | 796 | 5029 | 0 |

## Distractor set

The database for retrieval experiments is formed by the reference set and the distractor set, which consist of content unrelated to the relevant items. Table 6 summarises the distractor set.

Table 6. Distractor set.

|  |  |  |
| --- | --- | --- |
| Source | Videos | Type |
| MediaEval Blip | 4701 | UGC |
| OpenImages.eu (various collections) | 3789 | Broadcast, archival, education |

## Annotations

Temporal localisation annotations are provided for all videos in the query and reference set at the shot or group of shots level.

Spatial localisation annotations are not provided in version 1 of the evaluation framework, but may be provided in subsequent releases.

Experiment setup

The performance of proposals will be evaluated using two types of experiments:

* retrieval experiment, and
* pairwise matching experiment.

## Retrieval experiment

The retrieval experiment is intended to assess performance of proposals in the context of a video retrieval system. The block diagram of this process is shown in Figure 1.

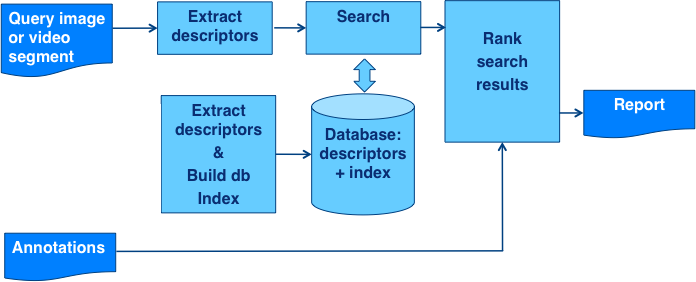


Figure 1. Retrieval experiment setup.

Descriptors for the query set are extracted. The reference set and the distractor set are used to create a database index. The database index and a descriptor of each query are used to perform retrieval operations. The pipeline returns results in the form of a ranked list of matches in the database. This list as well as ground truth information are used to compute retrieval performance.

## Pairwise matching experiment

The pairwise matching experiment is intended for assessing performance of proposals in the context of an application that uses descriptors for the purpose of video/image content matching, that is detection of relevant object in video/image. The overall scheme of this experiment is shown in Figure 2.



Figure 2. Pairwise matching experiment setup.

Pairwise matching is performed for videos in the query set against matching and non-matching videos in the reference set. The result of pairwise matching is a match / no match decision, and is then compared to ground truth. If ground-truth spatial and temporal localization data are provided, the accuracy of localization is also assessed.

# Performance measurements

This section describes performance measures to be used in the CDVA Evaluation Framework. To ensure fair and consistent comparison, MPEG is preparing automated evaluation scripts, which shall be released as the output from 113 MPEG meeting, by the end of October 2015.

## Descriptor lengths

The performance of proposals will be assessed for the following operating points (upper bounds on average descriptor lengths across all query clips):

* 16KB, 64KB and 256KB per second of video content

It must be understood that descriptors generated at any one of these operating points should allow retrieval and matching operations with descriptors generated at different operating points.

This will be tested using pairwise matching experiments with descriptors generated at the following combination of lengths:

* 16KB against 256KB

All reported results will be considered in assessing overall performance of proposal.

Average lengths of descriptors at each operating point shall be reported. The worst case length of descriptors shall also be reported.

## Performance characteristics

The information about performance of proposals shall be provided using the following set of measures.

**Retrieval performance**

* Mean Average Precision
* Success rate for top match
* Recall at k, where k is the number of reference clips for each query

**Pairwise matching performance**

* Success rate at given false alarm rate.

where:

* for the NMF (segments not containing query object):
  + **False positive rate (false alarm rate)** = (FP) / (FP+TN)
* for the MF (segments matching the query object):
  + **True positive rate (success rate)** = TP / (TP + FN)

where: TP the number of true positives FP the number of false positives

TN the number of true negatives FN the number of false negatives

For initial evaluation of proposals the target false alarm rate is set to 1%.

It must be understood that for some applications CDVA descriptors should support operation at lower false alarm rates. Therefore evaluations at lower false alarm rates may be introduced during core experiment stage.

## Complexity measurements

Complexity shall be measured by proponents using retrieval and verification experiments, reporting the following characteristics:

* average time of descriptor extraction
* average time of retrieval
* average time of pairwise matching

These numbers shall be reported for all bitrates used for evaluation of performance. This report shall include specification of the machine used for experiments.

The timing shall be measured when execution (affinity mask) is constrained to a single thread/ single CPU operation. Time measurements functions to be used are provided in C/C++ example modules in the evaluation software.

All time measurements will be verified on-site during 114th MPEG meeting by running delivered software modules on a reference platform. The reference platform shall consist of a desktop workstation with a single i7 6-core CPU (Intel 5930K), 32GB of RAM and 1TB SSD/8TB HDD. For time measurements processing should be performed only on the CPU without GPU acceleration.

Timings will be evaluated on appropriately selected subsets of data, which will be provided with the evaluation scripts.

The following limits for average time numbers on the reference platform must be met:

* extraction time: must not exceed 10 seconds per second of decoded video content.
* pairwise matching time: must not exceed 1 second per pair.
* retrieval time: must not exceed 60 seconds per query.

Only proposals that meet these limits will be further considered, and the decision on which one to select will be based mostly on accuracy characteristics. Complexity information will be used to decide among best performing proposals but only as a secondary characteristic.

# References

[1] Call for Proposals for Compact Descriptors for Video Analysis - Search and Retrieval (CDVA), ISO/IEC JTC1/SC29/WG11 N15339, Warsaw, Poland, June 2014.