**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/N17123**

**October 2017, Macao, CN**

**Title: Description of Core Experiments on CDVA**

**Source Video Subgroup**

**Status: Approved**

**Editors: Werner Bailer, Miroslaw Bober, Lingyu Duan, Yihang Lou, Yan Bai**

# Introduction

The purpose of this document is to provide descriptions of six core experiments on MPEG Compact Descriptors for Video Analysis (CDVA) [1]. The previous six CEs are being continued.

The results of experiments will be discussed on the reflector before the 121th MPEG meeting.

The report of each CE should include; (1) a comparison between the tested solutions, (2) recommendation to the AhG based on the results of the CE.

The results according to the evaluation criteria described below shall be reported (in particular those defined by the evaluation framework [2]), but results based on other evaluation criteria may be included.

If learning-based methods are used, then the CDVA data set must not be included for training. The data sets used for training shall be reported. The details of the network configuration being used shall be included in the CE response (or as a reference to an input document that contains these details).

# Timeline

* Dec. 1, 2017: CXM 3.0 released
* Jan. 8, 2018: Submission of responses to CEs and crosschecks
* Jan. 15, 2018: Submission of report on CEs

# CE1 Temporal Sampling

## Background

Proposals submitted in response to the CDVA CfP [6,7] make use of some key frames, which provide a self-contained feature description, or are used to predict global and local features of other frames in the sequence. Two proposals further use global descriptor similarity for temporal segmentation.

The CXM is built around these common components of the different proposals, and will be used in the CE with different parameterisation concerning the temporal sampling rate and discarding sampled frames based on visual similarity.

Results submitted in response to CE1 at the 117th  meeting [18] give some insight on the impact of temporal sampling rates and similarity thresholds and the maximum matching and retrieval performance that can be reached, as well as on the differences between visual activity and medoid based sampling. However, the results currently are in only slightly overlapping size ranges. More experiments are needed to study the differences of the two approaches at different size configurations.

## Purpose

In order to better understand the impact of the number of key frames and the selection of frames from one segment to be included, the CE aims to

* Compare different key frame densities (varying initial subsampling and similarity based dropping of frames)
* Compare different strategies of selecting the frames of one segment to be included in the descriptor (visual activity vs. medoid based)

In particular, the range in which size vs. performance curves of the visual activity based and medoid based methods overlap is of interest.

## Approach

The CE will use the CXM v3.0 to perform the experiment. The CXM uses the CDVS TM14.2 to extract the descriptors for key frames. The CE has to be performed using CDVS mode 0 for the extraction, changing only the key frame selection parameters (thresholds). Thus, only these parameters will be used to extract descriptors at different operating points, while the size of a key frame descriptor will be fixed.

The CE will change the key frame density in the visual activity based and medoid based methods by varying the following parameters

* skip\_before and skip\_after (keeping the before and after skip rate at the same values)
* drop\_frame\_th
* for the medoid based method, encode\_th

For each run of the experiment, the same temporal sampling, similarity thresholds and CDVS mode parameters must be applied to references and queries in pairwise matching and retrieval experiments, and to distractors in retrieval experiments.

## Measures

The evaluation of the results is to be performed as defined by the evaluation framework. In particular, the results for the following measures shall be reported:

* Size of the resulting bitstream (Bps).
* Matching (including temporal localisation) and retrieval performance.

Runtime and memory consumption of extraction, matching and retrieval shall be reported for all experiments as defined in the evaluation framework. The runtime/memory measurements for all experiments in one response to the CE (including the measurements for CXM v3.0) shall be performed on the same machine using CPU only.

## CE Coordinator

Massimo Balestri (Telecom Italia, [massimo.balestri@telecomitalia.it](mailto:massimo.balestri@telecomitalia.it))

AIT is responsible for performing crosschecks on this CE.

# CE2 Segment-level representation for matching and retrieval

## Background

Choosing descriptors that represent a segment have the potential to speed up the matching process, as they can help to make an early decision about whether any of the key frames in the segment is likely to match or not. The actual performance gain has however not been measured so far.

## Purpose

The experiment shall assess whether alternative matching methods making use of one or few representative descriptors per segment (e.g. medoid) can improve matching and retrieval efficiency, and determine any related change in performance.

## Approach

The CE will use the CXM v3.0 to perform the experiment. The CXM uses the CDVS TM14.2 to extract the descriptors for key frames. The CE has to be performed using CDVS mode 0 for the extraction, changing only the key frame selection parameters. Thus, only these parameters will be used to extract descriptors at different operating points, while the size of a key frame descriptor will be fixed.

The CE will use a method also evaluated in a response to CE1, and determine the improvement in matching/retrieval complexity when making use of segment-level representations.

For each run of the experiment, the same temporal sampling, similarity thresholds and CDVS mode parameters must be applied to references and queries in pairwise matching and retrieval experiments, and to distractors in retrieval experiments.

## Measures

The evaluation of the results is to be performed as defined by the evaluation framework. In particular, the results for the following measures shall be reported:

* Size of the resulting bitstream (Bps).
* Matching (including temporal localisation) and retrieval performance.
* Fraction of segment descriptors that have been matched completely and matched partly (specifying the number of local/global descriptor pairs matched between those descriptors).
* The size of descriptors that are discarded versus the size of descriptors that are not discarded also needs to be analysed to make an analysis of the complexity reduction.

Runtime and memory consumption of matching and retrieval shall be reported for all experiments as defined in the evaluation framework. In order to achieve reliable runtime measurements, measurements on larger sets as the timing set of the evaluation framework shall be performed. Where feasible, measurements of the actual execution time spent in the modified functions shall be reported, comparing the CXM 3.0 and modified version on the same machine. The runtime/memory measurements for all experiments in one response to the CE (including the measurements for CXM v3.0) shall be performed on the same machine using CPU only.

## CE Coordinator

Werner Bailer (JRS, [werner.bailer@joanneum.at](mailto:werner.bailer@joanneum.at))

AIT is responsible for performing crosschecks on this CE.

# CE3 Descriptor Temporal Prediction

## Background

It has been shown that temporal prediction using lossless and lossy methods provides size reductions of at least 50% without any performance loss, and much higher reductions at low to moderate performance reductions [19]. Different encoding options may also result in differences in encoding and matching complexity. Thus the impact on the computational costs has to be further studied.

In m40387 [22] a binarised representation of the NIP descriptor has been proposed. The lossless and lossy compression of NIP descriptors needs to be further studied.

## Purpose

* Study the bitrate reduction that can be achieved from predictive coding of local (CDVS) and/or global (CDVS or NIP) descriptors.
* Study the impact of lossy encoding on the retrieval and matching performance.
* Study the impact on the complexity of encoding, matching and retrieval.

## Approach

For experiments with CDVS descriptors, the starting point for the CE3 is CXM v3.0. The CE has to be performed using CDVS mode 0 for the extraction, and keeping the same parameters for key frame selection and temporal segmentation.

For experiments with NIP descriptors, the starting point is the method and parameterisation described in m40387. The CE has to be performed using the same parameters for key frame selection and temporal segmentation as CXM 3.0.

Any means of lossless or lossy predictive coding of global and/or local descriptors of a segment can be applied.

## Measures

The following measures for descriptor sizes and performance shall be reported:

* Descriptor sizes, matching, localisation and retrieval performance as defined by the evaluation framework.
* Fraction of descriptor size for global and local descriptors
* Fraction of predicted local and/or global descriptors per segment

Runtime and memory consumption of extraction, matching and retrieval shall be reported for all experiments as defined in the evaluation framework. The runtime/memory measurements for all experiments in one response to the CE (including the measurements for CXM v3.0) shall be performed on the same machine using CPU only.

## CE Coordinator

Werner Bailer (JRS, [werner.bailer@joanneum.at](mailto:werner.bailer@joanneum.at))

AIT is responsible for performing crosschecks on this CE.

# CE4 Combination of Deep-learning-based Descriptors with CDVS Descriptors

## Background

Contributions to the 116th meeting have proposed the use of descriptors based on deep learning [16], and further details have been provided in the response to CE4 at the 117th meeting [20]. The previous CE4 has been split into three different core experiments, each studying a specific question related to deep-learning based descriptors. CE4 addresses the combination of deep-learning based descriptors with CDVS descriptors. Initial results have been reported in [20], but results on a complete data set as specified by the evaluation framework as well as including runtime/memory complexity measurements are needed.

## Purpose

* Study the combination of deep-learning-based descriptors with CDVS local and/or global descriptors
* Study the impact of using deep-learning based descriptors on runtime/memory complexity

## Approach

For descriptor representation and combination, the starting point is CXM v3.0. The parameters for temporal sampling, temporal segmentation and CDVS extraction and matching are the same as in CXM v3.0. The aim is to compare the performance of the CXM with a pipeline that uses deep-learning based descriptors in addition to or in place of the global or local CDVS descriptors in CXM.

## Measures

* Descriptor sizes, matching/localisation/retrieval performance as defined by the evaluation framework.
* Runtime and memory consumption of extraction, matching and retrieval shall be reported for all experiments as defined in the evaluation framework. The runtime/memory measurements for all experiments in one response to the CE (including the measurements for CXM v3.0) shall be performed on the same machine using CPU only. If parts of the extraction, matching or retrieval use GPU, the runtime/memory measurements when using GPU may be reported in addition.

## CE Coordinators

Lingyu Duan (PKU, lingyu@pku.edu.cn), Yihang Lou (PKU, [yihanglou@pku.edu.cn](mailto:yihanglou@pku.edu.cn))

JRS is responsible for performing crosschecks on this CE.

# CE5 Representation of Deep-learning-based Descriptors

## Background

Contributions to the 116th meeting have proposed the use of descriptors based on deep learning [16], and further details have been provided in the response to CE4 at the 117th meeting [20]. The previous CE4 has been split into three different core experiments, each studying a specific question related to deep-learning based descriptors. CE5 addresses the issue of finding more compact representations of the descriptors obtained using deep learning.

## Purpose

* Study approaches for the compact representation of descriptors (e.g., binarisation) obtained from deep learning methods.

## Approach

For the descriptor representation, the starting point is the VGG16 NIP descriptor proposed in [16]. Methods for representing a single VGG16 NIP descriptor more compactly shall be studied, and the impact on matching, localization and retrieval performance shall be analysed.

For the purpose of runtime and memory complexity measurements, a version of the CXM with integrated extraction of NIP descriptors must be provided.

## Measures

* Descriptor sizes, matching/localisation/retrieval performance as defined by the evaluation framework.
* Runtime and memory consumption of extraction, matching and retrieval shall be reported for all experiments as defined in the evaluation framework. The runtime/memory measurements for all experiments in one response to the CE (including the measurements for CXM v3.0) shall be performed on the same machine using CPU only. If parts of the extraction, matching or retrieval use GPU, the runtime/memory measurements when using GPU may be reported in addition.

## CE Coordinators

Lingyu Duan (PKU, lingyu@pku.edu.cn), Yihang Lou (PKU, [yihanglou@pku.edu.cn](mailto:yihanglou@pku.edu.cn))

JRS is responsible for performing crosschecks on this CE.

# CE6 Compression of CNN Model Parameters

## Background

Contributions to the 116th meeting have proposed the use of descriptors based on deep learning [16], and further details have been provided in the response to CE4 at the 117th meeting [20]. The previous CE4 has been split into three different core experiments, each studying a specific question related to deep-learning based descriptors. CE6 addresses the compressed representation of the learned model parameters. Initial results have been reported in [21], but more results are needed in order to understand the influence of using a compressed network and to decide whether this aspect needs to be normative or not in order to ensure interoperability of descriptors.

## Purpose

* Study the efficient representation of the learned CNN model parameters to reduce model storage requirements.
* Study the interoperability of descriptors extracted from the original and compressed CNN, considering different quantisation of weights (in particular configurations with only moderate compression).
* Study the computational complexity of model compression and decompression.

## Approach

In order to address the issues of memory consumption of the stored model parameters, approaches for the efficient representation of these model parameters will be studied. The experiment will use the VGG16 NIP model after pruning the layers above the pool5 layer (as described in m39853 [21]) as a baseline.

During the experiment, the computational complexity of the compression and decompression of the model parameters shall be measured.

## Measures

* Matching/localisation/retrieval performance as defined by the evaluation framework when using the compressed model.
* Matching and retrieval performance shall be reported for cases in which the query descriptors have been extracted using a network that has been compressed, and the reference (and distractor) descriptors have been extracted using an uncompressed network.
* Runtime and memory consumption of model compression and decompression shall be measured (in analogy to the complexity measurements defined in the evaluation framework). The runtime/memory measurements for all experiments in one response to the CE shall be performed on the same machine using CPU only. For reference, the runtime/memory measurements for CXM v3.0 determined on the same machine as defined by the evaluation framework shall be reported.

## CE Coordinators

Lingyu Duan (PKU, lingyu@pku.edu.cn), Yihang Lou (PKU, [yihanglou@pku.edu.cn](mailto:yihanglou@pku.edu.cn))

Mitsubishi Electric is responsible for performing crosschecks on this CE.

# References

[1] N15339, “Call for Proposals for Compact Descriptors for Video Analysis (CDVA) - Search and Retrieval”, June 2015, Warsaw (PL)

[2] N15729, Evaluation Framework for Compact Descriptors for Video Analysis - Search and Retrieval – Version 2.0, October 2015, Geneva, CH

[3] N15040, Compact Descriptors for Video Analysis: Requirements for Search Applications, Oct. 2015, Strasbourg, FR.

[4] N14507, Compact Descriptors for Video Analysis: Objectives, Applications and Use Cases, Apr. 2014, Valencia, ES.

[5] m37880, BRIDGET Response to the MPEG CfP for Compact Descriptors for Video Analysis (CDVA) - Search and Retrieval.

[6] m37794, JRS Response to Call for Proposals for Technologies Compact Descriptors for Video Analysis (CDVA) - Search and Retrieval.

[7] m37636, PKU’s Response to MPEG CfP for Compact Descriptor for Visual Analysis.

[9] N16275, Description of Core Experiments on CDVA, May 2016, Geneva, CH.

[10] m38519, JRS Response to CDVA Core Experiment 1, May 2016, Geneva, CH.

[11] m38625, PKU’s Response to CDVA CE1, May 2016, Geneva, CH.

[12] m38664, BRIDGET report on CDVA Core Experiment 1 (CE1), May 2016, Geneva, CH.

[13] m39433, Improved temporal localization for CDVA, Oct. 2016, Chengdu, CN.

[14] m39172, JRS Response to CDVA Core Experiment 1, Oct. 2016, Chengdu, CN.

[15] m39173, JRS Response to CDVA Core Experiment 3, Oct. 2016, Chengdu, CN.

[16] m39219, Improved retrieval and matching with CNN feature for CDVA, Oct. 2016, Chengdu, CN.

[17] m39220, Suggestions on deep learning based technologies for the future development of CDVA, Oct. 2016, Chengdu, CN.

[18] m39918, JRS Response to CDVA Core Experiment 1, Jan. 2017, Geneva, CH.

[19] m39919, JRS Response to CDVA Core Experiment 3, Jan. 2017, Geneva, CH.

[20] m39852 Response to CDVA CE4: Combination of NIP and CDVS Descriptors, Jan. 2017, Geneva, CH.

[21] m39853, PKU’s Response to CDVA CE4: NIP Network Compression, Jan. 2017, Geneva, CH.

[22] m40387, PKU’s Response to CDVA CE5: Binarization for NIP descriptor, Apr. 2017, Hobart, AU.