|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| N42 | 2020/10/21 |
| Source | Roberto Iacoviello, RAI R&D, Email: [roberto.iacoviello@rai.it](mailto:roberto.iacoviello@rai.it) |
| Title | Collaborative Evidence Conditions (CEC) |
| Target | Contribution to MPAI-EVC |
| Purpose | Proposal |

# Abstract

The aim of this document is to define a common terminology, test conditions, (a strategy for training), conversion practices and software reference configurations to be used in the context of a series of experiments designed to verify the improvement achievable by selected one traditional video codec andadding deep neural networks who performance has already been published. Added technologies need not satisfy the future MPAI-EVC Framework Licence as the purpose is simply to assess the level of improved performance to justify the opening on the MPAI-EVC project. Moreover, the document provides a reporting template and guidance for analyzing the performance of the deep networks tools.

These evidence conditions are recommended for use in technical contributions for Collaboratibe Evidence Assessment phase of MPAI-EVC.

# Common Terminology

|  |  |
| --- | --- |
| MSE Mean Squared Error | MSE Mean Squared Error |
| MAE Mean Absolute Error | MAE Mean Absolute Error |
| PSNR Peak signal-to-noise ratio | PSNR Peak signal-to-noise ratio |
| MS-SSIM Multiscale Structural Similarity | MS-SSIM Multiscale Structural Similarity |
| Activation Function | A function (for example, ReLU or generalized divisive normalization, GDN) that takes in the weighted sum of all of the inputs from the previous layer and then generates and passes an output value to the next layer. |
| Backpropagation | Algorithm for performing gradient descent on neural networks. First, the output values of each node are calculated in a forward pass. Then, the partial derivative of the error with respect to each parameter is calculated in a backward pass through the graph. |
| Batch | The set of examples used in one iteration (one gradient update) of model training. |
| Batch Normalization | Normalization of the input or output of the activation functions in a hidden layer. |
| Batch Size | The number of examples in a batch. |
| Artificial Intelligence | A machine program that can solve sophisticated tasks.  Machine learning is a sub-field of artificial intelligence. Deep Learning is a subset of machine learning. |
| Deep Neural Network | A type of neural network containing multiple hidden layers. |
| Shallow Neural Network | It is a term used to describe a neural network that usually have only one hidden layer. |
| Cross-validation | A mechanism for estimating how well a model will generalize to new data by testing the model against one or more non-overlapping data subsets withheld from the training set. |
| Dataset | A collection of sequences |
| Epoch | A full training pass over the entire dataset such that each example has been seen once. |
| Fine tuning | Perform a secondary optimization to adjust the parameters of an already trained model to fit a new problem. |
| Generalization | Refers to your model's ability to make correct predictions on new, previously unseen data as opposed to the data used to train the model. |
| hyperparameter | The "knobs" that you tweak during successive runs of training a model. |
| inference | Refers to the process of making predictions by applying the trained model to unlabeled examples. In this document it is used as a synonym for testing to maintain consistency with video encoding. |
| learning rate | A scalar used to train a model via gradient descent. During each iteration, the gradient descent algorithm multiplies the learning rate by the gradient. The resulting product is called the gradient step. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Introduction

The general strategy to evaluate the improvement achieved by adding deep neural networks to a selected traditional video codec is depicted in Figure 1: The selected traditional video coding scheme is the Essential Video Coding (EVC) (baseline or main). The Video group will agree on Test and Training sequences. It is expected that all the members in the group will also agree on training and test conditions. In addition, the group will choose a common reporting method and the list of deep tools to be added. After this ‘preprocessing stage’ the proponents will integrate the deep tools into EVC and will calculate the gain (in terms of BD-RATE) compared to EVC (baselilne or main profile) and a cross-checker will verify the results. Finally, the percentages will be added up to obtain the final evidence.

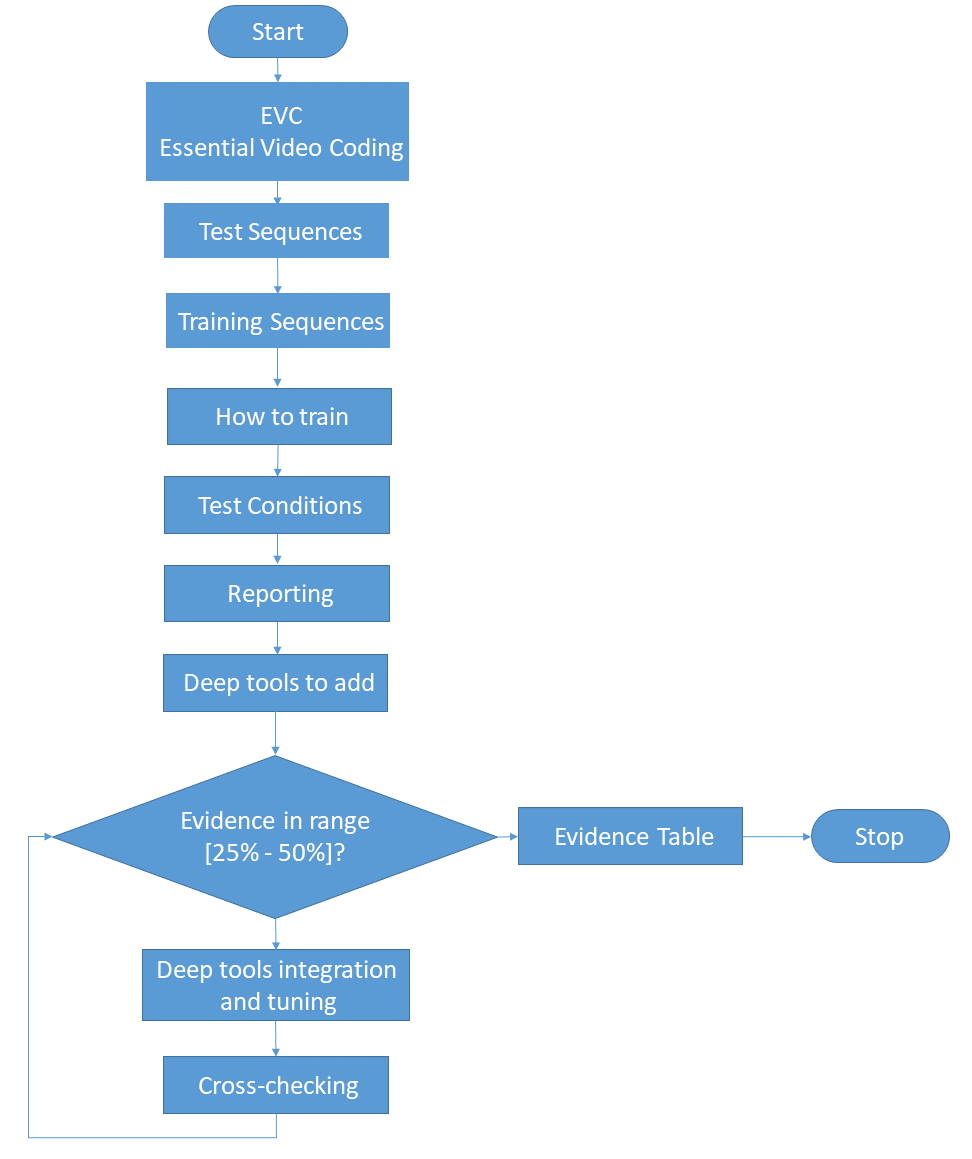


Figure 1: Working method flow chart

In this document will be evaluated two hybrid approaches [1]:

* Horizontal Hybrid approach which introduces AI based algorithms combined with traditional image video codec, trying to replace one block of the traditional schema with one deep learning based (Figure 2)
* Vertical Hybrid: an EVC base layer plus an enhanced layer based on deep learning (Figure 3)

The Figure 2 and Figure 3 represent the starting reference scheme adopted in the experiments.

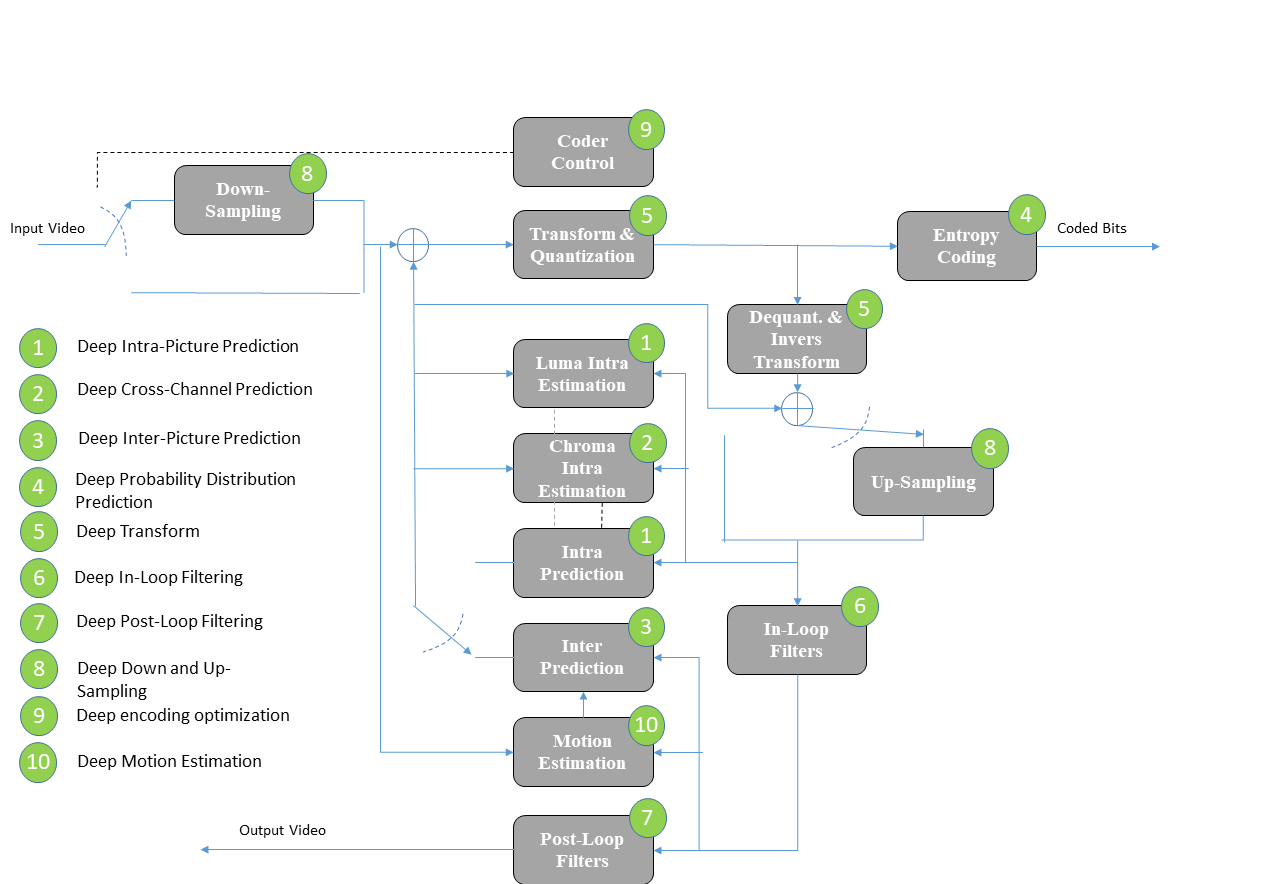


Figure 2: Essential Video Coding scheme (Horizontal Hybrid approach)

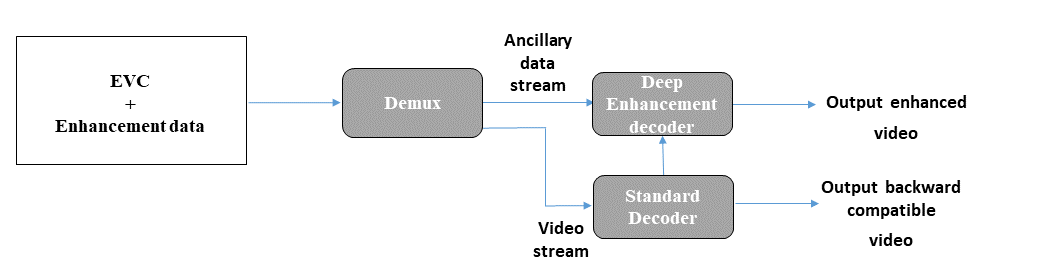


Figure 3: A reference diagram for the Vertical Hybrid approach

Additional components of EVC baseline may be added when appropriate conditions are verified.

The rest of this document is organized as follows. Section 3 describes the Test Sequences and Section 4 the Training Sequences. Section 5 the methodology for the training stage of neural network-based coding tools. Section 6 describes the methodology for the test (inference) stage. Section 7 highlights the reporting stage. In Section 8 the list of the deep tools to be added in the EVC codec. Section 9 describes the Evidence Assessment (EA). Section 9 describes the Anchor generation.

# Test Sequences

The test sequences are encouraged to belong to the following genres:

* sports (football, tennis, e.g. individual player teams…)
* music shows
* TV program
* action video games
* contents from archive
* computer generated graphics
* video conferencing
* 360 video

In Table 1 the list of test sequences to be used. All frames (as defined by frame count in the table) shall be encoded for all sequences.

The column corresponding to each configuration is interpreted as follows:

* “M” indicates that the test sequence is mandatory in the CEC for the given configuration
* “O” indicates that the test sequence is optional (but encouraged) in the CEC for the given configuration
* “-” indicates that the test sequence is not requested in the CEC or the given configuration

The test sequences are divided in Classes:

* Class A: 4K sequences
* Class B: HD sequences
* Class C: HDR (High Dynamic Range) sequences

Original versions of the test sequences in Table 1 are available as follows:

* Test sequences of classes X are available on ftp://seq@ftp.rai/testsequences/

Test sequences are only available to qualified participants.

Table 1. Test sequences

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Sequence name | Frame count | Frame rate | Original Bit depth | Random access |
| B | Down | 200 | 60 | 10 | M |
| B | Strange | 300 | 60 | 10 | M |
| B | Charm | 400 | 50 | 8 | M |
| C | Bottom | 500 | 60 | 8 | M |
| C | Top | 600 | 50 | 8 | M |

# Training Sequences (TBD)

Table 2 describes the set of training sequences to be used in the training process by a proposal. A list of training sets available on internet is provided in Annex A.

It is optional (but encouraged) to provide the time to train the network.

Table 2. Dataset for Training

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data base | Link | Description | Compressed? | Comment | License Terms and/or Link to Additional Information |
| RAI 4K | www.xxx.it | xxx video clips of 10 sec; 4K | uncompressed | big resolution | academic research only, no commercial use |
|  |  |  |  |  |  |

# How to train

All proponents are supposed to use the same training set described in section 6.1

## Methodology of training

It is optional to use:

* pretrained models
* retrain the network
* fine tuning
* continuous learning

# Test Conditions

Collaborative Evidence Conditions (CECs) are designed to conduct experiments in a well-defined environment and to facilitate comparison of experiment results.

This document defines only one test condition:

* Random access (10 bit): with a GOP of XX (group of pictures)

The following sections define, bitdepth, Quantization Parameter values (QP), anchors, evaluation metrics and configuration files.

Anyone bringing input contributions to MPAI Video meetings should provide a set of results that is as complete as possible and uses the test conditions that apply to the proposal.

## Bitdepth

The input and output bit-depth of the codec shall be 10-bit regardless of bit-depth of the input sequence. For the 8-bit sequences in Table 1, each 8-bit source sample x should be converted prior to encoding to a 10-bit value.

## QP Value

For the anchor and proposals with a quantization concept substantially similar to the anchor, results shall be provided using the 4 quantization parameter values: 27, 32, 37, 42. These values define the initial QP values that are specified as input QP of the anchor EVC software.

# Reporting

## Metrics

A common Excel sheet shall be used that contains a reporting template in which bitrate, PSNR, encoding and decoding time, and BD-rate results are reported for the tested configuration against the anchor.

As the work progresses metrics that better capture the influence of deep neural network based on quality of compressed video may be selected/defined.

## PSNR

PSNR shall be calculated as

where bitDepth is the bit-depth of the input video.

## Complexity

For the purpose of reporting it is encouraged to provide the time to train the network,

encoding and decoding running times; The anchor and proposal should be simulated on the same platform, e.g. similar CPU, GPU, FPGA configuration, to have reliable time comparison.

## Additional info

It is appreciated and high encouraged to provide additional information that includes (but is not limited to):

* Attributes of the testing environment, including CPU type, GPU type, GPU memory size, compiler, run time, peak memory usage and decoder configuration. Fields for providing this information are provided in the Excel template.
* Description of the process used to train the network, including the loss function, optimizer, and key hyperparameters. Example hyperparameters include the batch size, number of epochs, number of iterations, and the learning rate.
* Information about the network, including the size of the network, its architecture, parameter precision, and any changes in network configuration or weights required to generate the requested rate points.
* The framework used for implementing the technology (e.g., PyTorch, TensorFlow, etc.)

# Deep tools to add

In the following the list of deep tools to be added in the EVC schema:

* Deep Intra-Picture Prediction
* Deep Cross-Channel Prediction
* Deep Inter-Picture Prediction
* Deep Probability Distribution Prediction
* Deep Transform
* Deep In-Loop Filtering
* Deep Post-Loop Filtering
* Deep Down and Up-Sampling
* Deep encoding optimization
* Deep Motion Estimation

# Evidence Assessment (EA)

At this stage a deep tool is integrated in the EVC schema. The code is then cross-checked by one or more people inside MPAI Video group. The cross-checker verifies if the code is executed on the agreed video set in the agreed test conditions and calculates the PSNR metric.

After the ‘green light’ of the cross-checker, the gain percentage of that deep tool is entered in the Evidence Table (Table 3)

**Table 3** Evidence Table

|  |  |  |
| --- | --- | --- |
| **Deep-tool** | **Test condition** | **BD- Rate** |
| Intra | EVC baseline |  |
| Inter | EVC baseline |  |
| Cross-channel | EVC baseline |  |
| In-loop | EVC baseline |  |
| Down up-sampling | EVC baseline |  |
| Coder control- rate control | EVC baseline |  |
| Probability Distribution | EVC baseline |  |
| Post-loop filter | EVC baseline |  |
| **Total** | | xx% |

Finally, by adding all the percentages in Table 3, the MPA Video group will obtain the Evidence Assessment (EA) of the coding gain based on deep tools injected into the EVC.

If (EA in range [25% - 50%]) then

A new standard project is born

else

Houston, MPAI GA decides what to do

# Anchor

## Software

Version xx of the EVC software is expected to be used for the anchor.

## Configuration

Definition of the configuration files to be used for the anchor.

# Patent rights declaration(s)

**According to the Framework Licence.**

# References

[1] Roberto Iacoviello; Analysis of performance of AI based video codecs

# Annex A Dataset for Training

Table derived from Draft Test Conditions for DNNVC (m19583)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | Data base | Web | Description | Compressed? | Comment | Reported but Unconfirmed License Terms and/or Link to Additional Information |
| 1 | Subset of Vimeo90K | http://toflow.csail.mit.edu/ | 73,171 3-frame sequences with a fixed resolution of 448 x 256 | uncompressed | Small resolution but enough for training | Unknown |
| 2 | UGC | <https://media.withyoutube.com/> | ~ 1500 video clips of 20 sec; 360P, 480P, 720P, 1080P; 4K | uncompressed | Was used in CLIC-P | [CC BY](https://creativecommons.org/licenses/by/3.0/legalcode) |
| 3 | [SJTU](http://medialab.sjtu.edu.cn/index.html) | <http://medialab.sjtu.edu.cn/web4k/index.html> | 15 of 4k video | uncompressed | Campfire Party of JVET is from this set | Dataset is available for free and academic research only, no commercial use. Please cite our paper (L. Song, X. Tang, W. Zhang, X. Yang, P. Xia, [The SJTU 4K Video Sequence Dataset](http://medialab.sjtu.edu.cn/publications/2013/2013_QOMEX_SL.pdf), the Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013), Klagenfurt, Austria, July 3rd-5th, 2013) in any published work if you use those video sequences. |
| 4 | UVG Dataset | <http://ultravideo.cs.tut.fi/> | 4K 120fps test sequences | uncompressed | Appear  seen many times in different papers | [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/deed.en_US) |
| 5 | CDVL | <https://www.cdvl.org/> |  |  |  | [The ITU, VQEG, ATIS and other standards developing committees (SDO) can use CDVL R&D content for an internal project (e.g., a subjective test to validate objective video quality models).](https://www.cdvl.org/license/) |
| 6 | [CableLabs](http://www.cablelabs.com/resources/4k/) | <https://www.cablelabs.com/4k> | 11 4K videos |  |  | [Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License.](http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en_US) |
| 7 | UHD-1 TEST SEQUENCES | <https://tech.ebu.ch/testsequences/uhd-1> |  |  | Some EBU sequences were used for HDR tests in JCTVC/ JVET. | Unknown |
| 8 | EBU and SVT Public Test Sequences | <https://tech.ebu.ch/docs/hdtv/svt-multiformat-conditions-v10.pdf> | 5-13 UHD videos |  | Six minute program may still be available | [BY-NC-ND](http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode) |
| 9 | JVET FTP | [ftp.ient.rwth-aachen.de](ftp://ftp.ient.rwth-aachen.de)  [ftp.tnt.uni-hannover.de/testsequences/](ftp://ftp.tnt.uni-hannover.de/testsequences/) | Need only YUV420 or also 444? |  | Whatever was made available for JVET (including new HDR content from NHK). | License terms provided with content |
| 10 | [HMDB51](http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/) | <http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/> | HMDB: A Large Video Database for Human Motion Recognition |  | Action recognition  6766 videos  51 action classes | <http://serre-lab.clps.brown.edu/wp-content/uploads/2012/08/Kuehne_etal_iccv11.pdf> |
| 11 | UCF101 | <http://crcv.ucf.edu/data/UCF101.php> | UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild |  | Sports  13320 videos  101 action classes | <http://crcv.ucf.edu/papers/UCF101_CRCV-TR-12-01.pdf> |
| 12 | Sports-1M | <http://cs.stanford.edu/people/karpathy/deepvideo/> | The YouTube Sports-1M Dataset |  | Sports  1100000 videos  487 sports classes | <http://cs.stanford.edu/people/karpathy/deepvideo/deepvideo_cvpr2014.pdf> |
| 13 | Charades | <http://allenai.org/plato/charades/> | This dataset guides our research into unstructured video activity recognition and commonsense reasoning for daily human activities |  | Human activities  9848 videos  157 action labels, 27847 Free-text descriptions, action intervals, classes of interacted objects | <https://arxiv.org/pdf/1604.01753.pdf> |
| 14 | ActivityNet | <http://activity-net.org/> | A Large-Scale Video Benchmark for Human Activity Understanding |  | Human activities  28000 videos  203 classes | <http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.pdf> |
| 15 | Kinetics | <https://deepmind.com/research/open-source/open-source-datasets/kinetics/> | Kinetics is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human focused actions |  | Action Recognition  500000 videos  600 action classes | <https://arxiv.org/abs/1609.08675> |
| 16 | Youtube-8M | <https://research.google.com/youtube8m/download.html> | YouTube-8M is a large-scale labeled video dataset that consists of millions of YouTube video IDs and associated labels from a diverse vocabulary of 4700+ visual entities |  | Youtube Random Videos  8000000 videos  4716 classes | <https://arxiv.org/abs/1609.08675> |
| 17 | AVA | <https://research.google.com/ava/> | A Video Dataset of Spatio-temporally Localized Atomic Visual Actions |  | Atomic visual actions  57600 videos;  210k action labels, 80 atomic visual actions, spatio-temporal annotations | <https://arxiv.org/abs/1705.08421> |
| 18 | 20BN-SOMETHING-SOMETHING | <https://www.twentybn.com/datasets/something-something> | The 20BN-SOMETHING-SOMETHING dataset is a large collection of densly-labeled video clips that show humans performing predefined basic actions with every day objects |  | Human activities  108000 videos  174 classes | <https://arxiv.org/abs/1706.04261> |
| 19 | 20BN-JESTER | <https://www.twentybn.com/datasets/jester> | Human Hand Gestures Dataset |  | Hand Gestures  148000 videos  27 classes | <https://www.di.ens.fr/> |
| 20 | LSMDC | <http://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/vision-and-language/mpii-movie-description-dataset/> | Large-Scale Movie Understanding Dataset |  | Movie clips  118000 videos  Aligned captions | <https://arxiv.org/pdf/1605.03705.pdf> |
| 21 | DALY | <http://thoth.inrialpes.fr/daly/> | Daily Action Localization in Youtube videos |  | Spatio-temporal Action localization  8100 videos  3.6k spatio-temporal action annotation | <https://arxiv.org/pdf/1605.05197.pdf> |
| 22 | MPII-Cooking | <https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-2-dataset/> | MPII Cooking dataset |  | Cooking videos  273 videos  78 classes, 13k labelled instances | <https://www.mpi-inf.mpg.de/fileadmin/inf/d2/amin/rohrbach12cvpr.pdf> |
| 23 | VideoMCC | <http://videomcc.org/> | a New Benchmark for Video Comprehension |  | Video News Understanding  272000 videos  10 topics and Video captions | <https://arxiv.org/abs/1606.07373> |
| 24 | VGG Human Pose | <https://www.robots.ox.ac.uk/~vgg/data/pose/index.html> | The VGG Human Pose Estimation datasets is a set of large video datasets annotated with human upper-body pose |  | Human Pose Estimation  152 videos  Hours of human upper-body pose | <https://arxiv.org/abs/1511.06676> |
| 25 | YFCC100M | <http://yfcc100m.appspot.com/>? | YFCC100M: The New Data in Multimedia Research |  | Webly annotated Flickr videos  800000 videos  1570 tags, captions and diverse metadata | <https://arxiv.org/pdf/1503.01817.pdf> |
| 26 | DiDeMo dataset | <https://people.eecs.berkeley.edu/~lisa_anne/didemo.html> | the Distinct Describable Moments (DiDeMo) dataset consists of over 10,000 unedited, personal videos in diverse visual settings with pairs of localized video segments and referring expressions. |  | Captioning  10000 videos  40000 aligned captions | <http://arxiv.org/abs/1506.09215> |
| 27 | HACS | <http://hacs.csail.mit.edu/> | Human Action Clips and Segments Dataset for Recognition and Temporal Localization |  | Action recognition  520000 videos  200 action classes, 1.75M clip annotations | <https://arxiv.org/abs/1712.09374> |
| 28 | VLOG | <https://people.eecs.berkeley.edu/~dfouhey/2017/VLOG/index.html> | From Lifestyle VLOGs to Everyday Interactions: The VLOG Dataset |  | Action recognition  114000 videos | <https://arxiv.org/abs/1712.02310> |
| 29 | Moments in Time | <http://moments.csail.mit.edu/> | Moments in Time Dataset: one million videos for event understanding |  | Action recognition  1000000 videos  339 action classes | <https://arxiv.org/abs/1801.03150> |
| 30 | YouCook2 | <http://youcook2.eecs.umich.edu/> | YouCook2 is the largest task-oriented, instructional video dataset in the vision community |  | Cooking videos  2000 videos  15400 aligned captions | <https://arxiv.org/pdf/1703.09788.pdf> |
| 31 | EPIC-KITCHENS | <https://epic-kitchens.github.io/2018> | The largest dataset in first-person (egocentric) vision; multi-faceted non-scripted recordings in native environments - i.e. the wearers' homes, capturing all daily activities in the kitchen over multiple days. |  | Cooking videos  432 videos  39596 action segments, 323 object classes, 454,158 bounding boxes | <https://arxiv.org/pdf/1804.02748.pdf> |
| 32 | Oops! | <https://oops.cs.columbia.edu/> | Predicting Unintentional Action in Video |  | Classification  20723 videos  Intentionality label of action and localization of transition from unintentional to intentional | <https://arxiv.org/abs/1911.11206> |
| 32 | BVI-DVC | <https://vilab.blogs.bristol.ac.uk/?p=2375> | A Training Database for Deep Video Compression |  | 800 sequences at various spatial resolutions from 270p to 2160p | arXiv:2003.13552, 2020 |