Video Coding for Machines

Yuan ZHANG Rapporteur, ITU-T Q12/16
Chairman, MPEG AHG VCM
China Telecom

8 October, 2019

What is Video Coding for Machines (VCM)?

VCM – What to do

What human vision wanted

- High fidelity, large image size, high frame rate.

What machine vision wanted

- High accuracy, low latency, Object oriented, High level abstraction.

Combine Human/Machine Vision

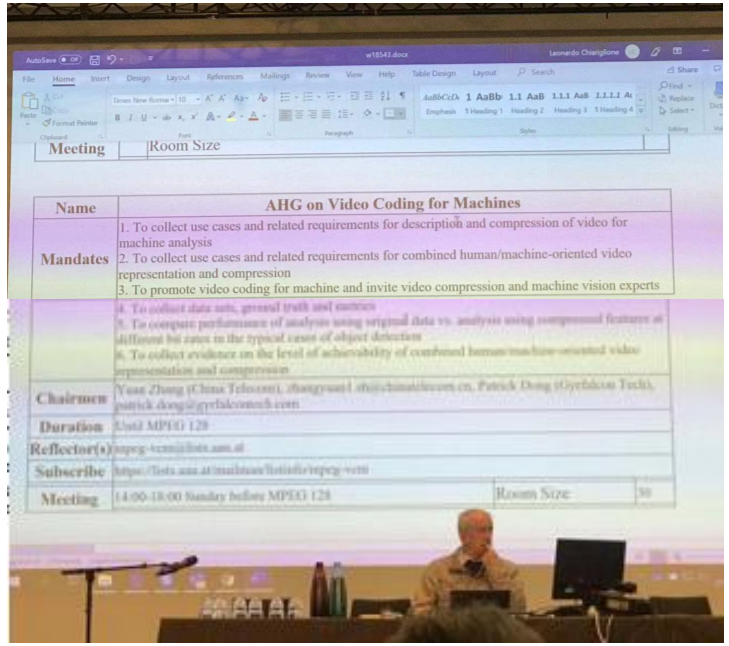
- Attention Mechanism
- Fidelity/Accuracy

How to evaluate the performance, metrics?

- PSNR?
- mAP, precision and recall
- Subjective and Objective evaluation

MPEG AHG on VCM

Name	AHG on Video Coding for Machines		
Mandates	1. To collect use cases and related requirements for description and compression of video for machine analysis 2. To collect use cases and related requirements for combined human/machine-oriented video representation and compression 3. To promote video coding for machine and invite video compression and machine vision experts 4. To collect data sets, ground truthand metrics 5. To compare performance of analysis using original data vs. analysis using compressed features at different bit rates in the typical cases of object detection 6. To collect evidence on the level of achievability of combined human/machine-oriented video representation and compression		
Chairmen	Yuan Zhang (China Telecom), zhangyuan1.sh@chinatelecom.cn Patrick Dong (Gyrfalcon Tech), patrick.dong@gyrfalcontech.com		
Duration	Until MPEG 128		
Reflector(s)	mpeg-vcm@lists.aau.at		
Subscribe	https://lists.aau.at/mailman/listinfo/mpeg-vcm		
Meeting	14:00-18:00 Sunday before MPEG	Room Size	30



m49181, “Requirements of video analysis and semantic compression” July 2019, Gothenburg, Sweden

m48429, “Fine grained feature” July 2019, Gothenburg, Sweden

Draft Work Plan of AHG VCM

AHG VCM will study the next generation of video compression standards - **machine vision oriented video compression and hybrid human and machine vision oriented video compression**. w18662 Requirements *Video Coding for Machines: Use Cases* is output at #127 MPEG.

Machine-oriented Analysis Use Cases

Smart Glasses, Unmanned store, Unmanned Warehouse/Store Robot, Smart Retailer, Industrial Production Line Detect Equipment, Smart Factory /Automatic Machinery, Smart fishery/ Smart agriculture, UAV, Automation ADAS, Autonomous Vehicles

Combined Machine and Human representation Use Cases

AR and Video Game Goggles, Sports Game animation, Surveillance

Preliminary Timeline

2019.07 Establishment of VCM, mailing list, draft use cases, Kickoff
2020.01 Draft EE, collect evidences, draft requirements
2020.07 Call for evidence, call for data set, call for application
2021.07 Cfp

Use Cases - Machine Vision

Case 1

- Smart factory & unmanned environment monitoring Industrial inspection in large-scale production
  Warehouse stocking / checking
  Industrial automated robot assistant

Case 2

- Autonomous Vehicles
  Advanced Driver Assistance System (ADAS) Connected cars (V2X)
  Drones

Case 3

- Retail Analysis
  Shopping center customer / group analysis
  Autonomous checkout
  Visual Machine Product Search
  Video surveillance security to protect staff and customers and assist loss prevention Store robot assistant
  People counting / customer traffic information

Use Cases - Hybrid Human&Machine Vision

Case 4

- Video surveillance
  Face recognition, person, vehicle re-identification Public safety / Law reinforcement
  Smart traffic monitoring
  Smart parking

Case 5

- Entranced media / entertainment
  VA / AR
  Collaborative gaming
  Ultra-high definition (UHD) content categorization/tagging Smart TV

Case 6

- Large-scale sensors
  Smart grid – utilities
  5G: Visual IoT (VIoT)
  Bayer pattern image / frame

Typical use case: Autonomous Driving

Machine vision requires different information from human vision. Autonomous Driving is a typical use case of VCM technology.

VCM (Video coding for Machines) is an enabling and fundamental technology for Autonomous Driving.

The performance of recognition base on machine vision should meets, or exceeds human recognition.

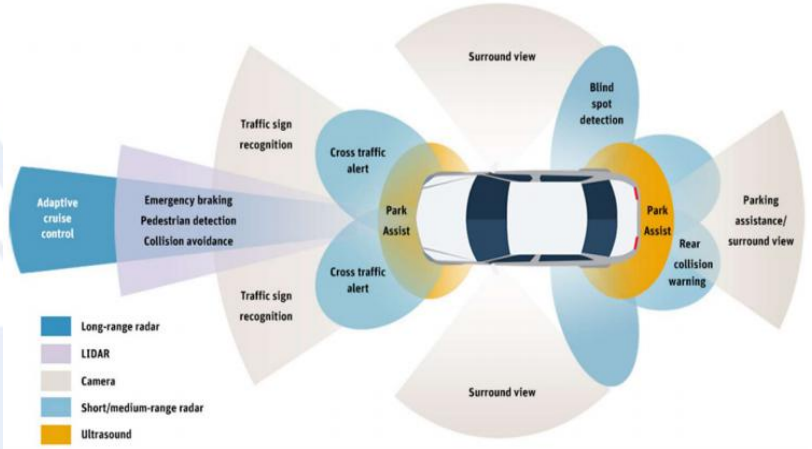
Besides error rate and accuracy, Turing Test is introduced as a measurement of autonomous driving. Generative adversarial networks are trained to confront the Turing Test.

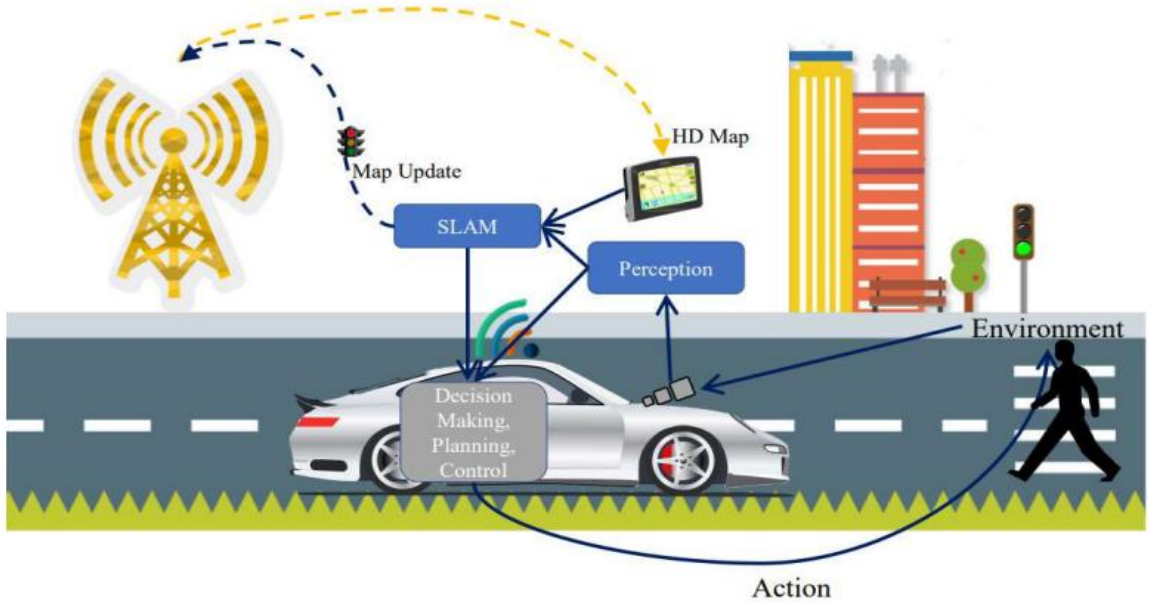
Perception, the first step of VCM, is to detect object, background, lane, vehicle, traffic sign, pedestrian, etc.

Key tasks include: Multi-object detection, Object segmentation, Object (Lane) Tracking, Activity recognition, Event prediction, Optical Flow.

Other enabling technologies include: 5G V2X, decision making (reinforcement learning), edge computing, deep learning, etc.

Sensors include camera, infrared ray, LiDAR, microwave radar, ultrasonic wave.





Key Tasks

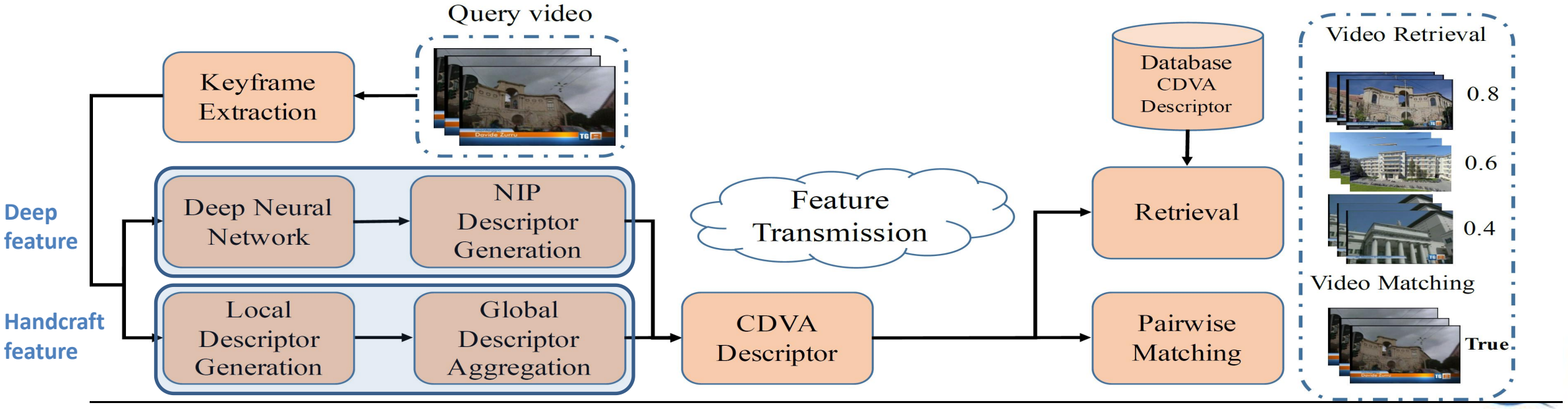
Key tasks	Sub-tasks
Segmentation	Object segmentation
Detection	Object detection, Multi-object detection, Emotion detection, Key point detection
Recognition	Gesture/movement recognition, activity recognition, face recognition, gait recognition
Classification	Object classification
Identification	Situation identification
Route planning (navigation)	
Traffic analysis	
Heat map generation	
Prediction	Event prediction, Pose estimation
Tracking	Object (Lane) tracking, Customer activity tracking
Optical flow generation	
Video Reconstruction	

Key tasks are identified based on the VCM uses cases, which could be categorized into segmentation, detection, recognition, classification, identification, navigation, heat map, tracking, prediction, optical flow, etc. Given that different tasks require different grained of features, resulting in different sizes of object video, detailed categorization of each specific task is to be defined.

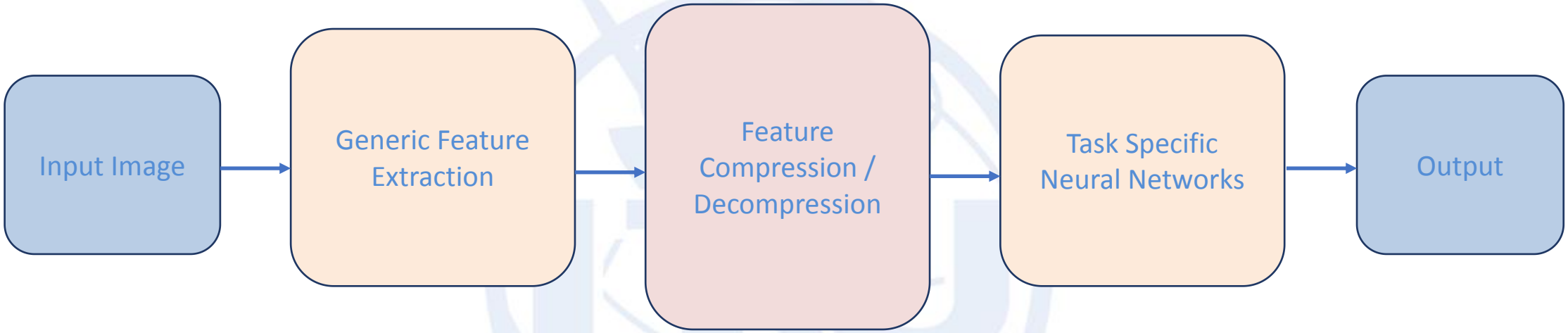
Exsiting Work - CDVS/CDVA

- **Backgroud:** compact feature descriptors instead of compressed video texture
- **Applications:** Media & Entertainment, Surveillance, Mobile video, Automotive, manufacturing, robotics, …
- **Target:** compact descriptors for video search and retrieval
  - enable design of interoperable object instance search applications;
  - minimize the size of video descriptors;
  - ensure high matching performance of objects (in terms of accuracy and complexity);
  - enable efficient implementation of those functionalities on professional or embedded systems.

CDVS, ISO/IEC 15938-13: 2015
CDVA, ISO/IEC 15938-15: 2019



Proposed VCM Framework



- The first stage extracts fine-grained features from the input image, which can be shared for multiple tasks.
- A task-specific network can be specifically trained for image object detection, segmentation, classification, etc.

Questions to be addressed

►1 The necessity of VCM standardization.

- a) Which applications require sharing features rather than video between different tasks? Is it more efficient to share feature than video, i.e., data rate efficiency?
- b) Which applications require standard features for specific analysis task among different vendors? Is it more efficient in data rate by sharing feature instead of coded video at the cost of some computational load in front-end devices?
- c) What is the most efficient way to define features to balance between general feature extraction and task specific analysis?

►2 What is the object to be standardized, i.e., the method of feature extraction, feature compression?

VCM EE framework

Accuracy and compression ratio differ as per different tasks, which enables different applications in intelligent analysis and understanding.

A common platform is required to test the technical metrics and corresponding models. Considering that feature compression may be lossy, the quality should be evaluated based on the accuracy and reliability of analysis and understanding tasks.

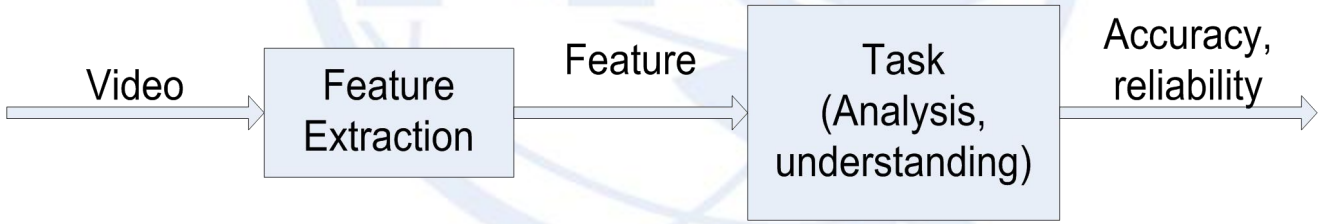
Results differ while using different models for a specific task, therefore, a common model for each task should be used for testing. Regarding the selection of models, using open source models is an option, while proponents are encouraged to contribute their own models.

Comparsion Framework

Exploration experiments should be done to compare the performance of:

1) Compressed video vs. compressed feature, and

2) Uncompressed feature vs. compressed feature,



based on the accuracy and reliability of analysis and understanding tasks.

Multiple bit rates and associated accuracy/reliability can be used to measure the performance.

Ongoing Works

- Smart tiling
- Smart sensing
- Shared backbone
- SuperCDVA
- CDVA with fine grained features
- Feature map compression for object detection and segmentation
- Joint image content understanding and compression
- Use cases refinement

Ongoing Work - Smart Tiling

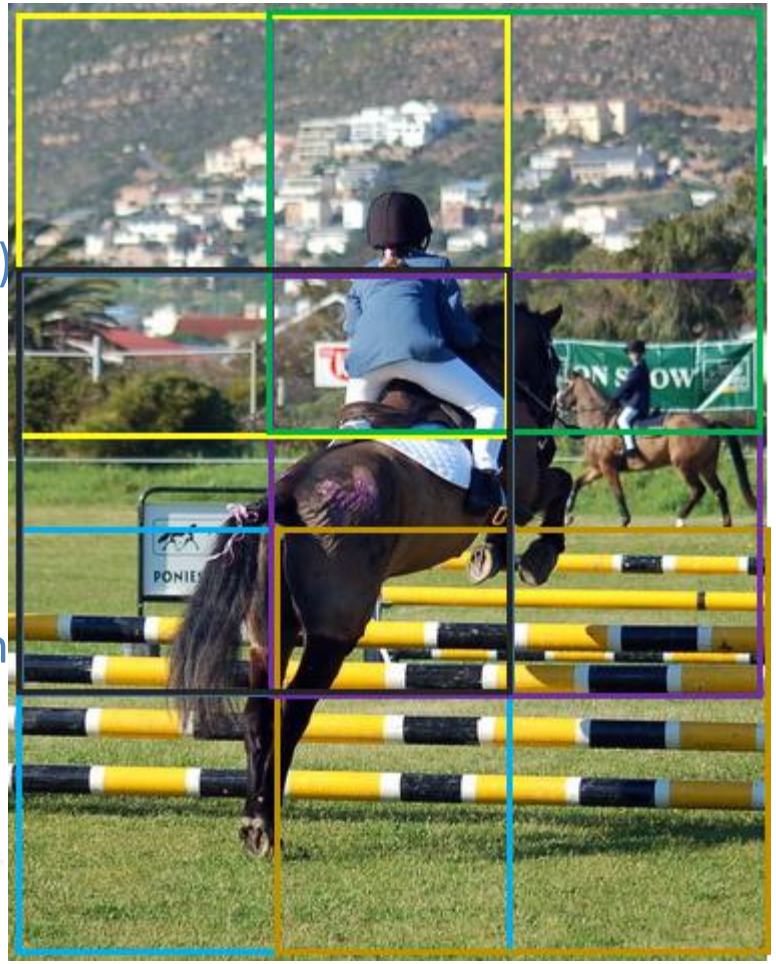
Tiled CDVA is a scalable extension of the MPEG CDVA standard to take advantage of higher resolution inputs without sacrificing latency. The method can be generalized to other tasks as well.

The CDVA standard relies on a VGG16 network based on VGA (640x480) input.

Tiled CDVA is a method introduced to handle input images/videos of any size with devices/networks that accept fixed, predetermined input sizes.

This method can be applied to CDVA, as well as other applications, such as object detection and segmentation, without compromising the accuracy incurred by device constraints.

With sufficient quantities of ASIC chips, each tile can be processed in parallel, regardless of the input size. This method scales to arbitrarily large images without needing arbitrarily large ASIC chips or memory buffers.



m50976, “Smart TiledCDVA” October 2019, Geneva, CH

Ongoing Work - Smart Sensing

Smart Sensing is an extension of the MPEG CDVA standard to directly process raw Bayer pattern data, without the need of a traditional ISP.

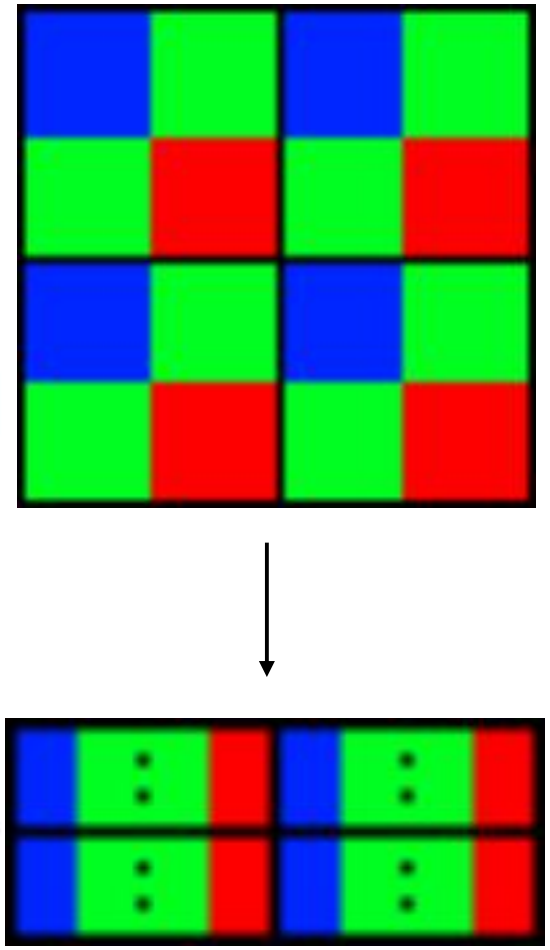
Traditionally, an ISP is needed to obtain RGB. This ISP is often accomplished via a dedicated ASIC chip, which is expensive, power hungry, and adds latency.

For deep learning applications like the CDVA standard, it is sufficient to adjust a pretrained CNN in a systematic manner, after which, the raw camera outputs can be used directly.

Typically, each 2x2 block of pixels consists of 1 red, 1 blue, and 2 green pixels.

Each of these 2x2 blocks of pixels can be arranged into a single pixel with 4 colors. As illustrated in the figure to the right, each of these 2x2 blocks is outlined in black; after the transformation, each of these blocks will become a single pixel with 4 colors.

The first layer of a CNN can be modified to average the 2 green values in addition to its original task (no need for retraining).



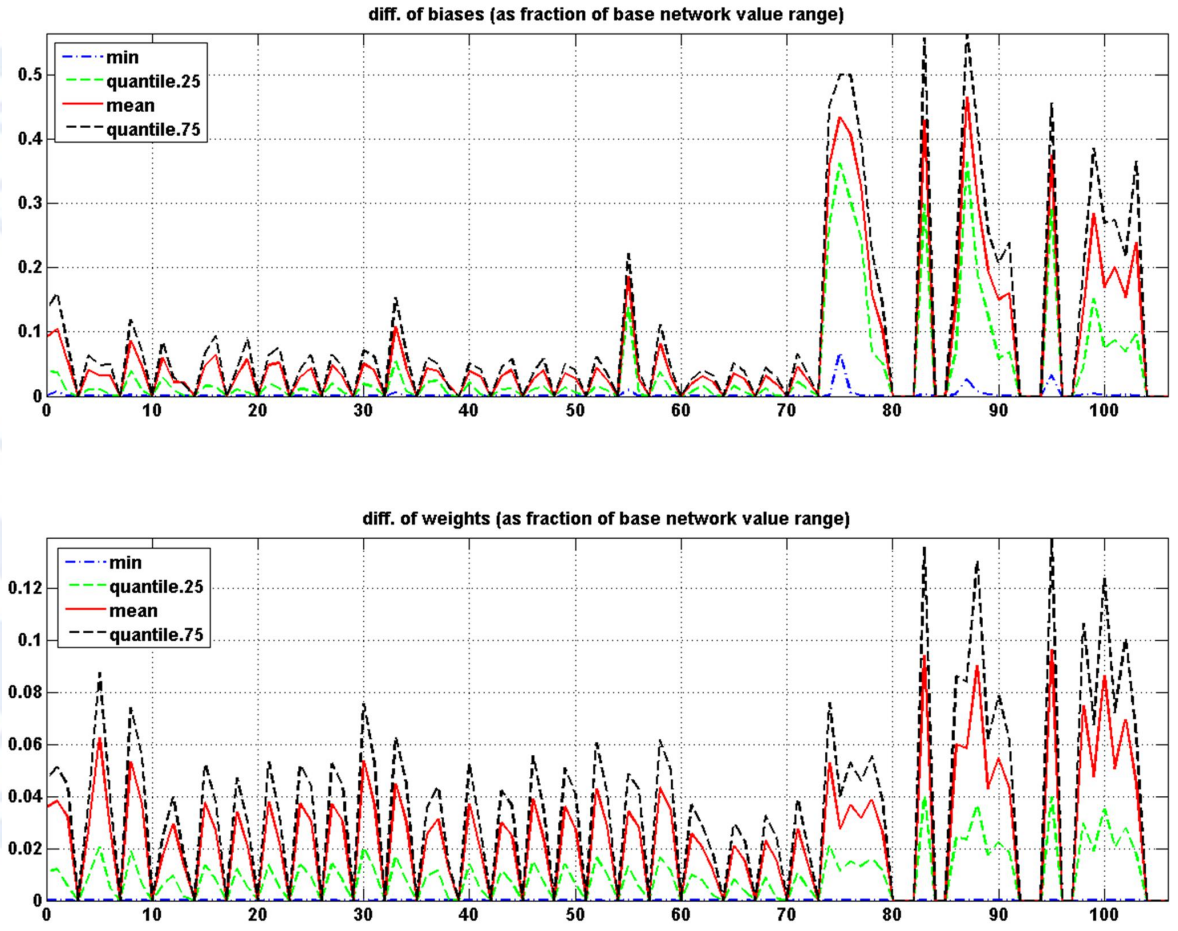
m50966, “Smart Sensing” October 2019, Geneva, CH

Ongoing Works - Shared Backbone

Sharing backbones of CNNs and working on a compressed representation of their outputs rather than on the input images could enable more resource efficient machine learning applications.

A common backbone network between neural networks is trained for different tasks, based on Yolo v3. Using different lengths of the common backbone, and retraining for the rest for the specific tasks, results are provided to show the possible performance and resource gain trade-offs.

We have analysed the resource savings that can be achieved by sharing backbone network between different object detection tasks. While significant resource savings can be achieved, the performance is reduced and cannot be easily moved to the original level without resorting to data of earlier layers.



Relative differences of biases and weights for the logo model

m49952, “Shared Backbone CNNs for Object Detection”, October 2019, Geneva, CH

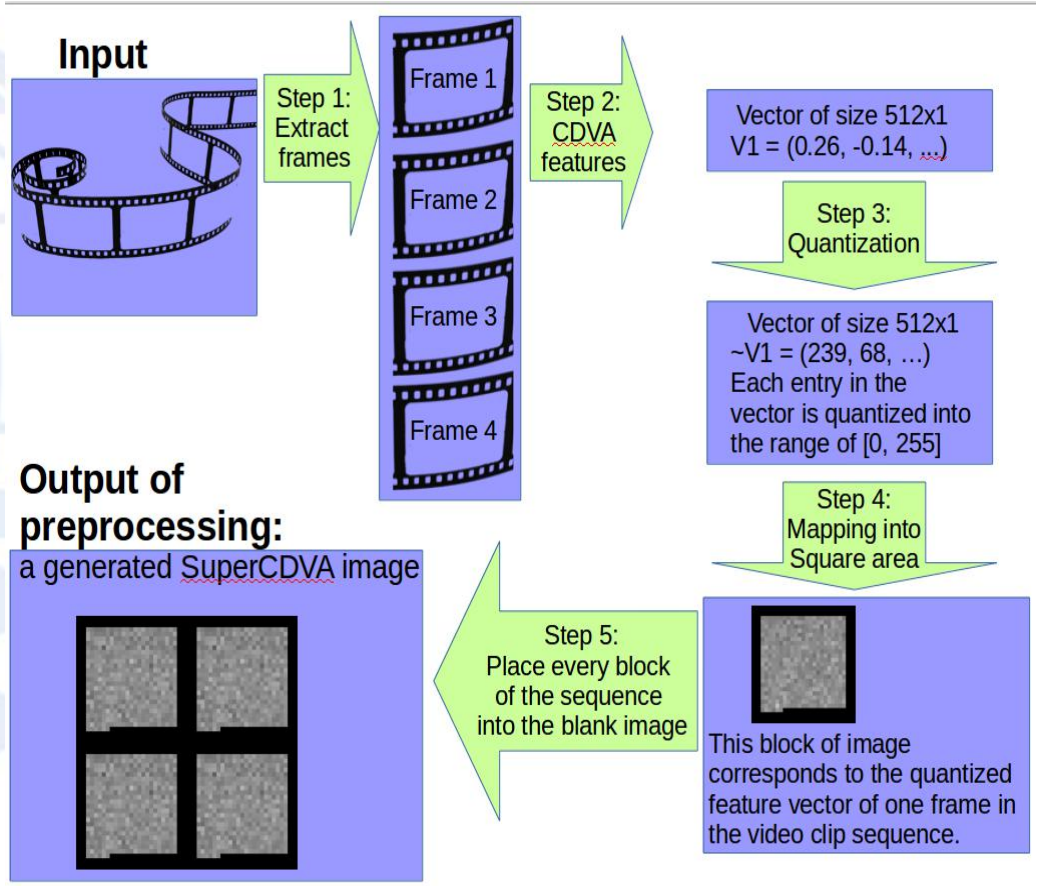
Ongoing Work - SuperCDVA

SuperCDVA is built upon the standard CDVA vectors, and in addition it includes the temporal information from the time dimension. Thus it has the advantage to understand a sequence of frames over the standard CDVA vectors which only understand the single static image. Super-CDVA is a two-step method:

Step 1: Extract CDVA vectors from each frame of the video clip. This step directly uses the exact MPEG7 CDVA standard.

Step 2: Embed the sequence of CDVA vectors into an image, and then this image is fed into a CNN model to classify, in order to understand the video.

For example, different swimming postures, such as breaststroke and front crawl can only be distinguished by watching a video clip, while a single static image is hard to tell apart these two swimming postures.



m50974, "SuperCDVA" October 2019, Geneva, CH

Potential Collaboration with ITU-T

Cases for collaboration between two SDOs:

- ❑ They have common vision and problems and overlapping/complementary expertise and experts resource
- ❑ They work together to develop a system of standards at different levels (lower level like PHY and higher level like protocols)
- ❑ One party A can contribute requirements and the other party B technical standards. B's yields will be integrated into A 's system

Potential collaboration in different optional forms (levels of collaboration strength):

1. Establish a joint working group with Q6/16
2. Develop standards jointly via joint meetings
3. Output Q12/16 and Q5/16 Recommendation to VCM as requirements, adopt and integrate VCM standards in Q12/16 and Q5/16 with relatively independent standards development process on both sides
4. Liaison-based approach like what we have in ITU known as JCA

Scan to subscribe
VCM mailing list