ISO /IEC JTC 1/SC 29 /WG 1 **N88038**
88th Meeting – Online – July 2020

ISO/IEC JTC1/SC29/WG1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE: Use cases and requirements for ISO/IEC 21122 (JPEG XS) v2.0

SOURCE: WG1 (Editor: Antonin Descampe, a.descampe@intopix.com)

PROJECT: ISO/IEC 21122 (JPEG XS)

STATUS: Approved

REQUESTED ACTION: For review

DISTRIBUTION: Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

1 Abstract

Image sequences have been transmitted and stored in uncompressed form in many cases, such as in professional video links (3G/6G/12G-SDI), IP transport (SMPTE2022 5/6 & proprietary uncompressed RTPs), Ethernet transport (IEEE/AVB), and memory buffers. A low-latency lightweight image coding system allows for an increased resolution and frame rate, while offering visually lossless quality with reduced amount of resources such as power and bandwidth at a reasonable level. This document provides use cases and requirements for such low-latency lightweight image coding system.

2 Introduction

Infrastructures and systems capabilities increase at a lower pace than content resolution and are therefore progressively becoming the bottleneck in many applications. As an example, NHK has announced plans to test 8K TV broadcast in 2016 and foresees a full service by 2020 Tokyo Olympics. Transmitting uncompressed 8K content over existing links or over soon-to-be-available alternatives channels is not feasible. A lightweight low-latency coding system appears to be a smart and affordable solution to meet market needs.

Even for transmitting content fitting currently available systems, the use of a lightweight and low-latency coding system could be beneficial as it allows for reduced bandwidth, and consequently results in lowering corresponding cost or enable longer cable runs (for example, usual 3G-SDI cable run is 14m while it reaches 21m for HD-SDI, 50m for SD-SDI).

In a nutshell, the key advantage of a lightweight and low-latency image coding system is to allow increasing resolution and frame rate in a cost-effective manner, i.e.

- **safeguarding all advantages of an uncompressed stream**
  - low power consumption (through lightweight image processing)
  - low-latency in coding and decoding
  - easy to implement (through low complexity algorithm)
  - small size on chip and fast software running on general purpose CPU with the use of SIMD and GPU.
- **without significant increase in required bandwidth**
  - low power consumption (through reasonable bandwidth interfaces)
  - longer cable runs
  - SRAM size & frequency reduction with a frame buffer compression
  - more adequate for current infrastructures

3 Use Cases

3.1 Transport over video links and IP network

In such use cases, a video link and transport protocol is employed to transport video streams at a higher throughput than its physical throughput, thanks to a lightweight compression with a compression ratio ranging from 2:1 to 6:1. Several examples are given in the table below.

Video stream	Video throughput[1]	Physical link	Available throughput	Comp. ratio
2K / 60p / 422 / 10 bits	2.7 Gbps	HD-SDI	1.33 Gbps	~ 2
2K / 120p / 422 / 10 bits	5.4 Gbps	HD-SDI	1.33 Gbps	~ 4
4K / 60p / 422 / 10 bits	10.8 Gbps	3G-SDI	2.65 Gbps	~ 4
2K / 60p / 422 / 10 bits	2.7 Gbps	1G Ethernet (SMPTE2022 1/2)	0.85 Gbps	~ 3
2K / 60p / 444 / 12 bits	4.8 Gbps	1G Ethernet (SMPTE 2022 6)	0.85 Gbps	~ 6
4K / 60p / 422 / 10 bits	10.8 Gbps	10G Ethernet (SMPTE2022 1/2)	8.5 Gbps	~ 1.3
3x [4K / 60p / 422 / 10 bits]	32.4 Gbps	10G Ethernet (SMPTE2022 6)	7.96 Gbps	~ 4
4K / 60p / 444 / 12 bits	19 Gbps	10G Ethernet (SMPTE2022 1/2)	8.5 Gbps	~ 2.2
2x [4K / 60p / 444 / 12 bits]	37.9 Gbps	10G Ethernet (SMPTE2022 6)	7.96 Gbps	~ 5
8K / 120p / 422 / 10 bits	85 Gbps	25G Ethernet	21,25 Gbps	~ 4

As shown in the table, the main applications targeted by these use cases are broadcast, digital cinema, and industrial vision applications.

3.2 Real-time video storage

Embedded devices such as cameras use internal storage to store large streams of images. These devices offer limited access rates (i.e. approx 500 MBytes/s (4Gbit/s) for SSD drives, approx 50-90 30MBytes/s (400-720 Mbit/s for SD cards). Lightweight compression would allow real-time storage of video streams with throughputs higher than these access rates.

[1] Gbps = Giga bit per second

Video stream	Video throughput	Physical storage	Access rate	Comp. ratio
UHD / 60p / 422 / 10 bits	10 Gbps	SSD Drive	~ 4 Gbps	2.5
HD / 30p / 422 / 10 bits	1.2 Gbps	SD card	~ 0.5 Gbps	2.4

3.3 Video memory buffer

Buffer compression reduces the system form factor's weight, decreases the number of interconnect wires and extends the battery life for battery powered systems

- Upscaler/downscaler
- buffer for high refresh rate displays (120~600 Hz, Triple Flash)
- storage and replay buffer for high speed camera
- key frame buffer for AVC/HEVC 4K decoder

3.4 Omnidirectional video capture system

Omnidirectional video capture systems are assembled from a multitude of cameras mounted on a platform. Each camera covers a certain field of view which tends to overlap with that of its adjacent cameras in order to facilitate image stitching.

The proposed use case addresses the challenge of concurrently transferring and storing the image streams from each camera to a front-end processing system. In order to reduce the required bandwidth and therefore allow multiple cameras to send their data over a shared physical link, a lightweight, real-time compression between 2:1 and 6:1 of the image data at the camera is desired. Furthermore, this compression should be configurable in a transparent manner. Applying such compression will furthermore reduce both the required storage size and throughput demands of the storage sub-system on the front-end processing system.

3.5 Head mounted display for Virtual or Augmented Reality (VR/AR)

Omnidirectional VR and AR content is highly suitable for viewing through head mounted displays ("HMD"). HMDs are either tethered (i.e. connected through a cable) or wireless in which case the display is battery powered. Furthermore, with omnidirectional content, the HMD will only show that portion of the media stream which is within the viewers field of vision. Given the computational (and power) constraints of such a display, it can not be expected to receive the full image stream and then locally perform the required filtering onto the viewers field of vision – this needs to be done upstream and based on spatial data received from the HMD.

From the viewer's perspective, the quality of experience is crucially tied to the latency with which the system reacts to changes in his spatial gaze. An immersive experience requires very high resolution video - well beyond HD. These requirements lead to the need for adaptive strategies which allow to transmit, switch between and decode multiple high resolution image streams (each covering a certain spatial region) while decoding the video streams with imperceptible latency.

3.6 Image sensor compression (raw-Bayer)

Image sensor compression has several advantages compared to the usual compression in the RGB domain.

First, in terms of complexity, less data needs to be processed : in a Bayer pattern coming from a sensor, there are 3 times less data to process than the subsequent RGB image obtained after debayering. This induces a drastic reduction in terms of buffering requirements. Moreover, performing the compression as close as possible from the sensor allows to bring all advantages of the JPEG XS compression to segments of dataflows that were still using uncompressed data so far. This makes JPEG XS beneficial for even more use cases and markets, as listed below.

The main target applications for image sensor compression are described below.

Broadcast and high-end cameras

The camera control unit (CCU) is an essential unit in a live television broadcast chain. It is responsible for powering the professional video caméra, handling signals sent over the camera cable to and from the camera, and can be used to control various camera parameters remotely. CCU usually converts the RAW data camera stream in an YUV 422 signals, by managing black level adjustment, denoiser, color conversion, white balance, debayering, gamma curve and chroma-subsampling. The video link between the camera and the CCU is currently always uncompressed RAW data, with a minimal latency (<<1ms). With the increase of resolution, the need of higher speed cameras and long distance remote CCU, there is a need of extremely low latency & low buffering raw compression with a compression ratio from 2:1 (supporting even mathematically lossless on some image) to 4:1 (visual transparency).

Prosumer and consumer cameras (including mobiles)

With the increase of the sensor resolution, the high volume market faces the challenge to always offer better high-quality products while keeping a low cost and low power approach. Decreasing information transfers by compressing the image-sensor data is an attractive solution to reduce power consumption. To be successful in this competitive environment, we must provide the smallest sensor-data image compression scheme, with the lowest memory as possible, to reduce the fabrication cost, and more importantly the power consumption and heat dissipation. In the sectors where the uncompressed image is the norm, the image quality is very sensitive and the compression must be transparent to be an appealing option. The required memory size of a combined encoder-decoder would usually be 8 lines of samples of an uncompressed raw Bayer image.

Machine vision

Machine Vision Systems typically have cameras for capturing images and powerful image processing systems for image analysis. The cameras itself are separated from the image processing systems due to mounting and dimension constraints. It is beneficial to use the cameras only as sensor devices with as few computation capabilities as possible. A debayering in the camera would increase the computational load and power of such devices. Only the compression should be added. For automatic analysis a debayering to RGB images is not necessary. In terms of latency, ultra low latency is not necessarily a requirement but it has however to be a fixed and constant value.

Automotive industry

One of the most important applications in the automotive industry is autonomous driving. For this application multiple sensors are combined and their data transmitted to a so-called ECU (Electronic central unit) in which the data from the sensors are jointly analysed and related actions calculated. Data from these sensors need to be processed with a maximum responsiveness, therefore implying a very low latency along the whole dataflow. In terms of implementation and given the number of sensors, power consumption needs to be constrained as much as possible because of thermal considerations and the necessary operation in all kind of climatic conditions. The cabling inside the car is critical, e.g. fiber is not likely to be used as they are not as robust as copper cable and can not be bent. High resolution cameras and limited interface bandwidths require compression to reduce the bandwidth during live transmission and/or aggregating multiple sensor streams on one port/cable. For automatic analysis a debayering to RGB images is not important. Given the infrastructure encountered in a car, multi-platform implementations of such compression scheme would be required, including GPU, FPGA, and ASIC.

The above-described use cases can use different types of interfaces, each with its own throughput. The table hereunder indicates the targeted interfaces and their corresponding application areas.

Interfaces	**Nominal throughput [Gbps]**	**Application areas**
MIPI A-PHY	2, 4, 8, 12, 16	● Vehicles
MIPI C/D-PHY	~6/lane	● Vehicles ● Mobiles ● Sensors ● Pro-AV
Ethernet Cat-5e/Cat-6e	up to 5 for Cat-5e	● Vehicles

	up to 10 for Cat-6e	● Broadcast ● Pro-AV ● Machine vision
DDR3/DDR4 access	~23	● Pro-AV ● Mobiles
PCI Express	~16-32/lane	● Mobiles ● Pro-AV
SATA	~3-6	● Mobiles ● Pro-AV
LVDS	~1-3	● Broadcast ● Vehicles ● Machine vision

4 Target markets

There are several target markets, among which:

- Broadcast applications and live production
- Live-production
- Digital Cinema applications
- Industrial vision
- Professional audio visual systems
- Consumer TV
- Mobile video applications
- Camera array based recordings
- Ultra high frame rate cameras
- Medical Imaging
- Video Surveillance and security
- Automotive Infotainment
- Camera manufacturers
- Set-top boxes
- Low-cost visual sensors in Internet of Things (IoT)
- HMD displays

5 Requirements

This Section presents the requirements that should be met by the proposals so as to be suited for the above described use cases. Requirements are split between “core coding requirements” and “optional features”.

5.1 Core coding requirements

5.1.1 Uncompressed image attributes

- *Image resolution*: from VGA (640x480) up to 10K
- *Component subsampling*: 420, 422, 444, 4224, 4444
- *Component type*: RGB, YCbCr, CFA (Bayer patterns)
    - Input type of the encoder shall match output type of the decoder.
    - Internal color space conversion is permitted .
- *Component bit-depth*: 8 to 16 bits per component (bpc), integer (up to 12 bits in the first phase of the specification).
- *Frame rate*: from 24 fps to 120 fps, progressive content.
- *Content*: natural, synthetic, screen content
- *Supporting different color spaces*, including Rec. BT 709 [1], Rec. BT2020 [2], P3D65 [3], LogC.

5.1.2 Compressed bitstream requirements

- High image picture quality.

  The general requirement for JPEG XS image picture quality is that the difference between the original image or image sequence and the same image or image sequence after compression and decompression is not detectable by a human observer under normal viewing conditions (*visually lossless quality*), or does not lead to a performance decrease in case of machine vision processing (*automatic-analysis-resilient quality*). Moreover, in a compressed video sequence, any compressed image frame from the sequence shall individually achieve such quality. For visually lossless quality, corresponding compression ratios to achieve this quality range from 2:1 to 10:1.

  For certain use cases, higher quality is required and shall also be achievable by JPEG XS:

    - *Mathematically lossless quality*.
    - *Flickering-resilient visually lossless picture quality*. In this case, the picture quality achieved after compression and decompression shall successfully pass the flickering test, as described in Annex B of ISO/IEC 29170-2:2015 [4]. Corresponding compression ratios to achieve this quality range from 2:1 to 6:1.
- Support of a variable bitrate (VBR) & constant bitrate (CBR) mode
- Ability to define a strict maximum compressed size per frame
- Guaranteed avoidance of target rate exceedance
- Self-contained compressed frame: a compressed frame shall contain all information required to completely recover the corresponding uncompressed frame.
- Robustness to multiple encoding-decoding cycles, equal to or above 7 cycles.
    - Live-acquisition requires several encoding/decoding cycles, with different kinds of intermediate processing operations (overlay, crop, editing, pan&scan).

5.1.3 Design requirements

- Low-latency:
  - a maximum algorithmic latency of 32 video lines for a combined encoder-decoder suite is required.
  - the codestream syntax shall be defined in such a way that it does not prevent an encoder and decoder implementation to fit in the targeted FPGA device with an end-to-end latency of 32 lines.
  - in the context of CFA data processing, a "line" has to be understood here as a line of "super-pixels". A "super-pixel", as defined in ISO/IEC 21122-1 2nd Edition, is a 2×2 arrangement of pixels in a CFA pattern array.
- Low complexity in hardware and software: the algorithm shall be defined in such a way to allow for low complexity implementations on multiple platforms. As an indication, to process real-time 4k 4:4:4 8bit 60p with a compression ratio compliant with the above requirement, neither encoder nor decoder should require more than 50% of an FPGA similar to Xilinx Artix 7 [5] or 25% of an FPGA similar to Altera Cyclon 5 [6]. The target of an optimized software implementation able to real-time process 4k 4:4:4 8bit 60p should be an i7 processor, or equivalent.
- Support of different kinds of end-to-end parallelization, for CPU, SIMD, GPU, FPGA and ASIC. Hence, there shall be no serial bottleneck in the encoding and decoding process.
- Implementation scalability: the resources required by the encoder and the decoder shall scale depending on required throughput.
- No external memory for hardware implementations (FPGA / ASIC).
- Multiple platform interoperability (FPGA / ASIC / GPU / CPU): for frequencies ranging from 100MHz to 3GHz, circuits of different type shall produce the same codestream and be interoperable, so as to enable massive adoption.
- The codestream syntax shall be defined in such a way that it does not prevent an encoder to deterministically generate a codestream that only depends on the currently encoded frame, while still meeting latency and complexity requirements.
- Configurability:
  - Image size, frame rate, bit-depth (bpc), component type, subsampling
  - Targeted compressed bitrate (bpp), VBR or CBR, maximum compressed size per frame
  - Option has to be given to disable optional features (see Section 5.2 hereunder), as disabling features might lead to smaller hardware or software footprint.

5.2 Optional features

- Robustness to post-processing operations (such as subsequent editing operations, color transform or gamma conversion that shall not induce visual artefacts).
- Handling of different transfer functions: the proposed algorithm should optimize its performance by taking into account the transfer function being used by the content to be processed.

- Avoidance of SDI forbidden values
    - According to SMPTE 292 and 425, (10 bit) video data values 000h – 003h and 3FCh – 3FFh are excluded and reserved for sync words (EAV,SAV and ancillary data start). These markers play a specific role for the descrambling and synchronization of the SDI data.
    - The image coding system should define more efficient SDI mapping operations.
- Robustness to error: independently from error protection mechanisms available at transport level, the image coding system should minimize the impact of random bit flips (in the case of transport over SDI links) and packet losses (in the case of transport over IP).
- Support for extracting multi resolutions from the generated codestream.

References

[1] ITU-R BT.709-5 (2002), “Parameter values for the HDTV standards for production and international programme exchange”. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-5-200204-I!!PDF-E.pdf.

[2] ITU-R BT.2020 (2012), “Parameter Values for ultra-high definition television systems for production and internal programme exchange”. http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2020-0-201208-I!!PDF-E.pdf

[3] SMPTE ST 428-1:2006 “D-Cinema Distribution Master – Image Characteristics”. P3 with D65 white point.

[4] ISO/IEC 29170-2, “Information technology -- Advanced image coding and evaluation methodologies -- Part 2: Evaluation Procedure for nearly lossless coding”.

[5] http://www.xilinx.com/products/silicon-devices/fpga/artix-7.html#productTable

[6] https://www.altera.com/products/fpga/cyclone-series/cyclone-v/features.html