ISO/IEC JTC 1/SC29/WG1 N91014
91st JPEG Meeting, Online, 19-23 April 2021

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG	**JPEG**
Joint Bi-level Image Experts Group	Joint Photographic Experts Group

TITLE:	JPEG AI Use Cases and Requirements
SOURCE:	WG1
STATUS:	Draft
REQUESTED ACTION:	Distribution
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 Introduction

The scope of JPEG AI is the creation of a learning-based image coding standard offering a single-stream, compact compressed domain representation, targeting both human visualization, with significant compression efficiency improvement over image coding standards in common use at equivalent subjective quality, and effective performance for image processing and computer vision tasks, with the goal of supporting a royalty-free baseline.

2 JPEG AI Framework

Learning-based image coding solutions have already shown that they can achieve substantially better compression efficiency than available conventional image coding solutions, namely by exploiting advanced machine learning tools, such as deep neural networks [1]. In particular, it has been shown that, when compared to JPEG, JPEG 2000 and HEVC Intra, learning-based coding solutions can provide better perceptual quality, for some target bitrates, both in terms of appropriate objective quality metrics and subjective assessment scores [2]. Besides their high compression efficiency, learning-based image coding solutions may be adapted with little extra effort to image processing and computer vision tasks without the need for full decoding, i.e., without performing image reconstruction. This contrasts with classical image codecs, which when used in image processing and computer vision pipelines, need to perform full decoding of the compressed bitstream to obtain a pixel-based representation and extract features from decoded images, thus eventually suffering from compression artifacts.

Figure 1 shows the high-level JPEG AI framework, highlighting the three pipelines. The input to the learning-based image coding framework is a digital image and the output bitstream may be processed for human visualization by performing entropy decoding and standard reconstruction, thus producing a standard decoded image. As shown in Figure 1, the standard reconstruction may be skipped since the latent representation produced by the encoder contains the necessary information not only for decoding but also to perform image processing and computer vision tasks at the decoder side (after entropy decoding). These tasks are carried out on the latent representation, directly extracted from the original image and not from the (lossy) decoded image. This intrinsically feature-rich latent representation can be used in two main ways: 1) to perform an image processing task, e.g. targeting the enhancement or modification of the image, where a processed image is produced, for example with increased resolution, contrast, etc.; and 2) to perform a computer vision task where high-level semantic information is extracted, e.g., in the form of classes, labels, regions, etc.

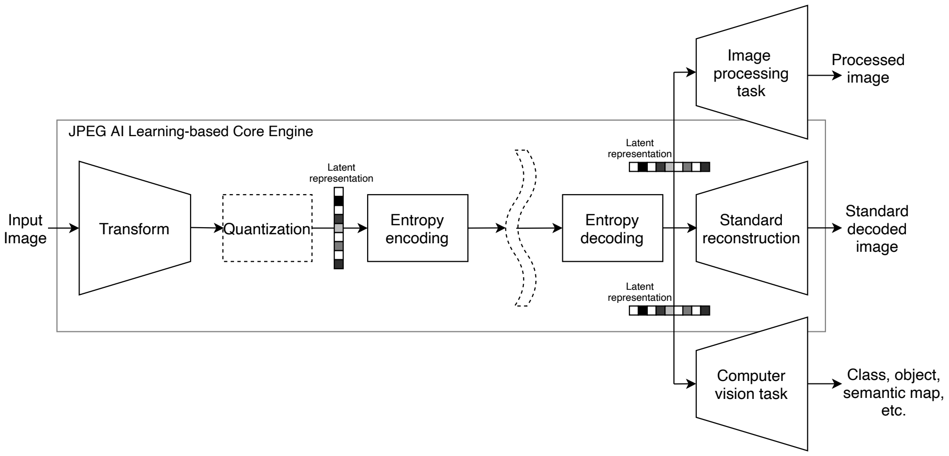

Fig. 1: JPEG AI learning-based image coding framework.

3 Key Tasks

Following the JPEG AI scope, the compressed bitstream will have a triple-purpose, thus offering compelling advantages for use cases where an image processing task aims to enhance or modify the image or where semantic (or higher-level) information needs to be extracted from large amounts of visual data. This may have a significant impact on image processing and computer vision tasks performance, which may be performed with lower complexity by using as input the compressed domain representation instead the original or decoded images; moreover, compressed domain features extracted from original and not lossy decoded images are used. Some examples of relevant image processing tasks are:

- Super-resolution
- Denoising
- Low-light enhancement
- Color correction
- Exposure compensation
- Inpainting

Relevant examples of computer vision tasks are:

- Image retrieval and classification
- Object detection, recognition and identification
- Semantic segmentation
- Event detection and action recognition
- Face detection and recognition

4 Use Cases

This section presents the use cases potentially targeted by the JPEG AI standard on learning-based image coding. These are used to motivate the requirements that are defined in Section 5.

4.1 Cloud storage

There has been an increasing number of images being stored in the cloud, due to the emergence of several online storage services. Several companies such as Tencent, Microsoft, Facebook and Google, often have thousands of billions of photos stored, which require considerable amount of resources, notably storage space, bandwidth or energy. Therefore, creating a highly efficient image coding solution for cloud storage is rather important to minimize costs, and even marginal savings in bitrate for some target quality may have a significant impact. The use of learning-based image compression may allow to optimize storage space, thus leading to high quality images at a fraction of the cost. In addition, high compression efficiency allows lower bandwidth costs, which translates into easy transmission and sharing of massive amounts of images. Besides compression efficiency, an efficient compressed domain representation is also desirable in this type of service, since it would facilitate several image processing tasks, such as super resolution or image denoising.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for image processing tasks
d. Perceptual quality optimization

e. Privacy preservation

4.2 Visual surveillance

Visual surveillance systems are widely deployed to perform video monitoring with several objectives, such as anomaly detection, detection of suspicious activity, provision of forensic evidence and intelligent control. Often, a very large number of cameras generate huge amounts of visual data that needs to be processed, compressed, analyzed and stored. Intelligent surveillance systems are often used to record relevant events not just as video but also as very high-resolution images. In some cases, non-visible light images (and even range maps) are also acquired. Considering the amount of data, the challenge is not only acquiring and compressing visual data but also understanding what is relevant and what can be ignored in an automatic way. Thus, image processing or computer vision tasks are often employed in order to allow efficient navigation and abnormal activity detection. Examples of such tasks are image classification, object recognition, action recognition and foreground-background segmentation. These tasks can be efficiently performed on the latent representation produced by learning-based image codecs.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for computer vision tasks
d. Effective compressed domain representation for image processing tasks
e. Low complexity encoding
f. Spatial random access, especially for very high resolution cameras
g. Privacy preservation

4.3 Autonomous vehicles and devices

Self-driving cars, drones and other autonomous devices generate a vast amount of visual data that must be analyzed and sometimes stored. Moreover, images collected from autonomous vehicles and devices may need to be processed offline and thus efficiently transmitted and/or stored. For example, drones carry cameras that are programmed to capture several Gigabytes of high-resolution aerial imagery which can be difficult to transmit over resource-constrained connections. Moreover, autonomous driving technology and other automated assistance systems may use several cameras for real-time analysis and decision, but the storage and transmission of key events allows other useful applications, such as traffic monitoring, accident investigation, etc. This scenario often involves several computer vision tasks, such as texture recognition, object recognition, foreground-background and other forms of semantic segmentation and event recognition. These tasks can be efficiently performed on the latent representation produced by learning-based image codecs.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for computer vision tasks
d. Effective compressed domain representation for image processing tasks
e. Low complexity encoding
f. Lossy to lossless coding

g. Privacy preserving

4.4 Image collection storage and management

With the wide deployment of smartphones and other consumer devices, every person has a digital camera which is used to acquire and store images of relevant events in photo albums. This collection of images is often backed up on online web storage to avoid their loss in the event of failure or loss of the smartphone or digital camera. Moreover, since these images usually have very high resolution, they require a significant amount of storage space and their storage has to be organized in a convenient way, to facilitate their search and consumption. In this scenario, texture and image classification, object and action recognition can be applied to facilitate the management and organization of images. These tasks can be efficiently performed on the latent representation produced by learning-based image codecs.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for computer vision tasks
d. Effective compressed domain representation for image processing tasks
e. Low complexity encoding
f. Privacy preserving

4.5 Live monitoring of visual data

Live streaming of visual data has significantly increased, from professional services such as online lectures, videoconferences and webcasts but also other entertainment services, such as video game live streaming and, short-form personal videos (see snack culture). Often, such visual data has to be analyzed in order to detect inappropriate content (as it is often done in social media networks) that may violate polices but also to provide additional information such as labeling of faces, emotions, gestures and so on. Also, computer vision tasks could be applied to live images/videos to perform intelligent review, understanding and distribution of this type of content. This means that compressed domain tasks such as foreground-background separation, semantic segmentation and action recognition could be applied for this type of content in an efficient way from the latent representation produced by learning-based image codecs.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for computer vision tasks
d. Low complexity encoding
e. Privacy preserving

4.6 Media distribution

Billions of user-generated images are captured and transmitted over the internet daily. These images are often uploaded and converted into multiple quality versions and formats, being stored on worldwide servers for distributions. In such scenario, efficient image compression solutions will allow to lower the storage and transmission cost and are especially relevant to users with low-bandwidth wireless connections. Progressive decoding may also be desirable, which allows for useful previews while the image is still being received. This may take the form of lower-resolution versions of images which are sufficient to display in displays with lower resolution, without requiring the resources needed for the entire high-resolution version. Moreover, for this use case, super-resolution and denoising from the latent representation produced by learning based codecs could be desirable as it would enhance decoding quality with a minor complexity increase.

Key requirements for this use case are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Effective compressed domain representation for computer vision tasks
d. Effective compressed domain representation for image processing tasks
e. Low complexity decoding
f. Perceptual optimization

4.7 Television broadcast distribution and editing

Interactive services are becoming popular in television broadcasts since they enable the participation of new actors. TV producers favor incorporating live streaming from remote locations in their shows, which may increase the size of the audience. Such remote live streams can be received either from a professional reporting field-crew or from a common television viewer who uses a smartphone to capture surrounding events or to give an interview. Besides the requirement of being efficiently compressed, such streams often need to be processed by many steps, e.g., foreground extraction to separate a person in the frame from an unpleasant background without using a green screen; or increasing the spatial resolution of the image to conform to broadcast quality standards. All above brings a number of challenges to television industry that can be addressed by efficient compressed domain processing which would eliminate expensive additional steps of transcoding. Thus, for this use case, compressed domain super resolution and foreground extraction is a highly desirable feature in such scenario.

Key requirements for this application are:

a. Tunable lossy compression/quality
b. High compression efficiency
c. Efficient compressed domain representation for image processing tasks
d. Low complexity decoding
e. Perceptual optimization

5 Requirements

This section presents the requirements that should be met by the standard so that it can be employed for the above described use cases. Requirements are split between "core requirements", which are essential for the standard, and "desirable requirements" which are not mandatory but are encouraged to be supported, as they might enlarge the possible use cases; the adoption will be decided depending on their complexity/cost burden.

5.1 Original image requirements

Core requirements

The standard shall at least support the encoding of images with the following attributes:

- Image resolution: from thumbnail-size images up to 8K, as minimum.
- Bit depth: 8-bit and 10-bit.
- RGB color space (three channels) and monochrome (one channel).
- Different types of content, including natural (photographs, aerial/satellite, document scans and synthetic (illustrations//UI elements/comics).

5.2 Compressed domain requirements

The standard shall cover at least the core requirements and may also cover desirable requirements as well.

Core requirements

- Effective compressed domain image processing.
- Effective compressed domain computer vision tasks.
- Significant compression efficiency improvement over coding standards in common use at equivalent subjective quality.
- Reconstructed images with both high subjective quality and high fidelity as measured by full reference objective quality metrics and double stimulus subjective assessment protocols.
- Hardware/software implementation-friendly encoding and decoding (in terms of parallelization, memory, complexity, and power consumption).
- Support for 8- and 10-bit depth.
- Support for efficient compression of images with text and graphics.
- Support for progressive decoding.

Desirable requirements

- Support for higher bit depth (e.g., 12 to 16-bit integer and floating-point HDR) images.
- Support for region of interest-based coding.
- Support for progressive decoding up to lossless.
- Support for lossless alpha channel/transparency coding.
- Support for animated image sequences.
- Support for wide color gamut coding.
- Support for different color representations.
- Support for very low file size image coding (e.g. 64×64 pixel images).
- Support for a low-complexity profile - low encode/decode time even on resource-constrained hardware (e.g., mobile devices).
- Minimal generation loss when lossy compression is applied multiple times.

6 Royalty-free goal

The royalty-free patent licensing commitments made by contributors to previous standards, e.g., JPEG 2000 Part 1, have arguably been instrumental to their success. JPEG expects that similar commitments would be helpful for the adoption of a learning-based image coding standard.