Quality-driven Variable Frame-Rate for Green Video Coding in Broadcast Applications

Glenn Herrou, Charles Bonnineau, Wassim Hamidouche, Patrick Dumenil, Jerome Fournier, Luce Morin

Quality-driven Variable Frame-Rate for Green Video Coding in Broadcast Applications

Glenn Herrou, Charles Bonnineau, Wassim Hamidouche, *Member, IEEE,*
Patrick Dumenil, Jérôme Fournier and Luce Morin

***Abstract*—The Digital Video Broadcasting (DVB) has proposed to introduce the Ultra-High Definition services in three phases: UHD-1 phase 1, UHD-1 phase 2 and UHD-2. The UHD-1 phase 2 specification includes several new features such as High Dynamic Range (HDR) and High Frame-Rate (HFR). It has been shown in several studies that HFR (+100 fps) enhances the perceptual quality and that this quality enhancement is content-dependent. On the other hand, HFR brings several challenges to the transmission chain including codec complexity increase and bit-rate overhead, which may delay or even prevent its deployment in the broadcast echo-system. In this paper, we propose a Variable Frame Rate (VFR) solution to determine the minimum (critical) frame-rate that preserves the perceived video quality of HFR video. The frame-rate determination is modeled as a 3-class classification problem which consists in dynamically and locally selecting one frame-rate among three: 30, 60 and 120 frames per second. Two random forests classifiers are trained with a ground truth carefully built by experts for this purpose. The subjective results conducted on ten HFR video contents, not included in the training set, clearly show the efficiency of the proposed solution enabling to locally determine the lowest possible frame-rate while preserving the quality of the HFR content. Moreover, our VFR solution enables significant bit-rate savings and complexity reductions at both encoder and decoder sides.**

***Index Terms*—High Frame-Rate (HFR), variable frame-rate, Ultra-High Definition (UHD), High Efficiency Video Coding (HEVC).**

I. Introduction

THE DEPLOYMENT of the latest Ultra-High Definition TV (UHDTV) system [1] aims to increase the user's Quality of Experience (QoE) by introducing to the existing High Definition TV (HDTV) system [2] new features such as higher spatial resolution, High Dynamic Range (HDR), wider color gamut and High Frame-Rate (HFR) [3], [4]. Technical definition of the UHDTV signal is available in the BT. 2020 recommendation of the International Telecommunication Union (ITU) [1]. The Ultra-High Definition (UHD)-1 Phase 2 specification enables to increase the video frame-rate from 50/60 frames per second (fps) in phase 1 to 100/120 fps.

Several papers [5]–[11] have investigated the HFR[1] video signal and shown its impact to enhance the viewing experience by reducing temporal artifacts specifically motion blur and temporal aliasing [5]. Authors in [9] have conducted subjective evaluations and shown that the combination of HFR (100 fps) and high resolution (4K) significantly increases the QoE when the video is coded at high bit-rate. The subjective evaluations conducted in [10] have also demonstrated the impact of HFR (up to 120 fps) to increase the perceived video quality. Moreover, this study shows that HFR improves the quality of video content with camera motion, while lower frame-rates are more convenient for sequences with complex motion such as dynamic textures. The study conducted in [11] by the British Broadcasting Corporation (BBC) showed that down-conversion of HFR video to 50 fps would inevitably result in considerable quality degradation. Mackin *et al.* in [12] have presented a new database containing videos at different frame-rates from 15 fps to 120 fps. Subjective evaluations performed on the video database have demonstrated a relationship between the frame-rate and the perceived video quality and confirmed that this relationship is content dependent.

The main limitation of the HFR in practical transmission chain is the significant increase in coding and decoding complexities. The complexity increase may prevent the deployment of the HFR since the recent SW/HW codecs do not support real-time processing of high resolution video at high frame-rate (+100 fps). This complexity overhead, estimated to 40% of the encoding and decoding times, is required to increase the frame-rate from 60 to 120 fps. Moreover, the HFR may increase both bit-rate and energy consumption compared to lower frame-rates without significant quality improvements depending on the video content, as shown in [10], [12].

A number of research works have investigated Variable Frame-Rate (VFR) [13]–[16]. These VFR solutions use different motion-related features with tresholding techniques [13], [16] or Machine Learning (ML) algorithms [14], [15] to select the desired frame-rate. The main limitations of these solutions are either a static sequence-level frame-rate adaptation, which greatly reduces the possible coding gains compared to a dynamic adaptation, or the target application with frame-rates lower than *30fps* or *60fps*, making them unusable for the recent HFR format without major updates and thorough testing.

Manuscript received April 13th, 2020; revised September 6th, 2020 and November 17th, 2020; accepted December 9th, 2020. This work has been funded by the French government through the ANR Investment referenced 10-AIRT-0007.

All authors are with the Institute of Research and Technology (IRT) b<>com, 35510 Cesson Sévigné, France.

G. Herrou, C. Bonnineau, W. Hamidouche and L. Morin are also with INSA Rennes, Institut d'Electronique et des Technologies du NuméRique (IETR), CNRS - UMR 6164, VAADER team, 20 Avenue des Buttes de Coesmes, 35708 Rennes, France, (e-mail: glenn.herrou@insa-rennes.fr).

C. Bonnineau is also with the Direction Technique, TDF, 35510 Cesson-évigné, France

P. Dumenil is also with Harmonic Inc., 35510 Cesson-Sévigné, France

J. Fournier is also with Orange Labs, 35510 Cesson-Sévigné, France

[1]In this paper high frame-rate video refers to video represented by 100 frames per second and more (+100 fps).

In this paper we propose a content-dependent variable frame-rate solution that determines the critical frame-rate of HFR videos. The critical frame-rate is the lowest possible frame-rate that does not affect the perceived video quality of the original HFR video signal. The proposed solution is based on a machine learning approach that takes as an input spatial and temporal features of the video to determine as an output the critical frame-rate. This can be considered as a classification problem to derive the critical frame-rate among three possible considered frame-rates: 30, 60 and 120 fps. The motivation behind these three frame rates is mainly for compliance with the frame rates specified by ATSC [17] and DVB [18] for use with various broadcast standards. The subjective results conducted on ten HFR video contents clearly show the efficiency of the proposed solution enabling to determine the lowest possible frame-rate while preserving the quality of the HFR content. This VFR solution enables significant bit-rate savings and complexity reductions at both encoder and decoder.

The rest of this paper is organized as follows. Section II gives a short overview of related works on HFR video, including coding, quality evaluation and rendering, followed by the objective and motivation of the paper. The proposed variable frame-rate solution is investigated in Section III as a classification problem with two binary Random Forest (RF) classifiers. The ground truth generation and features extraction, used to train the two RF classifiers, are described in Section IV. Section V gives details on the training of the two RF classifiers. The performance of the proposed variable frame-rate solution is assessed in Section VI in terms of perceived video quality, compression and complexity efficiencies. Finally, Section VII concludes the paper.

II. Related Work

A. High Frame-Rate Video

The UHDTV signal, defined in the ITU-R BT. 2020 recommendation [1], introduces a number of improvements over the HDTV [2] aiming at providing a better visual experience to the user. Along with a wider color gamut and an increased bitdepth, which allow to depict real colors and avoid ringing artifacts respectively, the key features of the UHDTV signal enabling a better depiction of live content are the higher spatial resolution - up to 3840x2160 and 7680x4320 pixels - and increased frame-rate - up to 120 fps. The different experiments that lead to the definition of each characteristic of the UHDTV signal are summarized in [3], [19].

Particularly, high frame-rate video has been an active field of research in the last decade, with the goal of avoiding well-known motion-related artifacts, namely flickering, motion blur and stroboscopic effect, which are present in traditional HDTV frame-rates of 60 fps and lower. Flicker is a phenomenon in which unwanted visible fluctuations of luminance appear on a large part of the screen and occurs at low refresh rates on non hold-type displays (e.g. Cathode Ray Tubes (CRTs)). Several studies [4], [20] have shown that flicker can be eliminated, for UHDTV signals, by simply using a frame-rate higher than 80 fps. The stroboscopic effect is the result of temporal aliasing, where the frame-rate is insufficient to represent smooth motion of objects in a scene causing them to judder or appear multiple times. At a given frame-rate, strobing can be reduced by lowering the shutter speed of the camera. However, a lower shutter speed also increases motion blur, which is caused by the camera integration of an object position over time, while the shutter is opened. Thus, strobing artifacts and motion blur can not be optimized independently except by using a higher frame-rate [21].

Based on previous studies by Barten [22] and Daly [23], Laird *et al.* [8] defined a spatio-velocity Contrast Sensitivity Function (CSF) model of the Human Visual System (HVS) taking into account the effect of eye velocity on sensitivity to motion. In [7], Noland uses this model along with traditional sampling theory to demonstrate that the theoretical frame-rate required to eliminate motion-blur without any strobing effect is 140 fps for untracked motion and as high as 700 fps if eye movements are taken into account. Since this theoretical critical frame-rate is not yet achievable, several subjective studies have investigated the frame-rate for which motion-related artifacts are acceptable for the HVS. In [24], Selfridge *et al.* investigate the visibility of motion blur and strobing artifacts at various shutter angles and motion speeds for a frame-rate at 100 fps. Their subjective tests showed that even at such a frame-rate, all motion-blur and strobing artifacts can not be both avoided simultaneously. Kuroki *et al.* [6] conducted a subjective test with frame-rates ranging from 60 to 480 fps, concluding that no further improvements of the visibility of blurring and strobing artifacts were visible above 250 fps. Recently, Mackin *et al.* [5] have performed subjective tests on the visibility of motion artifacts for frame-rates up to 2000 fps, achieved using a strobe light with controllable flash frequency. The study concluded that a minimum of 100 fps was required to reach tolerable motion artifacts.

For the purpose of the UHDTV signal definition, several studies further investigated the importance of HFR for television [9], [11], [25]. Emoto *et al.* [25] showed that increasing the frame-rate from the traditional 60 fps to high frame-rate of 120 fps provides a significant visual quality improvement. It is also stated that a further increase to 240 fps would also improve the motion portrayal but to a much lesser extent than the transition from 60 to 120 fps. Salmon *et al.* [11] have also studied HFR for television, showing that at least 100 fps is required for improvements over HDTV, especially for content with high motion such as sports. Recently, with one of the first 65 inches UHD HFR prototype displays, Hulusic *et al.* [9] studied the joint and independent contributions of 4K resolution and HFR. The subjective tests carried out showed that the 2160p100Hz format enables a significant increase in visual quality over other configurations - 1080p50Hz, 1080p100Hz and 2160p50Hz - but also that the improvements are strongly content dependent.

B. Compression of HFR content and Variable Frame-Rate

Since the adoption of HFR in the future television standard, through the second phase of the Digital Video Broadcasting (DVB) UHD standard [26], several studies of HFR content

compression have been carried out. Authors in [10] investigated the impact of high frame-rate on video compression, focusing on the perceptual quality of different motion types and frame-rates at several bit-rates. Using the test sequences of the public HFR dataset described in [12] compressed using an High Efficiency Video Coding (HEVC) encoder, it is shown that HFR is beneficial and desirable, especially at high bit-rates and even at the current HDTV broadcast data rates for sequences containing camera and/or simple motion, for which the encoder can make use of the increased temporal correlation to predict adjacent frames. Sugito *et al.* [27] showed that the overhead, in terms of bit-rate, introduced by the increase from 60 to 120 fps, is reasonable, with an optimal bit allocation of 6-7$\%$ of the total bit-rate for the additional frames needed to achieve HFR capability. However, one of the main limitation of doubling the frame-rate is the additional encoding complexity, with a near 40% increase in encoding time.

VFR, where the image frequency can be adapted based on the signal characteristics, is one of the solutions to cope with the complexity and bit-rate increases. Authors in [13] have studied the impact of both frame-rate and Quantization Parameter (QP) on the perceived video quality. They also developed an accurate rate model and quality model based on single layer and scalable video bitstreams. These models have been applied to frame-rate based adaptive rate control for single-layer and scalable video encodings. However, these models have been designed for frame-rate decisions between values ranging from *7.5fps* to *30fps*, which corresponds to a very low motion portrayal quality compared to *120fps* videos. In [14], a Support Vector Regression (SVR) is used to predict a satisfied user ratio - percentage of people who do not see the difference between original and lower frame-rates - which is then used to dynamically select the appropriate image frequency at a Group Of Pictures (GOP) or sequence level. The trained SVR uses complex and computationally demanding features, notably a visual saliency map and a spatial randomness map for each frame, thus making it unsuitable for real-time dynamic frame-rate selection. In addition, the training set is only composed of up to 60 fps content, limiting the frame-rate choice to *60fps*, *30fps* and *15fps*. With their design targeting lower than *120fps* maximum frame-rates, there is no guarantee that the features used by the VFR models proposed in [13] and [14] would also work on *120fps* content due to the different motion portrayal observed in HFR content.

More recently, VFR for HFR content has been investigated, aiming at offering a perceptually indistinguishable temporally downsampled video [15], [16]. Katsenou *et al.* train Bagged Decision Trees to predict the critical frame-rate at a sequence level [15]. The selected feature set is only composed of an Optical Flow (OF) for the temporal aspect, and Gray Level Co-occurrence Matrix (GLCM) for the spatial details contribution. In addition, the considered dataset consists of 22 test sequences with critical frame-rates of 60 or 120 fps. Thus, a good generalization of the VFR decision problem is hard to achieve with such few data to train and validate the model. In [16], a dynamic frame-rate adaptation is proposed based on the frame-rate dependent metric FRQM [28]. The temporal adaptation is coupled to a spatial resolution adaptation using kernel-based downsampling and a neural network based upsampling at respectively the pre and post-processing stages. The spatio-temporal adaptation model shows high coding gains through both objective and subjective tests. However, the authors indicates that the temporal adaptation has only been used for one sequence in their dataset, which only contained *60fps* sequences, making the performance evaluation of the VFR part of the solution difficult, especially for HFR content.

C. Motion Blur Rendering and Video Frame Interpolation

In a pipeline using a VFR video format to transport the video, several processing steps could be added to improve the perceptual quality of the output video. Indeed, on one hand, motion blur can be synthesized when the frame-rate is lowered to render a video close to what would have been captured with a camera at the lower frame-rate and its corresponding shutter speed. This would reduce the stroboscopic effect due to the frame decimation thus improving the visual quality of the VFR video. Motion blur synthesis has been extensively studied in an effort to render synthetic images as real as possible [29]. These techniques mostly rely on the perfect knowledge of the depth and motion of the scene and are thus not compatible for a live broadcast use-case. More recently, a motion blur rendering algorithm using only two consecutive images as inputs to produce a motion blurred output have been designed in [30]. These promising results are balanced by the computationally demanding algorithm, due to the underlying Convolutional Neural Network (CNN) architecture used to synthesize motion blur.

On the other hand, since most display devices do not support a variable frame-rate, the VFR video must be temporally upsampled to the original higher frame-rate before displaying it. Thus, frame interpolation methods can be used to improve the temporal upsampling step in order to obtain a visually better displayed video. Video frame interpolation is a well-studied field with several existing approaches to the problem. The classical approach interpolates intermediate frames from the optical flow field [31] of the scene. Interpolated frames, whose quality highly depends on the accuracy of the computationally expensive optical flow computation [32], typically suffer from motion boundaries and severe occlusions thus showing strong artifacts, even with state-of-the-art optical flow algorithms [33]. More recent promising works rely on neural networks to either predict convolution kernels for each pixel used to generate the interpolated frames [34] or leverage optical flow fields with exceptional motion maps [35]. However, these techniques involve a large number of convolutions, sometimes with large kernels (up to 41x41 for each pixel) to cope with large motion, thus making the computational demand unsuitable for real-rime use-cases.

D. Objective and Motivation

Most existing algorithms containing variable frame-rate have been designed purely for rate control in 30 fps video encoding schemes, with the goal of skipping frames when the bit budget constraint can not be met. Such behavior does not take properly into account the impact on perceptual

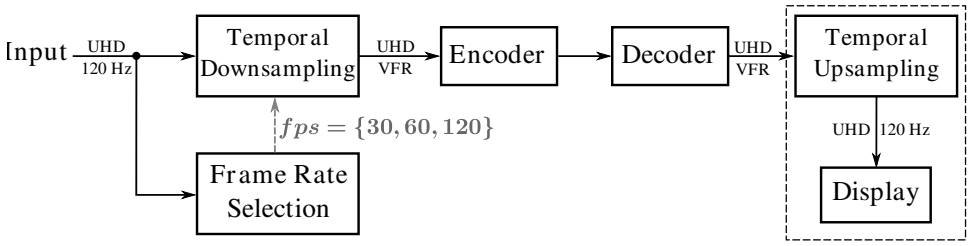


Fig. 1. Block diagram of the complete Variable Frame-Rate (VFR) coding scheme.

quality, making these solutions not suitable for HFR, which has been integrated to the UHDTV standard to improve motion portrayal.

Since the improvements brought by HFR capabilities are highly content dependent, several recent studies have based the frame-rate selection on perceptual factors to lower the frame-rate when there is no impact on visual quality. However, they use computationally expensive features and rely on a small dataset, not always composed of 120 fps content, to train and validate the variable frame-rate models. Thus, these solutions do not achieve good generalization of the problem. Moreover, they are not suitable for a real-time constraint, which is required for use-cases like live broadcast of, for example, sport events, that would periodically highly benefit from HFR.

In this paper, real-time variable frame-rate for HFR source content is addressed. The proposed system, depicted in Fig. 1, relies on two classifiers to predict the critical frame-rate. It has been designed with the following objectives:

- Design a dynamic frame-rate selection at a GOP level with perceptually invisible frame-rate changes.
- Low-complexity feature computation with low impact on the coding chain processing time.
- Training and testing on a well dimensioned dataset containing various types of 120 fps content.
- Perceptual validation of the obtained objective performance through subjective evaluation test.
- Assess the bit-rate and complexity savings within an HEVC encoding chain.

To meet the real-time constraint of live broadcast, state-of-the-art motion blur rendering and frame interpolation have not been integrated in the VFR pipeline. Instead, the proposed system has been designed with simple frame decimation (resp. duplication) as a temporal downsampling (resp. upsampling) tool. Additionaly, the RF algorithm has been chosen as classification technique thanks to a small benchmark comparing several ML techniques for VFR classification that showed a better trade-off between prediction accuracy and computational complexity for the RF algorithm.

III. Random Forest Classifier for Variable Frame-Rate

This section briefly presents Random Forests (RFs) as a classification tool and introduces the method proposed in this work to reduce the frame-rate with no visual impact, i.e. the VFR decision problem, as a combination of two binary classification problems.

A. Background on Random Forests

Random Forests [36] are a common ML tool used to solve classification problems. A RF classifier is able to predict the value of a target variable, i.e. a class, based on a set of input variables, i.e. input features, using the majority vote of an ensemble of nearly independent decision trees.

A decision tree is constructed by first partitioning the training dataset, i.e. the features and the associated class of each sample, into two different subsets, called nodes. This process is performed recursively until either all the node samples belong to a single class or a tree constraint has been reached. At each node, each available input feature is evaluated for all its possible values, in order to achieve the best separation of the classes in the subsequent child-nodes.

In this work, the criterion used to quantify the quality of a split, given the feature F and its threshold value t, is based on the Gini impurity measure, a common metric for Decision Trees [37]. It is computed as follows

$$I_G(D) = \sum_{c \in C} \mathbb{P}(c \,|\, D)\ (1 - \mathbb{P}(c \,|\, D))\,, \tag{1}$$

with D the sample set under consideration, C the set of possible class labels and $\mathbb{P}(c \,|\, D)$ the conditional probability of class c given the sample set D.

The best split is then obtained by finding the pair (F, t) that maximizes the Mean Decrease Impurity (MDI) ΔI_G defined by Equation (2).

$$\Delta I_G(D, F, t) = I_G(D) - \frac{|D_L|}{|D|}\, I_G(D_L) - \frac{|D_R|}{|D|}\, I_G(D_R), \tag{2}$$

with $D_L = \{x \in D, F(x) < t\}$ (resp. $D_R = \{x \in D, F(x) \geq \}$) the subset of sample set D for which each sample x has a value of feature $F(x)$ smaller (resp. larger) than threshold t and $|D|$ the cardinal of a set D.

To minimize the correlation between trees of the RF, the bootstrap aggregating, or bagging, technique [38] is used to construct the forest. This consists in training each tree T_i with a different subset D_i of the input data sample set D. Each D_i is obtained by a uniform sampling of D with replacement, i.e. replacing discarded samples by duplicates of a selected one. In addition and to further reduce the correlation between trees, only a random subset of the features, here $\sqrt{n}$ features with n the total number of input features, are evaluated at each node to find the best available split.

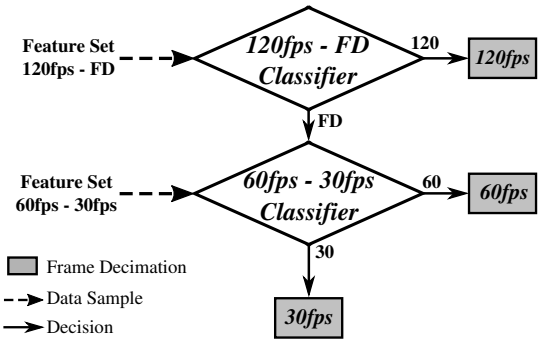


Fig. 2. Overall prediction scheme with cascaded binary RF classifiers.

B. VFR Classification Problem

The proposed solution aims at predicting when the frame-rate can be reduced, by discarding frames, without any perceptual impact on the quality of the original input HFR video. In an effort to keep the number of possible frame-rates reduced and obtain a regular frame decimation process, two frame-rates, *60 fps* and *30 fps*, were identified as potential candidates in addition to the original frame-rate of *120 fps*. The VFR decision problem thus becomes a three-class classification.

In this work, a combination of two successive binary RF classifiers has been chosen to solve the classification problem, as depicted in Fig. 2, instead of directly training a forest with multi-class outputs. This decision leads to a better overall performance by training both classifiers independently on separate datasets and features. Indeed, in addition to the specialization of each binary classifier, almost all samples of the database can be used for training either one or both classifiers while keeping balanced training datasets, as described in Section IV, thus increasing the accuracy of the overall model.

The first RF classifier, named *120fps - FD*, is specialized in deciding whether the frame-rate must remain *120 fps* or a Frame Decimation (FD) can be applied without impacting the visual quality. If the *120 fps* class is chosen by the first classifier, the frame-rate prediction process is stopped and no FD is applied. Otherwise, the prediction process continues by requesting the second classifier, named *60fps - 30fps*, which aims at selecting the appropriate lower frame-rate if a FD is applied on the input HFR video.

IV. Ground Truth Generation

One of the most crucial steps towards training a supervised RF model is to gather a dataset which will be used as ground truth, i.e. examples - features representing a sample and the sample class label - used by the model to learn to predict. It is thus important to have a ground truth that contains a good representation of all the real cases the model could encounter in order to achieve a good generalization of the problem. This section focuses on detailing the ground truth generation process, necessary to obtain the datasets used to train both RF classifiers. The HFR database is first presented, followed by the detailed methodology for subjectively determining the critical frame-rate labels. Then, the creation of the dataset is described, with the composition of the balanced training sets on one hand and the feature extraction process on the other hand.

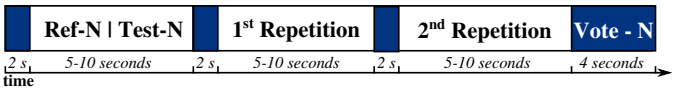


Fig. 3. Basic Test Cell (BTC) for the SDSCE evaluation method.

A. HFR Video Database

To the best of our knowledge, there is no publicly available dataset for the VFR classification problem except for the Bristol Vision Institute High Frame-Rate (BVI-HFR) database [12]. However, this database only contains 22 videos, which is rather small to train a reliable model. In addition, the temporal downsampling technique used to create the lower frame-rates in this database is frame averaging, whereas the chosen method in this work is frame decimation, which could lead to different decisions on the critical frame-rate.

Therefore, a new database has been gathered, composed of 375 native HFR video clips of 5 to 10 seconds. These clips are all uncompressed and stored in YUV format with 4:2:0 chroma subsampling and 8-bit depth. Their original frame-rate is 120 fps and spatial resolution 1920×1080 pixels (High Definition (HD)) - a downsampling has been performed for source videos of higher spatial resolution using Lanczos3 filter banks [39].

The database includes video clips from the BVI-HFR dataset [12], together with b<>com, Harmonic and other non-publicly available test sequences. In order to later evaluate the trained model on unseen data samples, 15 sequences, with heterogeneous spatio-temporal characteristics, have been extracted from the database, leaving 360 video clips to annotate before training both RF classifiers.

B. Critical Frame-rate Decision Methodology

The ground truth generation process requires each video of the database to be assigned a critical frame-rate chosen among the three considered ones in the VFR classification problem. A subjective test has thus been carried out with the objective of finding the lowest frame-rate for which no visual degradation can be observed compared to the original 120 fps video. Following the ITU-R BT.500-13 recommendation [40], the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) protocol has been used with a binary scale - either identical or visible difference - and two screens placed side-by-side. Therefore, subjects were asked, for each video of the database, if there was a visible difference between the two displayed videos, i.e. the known reference and the lower frame-rate, either 30 or 60 fps, test video.

The test was composed of 750 Basic Test Cells (BTCs), randomly divided into 30-minute test sessions. Each BTC is 40-second long and is composed of a 2-second message announcing the test video index followed by the side-by-side display of the reference and test videos with two repetitions.

A 2-second break displaying a mid-gray image has been added between each repetition. The BTC is concluded by a 4-second message asking the viewer to vote, as shown in Fig. 3.

TABLE I
DATABASE CRITICAL FRAME-RATE DISTRIBUTION.

	Critical frame-rate		
	30	60	120
# of shots w/ uniform motion	78	184	167
# of 4-frame chunks	3749	22327	19996

Due to the large duration of the subjective test, only five viewers, experts in video processing, participated in the whole database annotation. The final frame-rate decision is the lowest frame-rate for which the majority of expert viewers did not notice any visible difference with the reference 120 fps video.

The tests were conducted in a controlled laboratory environment, following the ITU-R Rec BT.500-13 [40]. Two identical 27-inch screens capable of displaying 120 fps content (Asus RoG Swift PG278Q) were used side-by-side, aligned and placed at a distance from the viewer position of three times the screen height. Each participant has been screened to ensure (corrected to) normal visual acuity and normal color vision.

C. Balanced Dataset Composition

The results of the expert subjective test are summarized in Table I. As can be seen, the sequences are not evenly distributed over the three possible frame-rates due to a large proportion of the available content being captured with HFR-capable devices containing high motion content, for which frame decimation downsampling to 30 fps is critical. Since the goal is to allow for a frame-rate adaptation at the lowest possible level, i.e. a frame-rate decision for every chunk of 4 frames to keep a regular frame decimation process, the sequence-level labels obtained via the subjective test have been extended to 4-frame chunk-level labels. Thus, for sequences with significant motion discontinuities, video shots with uniform motion can be identified and assigned different labels in a single sequence. Based on the observations made by the experts after the subjective test, a total of 429 video shots with uniform motion have been extracted from the 360 native HFR sequences. This refinement allows for a more accurate annotation of the database, thus avoiding chunks being annotated with inconsistent labels. For instance, a chunk not containing any movement associated to the *120fps* class. However, such cases could still remain in the ground truth due to the difficulty to identify the motion discontinuities with a precision of less than four frames.

From this ground truth, two different datasets were created to train the *120fps-FD* and *60fps-30fps* RF classifiers. The first one contains all samples of the *120fps* class as well as those from the *FD* class. The *FD* class is composed of all *30fps* samples and a random subset of the *60fps* samples. The amount of selected *60fps* samples has been chosen to produce a balanced dataset, i.e. to roughly obtain the same sample size for both the *120fps* and *FD* classes. The second dataset, used to train the *60fps-30fps* classifier, is comprised of the *30fps* samples and a random subset of the *60fps* samples, whose size has also been chosen to produce a balanced dataset. The choice of balancing datasets has been made because the unbalanced class distribution in the database does not necessarily represent the distribution of media content in a broadcast context, the chosen use-case for the proposed solution, but rather relates to the current difficulty to find HFR content with low motion. This is due to the fact that most of the currently available HFR content has been shot to demonstrate the gain in perceptual quality and motion portrayal brought by the technology.

D. Feature Extraction

The goal of a feature set is to gather the different metrics relevant to the considered classification problem that would help discriminate the output classes from one another. For the VFR classification problem, a first feature would intuitively be the motion information, e.g. the motion vectors between two consecutive frames. Indeed, high movement in a source HFR video will likely lead to visible temporal aliasing, i.e. stroboscopic effect, if a lower frame-rate is used after frame decimation. In addition, since motion blur is not added during the temporal downsampling process used in this work, lowering the frame-rate could introduce visible jerkiness in high motion videos.

For the three considered frame-rate classes, flickering, the other well-known motion-related artifact, can appear in highly textured areas where the local variation in luminance between two consecutively displayed frames would be visible at lower frame-rates. In an effort to capture this phenomenon in the feature set, the pixel luminance values and directional gradients can be used.

Based on the performed expert viewing sessions, it has been observed that small objects with high velocity, which would not necessarily be detected by the motion vectors depending on the used motion estimation algorithm, could induce visible artifacts at lower frame-rates. To take this observation into account, a simple metric capable of detecting both global displacements and small moving objects has been designed. This metric is based on the thresholding of the difference between two consecutive frames. First, the frame difference $D_n(i,j)$, i.e. the difference in pixel value of the luminance plane at the same location in space (i,j) between the n^{th} frame F_n and the preceding one F_{n-1}, is computed for each pixel using Equation (3)

$$D_n(i,j) = |\, F_n(i,j) - F_{n-1}(i,j)\,|. \tag{3}$$

Then, a thresholding operation is performed on the frame difference image, defined as follows

$$A_{n,Th}(i,j) = \begin{cases} 1 & if\ D_n(i,j) \geq Th \\ 0 & if\ D_n(i,j) < Th \end{cases}, \tag{4}$$

with $A_{n,Th}$ the resulting thresholding activation map for the n^{th} frame and a threshold Th. Fig. 4 depicts an example with both the original image and the resulting thresholded frame difference image.

The designed feature set is thus based on the following feature maps:

- **NormMV**, **HorMV**, **VerMV:** maps respectively representing the Motion Vectors (MVs) norm, horizontal coordinate and vertical coordinate.

(a) (b)

Fig. 4. Example of thresholded motion difference with (a) the original image of the *Jokey* sequence and (b) the thresholding activation map with threshold $Th = 25$.

- **ThreshDiffMap:** thresholded frame difference map as defined in Equation (4).
- **GradMag**, **GradHor**, **GradVer:** maps respectively representing the Sobel gradient magnitude, horizontal gradient and vertical gradient.
- **Luma:** pixel luminance map.

For each map, several scores have been computed, namely the mean value, the standard deviation, the maximum value and the mean of the 10% highest values, to produce a total of 32 different features that will serve as an initial feature set for the training of both the considered RF models.

V. Random Forest Training Process

Once the ground truth is available and the features computed, the RF models can be trained to solve the VFR classification problem. This section focuses first on the performance evaluation process, necessary to assess and optimize the quality of the model critical frame-rate prediction task. Then, a feature selection process, used to reduce the initial feature set to only the relevant features for each binary classifier, is presented. Finally, the classification results are presented and analyzed.

A. Model Evaluation Process

In order to optimize a ML classifier, it is necessary to use a metric capable of evaluating the model classification performance to find the best parameters. To do so, several common metrics, namely precision, recall and F1-score [41], can be used. First, the trained model confusion matrix has to be computed from the true and predicted labels of the dataset samples. Then, once the different quantities of the confusion matrix, namely True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN), are available either as number of samples or normalized probabilities, the precision, recall and F1-score can be computed as follows, with $C = \{c_1, c_2\}$ the set of classes for the binary classifier under test

$$precision(C) = \frac{1}{|C|} \sum_{c_i \in C} \frac{TP(c_i)}{TP(c_i) + FP(c_i)}, \quad (5)$$

$$recall(C) = \frac{1}{|C|} \sum_{c_i \in C} \frac{TP(c_i)}{TP(c_i) + FN(c_i)}, \quad (6)$$

$$F_1\text{-}score(C) = \frac{2}{|C|} \sum_{c_i \in C} \frac{precision(c_i)\, recall(c_i)}{precision(c_i) + recall(c_i)}, \quad (7)$$

As for any binary classification problem, the goal is to maximize the confusion matrix main diagonal values, i.e. the number of TP and TN representing the correct predictions. This can be achieved by maximizing precision, recall, or F1-score during the training process, depending on the considered classification problem and the criticality of each error type. For the VFR classification problem, the main goal is also to minimize the critical errors - predicted frame-rate lower than the ground truth - which would potentially induce visible temporal artifacts thus greatly reducing the output visual quality. To emphasize these critical errors and avoid them in the final model, another performance evaluation metric M_{crit} has been designed, as a combination of the precision of the lower frame-rate class and the recall of the higher frame-rate class, using Equation (8)

$$\begin{aligned} M_{crit}(C) &= \tfrac{1}{|C|} \left[precision(c_1) + recall(c_2) \right], \\ &= \tfrac{1}{|C|} \left[\tfrac{TP(c_1)}{TP(c_1)+FP(c_1)} + \tfrac{TP(c_2)}{TP(c_2)+FN(c_2)} \right], \\ &= \tfrac{1}{|C|} \left[\tfrac{TP(c_1)}{TP(c_1)+FP(c_1)} + \tfrac{TN(c_1)}{TN(c_1)+FP(c_1)} \right], \end{aligned} \quad (8)$$

with $C = \{c_1, c_2\}$ the set of ordered classes - frame-rate of c_1 lower than the frame-rate of c_2. This metric has been used together with the F1-score to assess the quality of RF models for both the feature selection process and hyper-parameter tuning described in the next sections.

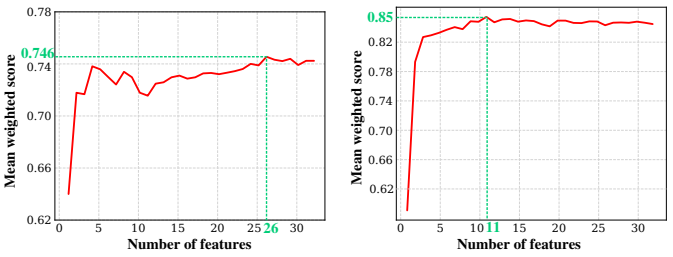
B. Feature Selection

In an effort to limit the model complexity and improve its performance, a dimensionality reduction algorithm has been used on the proposed initial feature set. Indeed, by only selecting the relevant features, thus removing features carrying useless information for the considered classification problem, both the feature computation time and training time are greatly reduced. Additionally, model over-fitting is also decreased when the size of the feature set is reduced due to the reduction of noise in the input data and the elimination of highly correlated features, i.e. features that would carry the same information about the target variable.

In this work, a Recursive Feature Elimination (RFE) process has been used to reduce the dimension of the initial feature set. It consists in recursively evaluating the model performance on a dataset and a feature in which the least important feature is removed after each iteration. The feature importance is computed in terms of mean decrease in Gini impurity, i.e. the average capacity of a feature to reduce the Gini impurity computed at a given tree node, using Equation (1). When the feature set size reaches the minimum tested dimension of 2, the feature set leading to the best model performance among all the tested dimensions is selected as the final feature set.

This process has been performed independently for both proposed RF models with the same initial feature set but leading to a different optimal feature set size for each binary RF classifier, respectively 26 and 11 features for the *120fps-FD* classifier and *60fps-30fps* classifier, as depicted in Fig. 5.

Fig. 6a shows the list of selected features for the *120fps-FD* RF classifier with their corresponding feature importance. As can be observed, the most relevant features to discriminate



(a) 120fps-FD RF model (b) 60fps-30fps RF model

Fig. 5. Recursive Feature Elimination with weighted (F1,M_{crit}) score.

samples from both classes are based on the two motion features *ThreshDiffMap* and *NormMV*. This correlates well with the observations made by the experts during the ground truth annotation subjective tests. Indeed, it was pointed out that above a certain amount of movement, either from a moving camera or an object with high velocity - which can be captured by both metrics -, a stroboscopic effect as well as jerkiness, due to lack of motion blur, could easily be detected with frame-rates lower than 120 fps. Most of the features based on spatial measures are present in the optimized feature set, with a significantly lower importance compared to the aforementioned motion features. This tends to indicate that flickering becomes an important criterion to keep a high frame-rate when the amount of movement does not induce other motion artifacts.

For the *60fps-30fps* RF model, the selected features and their importance are depicted in Fig. 6b. As for the first RF model, the features based on the motion vectors, *NormMV*, have a high capacity to discriminate samples from both classes. However, spatial features, based on the *Luma* and *GradHor* feature maps, hold a significantly higher importance compared to the first model, indicating that flickering mostly occurs at 30 fps for most of the videos of the training dataset.

C. Classification Results

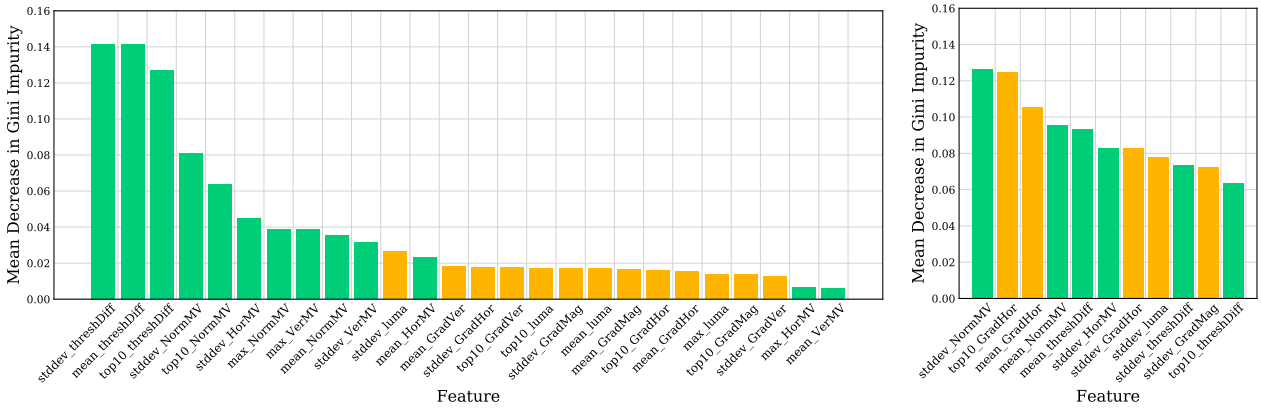
With the final feature sets, an optimization has been conducted on the maximum tree depth and number of trees hyper-parameters, leading to a final model with 200 trees of depth 7 for the *120fps-FD* classifier and 100 trees of depth 7 for the *60fps-30fps* classifier. An in-depth analysis of the prediction capability of the final models, both individually and combined to form the overall VFR prediction scheme, can then be conducted. It is important to note that the models have been trained using a 10-fold cross-validation so that the considered performance is a combination of the results from the validation fold of each iteration. This means that each tested sample prediction presented in the different confusion matrices has been obtained without using the validation sample for training. Additionally, the training set has not been shuffled, so that chunks from a same sequence could not be in the training and validation folds at the same time, thus avoiding a highly biased performance evaluation.

Figures 7a and 7b shows the resulting confusion matrices of both RF models, individually. For the *120fps-FD* classifier, error rates of 20% and 17% can be observed for the *120fps* and FD classes respectively, which represent a good performance considering the VFR classification problem and its imperfect ground-truth. Indeed, the frontier between annotating a sequence with a *120fps* label and a *60fps* label can be difficult to maintain consistent during the processing of the 360 videos of the training set. In addition, as detailed in Section IV-C, several sequences with high motion discontinuities have been separated into shots with different labels. Since these motion change frontiers could only be determined subjectively and not at a precise frame level, some dataset samples, i.e, 4-frame chunks, located at these frontiers could have been annotated with incorrect labels. Therefore, the error rate is likely to be over-estimated, leading to a visible motion artifacts rate lower than the observed 20%. For the *60fps-30fps* model, correct prediction rates are respectively 83% and 91% for the *30fps* and *60fps* classes. Critical errors, defined as critical frame-rate under-estimation, represent 9% of the *60fps* class samples. This rate can be problematic since the under-estimation with a frame-rate of 30 fps could lead to severe visible motion artifacts. However, the same aforementioned remark concerning imperfect ground truth applies to the training dataset of the *60fps-30fps* model. The proportion of frame-rate over-estimation errors, equal to 17%, does not impact the visual quality and is thus not as prejudicial as the critical errors.

The overall VFR prediction scheme confusion matrix, obtained by combining the cross-validation validation-fold predictions of both models, is depicted in Fig. 7c. It is important to note that only a subset of randomly chosen *120fps* class samples has been used to compute the overall prediction scheme confusion matrix so that the three classes have the same number of samples. The observed performance is consistent with the individual RF model prediction results, with good probabilities of correct prediction. The only significant change is the 69% correct prediction rate for the *60fps* class, which can be explained by the fact that it is the intermediate class, thus sharing characteristics with the other two classes which makes the discrimination of the class samples harder to generalize. In addition, the *extreme* errors, i.e. the critical under-estimation of a *120fps* sample with a *30fps* predicted label or the exact inverse, are rarely occurring with rates of 3% and 1%, respectively. This tends to bolster the hypothesis on the ground truth being imperfect due to possibly unstable/blurry annotation frontiers between adjacent labels. If this hypothesis is correct, the combined prediction model should lead to VFR output video sequences visually identical to the HFR input. However, the compression and encoding complexity gains should be slightly lower than with ground truth labels. The next section aims at verifying this statement.

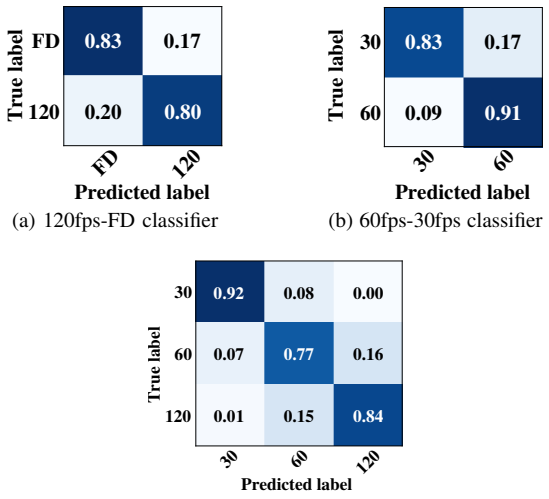
VI. Results and Analysis

Before analyzing the coding performance of the VFR coding scheme, the visual quality of the output VFR video must be evaluated to assess whether the RF model frame-rate decisions are preserving the perceptual quality compared to the HFR source video. This section first describes the characteristics of the test set sequences and the chosen subjective evaluation methodology. Then, the results of the subjective tests are



(a) 120fps-FD RF model (b) 60fps-30fps RF model

Fig. 6. Feature importance measured with Mean Decrease in Gini Impurity. *yellow: spatial features, green: motion features.*



(a) 120fps-FD classifier (b) 60fps-30fps classifier

(c) Overall scheme

Fig. 7. Individual classifier and overall scheme confusion matrices for a 10-fold cross-validation training with their respective datasets.

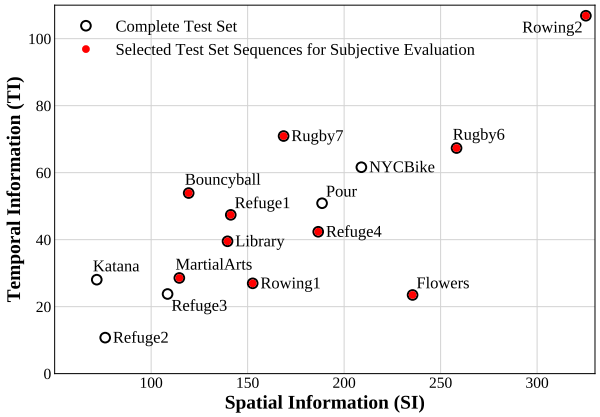


Fig. 8. SI-TI characteristics for test sequences of the three considered sets.

detailed and discussed for both uncompressed and compressed VFR videos. Finally, the coding performance of VFR coding scheme is presented in terms of bit-rate savings and complexity reduction.

A. Specific Test Datasets and Subjective Tests Motivations

A total of 15 sequences has been selected to validate the performance of the VFR model. These sequences are unknown to the model, i.e. they have not been used during the cross-validation training of both binary RF classifiers. The input frame-rate is 120 fps for all test sequences and their duration ranges between 9 and 13 seconds. Source content with a 3840x2160 original resolution have been downsampled to the 1920x1080 resolution with Lanczos3 filers [39] to ensure consistency during the subjective test.

The test set sequences have been selected from various sources to cover a wide range of spatio-temporal characteristics both in terms of temporal and spatial information (SI and TI), as recommenced in [42] and shown in Fig. 8. They also depict several use-cases, including sporting events and movie-type clips, in addition to the more common natural video content.

In order to generate the final VFR model, both RF classifiers have been retrained with their respective whole datasets as well as their feature sets and hyper-parameters determined via cross-validation. The prediction results for the test set are depicted in Fig. 9. As can be observed, correct prediction rates reach 92%, 77% and 84% for the *30fps*, *60fps* and *120fps* classes, respectively, showing the capacity of the model to generalize the VFR classification problem to unknown data. The slightly better prediction results for the test set compared to the cross-validation predictions presented in Section V-C may be explained by the more accurate ground truth labels obtained for the test set, thus minimizing the labeling issue previously raised. In addition, the low amount of samples falsely predicted with a *30fps* class label should lead to a good perceptual quality of the VFR output videos, very close to the HFR source content.

To verify this statement, a subjective test comparing the uncompressed HFR and VFR videos has been designed. The *30fps* and *60fps* versions, obtained by frame decimation, have also been introduced in the subjective evaluation to assess the interest of variable frame-rate compared to systematic temporal downsampling in terms of perceived quality.

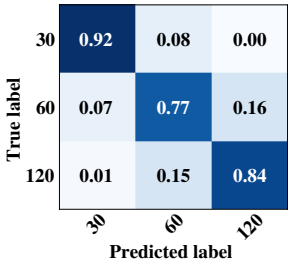


Fig. 9. Cascaded RF model prediction performance on test set.

B. Subjective Evaluation Methodology

The considered subjective evaluation aims at assessing the effect of a system, here the VFR coding scheme, on the visual quality. For this kind of test, the ITU-R BT.500-13 recommendation [40] proposes the Double Stimulus Continuous Quality Scale (DSCQS) method, which consists in showing the observer pairs of videos - the un-processed source content and the same sequence processed with the system under test - and asking the observer to rate the quality of both sequences. The grading scale is a continuous vertical scale divided into 5 equal parts corresponding to the common 5-level ITU-R quality labels: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*.

For each test session, a series of video pairs is presented to the observer in a random order, to distribute the degrees of quality impairments over the entire session. Each pair of videos is internally random, i.e. the observer is not aware of the position of the reference un-processed video (A or B), which is presented twice, successively. Fig. 10 depicts the structure of a BTC presenting a pair of videos to assess. As can be observed, each BTC begins with a 2-second message indicating the id number of the current test point and ends with a message asking to vote. In addition, each display of a 10-second sequence is preceded by a 1-second message indicating if the following video is A or B on the answer sheet, making the total duration of a BTC equal to 50 seconds.

A total number of 10 sequences have been selected within the test set for the subjective test, as indicated in Fig. 8. The sequence set has been formed to cover a wide range of spatio-temporal characteristics and content types. For each sequence, 4 frame-rate pairs have been evaluated by the observers: *120fps vs 120fps*, *120fps vs VFR*, *120fps vs 60fps* and *120fps vs 30fps*. Therefore, a total of 40 BTCs were presented to each observer, randomly divided into two 20-minute sessions separated by a 10-minute break. VFR video sequences have been obtained using the predicted frame-rates resulting from the proposed VFR model. Fig. 11 depicts the evolution of frame-rate decisions over the duration of each sequence of the subjective evaluation test set. As can be observed, the predicted frame-rates are highly dependent on the test sequence, as expected considering the wide range of spatial and temporal information characteristics for the selected sequence set. In addition, the predicted frame-rates also vary over-time for most of the test sequences, demonstrating the interest of the 4-frame level of granularity proposed for frame-rate decisions.

The test was conducted in a controlled laboratory environment, with a viewing distance fixed to 3 times the screen height. A 65-inch LG OLED B6 display with HFR capabilities and peak luminance of 340 cd/m^2 has been used for both subjective tests. During the whole duration of the tests, all internal post-processing were disabled to avoid any impact on the perceived quality. Each test sequence in raw format (YUV 4:2:0 and 8-bit precision) has been encoded using the *libx265* encoder at 100 Mbps in order to be presented to the TV set via USB3 interface. Special care has been taken to ensure that the encoding needed for display did not introduce any 'coding' artifacts. A total of 19 participants took part in the subjective test. They were aged between 20 and 53 with (corrected-to) normal vision acuity and color vision. A post-screening analysis of the results has been carried out, according to the method described in ITU-R Rec. BT.500-13, to detect and reject the outliers before computing the Mean Opinion Score (MOS) values.

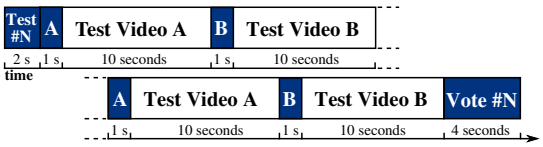


Fig. 10. Subjective test BTC structure for DSCQS evaluation method.

C. Subjective Visual Quality Results

Fig. 12 shows the results of the subjective test carried out to demonstrate the interest of variable frame-rate and evaluate the perceived quality of the proposed VFR model output. For each sequence of the test set previously presented, the Differential Mean Opinion Score (DMOS) values, computed using Equation (9), of each tested frame-rate are depicted together with their associated 95% Confidence Intervals (CIs). Since none of the participants were flagged as outliers after the post-screening analysis, the presented DMOS values have been obtained using the results from the 19 participants.

$$DMOS_f(s) = 100 - \frac{1}{N}\sum_{n=1}^{N} S_{n,120fps}(s) - S_{n,f}(s), \quad (9)$$

with N the total number of valid participants, $N = 19$ in this test, $DMOS_f(i)$ the DMOS value for sequence s at the tested frame-rate f, $f \in \{120fps, VFR, 60fps, 30fps\}$. The pair $(S_{n,120fps}(s), S_{n,f}(s))$ represents the scores attributed to sequence s at respectively the hidden reference *120fps* frame-rate and tested frame-rate f, i.e. both videos of a given BTC, by the n^{th} participant, $n \in \{1, .., N\}$.

The first statement that can be made by analyzing the results of the subjective test is that, as previously stated, the benefit brought by a frame-rate of 120 images per second compared to lower frame-rates is highly content-dependent. Indeed, for the sequences *Rugby7*, *library* and *Rugby6*, there is a significant difference between the DMOS values associated to the *120fps* frame-rate and those of the *60fps* and *30fps* frame-rates. The same trend can be observed for the *bouncyball*, *Rowing2* and *martials_arts* sequences. However, for these sequences, the CIs of the *120fps* and *60fps* DMOS are overlapping, thus a significant difference between the perceived quality of the

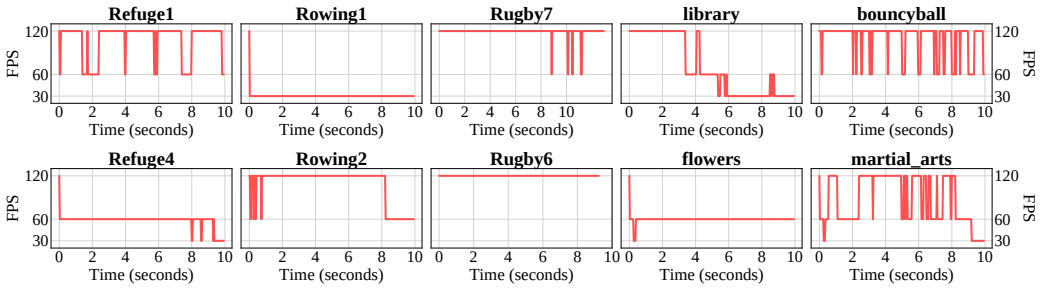


Fig. 11. Frame-rate decisions of the VFR algorithm for test set sequences.

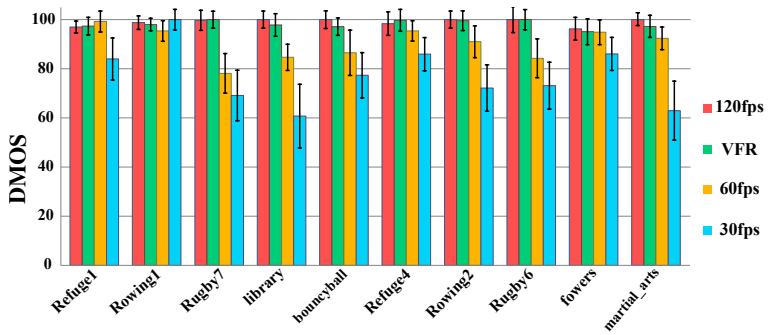


Fig. 12. Mean Opinion Score values with 95% confidence intervals for test set sequences and subjectively tested frame-rates.

two frame-rates cannot be confidently guaranteed for these sequences. For other sequences, namely *Refuge1*, *Refuge4* and *flowers*, the perceived quality of the *120fps* and *60fps* seem equivalent, with similar DMOS and highly overlapping CIs. Finally, the sequence *Rowing1* shows no visual difference even with a frame decimation down to *30fps*.

Comparing the perceived qualities of the VFR model outputs with their source HFR 120 fps counterpart, the DMOS values of both configurations appear to be equivalent for every sequence. This trend highlights the interest of variable frame-rate with its capacity to adapt to the quantity of movement possibly varying over time. For instance, the *library* sequence opens on a camera panning with a gradually slowing speed which then stops at the middle of the sequence on a stationary top spinning at high speed. The first part of the video requires 120 fps to correctly portray the camera panning, while lower frame-rates can be used without introducing artifacts as the speed of the camera gradually drops. For this sequence, participants attributed significantly lower scores to the *60fps* and *30fps* frame-rates due to the important motion artifacts present in the first part of the video at these frame-rates. On the contrary, the VFR model correctly lowers the frame-rate when the content permits it, resulting in a score identical to the one attributed to the source HFR video. However, despite highly correlated DMOS values and overlapping CIs, there is still a chance that the perceived qualities of the compared frame-rates are actually different.

To confirm these observations and confidently state that the VFR model output perceived quality is the same as for the source HFR content, a more rigorous analysis can be performed using a two-sample unequal variance Student's t-test with a two-tailed distribution (also called Welsch's t-test). This test allows to determine if indeed the perceived qualities given by the MOS values of each pair of tested frame-rates are "significantly" different or not. In this case, the null hypothesis, H_0, would be that the tested frame-rate f_{test} has the same perceived quality as the considered reference frame-rate f_{ref}. The alternate hypothesis, H_a, would be that there is a difference between the perceived qualities of f_{test} and f_{ref}. In order to test the similarity for each possible pair of frame-rates, the possible values for both frame-rates are: $f_{test} \in \{VFR, 60fps, 30fps\}$ and $f_{ref} \in \{120fps, VFR, 60fps\}$.

First, considering the sample populations from the scores attributed to a sequence s at the two frame-rates f_{test} and f_{ref} being compared, the t-statistic $t_{f_{test},f_{ref}}(s)$ can be used, expressed as follows

$$t_{f_{test},f_{ref}}(s) = \frac{\overline{S}_{f_{test}}(s) - \overline{S}_{f_{ref}}(s)}{\sqrt{\frac{\sigma^2_{f_{test}}(s)}{N_{f_{test}}} + \frac{\sigma^2_{f_{ref}}(s)}{N_{f_{ref}}}}}, \tag{10}$$

with $\overline{S}_{f_i}(s)$, $\sigma^2_{f_i}(s)$, N_{f_i} the sample mean, sample variance and sample population size for frame-rate f_i, $i \in \{test, ref\}$. In this test, $N_{f_{test}} = N_{f_{ref}} = N$, the number of observers that took part in the subjective test.

Then, by approximating the t-statistic with a Student's t-distribution, a value p, which indicates the degree of correlation between the means of the two sample populations, can be computed from the t-statistic. The higher the *p-value* is, the more significant the similarity between the distributions of the two populations is. A *p-value* lower than 0.05 indicates that there is statistical significance that the tested frame-rate f_{test} has a different perceived quality compared to the considered reference frame-rate f_{ref}. Indeed, in this case, there is a low probability of committing a type-I error, i.e. rejecting the null hypothesis when it is true, meaning that the null hypothesis

FPS	120	VFR	60
VFR	**0.77**		
60	**0.26**	**0.41**	
30	**0.02**	**0.03**	**0.00**

(a) Refuge1

FPS	120	VFR	60
VFR	**0.96**		
60	**0.14**	**0.14**	
30	**0.42**	**0.44**	**0.12**

(b) Rowing1

FPS	120	VFR	60
VFR	**0.56**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.01**

(c) Rugby7

FPS	120	VFR	60
VFR	**0.56**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.00**

(d) library

FPS	120	VFR	60
VFR	**0.41**		
60	**0.01**	**0.00**	
30	**0.00**	**0.00**	**0.01**

(e) bouncyball

FPS	120	VFR	60
VFR	**0.85**		
60	**0.27**	**0.27**	
30	**0.02**	**0.01**	**0.03**

(f) Refuge4

FPS	120	VFR	60
VFR	**0.45**		
60	**0.02**	**0.01**	
30	**0.00**	**0.00**	**0.00**

(g) Rowing2

FPS	120	VFR	60
VFR	**0.26**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.00**

(h) Rugby6

FPS	120	VFR	60
VFR	**0.99**		
60	**0.56**	**0.58**	
30	**0.00**	**0.00**	**0.01**

(i) flowers

FPS	120	VFR	60
VFR	**0.13**		
60	**0.00**	**0.03**	
30	**0.00**	**0.00**	**0.00**

(j) martial_arts

Fig. 13. *p-value* probabilities resulting from two-sample unequal variance bilateral Student's t-test on MOS values for each pair of tested frame-rates and each test set sequence. $p \geq 0.05$ *(green) means there is no significant difference between the MOS value of the row and column frame-rate labels while* $p < 0.05$ *(red) indicates that the MOS value of the row frame-rate label is significantly lower than the MOS value of the column frame-rate label.*

can be confidently rejected. On the contrary, if the *p-value* is greater than or equal to 0.05, the null hypothesis cannot be safely rejected and both frame-rates can be considered to have the same perceived quality.

Finally, the *p-value* does not give information on the probability of committing a type-II error, i.e. a failure to reject the null hypothesis when the alternate hypothesis is true, which is thus still a possibility. To ensure a low type-II error probability, and thus a statistically powerful test, the power β of the statistical test must be lower than 0.2. The power β has been computed for each possible pair of tested and reference frame-rates, resulting in an average β value of 0.044, showing that there is a lower than 5% chance, on average, to commit a type-II error. Therefore, the similarity assessment for each pair of possible frame-rates can be only based on the *p-values* resulting from the Student's t-test.

Fig. 13 depicts the *p-values* computed for each sequence and each possible frame-rate combination. Green-colored cells show the frame-rate pairs for which the associated *p-value* is greater than 0.05. Since every *VFR* vs *120fps* comparison falls within this category, it can be confidently concluded that the perceived quality of the VFR model output video is always the same as the original 120 fps frame-rate. This confirms that the under-estimated frame-rate predictions, identified in the confusion matrix depicted in Fig. 9, do not impact the perceived quality of the VFR videos. This also tends to validate the hypothesis made while analyzing training errors, stating that the ground truth is imperfect due to the coarse-grained nature of the ground truth annotations. Indeed, with its fine-grained decisions, the VFR model is capable of capturing smaller variations of critical frame-rates, thus resulting in predictions different from the ground truth, which are identified as prediction errors.

D. Compression Efficiency and Complexity Reduction

In order to evaluate the impact of VFR on coding performance, both the source HFR videos and VFR model outputs have been encoded using the HEVC reference software encoder HM16.12 [43]. The encoder was configured to use the HEVC Common Test Conditions (CTC) in Random Access (RA) configuration with a GOP size of 16 pictures and an intra-period of approximately 1 second to match with the considered broadcasting use-case. The quantization parameter was set to $QP = \{22, 27, 32, 37\}$ to cover a wide range of bit-rates and applications.

For the VFR encodings, the HEVC reference software encoder has been modified to handle the critical frame-rate decision coming from the proposed VFR module for each chunk of 4 input frames. Fig. 14 depicts the GOP structures needed to be supported in order to encode a VFR video sequence. Thanks to the built-in support of temporal scalability, removing the frames from upper Temporal Layers (TLs) does not break the coding dependencies. The VFR GOP structure is thus enabled in the reference software by simply skipping frames within the core encoding loop depending on the given frame-rate decision fed to the encoder and the current Picture Order Count (POC). This results in a VFR encoding with a lower bit-rate and reduced coding complexity while producing a bitstream decodable by the reference software decoder without any modification.

The bit-rate savings presented in this performance evaluation assume that, with the same QP, the perceived quality of the VFR decoded video is the same as the decoded original *120fps* video, as was demonstrated for the VFR and *120fps* uncompressed inputs. Indeed, on one hand, any frame decoded from the VFR bitstream is exactly the same as its corresponding frame in the *120fps* decoded video due to the use of identical GOP structures. On the other hand, the validity of the VFR frame-rate decisions on compressed data has been verified through an expert subjective test. This test aimed at evaluating the visual quality of the compressed *120fps* and VFR sequences, independently, for a subset of the VFR test set. The protocol used is the standard Degradation Category Rating (DCR) method [42] with an 11-grade scale. This subjective test resulted in the same MOS values for both the *120fps* and VFR decoded sequences, at every bit-rate tested, showing that the critical frame-rate decisions obtained with the VFR model trained on uncompressed data remain valid for compressed content. The subjective test also showed that the frame-rate could be further reduced in some cases due to the removal of details, by the encoding process, that justified a higher frame-rate in the original uncompressed video. The coding performance of the proposed model could thus be slightly improved by training it on compressed data. However,

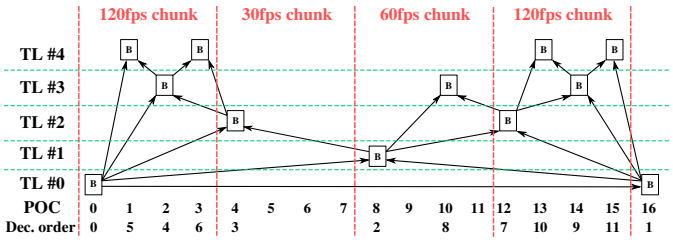


Fig. 14. Example of GOP structures of size 16 for a) source HFR 120 fps content and b) VFR with different frame-rates for each 4-frame chunk.

such an improvement would require the annotation of the entire database at several QPs, which would be a very time consuming task for a small coding gain. The encoding results presented in this work are thus obtained using the model trained on uncompressed videos.

Table II summarizes the performance results for the VFR encodings compared to regular *120fps* HEVC encodings, in terms of both bit-rate savings and encoding complexity reduction for the 15 sequences of the objective evaluation test set. The proportion of frames dropped by the VFR coding scheme are also added for information. Results for both the VFR model and ground truth decisions are presented to compare the performance at two different levels of granularity.

With the VFR model decisions, the VFR coding scheme offers 4.3% bit-rate savings on average, ranging from 0% to 15.4% for sequences where 120 and 30 frames per second are chosen for the whole sequence, respectively. For sequences with mostly *60fps* chosen or with temporally varying decisions, the bit-rate savings are generally around 4%. These bit-rate savings are not equal to the proportion of frames dropped by the VFR model due to the significantly lower amount of bits used to encode the frames of the upper TLs. Indeed, upper TL frames are coded using higher quantization steps and greatly benefit from the inter-picture predictions of the RA coding configuration. Thus, the amount of transmitted quantized residuals is lower for these frames, especially if the motion is easily predictable and if high spatial details are not present in the source content. For the complexity reduction brought by the VFR coding scheme, the results are close to the amount of frames dropped, with an average encoding complexity reduction of 28%, ranging from 0% to 70%. The per-sequence results follows the same trend as for bit-rate savings but with higher gain variations. The difference between the complexity reduction and frames dropped results mainly comes from the slightly reduced coding complexity of the upper TL frames compared to the kept frames of the lower TLs. Indeed, a higher number of residual coefficient to binarize and process with the entropy coding engine increases the encoding time. For the decoding complexity, the detailed results are not presented in this paper but the observed gains are highly similar to those observed for the encoding side, with an average decoding complexity reduction of 27.4% for the VFR coding scheme.

With the ground truth annotated frame-rates, the bit-rate savings and complexity reduction results are very close to the performance with the predicted frame-rates. This can be explained by the high correct prediction rate of the VFR model on the test set. The results are only significantly different for some sequences. *Refuge1* shows lower gains for the VFR model output due to the over-estimation of the required frame-rate, i.e. alternation between *120fps* and *60fps* prediction while the annotated ground truth frame-rate is *60fps* for the major part of the sequence. The opposite situation can be observed for the sequences *bouncyball*, *martial_arts* and *Katana*, where the VFR model allows for lower frame-rates more frequently than the ground truth, thus resulting in higher gains.

TABLE II
VFR HEVC ENCODING PERFORMANCE COMPARED TO 120FPS HEVC ENCODINGS FOR VFR PREDICTED LABELS (MODEL) AND GROUND TRUTH (G-T) LABELS ON THE TEST SET.

Sequence	bit-rate savings		Enc. Time Reduction		Frames Dropped	
	Model	G-T	Model	G-T	Model	G-T
Refuge1	-1.1 %	-4.9 %	7.8 %	39 %	10 %	50 %
Rowing1	-9.3 %	-9.3 %	60 %	60 %	75 %	75 %
Rugby7	-0.1 %	0.0 %	0.6 %	0.0 %	0.9 %	0.0 %
library	-5.0 %	-4.8 %	39 %	37 %	42 %	41 %
bouncyball	-2.5 %	0.0 %	6.7 %	0.0 %	8.7 %	0.0 %
Refuge4	-3.2 %	-3.5 %	47 %	49 %	52 %	55 %
Rowing2	-0.6 %	-0.4 %	6.1 %	5.4 %	9.7 %	8.7 %
Rugby6	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
flowers	-4.1 %	-4.1 %	41 %	40 %	50 %	50 %
martial_arts	-4.0 %	-0.5 %	23 %	5.9 %	28 %	7.6 %
Katana	-5.9 %	-1.2 %	35 %	11 %	38 %	13 %
NYCBike	-6.2 %	-5.6 %	27 %	25 %	32 %	29 %
pour	-1.6 %	-1.6 %	11 %	11 %	16 %	15 %
Refuge2	-15 %	-14.6 %	70 %	68 %	74 %	71 %
Refuge3	-5.8 %	-5.6 %	53 %	52 %	58 %	56 %
Average	**-4.3 %**	**-3.7 %**	**28 %**	**27 %**	**33 %**	**32 %**

VII. CONCLUSION

In this paper, a new variable frame-rate coding scheme is proposed for broadcast delivery of HFR (120 fps) contents. The proposed scheme incorporates a machine learning based VFR model capable of dynamically adapting the frame-rate of the video before encoding and transmitting it to the end receiver.

The VFR model relies on several spatio-temporal features extracted from each frame of the input video to predict the optimal lowest artifact-free frame-rate through two cascaded binary RF trained classifiers. The considered frame-rate adaptation is performed dynamically by choosing for each chunk of 4 consecutive input frames its associated critical frame-rate, among the three possible values: *30fps*, *60fps* or *120fps*. The model achieves an average critical frame-rate correct prediction rate of 84%, while keeping the frame-rate under-estimations error rate below 8%. The visual quality of the generated VFR videos has been carefully evaluated through formal subjective tests showing an identical perceived quality compared to the source HFR content.

From a coding performance perspective, the proposed VFR coding scheme provides average bit-rate savings of 4.3% in addition to average complexity reductions of 28% and 27.4% at the encoding and decoding sides, respectively. It should be noted that this work can be applied to other broadcast frame rates such us 25 and 50 and 100 fps by adopting the proposed algorithm.

The work proposed in this paper has been shown, at both the International Broadcasting Convention (IBC) 2019 and National Association of Broadcasters (NAB) Show 2019, through a real-time demonstration with both the input legacy HFR and processed VFR videos displayed synchronously on two HFR screens to demonstrate the equivalence in perceived quality. The demonstration includes a real-time software implementation of the VFR prediction model - 7.2 ms (138 fps) average runtime for feature computation (6.9 ms) and frame-rate prediction (0.3 ms) of HD sequences processed on a common consumer CPU.

The proposed solution is a practical candidate to lower the requirements for the broadcast delivery of the upcoming HFR services of the DVB UHD second deployment phase. Additionally, thanks to the its hardware-friendly solution (feature computation based on well-known h.264 encoding and existing RF hardware implementations [44]), the proposed VFR method can be considered by hardware video encoder manufacturers to enhance the quality of experience and reduce the energy fingerprint of their devices.

ACKNOWLEDGMENT

The authors would like to thank Franck Chi, Maxime Peralta and Clément Brossard who also contributed to this project.

REFERENCES

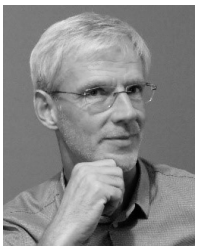
[1] ITU-R, "Recommendation BT.2020-1: Parameters Values of UHDTV Systems for Production and International Programme Exchange."
[2] ——, "Recommendation BT.709-5: Parameter Values for the HDTV Standards for Production and International Programme Exchange."
[3] M. Nilsson, "Ultra high definition video formats and standardisation," *BT Media and Broadcast Research Paper*, 2015.
[4] M. Sugawara and K. Masaoka, "UHDTV Image Format for Better Visual Experience," *Proceedings of the IEEE*, vol. 101, no. 1, pp. 8–17, 2013.
[5] A. Mackin, K. C. Noland, and D. R. Bull, "High frame rates and the visibility of motion artifacts," *SMPTE Motion Imaging Journal*, vol. 126, no. 5, pp. 41–51, 2017.
[6] Y. Kuroki, T. Nishi, S. Kobayashi, H. Oyaizu, and S. Yoshimura, "A psychophysical study of improvements in motion-image quality by using high frame rates," *Journal of the Society for Information Display*, 2007.
[7] K. Noland, "The application of sampling theory to television frame rate requirements," *BBC R&D White Paper*, vol. 282, 2014.
[8] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly, "Spatio-velocity csf as a function of retinal velocity using unstabilized stimuli," in *Human Vision and Electronic Imaging XI*, vol. 6057. International Society for Optics and Photonics, 2006, p. 605705.
[9] V. Hulusic, G. Valenzise, J.-C. Gicquel, J. Fournier, and F. Dufaux, "Quality of experience in uhd-1 phase 2 television: the contribution of uhd+ hfr technology," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*. IEEE, 2017, pp. 1–6.
[10] A. Mackin, F. Zhang, M. A. Papadopoulos, and D. Bull, "Investigating the impact of high frame rates on video compression," in *Image Processing (ICIP), IEEE International Conference on*. IEEE, 2017.
[11] R. Salmon, T. Borer, M. Pindoria, M. Price, and A. Sheikh, "Higher frame rates for television," *IBC Conference 2013*, 2013.
[12] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3407–3411.
[13] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, "Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 671–682, 2012.
[14] Q. Huang, S. Y. Jeong, S. Yang, D. Zhang, S. Hu, H. Y. Kim, J. S. Choi, and C.-C. J. Kuo, "Perceptual quality driven frame-rate selection (pqd-frs) for high-frame-rate video," *IEEE Transactions on Broadcasting*, vol. 62, no. 3, pp. 640–653, 2016.
[15] A. V. Katsenou, D. Ma, and D. R. Bull, "Perceptually aligned frame rate selection using spatio temporal features," in *Picture Coding Symposium (PCS), 2018*. IEEE, 2018, pp. 1–5.
[16] M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, 2018.
[17] *Advanced Television Systems Committee (ATSC) Standard*.
[18] "Specification for the use of video and audio coding in broadcast and broadband applications," *DVB, ETSI TS 101 154 V2.4.1*, 2000.
[19] "The present state of ultra-high definition television," *ITU-R Report BT.2246-6*, March 2017.
[20] M. Emoto and M. Sugawara, "Critical fusion frequency for bright and wide field-of-view image display," *Journal of Display Technology*, vol. 8, no. 7, pp. 424–429, 2012.
[21] R. Salmon, M. Armstrong, and S. Jolly, "Higher frame rates for more immersive video and television," *BBC White Paper WHP*, vol. 209, 2011.
[22] P. G. Barten, *Contrast sensitivity of the human eye and its effects on image quality*. Spie optical engineering press Bellingham, WA, 1999.
[23] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision Models and Applications to Image and Video Processing*. Springer, 2001, pp. 179–200.
[24] R. Selfridge, K. C. Noland, and M. Hansard, "Visibility of motion blur and strobing artefacts in video at 100 frames per second," in *European Conference on Visual Media Production (CVMP 2016)*. ACM, 2016.
[25] M. Emoto, Y. Kusakabe, and M. Sugawara, "High-frame-rate motion picture quality and its independence of viewing distance," *Journal of Display Technology*, vol. 10, no. 8, pp. 635–641, 2014.
[26] EBU, "Ebu policy statement on ultra high definition television," in *European Broadcasting Union, Grand-Saconnex, Switzerland.*
[27] Y. Sugito, S. Iwasaki, K. Chida, K. Iguchi, K. Kanda, X. Lei, H. Miyoshi, and K. Kazui, "A study on the required video bit-rate for 8k 120hz hevc temporal scalable coding," in *Picture Coding Symposium*. IEEE, 2018.
[28] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 300–304.
[29] F. Navarro, F. J. Serón, and D. Gutierrez, "Motion blur rendering: State of the art," in *Computer Graphics Forum*. Wiley Online Library, 2011.
[30] T. Brooks and J. T. Barron, "Learning to synthesize motion blur," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6840–6848.
[31] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.
[32] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, pp. 115–137, 2014.
[33] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE conference on computer vision and pattern recognition*, 2017.
[34] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 670–679.
[35] M. Park, H. G. Kim, S. Lee, and Y. M. Ro, "Robust video frame interpolation with exceptional motion map," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.
[36] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, 2001.
[37] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," 1984.
[38] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, 1996.
[39] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of applied meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.
[40] ITU-R, "Recommendation BT.500-13:Methodology for the Subjective Assessment of the Quality of Television Pictures."
[41] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th conference on Message understanding*. Association for Computational Linguistics, 1992, pp. 22–29.
[42] "Subjective video quality assessment methods for multimedia applications," *ITU-T Rec. P.910*, April 2008.
[43] "HEVC reference software version 16.12." [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.12/
[44] B. Van Essen, C. Macaraeg, M. Gokhale, and R. Prenger, "Accelerating a random forest classifier: Multi-core, gp-gpu, or fpga?" in *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2012, pp. 232–239.

Glenn Herrou received the Dipl.-Ing. (M.Sc.) degree in electrical and computer engineering and PhD in signal processing from the Institut National des Sciences Appliquées (INSA) de Rennes, France, in 2016 and 2019. From 2016 to 2019, he worked at the Institute of Research and Technology b<>com, Cesson-Sévigné, France, on projects focusing on adaptive spatio-temporal resolution for efficient video coding. Since 2020, he is a post-doctoral researcher in the VAADER team of the Institut d'Electronique et des Technologies du NuméRique (IETR), Rennes, France. His current research interests focus on video coding and applied Machine/Deep Learning.

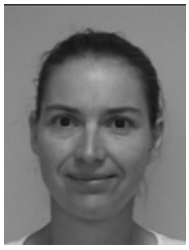
Wassim Hamidouche received Master's and Ph.D. degrees both in Image Processing from the University of Poitiers (France) in 2007 and 2010, respectively. From 2011 to 2013, he was a junior scientist in the video coding team of Canon Research Center in Rennes (France). He was a post-doctoral researcher from Apr. 2013 to Aug. 2015 with VAADER team of IETR where he worked under collaborative project on HEVC video standardisation. Since Sept. 2015 he is an Associate Professor at INSA Rennes and a member of the VAADER team of IETR Lab. He has joined the Advanced Media Content Lab of b¡¿com IRT Research Institute as an academic member in Sept. 2017. His research interests focus on video coding and security of multimedia contents. He is the author/coauthor of more than one hundred and twenty (+120) papers at top journals and conferences in Image Processing, two MPEG standards, two patents, several MPEG contributions, public datasets and open source software projects.

Jérôme Fournier received the Ph.D. in signal and image processing from the University of Rennes, France, in 1995. He started his career at Philips in the field of video communications. In 1997, he joined Orange Labs (formerly France Telecom) and worked on video codecs like MPEG-4 Part 2 and H.264. From 2004 to 2012, Jérôme focused on the deployment of the Orange TV services, HDTV and stereoscopic 3DTV, as well as on innovative 3DTV depth-based video formats. From 2012 to 2018, he was mainly involved in the subjective evaluation and the ITU-R standardization of Ultra HD video formats including HDR and HFR features. Now, he is contributing to b<>com studies on topics like VFR, view synthesis and 8K.

Charles Bonnineau received the Dipl.-Ing. degree in Computer Science at the Ecole Supérieure D'Ingénieurs de Rennes (ESIR) from the Université de Rennes 1, France, in 2018. He is currently a PhD Student in Signal and Image Processing jointly with the Institut d'Electronique et des Technologies du NuméRique (IETR), the Intitute of Research and Technology b<>com, and TDF. His current research include video processing and coding using deep-learning-based methods.



Patrick Dumenil received his Master's degree in Signal Processing from CentraleSupelec, Gif-sur-Yvette (France), in 1987. He started his career as a research engineer at Thomson Laboratoires Electronique, Rennes (France), where he developed new technologies for digital television. Then, he managed projects focused on designing first-generation digital television products and contributing to MPEG-2 and AVC standards. Since 2016, he leads the Codec Innovation team at Harmonic, Cesson-Sévigné (France), and is also involved in projects at the b<>com research institute. His current research focuses on developing Machine Learning solutions for codec efficiency improvement.



Luce Morin is a full professor at National Institute of Applied Sciences (INSA), Rennes, France, where she teaches computer science, image processing, and computer vision. She received the M.S. degree from ENSPS school in Strasbourg in 1989 and spent a 6 month internship at the NASA GSFC in Washington D.C. She then prepared a Ph.D. thesis supervised by professor Roger Mohr in the LIFIA laboratory, INP-Grenoble, on projective invariants applied to computer vision. From 1993 to 2008, she was an associate-professor at University of Rennes and a member of the Temics team in the IRISA/INRIA-Rennes laboratory. She is now a member of the Institut d'Electronique et des Technologies du NuméRique (IETR) laboratory, where she leads the VAADER research team. Her research activities deal with computer vision, 3D reconstruction, image and video compression, and representations for 3D videos and multiview videos.