



AI BASED SALIENCY-AWARE VIDEO CODING

S. Pelurson, J. Cozanet, T. Guionnet, M. Abdoli, T. Biatek

ATEME, France

ABSTRACT

The demand for video through OTT has been constantly increasing over the years. With the COVID situation, this demand is skyrocketing, hence the need for better video compression. The human visual system (HVS) can quickly select visually important regions in its visual field. Those regions are captured at high resolution, while other peripheral regions receive little attention. Saliency maps are a way to imitate the HVS attention mechanism. Recently, deep learning-based saliency models have achieved tremendous improvements. This paper aims at leveraging state-of-the-art deep learning-based saliency models to improve video coding efficiency. First, a saliency-based rate control scheme is integrated in an HEVC video encoder. Then, a saliency guided preprocessing filtering step is introduced. Finally, the two approaches are combined. Objective and subjective evaluations show that it allows lowering the bitrate from 6% to almost 30% while maintaining the same visual quality.

INTRODUCTION

Video demand on the internet has grown significantly in the past years, representing as of today 80% of all traffic while representing only 67% in 2016 as described by CISCO (1) report. With COVID situation, the demand for video services skyrockets, leading operators to degrade video quality to avoid network congestion and services interruptions. Although 5G deployment has started and Fiber-to-the-home (FTTH) is becoming more and more prevalent, the network congestion still remains a serious issue for operators.

Several solutions exist to address this issue using emerging technologies. On one hand, the optimization of video delivery, switching from unicast to multicast for high audience services, is a solution. This can be achieved, for example, by using recently standardized DVB-MABR delivery DVB (2). On the other hand, the source coding can be improved to further reduce the transported video files size, for instance by using VVC ITU-T (3) or AV1 AOM (4). Although these solutions address the bandwidth issue, their deployment might cause problems for operators. First, the investment cost required to upgrade and replace network equipment to support multicast is high. Second, the installed parc of receivers is not compatible with a new codec and would then not be compatible with upgraded video services.

Another solution to address this problem while maintaining compatibility with currently in-place ecosystem consists of improving coding efficiency of existing codecs. Recent encoders are using complex algorithms in order to achieve perceptually relevant compression performance. In order to push further the envelope, one must rely on even more advanced mechanisms, such as those proposed in this paper.



The human visual system (HVS) can quickly select visually important regions in its visual field. Thanks to the foveation mechanism, those regions are captured at high resolution, while other peripheral regions receive little attention at low resolution. This cognitive process enables humans to interpret complex scenes in real-time. Taking advantage of HVS is a challenge for encoder manufacturers, which try to mimic it to improve coding performance. The usage of saliency maps as a basis for making encoding decision is addressing this challenge but traditional approaches were not accurate enough to achieve substantial gains so far. Recently, deep learning-based models have achieved tremendous improvements, relying on the powerful end-to-end learning ability of neural networks and the availability of large-scale saliency datasets.

This paper aims at leveraging state-of-the-art AI-based saliency models to improve video coding efficiency. A saliency-based rate control scheme is proposed to modulate frame areas compression according to the predicted saliency map. In addition, the saliency data is used to guide pre-processing decision, improving video quality while lowering the bitrate. The efficiency of the proposed method is assessed using both objective and subjective criterions, reporting bitrate savings at a similar perceived video quality.

The rest of the paper is organized as follows. The first section provides an overview of related work on saliency estimation and its application to video encoding. The second section introduces the proposed method and how it is practically integrated in an industrial HEVC encoder. The third section reports experimental procedure and results. Eventually, the paper is concluded in the last section.

RELATED WORK

Visual attention modelling

Predicting the human visual attention is an intensive research area in cognitive sciences and computer vision since several decades. Automatically detecting and segmenting salient regions of an image or video has a lot of interest in many computer vision tasks. For example, Ma et al (6) and Lee et al (7) use it for visual tracking, Xu et al (8) for image captioning, Qin et al (9) and Fu et al (10) for image/video segmentation. In order to predict where people look in a scene, two kinds of models can be used: fixation prediction (FP) models, or salient object detection (SOD) models. The first ones aim to attend where an observer may fixate within few seconds of free viewing. The second ones detect and segment the entire extent of the salient regions/objects in the scene. Although promising results have been achieved with SOD models, they often fail to detect most salient objects in complex cluttered scenes containing several objects. Li et al (11) demonstrated that this is mostly due to a heavy bias in many widely used SOD datasets that have only a few obvious objects in the scene. Moreover, the human attention prior from FP models is more consistent with visual processing of human visual system because of its origin from cognition and psychology research community. FP are therefore more suited to the video coding use case.

Existing FP models can be categorized into two classes: conventional and deep learning-based solutions.

Conventional solutions

These kinds of solutions mainly rely on bottom-up mechanisms and cognitive assumptions about visual attention. They use biologically-inspired hand-crafted features such as color,



intensity or orientation to create feature maps that are then weighted and summed to create a saliency map. A saliency map is a gray-scale image in which the brightness of a pixel is proportional to its saliency. Kosh et al (12) proposed a theoretical model that has been the basis of many later models. When published, it was not yet implemented but provided the algorithmic reasoning for later implementations.

All these visual attention models can be categorized into different classes, such as cognitive models, Itti et al. (13), Bayesian models, Zhang et al. (14), decision theoretic models, Gao et al. (15), information theoretic models, Bruce et al. (16), graphical models, Harel et al. (17), pattern classification models, Judd, T. et al. (18), etc. Borji and Itti (19) present a detailed overview of these models. Judd et al (20) show that most of them cannot compete with a simple center-prior model, which limits their applicability to the video coding field. Thanks to the creation of large-scale eye tracking datasets, deep learning-based models have been proposed and generally offer better performances.

Deep learning-based solutions

Vig et al (21) was an early model that automatically learns deep representation of FP. It uses a set of shallow CNN to search optimal deep features and feed them to an SVM for fixation prediction. Later, deeper and more complex models were proposed for example by Kruthiventi et al (22), Huang et al (23) and Liu et al (24). All of them rely on deep CNN and use pretrained architectures such as VGG16 by Simonyan et al (25) to extract visual features used to predict fixation points.

Other models are based on autoencoder architectures to directly construct the saliency map such as works by Simonyan et al (26) and Pan et al (27). These models use similar feature extractor backbones but also learn to reconstruct the saliency map to its final size with convolution and upsampling layers while previous works has commonly used bilinear interpolation for that task.

All presented models have been designed for static visual saliency, but some works have focused on dynamic saliency models. Most of them rely on two-stream networks such as works done by Fang et al (28), Bak et al (29), or 3D CNN as Sun et al (30) did, that respectively extract visual and temporal features in parallel or simultaneously. While being much more computationally expensive than static models, it has been proven by Tangemann et al (31) that dynamic models failed to capture important temporal patterns for saliency prediction. This is mostly due to a deficiency of datasets used to train video saliency models.

Saliency-based video coding

Use of saliency data in video coding is a way to imitate the foveation mechanism of the human visual system (HVS) that capture visually important region at high resolution while other peripheral regions receive little attention at low resolution. So, it seems to be a good solution to reduce video bitrate without sacrificing quality of regions of interest. Different works use saliency data in video coding field, and two categories of works can be distinguished: implicit and explicit methods.

Implicit methods use saliency data before feeding the video to the encoder. For example, Itti et al (32) and Polakovič et al (33) apply non-uniform blur according to a saliency map to the input frame to force the desired bit allocation. They show that pre-processed videos have a better visual quality than the unprocessed ones at the same bitrate.



Explicit methods modify internal encoder data according to the saliency map to change bit allocation. The most commonly used strategy is to modify quantization parameter for each macroblock according to its mean saliency value as proposed by Gupta et al (34), Li et al (35), Zhu et al (36) and Lyudvichenko et al (37). While these works directly adapt bit allocation modifying quantization parameters, some works use saliency data to adapt computational resources. Wei et al (38) propose a fast CU mode decision using saliency as an approximation of texture complexity and movements.

While some works show quality improvement using saliency data (35)(36), others highlight bitrate reduction with equivalent perceived quality (32)(37)(38).

In this work, relying on a deep learning-based saliency model, we aim to leverage saliency data both in an implicit and explicit way. This is integrated in an optimized HEVC encoder, using realistic video delivery bitrates to match actual industrial use cases.

PROPOSED SOLUTION

Proposed architecture

A video encoder is usually not limited to the implementation of a core codec, such as HEVC. Many other parts are combined to maximize coding efficiency. The black part of Figure 1 summarizes such a system. The quality of pre-processing, pre-analysis and rate control play an equally important role as the encoding core regarding resulting video quality.

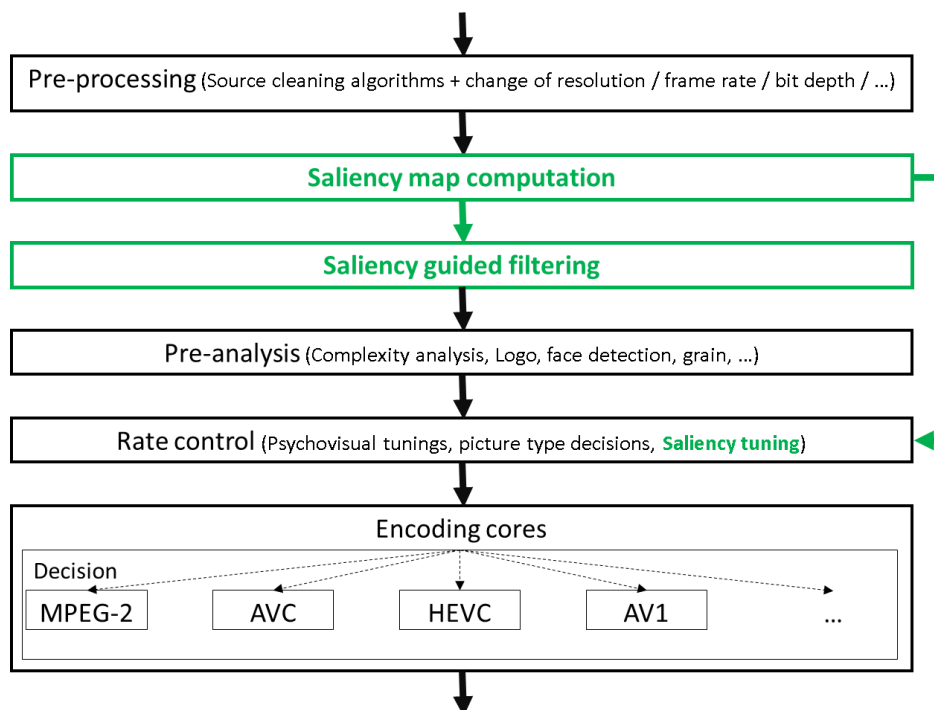


Figure 1: Proposed video encoder architecture, with saliency process indicated in green.

In this paper, it is proposed to add the saliency information to the coding structure, as depicted in green on Figure 1. A saliency map is computed using the neural network described by Kroner et al (27). It is an autoencoder characterized by a lightweight architecture combined with a multi-scale feature extractor. It has been selected because it achieves fixation prediction accuracy at the level of current state of the art approaches on several public datasets while being computationally less expensive. The network is used as



provided by the author and has not been retrained, so some of its drawbacks may potentially impact our results.

Each video frame of the processed video stream is downscaled to 320x240 pixels using a Lanczos kernel prior to being input to the neural network. The output of the network is a saliency map, which for each pixel provides a saliency score between 0 and 255. The saliency map is rescaled back to the original frame resolution through simple bilinear interpolation and transmitted to the other processing blocks. The saliency guided filtering block is a pre-processing block, implementing the implicit use of saliency data. The explicit use of saliency data is integrated in the rate-control process.

Implicit solution: pre-processing

Even though encoders are built in order to preserve the most useful information, and get rid of the less visible details, it has been demonstrated that it is possible to improve coding performance by filtering data from the video prior to encoding. It is a video pre-processing which can be referred to as simplification. One most recent example is the “GOP based filter” which has been proposed and implemented along with the VVC standardisation process. Latest results as the one by Wennersten et al (5) demonstrate a 5.5% coding efficiency gain on top of the already excellent VVC. Many strategies are possible for video simplification. As shown in (5), better performance is achieved when knowing the nature of compression. Thus, the filter is adapted temporally, depending on the encoding type of each frame. Rather than adapting a simplification filter temporally, it is proposed here to adapt it spatially, based on the information provided by salience analysis.

The goal here is to demonstrate how saliency analysis can be useful to video compression. The optimization of the simplification filter itself is out of the scope of this paper. Therefore, a simple low pass filtering strategy has been implemented. A set of short linear filters has been defined, as illustrated by Table 1. These filters are classical windowed sinus cardinal based low pass. The set is generated by varying the filter synthesis cut-off frequency. Each filter is characterized by the frequency bandwidth which is preserved, ranging from full bandwidth (no filtering) to 13% of the bandwidth (strong filtering effect). The main benefit of these filters is their low complexity, which make them easier to apply selectively at the pixel level.

Table 1: Integer digital low pass filters bank for video simplification.

Filter Cut-off Frequency	Filter coefficients						
100% bandwidth (identity)	0	0	0	1024	0	0	0
93% bandwidth	15	-31	48	960	48	-31	15
87% bandwidth	25	-57	94	900	94	-57	25
		...					
25% bandwidth	25	104	220	326	220	104	25
19% bandwidth	44	113	206	298	206	113	44
13% bandwidth	57	117	192	292	192	117	57

At this stage, the saliency of each pixel is known, and a set of filters is available. A simple linear mapping is defined between the saliency value and the filter index in the filter table. The maximal filtering is a parameter of the linear function.



Explicit solution: QP-adaptation

In HEVC, a quantization parameter (QP) is used to determine the quantization step. The higher the QP, the higher the quantization, and the lower the number of bits allocated. Thus, the main idea is to increase the QP on non-salient regions to reduce bitrate and decrease it on salient regions in order to preserve the subjective visual quality.

The QP is managed at macroblock level. So, based on the saliency map, we defined a saliency value for each macroblock, which is defined as the mean value of every pixels inside this macroblock. The saliency value is defined between 0 and 255, 255 corresponding to a very salient region. A rate control module uses this saliency value to adapt the QP for each macroblock. In order to do that, the macroblock saliency value is first normalized between 0 and 3 to define 4 saliency levels. Level 3 defines most salient areas while level 0 defines less salient ones. Then, a saliency factor is defined as below and is applied to the initial QP value.

$$SaliencyFactor = \begin{cases} a, \text{ where } a > 1.0 & \text{if saliencyLevel} = 3 \\ x, \text{ where } x < 1.0 & \text{if saliencyLevel} < 3 \end{cases}$$

a is a customizable constant value. The lower the saliency level, the lower the x parameter. Simply lowering the QP value proportionally to the saliency level tends to degrade the quality too much on low bitrate. That is why we defined the x parameter as follow, where b is a constant factor fixed for each saliency level, and $qualityFactor$ is a ratio between a specified quality index (integer between 0 and 255), and its mean value (128). This quality index is used by the encoder to tune the internal parameters and make the coding decisions. $qualityFactor$ is less than 1 for high quality encoded sources, and higher than 1 for low quality ones.

$$x = \begin{cases} b * qualityFactor, & \text{if } qualityFactor < 1.0 \\ b, & \text{if } qualityFactor \geq 1.0 \end{cases}$$

Moreover, we also considered the percentage of macroblocks with saliency level 3 over the total number of macroblocks inside a frame. Indeed, our hypothesis is that the larger this value, the more we can degrade the quality of non-salient macroblocks without losing subjective quality. Finally, the whole process summarizes as:

$$saliencyFactor = \begin{cases} a \text{ where } a > 1.0, & \text{if saliencyLevel} = 3 \\ \frac{b * qualityFactor}{saliencyPercent}, & \text{if saliencyLevel} < 3 \text{ and } qualityFactor < 1.0 \\ \frac{b}{saliencyPercent}, & \text{if saliencyLevel} < 3 \text{ and } qualityFactor \geq 1.0 \end{cases}$$

Combined solution

Our final solution simply combines the two previously presented solutions by feeding the output of the saliency-based pre-processor as the input of the saliency-based encoder. Doing that, generated saliency data is used twice in the encoding process. Saliency-based pre-processor highlights important areas of each frame in order to implicitly improve their global coding efficiency modifying encoding decisions. Moreover, the saliency-based encoder specifically considers saliency data in rate control module, which explicitly allocates more bits for areas of interest. The goal of this solution is to improve the effect of both individual solutions in order to reduce even more bitrate while keeping the same perceived quality.

EXPERIMENTS

In this section, the performance of the proposed method is evaluated under different test conditions. To this end, a reference encoder setting is selected as the basis. On top of this encoder, the proposed saliency-based solutions are implemented and used for compression of a set of test sequences. Both objective and subjective evaluations are conducted.

Encoder settings


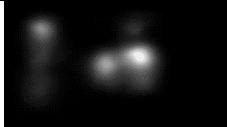

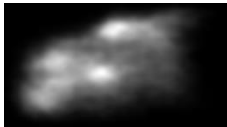

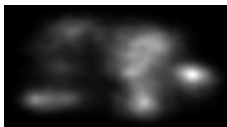

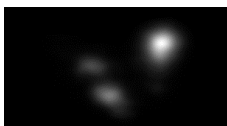

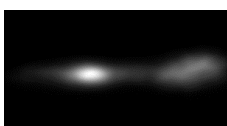
The base encoder for integration of the proposed methods is the Titan solution of ATEME, which contains different state-of-the-art codecs. For the experiments of this paper, the HEVC codec of the Titan has been used. Test sequences used in the experiments have been chosen to represent content diversity.

Table 2 Selected sequences for the performance evaluation of the proposed method summarizes the sequences that are used in our experiments, with snapshots of their first frame along with their corresponding saliency map. As can be seen, the resolution of sequences is limited to 1080p, as this is a frequently used resolution in broadcast and streaming applications.


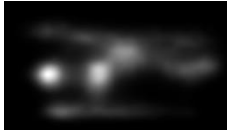



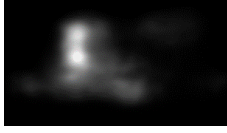



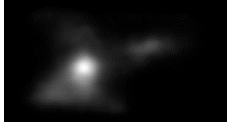
Four configurations are possible:

1. Baseline: Titan in default configuration
2. Baseline + implicit saliency (pre-processing)
3. Baseline + explicit saliency (encoder QP control)
4. Baseline + combined implicit and explicit saliency

Table 2 Selected sequences for the performance evaluation of the proposed method

Sequence	Signal description	First frame	Saliency map	Baseline bitrate
Basketball	1920x1080p 50Hz 8 bit			5523,5 Kb/s
BQTerrace	1920x1080p 60Hz 8 bit			2514,4 Kb/s
Cactus	1920x1080p 50Hz 8 bit			4153,5 Kb/s
Extrait	1920x1080p 25Hz 10 bit			644,2 Kb/s
Jaguar	1920x1080p 50Hz 8 bit			3984,8 Kb/s



PGM1	1920x1080p 30Hz 8 bit			4985,7 Kb/s
MarketPlace	1920x1080p 60Hz 10 bit			2287,7 Kb/s
RitualDance	1920x1080p 60Hz 10 bit			8158,8 Kb/s
Rugby	1920x1080p 50Hz 10 bit			4448,6 Kb/s
Explorer	1920x1080p 50Hz 10 bit			6298,7 Kb/s

Objective test

The objective performance of the proposed solutions has been assessed, using Peak Signal-to-Noise Ratio (PSNR) and Video Multimethod Assessment Fusion (VMAF) as described by Netflix (40). Figure 2 shows the Rate-Quality curves of the corresponding metrics for one of the test sequences. As can be seen, compared to the baseline method (configuration 1), a moderate performance reduction is caused by the proposed combined method (configuration 4). This is not surprising as the proposed saliency-based solutions basically disrupt the internal process of the encoder, which optimizes the objective-based quality. In other words, the abovementioned objective loss is the cost of subjectively optimizing the encoder in order to maintain a higher quality for salient areas that are supposedly more attractive. As a conclusion, objective evaluation is useful to ensure that no dramatic damage has been caused but is not a relevant quality assessment. Therefore, a set of subjective tests has also been conducted.

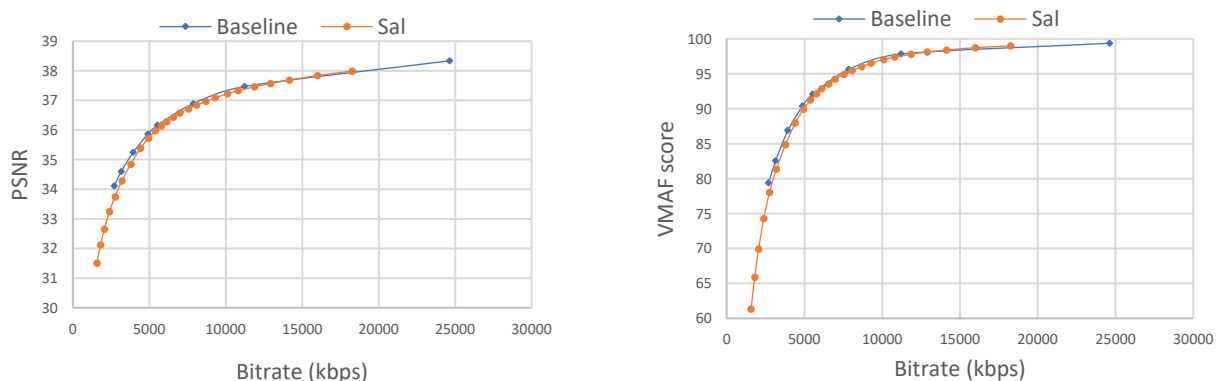


Figure 2 – Rate-Quality curves of BasketballDrive sequence, using PSNR and VMAF.



Subjective test

Subjective tests have first been designed and performed by video experts. Encodings have been carried out in the constant quality mode, where an integer input parameter between 0 and 255 defines desired quality level (the same parameter is used to control QP adaptation). The test sequences described in Table 2 have been encoded with all the configurations at five quality levels covering a target range of perceived quality levels and bitrates corresponding to conventional broadcast and streaming applications. These encodings allow the constitution of stream pairs suitable for visual comparison. Each pair is composed of the baseline and one of the other 3 configurations. The highest quality pairs have been discarded, as they cannot be discriminated visually. A selection of bitrates for the baselines is provided in Table 2.

Visual inspection of the selected pairs led to the following conclusions: All configurations provide a bitrate gain at equivalent visual quality. The degradation of the quality on the non-salient areas has been observed, as well as the improvement of important areas such as faces. The gain of the explicit method (configuration 3) is found to be higher than the gain of the implicit method (configuration 2). Finally, the gain of the combination (configuration 4) is found to be greater than the gain of the explicit method. In other words, the gains of the two methods are cumulative.

In summary, the interest of the approach has been demonstrated. Saliency information is useful to enhance video coding performance. Moreover, implicit and explicit approaches are cumulative. To further validate this result, a formal subjective test has been conducted. Due to time and practical constraints, the test has been reduced to the implicit and explicit combination (configuration 4) versus baseline comparison. The pairs already selected for the first evaluation step have been used.

The viewing sessions of the subjective tests followed the recommendation described in ITU-T P.910 (39). This double stimulus methodology, called pair comparison, consists of a same sequence being presented first through one system under test, and then through another system. In our case, the two systems are HEVC encodings with and without the proposed method, respectively. A set of pairs, corresponding to same combinations under the two systems, are generated. A pair (A,B) represents a given combination of sequence and quality, that are encoded with the two encoder systems. Their perceived qualities are meant to be comparable. Given the pair (A,B), each subject is asked to make a decision between four options: Option 1) A is much better than B, option 2) A is better than B, option 3) B is better than A, and finally, option 4) B is much better than A. It is noteworthy that subjects are kept unaware of the details of the systems and the random order in which the two sequences of each pair are shown. The test has been conducted on 30 naïve subjects.

PERFORMANCE ANALYSIS

Figure 3 summarizes the results obtained from the subjective test. In this figure, the left y-axis presents the average subjective score of the tested sequences with their confidence intervals. As can be seen, the average subjective score of the tested sequences indicates that the perceived quality of the videos encoded by the two systems were somewhat similar (i.e. Saliency = Baseline). Moreover, the green bars represented by the right y-axis of this figure demonstrate the amount of bit-rate saving that the proposed method can offer in the given subjective quality score. As can be seen, an average of 17% bit-rate reduction is



achieved by the proposed method, whereas the perceived quality has remained almost unchanged.

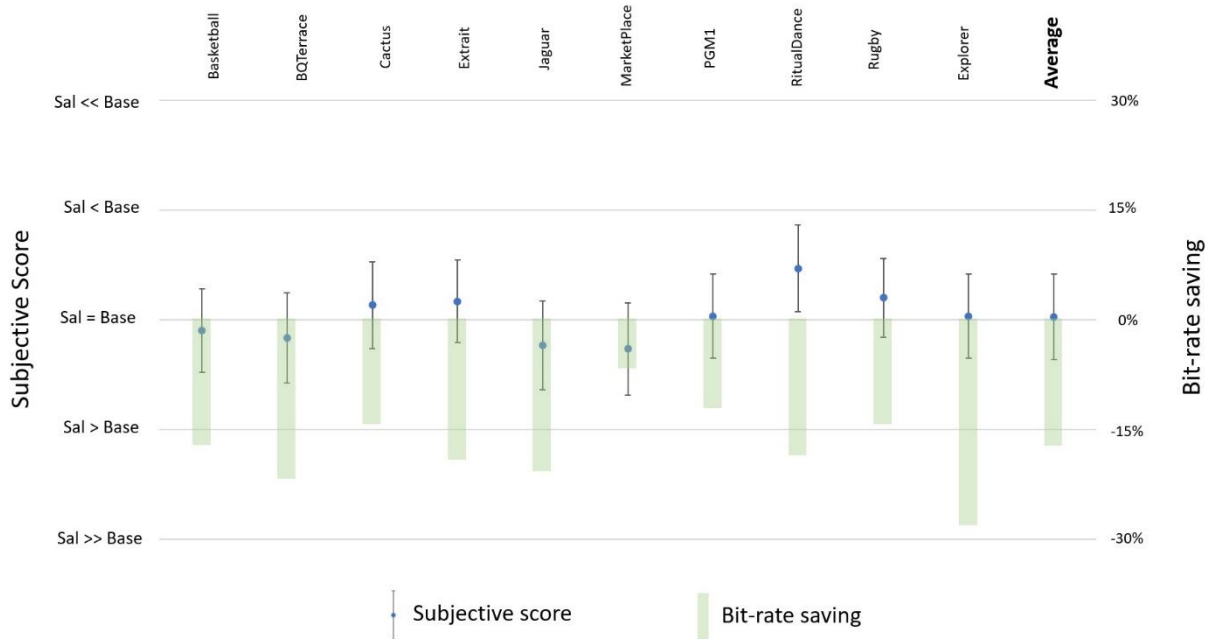


Figure 3: Subjective scores of the tested sequences as well as their bitrate saving in the given subjective quality level.

The RitualDance sequence shows poorer results as the mean subjective score tends towards a baseline preference. This score is mainly due to the saliency prediction quality. Indeed, as it can be seen in Table 2, the man in the foreground is not detected by the saliency model, while it is an important area of the sequence. Therefore, the quality of this area is degraded and this loss in quality is perceived by several viewers. As this man is clearly distinguishable in the sequence, the saliency model failure can be explained by the strong light area present in the middle of each frame. This is a specific case that the model may not have seen during the training phase, and that can be improved using data augmentation techniques reproducing this kind of case.

CONCLUSION

In this paper, a saliency-based video coding framework is introduced. This framework proposes that internal modules of encoder treat regions of input video differently based on their likelihood of being attractive for viewers. To apply this concept, saliency maps of video frames are extracted using a CNN-based method. These maps are then fed to two modules of pre-processing and rate-control, in order to put more importance on salient regions and encode them accordingly. The subjective tests show that the proposed method can provide 17% bit-rate saving on average to an HEVC codec while maintaining the same level of perceived quality. As a future track, one can design more adapted saliency map detection algorithms. Since the currently used network is designed to create saliency maps on single images, a video-specific network that detects salient zones of a frame based not only on the pixels of this frame but also on some video-specifics data (such as motion vectors) might be able to obtain more accurate saliency maps and thus improve the overall effect of our



solution. Moreover, data augmentation techniques reproducing video-specific features such as blur or scale changes could improve saliency prediction accuracy. Finally, the saliency model could be combined with specific detection models such as people detection that are very frequently areas of interest.

REFERENCES

1. Cisco, VNI Report 2021, https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf
2. DVB, "DVB-MABR: Adaptive Media Streaming over IP Multicast", <https://dvb.org/?standard=adaptive-media-streaming-over-ip-multicast>
3. ITU-T, "H.266: Versatile Video Coding", <https://www.itu.int/rec/T-REC-H.266>
4. Alliance for Open Media, "AV1 Bitstream & Decoding Process Specification", <https://aomedia-codec.github.io/av1-spec/av1-spec.pdf>, 2019-01-08.
5. P. Wennersten, *et al.* "GOP-based temporal filter improvements", Joint Video Experts Team (JVET), 22nd Meeting, by teleconference, 20–28 Apr. 2021, JVET-V0056
6. Ma, C. *et al.* A saliency prior context model for real-time object tracking. *IEEE Transactions on Multimedia*, 19(11), 2415-2424.
7. Lee, H. *et al.* Salient region-based online object tracking. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1170-1177). IEEE.
8. Xu, K., *et al.* Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
9. Qin, C. *et al.* Integration of the saliency-based seed extraction and random walks for image segmentation. *Neurocomputing*, 129, 378-391.
10. Fu, H. *et al.* Object-based multiple foreground segmentation in RGBD video. *IEEE Transactions on Image Processing*, 26(3), 1418-1427.
11. Li, Y. *et al.* The secrets of salient object segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 280-287).
12. Koch, C. *et al.* Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115-141). Springer, Dordrecht.
13. Itti, L. *et al.* A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.
14. Zhang, L. *et al.* SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7), 32-32.
15. Gao, D. *et al.* Discriminant saliency for visual recognition from cluttered scenes. *Advances in neural information processing systems*, 17, 481-488.
16. Bruce, N. *et al.* Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 5-5.
17. Harel, J. *et al.* Graph-based visual saliency.
18. Judd, T. *et al.* Learning to predict where humans look. In 2009 IEEE 12th international conference on computer vision (pp. 2106-2113). IEEE.
19. Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 185-207.



20. Judd, T. *et al.* A benchmark of computational models of saliency to predict human fixations.
21. Vig, E. *et al.* Large-scale optimization of hierarchical features for saliency prediction in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2798-2805).
22. Kruthiventi, S. *et al.* Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9), 4446-4456.
23. Huang, X. *et al.* Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 262-270).
24. Liu, N. *et al.* Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE trans. on neural networks and learning systems*, 29(2), 392-404.
25. Simonyan, K. *et al.* Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
26. Pan, J. *et al.* Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*.
27. Kroner, A. *et al.* Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129, 261-270.
28. Fang, Y. *et al.* A video saliency detection model in compressed domain. *IEEE transactions on circuits and systems for video technology*, 24(1), 27-38.
29. Bak, C. *et al.* Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7), 1688-1698.
30. Sun, Z. *et al.* Real-time video saliency prediction via 3D residual convolutional neural network. *IEEE Access*, 7, 147743-147754.
31. Tangemann, M., Kümmerer, M., Wallis, T. S., & Bethge, M. (2020, August). Measuring the Importance of Temporal Features in Video Saliency. In *European Conference on Computer Vision* (pp. 667-684). Springer, Cham.
32. Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing*, 13(10), 1304-1318.
33. Polakovič, A., *et al.* An approach to video compression using saliency based foveation. In *2018 International Symposium ELMAR* (pp. 169-172).
34. Gupta, R., *et al.* A scheme for attentional video compression. In *International Conference on Pattern Recognition and Machine Intelligence* (pp. 458-465). Springer, Berlin, Heidelberg.
35. Li, Z., *et al.* Visual attention guided bit allocation in video compression. *Image and Vision Computing*, 29(1), 1-14.
36. Zhu, S., *et al.* Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275, 511-522.
37. Lyudvichenko, V., *et al.* Improving video compression with deep visual-attention models. In *Proceedings of the 2019 International Conference on Intelligent Medicine and Image Processing* (pp. 88-94).
38. Wei, H. *et al.*, Visual saliency based perceptual video coding in HEVC. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 2547-2550).
39. ITU-R, “Methodologies for the subjective assessment of quality of television images”,
40. Netflix, Toward A Practical Perceptual VideoQuality Metric <https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652>.