Noname manuscript No.
(will be inserted by the editor)

Delivery of Omnidirectional Video using Saliency Prediction and Optimal Bitrate Allocation

Cagri Ozcinar · Nevrez İmamoğlu · Weimin Wang · Aljosa Smolic

Received: date / Accepted: date

Abstract In this work, we propose and investigate a user-centric framework for the delivery of omnidirectional video (ODV) on VR systems by taking advantage of visual attention (saliency) models for bitrate allocation module. For this purpose, we formulate a new bitrate allocation algorithm that takes saliency map and non-linear sphere-to-plane mapping into account for each ODV and solve the formulated problem using linear integer programming. For visual attention models, we use both image and video-based saliency prediction results; moreover, we explore two types of attention model approaches: i) salient object detection with transfer-learning using pre-trained networks, ii) saliency prediction with supervised networks trained on eye-fixation dataset. Experimental evaluations on saliency integration of models are discussed with interesting findings on transfer-learning and supervised saliency approaches.

Keywords 360°video streaming · attention based bitrate allocation · saliency maps with transfer learning and supervision

1 Introduction

Recent technological advancements in media streaming networks and virtual reality (VR) devices have made it feasible to deliver *omnidirectional video* (ODV) with high quality. A live ODV streaming service via the 5G network at the world of the Olympic Winter Games [10] is a decisive proof of relevance. ODV technology provides an interactive VR video experience than that available through traditional 2D video displayed with a flat-screen. ODV covers 360° of a scene and can be viewed through a head-mounted display (HMD), that allows viewers to look around a scene from a central point of view in VR. ODVs are stored in 2D planar representations, equirectangular projection (ERP), to be compatible with the existing video technology systems. Thanks to its immersive and interactive nature, ODV can be used in different applications such as entertainment, e-commerce, social media, and even job training.

Despite recent technological improvements and ODV streaming works, the establishment of a cost-efficient ODV streaming service is a challenging task because of its requirement of transmission of high volume of data size. This requirement can be characterized by high-resolution size and fidelity to ensure a high quality of experience (QoE). At any given time, HMDs render only a portion of the captured 360°scene, known as the viewport. Hence, there is a need for a user-centric VR system to optimize the use of ODV streaming in order to maintain QoE.

There have been various ODV streaming works proposed in the literature. For a comprehensive literature review, we refer the reader to [30] and [5]. The quality of viewing experience in VR can be improved by benefiting visual attention. A tiled-based system, for instance, was introduced by Ozcinar *et al.* [18], in which

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760, V-SENSE, Trinity College Dublin, Ireland. This paper is partly based on the results obtained from a project commissioned by Public/Private R&D Investment Strategic Expansion Program (PRISM), AIST, Japan. Cagri Ozcinar and Nevrez İmamoğlu equally contributed to this work.

Cagri Ozcinar* · Aljosa Smolic
V-SENSE, Trinity College Dublin, Dublin, Ireland
E-mail: (c.ozcinar, smolica) @tcd.ie

Nevrez İmamoğlu* · Weimin Wang
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
E-mail: (nevrez.imamoglu, weiming.wang) @aist.go.jp

a visual-attention based quality metric was introduced to find optimal tiling schemes for ODV streaming. Petrangeli *et al.* [22] modeled the evolution of the user's ODV viewing trajectory and predicted the future viewport position of a new user in the ODV system. Furthermore, Nguyen *et al.* [16] presented an adaptation logic for ODV streaming to decide an optimal version of each tile according to users' head movements and network bandwidth. Recently, Kan *et al.* [12] developed a deep reinforcement learning-based rate adaptation algorithm to control the ODV bitrates to achieve an optimal trade-off between switching latency and bandwidth efficiency. Their simulation result shows that the proposed algorithm keeps the buffer occupancy level to a low level while improving the overall QoE.

This work extends the streaming model in Ozcinar et al. [18], where visual attention is used as the focus area of users HMD viewports. In particular, this previous model requires several subjects to watch the same ODVs estimated by an averaged visual attention map from collected data before streaming. This data can be seen as a ground-truth attention area for streaming, but in a real-world scenario for large-scale datasets or VR videos online, this type of subject-based fixation area may not be available. Saliency prediction models has been used in various multimedia compression applications [8], re-targeting [6], including QA [13, 28] too. Therefore, saliency prediction models can be complementary to the proposed idea in [18], and as in this work, the collected subject data can be used to evaluate the quality of streaming based on the viewport position for specific frames in each video. Besides, the algorithm in [18] distributed the target bitrate to the tiles using only the value of the visual attention, and the nearby areas of the salient pixels were not considered for the formulation of bitrate allocation. To address this drawback, in this work, we base on the idea of the viewport-aware representation selection in [19], and extended it with saliency prediction and viewport proposal modules. Hence, in addition to providing high importance to the high salient regions by selecting most probable viewports, the new formulation also assigns more importance to the tiles that are closer to the selected viewports by the viewport proposal module. The bitrate of the tiles is gradually reduced based on the distance between the selected viewports and each tile.

In this work, we proposed a user-centric framework for the delivery of ODVs for VR systems (e.g., HMD) using saliency maps. For this purpose, we first formulate a new bitrate allocation algorithm that considers saliency map and non-linear sphere-to-plane mapping using the WS-PSNR metric [27] for each ODV. The bitrate allocation optimization problem is then solved with linear integer programming (ILP). Then, we examine how image and video saliency maps (i.e., video saliency is the fusion of motion and image saliency maps) affect the proposed ODV streaming system. Moreover, regarding image saliency on optimal bitrate allocation process for video streaming, we have evaluated the effect of two types of computational visual attention (saliency map) approaches: i) *finding what to look (pre-trained networks for weakly-supervised detection [9]):* using transfer-learning as means of taking advantage of existing knowledge of object representations learned through the large-scale object-recognition task (i.e., ImageNet image recognition dataset [23]); ii) *learning where to look (fully supervised model [4]):* supervision of a neural network to generate saliency map by using the eye-fixation dataset (i.e., SALICON saliency dataset [11]). In this case, we also investigate the effect of image saliency from computation through both merging 2D image projections from spherical representation and ERP.

The remainder of this paper is organized as follows. In Sec. 2, we describe our proposed framework. Then, we present evaluation results in Sec. 3. Finally, we conclude this paper in Sec. 4.

2 System Model

We consider a tile-based adaptive streaming pipeline to deliver ODVs over internet networks, as depicted in Fig. 1. The server side of the proposed system contains tiling and encoding, saliency prediction, viewport proposal, bitrate allocation, and packing for adaptive streaming. The client-side of the proposed system utilizes a tile-based adaptive streaming player, dynamic adaptive streaming over HTTP (DASH) [19] to navigate through delivered cost-effective ODV streams and consume with HMDs. We extended the bitrate allocation algorithm proposed in [18,19] by integrating saliency models to achieve a cost-efficient performance, considering the target bitrate and a predicted saliency map for each c-th chunk. This bitrate allocation algorithm can be named as saliency-aware bitrate allocation.

At the server side, we partition each ODV into tiles where each tile is encoded at various bitrate levels. In this framework, each tile is considered as an independent bitstream so that the viewport of the client can be processed independently from the rest of the video. For adaptive streaming, each encoded tile is also divided into temporal chunks and stored on the HTTP server.

The proposed system prepares an optimal DASH representation for each ODV at a given target bitrate budget. For this, we propose a saliency-aware bitrate allocation module which consist of saliency prediction and bitrate allocation modules. The saliency prediction

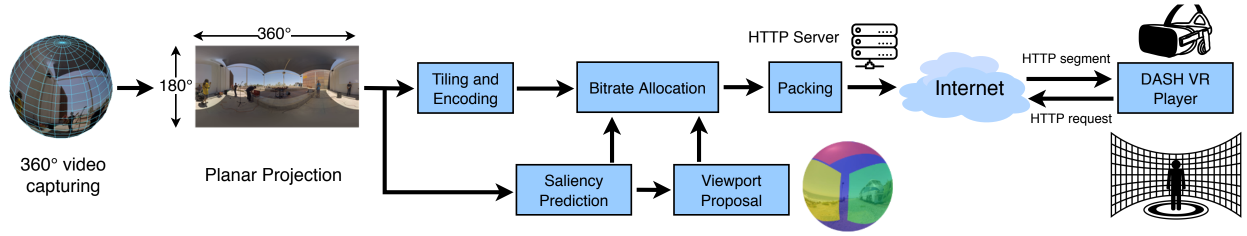

Fig. 1: Schematic diagram of the proposed ODV streaming system with saliency prediction algorithm.

algorithm estimates a saliency map per chunk of each ODV. The predicted saliency map is then used with the viewport proposal algorithm to estimate N number of possible viewports to be used for the proposed bitrate allocation algorithm. Our bitrate allocation algorithm aims to generate cost-optimal DASH representation for each ODV. Hence, it depends on the target bitrate for each c, R^c, and the predicted viewport proposals, $\widetilde{V}_n$ and $n \in \{1, 2, \ldots, N\}$.

2.1 Bitrate allocation

We formulate a new bitrate allocation algorithm for DASH systems on the server-side that take saliency map and sphere-to-plane mapping into account for each ODV. Our objective is to provide a high-quality ODV streaming experience with a given target bitrate of the c-th chunk, R^c. To this end, as in [18, 19], we focus on the decision of an optimal bitrate r for each tile t with $t \in \mathcal{T}$, denoted as $\mathcal{T}^*_{R^c}$ at c-th chunk by formulating the optimization problem with ILP as follows:

$$\mathcal{T}^*_{R^c} : argmax \sum_{t \in \mathcal{T}} q_{tr} a_{tr} w_t \quad \forall r \in \mathcal{R}, w \in \mathbf{W}, \tag{1}$$

with

$$\sum_{r \in \mathcal{R}} a_{tr} \leq 1 \quad \forall t \in \mathcal{T}, \tag{2}$$

and

$$\sum_{r \in \mathcal{R}} \sum_{t \in \mathcal{T}} b_{tr} a_{tr} \leq R^c, \tag{3}$$

where q_{tr} and a_{tr} are the quality value in terms of WS-PSNR [27] and decision variable for the t-th tile of the r-th rate. The decision variable $a_{tr} = \{0, 1\}$ indicates whether the t-th tile of the r-th rate is included *or* excluded for $\mathcal{T}^*_{R^c}$, which is able to reconstruct the ODV at R^c. The data rate is represented with b_{tr} for the r-th rate of the t-th tile. Also, w_t is a weight of the t-th tile, $\mathbf{W} = [w_1, w_2, \ldots w_T]$ and T is the total number of tiles, $T = |\mathcal{T}|$, that is calculated as follows:

$$\mathbf{W} = \sum_{n}^{N} p^n \widehat{\mathbf{L}^n} \quad \text{with} \quad \widehat{\mathbf{L}^n} = \frac{\mathbf{L}^n}{\sum_{t \in \mathcal{T}} \mathbf{L}^n_t}, \tag{4}$$

where p^n represents the probability that n-th viewport be watched according to the estimated saliency map, and $\sum_n^N p^n = 1$. Also, $\widehat{\mathbf{L}^n}$ is a normalized array for weights of $\mathbf{L}^n$ which can be estimated using a linear equation as follows:

$$\mathbf{L}^n = \mathbf{W}^n_{in} \beta + \mathbf{W}^n_{out}(1 - \beta), \tag{5}$$

where $\mathbf{W}^n_{in}$ and $\mathbf{W}^n_{out}$ be the sets of tile weights inside the n-th viewport (or overlapping with it), and outside the n-th viewport, respectively. The constant parameter of β defines distribution between $\mathbf{W}^n_{in}$ and $\mathbf{W}^n_{out}$, and $\beta = [0, 1]$. Here, $\mathbf{W}^n_{in}$ contains a weight for each t tile, $\mathbf{W}^n_{in} = [k^n_{in,0}, k^n_{in,1}, \ldots, k^n_{in,T}]$, that overlap with the area of n-th predicted viewport, $\widetilde{V}^n$, and can be estimated as follows:

$$k^n_{in,t} = \begin{cases} \frac{f(\widetilde{V}^n \cap \mathcal{S}^n_{in})}{f(\widetilde{V}^n)} & \widetilde{V}^n \cap \mathcal{S}^n_{in} \neq \emptyset \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where f is the function that estimates the corresponding number of pixels on the 3D sphere, the sets of $\mathcal{S}^n_{in}$ and $\mathcal{S}^n_{out}$ represent tiles inside of the viewport, $\widetilde{V}^n$, and outside of the viewport, respectively.

Next, we estimate $\mathbf{W}^n_{out}$, which has weights for the outside-viewport tiles, $\mathcal{S}^n_{out}$. For this purpose, we use the Euclidean distance, δ_t, measure the distance between the spherical center location of $\widetilde{V}_n$ and the spherical center location of each t-tile in $\mathcal{S}^{out}$. Each weight, $k^n_{out,t} \in \mathcal{S}^n_{out}$, is estimated as follows:

$$k^n_{out,t} = \frac{\kappa_t}{\Sigma_t \kappa_t} \quad \text{with} \quad \kappa_t = \frac{\max\{\delta_t\}}{\delta_t}. \tag{7}$$

2.2 Saliency prediction model

The general flowchart for the image/video saliency used in this work (Saliency Prediction block in Fig.1) can be seen in Fig.2. In VR applications, scenes in 360°are

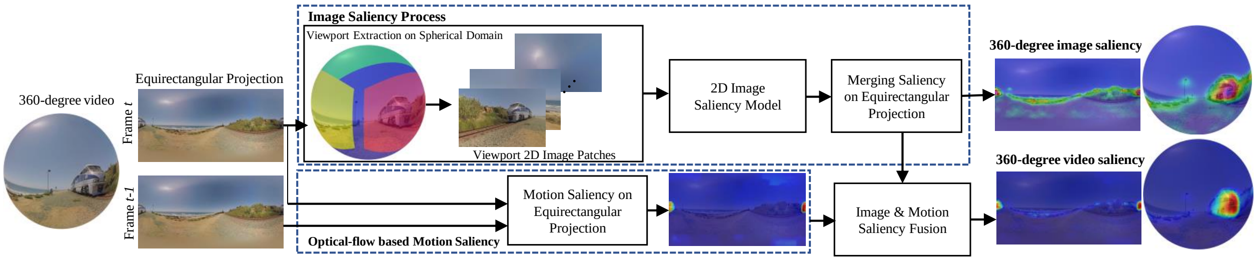
360-degree video
Equirectangular Projection
Frame t
Frame t-1
Image Saliency Process
Viewport Extraction on Spherical Domain
Viewport 2D Image Patches
2D Image Saliency Model
Merging Saliency on Equirectangular Projection
360-degree image saliency
Motion Saliency on Equirectangular Projection
Optical-flow based Motion Saliency
Image & Motion Saliency Fusion
360-degree video saliency

Fig. 2: Flowchart of 360°image and video saliency prediction block of Fig.1.

not visible to the observers fully except for the HMD viewport field-of-view, and without any cue of spatial audio or without seeing the whole scene in ERP, it is hard to pay attention to motion specific salient regions out of HMD viewport. To consider these aspects, along with video saliency, we also focus on the image saliency models for ODV frames separately.

To obtain image saliency maps for each video frame, we investigate two types of deep learning approaches: i) a transfer-learning based saliency model as VGG-TL proposed in [9]), and ii) a supervised saliency detection model as VGG-MLNet introduced in [4]). To compare two saliency approaches in ODV delivery as in Fig.1, VGG-TL [9] and VGG-MLNet [4] are specifically chosen due to having the same core network structure and similar saliency detection process to compare their effect on the proposed ODV streaming process. Both [9] and [4] employ the VGG-16 deep convolutional neural network (CNN) introduced in [26] as the core of their saliency prediction models, and they both apply saliency extraction through the fusion of different layers in the CNN. However, one requires supervision with relatively large eye-fixation data, and the other uses image recognition task based pre-trained network.

We extend the use of VGG-TL [9] and VGG-MLNet [4] on our ODV image saliency computation as in Fig.2 by using it in 2D Image Saliency Model block for the extracted image patches. For both VGG-TL and VGG-MLNet, our framework first need to create various overlapping 2D projections of scene using spherical coordinate at fixed image resolutions (i.e. 512×512 pixels and 480×640 pixels for VGG-TL and VGG-MLNet models respectively) with $\theta^{FoV} = 120$°at $\theta^{viewport-center} =$ [0, 15 , ..., 330, 345] and $\phi^{viewport-center} =$ [-90, -75, ..., 0, ..., 75, 90] similar to various works on 360°data processing [13–15, 31]. Then, each 2D image patch from the scene is used in VGG-TL or VGG-MLNet to obtain saliency maps of each extracted view-ports. Finally, saliency values at each pixel location of ERP is computed from the nearest k pixels back-projected through 2D image patches by weighted averaging the saliency values based on the inverse proportion of the distance.

In the following part, we explain VGG-TL as transfer learning based saliency prediction and VGG-MLNet as the supervised saliency detection model. Then, video saliency map is also described briefly, as a combination of motion saliency with the image saliency from models (VGG-TL, VGG-MLNet, or VGG-MLNet on ERP).

2.2.1 VGG-TL: 360°image saliency with transfer learning

In weakly-supervised object detection models, it is possible to localize the class specific objects or semantics of the scene using class activation mapping by class specific weighted average of features or gradient of the semantic features (e.g. CAM [32], Grad-CAM [24], DCSS [25]). Inspired by these approaches, but instead of class specific attention, the existing knowledge of network can also be used to generate a general attention framework to find what to look. To be able to obtain attention through object recognition task, VGG-TL proposed by Imamoglu et al. [9] uses learned representations of objects and gradient features (i.e. combining forward representation feature and backward gradient features at different convolution layers averaged by empirically selected weights). VGG-TL [9] takes advantage of pre-trained VGG-16 CNN [26] trained to handle recognizing 1000 objects in a large-scale image recognition dataset called ImageNet [23].

2.2.2 VGG-MLNet: 360°image saliency with supervised model

To compare transfer-learning based saliency VGG-TL to an end-to-end supervised saliency model, we selected fully-supervised model proposed in [4] as VGG-MLNet, which also utilizes VGG-16 CNN [26] without the fully connected layers in original model. In addition to backbone network, VGG-MLNet includes two more convolution layers to be able to learn optimal fusion of extracted features from the last three convolution block representations of VGG-16 backbone structure.

In [4], the whole model is fine-tuned using SALICON saliency dataset [11] based on the eye-fixations collected from subjects. Before fine-tunning the model

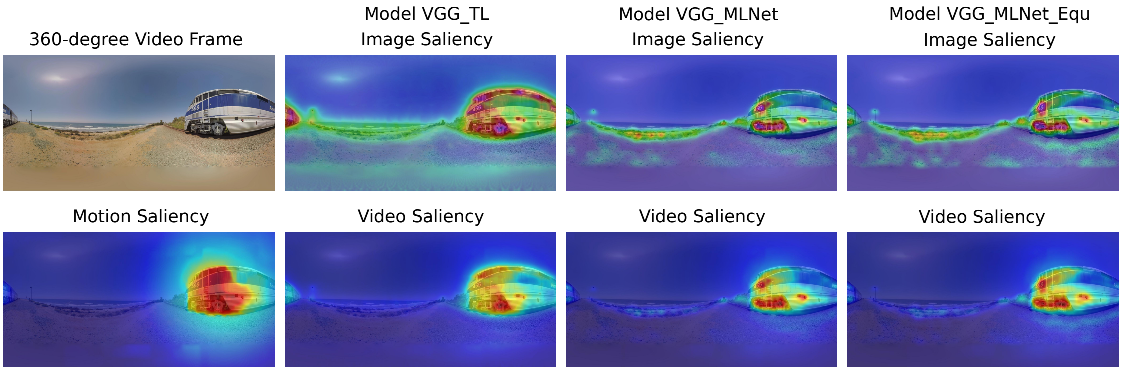

Fig. 3: Sample saliency maps predicted by the evaluated models.

in [4], VGG-16 CNN part of model is initialized with the pre-trained weights learned with ImageNet dataset [23]. However, in this work, we train VGG-MLNet with random initialization for the whole network on SALICON dataset [11] since we want model to learn soley where people looked without the prior information learned regarding to objects representations from Imagenet dataset [23].

We also use VGG-MLNet to evaluate the performance of saliency from the ERP as 2D image input to the CNN. The main objective of this is to investigate saliency performance of ERP for ODV streaming compared with the saliency obtained through merging 2D images patches extracted as the viewports from spherical coordinates as demonstrated in Fig.2. Here, since the VGG-MLNet approach uses saliency location prior which decreases the saliency information at the image boundaries, we shift ERP images on horizontal axis to bring the boundary scene to the center. Therefore, VGG-MLNet is applied on two ERP images then the results are merged by selecting the maximum values to provide the boundary connectivity or continuity for saliency computation on ERP input data.

2.2.3 360°video saliency

In addition to the image saliency for each video frame, we also obtain video saliency by computing motion saliency and combining it with image saliency results as in Fig.2. Motion saliency of frames is computed in ERP by taking advantage of optical flow distribution in the scene by using the approach in [7], then as for video saliency, image saliency and motion saliency is combined by averaging summation and element-wise multiplication of image and motion saliency maps as in [7]. It should be noted that all the saliency computations are done on 2K down-sampled from 8K images for computational efficiency, especially for motion saliency.

In Fig.3, a sample frame from the *Train* video in ERP is given with its respective motion saliency result, which is combined with the image saliency models evaluated in this work resulting the video saliency for each model. In summary, we experimented and evaluated the proposed ODV streaming in Fig.1 on six videos with the following saliency models:

1. Image saliency with VGG-TL (VGG-TL-Image);
2. Image saliency with VGG-MLNet (VGG-MLNet-Image);
3. Image saliency with VGG-MLNet on ERP (VGG-MLNet-Equ-Image);
4. Video saliency with VGG-TL-Image (VGG-TL-Video);
5. Video saliency with VGG-MLNet-Image (VGG-MLNet-Video);
6. Video saliency using VGG-MLNet-Image on ERP (VGG-MLNet-Equ-Video).

3 Results

3.1 Experimental Setup for ODV Streaming

We used the following six ODVs from the JVET and MPEG video coding exploration experiments: $\mathcal{V} = \{$ *Harbor*, *Gaslamp*, *KiteFlite*, *Train*, *Basketball*, *Dancing* $\}$. Each ODV is in *uncompressed* and 8K×4K ERP resolution with YUV420p format of 10 *sec.* length. These ODVs were selected from the videos of the joint video exploration team of ITU-T VCEG and ISO/IEC MPEG [1–3].

We divided each ODV frame into six independetly encoded tiles based on the described tiling structure in [20]. This structure consists two tiles at the poles and four tiles at the equatorial region. The reason of this as the existing HMDs have an approximate 90° of field-of-view (such as Oculus Rift). In addition, most available ODVs have the most salient objects in the area of the equatorial segment that spans 90° longitude. Hence, we consider the equatorial region as a 90° longitude segment. The equator was further split into 90° tiles as in [20].

For encoding, we used the HEVC standard [17] to compress each tile of ODVs. For this, the FFmpeg software (*ver.* N-85291) [29] was utilized with two-pass and 150 percent constrained variable bitrate configurations to encode each tile. In addition, we used a set of bitrate level for each tile $\mathcal{B} = \{$ 334, 467, 654, 915, 1281, 1793, 2510, 3514, 4920, 6887$\}$ (in terms of *kbps*) and ($b_{tr} \in \mathcal{B}$) to encode each ODV content. Each bitstream was divided into 2 *sec.* chunks.

We calculated PSNR and SSIM quality within viewport with the collected viewing trajectories. For this purpose, we randomly selected ten user viewport trajectories from the published users' viewport trajectories datasets in [18, 21]. The quality scores were measured on 2D rectilinear viewport image in a frame-by-frame manner. In quality measurement, the viewport rendered from the reconstructed ODV is compared with the viewport rendered from uncompressed ODV.

3.2 Evaluation of Saliency Integration on ODV Streaming

In Fig.4 and Fig.5, for each ODV, we demonstrated the PSNR and SSIM results on each model with streaming at 15Mb/s, 10Mb/s, 5Mb/s, and 3Mb/s bitrates. We will discuss our findings in detail for each ODV, and then, demonstrate the average evaluation results for all ODVs which is given in Table 1. As expected, there is a tendency of decreased performance both in PSNR and SSIM metrics using each models tested on all videos while bitrates are decreasing (see Fig.4 and Fig.5).

One interesting finding is observed in using the transfer learning or object-knowledge based image saliency approach with VGG-TL [9], which is proposed as a salient object detection model rather than trying to predict saliency for eye-fixation regions. In all bitrates, the average video quality using image only saliency *VGG-TL-Image* seems, in most cases (except for the *Train* and *Dancing* ODVs), to be performing better than or on par with using video saliency model, *VGG-TL-Video*. These two ODVs contain highly salient moving objects, such as a fast-moving train and dancing girls. The *VGG-TL-Video* method, contains transfer-learning using object knowledge and optical flow distribution, performing better on the ODVs that have fast-moving salient objects. Differently, image only saliency representation is correlated with a static scene, which contains low motion information, performing better saliency distribution in favor of increasing streaming quality.

In addition, the used motion saliency algorithm in [7] normalizes the saliency output to be more distinctive, so that if there are multiple moving objects, some other moving objects will have much less saliency values depending on the motion distribution and motion similarity/dissimilarity values. This normalization procedure can be beneficial on standard 2D video since viewer can see the whole scene where global motion can be observed. However, in ODVs, any motion cues can be equally important depending on the viewport observed by the viewer. Because during interactive look-around viewing, subjects may not observe the motion cues without checking the whole 360°scene.

On the other hand, VGG-MLNet image saliency, which is saliency prediction model supervised with an eye-fixation dataset, enhanced the overall video average streaming quality in almost all bit-rates when it is combined with the motion saliency referred as VGG-MLNet-Video (see. Table.1, Fig.4, and Fig.5 results). So, regarding VGG-MLNet video saliency, the results suggest that motion saliency is complementary to VGG-MLNet image saliency in favor of enhancing the overall perceived streaming quality. However, as in VGG-TL model case, we do not see same tendency in low bit streaming scenarios while MLNet applied to ERP image as VGG-MLNet-Equ-Image and VGG-MLNet-Equ-Video (Table.1, Fig.4, and Fig.5).

Moreover, VGG-MLNet-Equ either only image or video saliency seems to resulting in best average PSNR and SSIM scores in all bit-rates. It may be because of the fact that VGGMLNet model has a location-prior learned through training, which decreases the saliency values at the edges or corners. Especially, suppressing saliency values upper and lower boundaries of frames in our case may be the main reason in improved results when using the ERP. However, the average PSNR and SSIM values does not have big differences. In addition, transfer learning results might be better with model which does not require down-sampling. In that case, it would be also possible to test it in ERP. It should be noted that our primary aim in this work was to analyze transfer-learning saliency (using object knowledge) and supervised saliency (learning based on eye-fixation data) models rather than providing state-of-the-art streaming approach though the results seems to be fairly good for all saliency models integrated to ODV streaming process.

4 Conclusion

In this paper, we studied an omnidirectional video (ODV) streaming system by introducing a new bitrate allocation and saliency prediction modules. The proposed bitrate allocation algorithm takes the saliency map computed by learning-based saliency methods, predicts several most probable viewports, considers the distortion

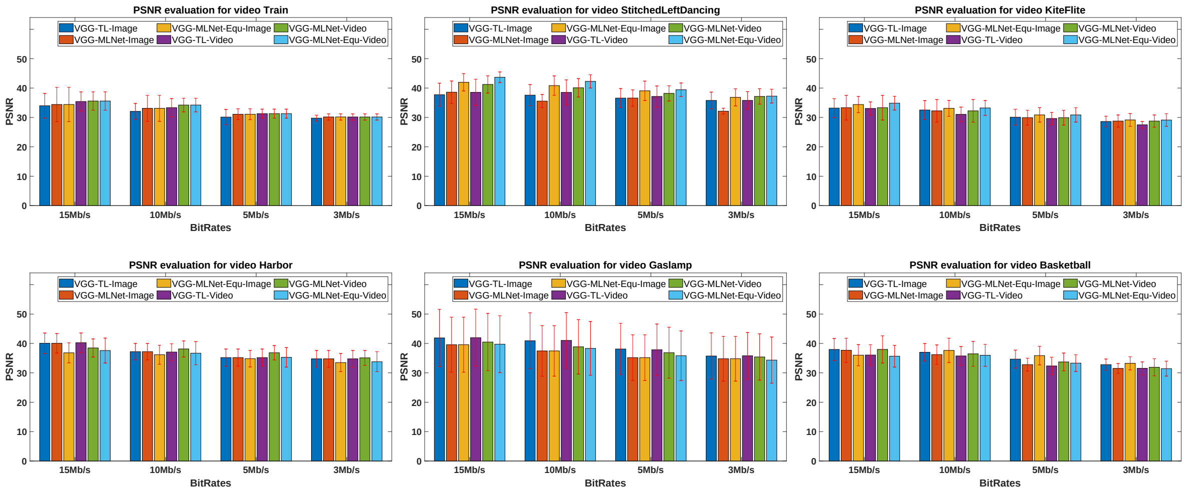

Fig. 4: Model comparisons on each ODV based on PSNR.

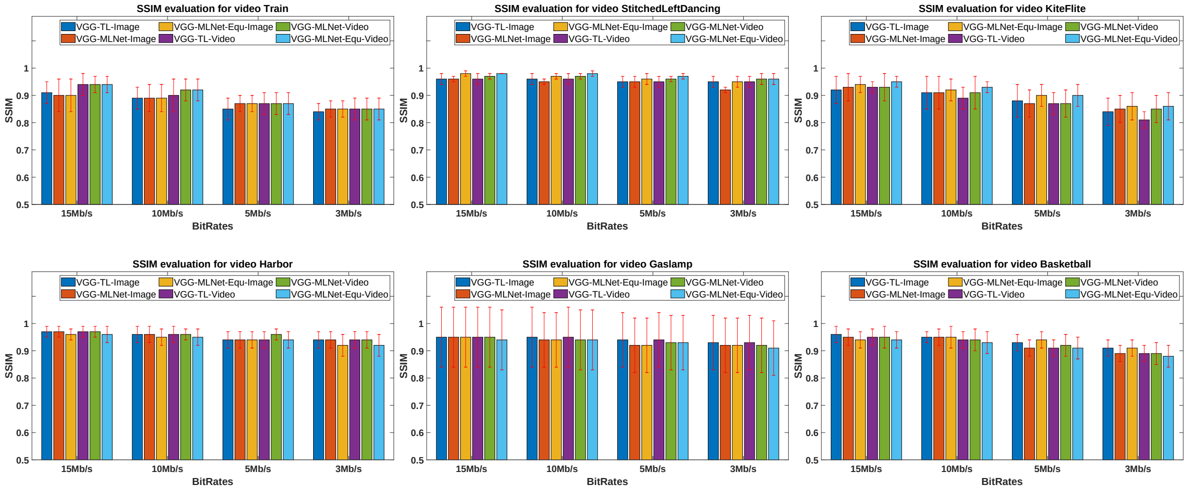

Fig. 5: Model comparisons on each ODV based on SSIM.

Table 1: Average PSNR and SSIM values for ODVs at defined target bitrates.

Data Bitrate / Saliency Model		Average PSNR				Average SSIM			
		15Mb/s	10Mb/s	5Mb/s	3Mb/s	15Mb/s	10Mb/s	5Mb/s	3Mb/s
Image	VGG-TL	37.47	36.22	34.13	32.91	0.9450	0.9367	0.9150	0.9017
	VGG-MLNet	37.28	35.30	33.47	32.03	0.9433	0.9333	0.9100	0.8950
	VGG-MLNet-Equ	37.20	36.39	34.48	32.95	0.9450	0.9367	0.9217	0.9017
Video	VGG-TL	37.55	36.13	33.91	32.59	0.9500	0.9333	0.9133	0.8950
	VGG-MLNet	37.84	36.67	34.47	33.08	0.9517	0.9400	0.9183	0.9017
	VGG-MLNet-Equ	37.86	36.78	34.34	32.68	0.9517	0.9417	0.9200	0.8967

of sphere-to-plane mapping, and determines optimal target bitrates using the formulated linear integer programming problem. In learning-based saliency, we investigated two different learning-based saliency methods with an aspect of the delivery of ODV. One technique focused on salient object detection with transfer-learning using pre-trained networks. Another approach considered supervised networks trained on a well-known eye-fixation dataset. The proposed delivery system was then investigated with these visual attention models with an aspect of content types and bitrate levels. Experimental evaluations on saliency integration of models were discussed with the finding of transfer-learning and supervised saliency approaches. As future work, we plan to extend the proposed system by developing a new computationally ODV saliency map prediction model with the findings of this paper and considering the learning-based bitrate allocation model in ODV streaming.

References

1. Abbas, A., Adsumilli, B.: AhG8: New GoPro test sequences for virtual reality video coding. Tech. Rep. JVET-D0026, JTC1/SC29/WG11, ISO/IEC, Chengdu, China (2016)
2. Asbun, E., He, H., Y., H., Ye, Y.: AhG8: InterDigital test sequences for virtual reality video coding. Tech. Rep. JVET-D0039, JTC1/SC29/WG11, ISO/IEC, Chengdu, China (2016)
3. Bang, G., Lafruit, G., Tanimoto, M.: Description of 360 3D video application exploration experiments on divergent multiview video. Tech. Rep. MPEG2015/ M16129, JTC1/SC29/WG11, ISO/IEC, Chengdu, China (2016)
4. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A Deep Multi-Level Network for Saliency Prediction. In: International Conference on Pattern Recognition (ICPR) (2016)
5. Fan, C.L., Lo, W.C., Pai, Y.T., Hsu, C.H.: A survey on 360° video streaming: Acquisition, transmission, and display. ACM Comput. Surv. (2019)
6. Fang, Y., Chen, Z., Lin, W., Lin, C.W.: Saliency Detection in the Compressed Domain for Adaptive Image Retargeting. IEEE Transaction on Image Processing **21**(9), 3888 – 3901 (2012). DOI 10.1109/TIP.2012.2199126
7. Fang, Y., Wang, Z., Lin, W., Fang, Z.: Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Transactions on Image Processing **23**(9), 3910—3921 (2014). DOI 10.1109/tip.2014.2336549. URL https://doi.org/10.1109/TIP.2014.2336549
8. Hadizadeh, H., Bajić, I.V.: Saliency-Aware Video Compression. IEEE Transaction on Image Processing **23**(1), 19–33 (2014). DOI 10.1109/TIP.2013.2282897
9. Imamoglu, N., Zhang, C., Shimoda, W., Fang, Y., Shi, B.: Saliency detection by forward and backward cues in deep-cnns. In: 2017 IEEE International Conference on Image Processing (ICIP) (2017)
10. Intel: Experience the Future of the Olympic Games with Intel (2018). URL https://www.olympic.org/news/experience-the-future-of-the-olympic-games-with-intel
11. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
12. Kan, N., Zou, J., Tang, K., Li, C., Liu, N., Xiong, H.: Deep reinforcement learning-based rate adaptation for adaptive 360-degree video streaming. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4030–4034 (2019)
13. Li, C., Xu, M., Jiang, L., Zhang, S., Tao, X.: Viewport Proposal CNN for 360° Video Quality Assessment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. Mazumdar, P., Lamichhane, K., Carli, M., Battisti, F.: A feature integrated saliency estimation model for omnidirectional immersive images. Electronics **8**(12), 1538 (2019). DOI 10.3390/electronics8121538. URL http://dx.doi.org/10.3390/electronics8121538
15. Monroy, R., Lutz, S., Chalasani, T., Smolic, A.: Salnet360: Saliency maps for omni-directional images with cnn. Signal Processing: Image Communication **69**, 26 – 34 (2018)
16. Nguyen, D.V., Tran, H.T.T., Pham, A.T., Thang, T.C.: An optimal tile-based approach for viewport-adaptive 360-degree video streaming. IEEE Journal on Emerging and Selected Topics in Circuits and Systems (2019)
17. Ohm, J.R., Sullivan, G.: Vision, applications and requirements for high efficiency video coding (HEVC). Tech. Rep. MPEG2011/N11891, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland (2011)
18. Ozcinar, C., Cabrera, J., Smolic, A.: Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality. IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**(1), 217–230 (2019). DOI 10.1109/JETCAS.2019.2895096
19. Ozcinar, C., De Abreu, A., Smolic, A.: Viewport-aware adaptive 360 video streaming using tiles for virtual reality. In: 2017 IEEE International Conference on Image Processing (ICIP17) (2017)
20. Ozcinar, C., De Abreu, A., Smolic, A.: Viewport-aware adaptive 360 video streaming using tiles for virtual reality. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2174–2178 (2017)
21. Ozcinar, C., Smolic, A.: Visual attention in omnidirectional video for virtual reality applications. In: 10th International Conference on Quality of Multimedia Experience (QoMEX 2018). Sardinia, Italy (2018)
22. Petrangeli, S., Simon, G., Swaminathan, V.: Trajectory-based viewport prediction for 360-degree virtual reality videos. In: 2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), pp. 157–160 (2018). DOI 10.1109/AIVR.2018.00033
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). DOI 10.1007/s11263-015-0816-y
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2019). DOI 10.1007/s11263-015-0816-y
25. Shimoda, W., Yanai, K.: Distinct Class-Specific Saliency Maps for Weakly Supervised Semantic Segmentation. In: European Conference on Computer Vision (ECCV) (2016)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
27. Sun, Y., Lu, A., Yu, L.: Weighted-to-spherically-uniform quality evaluation for omnidirectional video. IEEE Signal Processing Letters (2017)
28. Tang, L., Wu, Q., Li, W., Liu, Y.: Deep Saliency Quality Assessment Network With Joint Metric. IEEE Access **6**, 913–924 (2017). DOI 10.1109/ACCESS.2017.2776344
29. x265: x265 HEVC Encoder / H.265 Video Codec. http://x265.org/ (2018)
30. Xu, M., Li, C., Zhang, S., Le Callet, P.: State-of-the-art in 360° video/image processing: Perception, assessment and compression. IEEE Journal of Selected Topics in Signal Processing pp. 1–1 (2020). DOI 10.1109/JSTSP.2020.2966864
31. Xu, M., Li, C., Zhang, S., Le Callet, P.: State-of-the-art in 360° video/image processing: Perception, assessment and compression. IEEE Journal of Selected Topics in Signal Processing **PP**, 1–1 (2020). DOI 10.1109/JSTSP.2020.2966864
32. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)