Versatile Video Coding of 360-Degree Video using Frame-based FoV and Visual Attention

J. Carreira, Sérgio M. M. de Faria, Luis M. N. Távora, António Navarro, Pedro A. A. Assunção

Instituto de Telecomunicações
Instituto Politécnico de Leiria
Universidade de Aveiro
Portugal

21st IEEE International Symposium on Multimedia, San Diego, USA

December 10, 2019

Summary

- Context and motivation
- Existing 360° video coding schemes
- Proposed Frame-based FoV coding using visual attention
- Performance evaluation
- Conclusions and future directions

Context and motivation

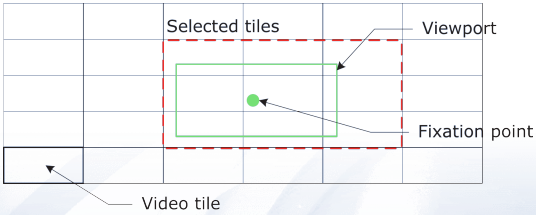
- 360° video popularity has been increasing over the last years:
  - diverse solutions for 360° video acquisition;
  - high range of head-mounted displays (HMDs);
- Availability in well-known video sharing services (e.g., Youtube, Facebook).

Motivation

- High resolution required by applications poses new challenges in by 360° video streaming:
  - Live 360° video streaming, surveillance and monitoring.
- High demand for flexible video coding and streaming schemes that can adapt to the users' needs:
  - Spatial/temporal resolution;
  - Bitrate/quality;
  - Field-of-View (FoV).

Existing 360° video coding schemes

- Video partitioning using the tilling concept.
- Adaptive streaming where the receiver requests the desired view-port based on the viewer's orientation.
- Asymmetric view-port coding and support for different levels of quality.



Existing 360° video coding schemes

Disadvantages

- Requires constant feedback from the viewer's orientation.
- Server-side complexity significantly increases with the number of streams.
- High complex streaming is not suitable for real-time video systems.

Proposed Frame-based FoV coding using visual attention

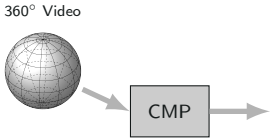
Problem

- How to efficiently encode 360° video using VVC, enabling independent Field-of-View (FoV) decoding?

Solution

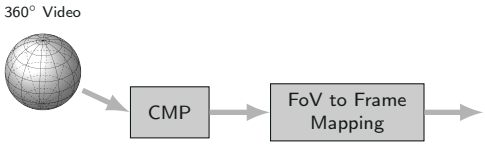
- Take advantage of implicit temporal scalability of VVC.
- Use visual attention based quantisation to reduce the overall bitrate.

Proposed Frame-based FoV coding using visual attention



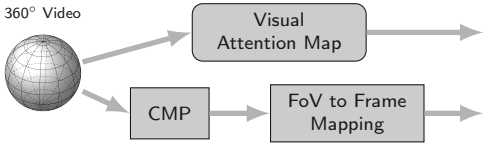
- The proposed approach is based on six 90° FoVs obtained from the Cube-Map Projection (CMP).

Proposed Frame-based FoV coding using visual attention



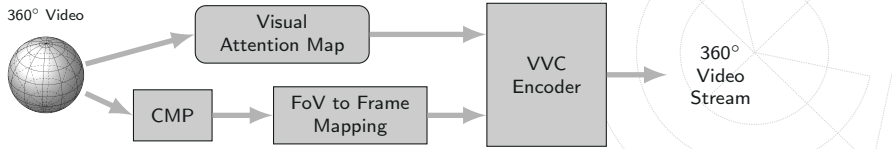
- The proposed approach is based on six 90° FoVs obtained from the Cube-Map Projection (CMP).
- FoV-to-frame mapping to obtain each FoV as a frame sequence.

Proposed Frame-based FoV coding using visual attention



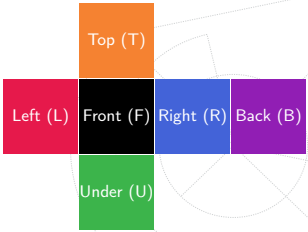
- The proposed approach is based on six 90° FoVs obtained from the Cube-Map Projection (CMP).
- FoV-to-frame mapping to obtain each FoV as a frame sequence.
- Visual attention based quantisation for perceptual bitrate allocation.

Proposed Frame-based FoV coding using visual attention

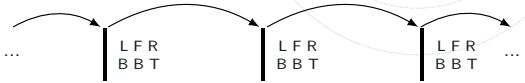


- The proposed approach is based on six 90° FoVs obtained from the Cube-Map Projection (CMP).
- FoV-to-frame mapping to obtain each FoV as a frame sequence.
- Visual attention based quantisation for perceptual bitrate allocation.

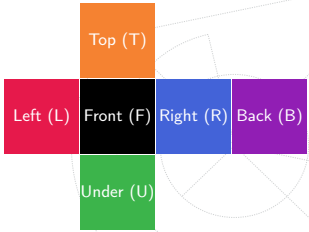
FoV-to-frame mapping



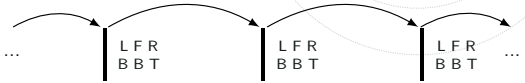
1. Conventional approaches encode all FoVs in one frame.



FoV-to-frame mapping

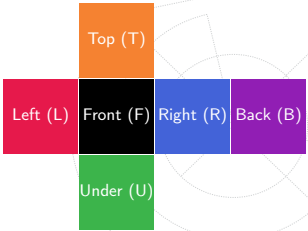


1. Conventional approaches encode all FoVs in one frame.

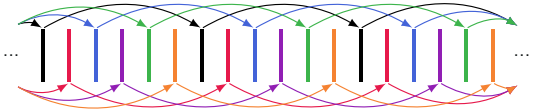


2. FoV-to-frame mapping is used in order to allow partial decoding (i.e., only few FoVs).

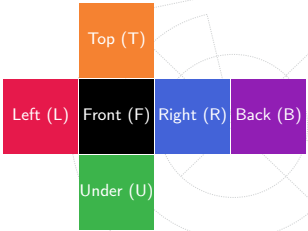
FoV-to-frame mapping



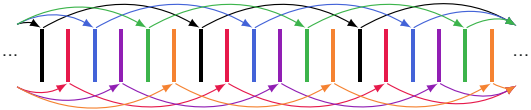
3. Temporal predictions are disabled across FoVs (only intra-FoV predictions).



FoV-to-frame mapping



3. Temporal predictions are disabled across FoVs (only intra-FoV predictions).



Adaptively decoding of FoVs is allowed!

Visual attention based quantisation

- The quantisation parameter (QP) for each FoV (f) is obtained based on the energy of the visual attention map (A):

$$E(f) = \sum_{\mathbf{p}} \mid A(f, \mathbf{p}) \mid^2, p \in \Omega$$

- The level of visual attention in each FoV is given by the relative energy of its attention map:

$$E_R(f) = \frac{E(f)}{\sum_{i=1}^{6} E(i)}$$

Visual attention based quantisation

Proposed approach

- FoVs (i.e., frame) with low level of E_R are assigned higher QPs.
- The ΔQP are calculated as follows:

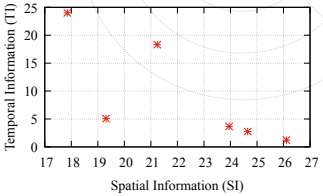
$$\Delta QP(f) = \alpha \left(1 - \frac{E_R(f)}{max\{E_R(f)\}}\right)$$

- $\Delta QP(f) = 0$ for the FoV with the highest visual attention.
- The α parameter controls the maximum QP variation.

Experimental setup - Test material

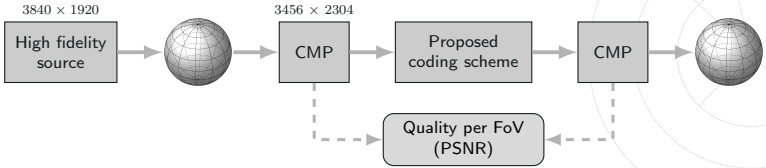
- 360° video sequences and attention maps: from Salient360 dataset.

Sequence	Description
Abbottsford	Dorm room with one student talking with moderate motion
Cockpit	Cockpit footage with high camera vibrations
PlanEnergyBioLab	Laboratory with several people moving around
PortoRiverside	Riverside images with two standing guys and low overall motion
TeatroRegioTorino	Orchestra concert with high different people playing instruments
Turtle	Two women helping a turtle return to the ocean



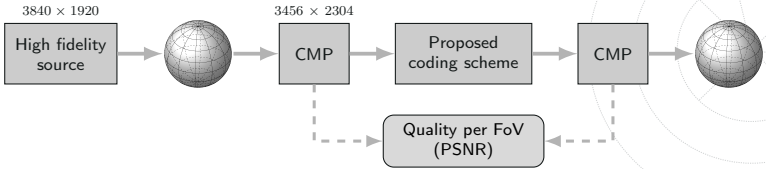
J. Gutiérrez, E. David, A. Coutrot, M. Perreira Da Silva, P. Le Callet, "Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360 contents", International Conference on Quality of Multimedia Experience (QoMEX), Sardinia, Italy, May. 2018.

Experimental setup - Test conditions



- VVC reference software (VTM) version 4.0.
- Spatial resolution: 3456×2304 pixels (CMP).
- Quality evaluation is performed between the original and reconstructed CMP.
- Quality of the most relevant FoV is considered.

Experimental setup - Tested methods



- Proposed frame-based FoV coding using visual attention **(Prop)**:
  - Two values of α were tested (5 and 10).
- Frame-based FoV coding using constant QP, i.e., homogeneous coding **(HC)**.
- Conventional CMP encoding, i.e., all FoVs in a single frame **(Ref)**.

Performance evaluation - Bjøntgaard metric

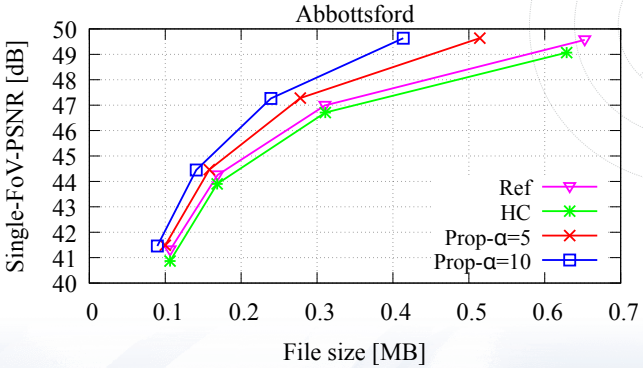
- Single-FoV quality vs total bitrate (all six FoVs).

Sequence	**BD-Single-FoV-PSNR** (FoV with higher E_R)		
	HC	Prop-$\alpha = 5$	Prop-$\alpha = 10$
Abbottsford	-0.33	0.69	1.35
Cockpit	-0.16	1.05	2.03
PlanEnergyBioLab	-0.37	0.54	1.06
PortoRiverside	-0.24	0.86	1.68
TeatroRegioTorino	-0.12	0.86	1.55
Turtle	-0.26	0.90	1.67
Average	-0.24	0.82	1.56

- Higher rate-distortion performance in the most relevant FoV is achieved.
- Increasing the value of α leads to a higher quality in the relevant FoV.

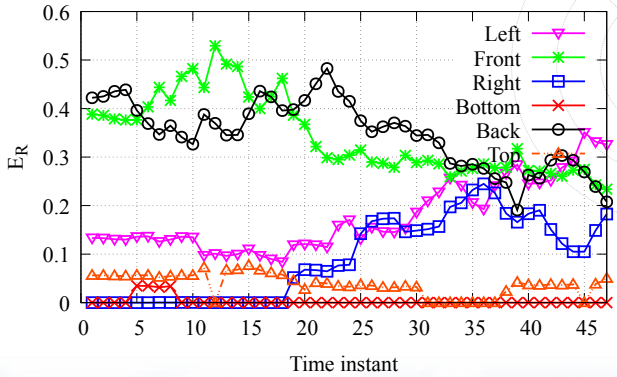
Performance evaluation - Bjøntgaard metric

- Single-FoV quality vs total bitrate (all six FoVs).

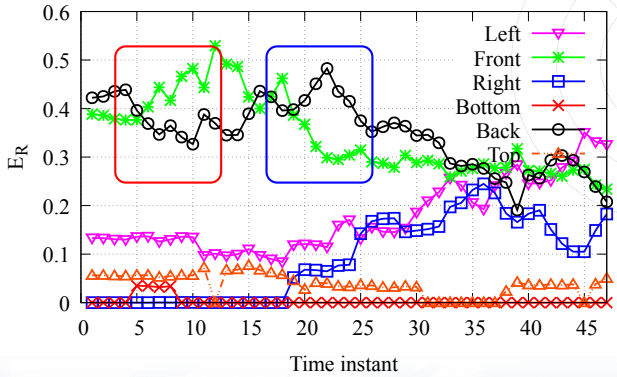


- Coding gains of the proposed method increase for higher bitstream sizes.

Performance evaluation - Temporal analysis

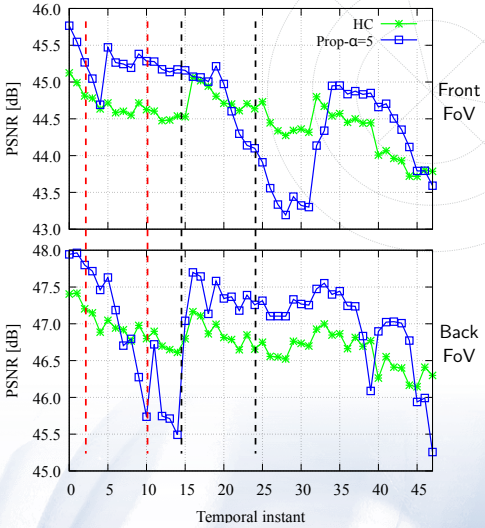


Performance evaluation - Temporal analysis



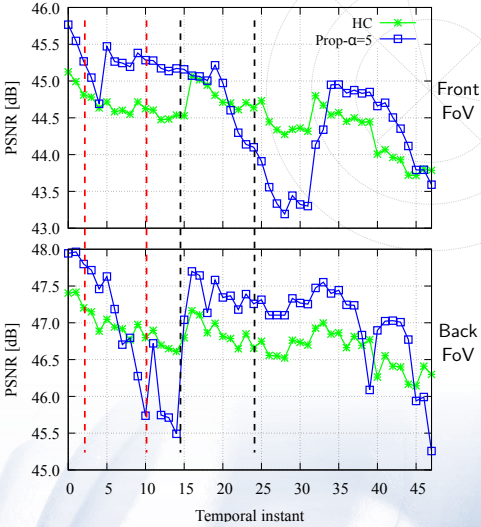
- Results show that the relevant FoV is not constant across all frames.

Performance evaluation - Temporal analysis



Performance evaluation - Temporal analysis

- The proposed method is able to adjust the quality of each FoV based on the energy of the visual attention map.



Conclusions

- A flexible coding approach for 360° video was devised:
  - Frame-based mapping enables independent streaming/decoding of each FoV.
  - Visual saliency based quantisation for efficient perceptual bitrate allocation.
- Results show that using visual saliency information allows for reduced bitrate while maintaining the quality of the relevant FoVs.
- The proposed coding method is an efficient approach to enable partial delivery and decoding of 360° video streams.

Future work

- Extend the proposed method using spatial scalability to dynamically adjust the FoV resolution/quality:
  - Extend VVC to include spatial scalability.
- Combine visual attention with object detection to define regions of interest with a narrow FoV:
  - Use both FoV and tile partitioning.
- Further investigate this architecture to develop smart 360° surveillance systems.

Thanks for your attention!

João Carreira
jcarreira@co.it.pt