Offline H.264 encoding method for omnidirectional videos with empirical region-of-interest

RÉMI SORMAIN
Offline H.264 encoding method for omnidirectional videos with empirical region-of-interest

Rémi Sormain

Supervisors: Assoc. Prof. Fredrik Kilander
Tiphaine Rousseau (RE’FLEKT GmbH)

Master Degree in Information and Communication Technology

School of Information and Communication Technology
KTH Royal Institute of Technology
Stockholm, Sweden
June 2017
Abstract

Panoramic virtual reality is an emerging technology that has recently gained the attention of both the research community and regular consumers. It allows the users to immerse themselves in omnidirectional videos with the help of a virtual reality headset: thanks to an increasing amount of affordable head-mounted-displays, any recent smartphone can offer a decent panoramic virtual reality experience. However, since omnidirectional videos are videos with a large field-of-view that covers the entire sphere around the camera, they require large resolutions and thus high bitrates.

This master degree project conducted at RE’FLEKT GmbH is an exploratory work that seeks to reduce the panoramic video bitrate. Because of the nature of omnidirectional videos, the user can only see a subpart of each video frame, and thus some zones of the video can attract more attention than others. The purpose of this study is to introduce the concept of region-of-interest encoding in panoramic VR. The main contribution is a method to encode panoramic videos in an H.264 video format stream with a space-variant level of details depending on the zones that attract the most the viewers’ interest. First, the region-of-interest are detected through a head-tracking module combined with a Gaussian attention model. Then, the reference video is encoded with the open source x264 encoder, with a quantization step adjusted to the region-of-interest information. The International Telecommunications Union standard subjective tests show that this method can perform better than classic H.264 encoding only in specific cases.

Keywords: virtual reality, 360 degree video, omnidirectional video, video encoding, gaze-contingency, panoramic virtual reality, region-of-interest.

Glossary

**API** Application Programming Interface. 25, 35
**AR** Augmented Reality. 7, 9
**CCR** Comparison Category Rating. 26, 39
**CGI** Computer-Generated Imagery. 26
**CPU** Central Processing Unit. 21
**DCT** Discrete Cosine Transform. 13
**DMOS** Differential Mean Opinion Score. 28, 31, 34, 35
**FOV** Field Of View. 9–11, 21–23
**HEVC** High Efficiency Video Coding. 12, 15, 16, 25, 35
**HMD** Head-Mounted Display. 7–11, 15, 17, 23, 24, 26, 27, 35, 37, 39
**ITU** International Telecommunication Union. 12, 16, 35, 39
**MB** Macro-Block. 13, 14, 16, 21, 25, 29, 33–35, 39
**MJPEG** Motion JPEG. 13
**PC** Pair Comparison. 26
**PDF** Probability Density Function. 24
**PR** Projected Reticule. 19, 21
**PSNR** Peak Signal to Noise Ratio. 16, 17
**QP** Quantization Parameter. 13, 16, 25, 29, 33, 34, 39
**ROI** Region-Of-Interest. 11–16, 19, 21, 24–27, 29–37, 39
**VR** Virtual Reality. 7–15, 17, 19, 24–27, 31, 35, 36, 39
Chapter 1

Introduction

Over the years, the desire to immerse oneself into a virtual environment has aroused great scientific interest, as early as in the 50s with pioneering prototypes such as the Telesphere Mask, created by Morton Heilig in 1957 and patented in 1960 [22]. The device consists of a head-mounted display (HMD) described by Heilig in his patent as a **telescopic television apparatus for individual use**, and it was able to play stereoscopic images, stereo sounds and even, according to the patent, "conveying to the head of the spectator, air currents of varying velocities, temperatures and odors". Heilig could only build a prototype, but figure 1.1 shows how the Telesphere Mask is a ground legacy work that inspired modern technologies that we nowadays associate with the field of Virtual Reality (VR).

Half a century later, a large range of means to mix virtual and real worlds have been developed, to which we generally refer to as a "virtuality continuum", a concept described by Paul Milgram and Fumio Kishino in *A Taxonomy of Mixed Reality Visual Displays* [27] as seen in figure 1.2. Indeed, one has to distinguish Augmented Reality (AR) from VR, for both are often introduced alongside. AR consists of bringing virtual elements into the real surrounding of the user, whereas VR intend to bring the user inside a virtual environment. Unlike AR, VR requires a realistic illusion in order for the user to feel present in the virtual world, a requirement divided into *Extent of Presence Metaphor* and *Reproduction Fidelity* scales in Milgram and Kishino’s study. Heilig explained already in 1962, in the patent of the Sensorama simulator [15] (a bigger version of the Telesphere concept), that his invention was especially conceived to improve the realism and thus, the impact of a virtual experience.

Although real-time computer generated graphics have been pushed further in the past years, real-life photo and video footages are still hard to beat on mobile platform when it comes to to realistic VR experiences. This might explain the economical blossom of the research area that is investigated in this paper: immersive panoramic videos, also called cinematic virtual reality. In the cinematic VR paradigm, the virtual environment presented to the user is not a real-time computer generated environment but an omnidirectional video or photo of a real environment, possibly enhanced with virtual objects and interactions. "Omnidirectional" means literally "in all direction", thus omnidirectional pictures refers to images that captures the whole field of view around a point in space, 360 degrees horizontally and 180 degrees vertically. Again, as it can be seen in Heilig’s work, the concept is not new, but VR technologies are finally reaching to a large audience, with big companies such as Facebook and Google providing more frameworks to develop omnidirectional immersive experiences ([7], [21],[2] and [3]).

1.1 Benefits and company’s interest

This section describes the motivation behind the presented exploratory work, together with the underlying context in which it was developed.
Figure 1.1: Figure extracted from the "Stereoscopic-television apparatus for individual use" patent [15], that we could describe as an ancestor of our modern Head-Mounted Display (HMD) virtual reality devices.

Figure 1.2: Representation of the reality-virtuality continuum of Paul Milgram and Fumio Kishino. This thesis focuses on cinematic VR, a technology that belongs to the right end of the spectrum (immersive display of virtual images [27]).
1.1.1 Stakeholder

This master degree project was developed at RE’FLEKT GmbH, a German company specialized in VR and AR for business applications, ranging from marketing to industry. They provide software solutions to other companies that take advantage of those technologies in their processes, for example:

- **REFLEKT One**, a tool to generate AR documentation for industrial machines or medical devices, integrated with standard documentation frameworks.

- **REFLEKT Remote**, a solution for remote customer assistance service supported by AR.

- **REFLEKT 360**, a panoramic VR storytelling platform that is further described in the following section.

1.1.2 Benefits

One of RE’FLEKT’s software solutions is REFLEKT 360 [30], a tool that makes possible for other companies to easily create mobile applications with custom panoramic and interactive experiences: the end user, possibly equipped with a VR headset such as a simple Google Cardboard [12] or the more advanced Samsung Gear VR [31], is able to view 360° medias (photo or video) and to interactively trigger reactions when gazing at pre-defined interaction points. This opens a wide range of applications:

- virtual tour of remote and potentially inaccessible locations: museums, caves, international space station, etc.

- demonstration and marketing for big products such as cars, houses and apartments etc.

- attending events such as concerts and theatrical performances.

- documentary and storytelling from panoramic footages. The German newspaper Süddeutsche Zeitung is using REFLEKT 360 to produce documentaries such as the one described in figure 1.3. See reference [35] for more information or to download the application.

- prototyping AR applications. pARnorama [6] demonstrates how panoramic VR can emulate expensive AR headsets such as Microsoft’s Hololens [26].

- games and other interactive experiences, for example *Fort McMoney*, a 2013 documentary realized by David Dufresne in the form of a VR cinematic serious game [8], [10].

Although REFLEKT 360 is the main focus of this study and shapes the constraints of the prototype, the results presented here can benefit the whole range of applications above.

1.2 Problem formulation

The problem at the heart of this study is the bitrate and byte size of the panoramic videos. Most smartphones now offer a Full HD (1920 per 1080 pixels) resolution to display, if not QHD (2560 per 1440 pixels) for high-end phones like the Google Pixel or Samsung Galaxy S7. Depending on the lenses used in the HMD, the limited Field Of View (FOV) (around 100 degrees for most Google Cardboard headsets) implies that only a subset of the pixels of the video is displayed on those high resolution screens. Thus, a comfortable VR experience requires high resolution medias, from Full HD to 4K to avoid pixelized and blurry results, and twice as much when it comes to stereoscopic spherical videos.

In the example given in figure 1.3, allowing several different panoramic points of view means having several high resolution 6 minutes videos for one single experience.
Figure 1.3: Screenshots of the Süddeutsche Zeitung android application, made with REFLEKT 360. (a) is in 2D mode, meant to be used without any HMD, whereas (b) has to be used with a headset such as the Google Cardboard. In this VR documentary - a rehearsal concert of the Münchner Philharmoniker, the audience can gaze at the camera interaction points to change the location of the camera, and thus have the ability to "move" in the scene and change the point of view. Note that both (a) and (b) result in the same Field Of View (FOV) thanks to the headset’s biconvex lenses, although not the same resolution.
Each of these are in average 350 Mb videos in low quality (1920 per 960 pixels), and about 500 Mb in higher quality (4086 per 2048 pixels). As there are 4 points of view, we have a total of 1.5 to 2 Gb of data needed for a 6 minutes monoscopic experience. It is then easy to imagine how several similar omnidirectional experiences require for the users to either stream or download several GB on their devices.

However, unlike flat videos, panoramic VR allows the user to see only a subset of the video, since the user’s FOV is limited by the HMD (usually to 100 degrees horizontally). The core idea of this exploratory work lies in the fact that some regions of the video frames are more attractive to the user. For instance in figure 1.3, the playing orchestra will probably be more interesting for the viewer than the empty seats in the room.

1.2.1 Research question

This study examines the possibility to measure the distribution of interest across each video frame and to encode omnidirectional video given this distribution. It means allocating more quality to the most viewed areas of the video, and less quality to the least viewed ones in an offline process. The choice of an offline process is further discussed in section 1.6. The research question is to determine the qualitative impact of space-variant offline compression using head-tracking data on immersive panoramic videos, compared to the classic H.264 video encoding technique.

1.2.2 Hypothesis

The hypothesis for the presented method is divided into two parts:

1. The users of panoramic VR do not have a uniform interest across the video, both spatially and temporally.

2. By allocating more quality to the regions attracting more interest, it is possible to obtain a better subjective quality than the regular H.264 encoding with the same bitrate.

1.3 Purpose

The purpose of this thesis is to propose and explain the process of head-tracking based offline video encoding when applied to panoramic experiences, and discuss its benefits and drawbacks.

1.4 Goal

The goal of this master degree project is to deliver a prototype implementation of the aforementioned process and evaluate its qualitative performances, so that RE’FLEKT has a proof-of-concept for providing better quality experiences with reduced video bitrate and file size (and thus reduced storage and bandwidth usage).

1.5 Research method

As this thesis attempts to answer to a known and practical problem, the method used here is applied research. The encoding prototype was developed based on previous works in the area of Region-Of-Interest (ROI) video encoding presented in section 2.2. The prototype provides an estimation of the interest across each video frame, and outputs an H.264 ROI encoded video.

Its performances were then evaluated in terms of achieved bitrate and subjective quality compared to classic H.264 encoding (with a similar bitrate). Both qualitative and quantitative results are assessed on a limited set of data, thus the research approach needed to answer the problematic is inductive. The qualitative results of the proposed method are collected through standard quality assessment tests described in
section 3.3 and provided by the International Telecommunication Union (ITU). Those results are transcribed into quantitative results: this analysis method belongs to the coding data analysis pattern methods described in Portal of Research Methods and Methodologies for Research Projects and Degree Projects [14].

1.6 Delimitation

The context given in section 1.1 limited the scope of the research presented here. A custom version of the REFLEKT 360 application was developed for the needs of the prototype, yet those changes had to be minor alterations because of the time constraints of this master degree project. In particular, it was chosen to use an offline pipeline, which means that the steps described in the method chapter could be triggered independently, as opposed to real-time encoding where most steps should happen sequentially. A real-time encoding technique would not only put strong requirements on the encoding performance but would also require a modification of the back-end endpoints, which was not possible for this project. Real-time techniques are however mentioned in section 2.2.

As further discussed in section 3.2, the target platforms of REFLEKT 360 also limited the encoding format to H.264, since more recent formats such as High Efficiency Video Coding (HEVC) are not supported by the video decoder on iPhones at the time of this work.

1.7 Ethics and Sustainability

This project does not focus on sustainability issues, although it could have a positive environmental impact by allowing less storage consumption. To be able to compress panoramic content as described in the method section, it is needed to collect head-tracking data. All test subjects were informed about this collection, and they could freely decline the collection. Moreover, the collected head orientation is averaged and anonymized.

Screenshots and short video sequences of existing panoramic experiences from the Süddeutsche Zeitung are used with their authorization.

1.8 Outline

The first chapter of this master degree thesis introduces the reader to the concept of panoramic VR and describes the problem that this project attempts to solve. It provides the context in which the prototype was developed and on which hypothesis it is based.

The second chapter presents the academic work in related areas, from the classic video encoding techniques, to the ROI encoding research area. It shows how ROI coding can save bitrate and thus why it can benefit panoramic VR.

The third chapter explains the ROI encoding process implemented in the prototype, starting from the interest heatmaps generation to the H.264 encoding. It also described how the perceived quality is evaluated.

The fourth chapter contains the results useful to evaluate the ROI encoding prototype. Those results are succinctly presented then discussed.

The fifth chapter concludes this thesis, summarizing the outcomes of the tests and pointing out directions for future work on the topic of panoramic VR.
Chapter 2

Background and related work

This chapter presents the related research areas that aim at solving similar issues. First a general overview of video encoding techniques is given, then the area of ROI encoding is investigated, first in terms of usages, then in terms of implementation, before the topic of ROI encoding evaluation is tackled. In section 2.5, a brief overview of the latest works in the area of panoramic VR is given.

2.1 Theoretical background on video encoding

Video encoding refers to the process of representing a moving visual media in order to store or transmit it. The resulting digital video is encoded to meet proper formats and specifications, in order to be then played by a video decoder. Because it belongs to data compression techniques, it is also referred to as video compression, but should not to be confused with a video codec, which relates to a specific implementation of a coding format specifications. Two broad types of video coding formats can be distinguished: Intra-frame and inter-frame coding formats.

2.1.1 Intra-frame video coding formats

Those formats are close to what a video is, i.e. a succession of frames. In other words, intra-frame coding is storing all frames of the video independently, as if they were single images. Motion JPEG (MJPEG) stand as an example of a coding standard where every frame in the video stream is a JPEG image: no temporal processing is performed outside of each picture (thus, the adjective intra).

For each frame, the JPEG coding process consists of the following main steps (in details in the JPEG still picture compression standard [39]):

- Decomposing the frame into 16x16 blocks called Macro-Block (MB) - sometimes also 8x8.
- Converting each MB of each color component (Y, Cb and Cr see [39]) to a frequency-domain representation, using a two-dimensional Discrete Cosine Transform (DCT). The obtained matrix contains the DCT coefficients.
- Quantizing each DCT coefficients uniformly. Each component is divided (in the frequency domain) by a constant for that component, called Quantization Parameter (QP), then rounded to the nearest integer. The dividers are taken from a quantization matrix: the bigger the QP, the closer to 0 for the rounded coefficients, the more information loss.
- Entropy coding of the quantized coefficients.

The advantage of intra-frame coding is the resulting quality of the media (as described in Low-bitrate motion JPEG using differential encoding [18]), and it is often used as an output format for video cameras, such as Apple ProRes codecs. But while having the benefit of a low decoding computational complexity, intra-frame
formats create heavy video file, which is not in line with the goal of this study. With 10427 frames in each video, the example in figure 1.3 would require each JPEG frame (4092x2048) to weigh less than 48 Kb in order to obtain the same video bitrate as in the H.264 version. And it would result in a blurry and full-of-artifacts video.

### 2.1.2 Inter-frame video coding formats

Those coding formats are build to exploit the temporal redundancies between frames to save a significant amount of video bitrate. It is based on the assumption that most consecutive frames within a sequence are very similar to the frames not only before but also after the current frame. One of the most widespread encoding pattern is H.264 (or MPEG-4 AVC for Advanced Video Coding) [40], with one of its key algorithm, motion compensation. This technique allows, given a reference frame, to predict if and where each MB of the reference frame will be shifted in the "neighboring" frame (it can be the following frame but also the previous one). It is then enough to encode this shift (motion vector) and only store the remaining difference between frames with the intra-coding technique previously described. H.264 still needs some frames to be entirely coded so they can be used as references: thus, some frames are intra-coded (called I frames), while the other frames are coded using motion estimations in one direction (P frames) or in both back and forward direction (B frames).

The advantage of the inter-frame format is the size of the encoded video compared to the intra-frame techniques, but it comes with a higher encoding and decoding complexity. Fortunately, as recent mobile devices have dedicated hardware decoding units to this purpose, this is not a major concern. It is thus the encoding format used in this thesis.

### 2.2 Usage and effects of region-of-interest encoding

Before the democratization of VR headsets and even before the H.264/AVC format, there had been research on the topic of non uniform video compression, depending on what the user is looking at. In Static and dynamic spatial resolution in image coding: an investigation of eye movements [33], Lew B. Stelmach and his team demonstrate the existence of a general agreement across human observers, both in viewing behavior and attracting regions in given regular video scenes, often referred to as ROI.

From those fundamental results, attempts to make use of the point of user attention can be classified according to two characteristics:

- **Empirical or predictive encoding.** The empirical one advises to use a panel of human viewer to measure and collect a representative set of ROI, while the predictive encoding relies on a static and algorithmic analysis of the frames.

- **offline or dynamic encoding.** The former refers to a classic encoding technique, when the media is compressed before it is played to the user, while the latter imposes a just-in-time encoding of the video stream.

In all cases those attempts are mainly gathered under the term of gaze contingency displays, because they take into consideration the distinctive characteristics of the human eye: rapid movements, non-uniform spatial resolution of the retina (called foveation encoding) etc. For example, M. Nyström and K. Holmqvist [29] have analyzed the perceptual effects of offline compression based on multiple free viewing records. Having designed an empirical offline technique, they managed to gain a minimum of 20% less bitrate with a perceived quality only slightly inferior to the original video clip. Interestingly, when the users have to assess the quality several times, the subjective quality is decreasing: as M. Nyström and K. Holmqvist point out, it is likely that several viewings imply a change in the gaze patterns of the users.

Another issue pointed out in [29] and in multiple studies (for example Perceptual Comparison of Empirical and Predictive Region-of-Interest Video [13]), is the collection of eye-tracking data. The eye tracking device (see figure 2.1) can be quite distracting for the testers, and the whole process long and costly. Thus, in their
Figure 2.1: An eye tracker, similar to the one used by M. Nyström and K. Holmqvist [29] to track ROI while users watch a video clip. However the experience of viewing video content through this device is arguably not natural.

study, Stephen R. Gulliver and Gheorghita Ghinea [13] demonstrate that predictive encoding can achieve quasi similar results when compared to empirical encoding. This is in particular very efficient for surveillance cameras [25] or video conferencing system, where faces and moving objects are automatically detected as ROI. Region-of-interest-based rate control scheme for high-efficiency video coding [23] gives an application example with video conferencing, based on HEVC reference implementation 10 (HM10), achieving better quality in the ROI while keeping a regular bitrate. However some other realms of application tend to produce deceiving results when no human input is given to compute the ROI, such as remote surgical application. In Region-of-interest video coding for enabling surgical telementoring in low-bandwidth scenarios [20], Sourabh Khire and his team point out that determining the proper ROI in surgical videos is better left to the high expertise of the surgeon.

It has to be kept in mind that the previously mentioned articles all relate to videos meant to be seen on a regular screen, with an empirical or computational model of gaze attention, while the problem tackled by this master degree thesis relates to panoramic frames which the user can never see entirely. A closer problem to this study could be the works of A. Kenny, and his team, who published an investigation [19] of eye gaze behaviors in first-person shooter games. Even though the study was not conducted with a HMD, first-person shooter games share the specific feature of having a virtual camera placed as the character’s head, so that the player can look around with a controller and feel as if he or she was the actual character. This corresponds to the fundamental idea behind panoramic VR: having the user feel present and involved in the virtual environment, with a similar eye-windowed view. Most VR applications feature indeed a reticle in the center of the screen, just as first-person shooter games do (see RE’FLEKT application figure 1.3). Kenny observed that 86% of the eye’s fixation time is spent around the screen’s reticle. This result was later confirmed and expanded in Depth-of-Field Blur Effects for First-Person Navigation in Virtual Environments [16], with about 55% of the time spent gazing at less than one fourth of the screen’s height from the reticle. More details are given in section 3.1.2.

2.3 Region-of-interest encoding implementation

When it comes to the actual implementation of the ROI coding, we can discriminate three general types of implementations:
• New encoding formats or customized codecs that require a customized decoder before displaying each frame, as presented in *A real-time foveated multiresolution system for low-bandwidth video communication* [11]. This is mostly avoided, since using a custom decoder limits the adoption among users. Moreover, it requires a fine knowledge of video encoding parameters to propose a custom media format or custom decoder that outperforms the current standards such as HEVC [34], at least for standard videos.

• **Codec-independent techniques with either or both pre-processing and post-processing steps.** This is the usual implementation, generally with a post-processing step that receives both the decoded video stream and the ROI information to compose the final frame - more details can be found in [24] or [25]. Nyström and Holmqvist [29] use a pre-processing step to blur the frames before feeding them to the video codec. The advantage is that the ROI video coding can theoretically be done with any video encoding format, although not all will have the same performances.

• **Codec-dependent techniques compliant with standard encoding formats.** While those are limited to what the codec implementation offers, they have the benefit of not requiring any changes in an existing application, and enable a wide compatibility. This is used in [20] with the open x264 codec: QP values for each MB are increased or decreased on top of the rate control algorithm. The video conferencing example in [23] is more intricate but relies on a modification of the rate control algorithm used in HEVC, thus still producing compliant HEVC stream.

2.4 Region-of-interest quality assessment

As pointed by Nyström and Holmqvist [29], the evaluation of foveated or ROI video encoding with traditional methods for video quality assessment gives biased results. On one side, there are automatic objective methods such as the measure of the Peak Signal to Noise Ratio (PSNR), defined for two images \(I_o\) and \(I_r\) of size \(m \times n\) as:

\[
PSNR = 10 \cdot \log_{10} \left( \frac{d^2}{MSE} \right) \tag{2.1}
\]

with \(MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I_o(i,j) - I_r(i,j))^2\)

\(d\) is the maximal value for a pixel, which is 255 in the common 8 bits case. The PSNR is useful for measuring the similarity of a compressed image relative to the reference signal, pixel wise. However in the case of ROI coding, some regions are explicitly more important than others, which is not taken into account in the formula 2.1. Thus, most ROI studies such as *Region-of-interest-based rate control scheme for high-efficiency video coding* [23] or *Codec independent region of interest video coding using a joint pre- and postprocessing framework* [24] measure not only the global PSNR of the frame, but also the PSNR reduced to the ROI. This does not represent a satisfying metric for two reasons: first, the PSNR is a purely spatial measure, and it does not take into account the temporal characteristics of a video (see formula 2.1). Second, if the PSNR is reduced to the ROI pixels, the resulting score ignores the overall perceived effect of having varying spatial resolution in a same frame. For example, it has been repeatedly pointed out that sharp ROI edges harm the perceived quality of the video, an issue that PSNR barely captures.

On the other side and to improve the assessment of perceived video quality, subjective evaluation methods are defined, such as evaluation recommendations from the ITU P.913 [38] and P.910 [36]. Applying those standards enables comparison between research studies, yet the immersive videos studied here are not conventional video content. Both P.913 [38] and P.910 [36] present several test methodologies, one of them being a pairwise comparison of video sequences. The method description includes
setup parameters, the assessment procedures as well as the descriptions and scales used by the test subjects to describe the perceived quality. But unlike P.910 [36], ITU-R P.913 allows to deviate from test conditions that don’t fit innovative viewing technologies such as VR (e.g. lighting, types of stimuli (video) etc.).

This is a crucial point, as viewing conditions of omnidirectional videos are completely different from regular TV or mobile videos viewing, as described in the study A Framework to Evaluate Omnidirectional Video Coding Schemes [41]: the viewer will see the video not on a flat screen, but mapped onto a sphere, and this is this final rendering that should be evaluated. The framework presented in [41] is a first step toward objective evaluation, but it still uses PSNR on the video, while subjective tests allow an "in conditions" quality assessment, through the lenses of the HMD and not just from the bits of the video (see the analysis in 4.4).

Thus, the P.913 recommendation appears to be the most appropriate evaluation method to evaluate the outcomes of this work.

2.5 Latest work about encoding for panoramic VR

As mentioned in the introduction, VR and especially panoramic VR is reaching a large audience thanks to the adoption of panoramic contents by large companies like Google of Facebook, which raises concerns for the bitrate of the needed video. Engaging Immersive Video Consumers: Challenges Regarding 360-Degree Gamified Video Applications [4] points out the growing interest as well as the file size challenge for panoramic VR applications. There are valuable contributions to improve the streaming of panoramic content. One of those solutions uses MPEG DASH SRD (spatial relationship description [28]) and consists of splitting the video in tiles in order to be able to request only necessary tiles. The process is detailed into Adaptive 360 VR Video Streaming: Divide and Conquer! [17]. However this does not change the overall bitrate and size of the video, as REFLEKT 360 applications can require to download (and not stream) content.

Some other recent works tackle the video projection (see Ultrahigh-Resolution Panoramic Imaging for Format-Agnostic Video Production [32] for an extensive overview of panoramic imaging). An equirectangular projection is generally used [32], whereas it has been regularly described as a sub-optimal solution by large content platforms such as Facebook [21] and Google [2], and tends now to be replaced by a cubic projection, as explained by Google’s engineer Chris Brown [2]. Those improvements do not stand in contradiction with this study and could be combined to reach better perceived quality and lower file sizes.
Chapter 3

Method

The idea behind the presented prototype lies in the fact that, unlike conventional videos, panoramic VR videos require the viewing device to access the user’s head orientation. This is necessary in order to display a subset of the video projected onto a sphere. It is thus possible to retrieve this orientation information and use it to delimit ROI within the viewed omnidirectional video.

This section presents the panoramic video encoding process based on empirical ROI determination. First, the collection of ROI through head-tracking is described, followed by the detailed video encoding using the open-source x264 library. The metrics and test environment used to assess the performances of the prototype are presented in a third part. Figure 3.1 gives an overview of the proposed prototype.

3.1 Empirical head-tracked regions-of-interest

This section presents the first part of the empirical ROI encoding prototype (the upper part of the diagram in figure 3.1) which aims at producing an attention heatmap for each frame of the video.

3.1.1 Head-tracked points collection

The first step is the collection of head orientation during the video playback. All VR applications are capable of computing the head orientation from the phone’s gyroscope. In this prototype, the set of Euler angles describing the head orientation is used to compute the projection of the user’s reticule onto the video frame. Since REFLEKT 360 content as well as most cinematic VR content [32] use an equirectangular projection, the described prototype uses this specific projection, although the ROI encoding method could easily be extended to all projections such as cubic projection.

In the case of the equirectangular projection, it is trivial to map a 3D orientation to the projection 2D coordinates. If the latitude \( \varphi \) and longitude \( \lambda \) (both in radians) are defined as in figure 3.2, those angles can be mapped to 2D coordinates \((x, y)\) when the center of the frame corresponds to angles \((0, 0)\):

\[
x = \frac{\lambda}{2\pi} \quad \text{and} \quad y = \frac{\varphi}{\pi}
\] (3.1)

Those coordinates \((x, y)\) correspond to the projection of the reticule on a video frame (figure 3.2 (b)), and will later be referred to as the frame’s Projected Reticule (PR) point.

In this prototype, the PR point was recorded for each frame. As most VR applications, REFLEKT 360 uses a 3D engine to project the video onto a 3D mesh (a sphere in most cases, see [32]) : the issue to consider is that the frame-rate of the 3D engine does not match the frame rate of the video, and thus the head position is not recorded synchronously with the video frame, as illustrated in figure 3.3 : The
Figure 3.1: Overview of the prototype architecture. In a production environment, the QP offset generation (and the following steps) could happen in a periodic manner or after a predefined number of attention heatmaps is aggregated.

Figure 3.2: Geometrical representation of the equirectangular projection.

(a) The spherical coordinates (latitude $\varphi$ and longitude $\lambda$) of the reticule, representing the head orientation.

(b) The corresponding equirectangular projection on the frame’s plane.
solution is to compute an average PR point when the video frame rate is inferior to
the engine’s one, and to interpolate PR points in the opposite situation.

Indeed in most cases, the engine frame-rate is higher than the video frame-rate,
as with the video frame 0 in figure 3.3. Two gaze points will be recorded during
engine frame F0 and F1, then averaged to have one average gaze point for the video
frame 0. In this prototype, the engine frame F2 was not included in the average for
video frame 0, but will be associated with frame 1. In a more unusual situation,
the engine frame-rate could temporarily become smaller (the display thread could be
slowed down by a sudden heavy Central Processing Unit (CPU) usage), and thus the
gaze point in video frame 2 will consist of an interpolation between the gaze points
in the engine frames F2 and F3.

As a result, the head-tracking module outputs the coordinates of the viewer’s
PR for each video frame. Since its coordinates \((x,y)\), in equation 3.1, represent the
coordinates as a fraction of the frame’s width and height, it is straightforward to
compute the corresponding coordinates in pixels or in MBs.

### 3.1.2 Attention model

Although the head-tracking module delivers a good approximation of the ROI, this
information is not as precise as with a gaze tracker. It is only known which subset of
the frame the user is seeing (the area around the PR point, delimited by the FOV),
and not exactly the point that attracts interest.

But instead of considering the full view as of equal interest for the viewer, the
ROI can be further refined under the assumption that the viewer’s attention is mainly
focused on the reticule. According to the investigation of A. Lécuyer and his team [16],
users exploring a virtual environment from a first-person perspective spend around
55% of the time gazing in a zone centered on the reticule that is 8 times smaller than
the total surface of the screen. To compute the user’s attention across the screen, this
prototype uses a 2D Gaussian model based on A. Lécuyer’s results: here, the reticule
is the output of the head tracking module (PR point), and thus it can be assumed that
the actual point of interest has 55% chances to be in a zone around the head-tracked
PR point which is a quarter of the entire screen height high (respectively, a quarter
of the entire screen width wide, see figure 3.4). In probabilistic words, let \((X, Y)\) be
a set of two continuous random variables representing the screen space coordinates of
the point actually gazed at by the viewer, relative to the reticule in the current screen.
We can normalize those coordinates into \((x, y)\), so that \((x,y) = (1,1)\) corresponds to
the top right corner of the screen and \((x,y) = (-1,-1)\) to the bottom left one:

\[
    x = \frac{2X}{FOV_{width}}
\]

\[
    y = \frac{2Y}{FOV_{height}}
\]

Then A. Lécuyer’s results can be reformulated:
Figure 3.4: Geometrical representation of A. Lécuyer’s results [16]: In the screen space centered around the reticule \( R \), the point \( G \) gazed at by the user has 55\% chances to be in the gray area \( A \).

\[
P(|x| \leq \frac{1}{4}) \approx 0.55
\]

\[
P(|y| \leq \frac{1}{4}) \approx 0.55
\] \hspace{1cm} (3.3)

From the conditions in equation 3.3 and using normal distribution tables [1], the standard deviation \( \sigma \) can be inferred to model the user attention across the screen:

\[
\sigma \approx 0.33
\]

Figure 3.5 shows the resulting distribution of interest across the normalized screen space (reticle in \((0,0)\), top right corner in \((1,1)\)), while figure 3.6 gives a example of this distribution projected in a video frame. The level of attention, or interest, is normalized so that the highest interest corresponds to 1 and the lowest, 0. Since the Gaussian distribution is projected using the an equirectangular projection, the distribution is distorted towards the pole: the distortion factor (the horizontal sampling density) at latitude \( \varphi \) is given by \( \frac{1}{\cos(\varphi)} \).

Of course this distribution is a first and naive approximation of the gaze point on the screen, this will be discussed in section 4.4. However the Gaussian distribution is the best first approximation model to represent the user’s gaze point. Gaussian models are indeed often used to describe the coordinates of a 2D point, for example, the dispersion of a physical gunshot from a target [9] - unlike an exponential model which is more suited to describe time-based events.

In conclusion, this module produces an attention heatmap for each viewing and for each frame of the video, using three inputs:

- the position of the gaze point in the frame (obtained from the head-tracking module).
- the attention model, which, in this prototype, is implemented as a Gaussian distribution around the gaze point.
- the FOV of the device, which delimits the subset of the frame that the user is seeing.
Figure 3.5: Representation of 2D (or bivariate) Gaussian distribution used to model the user attention across the HMD's screen. $x$ and $y$ are defined in equation 3.2.

Figure 3.6: Interest distribution for a specific frame and a single record. Darker shades denote a higher level of interest, while the surrounding black delimits the part of the frame currently displayed, the FOV.
Figure 3.7: Results of the attention heatmaps aggregation. (a) shows the aggregation for a specific frame and 15 viewings, which is overlaid in (b) on the corresponding frame in the Philharmonie (see table 3.2) video. The viewers, at that specific point in time, are mainly looking at the orchestra and its chief, although some of them are looking sideways in the audience.

### 3.1.3 Attention heatmap aggregation

The strength of the empirical ROI determination lies in the many recordings that can be gathered for a single video. In this prototype, heatmap aggregation simply consists of adding up attention (or interest) heatmaps from several viewings and normalizing them in a Probability Density Function (PDF). Figure 3.7 gives an example of the aggregation results.

Using this method, the costly and unnatural data collection disadvantages encountered with a gaze tracker (pointed out in [29] and [13]) become irrelevant, as head-tracking is an inherent feature of VR HMDs. In this prototype, those data were collected and recorded locally, but a production environment could collect the head-tracking information through an existing analytics system, as Facebook is already doing for its 360 content [5]. This would enable the collection of thousands of viewings and help reach a cheap and seamless empirical ROI detection.

### 3.2 Video offline encoding

The next step in the process shown in figure 3.1 is the actual video encoding. For both time constraints and the need to be implemented within REFLEKT 360, this prototype has to use video encoding standards and cannot use any post-processing step on the decoder side, as this would mean modifying the viewer application and performing tests over many hardware devices (phones) to ensure that the quality of the decoding is not harmed by the introduced algorithm. This also establishes the need of using an offline process, such as the one described in [29]: a dynamic ROI encoding would require an exchange of information between the back-end video
services and the viewing application with high reliability, which is out of the scope of this master degree project.

This reduces the possibilities to offline, codec-dependent techniques described in section 2.3. As briefly presented in section 2.1, there are many encoding formats available, yet not all can be natively decoded with a dedicated hardware resource. In particular, the HEVC standard used in [23] is still not supported by most iPhone platforms, which leads to the choice of H.264 for this prototype, an encoding format which, while acknowledged as less efficient than H265/HEVC or Google’s VP9, is widely supported across many platforms.

The presented prototype uses x264, a free H.264 compliant encoder that exposes a convenient Application Programming Interface (API) for ROI encoding. Indeed, x264 can be provided with a QP offset matrix for each frame. Each macroblock in the frame corresponds to a quantizer offset in the matrix, which will be added to the QP decided by x264’s rate control algorithm. A negative value leads to a smaller QP and better quality, and vice versa, as described in section 2.1. A similar method is used in Region-of-interest video coding for enabling surgical telementoring in low-bandwidth scenarios [20] to encode surgery videos.

To obtain those quantizer offsets, one simply has to define minimum $QP_{\text{min}}$ and maximum $QP_{\text{max}}$ values, and determine each QP offset from the aggregated attention heatmap of the previous module. Let $A$ be the attention matrix, where $A_{ij}$ is the normalized aggregated attention level for the MB of coordinates $(i,j)$, and $O$ the matrix of QP offsets:

$$\forall (i,j) \in \text{frame}, O_{ij} = (QP_{\text{min}} - QP_{\text{max}}) \times A_{ij} + QP_{\text{max}}$$

Determining $QP_{\text{max}}$ and $QP_{\text{min}}$ is not the core problematic of this work. Just as with classic encoding techniques, reaching a specific target bitrate can be achieved with several passes. In the following tests, the following values were used, as they offered similar bitrate gains as in [29]:

- $QP_{\text{max}} = 4$
- $QP_{\text{min}} = -2$

The resulting bitrate is documented in chapter 4 and depends on x264’s QP decisions, and thus is different for each video. Using raw video in pixel format YUV420 and the generated QP offsets as inputs, the H.264 encoding module outputs a raw H.264 stream that is then muxed (combined) with the original audio stream (AAC) within an MP4 container (see figure 3.1). The advantage of using a standard codec such as x264 is that all H.264 optimizations and quality preservation policies are used, for example the QP step limit, which prevent noticeable quality jumps between frames by restricting the range of reachable QP for a MB between two consecutive frames.

3.3 Evaluation

This section describes the experimental setup that was assembled to verify the hypothesis, by verifying that empirical ROI encoding using the proposed method produces better perceptual quality than standard encoding. In order to assess the performances of the proposed encoding scheme, a two-phases test process was needed:

1. The first phase is the recording session used to collect attention heatmaps as described in section 3.1.

2. The second phase is the quality assessment session, where test users were asked to evaluate and compare the quality of the videos.

3.3.1 Test environment

To perform those two phases, the commercially available Homido VR headset was used with a recent smartphone described in table 3.1, together with REFLEKT 360
application. In both phases, most users (83%) had no previous VR experience and were given a short introduction within the REFLEKT 360 demonstration application. The Homido HMD let them adjust the distance between both lenses, as well as the distance from the screen to the lenses, allowing far- and short-sighted viewers to use the headset without additional glasses.

<table>
<thead>
<tr>
<th>Smartphone</th>
<th>OnePlus 3T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen resolution</td>
<td>1920x1080 pixels</td>
</tr>
<tr>
<td>Screen type</td>
<td>5.5 inch AMOLED</td>
</tr>
<tr>
<td>GPU</td>
<td>Adreno 530 (624 MHz)</td>
</tr>
<tr>
<td>CPU</td>
<td>Snapdragon 821 (2.4 GHz)</td>
</tr>
<tr>
<td>RAM</td>
<td>6GB LPDDR4</td>
</tr>
<tr>
<td>Operating system</td>
<td>Android 7.0</td>
</tr>
<tr>
<td>Headset and FOV</td>
<td>HomidoV2, 100°</td>
</tr>
</tbody>
</table>

Table 3.1: Hardware specifications used in both test phases.

Five video clips extracted from publicly available 360° videos (on Youtube) and REFLEKT 360 application were used to answer the problem stated in section 1.2. As recommended in the ITU-T P.913 standard [38], they were specifically chosen because they form a representative set of various usages of immersive videos with different characteristics: resolution, bitrate or more subjective traits, such as the presence of moving objects. Those components are gathered in table 3.2. In particular, Warcraft is a Computer-Generated Imagery (CGI) video clip, and it shows the advantage of using offline generated video rather than dynamically generated 3D environments, which are still behind what ray-tracing techniques can accomplish. The sequences feature a fading at the beginning and at the end of the video, which leaves between 20 and 25 seconds of non-black frames: this is longer than the recommended stimuli duration in [38], but during pilot tests, participants raised the concern that the samples were too short to evaluate their subjective quality.

All participants could, in both phases, either stand or sit depending on their preferences. All of them wore headphones so that they could listen to the audio soundtracks accompanying the videos. The test environment is thus fairly close to what a VR consumer could experience in real life conditions.

3.3.2 Recording session

During the first phase, 16 volunteers from age 20 to 61 (mean 29.5 ± standard deviation 13.68) were asked to watch the original video samples within the described test environment 3.3.1. No specific instruction was provided as the goal of this step is to achieve a natural viewing of the video. Most volunteers chose to view some clips several times, which led to, on average, 23 attention heatmaps per video to aggregate (117 recordings in total). This corresponds to the actual usage of the REFLEKT 360 application, since users watch a single omnidirectional experience several times.

3.3.3 Encoding and quality assessment session

Once all attention maps were collected, the encoding pipeline shown in figure 3.1 was used to re-encode the sample videos on a regular computer. Since this is offline encoding, no specific attention was paid to encoding time and/or computational complexity optimizations. The first quantitative metric used to evaluate the prototype is the output video bitrate. In order to have a comparison reference, a second video was encoded with x264 with standard parameters, in two passes so that it has the same bitrate as the output video from the proposed method. To evaluate this prototype according to standard recommendations [38], a Pair Comparison (PC) method was designed, also known as Comparison Category Rating (CCR). This implies that the test sequences are presented in pairs, one being the reference sequence encoded through a classic x264 process, the other encoded through the ROI encoding pipeline.
Table 3.2: Video sequences used to assess the quality of the proposed method.

<table>
<thead>
<tr>
<th>Video</th>
<th>Resolution</th>
<th>Frames</th>
<th>FPS</th>
<th>Video bitrate</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpes</td>
<td>3840x2048 pixels</td>
<td>771</td>
<td>25 Hz</td>
<td>7.34 Mb/s</td>
<td>At night outside, still with moving starry sky</td>
</tr>
<tr>
<td>Maquinna</td>
<td>3840x1920 pixels</td>
<td>926</td>
<td>30 Hz</td>
<td>15.50 Mb/s</td>
<td>Indoor, still with moving characters</td>
</tr>
<tr>
<td>Philharmonie</td>
<td>4096x2048 pixels</td>
<td>929</td>
<td>29.97 Hz</td>
<td>6.98 Mb/s</td>
<td>Indoor, still with still characters</td>
</tr>
<tr>
<td>Sharks</td>
<td>3840x2048 pixels</td>
<td>924</td>
<td>30 Hz</td>
<td>17.00 Mb/s</td>
<td>Underwater, slightly moving with moving fishes</td>
</tr>
<tr>
<td>Warcraft</td>
<td>2560x1440 pixels</td>
<td>953</td>
<td>30 Hz</td>
<td>7.87 Mb/s</td>
<td>Outdoor, flying above still landscape, CGI</td>
</tr>
</tbody>
</table>

The viewer was then asked to judge the quality of the test sequence compared to the reference sequence using the ITU-R BT.500-13 comparison scale reported in table 3.3.

Seven volunteers, aged from 21 to 61 (mean 30.4 ± standard deviation 14.9) took part in this second phase (in the same test environment 3.3.1). This is less than the recommended 15, yet as advised in ITU-R BT.500-13 [37], it is acceptable for a study of exploratory nature, and over the absolute minimum of four [36]. Out of the seven, four participants were volunteers from the first phase and already knew the video samples, but all viewers followed the same instructions. This reflects the actual usage of VR applications where users naturally want to explore an omnidirectional experience again.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>Much worse</td>
</tr>
<tr>
<td>-2</td>
<td>Worse</td>
</tr>
<tr>
<td>-1</td>
<td>Slightly worse</td>
</tr>
<tr>
<td>0</td>
<td>The same</td>
</tr>
<tr>
<td>1</td>
<td>Slightly better</td>
</tr>
<tr>
<td>2</td>
<td>Better</td>
</tr>
<tr>
<td>3</td>
<td>Much better</td>
</tr>
</tbody>
</table>

Table 3.3: ITU-R BT.500-13 [37] double stimulus comparison scale used in pair comparison subjective tests. With this scale, a participant can indicate to the tester how he or she subjectively evaluates a video sequence compared to a reference one. In this study, the scale is used to compare a regular H.264 encoding with the ROI encoding technique.

In all cases, participants tested the prototype with the following protocol:

1. The participant is explained the purpose of this study, and what is expected from him or her: in particular, they are given the scale 3.3. They are informed about data collection, but they are not aware of the specific details of the ROI encoding process, as it might influence their natural gaze patterns (see [29]). The user is encouraged to watch the video as naturally as possible.

2. With the HMD on, the participant runs through a test session so that he or she understands how the rest of the session will take place. This test is also used to verify if the user presents any hearing or visual impairments or disabilities before running the actual assessment.

3. Then for each video sample in table 3.2, the user is presented the two sequences one after the other, and is asked to rate the second one with the scale 3.3. This assessment is recorded and the participant is presented another pair of sequences.

4. End of the test.
Every five minutes the participant had a break, and was able to drink and rest in order to avoid fatigue and boredom effects. They could also freely share comments on the experience, which will be taken into account in the discussion section 4.4. The video pair sequences presentation order was randomized for each subject, as was the order in each pair: this increases the reliability of the collected votes.

All ratings for each video are then aggregated in a Differential Mean Opinion Score (DMOS). As described in ITU:P.913[38], the DMOS score of an encoded sequence is given by equation 3.5: the randomized order was of course removed, and subjects who saw the reference video (regular H.264) second had their opinion multiplied by $-1$ to obtain a proper DMOS.

$$DMOS = \frac{\sum_{n=0}^{N} \delta_n R_n}{N}$$

where:

- $N$ is the number of participants
- $R_n$ is the rating of participant $n$, $n \in [0, N]$ according to the discrete scale 3.3.
- $\delta_n = \begin{cases} 1, & \text{if participant } n \text{ saw the reference video first} \\ -1, & \text{otherwise} \end{cases}$
Chapter 4

Results

This section explores the results of the experiment described in the previous section 3.3, first the objective, then the subjective ones.

4.1 Objective results

The first quantitative results to examine in this exploratory work are the impact of ROI encoding on the video bitrate, as this is the crux of this degree project. On average, with $QP_{\text{min}} = -2$ and $QP_{\text{max}} = 4$, there is a 30.4% gain in the bitrate, with a low standard deviation of 3%. The output bitrates are detailed in table 4.1.

<table>
<thead>
<tr>
<th>Video sequence</th>
<th>Original bitrate</th>
<th>Obtained bitrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpes</td>
<td>7.34 Mb/s</td>
<td>5.18 Mb/s</td>
</tr>
<tr>
<td>Maquinna</td>
<td>15.50 Mb/s</td>
<td>10.70 Mb/s</td>
</tr>
<tr>
<td>Philharmonie</td>
<td>6.98 Mb/s</td>
<td>5.09 Mb/s</td>
</tr>
<tr>
<td>Sharks</td>
<td>17.00 Mb/s</td>
<td>11.00 Mb/s</td>
</tr>
<tr>
<td>Warcraft</td>
<td>7.87 Mb/s</td>
<td>5.59 Mb/s</td>
</tr>
</tbody>
</table>

Table 4.1: bitrate obtained for each ROI encoded video with the parameters $QP_{\text{min}} = -2$ and $QP_{\text{max}} = 4$, for each test sequence.

An example of the resulting ROI encoding compared to regular H.264 compression is given in figure 4.5, for one frame of the Philharmonie experience. The green overlay illustrates how some MB receive lower QP values in the ROI encoded frame. As emphasized in the method section, the attention map is used to locally add QP variations on top of x264’s decisions and limits, thus the QP values are not necessarily matching the interest heatmap.

Figure 4.3 shows the aggregation of all frames’ attention heatmaps per video. This representation implies a loss in information, as it is not as good as a video showing the ROI for each frame, yet two points can be observed:

1. In all cases the attention is mainly concentrated around the center of the frames
2. Apart from this common central point, video sequences present different interest distribution.
Figure 4.1: Ratio between the original H.264 bitrate and the ROI encoded video one, for each test sequence, with the parameters $QP_{\text{min}} = -2$ and $QP_{\text{max}} = 4$.

Figure 4.2: Evolution of the bitrate ratio between the original H.264 bitrate and the ROI encoded video, depending on the maximum $QP_{\text{max}}$ value, for the clip Warcraft and with $QP_{\text{min}} = -2$. 
Figure 4.3: Aggregated interest heatmaps for all frames and all viewings, for each test sequences.

4.2 Perceived quality

According to the protocol described in the method section 3.3, the ROI was subjectively evaluated by non-professional users. The results are aggregated in a form of a DMOS in figure 4.6. The DMOS score corresponds to the scale defined in table 3.3 and to the equation 3.5. An insight of the resulting quality difference between ROI coding and regular coding is given in figure 4.4.

In average, the DMOS score in figure 4.6 does not depict a clear preference for either the ROI-encoded or the regular H.264 encoded version. Only the Sharks clip shows a marginally higher score for the ROI encoded version, but the average score remains below the slightly better evaluation in the scale 3.3. Figure 4.4 shows that the difference between the two versions in the ROI is indeed moderate.

4.3 Analysis

The aggregated interest heatmaps in figure 4.3 confirms the main hypothesis of this study: As expected from the previous works in related areas (section 2.1), panoramic VR users do not have a uniform interest across omnidirectional videos. Some zones attract the viewer’s attention more than others: it seems that the centers of the frames, in all sequences, are always attracting attention. This is mainly due to the convention that the center of the frame correspond to the head orientation $(\lambda, \varphi) = (0, 0)$ (latitude and longitude, see figure 3.2). Thus, most omnidirectional videos are composed so that the characters and other points of interest are in the center of the frame at the beginning of the video. When those points of interest are not moving, like in the Philharmonie video, this leads to an interest concentrated on the middle of the frame for all frames. Conversely, the Sharks sequence features sharks moving around the viewer, thus the ROI is spread over the frames in a similar way.

This explains why video sequences are watched with different viewing pattern, while having this common central point of interest. In the Alpes sequence, the par-
Figure 4.4: Zoom inside a frame of the *Philharmonic* sequence, inside a ROI zone. (a) is encoded with the regular H.264 method while (b) used the ROI coding process. The two versions are rather similar except for some artifacts that appear on the classic H.264 frame (hand and nose circled in blue).
Figure 4.5: Quantization Parameter (QP) values in a frame of the *Philharmonie* sequence. (a) is encoded with the regular H.264 method while (b) used the ROI coding process, (c) the corresponding interest heatmap. Greener MBs correspond to lower values of QP and thus higher quality.
Figure 4.6: Differential Mean Opinion Score (DMOS) results from the subjective tests comparing H.264 regular encoding (reference) to ROI encoding, with the associated standard deviation. According to the scale 3.3, higher DMOS values denote a preference for the ROI encoding, while negative values indicate a lower perceived quality than the classic H.264 encoding.

Participants tended to look on the upper part of the frame, as this experience presents a starry sky, while Warcraft features a flight over a town, thus users had more interest in looking down, at the lower part of the frame.

Moving to the perceived quality, the figure 4.6 reveals dissimilar results that demonstrate how ROI coding technique can have a positive as well as a negative impact on the video quality. In particular, the fishes in the Sharks sequence attracted the user’s focus in most cases, according to the aggregated attention heatmap. Thus, the MB composing the fishes received a lower QP, on top of x264’s rate control decisions. This enabled a slightly better perceived quality, as denoted in figure 4.6. On the opposite side, Alpes featured a still landscape at night, where attention was not all gathered on a few points, but rather dispersed across the video. With few test subjects and relatively few aggregated heatmaps, this led to participants gazing at areas with high QP values and thus, lower details.

To explain the relatively low improvement over regular H.264 encoding, one has to take the impact of movements in the test video into account, as developed in section 2.1. When a video feature a predictable movement, such as the camera moving forward in Warcraft, or the still background from Philharmonic, a lot of MB in a frame are recognized as MB from a previous (or following) frame, and thus encoded only as reference, and not with a quantized frequency domain representation. For those areas in B or P frames (called slices), increasing the QP value of a MB has no effect and the quantization only depends on the QP value assigned to this specific MB in the frame where it is fully encoded. In other words, the proposed method is more efficient when the inter-frame coding is weak and unable to identify similarities between frames: for example, a character or object turning on itself to present a hidden face, from which no MB cannot be predicted.
4.4 Discussion

The results help answering the problem in section 1.2: it is possible to reduce the bitrate of a video with ROI coding while maintaining better perceived quality than a regular H.264 encoding. However, the perceived quality improvement depends on the video content and performs better when the H.264 inter-frame coding is weak, for instance when there are few predictable movements in the video clip. It seems that the H.264 encoding method presented here and similar to [20] is better suited to teleconferencing or telementoring application than to omnidirectional video experiences.

It has to be emphasized that the DMOS results really depend on the quality seen by the user, which is, with HMDs, quite different from the quality of the video as an output of the ROI coding process in figure 3.1. As presented in figure 4.7, a Full HD resolution (1920x1080 pixels) for both eyes leads to visible pixels through the HMD’s lenses. In such conditions, the lack of detail in some MBs is not always detectable, as reported by some participants, and thus, in the case of VR, hardware settings stand as a limit to the benefits of ROI. This especially highlights how PSNR-based objective metrics have to be used with precaution. Some users reported to be more disturbed by a low playback framerate in situations: having a staggering video breaks the immersion and can be the source of motion sickness, which is a far worse experience than a low level of details in a subset of a video.

It also reflects the lack of a new reference recommendation to evaluate panoramic VR technologies. The ITU recommendations [38] [37] [36] contain a thorough description of the test settings, from lighting conditions to video sequences content, which ensure a strong confidence in and replicability of the results. Yet in the case of VR, there are no subjective quality assessment standards, nor is there reference omnidirectional video sequences that can be used across VR research works ([17] also uses publicly available videos on Youtube as test sequences for example). In the case of this study, only 4K and already H.264 encoded clips were available for testing, while having higher resolution (8K) raw video sequences could have potentially lead to a better DMOS results when compressing to 4K video samples with the proposed method. Indeed, H.264 encoded videos already lose high resolution details during the first quantization step.

4.4.1 Limitations

The proposed method can be implemented in a panoramic VR production environment. There are however a few points that limit its efficiency:

- As stated in the previous section 4.4, the perceived quality is highly dependent on the hardware used, and its influence on the playback framerate.

- As advised in the ITU recommendations, the video test sequences used in this study contain only one scene. If the prototype is to be implemented in a production environment, one has to keep in mind that having several scenes in a video could lead to a different perceived quality, depending on the scene cuts (fade to black etc.).

- While it allows a quick and efficient ROI encoding implementation, the x264 library limits the ROI encoding possibilities to the offered API. Moreover, although most of the video content is nowadays H.264-encoded, it is likely that a new encoding standard such as HEVC will establish itself as the new reference in the future. Therefore using a codec-independent method with a pre- and post-processing step could lead to better results (similar to [29]), since it allows direct control over the input and output video stream. Such a method is proposed in the further work section 5.1.

- The empirical ROI detection is not a 100% accurate. A subjective analysis of the attention heatmaps, for example in figure 4.5, shows that the head tracking combined to the attention model provides a good indication of the interest distribution across a frame. But those distributions are based on the assumption
that the user is mostly looking in areas around the reticule, as with a classic 2D display. Yet this was not verified in VR and thus the prototype’s attention model does not fit all viewing patterns. A way to evaluate the accuracy of head-tracked ROI versus eye-tracked ROI is proposed in the further work section 5.1.
Figure 4.7: A frame of the Philharmonie video seen through the HMD in (a), zoomed-in in (b). The viewer is able to see pixels and thus the additional level of details in ROIs is hardly discernible.
Chapter 5

Conclusion

In this master degree thesis, a method to encode panoramic Virtual Reality (VR) videos with empirical Region-Of-Interest (ROI) has been explored. In this method, ROIs are empirically detected from several viewings via the head tracking module required in all VR applications. To reach an accurate ROI detection, a Gaussian-based attention model is proposed. It was shown that viewers share similar ROI and the distribution of interest is temporally and spatially non uniform. With the help of the interest heatmap for each frame, the source video is then re-encoded using x264 and a matrix of Quantization Parameter (QP) offsets. This enables a space-variant level of detail depending on the level of interest at each point of a video frame, for each frame.

The perceived quality of the ROI encoded video was assessed accordingly to ITU evaluation standards. The results presented in this work show that this encoding method can achieve better perceived quality than a regular H.264 encoding, when the video has few predictable Macro-Block (MB). When many slices of the video are inter-frame coded, the performance of the proposed method is equivalent to H.264 but does not bring subjective quality improvements.

5.1 Further work

As stated in the discussion section, there is a dire necessity to define new subjective quality assessment standards for the growing number of academic works dealing with panoramic VR. In particular, the method presented here could be tested again with standard raw YUV420 videos as source video, instead of the H.264 sample used here. Two points could then be further explored:

- On the empirical ROI detection side, the attention model presented in this work could be further refined with the help of a HMD like the Fove 0, a VR headset that includes an eye-tracker. Such a hardware could help to build an empirical and more accurate attention model in a VR viewport, and thus more accurate ROI detection.

- On the encoding side, a pre-processing blurring method similar to the experience of Nyström and Holmqvist [29] could be tested with the same CCR assessment against regular H.264. Although the blur effect could have a negative impact on the quality, this would lead to a codec independent improvement for panoramic VR.
References


