Omnidirectional video encoding and delivery using advanced processing techniques

Ali Deniz Aladagli

2nd September 2016

Supervised by Erhan Ekmekcioglu

Institute for Digital Technologies
Loughborough University London

Contents

List of Figures **3**

List of Tables **4**

1 Introduction **5**
1.1 Background 5
1.2 Problem Definition and Objectives 6

2 Literature Review **8**
2.1 Quality Assessment 8
2.2 Projection 9
2.3 Tiling for Partial View 11
2.4 Summary 16

3 Work Plan **17**
3.1 Survey and Comparison 17
3.2 Novel Method 17

References **19**

List of Figures

1 VR Workflow 5
2 Heatmap statistics over a dataset of videos with interest represented in increasing order by blue, yellow and red from [24] 8
3 Equirectangular projection from [25] 9
4 Catadioptric camera example frame from [28] 10
5 Catadioptric camera example frame projected into 2D plane with cylindrical projection from [28] 10
6 Spherical subdivision and visibility sensors from [31] 11
7 Example tiling from [25] 15
8 Example tiling and rectangular layout from [46] 16

List of Tables

1 Work plan Gantt chart 18

1 Introduction

Even though the concept of virtual reality (VR) with the use of head mounted displays (HMD) is not new, the increasing availability of consumer grade HMDs such as Oculus Rift [1], HTC Vive [2] and mobile phone mount type VR setups like Samsung Gear VR [3] and Google Cardboard [4] have created a resurgence for the virtual reality medium in the recent years. While such devices might be associated with computer generated VR experiences, such as games, another trend that is becoming popular in their use are VR videos. This can be seen from popular multimedia companies such as YouTube [5] and Facebook [6] making VR videos available in their services and broadcasters such as BBC also experimenting with VR videos [7]. These type of videos are known by many different names in both the industry and the literature such as 360° videos, omnidirectional videos, spherical videos and spherical panoramas. VR videos are videos that have content available in many different directions, optimally in all directions. When viewed through HMDs with head tracking capabilities, VR videos provide a level of interactivity to the user by providing real time visuals for whichever direction the user is looking at and thereby creating a sense of immersion.

1.1 Background

The typical workflow for serving a VR video can be roughly represented by the flowchart in Figure 1.

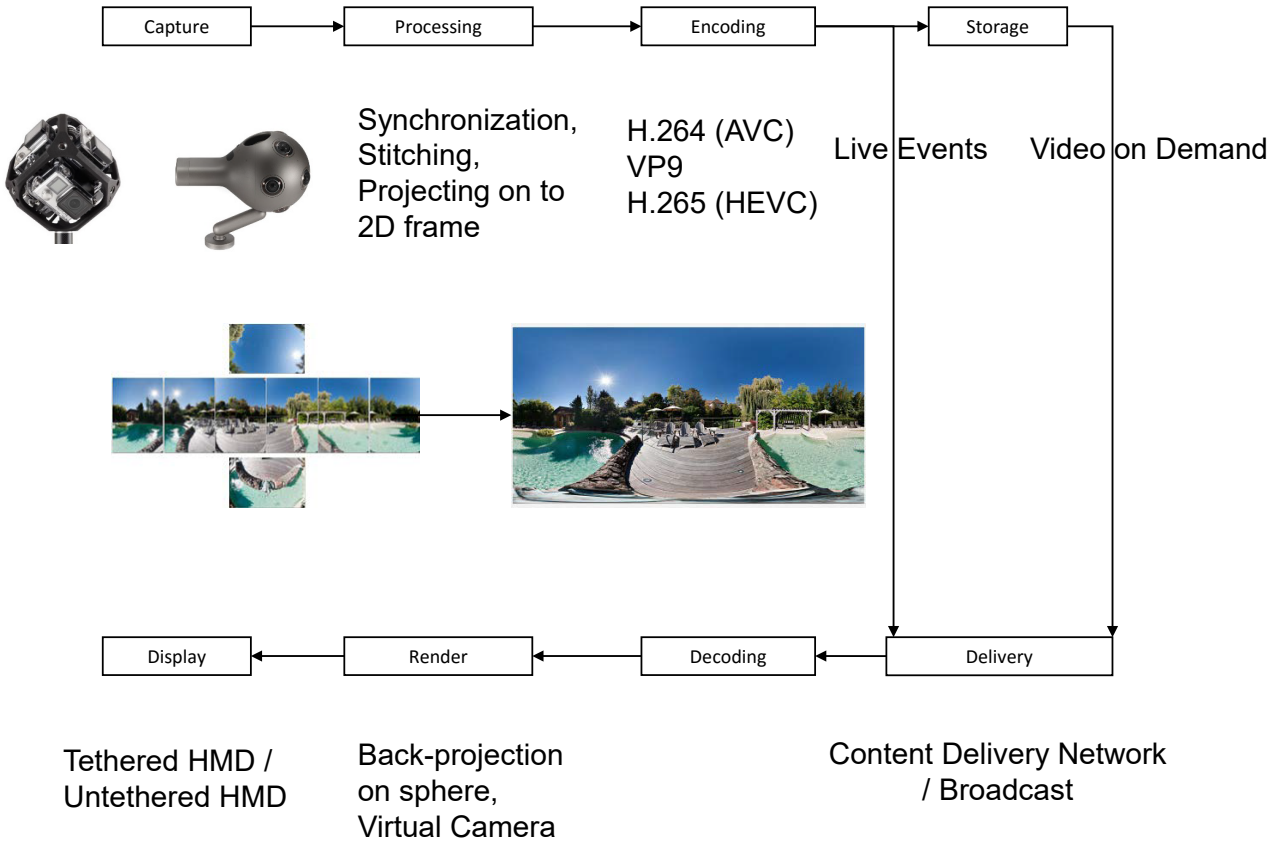


Figure 1: VR Workflow

.

Firstly, the appropriate omnidirectional footage needs to be created. While this is less problematic for computer generated animated content, real world cameras have limited field of views, even with wide-angle fish eye lenses. Therefore, multiple camera systems need to be employed to capture a fully spherical view. One solution for capturing omnidirectional videos is proposed in [8] where two hemispherical cameras sharing a viewpoint can be placed back-to-back to capture a fully spherical view. The cameras can capture hemispherical views through the use of an orthographical lens and a

paraboloid mirror. Since the images are being captured through a mirror, the camera aperture is in the field of view. This creates a spherical cap shaped blind spot in the footage. However, this also enables them to share a single viewpoint which help eliminate possible parallax artefacts when the hemispherical views are to be combined. Current practical techniques for capturing omnidirectional videos, however, employ multiple cameras pointing outwards from each other to capture the whole spherical view at the same time. Several different examples of these setups are Samsung Gear 360 [9], Orah 4i [10] and Nokia OZO [11].

Secondly, these multiple footages need to be stitched together into a singular representation to merge overlapping areas and provide seamless switching between different parts of the video. Different stitching methods and related algorithms have been studied in depth in [12]. The result of the stitching is a spherical representation of the environment. However, there are no standards for the encoding of omnidirectional content. Therefore, to make use of the widespread availability of standardized codecs such as H.264/AVC [13], VP9 [14] and the newer H.265/HEVC [15], the videos are required to be transformed into a planar representation. This transformation is done through a sphere to plane projection scheme, with the most popular and widely used one being the equirectangular projection. This is also the projection being used by YouTube [5] and Orah 4i live capture and streaming system [10]. Multiple research considers different mapping methods in order to optimize for conformation with encoding techniques while aiming to minimize distortion and redundant over-sampling that happens during the projection process.

Once the content is encoded, it can either be stored for later delivery or be transmitted through the use of existing transmission schemes such as broadcast, HTTP Live Streaming (HLS) or Dynamic Adaptive Streaming over HTTP (MPEG-DASH) just as other types of video content would be. Upon retrieval, the client device can decode the video and select the view since it has the whole representation available, which can be done with HMD head tracking capabilities or other interaction interfaces such as mobile handheld device touchscreens. Here, the HMD can be un-tethered like the Gear VR which is supported by a mobile touchscreen device with limited decoding capabilities or it can be a tethered HMD device like the Oculus Rift or HTC Vive connected to a computer which has much higher computation power. To recreate the selected view, the video is wrapped around a sphere with the corresponding back projection and viewed on the final display by covering the display with the boundaries of the selected view. This final step can be accomplished in a 3D rendering environment, such as a game engine, with a perspective camera in the selected viewing direction by using the decoded video frames as textures on a rendered sphere, with the appropriate texture mapping which reverses the initial planar projection.

1.2 Problem Definition and Objectives

In the current VR video ecosystem, as explained in 1.1, the whole spherical representation is delivered to the user devices in encoded planar video format. According to [16], the ideal VR experience with stereoscopic 3D vision would require a frame rate of 60 frames per second at 6K resolution per eye for the whole sphere to make full use of the HMDs currently available such as the Gear VR. Such video would need to be encoded with 20 to 40 megabits per second (Mbps) depending on the quality as mentioned in [16]. The delivery of this type of content is problematic in terms of both the bandwidth requirements for streaming and the decoding device capabilities. In terms of bandwidth, it is stated in [17] that the average internet connection speed across the globe was 5.1 Mbps as of 2015, while around 65% of the connections had a speed of at least 4 Mbps. At the same time the mobile devices used with the un-tethered Samsung Gear VR currently support decoding for 4K UHD at 30 frames per second at the most. With these limitations, it becomes apparent that specialized encoding and delivery techniques are required for high quality VR video services.

If a VR video is thought as a sphere surrounding the user's head in a virtual environment, as explained above, only a limited part of this sphere can be viewed at any time because of the limited field of vision of the human visual system. According to [18], human visual field has a field of view of around 140° vertically and around 180° horizontally with vision from two eyes combined. The area $A_{visual\ field}$ that this visual field expands on the surrounding sphere with radius R can be calculated by adjusting the surface area A_{zone} of the spherical zone with height h as given in equation (1) from

[19] with the vertical and horizontal viewing angles in degrees α and θ as shown in equation (2).

$$A_{zone} = 2.\pi.R.h \tag{1}$$

$$\begin{aligned} h &= 2.R.(sin(\alpha/2)) \\ A_{visual\ field} &= \frac{4.\pi.R^2.(sin(\alpha/2)).\theta}{360} \end{aligned} \tag{2}$$

$$\frac{A_{visual\ field}}{A_{sphere}} = \frac{sin(\alpha/2).\theta}{360} \tag{3}$$

Since the surface area of a sphere is $4.\pi.R^2$, the ratio of the visual field to the surface area of the sphere is can be calculated as in equation (3). Putting in 140° for vertical and 180° for horizontal viewing angles, the percentage of the visible area to the whole sphere can be calculated as 46.9%. However, the currently available HMDs like the Oculus Rift have horizontal viewing angles around 110° which actually change with factors like eye distance to lens of the HMD. Readjusting for this, approximately 28.7% of the sphere will be visible on the HMDs that are currently available.

Significant bandwidth savings can be achieved if only the required parts of the content were to be transmitted. Because of the similarity in applications, techniques that take advantage of partial visibility have been developed for panoramic video streaming applications [20]. One of these different methods that stand out is the tile based delivery system, where the video is segmented into tiles and only those that intersect with the current view are transmitted. However, the adaptation of these schemes to HMD viewing applications is not straightforward because of the different projection schemes and interactivity level involved. To minimize the interaction delay and its adverse affects, predictive delivery approaches and out of view contingency mechanisms have been proposed.

Such partial delivery mechanisms where every user requires a personalized delivery would consume huge amounts of bandwidth in a live event coverage setting such as sports, concerts or other events. However, users might be enjoying such content in clusters, for example, in the same household. Given that users' have some overlapping areas in their viewing directions, significant bandwidth savings can be achieved by creating a managing entity which manages the content delivery to the household and its distribution to the users. Therefore, the objectives of this research can be explained in three parts:

- The development of novel video processing methods which can be used for efficient delivery of VR videos that also incorporate the prediction of the likelihood of users watching specific viewing angles, by making use of visual attention models and properties of human vision system.
- The design and simulation of a scalable VR video delivery system together with a managing entity which coordinates the retrieval of required video data and the delivery of unicast, multicast and/or broadcast packets for users in a shared household, thereby smoothly facilitating the changes in view directions.
- The measurement of the proposed system's success with the appropriate metrics, evaluating the quality of experience, the data rate required, the reduction in latency achieved.

2 Literature Review

Streaming of omnidirectional videos cover a wide area of research because of the many parts that need to work together to make it functional. Therefore, the areas being researched can be roughly categorized as sphere to plane projection methods, optimized segmentation algorithms, predictive and compensating tile delivery selection algorithms and omnidirectional video evaluation methods.

2.1 Quality Assessment

As in normal videos, PSNR is one of the common distortion measurement tools used while BD-rate [21] is used the compare rate-distortion over a range of bitrates. Due to omnidirectional videos having multiple representations in multiple domains, there is no one way to evaluate the quality of a video. For example, the overall quality in all areas of the video might be favoured for a storage scenario while it might make more sense to evaluate what the user is seeing to measure the quality of experience. To measure the quality from the user's perspective, the PSNR at the rendered viewport is measured in [22] and [23]. In [22], sPSNR is also proposed to measure the overall fidelity of a projection in a video, where the error at each point is multiplied by its solid angle (area the pixel covers on the sphere) to account for the differences of the projection schemes.

In [24], two metrics, namely S-PSNR and weighted S-PSNR are described. S-PSNR measures the PSNR value in the spherical domain by mapping the 2D planar ground truth and processed videos back onto spheres. The error is computed over the same set of points sampled on the spheres. S-PSNR makes it possible to measure the distortion between different representations over the whole content, which is also useful for comparing projection methods. The weighted S-PSNR, on the other hand, can assign different levels of importance to different points by multiplying the error by a weight. This is inspired by the viewing statistics optained from experiments performed over an omnidirectional video data set. As shown in Figure 2 from [24], it is concluded that users spend more time viewing the equator rather than the poles of an omnidirectional video. Therefore, it is theorized that S-PSNR weighted according to latitude access frequencies trained from a dataset would be able to predict the PSNR at the user viewport, which would enable measuring quality at user side without the need for viewings. Experiments showed that the weighted S-PSNR differed less than 7% on average in BD-rate from the actual PSNR at the viewport, while PSNR at the equirectangular domain and the normal S-PSNR are shown to have more difference from the viewport PSNR. It is not clear if the latitude frequency dataset and the testing dataset are the same, however this may cause a bias in the measurements, favouring weighted S-PSNR.

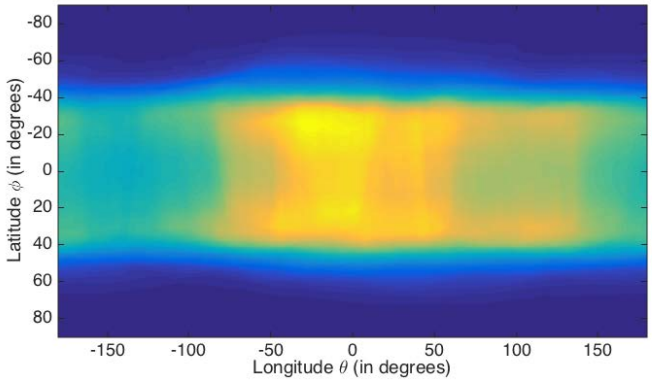


Figure 2: Heatmap statistics over a dataset of videos with interest represented in increasing order by blue, yellow and red from [24]

.

2.2 Projection

As explained in section 1.1, omnidirectional videos need to be mapped before being encoded with standard codecs. There have been multiple projection methods proposed in both in the industry and the literature. The current basic projection method most commonly used is the equirectangular projection. In this projection, sampling in latitudes and longitudes are done with the same angular frequency. The arc distance on the sphere between two longitudes actually decrease when latitudes get closer to the poles. This causes distortions and increasing oversampling approaching the poles of the sphere. To achieve the projection in shown in Figure 3, (x, y), the horizontal and the vertical coordinates respectively, corresponding to the spherical coordinates (ϕ, θ), in a video with resolution $(width, height)$ and upper left corner at $(0, 0)$ can be calculated as shown in equation (4), with latitude ϕ and longitude θ given in radians.

$$\begin{aligned} x &= \frac{width}{2} + \frac{\theta.width}{2.\pi} \\ y &= \frac{height}{2} + \frac{\phi.width}{\pi} \end{aligned} \tag{4}$$

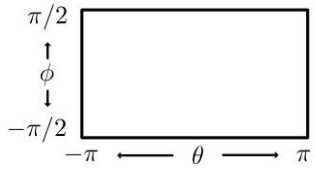


Figure 3: Equirectangular projection from [25]

.

Facebook, in [16], [26], [27], has shared experiences with multiple projection methods. In [26], it is explained that when compared to the equirectangular projection, the cubic projection has more uniformly distributed sampling and less distortion which makes cubic projection conform more to the standardized codecs which expect linear motion. In [27], a pyramid projection is presented for an adaptive bit rate streaming approach. The pyramid projection is a view based projection where a pyramid is faced with its base parallel to the assumed viewport facing forward while its base point is in the exact opposite direction. This setup provides high quality in the viewing direction and a uniformly degrading quality on areas further away from the viewing direction due to the shapes of the pyramid's faces. Since the video needs to be viewable from all directions, multiple representations with the pyramid projection are required to make these directions available at high quality, while each of these representations require different bitrate encodings to fit into a dynamically adaptive bitrate streaming scheme. In the presented case, each video is encoded in 5 different qualities with 30 different viewports. It is stated that, when compared to equirectangular projection, the cubic projection requires 25% less file size while the pyramid projection requires 80% less file size. Finally, in [16] the complicated computation required for pyramid projection is noted while off-set cubic projection (or a truncated pyramid) is offered as an alternative. This method is shown to be computationally simpler than the pyramid projection while having a smoother quality degradation as the view moves out of alignment.

The challenges and shortcomings of compressing omnidirectional videos with H.264/AVC that were captured with catadioptric systems (similar to ones presented in [8]) are discussed in [28]. While the videos being considered have 360° FOV horizontally they are lacking the spherical caps at the top and the bottom that fully spherical omnidirectional videos have. The high computational intensity involved in recreating a perspective view from the video captured with such systems is noted. It is stated that if this footage is compressed directly with H.264/AVC, the compression integrity is lower in these kind of videos because of several reasons. Firstly, it is explained that jagged vertical edges in the recreated perspectives of the video are formed by the artefacts caused by the block processing aspect of H.264/AVC. Later it is described that, camera movement or motion in the scene results

in orientation or scale changes in the captured footage because of the distortions created by the hyperbolic mirror which do not coincide with the translational motion expected and taken advantage of by the motion estimator of H.264/AVC. To overcome these problems, a pre-processing stage is proposed to be performed before encoding the video where the footage is converted to panoramic images with cylindrical projection, as shown in Figure 5. It is stated that rendering a perspective view is computationally simpler from a cylindrical projection and that this projection reverses the distortions in the linear motion trajectories inside the video to some extent while also compensating for the scaling effects. The PSNR values at the rendered perspective views created from the proposed method and the direct encoding method are measured at different rates and compared, using the rendered perspective views created from the uncompressed video as the ground truth. It is reported that the proposed method performed better in lower and higher segments of the rendered perspective views and that the player was 30% faster in navigating frame by frame with the proposed method compared to the direct encoding method.

Figure 4: Catadioptric camera example frame from [28]

.

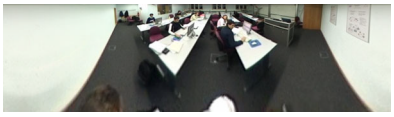


Figure 5: Catadioptric camera example frame projected into 2D plane with cylindrical projection from [28]

.

In [22], the disadvantages of previous sphere to plane mapping schemes are discussed in the domain of omnidirectional video and a novel method called the rhombic dodecahedron map is proposed. The rhombic dodecahedron is a convex polyhedron formed of twelve identical rhombus shaped faces. After the rhombus faces are projected onto a sphere using gnomic projection, the formed spherical rhombic segments are subdivided into a grid where each cell represents a pixel in the planar domain to complete the mapping. The proposed mapping method is evaluated in terms of uniform sampling, area deviation and stretching and shown to be better than cubic, equirectangular and equal area cylindrical projections in these aspects. Experimental results also showed that the proposed projection scheme performed better than the cubic projection in both compressed and uncompressed domains based on measurements performed with PSNR values averaged over arbitrarily chosen view directions. The measurements with sPSNR in order measure quality in the whole video, as explained in subsection 2.1, also showed consistent results.

2.3 Tiling for Partial View

The tiling approach seems to be effective in reducing the bandwidth requirement for transmission and the workload on the decoding device [23], [29]. However, in a naive delivery scenario where only the tiles covering the requested view are transmitted to the client, the latency between the video server and the client causes discrepancies in the rendered view when the viewing direction changes [23]. This is due to clients' cache containing only the previously required tiles' data required for continuous playback. Therefore, when the viewing direction changes so much that different tiles are required, the situation is similar to seeking to a not yet loaded part of a video during a normal playback stream and the required data is missing until the new request can be delivered. Several previous works seek to overcome this problem by modifying the tiling and the transmission schemes in different ways.

In [29], the compression and the transmission of panoramic videos are discussed. For the compression, a tiled encoding system is proposed where a 360° panoramic video is divided into six vertical tiles and encoded separately with MPEG-2 standard [30]. In the proposed system, at most two tiles is enough to create the desired view on the viewer side, therefore it is sufficient to deliver and decode at most two tiles which decreases both the bandwidth requirement and the load on the decoder. In this system, the tiles must be encoded with a GOP structure that synchronizes their I-frames, which makes switching tiles easier when the viewing angle is changed by performing the switch at the next I-frame of the next GOP. The importance of adjusting the frequency of I-frames is noted, trading off between compression efficiency and response time delay on a tile switch. In terms of transmission, a VOD stream sharing system is proposed where multiple users are on the same local network. Users make primary requests for the current video segments and secondary requests for the future video segments. The video server delivers the overlapping requests with multicast or broadcast instead of unicast, thereby reducing disk read requirements and the consumed bandwidth. It is a requirement that users maintain a larger cache than a normal singular TCP delivery scheme so as to accommodate the future video segments that arrive from the shared streams. This sharing scheme is tested in setup with a video server and eight user computers connected to a LAN. Presented results show that the proposed sharing scheme succeeds in using both less disk bandwidth and less network bandwidth when compared to the video server sending the video to users one by one periodically. While only sending the requested tiles is discussed and is possible in the scheme, this is not tested and every tile is sent to the viewers. A broadcast perspective of transmission is briefly discussed for panoramic videos where different tiles can be delivered from different channels while the user's set-up box needs to be modified to be able to determine the needed tiles based on user input, decode them and present them in final view.

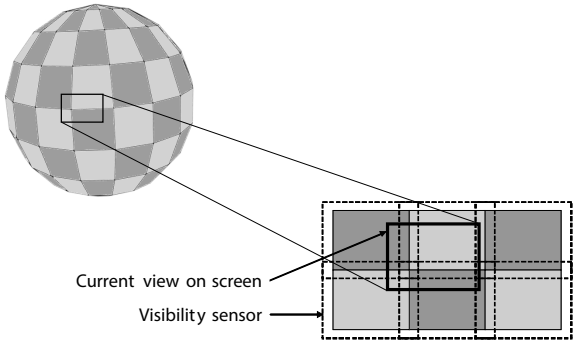


Figure 6: Spherical subdivision and visibility sensors from [31]

.

The interactive viewing of omnidirectional videos through HMDs with head tracking is presented in [31]. Once again a tiling mechanism is proposed where the tiles are encoded individually with MPEG-4, however, the subdivision into tiles is performed on the spherical domain rather than in the planar domain. To determine which tiles need to be decoded, tiles are associated with areas on the sphere called visibility sensors (which can be of different sizes). These sensors are tested against the

view requested by the head tracking of HMD to determine if the associated tiles need to be decoded. When the visibility sensor area is larger than the area of the associated tile, this mechanism also acts as a prefetcher, ensuring smooth changing of viewing direction by preparing tiles just before they are actually needed by the view. The subdivision and visibility sensors are shown in Figure 6. Experiments evaluating the data overhead while using different tile size selections are performed. It is shown and concluded that, there is a trade-off between data overhead of large tile sizes and the delay caused by the number of operations required to process a high number of smaller tiles. The differences in encoding performance and effects of different visibility sensor overlapping rates are not evaluated.

In [32], possible incompatibilities that tile based transmission systems have with fully spherical videos that are stored in equirectangular projections are discussed. It is explained that in a fully spherical equirectangular video, both the shape and the size of the area required from the source video to render a view changes drastically depending on the requested view direction on the sphere. Due to this variation, tiles required to form a view from a video tiled with low granularity carry redundant data. While increasing the granularity of tiling can counteract this, it is stated that this will reduce the performance of the compression. A transcoding based approach is proposed where the smallest rectangular area covering requested view is extracted, encoded with fast motion re-estimation and transmitted. It is explained that, in a transcoding scenario the fast motion re-estimation takes advantage of the already calculated motion vectors in the previously encoded video and therefore is computationally more efficient than encoding the video from scratch. It is stated that, depending on the viewing angle, this method provides bitrate savings of 42.4% to 86.3% when compared to a tiled scheme which has twelve (six by two) uniform tiles in total. In the experiments performed, the fast motion re-estimation approach presented an average PSNR value of 0.57 dB worse when compared to full search motion estimation, this is explained as an acceptable trade-off for the computation time saved. The latency or real-time aspects of the proposed approach are not discussed.

To address the viewing discrepancy occurring from network latency, an adaptive tile delivery based scheme is proposed in [23]. In the proposed scheme, all tiles are encoded at multiple different bitrates and a utility value (U) and bitrate cost (C) is assigned to each encoding of each tile. JPEG2000 is used to encode the video with an Intra-only scheme in order to focus the study on the evaluation of the adaptation of the scheme which also ensures random access is available for each frame in the video. Assuming that the transmission delay (d) and the speed of the viewing direction change (s) is known, the probability of each tile being required in the next view (P) is calculated by passing a Gaussian filter with a standard deviation of $\sigma = d.s$ over the 2D visibility mask of the tiles in the last view. For the tile selection process at each frame, the tile-encoding pairs are sorted in increasing order of their expected utility over cost ratio $U.P/C$ and pairs are selected in order from this list until a given bandwidth budget is reached. Experiments are performed with a viewport that is moving at a constant speed at 120° per second between randomly sampled points on the sphere and the PSNR at the viewport is measured for comparison. The proposed delivery is tested under different network delays ranging from 40 ms to 1280 ms and different bandwidth budgets. The implementation is also compared to three other implementations of omnidirectional video delivery in the same aspects; a constant rate complete delivery as in [33] and [5], a naive tile delivery where only the tiles covering the requested view are transmitted and a double quality delivery where the tiles covering the requested view are transmitted with high quality while the tiles out of view are sent with lower quality. When the measured PSNR values are averaged over forty seconds of video the proposed method is shown to be never worse than the constant rate approach and consistently better than all methods for increasing delay values larger than 160 ms. In this study, the utility of a tile-encoding pair is dependent on the MSE difference the tile has with respect to the previous quality setting of the same tile. Since an equirectangular projection is used in the proposed scheme, this utility function does not take into account the fact that tiles in different locations on a uniform grid in an equirectangular video contribute different amounts to the viewport they are visible in. This could be countered by adjusting the MSE of the tiles with a projection area ratio or calculating the MSE difference of changing tiles at the viewport.

In [34], a partial decoding enabled region of interest panorama player is proposed. In the proposed scheme, the user selects an area from a down sampled version of the original video, called the navigation video, at which point they are presented a high quality version of that region. The navigation video

is selected as the base view required for multiview video coding [35] of H.264/AVC. MVC is chosen for being able to decode synchronized multiple views. The high quality representations tiled from the original panorama are encoded with MVC together with the navigation video while across view prediction disabled. Only the tiles required for the region of interest are decoded. When choosing the required tiles, the region of interest is extended with a support zone for smooth switching since inter frame prediction is used and random access is only available in I-frames. The final view is rendered out of these tiles. This system is extended into an adaptive streaming approach in [36]. In this approach the tiles are encoded with multiple bitrates and combined in a single MVC stream with the navigation video. The client is initially presented with the navigation video and is subsequently presented with the PSNR values of the tiles for each different bitrate level. The client determines the needed tiles and chooses higher bitrate levels for each selected tile until bandwidth budget consumed. The client requests the required tiles at decided bitrates from the server, to which the server replies by extracting the required tiles from the complete MVC stream and delivering them. A complete capture to panoramic region of interest streaming system, using the tiling and the transmission scheme in [36], is presented in [37]. End-to-end transmission was performed with bitrates below 15 Mbps in a best effort network with a maximum bandwidth of 100 Mbps, where the client platform is a PC and the system supports both horizontal and vertical panning together with zooming in and out through the use of a game controller. This transmission scheme is evaluated in both [38] and [39] with subjective questionnaires. Uniform delivery of the complete video is compared to the partial delivery scheme in [38], where subjective questions evaluating image clarity, crispiness and block-noise annoyance showed users favoured the partial delivery scheme over the complete delivery scheme. The sense of presence and enjoyableness of the interactive content made available with the partial delivery scheme is subjectively evaluated with two questions in [39], with users giving an average score of 3.91 and 4.01 respectively over a 5 point scale.

This tiling scheme in [37] is further extended in [40], where streaming of region of interest panoramic video to mobile devices with less decoding capabilities and less bandwidth are considered. In order to achieve this, the number of tiles that need to be decoded is decreased to two, one for low quality navigation video and one for the high quality region of interest. To make the region of interest available through only one tile, the tile sizes are enlarged to the point where they overlap with each other when the original tiling is performed. The delivery to the mobile device is done through the Long Term Evolution (LTE) network. In [41], this two-tile delivery scheme is used to stream video to HMDs, where a test is performed using Oculus Rift as the HMD. In this scheme, the navigation video from [40] acts as the background video while the area the user is looking at gets increased quality with the high quality tile. The user never sees a black area in the video however the quality increase is only initiated once the users stop moving their heads. It is reported that with a low quality background of 0.5 Mbps and high quality region of interest of 2 Mbps, it takes 2.6 seconds after the request is made on average to switch to a new high quality tile. It should be noted that the video is encoded in the equirectangular format and it is not mentioned if the tile sizes are adjusted for the pole areas. This stream to HMD scheme is applied in a live streaming scenario in [42] with 4762 unique viewers. To handle the tile encodings in real-time, multiple encoders are used following the real-time stitching, later the encoded tiles are delivered to content delivery networks to handle the mass delivery to users.

Spatial Relationship Description (SRD) extension to MPEG DASH standard is presented in [43]. SRD integrates tile based schemes into adaptive bitrate streaming by describing the relationship of different tiles, such as the area they represent on the video. A live tiled streaming system based on SRD has been presented in [44] together with the results of a field trial. This enables mobile tablet users to participate in viewing a high resolution panoramic coverage of a live event by choosing their own regions of interest through interacting with the mobile application. This is accomplished similar to as it was done in [40], with overlapping tiles so that only the decoding of one tile is necessary for the complete mobile device view, however, with the exception of the inclusion of a low quality background image. The system provides multiple resolution levels where the lowest resolution layer is not tiled and acts as the most zoomed out level, while increasing resolution layers are tiled more granularly. The tiled representations of the live footage are encoded in multiple nodes and delivered to users using content delivery networks to keep up with the live constraint. Two types of viewing

experiments were performed where in one the users were directed with a number of interaction tasks to complete while in the other they were free to interact as they wished. For objective measurements, metrics such as buffering time and region of interest switching time are recorded while user experience is measured through subjective questionnaires. The correlation between the objective metrics and the user experience is analysed. These analyses showed that during the experiments where the user is directed, the user experience was observed to be non-significantly affected by the interactivity features of the application, which might show that these interaction features are beneficial. However, the same is not observed in the free viewing experiments while the user experience in the continuities of zooming and of navigation during the free viewing experiments were observed to be negatively correlated with the buffering length as expected.

The unicast/multicast management of tiled videos in wireless local area networks are discussed [45]. The case is made that, in a scenario where multiple users in the same wireless network access the same tiled video service simultaneously, some users might have overlapping regions of interest and therefore might be requesting the same tiles in the video. Normally, each requested tile would be sent over unicast to each user that requested it, which puts a high strain on the network with increasing number of users. To optimize bandwidth usage, five tile delivery assignment algorithms are proposed which direct the delivery of tiles that multiple users request to multicast instead of unicast, while keeping to the transmission deadlines of the users. First of the algorithms minimizes the transmission delay by sending a tile from multicast if multicast delivery time for that tile is less than time required for the unicast delivery of the tile to all users who need it. This algorithm is noted to support the maximum number of users. Given n, the number of users, n_t, number of users requiring a tile t and p, the probability that a multicast tile will reach all users, the second algorithm maximizes the total utility for assignments that keep to the transmission deadline. The utility for the unicast and the multicast tiles, $\psi_u(t)$ and $\psi_m(t)$ respectively, are calculated as shown in equation (5). The term $\frac{(n.p-n_t)}{n}$ denotes the penalty for users that did not need the tile receiving it because of the multicast delivery. The utility maximization algorithm is noted to complete in the order of minutes, which make it unusable in a real-time scenario. The third algorithm called threshold, delivers a tile from multicast if the utility of it being multicast is positive, which is noted to run in linear time while it may fail to provide a valid assignment due to it not considering the transmission delays. The fourth algorithm, called greedy, starts with the solution from minimizing delay and converts multicast assignments with the minimum number of n_t into unicast until the deadline is violated, which increases the total utility since $\psi_u(t)$ is always greater than $\psi_m(t)$. The final algorithm, called expectation, uses the statistics from previous viewings to modify the utility maximizing algorithm, which means this algorithm has to work offline. The experiments performed with simulated network conditions showed that the greedy algorithm was the most reasonable one to use in real-time, giving valid assignments for the most number of users. It is also explained that the expectation algorithm performed with low performance because of the valid assignments being invalidated when the views were differing from the statistics.

$$\begin{aligned} \psi_u(t) &= \frac{n_t}{n} \\ \psi_m(t) &= \frac{n_t.p - (n.p - n_t)}{n} \end{aligned} \tag{5}$$

In [25], a content adaptive tile based projection scheme is proposed in order to negate the redundancies that exist in the equirectangular projection. In this method, an equirectangular source video is tiled horizontally while allocating different number of samples (resolutions) and bitrates to each horizontal tile. Given a number of tile sample sizes and a number of bitrate settings to choose from, a Lagrangian optimization algorithm that minimizes the distortion is proposed to select the best tiling configuration. The distortion calculation method used is L-PSNR proposed in [24] by the same authors. An example configuration can be seen in Figure 7. The edges formed at the boundaries of neighbouring tiles caused by the sampling difference is noted. To counteract this, the tile heights are enlarged by a small percentage and alpha blended at the overlapping regions. An average tiling configuration is also computed over a dataset of videos to evaluate if the content adaptation is more effective than using a single configuration. The average configuration and the content optimized tiling

for each video are compared to equal area cylindrical projection and equirectangular projection methods using L-PSNR [24]. The evaluations show that the equirectangular projection performs worse than all the other methods consistently. On a dataset of ten videos, when the RD-curves are compared to the equal area cylindrical projection method with the BD-rate, the content based approach performs better in eight videos while the average method performs better in six of them. The optimized approach always performs better than the average method. It should be noted that the optimization algorithm is performed for the first frame of the video which can affect performance. The tile selection for transmission of video has not been studied. The encoding method used in the tests is stated to be H.264/AVC, however the encoding settings such as the I-frame frequency have not been included in the study.

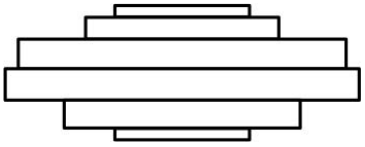


Figure 7: Example tiling from [25]

.

[46] extends the horizontal tiling discussed in [25], where a method to find a segmentation that minimizes the ratio of the planar projection area to the spherical surface area is presented. This ensures the least amount of redundant pixels are used in the projection. Instead of using all horizontal rectangular tiles, the tiling is modified to project the spherical caps onto squares or circles. For the segmentation latitudes given in increasing order $\theta_0, \theta_1, ..., \theta_n \in (0, \frac{\pi}{2})$, the total area of the tiling projection for one hemisphere $S_{hemisphere}$ is given in equation (6). The area for the square and the circle areas used for the projection of spherical caps S_{pole} is given in equation (7). The area used for the projection of the spherical caps $S_{pole,Yu}$ in [25] is given in equation (8). Since the area in equation (8) is always larger than the area in equation (7), the configuration proposed in [46] is proven to require less area than [25] under the same minimum sample density constraint. The overlapping method is used with alpha blending to eliminate the edges forming at the neighbouring tile boundaries and the area calculations are adjusted accordingly and used as the minimization functions for the segmentation. For codecs that expect a single video stream, the tiles are also shown to be able to be arranged into a single rectangular layout as shown in Figure 8. The optimized segmentation latitudes are presented for three and five horizontal tiles and used for comparison with the method from [25] and cubic projection using equirectangular projection as the source video. The representations are encoded with H.265/HEVC intra-only coding and compared with S-PSNR and L-PSNR [24]. The BD-rates averaged over twelve video sequences showed that the proposed method achieved savings over both cubic and equirectangular projections and the method proposed in [25]. The performance for transmission of videos using this tiling mechanism have not been studied.

$$S_{hemisphere} = S_{pole} + \sum_{i=1}^{n} 2.\pi.r^2.cos\theta_{i-1}.(\theta_i - \theta_{i-1}) \tag{6}$$

$$S_{pole} = \begin{cases} \pi.(\frac{\pi}{2} - \theta_p)^2.r^2 & , CirclePole \\ 4.(\frac{\pi}{2} - \theta_p)^2.r^2 & , SquarePole \end{cases} \tag{7}$$

$$S_{pole,Yu} = 2.\pi.r^2.(\frac{\pi}{2} - \theta_p).cos\theta_p \tag{8}$$

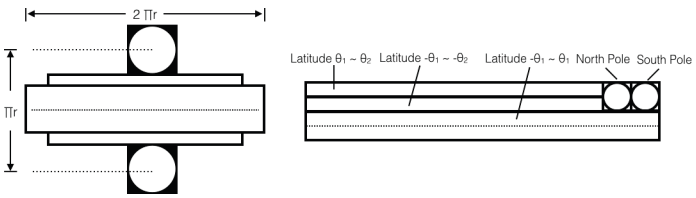


Figure 8: Example tiling and rectangular layout from [46]

.

2.4 Summary

It is clear from the literature that many approaches for the partial delivery of a video have been proposed, with the tiled approach being mentioned most commonly. However, these tiled approaches either do not consider the immediate interaction and the loss of vision associated with HMDs or have low performance in getting a new high quality view when the viewing direction changes. While different projection methods have been proposed, the most commonly used one is still the equirectangular format, which is consistently proven to have lower performance compared to other projection methods.

There are several unstudied aspects in this area. In most situations the tiling is performed uniformly on the projected planar video. In this case, the views required by the user map to different shapes in the planar domain, which means different views will require different number tiles where the tiles have different amounts of information of the sphere. Therefore, non-uniform tiling schemes that depend on the projection method must be studied as these can eliminate redundant information being delivered and improve performance by decreasing the number of tiles required.

Viewing through HMDs mean any view presented to the user will cover their field of vision completely. However, the human vision has variable resolution for areas further away from the fovea. Some HMDs might also have varying pixel layouts to account for this while their lenses cause distortions in certain areas of the view [47]. Adapting the quality of the tiles accordingly might result in bitrate savings, this needs to be studied while keeping in mind the users might just change their eye fixation points instead of turning their heads. Studying the users HMD usage patterns might be helpful in this aspect.

3 Work Plan

The work plan for the next two years will form of implementation of base techniques and their comparison, followed by the design, implementation and evaluation of a novel method. The last year will be focused on writing of the thesis. A subdivision of these tasks as shown in Table 1 are explained more in detail below.

3.1 Survey and Comparison

The previous works reviewed here present improvements in different parts of the VR video delivery workflow. Some of the methods presented are more restrictive such as being applicable only for certain projection schemes or delivery methods while others might not be suitable for use with deliveries into HMDs at all. Also many of these works are evaluated with different metrics, which make a comparison between methods problematic. The first step to take will be acquiring/creating implementations of the prominent methods proposed in the previous works so that they can be compared under the domain of VR videos delivered to HMDs through the network.

One of the most adaptable of these techniques seems to be [23] with a bandwidth budget based approach which can fill out-of-view directions to account for head movement. However, this method might not be suitable for use with live event coverages. Another notable method is [41] which has been field tested with a live to HMD streaming scenario [42]. One more live tiling system, albeit without HMD interaction, is evaluated in [44]. It should be noted that all of these systems use equirectangular videos as the source video and therefore their adaptations using other projection methods will also be implemented.

Later the implementations will be evaluated under different latency and bandwidth conditions simulated with network simulator [48] using a dataset of videos. The complete equirectangular delivery method will be used as the baseline. The metrics that will be used for evaluations will include omnidirectional video specific methods such as those discussed in [24] and perceptual visual quality metrics such as those discussed in [49]. Through these comparisons, an extensive survey of the VR video field will be created, which will be the first publication targeting the seventeenth month.

3.2 Novel Method

The current literature for VR video delivery does not utilise perceptual visual quality metrics for optimization or evaluation. Optimizing for a higher value in perceptual visual quality metrics instead of signal fidelity metrics might yield a method that is optimized more towards user experience for the same bandwidth. While some approaches deliver lower bitrate encodings for out of view areas, this is achieved either through either spatial down sampling or higher quantization. Temporal down sampling methods have not been investigated. Further investigation on these subjects and view prediction from visual attention will be performed. Combining the investigation results together with the analysis of the survey, the VR video streaming system will be designed together with the household stream managing entity.

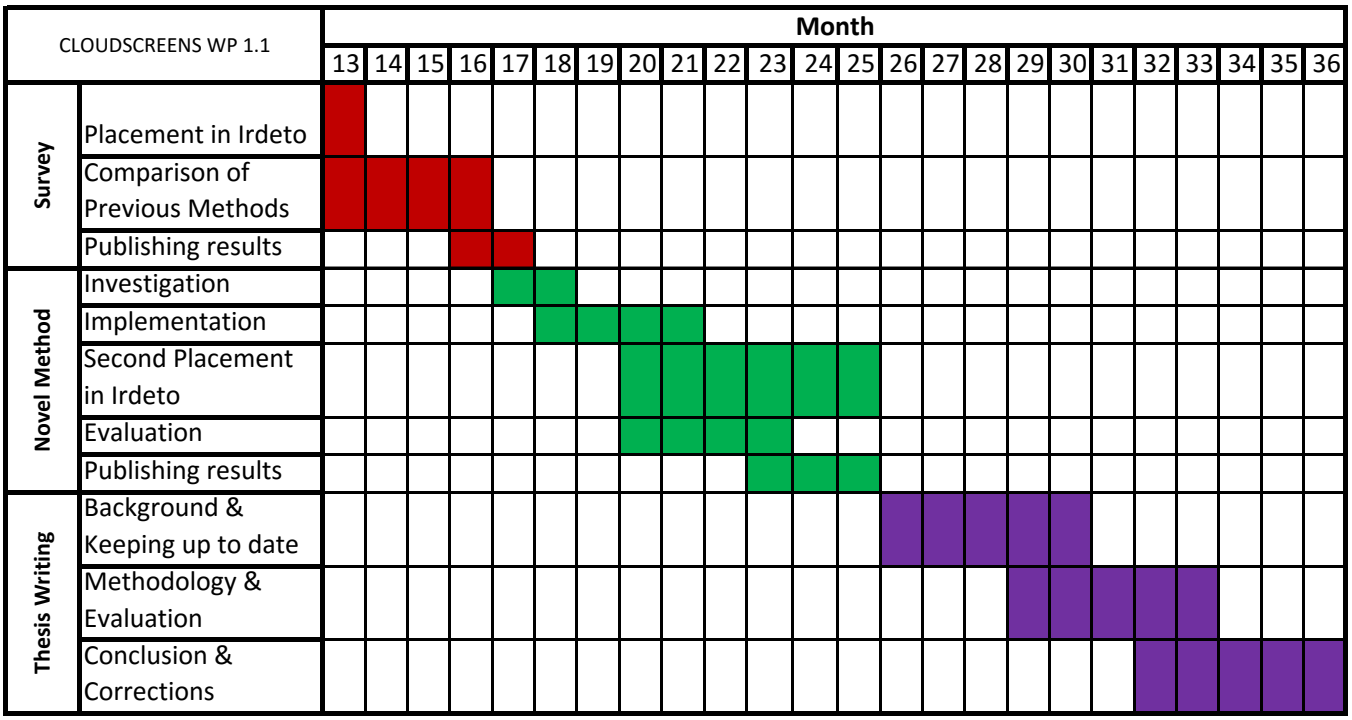
On the content preparation part of the workflow, the projection method most suitable for tiling will be selected based on the investigation. The uniform tiling problem mentioned in subsection 2.4 can be countered in two ways. One of them is to perform the tiling in the spherical domain, where each tile is a distinct projection by itself that can represent a specific view. The other one is changing the tile sizes and shapes based on the areas they represent based on the projection. These two methods will be implemented and compared. If a single tile streaming scheme is pursued, the human visual system properties may be taken advantage of here. For example, the edges of the tiles might be encoded in lower bitrates to take advantage of low spatial resolution in peripheral vision, lens distortions and differing pixel densities on the display, however these need to be studied first.

On the content delivery part, a management entity will be added following the transmitting server to manage the tiles requested from the server and deliver them to the requesting HMDs sharing the connection and viewing the same content simultaneously. Here a prediction algorithm will be selected or developed to predict and pre-fetch tiles based on view directions, view change speeds and the

content itself. While similar prediction methods are proposed in [50] and [51], these methods only perform prediction per single user and are low performance. Since the managing entity is managing the delivery for multiple users, other options will need to be explored here.

After the implementation is complete, it will be compared and contrasted to the techniques in the survey again by simulating network conditions with network simulator with a dataset of videos. A second experiment will be performed by increasing the number of users in the system in order to measure the scalability of the system. The results will be published targeting twenty-fifth month.

Table 1: Work plan Gantt chart



References

[1] Oculus rift, [Online]. Available: https://www3.oculus.com/en-us/rift/ (visited on 28/08/2016).

[2] Htc vive, [Online]. Available: https://www.htcvive.com/ (visited on 28/08/2016).

[3] Samsung gear vr, [Online]. Available: http://www.samsung.com/global/galaxy/gear-vr/ (visited on 28/08/2016).

[4] Google cardboard, [Online]. Available: https://vr.google.com/cardboard/ (visited on 28/08/2016).

[5] Youtube help - upload 360-degree videos, [Online]. Available: https://support.google.com/youtube/answer/6178631?hl=en (visited on 28/08/2016).

[6] Facebook help centre - what is a 360 video?, [Online]. Available: https://www.facebook.com/help/851697264925946 (visited on 28/08/2016).

[7] 360 video & vr - immersive news, [Online]. Available: http://bbcnewslabs.co.uk/projects/360-video-and-vr/ (visited on 28/08/2016).

[8] S. K. Nayar, "Catadioptric omnidirectional camera", in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, IEEE, 1997, pp. 482–488.

[9] Gear 360, [Online]. Available: http://www.samsung.com/global/galaxy/gear-360/ (visited on 28/08/2016).

[10] Orah 4i, [Online]. Available: https://www.orah.co/ (visited on 28/08/2016).

[11] Nokia ozo, [Online]. Available: https://ozo.nokia.com/ (visited on 28/08/2016).

[12] R. Szeliski, "Image alignment and stitching: A tutorial", *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.

[13] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra, "Overview of the h. 264/avc video coding standard", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, 2003.

[14] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins and Y. Xu, "A technical overview of vp9the latest open-source video codec", *SMPTE Motion Imaging Journal*, vol. 124, no. 1, pp. 44–54, 2015.

[15] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[16] (2016). Optimizing 360 video for oculus, [Online]. Available: https://developers.facebook.com/videos/f8-2016/optimizing-360-video-for-oculus/ (visited on 28/08/2016).

[17] D. Belson, *Q3 2015 report, akamai's state of the internet*, 2015.

[18] J. J. Gibson, *The ecological approach to visual perception: Classic edition*. Psychology Press, 2014.

[19] E. W. Weisstein. Zone, [Online]. Available: http://mathworld.wolfram.com/Zone.html (visited on 28/08/2016).

[20] S. Chen, "Quicktime vr: An image-based approach to virtual environment navigation", in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 29–38, ISBN: 0897917014.

[21] G. Bjontegaard, "Calcuation of average psnr differences between rd-curves", *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April 2001*, 2001.

[22] C.-W. Fu, L. Wan, T.-T. Wong and C.-S. Leung, "The rhombic dodecahedron map: An efficient scheme for encoding panoramic video", *Multimedia, IEEE Transactions on*, vol. 11, no. 4, pp. 634–644, 2009.

[23] P. R. Alface, J.-F. Macq and N. Verzijp, "Evaluation of bandwidth performance for interactive spherical video", in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, IEEE, 2011, pp. 1–6.

[24] M. Yu, H. Lakshman and B. Girod, "A framework to evaluate omnidirectional video coding schemes", in *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*, IEEE, 2015, pp. 31–36.

[25] ——, "Content adaptive representations of omnidirectional videos for cinematic virtual reality", in *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, ACM, 2015, pp. 1–6.

[26] E. Kuzyakov and D. Pio. (2015). Under the hood: Building 360 video, [Online]. Available: https://code.facebook.com/posts/1638767863078802/under-the-hood-building-360-video/ (visited on 28/08/2016).

[27] ——, (2016). Next-generation video encoding techniques for 360 video and vr, [Online]. Available: https://code.facebook.com/posts/1126354007399553/next-generation-video-encoding-techniques-for-360-video-and-vr/ (visited on 28/08/2016).

[28] I. Bauermann, M. Mielke and E. Steinbach, "H. 264 based coding of omnidirectional video", in *Computer Vision and Graphics*, Springer, 2006, pp. 209–215.

[29] K.-T. Ng, S.-C. Chan and H.-Y. Shum, "Data compression and transmission aspects of panoramic videos", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 82–95, 2005.

[30] P. Tudor, "Mpeg-2 video compression", *Electronics & communication engineering journal*, vol. 7, no. 6, pp. 257–264, 1995.

[31] S Heymann, A Smolic, K Mueller, Y Guo, J Rurainsky, P Eisert and T Wiegand, "Representation, coding and interactive rendering of high-resolution panoramic images and video using mpeg-4", in *Proc. Panoramic Photogrammetry Workshop (PPW)*, 2005.

[32] F. Dai, Y.-d. Zhang, Y.-f. Shen and S.-x. Lin, "Transcoing-based data transmission of sphere panoramic videos", in *2006 8th international Conference on Signal Processing*, IEEE, vol. 2, 2006.

[33] A. Smolic and P. Kauff, "Interactive 3-d video representation and coding technologies", *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110, 2005, ISSN: 0018-9219. DOI: 10.1109/JPROC.2004.839608.

[34] H. Kimata, S. Shimizu, Y. Kunita, M. Isogai and Y. Ohtani, "Panorama video coding for user-driven interactive video application", in *Consumer Electronics, 2009. ISCE'09. IEEE 13th International Symposium on*, IEEE, 2009, pp. 112–114.

[35] A. Vetro, T. Wiegand and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the h. 264/mpeg-4 avc standard", *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011.

[36] M. Inoue, H. Kimata, K. Fukazawa and N. Matsuura, "Interactive panoramic video streaming system over restricted bandwidth network", in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1191–1194.

[37] H. Kimata, M. Isogai, H. Noto, M. Inoue, K. Fukazawa and N. Matsuura, "Interactive panorama video distribution system", in *Telecom World (ITU WT), 2011 Technical Symposium at ITU*, IEEE, 2011, pp. 45–50.

[38] M. Inoue, H. Kimata, K. Fukazawa and N. Matsuura, "Partial delivery method with multi-bitrates and resolutions for interactive panoramic video streaming system", in *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, IEEE, 2011, pp. 891–892.

[39] M. Inoue, H. Noto, Y. Tanaka, K. Fukazawa, H. Kimata, T. Mukouchi and N. Matsuura, "Field trial of interactive panoramic video streaming system", in *2011 IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, IEEE, 2011, pp. 104–107.

[40] H. Kimata, D. Ochi, A. Kameda, H. Noto, K. Fukazawa and A. Kojima, "Mobile and multi-device interactive panorama video distribution system", in *Consumer Electronics (GCCE), 2012 IEEE 1st Global Conference on*, IEEE, 2012, pp. 574–578.

[41] D. Ochi, Y. Kunita, K. Fujii, A. Kojima, S. Iwaki and J. Hirose, "Hmd viewing spherical video streaming system", in *Proceedings of the ACM International Conference on Multimedia*, ACM, 2014, pp. 763–764.

[42] D. Ochi, Y. Kunita, A. Kameda, A. Kojima and S. Iwaki, "Live streaming system for omnidirectional video", in *Virtual Reality (VR), 2015 IEEE*, IEEE, 2015, pp. 349–350.

[43] O. A. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual and S. Y. Lim, "Mpeg dash srd: Spatial relationship description", in *Proceedings of the 7th International Conference on Multimedia Systems*, ACM, 2016, p. 5.

[44] J. Redi, L. D'Acunto and O. Niamut, "Interactive uhdtv at the commonwealth games: An explorative evaluation", in *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, ACM, 2015, pp. 43–52.

[45] R. Guntur and W. T. Ooi, "On tile assignment for region-of-interest video streaming in a wireless lan", in *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*, ACM, 2012, pp. 59–64.

[46] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo and J. Wen, "Novel tile segmentation scheme for omnidirectional video", in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 370–374.

[47] D. Pohl. (2015). Smart rendering for virtual reality, [Online]. Available: http://blogs.intel.com/intellabs/2015/02/23/smart-rendering-virtual-reality/ (visited on 28/08/2016).

[48] Ns-3, [Online]. Available: https://www.nsnam.org/ (visited on 28/08/2016).

[49] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey", *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[50] A. Mavlankar and B. Girod, "Pre-fetching based on video analysis for interactive region-of-interest streaming of soccer sequences", in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, IEEE, 2009, pp. 3061–3064.

[51] D. Pang, S. Halawa, N.-M. Cheung and B. Girod, "Mobile interactive region-of-interest video streaming with crowd-driven prefetching", in *Proceedings of the 2011 international ACM workshop on Interactive multimedia on mobile and portable devices*, ACM, 2011, pp. 7–12.