Thèse n° 7172

Visual Attention and Perceptual Quality in Omnidirectional Imaging

Présentée le 31 janvier 2020

à la Faculté des sciences et techniques de l'ingénieur
Groupe Ebrahimi
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Evgeniy UPENIK

Acceptée sur proposition du jury

Prof. J.-Ph. Thiran, président du jury
Prof. T. Ebrahimi, directeur de thèse
Dr I. Ivanov, rapporteur
Prof. J. Wen, rapporteur
Dr J.-M. Vesin, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

"Well, in our country," said Alice, still panting a little,
"you'd generally get to somewhere else—if you run
very fast for a long time, as we've been doing."
"A slow sort of country!" said the Queen. "Now, here,
you see, it takes all the running you can do, to keep
in the same place. If you want to get somewhere
else, you must run at least twice as fast as that!"
— Lewis Carroll

. . . Assume a spherical cow, uniformly
emitting milk in all directions. . .
— a joke popular among physicists

Acknowledgements

First of all, I would like to sincerely express my boundless gratitude to Prof Dr Touradj Ebrahimi, my advisor and mentor, who has been guiding me on this thorny path of working toward my doctorate during previous four years. Prof Ebrahimi has not only supervised my scientific research, but helped me also to understand and to get adjusted to the academic environment after my years in industry. This work would not be done without his advice and support.

I would like also to gratefully thank the members of the jury: Prof Dr Jean-Philippe Thiran and MER Dr Jean-Marc Vesin from the neighboring signal processing laboratories at EPFL, Prof Dr Jiangtao Wen from Tsinghua University (Beijing), and Dr Ivan Ivanov from ELCA Informatique SA, for kindly agreeing to read these pages, provide their feedback, and listen to my presentation during the theoretical examination.

The work presented in this dissertation was funded by Swiss State Secretariat for Education, Research and Innovation SERI in the framework of ImmersiaTV project under the European Union Horizon 2020 research and innovation program (grant agreement no. 688619), and by The Swiss Commission for Technology and Innovation (CTI) under the grant 27403.1 PFES-ES.

Lausanne, 30 December 2019 E. U.

Abstract

Omnidirectional imaging has reached a level of widespread availability driven by recent advances in integrated circuit technology, image sensors, and computer graphics which now allow capturing, rendering and displaying of such type of immersive content in spatial resolutions sufficient to convey visual information directly to humans, as opposed to its previous use almost solely in computer vision for robotics and surveillance. In addition to its main property of covering full spherical field of view, omnidirectional imaging nowadays is an interactive multimedia; and, when experienced by means of virtual reality head-mounted displays, it achieves a remarkably high level of immersiveness. The paradigm, thus, has shifted toward human consumption of omnidirectional images and video.

Automatic prediction of salient regions in images is a well-developed topic in the field of computer vision. Yet, omnidirectional imaging brings new challenges to it, due to a different representation of visual information and additional degrees of freedom available to viewers. Having a model for visual attention in omnidirectional imaging is important to continue research in this subject. We develop such a model for interpreting experimental head-direction trajectories with a goal to construct a visual attention heat-map representing salient regions of an omnidirectional image. The developed model is further used in objective assessment of perceptual visual quality of omnidirectional visual content.

The problem of objectively measuring perceptual quality of omnidirectional visual content arises in many immersive imaging applications; and it is particularly important for compression and delivery. The interactive nature of this type of content limits the performance of earlier methods designed for static images or for video with a predefined dynamic. We aim to address a non-deterministic impact by using a statistical approach. To be specific, we attempt to describe and analyze viewer interactions in omnidirectional imaging through estimation of visual attention. We propose an objective metric to measure perceptual quality of omnidirectional visual content considering visual attention information.

Additionally, we explore certain related extensions and applications of omnidirectional imaging. Firstly, we investigate a possible extension to 3+ degrees of freedom by considering an individual case of rendering narrow baseline light filed images with limited translational interactions. We also provide results of extensive analysis of these iterations, including: circular histograms of directions of head movements, average vectors for a next perspective view, and charts of time spent on a view. Secondly, we look into privacy protection which is yet another field drawing more attention with the advances in image processing, visual and social media. We present a method for protecting user privacy in omnidirectional media, by removing parts

of the content selected by the user, in a reversible manner. Results on distinct contents indicate that our object removal methodology in the viewport domain enhances perceived quality, thereby improves privacy protection as users are able to hide objects with less distortion in the overall image.

Keywords: omnidirectional imaging, 360-degree images, visual attention, saliency maps, perceptual visual quality, objective metrics, virtual reality, image processing, multimedia signal processing

Résumé

L'imagerie omnidirectionnelle s'est beaucoup répandue grâce aux progrès récents de la technologie dans les domaines des circuits intégrés, des capteurs d'images et de l'infographie. Ces progrès permettent désormais de capturer, de rendre et d'afficher ce type de contenus immersifs avec des résolutions spatiales suffisantes pour transmettre des informations visuelles directement aux humains, alors qu'il était auparavant utilisé presque exclusivement en vision par ordinateur pour la robotique et la surveillance. En plus de son objectif principal qui est de couvrir tout le champ de vision sphérique, l'imagerie omnidirectionnelle est aujourd'hui un multimédia interactif; et, lorsqu'elle est expérimentée au moyen des casques de réalité virtuelle, elle atteint un niveau d'immersion remarquablement élevé. Le paradigme s'est ainsi déplacé vers la consommation humaine d'images et de vidéos omnidirectionnelles.

La prédiction automatique des régions saillantes dans les images est un sujet bien développé dans le domaine de la vision par ordinateur. Pourtant, l'imagerie omnidirectionnelle apporte de nouveaux défis à ce sujet, en raison d'une représentation différente des informations visuelles et des degrés de liberté supplémentaires disponibles pour le spectateur. Il est important de disposer d'un modèle d'attention visuelle en imagerie omnidirectionnelle pour poursuivre les recherches sur ce sujet. Nous développons un tel modèle pour interpréter les trajectoires de direction de la tête observées expérimentalement dans le but de construire une carte de fréquentation de l'attention visuelle représentant les régions saillantes d'une image omnidirectionnelle. Le modèle développé est en outre utilisé dans l'estimation objective de la qualité perceptuelle du contenu visuel omnidirectionnel.

Le problème de la mesure objective de la qualité perceptuelle de contenu omnidirectionnel se pose dans de nombreuses applications d'imagerie immersive, et il est particulièrement important dans la compression et la livraison. La nature interactive de ce type de contenu limite les performances des méthodes antérieures conçues pour les images statiques ou pour la vidéo avec une dynamique prédéfinie. Nous visons à aborder l'impact non déterministe en utilisant une approche statistique. En particulier, nous tentons de décrire et d'analyser les interactions des spectateurs en imagerie omnidirectionnelle en estimant l'attention visuelle. Nous proposons une mesure objective pour évaluer la qualité perceptuelle du contenu visuel omnidirectionnel en tenant compte des informations d'attention visuelle.

De plus, nous explorons certaines extensions et applications associées à l'imagerie omnidirectionnelle. Premièrement, nous étudions une extension éventuelle à 3+ degrés de liberté en considérant un cas individuel de rendu d'images plénoptiques de base étroite avec des interactions de translation limitées. Nous fournissons également les résultats d'une analyse

approfondie de ces itérations, notamment : des histogrammes circulaires des directions des mouvements de l'utilisateur, des vecteurs moyens pour une prochaine vue de perspective et des graphiques du temps passé sur une vue. Deuxièmement, nous nous penchons sur la protection de la vie privée, qui est un autre domaine qui attire davantage l'attention avec les progrès du traitement d'image, des médias visuels et sociaux. Nous présentons une méthode pour protéger la confidentialité des utilisateurs dans les médias omnidirectionnels, en supprimant de manière réversible des parties du contenu sélectionné par l'utilisateur. Les résultats sur des contenus distincts indiquent que notre méthodologie de suppression d'objet sur la fenêtre améliore la qualité perçue, améliorant ainsi la protection de la vie privée, car l'utilisateur est capable de masquer les objets avec moins de distorsion dans l'image globale.

Mots-clés : imagerie omnidirectionnelle, images à 360 degrés, attention visuelle, cartes de saillance, qualité visuelle perceptuelle, mesure objective, réalité virtuelle, traitement d'images, traitement du signal multimédia

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Omnidirectional imaging has reached a level of widespread availability following recent advances in integrated circuit technology, image sensors, and computer graphics which now allow capturing, rendering and displaying of such type of immersive content in spatial resolutions sufficient to convey visual information directly to humans, as opposed to its previous use almost solely in computer vision for robotics. Besides covering full spherical field of view, omnidirectional imaging nowadays is an interactive multimedia; and, when experienced by means of virtual reality head-mounted displays, it achieves a remarkably high level of immersiveness. The paradigm, thus, has shifted toward human consumption of omnidirectional images and video. This shift has occurred very recently, at once creating a great number of new research problems in this field.

Automatic prediction of salient regions in images is a well-developed topic in the field of computer vision. Yet, omnidirectional imaging brings new challenges to it, due to a different representation of visual information and additional degrees of freedom available to viewers. Analyzed previously only with estimation of eye gaze, now, head-movements must be also accounted, as well as eye-head coordination in humans. Having a model for visual attention in omnidirectional imaging is important to continue research in this subject. We aim to develop such a model for interpreting experimental head-direction trajectories with a goal to construct a visual attention heat-map representing salient regions of an omnidirectional image.

The problem of objectively measuring perceptual quality of omnidirectional visual content arises in many immersive imaging applications; and it is particularly important in compression and delivery. The interactive nature of this type of content limits the performance of earlier methods designed for static images or for video with a predefined dynamic. We aim to address a non-deterministic impact by using a statistical approach. More specifically, we attempt to describe and analyze viewers' interactions in omnidirectional imaging through estimation of visual attention. We propose an objective metric to measure perceptual quality of omnidirectional visual content considering visual attention information.

Additionally, we explore certain related extensions and applications in omnidirectional imaging. Firstly, we investigate a possible extension to 3+ degrees of freedom by considering an individual case of rendering narrow baseline light filed images with limited translational interactions. We provide also results of extensive analysis of those iterations, including: circular histograms of directions of user movements, average vectors for a next perspective view, and charts of time spent on a view. Secondly, we look into privacy protection which is yet another field drawing more attention with the advances in image processing, visual and social media. Photo sharing is a popular activity, which also brings the concern of regulating permissions associated with shared content. We present a method for protecting user privacy in omnidirectional media, by removing parts of the content selected by the user, in a reversible manner. Object removal is carried out using three different state-of-the-art inpainting methods, employed over the mask drawn in the viewport domain so that the geometric distortions are minimized. The perceived quality of the scene is assessed via subjective tests, comparing the proposed method against inpainting employed directly on the equirectangular image. Results on distinct contents indicate our object removal methodology on the viewport enhances perceived quality, thereby improves privacy protection as the user is able to hide objects with less distortion in the overall image.

1.2 Contributions

The present work includes the following contributions made by author to the fields of omnidirectional imaging and multimedia signal processing.

A method to obtain human visual attention data in virtual reality for omnidirectional content. We have proposed a method to obtain fixation locations and continuous fixation maps from head-direction trajectories for omnidirectional content in head-mounted virtual reality. The model incorporates analysis of head angular velocity and provides the idea of a generic solution to produce continuous fixation maps for omnidirectional images represented in panoramic projections. The saliency maps obtained from head position data can be a suitable first-order approximation when eye tracking data is not available. They can also be self-sufficient for interaction analysis and prediction. The results were published in a peer-reviewed conference paper [109].

A testbed for subjective evaluation of omnidirectional visual content. We proposed and demonstrated a testbed for subjective quality evaluation of omnidirectional visual content. The testbed allows researchers to perform experiments using different methods for subjective quality evaluations. The software implementation has a customizable storyboard and immersive menus for rating. Experimental data that can be obtained with this testbed includes subjective mean opinion scores, time spent on stimulus, and view direction tracks. The results were published in a peer-reviewed conference paper [111].

Evaluation of performance of objective metrics for omnidirectional visual content. We performed a subjective evaluation experiment on omnidirectional images. A total number of 45 observers were involved in the study, including 40 naïve and 5 expert participants. Subjective evaluation scores were obtained for 104 test stimuli. Seven objective metrics, among which three are specifically designed to assess omnidirectional visual content, were calculated for each stimuli. Rigorous performance evaluation has been carried out for objective quality metrics for omnidirectional visual content. Analysis of the obtained subjective and objective scores indicates moderate performance of investigated metrics for omnidirectional visual content. Being PSNR based, these metrics do not outperform significantly their ancestor in predicting visual quality of omnidirectional content. The results were published in a peer-reviewed conference paper [112].

Visual attention based objective metric for omnidirectional content. We proposed a new method called VA-PSNR which estimates perceptual quality of omnidirectional content considering visual attention. We validated our method against subjective MOS and benchmarked it against state-of-the-art objective metrics. VA-PSNR shows performance which is as high as the best alternative approaches based on PSNR. The results were published in a peer-reviewed conference paper [110] and presented at International Conference on Image Processing.

Coding of Omnidirectional Visual Content. We proposed OmniJPEG, a JPEG backward compatible solution to encode omnidirectional images. In order to ensure the JPEG backward compatibility, OmniJPEG extracts predefined regions of interest from omnidirectional images, as well as properties of equirectangular projection, while at the same time also keeping complete equirectangular information to preserve the capability of correctly rendering an omnidirectional image with appropriate devices and software. The results were published in a peer-reviewed conference paper [92].

Privacy issues in omnidirectional images. We presented a method for reversible object removal in omnidirectional images, which is targeted for privacy protection in immersive media. We show by performing subjective quality evaluation involving 16 naive subjects that viewport extraction can enhance the performance of state-of-the-art inpainting algorithms in omnidirectional images. The results were published in a peer-reviewed conference paper [108].

Investigation toward an extension of omnidirectional imaging to 3+ DoF. We developed a solution for rendering narrow baseline light field images in a virtual reality environment which allows interactions with their perspectives. A pilot subjective quality evaluation experiment for light field in virtual reality was conducted with 17 subjects participating in the assessments. The results of extensive analysis of the iterations include: circular histograms of directions of user movements, average vectors for a next perspective view, and charts of time spent on a view. The results were published in a peer-reviewed conference paper [113].

1.3 Organization of the thesis

The remainder of this dissertation is organized as following:

In Chapter 2 "Omnidirectional Imaging", we provide an introduction to the main concepts of omnidirectional imaging. We follow, then, with a review of the state of the art on acquisition, representation and visualization of omnidirectional images and video. And, finally, we formally describe the modalities of omnidirectional content and the relations between them.

In Chapter 3 "Visual Attention", firstly, we review prior research in visual attention and saliency for conventional images. We provide some insights, afterwards, on eye-head coordination in humans. Then we follow with a review of the state of the art on saliency in omnidirectional imaging. The chapter continues with a description of the proposed method for estimation of visual attention in virtual reality environment from experimental data. Then we describe experiments which were performed to estimate visual attention with the proposed method and present the results.

In Chapter 4 "Subjective Evaluation of Perceptual Visual Quality", we present a methodology to perform subjective evaluation of perceptual visual quality of omnidirectional visual content in head-mounted virtual reality. We describe, then, a testbed designed for such evaluations. The results of multiple experiments performed in order to assess subjective perceptual quality of omnidirectional visual content including mean opinion scores and viewing time statistics are finally presented in this chapter.

Chapter 5 "Objective Metrics for Perceptual Visual Quality" starts with a brief review of the state of the art in objective quality assessment for omnidirectional visual content. We describe, then, a methodology of computing existing objective metrics on a dataset of compressed omnidirectional images and provide results of benchmarking of those metrics against subjective mean opinion scores, followed by a discussion. In the second part of this chapter, we describe a novel approach to incorporate visual attention data into a full-reference objective perceptual visual quality measurement. We propose a metric which takes into account ground-truth viewer's visual attention information in order to make image quality assessment selective with respect to regions of interest. We continue with the results of validation of our method using subjective mean opinion scores and of benchmarking it against the existing metrics, followed by a conclusive discussion.

Chapter 6 "Coding of Omnidirectional Visual Content" begins with a brief review of coding and compression of omnidirectional visual content. We follow then with a description of a method for JPEG backwards compatible coding of omnidirectional images, and provide details on its architecture and implementation. Thereupon, the performance of the proposed coding scheme is evaluated.

Chapter 7 "Applications and Extensions of Omnidirectional Imaging" addresses the topics of privacy protection in omnidirectional imaging by improving perceptual visual quality of object

removal, and of an extension of omnidirectional imaging to 3+ degrees of freedom by rendering narrow baseline light field and analyzing viewers interactions.

Finally, Chapter 8 concludes this dissertation by summarizing the accomplishments of the present work and by discussing the directions of future research in this subject.

2 Omnidirectional Imaging

In this chapter, we provide an introduction to the main concepts of omnidirectional imaging. We follow then with a review on acquisition, representation and visualization of omnidirectional images and video. Finally, we formally describe the modalities of omnidirectional content and the relations between them.

2.1 Basic concepts

Omnidirectional imaging, as a subject in the field of multimedia signal processing, studies omnidirectional visual content, its acquisition, representation, rendering, processing, and analysis. This subject finds its roots in panoramic photography, which has been known and used for two hundred years and traces back to the first part of XIXth century [9].

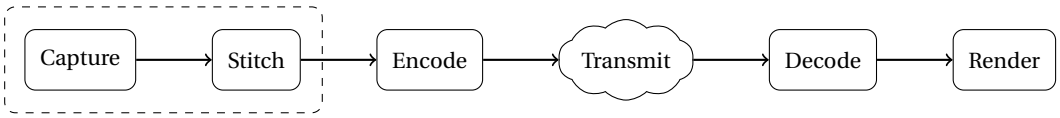


Figure 2.1 – General workflow of omnidirectional visual content

Omnidirectional visual content is an image or video signal which carries visual information for all directions observed from a single point of view when looking outwards. The entire field of view for omnidirectional images and video must cover 360 degrees horizontally and 180 degrees vertically, or a solid angle of 4π steradians. Colloquially, omnidirectional visual content is also referred to as 360-degree images and video. In some cases, panoramic visual content covering only a part of a full sphere may also be considered as an object of study in omnidirectional imaging.

The main stages of the omnidirectional imaging workflow are depicted in Figure 2.1. It starts with acquisition consisting of capturing and stitching of omnidirectional visual content. This stage is typically followed by encoding, which is a crucial step required for transmission or storing of any kind of information today. Decoding and rendering are performed during consumption of omnidirectional visual content. One can notice here two stages which are not usually present in conventional imaging: namely, stitching and rendering. The former is required in many cases in order to obtain an omnidirectional image, and the latter is an unavoidable step in displaying due to the interactive nature of this type of multimedia.

When working with digital images which are intrinsically spherical we need to define a representation and an arrangement of pixels in a data structure. Historically, in image processing we deal with rectangular pictures. Hence, the legacy of all algorithms and approaches in the field obliges us to treat omnidirectional images as matrices. Thus, a **planar rectangular representation** (also called planar projection) is the most suitable for storing and interpreting these data. There are many approaches to map a spherical surface to a plane including certain of them which are two thousand years old. In omnidirectional imaging, two projections are most commonly used: namely, an equirectangular and a cube map. The former has its origin in cartography and can be easily recognized by one as a world map. The latter comes from the field of computer graphics where it has been used for environmental mapping. See Section 2.3 for more details. In this work, we consider a planar rectangular form of an omnidirectional image (such as equirectangular or cube map) to be the given input signal. The preceding stages of acquisition and stitching are reviewed further in this chapter in Section 2.2; they fall, however, outside of the scope of the present study.

Consumption of omnidirectional visual content differs rather notably from that of conventional planar images. The human visual field of recognition is restricted by 46° of eccentricity horizontally and 32° vertically [100, Fig. 14], which results in an effective field of view of $92° \times 64°$. This is not enough to cover a full sphere, thus an observer, while consuming omnidi-

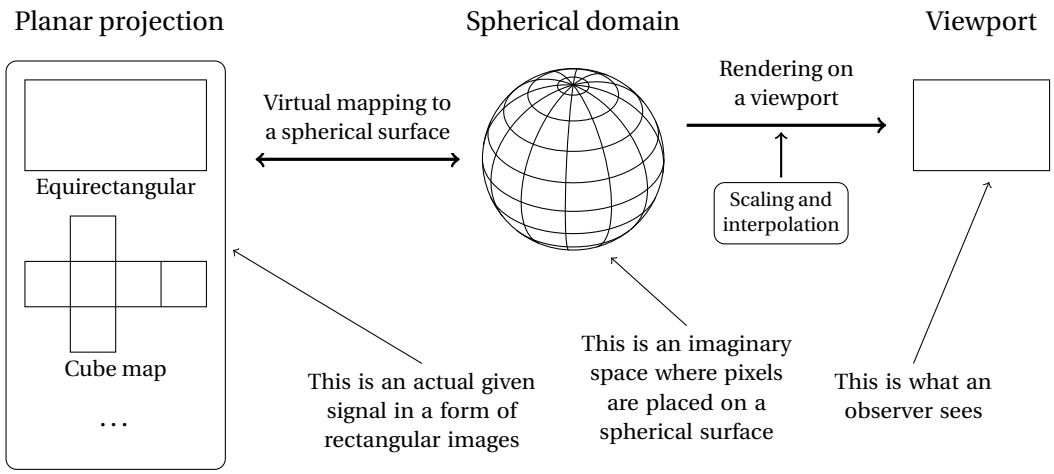


Figure 2.2 – Main concepts in omnidirectional imaging

rectional visual content, only sees a part of it at one moment, whether it is rendered on a full dome or on a smaller screen. Typically, omnidirectional visual content is viewed by means of a head-mounted display, a hand-held device, or a regular screen with assisting interaction controls. The part of an entire omnidirectional image which is exposed to a viewer at one moment in time and its viewing window are called a **viewport**. It is commonly rendered using computer graphics software and hardware. More details about displaying omnidirectional visual content can be found in Section 2.4.

Figure 2.2 depicts the relations between the planar projection, the spherical representation, and the viewport of an omnidirectional image. One should notice that an input signal in the form of a planar projection can be reversibly mapped to a virtual spherical domain, in order to simplify understanding of processing. Rendering on a viewport, on the contrary, is irreversible, due to the loss of information during possible re-scaling and the noise added by interpolation.

2.2 Acquisition and stitching

In this section, we review the state of the art in acquisition and stitching stages of omnidirectional imaging workflow. Being a specific type of visual content, omnidirectional images are acquired with the help of a particular type of cameras and camera systems which may require extensive post-processing in order to obtain a final picture.

2.2.1 Omnidirectional cameras and camera systems

Early predecessors of today's omnidirectional cameras are panoramic photo cameras [9]. The main technique employed to acquire panoramic photographs was rotation of the camera by means of a pan-tilt mechanism in order to keep the optical axes aligned. This technology only

allowed to capture static pictures, and thus could not be applied for video. The complexity of post-processing which required manual adjustments was also a significant drawback.

Modern omnidirectional cameras first attracted the attention of the academic community working in computer vision and image processing during the last decade of the XXth century [84, 89, 42], when availability of extra-wide-angle and omnidirectional optics, digital sensors, and sufficient computational resources finally allowed capturing of digital images in panoramic projections and manipulating them within reasonable time. Research on omnidirectional imaging and vision was mostly targeting such applications as surveillance, robotic vision, and video conferencing, due to the still existing limitations of those days technology.

Cameras used in acquisition of omnidirectional images and video can be divided into three main classes: catadioptric, dioptric extra-wide-angle, and polydioptric or multi-camera systems. Catadioptric cameras along with lens use a mirror of parabolic, hyperbolic or elliptical shape. They provide 360-degree in horizontal plane and up to 270 degrees in altitude angle. Dioptric cameras use a single lens with an extremely wide angle of view, which is colloquially known as a fish-eye lens. Polydioptric and multi-camera systems use multiple lenses or cameras to capture images in all directions with overlapping regions among different sensors.

A concept of a **catadioptric omnidirectional camera** covering a hemispheric field of view in application to television was patented [90] by Rees in 1970 and later extended by Nayar [84] to cover the full spherical field of view in 1997. The term catadioptric refers to an optical system which forms an image by means of both lenses and mirrors: refraction and reflection; such technique is often used in telescopes. A thorough review of a class of single-mirror catadioptric omnidirectional cameras was presented by Baker and Nayar in [7]. Bruckstein and Richardson independently summarized their findings on the same topic in [17]. It is important to notice that catadioptric cameras require a sophisticated procedure of calibration [37, 130]. This became a significant disadvantage in comparison to alternative acquisition systems for omnidirectional imaging with the advances in lens production technology and image processing.

The use of **extra-wide-angle refractive-only lenses**, known also as fish-eye lenses, is an alternative approach to capture omnidirectional visual content with a single-lens camera. It simplifies the configuration of the optical part of the acquisition workflow and does not increase computational complexity in the image processing part. The widest known angle of view achievable by a refractive-only optical system was reported by Martin in [81]. His hyperfield fish-eye lens has an unvignetted field of view of 310 degrees. Other single lens refractive-only systems include a 270-degree wide fish-eye lens patented by Nikon [57], as well as, commercially available today, lenses with 250-[1] and 270-[2] degree field of view. The main drawback of fish-eye-lens systems is the impossibility of covering a full spherical field of view, even though the blind area can be very small. Geometrical distortions and low optical

[1] https://products.entaniya.co.jp/en/products/hal-250200/
[2] https://www.lensrentals.com/blog/2019/02/assembling-the-c-4-optics-hyperfisheye-prototype/

resolution at the edge of the field of view impair also the quality of a final image.

Finally, the most recently developed technology for capturing omnidirectional visual content is **multi-camera systems**. An acquisition apparatus of this class consist of several mechanically fixed directional cameras installed on a mount and facing outwards in order to cover full spherical field of view. Such a system can be either a one-piece device or manually assembled equipment. Early examples of modern omnidirectional multi-camera systems can be traced back to the last decade of the XXth century. A seamless multi-camera omnidirectional imaging system was patented by Henley in 1997 [50]. Six years later, Ikeda et al. in [56], introduced a telepresence solution based on a six-lens omnidirectional multi-camera system. Other examples of omnidirectional multi-camera systems can be found in [39, 96].

The most important advantage of using omnidirectional multi-camera systems consist in significantly higher spatial resolution of the final image, due to merging of multiple pictures from multiple image sensors. The main drawback, however, is the necessity of computationally heavy post-processing required to stitch together images from all the cameras.

2.2.2 Stitching as part of omnidirectional acquisition

As it was described in Section 2.2.1, there exist multiple ways to capture omnidirectional visual content. The most preferred one today involves employing multi-camera systems, for the reason that it provides the highest possible spatial resolution and a true full spherical field of view. Images from individual directional cameras are combined in order to obtain a final omnidirectional picture. This process is called **image stitching**.

Methods for seamless alignment and stitching of images into photo-mosaics are among the oldest in topics of image processing and computer vision [104]. The goal is to merge several photographs which have an overlapping field of view into a single panoramic image while minimizing the amount of artifacts. The differences in illumination, color and geometry of the pictures must be compensated prior to stitching *per se*. The images then need to be aligned and seam paths should be defined. Blending is finally performed in order to smooth seam borders.

When applied to omnidirectional imaging, stitching includes the following essential stages [78]. Firstly, pictures from each directional camera undergo a compensation of lens distortion [104]. Feature detection and matching is used then to find corresponding points in the overlapping regions of the images [75, 8]. Next, each input picture is warped to conform to a target projection [128, 21], which can be equirectangular, cube map or other. Exposure compensation and color correction are then applied in order for all the images to have uniform brightness level and matching color temperature [114, 120, 36]. Afterwards, several different methods can be used in order to find an optimal seam path in the overlapping region [65, 24]. Image blending is finally applied to the seams in order to smooth residual borders [18, 87, 35, 105].

The final result of stitching is an omnidirectional image or a video frame represented in a

(a) Spherical view

(b) Equirectangular projection

Figure 2.3 – Example of an omnidirectional image represented in equirectangular projection.

targeted panoramic projection.

2.3 Planar representations

Omnidirectional visual content, despite being intrinsically spherical, must be represented in the form of a planar rectangular image in order to be stored, processed and transmitted. Such representations are called planar projections. The history of representing a sphere as a plane goes back to first attempts in cartography to draw a map of the globe on paper [98]. The main challenges in finding a perfect way to project a sphere to a plane are the introduced geometrical distortions resulting in alternated shapes of objects and mismatching distances and areas [43]. Only two planar projections are widely used in omnidirectional imaging: namely, equirectangular projection and cube map projection.

2.3.1 Equirectangular projection

Equirectangular projection has been known in cartography for almost two thousand years with a first mention *circa* 100 CE [98]. It is a cylindrical equidistant projection (also called plate carré) where the horizontal coordinate is the longitude and the vertical coordinate is the latitude. Coordinates in equirectangular projection have their origin at $0^\circ N\, 0^\circ E$ in terms of geographical position. Figure 2.3 shows an omnidirectional image wrapped around a sphere and its equirectangular representation.

Due to distortions introduced by equirectangular projection, it is not typically used in navigation or in real-estate land recording nowadays. Equirectangular projection has found its place, however, in geographic information systems, where it is now a standard *de facto* for displaying and storing data [80], on account of a simple correspondence between pixels in images and geographic coordinates.

In Figure 2.4, one can see a visualization of the distortion introduced by equirectangular projection depicted with the help of the Tissot indicatrix, an imaginary circle of a perfect symmetrical shape on the surface of a sphere which stretches to a form of ellipse when projected onto a plane. One can see that the circles in the near-equator area are almost perfectly symmetrical, whilst when moving close to the poles they become ellipses with a noticeable eccentricity.

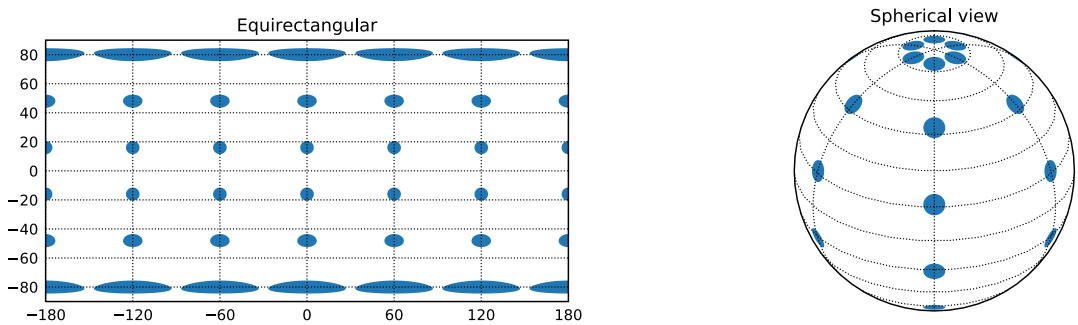


Figure 2.4 – Equirectangular projection (left) and spherical view (right). Tissot indicatrix (blue) shows geometrical distortions in equirectangular projection. Longitude $\theta \in [-180, 180]$ degrees. Latitude $\varphi \in [-90, 90]$ degrees.

Equations 2.1 and 2.2 describe how to obtain Cartesian coordinates on the plane from spherical coordinates for the more general case of equidistant cylindrical projection and for equirectangular projection, respectively.

$$\begin{aligned} x &= (\theta - \theta_0)\ \cos\varphi_0 \\ y &= \varphi \end{aligned} \tag{2.1}$$

$$\begin{aligned} x &= \theta \\ y &= \varphi \end{aligned}, \quad \text{for } \varphi_0 = 0 \text{ and } \theta_0 = 0 \tag{2.2}$$

where x, y are the Cartesian coordinates on the plane for equirectangular representation, and θ, φ are the longitude and latitude, respectively, in spherical coordinates. The parameters θ_0, φ_0, in Equation 2.1 specify the shift of the origin if the map is not centered at the crossing of the equator and zero-meridian.

The inverse transform from Cartesian coordinates on a plane to equidistant cylindrical and

equirectangular projections, respectively, is performed according to Equations 2.3 and 2.4.

$$\begin{aligned} \theta &= \theta_0 + \frac{x}{\cos\varphi_0} \\ \varphi &= y \end{aligned} \tag{2.3}$$

$$\begin{aligned} \theta &= x \\ \varphi &= y \end{aligned}, \quad \text{for } \varphi_0 = 0 \text{ and } \theta_0 = 0 \tag{2.4}$$

The equirectangular projection is the most widely used representation for omnidirectional visual content for the moment. Despite its disadvantages of introducing the geometrical distortion and overusing the pixels in the near-pole areas, it has been universally adopted by industry and is now a default format for the vast majority of capturing, coding and displaying software-hardware solutions. The equirectangular representation owes its success to the comprehensive visualization of a scene as a panoramic picture and to the continuity of the visual information in the frame which is beneficial for compression and rendering.

2.3.2 Cube map projection

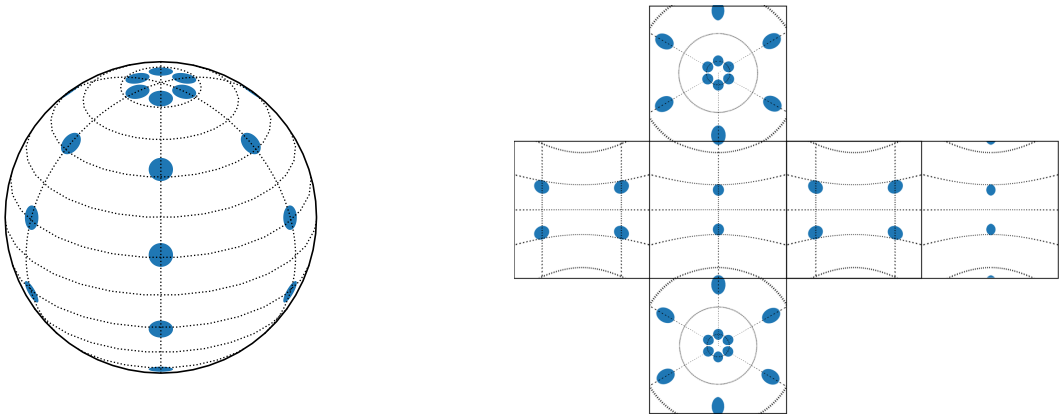
The cube map projection in omnidirectional imaging takes its name and origin from the field of computer graphics, where it appeared in 1984 as a solution to the problem of environmental texture mapping, as opposed to a more complex way of achieving the result using ray tracing. An unfolded cube projection of a 3-dimensional environment was proposed by Greene in [44].

(a) Spherical view (b) Cube map projection

Figure 2.5 – Example of an omnidirectional image represented as a cube map

Essentially, due to the fact that rendering of omnidirectional images is a problem of computer

graphics, cube map projection was an obvious solution to display and therefore store omni-directional visual content. It is also in its favor that all graphic processing units of modern consumer computers have a hardware support of cube map rendering.



(a) Spherical view

(b) Cube map projection

Figure 2.6 – Cube map projection (right) and a spherical view (left). Tissot indicatrix (blue) shows geometrical distortions in cube map projection.

Geometrical distortions of cube map projection are far less severe that the ones introduced by equirectangular projection as one can clearly see from Figure 2.6: only on the sides of the cube facets Tissot indicatrices are stretched, and to much lesser extent than we observe in Figure 2.4 for the equirectangular projection.

Formulae for obtaining a cube map from a spherical surface derive from the Rectilinear (also called Gnomonic) projection [98, 118]. The transform of the coordinates for each cube face is performed according to Equation 2.5:

$$\left.\begin{aligned} x &= \frac{\cos\varphi \sin(\theta-\theta_0)}{\cos c} \\ y &= \frac{\cos\varphi_0 \sin\varphi - \sin\varphi_0 \cos(\theta-\theta_0)}{\cos c} \end{aligned}\right| \quad \theta \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right] \text{ and } \varphi \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right] \tag{2.5}$$

where

$$\cos c = \sin\varphi_0 \sin\varphi + \cos\varphi_0 \cos\varphi \cos(\theta-\theta_0)$$

and θ, φ are the longitude and latitude, respectively, in spherical coordinate system; and θ_0, φ_0 define a vector pointing from the center of the sphere to the center of a current cube face.

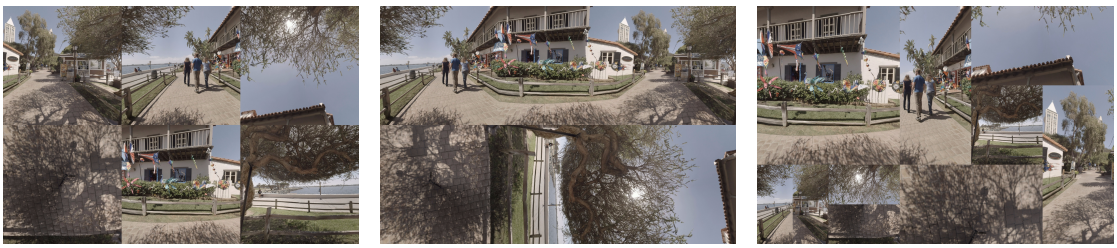
The inverse mapping is described in Equation 2.6.

$$\begin{aligned} \varphi &= \arcsin\left(\cos c\ \sin\varphi_0 + \frac{y\ \sin c\ \cos\varphi_0}{\rho}\right) \\ \theta &= \theta_0 + \arctan\left(\frac{x\ \sin c}{\rho\cos\varphi_0\ \cos c - y\ \sin\varphi_0\ \sin c}\right) \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2} \\ c &= \arctan\rho \end{aligned}$$

In practice, however, Equations 2.5 and 2.6 can be simplified by using a two-argument arctangent function, namely, *arctan2(y,x)*. The Rectilinear projection can also be interpreted as a *pinhole camera* forming an image on a cube face. In this case, the transforms are done in two steps: firstly, points are mapped from the spherical coordinate system to the Cartesian one; and, afterwards, they are projected to the image plane using the pinhole camera model. The latter method is a typical solution used for software implementations.



(a) Cube map composed as 3x2 (b) Cube map with rotated faces (c) Half-resolution back

Figure 2.7 – Variations of cube map projection

The cube map projection has many variations which derive from rearranging and rotating cube face in a final rectangular image. In Figure 2.7 one can see examples of different modifications of cube map projection. Assessment of an impact of a type of cube map projection on codding efficiency can be found in [132].

The cube map projection has much lighter distortions (see Figure 2.6) comparing to equirectangular projection. Moreover, cube mapping is implemented in hardware in all modern graphic processing units, thanks to its long history in computer graphics. All these advantages, nonetheless, did not allow the cube map projection to become a dominant format for omnidirectional visual content. The reasons may not be obvious at a fist glance, but the drawbacks are the following: discontinuity in the planar image reduces performance of compression methods designed for conventional images and video; the visual representation is less comprehensive than in the case of equirectangular projection, which is an important property for

production, monitoring and editing omnidirectional visual content.

2.3.3 Other projections

There exist other planar representations of omnidirectional visual content besides equirectangular projection and cube mapping. Those representations, however, are rarely used in practice. We attempt to provide a short review of the most known of them in the following paragraph for the sake of completeness.

An extensive and thorough review of planar representations in omnidirectional imaging can be found in [23], where Chen et. al. propose their classification of projections and provide a wide variety of examples. Polyhedron-based projections map a spherical image to the faces of different convex polyhedra with the goal to reduce pixel redundancy. Examples of such approaches can be found in [38, 71, 72, 5]. Tile-based projections divide a frame into several stripes or tiles; typically it is performed in order to efficiently use a planar frame while preserving quasi-constant pixel density in spherical domain. Tile-based projections has been proposed in [125, 70, 117, 129]. Viewport dependent projections intend to exploit intentionally introduced anisotropy in spatial resolution by giving more pixels to presumably more probable viewing directions [3, 30].

2.4 Displaying omnidirectional visual content

Omnidirectional visual content is a particular form of immersive multimedia which extends conventional image and video sensations to a three-dimensional space by providing full-spherical coverage of field of view and allowing change-of-sight interactions. This type of content is typically consumed using virtual reality head-mounted displays, hand-held devices, and, less frequently, conventional displays of personal computers. Viewers perform interactions by moving their heads, displacing and rotating an accelerometer-equipped device or by means of direct controllers such as computer mice, trackpads and touchscreens.

The interactivity is a property of omnidirectional visual content as well as other immersive multimedia which distinguishes them drastically from conventional images and video. The information acquired by a viewer during consumption of such content may be affected by their interaction, thus, making the experience personal.

2.4.1 Displays and level of immersiveness

The immersive quality of omnidirectional imaging, most certainly, originates from its displaying and rendering technologies. An increase in the power of commonly available computational resources, a significant breakthrough in LED and OLED display technology, and the arrival of consumer devices equipped with motion and direction sensors led to a dramatic raise in the amount of hardware which are capable of not only outputting visual information,

but also of instant acquisition of signals from user interactions.

Omnidirectional visual content currently can be consumed by means of the following three visualization technologies: namely, conventional displays, hand-held devices, and head-mounted displays, placed here in the order of an increasing level of immersiveness.

Conventional displays are the least immersive of the technologies capable of rendering omnidirectional visual content. A personal desktop computer or a laptop can run certain software in order to visualize an omnidirectional image. The viewport is represented as a dynamic window which includes only a part of an entire image corresponding to the current field of view. Thanks to the fact that every personal computer is equipped with direct input devices, such as a keyboard, a mouse, a touchpad, or a touchscreen, a user is able to manually change the direction of sight for a current viewport. Conventional displays provide three degrees of freedom. However, the viewer is not isolated from the environment and must control the interactivity explicitly using input devices.

Hand-held devices act as a window to look through which users can move around themselves in order to observe different parts of an omnidirectional image. The range in size of this type of devices expands from 5-inch-screen mobile phones up to 13-inch-screen tablet computers. Virtually all hand-held devices carry an on-board system for accurate enough estimation of attitude, absolute orientation, and position. This typically can be achieved by means of the following types of embedded sensors: accelerometers, gyroscopes, and magnetometers. Usually, hand-held devices do not have any local reference of position and orientation other than a GPS module and a compass. Hence, even though, in general, hand-held devices provide six degrees of freedom, the translational motion cannot be accurate. Moreover, similarly to conventional displays, hand-held devices do not isolate the viewer from the environment, which decreases the level of immersiveness.

Head-mounted displays provide the most immersive abilities for consuming omnidirectional visual content. Displays of this type are worn by a user and are able to accurately capture the position and the orientation of the viewer's head. From an engineering perspective, head-mounted displays have different designs. They can be tethered or untethered (wireless). Devices of the former class act typically as an external display and are driven by a graphics processing unit of a personal computer. Thus all the computations for rendering are performed externally with respect to the device. This gives this type of HMD a significant advantage in terms of computational performance.The latter type of devices can have an embedded screen or can exploit a hand-held device as one. Properties of hand-held based HMD are similar to the hand-held devices themselves, with an exception that the user is isolated from the environment. Moreover, most of the HMD exploit locally referenced positioning system, which improves accuracy. This helps to provide six degrees of freedom with better user experience. HMD fully isolate viewers from the environment and put them into a virtual reality. The interactivity is intrinsically

implemented and is completely transparent for users, which ensures the highest level of immersiveness among all the devices capable of rendering omnidirectional visual content.

2.4.2 Types of interactions and degrees of freedom

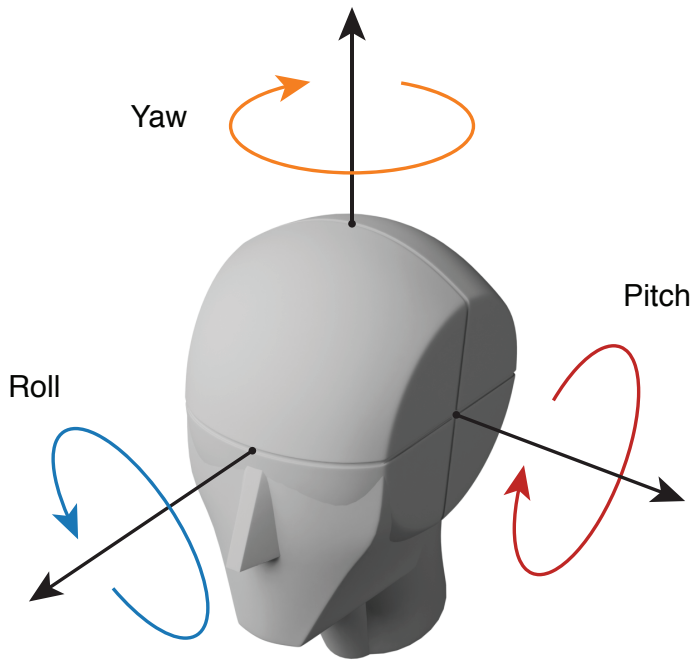


Figure 2.8 – Head rotational position: yaw, pitch, and roll.

Let us list and classify the types of interactions possible when consuming omnidirectional visual content. Being a dynamic type of multimedia, omnidirectional images and video require an instant set of parameters at each moment in time in order to render the viewport representing a current field of view.

Head rotational position is crucial and non-optional information required to render any omnidirectional image of video. All devices must have the ability to capture head rotational position either directly or indirectly by means of the user input. Head rotational position is represented in three coordinates: namely, yaw, pitch, and roll. Yaw is a horizontal angular displacement from the initial position. Pitch is an angle of elevation from the horizon. And Roll is the angle of rotation about the vector representing the person's direction of sight (See Figure 2.8). This set of three coordinates along with information about available field of view defines unambiguously the instant viewport of an omnidirectional image or video. Head rotational position provides three degrees of freedom.

Eye movements are optional information, which can be used in, for example, foveated rendering, or for statistical analysis of user interactions. In order to capture eye movements, additional equipment is required, since most of the devices do not provide this option by default. The process of eye movements acquisition is non-trivial and needs special calibration procedures for every use, which restrict its application to only controlled laboratory environments. Eye movements add two optional degrees of freedom to a required head rotational position by describing a vector of eye gaze.

Range-limited head movements with a stationary body position are used in omnidirectional imaging enhanced, for instance, with a depth map. This type of immersive omnidirectional content is considered to provide so called 3+ degrees of freedom, because the translational movements of the head are limited to a certain extent which may not exceed the range achieved with a fixed position of the viewer's body.

2.5 Formal definitions

In this section, we aim to formally define the modalities in omnidirectional imaging and the relations between them. This includes describing the planar representation, the spherical domain, and the viewport domain. We also provide insights on the irreversible changes of the signal during rendering.

2.5.1 Planar projection as effective signal and its mapping to sphere

Despite the fact that omnidirectional images and video are intrinsically spherical, they are stored and transmitted in a form of planar rectangular pictures. That is to say, our input signal is a normal conventional image, and the only difference occurs in the way how we operate with it.

Let us define a **planar representation** of an omnidirectional image as a tensor:

$$\mathbf{I}^P \in \mathbb{R}^{M \times N}, \quad M, N \in \mathbb{N} \tag{2.7}$$

The elements of the tensor $\mathbf{I}^P$ can be defined by a continuous function $f_P(x, y)$:

$$\mathbf{I}^P_{m,n} = f_P(x_m, y_n), \quad \text{where} \quad f_P : \mathbb{R}^2 \to \mathbb{R}, \quad \text{and} \quad m \in \{1, \dots, M\},\ n \in \{1, \dots, N\} \tag{2.8}$$

M and N are the dimensions of the planar image; x_m and y_n are the coordinates of a pixel $(m,\ n)$ in the continuous domain. For the sake of simplicity, here, we only consider monochrome images.

We should not forget, however, that omnidirectional visual content is spherical by nature, and, thus, we may benefit from switching to the spherical domain in order to simplify notations of processing algorithms.

The **spherical representation** of the image defined in Equation 2.7 will be:

$$\begin{aligned} &\mathbf{I}^S \in \mathbb{R}^{M\times N} \\ &\mathbf{I}^S_{m,n} = f_S(\theta_m, \varphi_n), \quad \text{where} \quad f_S : \mathbb{S}^2 \to \mathbb{R},\ (\theta,\varphi) \mapsto f_S(\theta,\varphi), \end{aligned} \tag{2.9}$$

where

$$\mathbb{S}^2 := \left\{(x,y) \in \mathbb{R}^2 \,\middle|\, x^2 + y^2 = 1\right\} \tag{2.10}$$

is a 2-dimensional manifold, or a spherical surface, and $(\theta,\ \varphi,\ \rho = 1) \in \mathbb{S}^2$ are spherical coordinates of all the points in $\mathbb{S}^2$.

The spherical coordinates of each point in $\mathbf{I}^S$ are calculated as:

$$\begin{aligned} &\theta_m = \theta_m(m), \quad m \in \{1,\dots,M\} \\ &\varphi_n = \varphi_n(n), \quad n \in \{1,\dots,N\} \end{aligned} \tag{2.11}$$

For the equirectangular projection case we have:

$$\begin{aligned} &\theta_m = 2\pi\frac{m}{M}, \quad m \in \{1,\dots,M\} \\ &\varphi_n = \pi\frac{n}{N}, \quad n \in \{1,\dots,N\} \end{aligned} \tag{2.12}$$

2.5.2 Viewport domain and pixels on screen

The viewport is a rectilinear projection of the spherical image to a screen, and it is defined by two pairs of parameters: namely, the direction of sight (θ_0, φ_0) (or a viewport position) and the field of view $(\alpha_{\mathrm{H}}, \alpha_{\mathrm{V}})$ (or a viewport coverage).

An image representing a viewport when it is displayed on a screen is defined as:

$$\mathbf{I}^V \in \mathbb{R}^{P\times Q}, \quad \text{where} \quad P, Q \in \mathbb{N} \tag{2.13}$$

with P and Q being the dimensions of a viewport in pixels. Thus, if $f_V(u,v)$ is a continuous representetion of the viewport image, we have:

$$\mathbf{I}^V_{p,q} = f_V(u_p, v_q) \quad \text{where} \quad f_V : \mathbb{R}^2 \to \mathbb{R},\ (u,v) \mapsto f_V(u,v) \tag{2.14}$$

The viewport image $f_V(u,v)$ for a direction of sight (θ_0, φ_0) and a window angular size of $(\alpha_{\mathrm{H}}, \alpha_{\mathrm{V}})$ can be obtained from a spherical image $f_S(\theta,\varphi)$ as:

$$f_V(u,v) = \mathscr{F}_V(f_S(\theta,\varphi), \theta_0, \varphi_0, \alpha_{\mathrm{H}}, \alpha_{\mathrm{V}}) \tag{2.15}$$

where operator $\mathscr{F}_V$ maps virtual pixels on a sphere to virtual pixels in the viewport, and (u,v) are the continuous coordinates in the viewport domain $u \in [0,1],\ v \in [0,1]$.

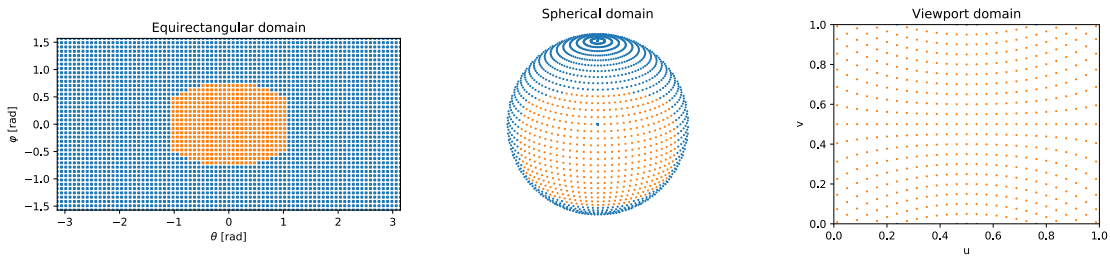
The correspondence between (θ,φ)-domain and (u,v)-domain must be known in order to perform the transform (2.15) and is described in Equations 2.16 and 2.17.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & \frac{k}{x} & 0 \\ 0 & 0 & \frac{m}{x} \end{bmatrix} R_Z^{\varphi_0} R_Y^{\theta_0} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \Bigg|_{x>0} \tag{2.16}$$

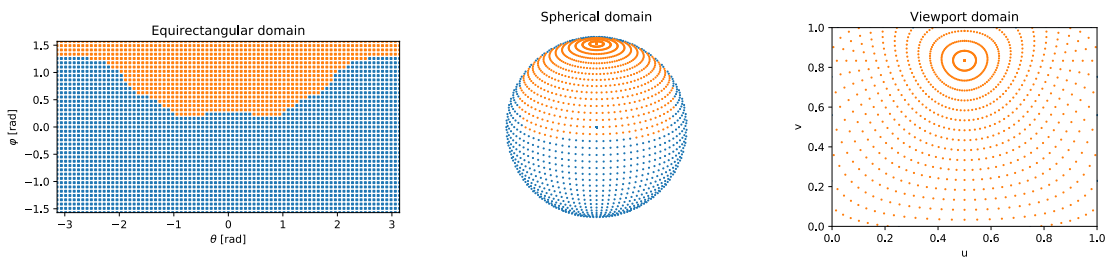
where k and m are scaling coefficients for viewport coordinates, $R_Y^{\theta_0}$ and $R_Z^{\varphi_0}$ are rotations corresponding to yaw and pitch respectively, and vector $[x\ y\ z]$ represents Cartesian coordinates of a point on the sphere, which are derived from the spherical coordinates as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin\varphi\cos\theta \\ \sin\varphi\sin\theta \\ \cos\varphi \end{bmatrix} \tag{2.17}$$

assuming that the radius of the sphere equals 1.



(a) Pixels of equirectangular image in the viewport in the near-equator area.

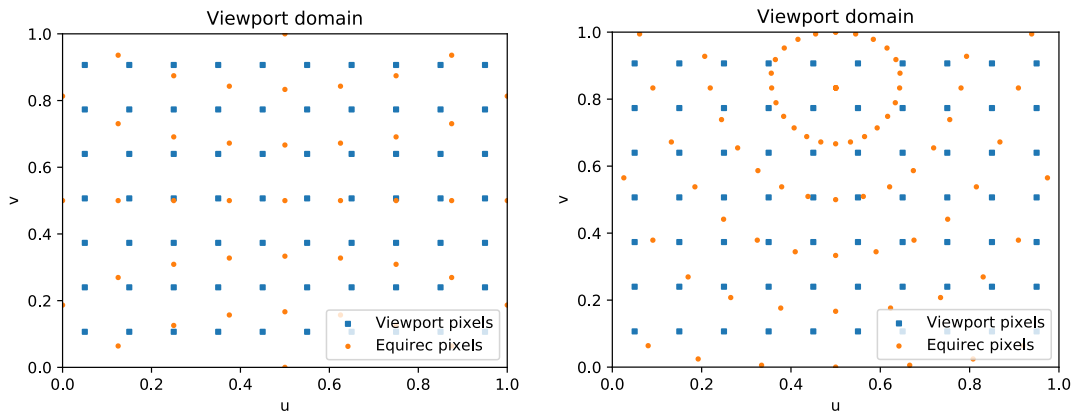


(b) Pixels of equirectangular image in the near-pole area.

Figure 2.9 – Changes in pixel density in the viewport.

It is important to notice that the given input signal is always a rectangular digital image in discrete coordinates. That is to say, not every point is defined for the planar representation $f_P(x, y)$. Let us consider an example where an equirectangular image is given, and we would like to extract a viewport from it. There can be two extreme cases for viewport positions: namely, near-equator and near-pole. In Figure 2.9, one can see how these two extreme cases affect the amount of available information after we move to the viewport domain. The density of original pixels is notably higher if an observer looks up, so the viewport holds a near-pole position.

Another aspect of viewport rendering which cannot be left unattended is interpolation. A viewport is displayed on a screen whether it is an HMD, a hand-held device or a regular monitor. Therefore, this screen has a finite discrete pixel grid which virtually never corresponds exactly to the positions of mapped original pixels. Thus, the values of the pixels in the grid of a viewport screen must be estimated from the original pixels or, in other words, interpolated. Figure 2.10 shows an example of two viewports with screen pixel grid depicted in blue and original pixel values depicted in orange. One can see from this figure how a viewport position can affect a rendered image. The step of interpolation is irreversible. One cannot reconstruct the original image (or its part) from a viewport.



(a) near-equator area

(b) near-pole area

Figure 2.10 – Interpolation during rendering of a viewport. Input pixels of equirectangular image (orange) in comparison to the rendered pixels of a viewport (blue).

3 Visual Attention

Automatic prediction of salient regions in images is a well developed topic in the field of computer vision. Yet, virtual reality omnidirectional visual content brings new challenges to this topic, due to a different representation of visual information and additional degrees of freedom available to viewers. Having a model for visual attention is important to continue research in this direction. In this chapter, we develop such a model for head direction trajectories. The method consists of three basic steps: Firstly, a computed head angular speed is used to exclude the parts of a trajectory where motion is too fast to fixate viewer's attention. Secondly, fixation locations from different subjects are fused together, optionally preceded by a re-sampling step to conform to the equal distribution of points on a sphere. And finally, a

Gaussian based filtering is performed to produce continuous fixation maps. The developed model is used to obtain ground truth experimental data when eye tracking is not available.

3.1 Review of visual attention in VR

Omnidirectional images and video are typically consumed using a virtual reality (VR) head-mounted display (HMD). Visual content represented in one of the projections is rendered on a viewport of an HMD where data from acceleration and orientation sensors is used to define which part of the content is to be displayed. This data, if stored, may then be used for analysis of human visual attention in VR imaging.

3.1.1 Prior research in visual attention and saliency

Computational prediction methods for human visual attention have been studied for decades in conventional flat images. The first theoretical computational model of human visual attention was introduced by Koch and Ullman in [61], and the first practical implementation was presented by Clark and Ferrier in [25]. Detailed descriptions and classifications of state-of-the-art visual attention models can be found in [13, 14, 59]. There exist two main approaches for modeling human visual attention, namely, bottom-up and top-down. The former starts by computing different features in images, typically intensity, color and orientation characteristics. These features are then fused together to produce a saliency map. The latter approach takes into account certain high level information about the scene which is used, for example, by incorporating face, object, and text detection. Top-down methods are often combined with bottom-up models.

Experimental visual attention data, unlike prediction models, does not provide saliency maps. After initial processing, one can obtain fixation locations, i.e. points in the image where observers fixated their attention. This data can be further processed to produce continuous fixation maps. The first step is to analyze eye movements using one of the methods based on velocity and distance criteria. Methods to obtain fixation locations are described in [94, 34, 51]. Typically the next step is to produce a continuous fixation map by applying to fixation locations a Gaussian filter with a certain standard deviation corresponding to the high acuity vision area [82].

3.1.2 Eye-head coordination in humans

In virtual reality environment, in addition to eye movements, the observer's head direction must be taken into consideration. One can find studies on eye-head coordination in humans during different tasks in [45, 32]. The main findings in these studies support a hypothesis that the human eye movement range is restricted not physiologically but neurologically and this range is narrower when the subject's head is not fixed. Nonetheless, there is no commonly

adopted model for interpretation of eye-head position data in visual attention fixation maps for omnidirectional visual content.

3.1.3 State of the art in saliency for virtual reality

Current research on visual attention in omnidirectional images and virtual reality is mainly represented by two trends: one concerns obtaining visual attention information from experimental data involving human viewers, whilst another concentrates on prediction of salient regions using algorithmic approaches. The problem of obtaining visual attention empirically is investigated by researchers in [109, 97, 33, 85, 88, 4]. These works provide analysis of eye and head movements during consumption of VR content and propose several methods to process raw experimental data in order to obtain saliency maps. Prediction of salient regions using the algorithmic approach is studied in [11, 12]. Bogdanova et al. propose bottom-up methods to obtain saliency maps from omnidirectional images for static and dynamic cases. Features are computed and fused in a spherical domain. In [99, 73] authors advocate mostly adaptation of earlier conventional saliency prediction methods described in [13, 20]. Deep learning approaches to predict visual saliency in omnidirectional visual content are presented in [131, 83, 15].

3.2 Visual attention in VR with head-direction data

Typically, if one needs to obtain experimental data on visual attention, different types of eye-tracking devices are exploited. This equipment is complex and requires many steps of prior calibration. Thus, eye tracking in VR can only be performed in a controlled laboratory environment. For the moment, there are no consumer devices available which provide accurate tracking of eye movements. Nonetheless, the head tracking, being an intrinsic part of any HMD and any hand-held device, is always available. Therefore, there is a need for a method to estimate visual attention statistically using only head direction data.

We propose a simple approach to treat raw experimental head direction trajectories in virtual reality content. The proposed approach implies three basic steps: First, a computed head angular speed is used to exclude the parts of a trajectory where motion is too fast to fixate viewer's attention. Second, fixations of different subjects are fused together. If needed, this step is preceded by re-sampling track coordinates in order to conform to the equal distribution of points on a sphere. Finally, a Gaussian based filtering is performed to produce continuous fixation maps.

3.3 Fixation locations

In this section, we describe a method to obtain viewer's fixation locations from head direction trajectories recorded when observing omnidirectional images using an HMD.

Fixation locations, contrary to continuous fixation maps, consist in a set of discrete points, each corresponding to a center of an area on an image where viewers fixated their attention.

Given head direction tracks, which are essentially sequences of coordinates with associated timestamps, we start the analysis by computing a rotational speed at each point. An angular velocity of observer's head evidently impacts his ability to fixate attention. Although the fact that human visual perception depends on motion is well known, the impact of the ocular-vestibular reflex can decrease its effect. Nonetheless, we make the assumption that there exits a threshold head angular velocity beyond which users are not able to focus their attention on any object. A value of 15-20 degrees per second has been chosen as the upper boundary based on observations during subjective experiments. However, we would like to point out that this is a parameter and additional experimental data is needed to determine the optimal threshold angular velocity.

3.3.1 Head angular velocity

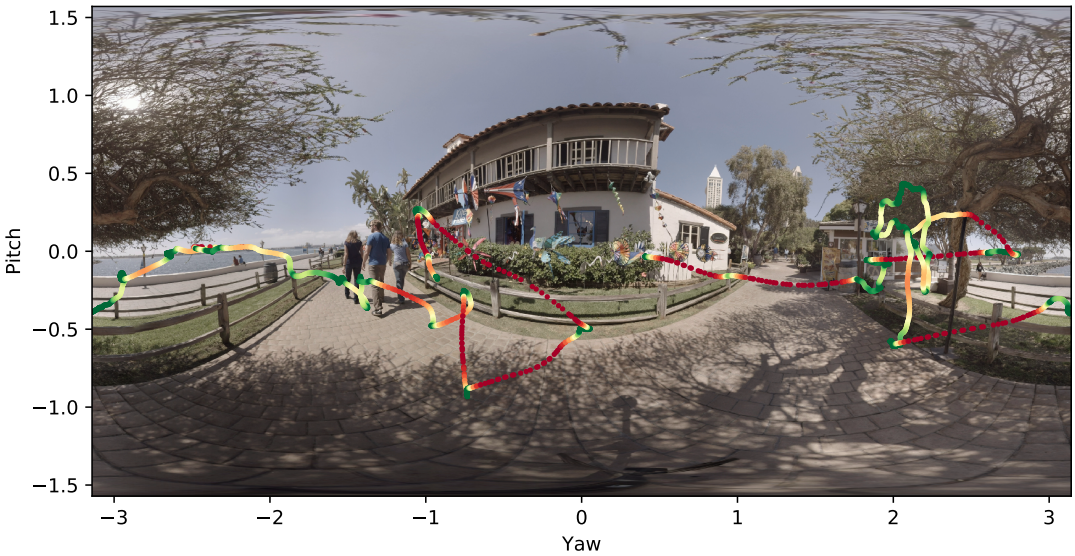
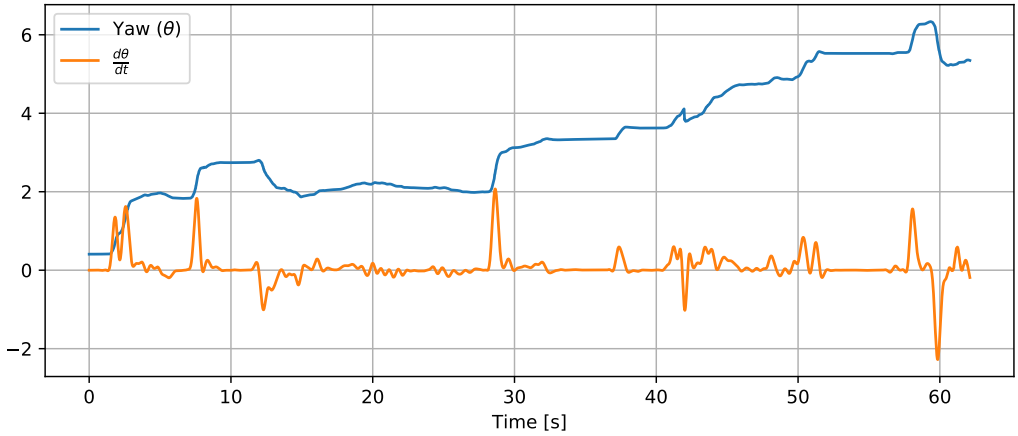


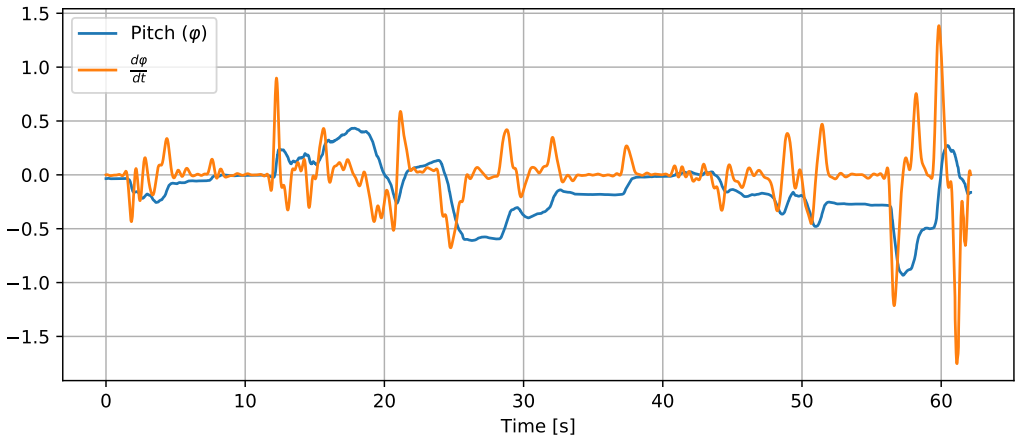
Figure 3.1 – Example of a head direction trajectory. The color depicts the speed of rotational movement, where green is lower and red is higher than the threshold.

We define observer head rotational position as a vector $[\theta\ \varphi]$, where θ and φ represent *yaw* and *pitch* respectively. Values of yaw and pitch in degrees over time are presented in Figure 3.2ab. In order to obtain the head angular velocity, we compute a first order derivative of the vector as following:

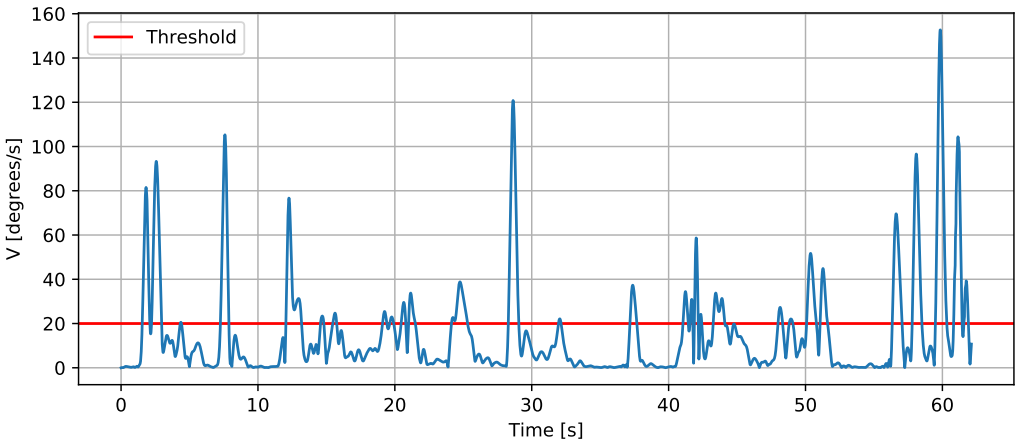
$$\mathbf{V}_{ang} = \begin{bmatrix} V_\theta \\ V_\varphi \end{bmatrix} = \begin{bmatrix} \frac{d\theta}{dt} \\ \frac{d\varphi}{dt} \end{bmatrix} \tag{3.1}$$



(a) Yaw value in time (blue) and its time derivative (orange)



(b) Pitch value in time (blue) and its time derivative (orange)



(c) Speed head angular movement in time (blue) and a visual fixation threshold (red)

Figure 3.2 – Head angular position and speed of a single head direction trajectory.

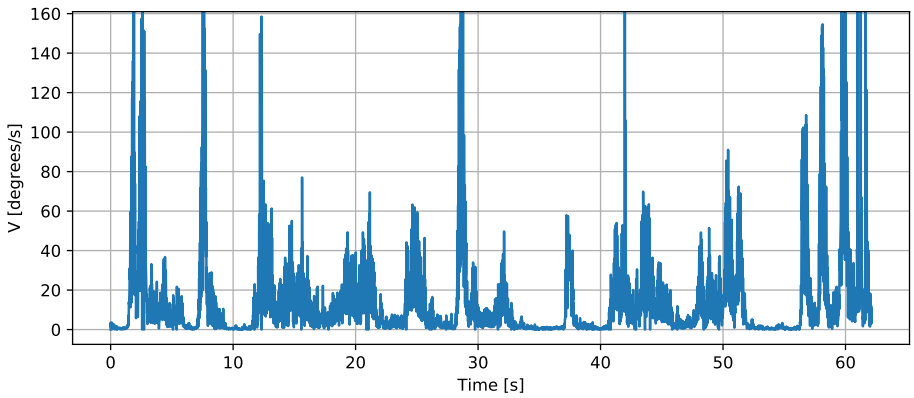
Since we only consider the velocity magnitude, or in other words speed, depicted in Figure 3.2c, we take the norm of the vector as:

$$\|\mathbf{V}_{ang}\| = \sqrt{\left(\frac{d\theta}{dt}\right)^2 + \left(\frac{d\varphi}{dt}\right)^2} \tag{3.2}$$

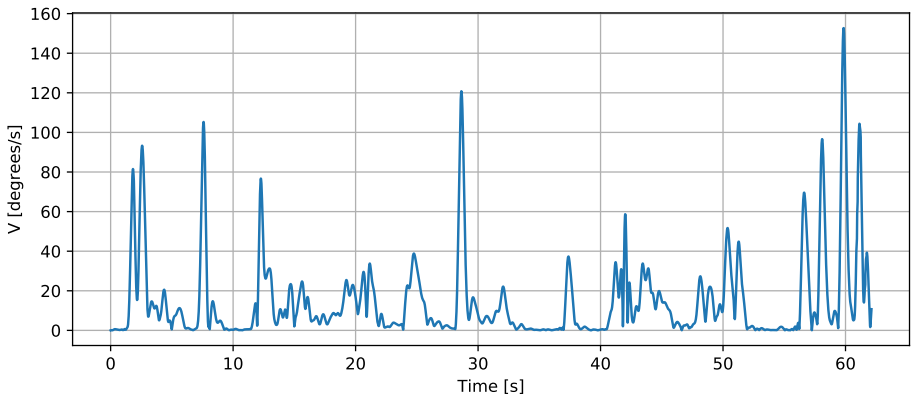
The yaw and pitch data is represented in digital format. Thus we compute a derivative using a standard method of numerical differentiation. For each signal sample the difference with its next value is obtained and divided by the sampling period:

$$s'_n = \frac{s_n - s_{n-1}}{T_{sampl}} \tag{3.3}$$

Then a 2nd order Butterworth low-pass filter with cutoff frequency of f_c = 2 Hz is applied



(a) No low-pass filter applied.



(b) Butterworth low-pass filter with $f_c = 2Hz$ applied.

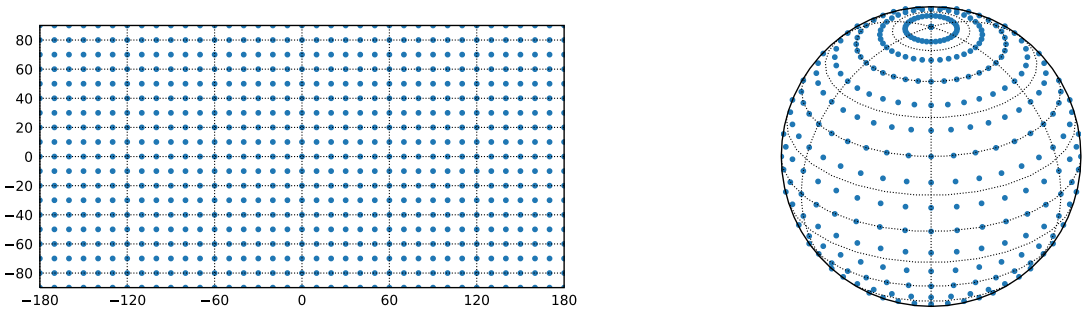
Figure 3.3 – Applying low pass filter when computing angular speed.

separately to V_θ and V_φ in order to remove digital differentiation noise. We use a forward-backward numerical implementation of the filter to avoid a group delay in the signal [46]. Figure 3.3 shows head angular speed of a typical track before and after applying a low-pass filter. The resulting head angular velocity over time is depicted in Figure 3.2c. All the head direction trajectory data with speed above the threshold (red line in Figure 3.2c) is discarded from further analysis.

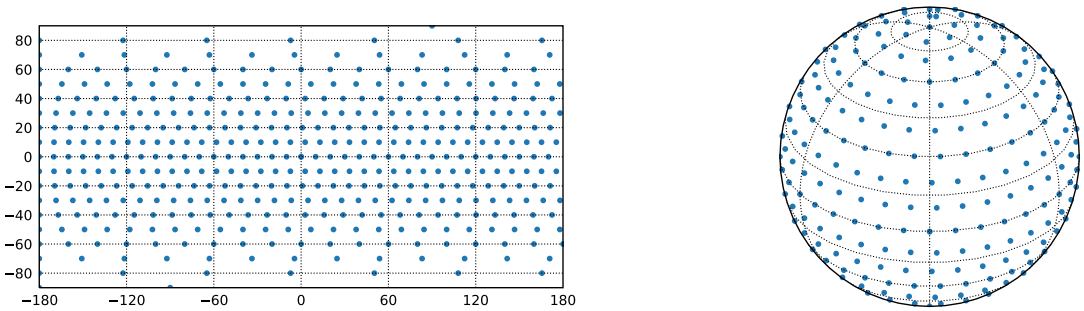
Figure 3.1 shows a typical head direction trajectory. The color of the trajectory reflects the head angular velocity. Only the regions colored green are considered as fixations of attention.

3.3.2 Equal distribution of points on sphere

We assumed in Section 3.3.1 that the points in head-direction-track series are represented in a uniform discrete coordinate system on the surface of a sphere. There exist, however, cases when, under the head angular velocity restrictions, a resulting track requires an additional step of processing before becoming a set of viewers' fixations. Depending on the device used to obtain the raw data, the discrete domain of coordinates can distribute points in a non-



(a) Points equally distributed in the equirectangular domain becoming more dense closed to the poles.



(b) The density on points is the same on the entire surface of a sphere.

Figure 3.4 – Distribution of points on sphere

equidistant manner on the surface of a sphere. If so, a re-sampling needs to be performed on the data in order to comply with equal uniform spatial distribution.

Thanks to the discrete nature of the signal, we can treat each line of the equirectangular coordinate grid independently. Being a function of latitude, the sampling rate for each line must be adjusted considering the length of the latitudinal circumference. For each latitude level one re-samples the longitude signal $s(n)$ defined on $n \in N$ to the signal $g(m)$ defined on $m \in M$, where $M = N\cos(\phi)$ and $\phi \in (-\pi/2, \pi/2)$.

Figure 3.4 (a) shows how points which are equally distributed in the equirectangular domain (left), when wrapped around a sphere (right), do not preserve constant density. In the near pole area the are more points per unit area than at the equator. Figure 3.4 (b) provides an example of the distribution of points on the surface of the sphere (right) after applying re-sampling procedure, and the correspondent equirectangular projection view (left).

3.3.3 Fusion of fixation points

Thus far, in Section 3.3, we were talking about extracting fixation points from an individual track obtained from a single viewer. However, this is not the final result one would need to describe human visual attention. The same stimuli must be shown to multiple observers under the same conditions and an identical task. Afterwards, fixation locations obtained from different subjects must be fused in order to derive statistical information. Several methods to perform the fusion can be considered.

Simple conjunction adds all the fixation points from each subject as *unity values* to a resulting set of overall fixations. One should keep in mind that simple re-sampling might be needed to adapt the coordinates to the effective resolution of the equirectangular image. The resulting set of fixation locations is binary, meaning that any point can be either one (a fixation point) or zero (not a fixation point).

Simple histogram Sums-up all the points in a predefined number of cells with specified size producing a weighted set. The cells are organized as a two-dimensional array with indexes corresponding to the yaw and pitch coordinates of an equirectangular image. Because of the wide range of values, the resulting data might need to be represented in logarithmic scale.

Thresholded histogram method only adds points if a certain number or a percent of subjects fixated at a location falling into a cell is reached. The thresholding is binary thus the resulting matrix only contains ones and zeroes corresponding to fixating or not fixating, respectively, in the area covered by a cell.

We use the second method (simple histogram) to produce fixation locations further in the present work because of its moderate computational complexity and its satisfactory accuracy.

3.4 Continuous fixation map

In this section, we provide a description of the calculation of continuous fixation maps from fixation locations in an equirectangular image.

Fixation location data does not typically allow to properly depict the regions of visual attention. Because of its discrete nature, this information is not consistent even among human subjects. Indeed, very rarely a person will fixate their attention at the same exact point as another. Thus, there is a need to introduce statistical areas of fixations. For conventional images, typically, a Gaussian filter is applied to model a human acuity vision region of 1-2 degrees. In case of head direction fixations we assume, considering our observations and research in human physiology [45, 32], that the region of possible attention is 30 degrees. As in Section 3.3.1, this value is a parameter and may be changed after further experimentation.

3.4.1 Gaussian filter in viewport domain

Omnidirectional content is consumed using an HMD. An observer sees a part of a panoramic picture rendered in the viewport. Therefore, to highlight the viewing area angle we need to apply a Gaussian filter in the viewport domain.

The kernel for the filter is defined as follows:

$$G(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}} \tag{3.4}$$

where u and v are viewport coordinates. However, one normally works with an equirectangular or other panoramic representations of omnidirectional image or video. Thus, in the equirectangular domain, the kernel becomes:

$$G_{eqr}(\theta,\varphi) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2(\theta,\varphi)+v^2(\theta,\varphi)}{2\sigma^2}} \tag{3.5}$$

Functions $u(\theta,\varphi)$ and $v(\theta,\varphi)$ are calculated using the following expression:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & \frac{k}{x} & 0 \\ 0 & 0 & \frac{m}{x} \end{bmatrix} R_Z^{\beta} R_Y^{\alpha} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \Bigg|_{x>0} \tag{3.6}$$

where k and m are the scaling coefficients for viewport coordinates, R_Z^{β} and R_Y^{α} are rotations

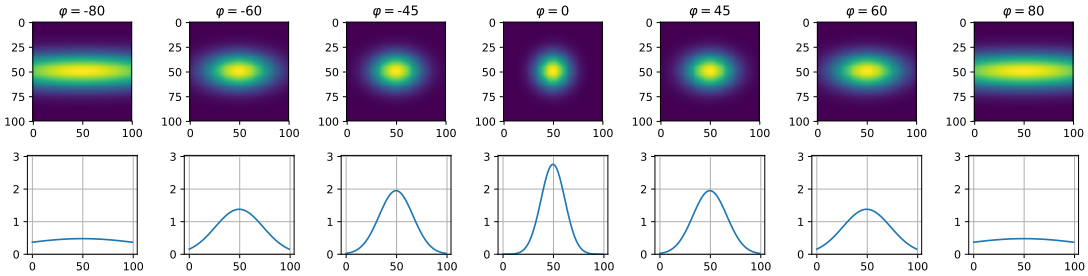


Figure 3.5 – Modified Gaussian kernel (Eq. 3.8) at different latitudes φ.

for yaw and pitch respectively, and vector $[x\ y\ z]$ represents the Cartesian coordinates of a point on the image sphere, derived from yaw and pitch as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin\varphi\cos\theta \\ \sin\varphi\sin\theta \\ \cos\varphi \end{bmatrix} \tag{3.7}$$

assuming that the radius of the sphere equals 1.

The result of applying kernel $G_{eqr}(\theta,\varphi)$ to filter the image directly in the equirectangular format is shown in Figure 3.6 (b).

Another approach to perform Gaussian smoothing in an equirectangular picture is to apply the filter in the rendered viewport and then project it back. However, the drawback of this method is the interpolation noise added during the transformations.

3.4.2 Modified Gaussian kernel in equirectangular domain

The filtering method proposed in Section 3.4.1 is computationally very heavy. To simplify the calculations we propose a modified Gaussian kernel.

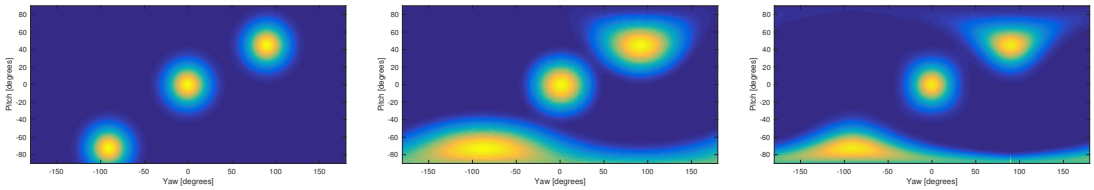
$$G_{mod}(x,y) = \frac{1}{2\pi\sigma_y^2} e^{-\frac{x^2}{2\sigma_x}} e^{-\frac{y^2}{2\sigma_y}} \tag{3.8}$$

where

$$\sigma_x = \frac{\sigma_y}{\cos(\varphi)} \tag{3.9}$$

and σ_y is a constant value. In the denominator normalization coefficient, one can use σ_y^2 instead of $\sigma_x\sigma_y$ to prevent the change of the amplitude with x.

Figure 3.6 (c) shows an equirectangular image filtered using kernel $G_{mod}(x, y)$. As can be seen by comparing Figures 3.6 (b) and (c), the Gaussian filter in the viewport domain and the Modified Gaussian kernel give visually similar results.



(a) (b) (c)

Figure 3.6 – Different Gaussian based filters on equirectangular image: (a) Gaussian filter applied in equirectangular domain. (b) Gaussian in viewport domain. (c) Proposed Modified Gaussian. Filters are applied to equirectangular image containing three unity points at (-90,-72), (0,0), and (90,45) degrees.

3.4.3 Generic statistical kernel in equirectangular domain

Faced with a lack of statistical data on eye-head relative movements, we assumed a Gaussian distribution of eye fixations around the center of the viewport. However, if such statistics is available, it can be used to form a kernel in the viewport domain:

$$K \equiv f(u, v) \tag{3.10}$$

where $f(u, v)$ is a probability density function on $(u, v) \in \mathbb{R}^2$, which can be estimated from statistical frequency distribution of eye fixations in the viewport by applying a regression to its two-dimensional histogram $m_{i,j}$ with k^2 being the number of bins:

$$f(u, v)\Big|_{\substack{u=(i-k/2)w \\ v=(j-k/2)w}} \cong \frac{m_{i,j}}{\sum\limits_{i,j\in\mathbb{N}} m_{i,j}} \tag{3.11}$$

where $i, j \in \mathbb{N}$ are the indexes of each histogram bin, and $w \in \mathbb{R}^+$ is the bin width. The

histogram is calculated as:

$$m_{i,j} = \sum_{\substack{(i-1-k/2)w<u_p\leq(i-k/2)w \\ (j-1-k/2)w<v_p\leq(j-k/2)w}} X[u_p, v_p] \Bigg|_{\substack{i\in[1,k] \\ j\in[1,k]}} \tag{3.12}$$

where $X[u_p, v_p]$ is the relative frequency distribution of fixation locations $(u_p, v_p) \in \mathbb{R}^2$, which are determined as a shift from the viewport center for $p \in [1, M]$, $p \in \mathbb{Z}$ and M is finite. The number of bins must be chosen according to one of the criteria described in [101, 31] depending on the distribution law.

Moving to the equirectangular domain can be performed as in Section 3.4.1:

$$K_{eqr}(\theta, \varphi) = K(u(\theta, \phi), v(\theta, \varphi)) \tag{3.13}$$

A filter with the kernel $K_{eqr}(\theta, \varphi)$ can be applied to fixation locations directly in the equirectangular domain.

3.5 Experimentally obtained visual attention data

The data used in the current section of this work has been obtained during a subjective quality evaluation experiment [112] on omnidirectional images.

3.5.1 Head direction tracks

Raw data of a head direction trajectory contains an array of yaw and $pitch$ coordinates along with their time-stamps. The tracks were recorded for each assessed stimulus. However, only the trajectories obtained from images rated "Fair" and higher have been selected for the current analysis. Each presented content has head-direction tracks from 40 subjects. Two seconds of data in the beginning of each track were dropped, in order to compensate the initial head position impact on calculating user gaze fixations.

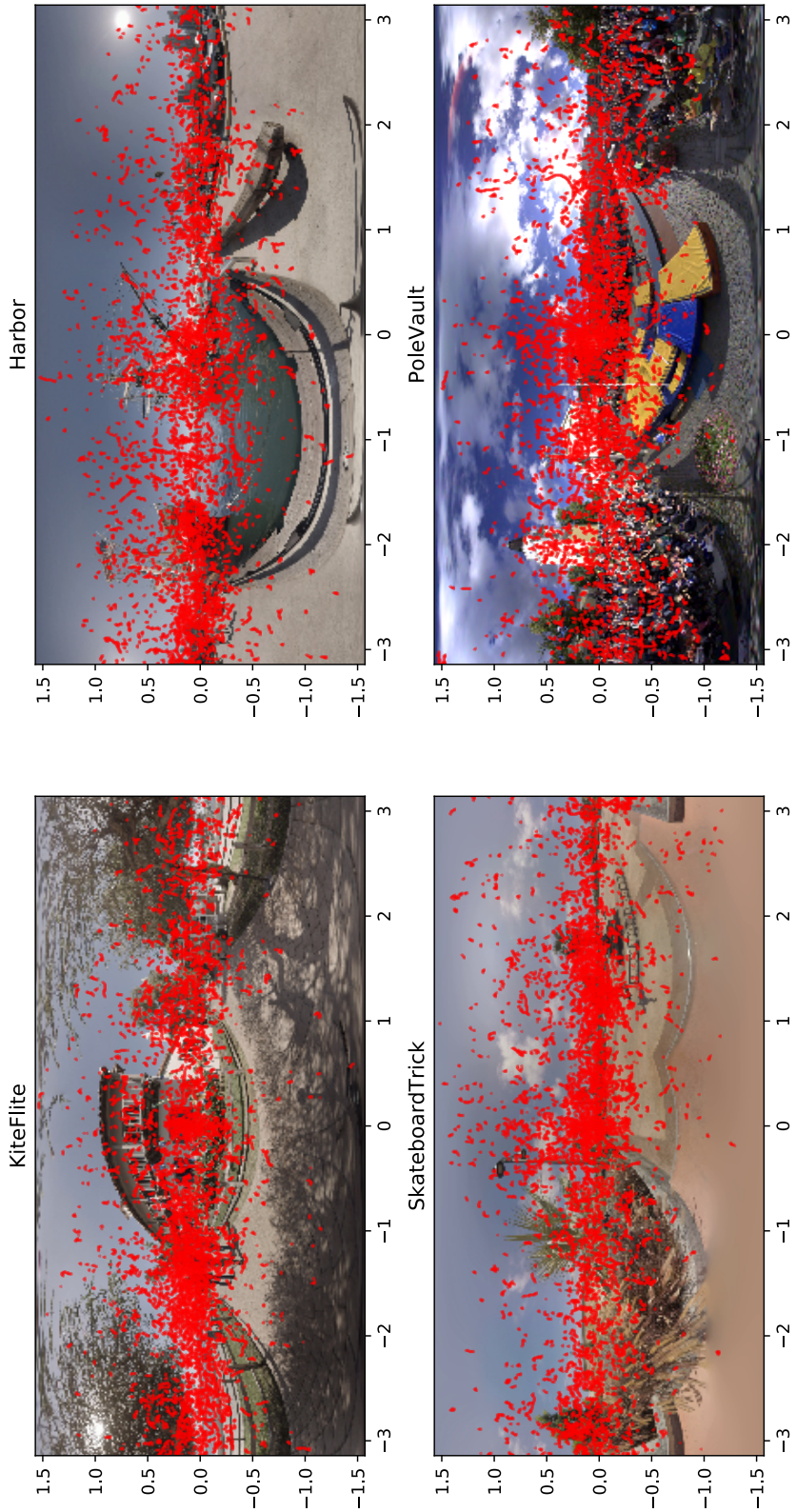


Figure 3.7 – Fixation locations obtained experimentally.

3.5.2 Experiment

Observers were asked to assess visual quality of four different omnidirectional images represented in the equirectangular projection and compressed with different quality parameters and different codecs. In particular, viewers were instructed to search for compression artifacts. Overall, 40 subjects participated in the experiment, 25 male and 15 female subjects, between 18 and 32 years old with the average and the median of 24.9 and 24.8, respectively. All participants were tested for correct color vision and visual acuity using Ishihara and Snellen charts respectively.

The experiment was conducted using the testbed for subjective evaluation of omnidirectional visual content proposed in [111]. This software has been developed for iOS and is publicly available for download[1]. During the experiment, subjects were wearing an HMD composed of a VR head-mount with buttons[2] and a mobile device installed inside as a screen. iPhone 6 was used to display the images. The overall resolution of the phone's screen was 1334×750 pixels, which gives 667×750 pixels per eye. The vertical field of view provided by the hardware-software solution was 90 degrees, which corresponds to 8.33 pixels per degree. All the subjects were sitting on a spinning chair during the experiment.

3.5.3 Results and discussion

We apply the proposed approach to compute fixation locations and continuous fixation maps as interpretation of the raw experimental data described in [112]. A head angular velocity threshold equal to 15 degrees per second was used. The Gaussian filtering was performed using $\sigma = 15$ in the base function. In order to fuse individual fixation locations, the points ware summed up in cells of 1×1 degree. The modified Gaussian kernel $G_{mod}(x, y)$ was used to filter the data in the equirectangular domain. Figure 3.7 shows the fixation locations for four contents used in the experiment. The generated continuous fixation maps are presented in Figure 3.8.

In the present work, we applied Gaussian filtering in the equirectangular projection. However, the proposed approach can be easily generalized to cope with other panoramic representations of omnidirectional visual content, such as cube mapping and other convex polyhedron projections. Only the calculation of $u = u(x, y)$ and $v = v(x, y)$ must be changed to comply with a new projection.

In more theoretically oriented work [11, 12], the authors developed a mathematical model for Gaussian filtering in the geometry of the two-dimensional surface of a sphere. We consider these to be unnecessary complications, due to the fact that an observer sees only a rendered rectilinear viewport of an omnidirectional content and not the entire image. Thus, applying

[1] https://github.com/mmspg/testbed360
[2] https://mergevr.com

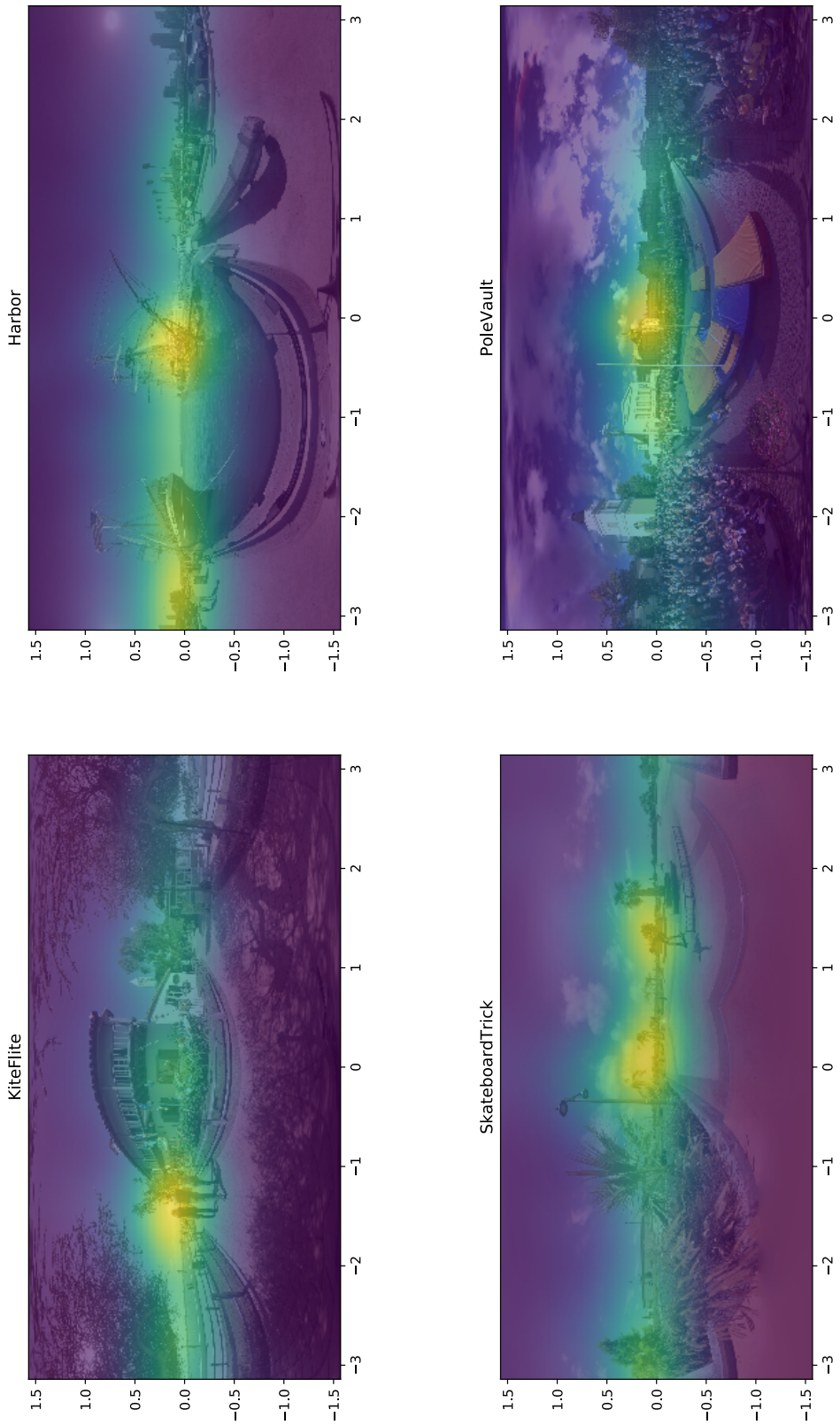


Figure 3.8 – Continuous fixation maps obtained experimentally.

vision field range models in the viewport geometry appears to reflect better user experience and perception.

Head motion information is typically available without any additional cost during rendering of omnidirectional visual content in VR environments. For instance, during broadcasting, a content provider can obtain anonymized head direction trajectory statistics of consumers. This information can be further used to adapt compression parameters when adaptive coding is applied. An example of such an adaptive coding method has been proposed in [52] for conventional images.

4 Subjective Evaluation of Perceptual Visual Quality

In this chapter, we, firstly, describe methodologies for subjective assessment of omnidirectional visual content. Then, we present a testbed designed as a solution for performing such evaluation experiments, including the descriptions of its architecture and implementations. We follow, afterwards, by presenting results of subjective evaluations with mean opinion scores and time statistics, which are used further in Chapter 5 for benchmarking and validating of objective metrics.

4.1 Subjective evaluation of omnidirectional content

Evaluation of perceptual visual quality is crucial for many applications in image processing. It plays an important role in encoding, compression and transmission of images and video. Subjective evaluation of visual quality is typically applied for establishing the ground truth for objective quality prediction metrics, as well as for assessing performance of newly proposed methods for compression and coding. A generic subjective evaluation experiment consists in showing visual content to human observers and asking them to rate visual quality. The ratings are used afterwards for computing mean opinion scores (MOS).

Omnidirectional visual content has additional features, such as interactivity and immersiveness, comparing to conventional images and video. Thus, even though conditions, requirements, and procedures for subjective evaluation of perceptual visual quality have been already developed and validated for conventional visual content [1, 2], they need to be extended in order to cope with omnidirectional images and video. Very few studies have been published on this subject: In [68], authors present new strategies for assessing the quality of composite video streams focusing on videoconferencing applications only.

4.1.1 Approaches

Omnidirectional visual content can be consumed by means of different technologies (see Section 2.4 in Chapter 2), which provide different level of immersiveness. Thus, an experimenter needs, fist of all, to choose a type of display. We argue in favor of head-mounted virtual reality, because it isolates subjects from their surroundings and simplifies the setup of an experiment by eliminating the impact of room lighting conditions and other distractions.

Interactivity of omnidirectional imaging can also be approached differently in experiments on subjective visual quality evaluation:

Free exploration is an approach where subjects are not restricted or only restricted by time in their interactions with an omnidirectional image or video. The task is defined and explained prior to the viewing sessions. Participants need to understand that they are evaluating visual quality; and the expected artifacts must be pointed out. The flaw of this method is that every subject sees different parts of the entire omnidirectional image and the complete coverage by a single observer cannot be guaranteed. This can be, however, compensated by increasing the number of subjects.

Prerecorded interactions is another way to cope with interactive content. Following this approach, subjects are not given the ability to interact freely, but are exposed to a prerecorded scenario, as if they were watching a video sequence. The main flaw of this approach is its non-compliance with the natural way of consuming immersive content, and the lack of evidence that prerecorded interactions are perceptually equivalent to free exploration. Even in those cases, where prerecorded interactions conform to a statistical average path, the latter can only be obtained from prior subjective evaluations with free exploration conditions.

Further in our work, we only perform free exploration experiments on subjective evaluations of omnidirectional visual content, due to advantages of this approach compared to the prerecorded interactions.

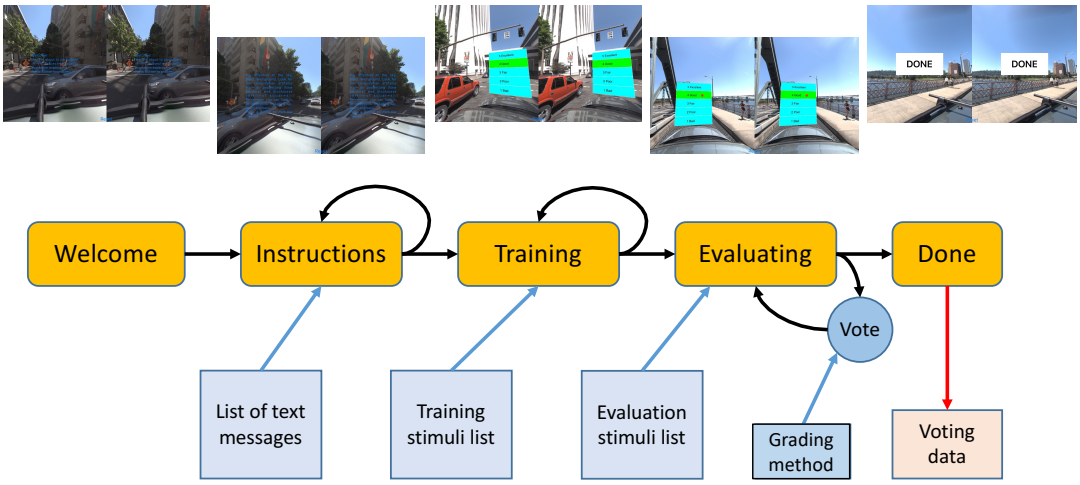


Figure 4.1 – The workflow of a subjective evaluation implemented in the proposed testbed

4.1.2 Workflow and conditions

Let us discuss the general workflow of an experiment on subjective evaluation of omnidirectional visual content. This concerns the following aspects: equipment and software, introduction, prior instructions, and explanations for subjects, the sessions in virtual reality environment, presentation of stimuli, and collection of data.

In order to fully exploit immersive capabilities of omnidirectional visual content, it is favorable to use a head-mounted display for subjective quality evaluations. Although, hand-held devices and personal computers can render omnidirectional images and video, they both lack the intrinsic ability to isolate a subject from the environment, which would complicate the setup. Moreover, disconnection from the environment reduces the impact of external factors on the results of subjective evaluation. It is important also to point out that one can only benefit from a HMD setup if training and voting is performed within virtual reality.

For safety reasons subjects must be in a seated position during the entire experiment. Since omnidirectional visual content requires only three rotational degrees of freedom with possible range-limited head movements this does not affect the ability to perform any possible interactions. The chair, however, must be a revolving one. All subjects should pass visual acuity and color vision tests. Prior to the evaluation and training in virtual reality environment, an experimenter should orally explain to subjects how to use the equipment and the software, and inform them about the task and the types of expected visual artifacts which could be found in impaired omnidirectional images.

Figure 4.1 shows the workflow of the VR part of a generic subjective visual quality evaluation experiment. It begins with a welcome message which is shown to a subject first. At this step the HMD position can be adjusted. The welcome screen is followed by instructions

rendered as text inside the immersive virtual reality environment. After finishing reading the instructions subjects proceed to a training session during which they are exposed to three or more omnidirectional images with degradation to different extent. These images have pre-selected ratings in order to familiarize subjects with expected levels of visual quality. Next, follows an evaluation session, the payload of the experiment, where subjects give their opinion about perceptual visual quality of tested omnidirectional images or video according to a selected type of comparison and a scale. The voting is performed inside the virtual reality. When the last stimulus is graded, the score data and the interaction data are stored. After the data are collected from all subjects, it is screened offline in order to exclude possible outliers.

4.1.3 Testbed for subjective evaluation

In this subsection, we describe a testbed for subjective evaluation of omnidirectional visual content. The main purpose of this testbed is to provide researchers with a tool to perform subjective assessment of omnidirectional images and video. The testbed consists in a software application for the hand-held and mobile platforms and can be used with head-mounted displays. The application is able to visualize omnidirectional images and video represented in different projections. The set of supported projections or geometrical representations can be easily extended. Moreover, the testbed provides a special customizable graphical user interface and a storyboard for subjective quality evaluation experiments.

The application allows viewing images and video using head-mounted displays, available from many manufacturers. These mounts do not contain the display itself and a mobile device with a screen is inserted inside in order to use them. Additionally, images and video can be displayed by the application without HMD using only hand-held devices, i.e. tablet computers and mobile phones. The developed testbed supports visualization for different types of geometrical representations or projections of omnidirectional visual content. The current set includes equirectangular, cubic, half-back cubic, and cubic with rotated facets projections. However, thanks to the fact that the interfaces are open and scalable, other projections can be easily implemented and added in the future by authors or by third party contributors from the open source community. Running with an HMD mount or on a single hand-held device, the application allows one to freely change the direction of sight by means of the motion control. This feature is implemented by obtaining the attitude data from built-in accelerometer and magnetometer sensors in the mobile device or tablet computer. In the hand-held mode it is also possible to configure the application to use finger gestures in order to control the direction of view.

In Figure 4.2 one can find a block-diagram of the testbed application architecture. The software is logically divided into three modules: namely, Control, View, and Model. The application takes image files and auxiliary data, such as user instructions and evaluation methodology type, as an input. After the experiment is finished, the acquired data is stored as an output. The Model module keeps the current state of the application, according to which it takes signals

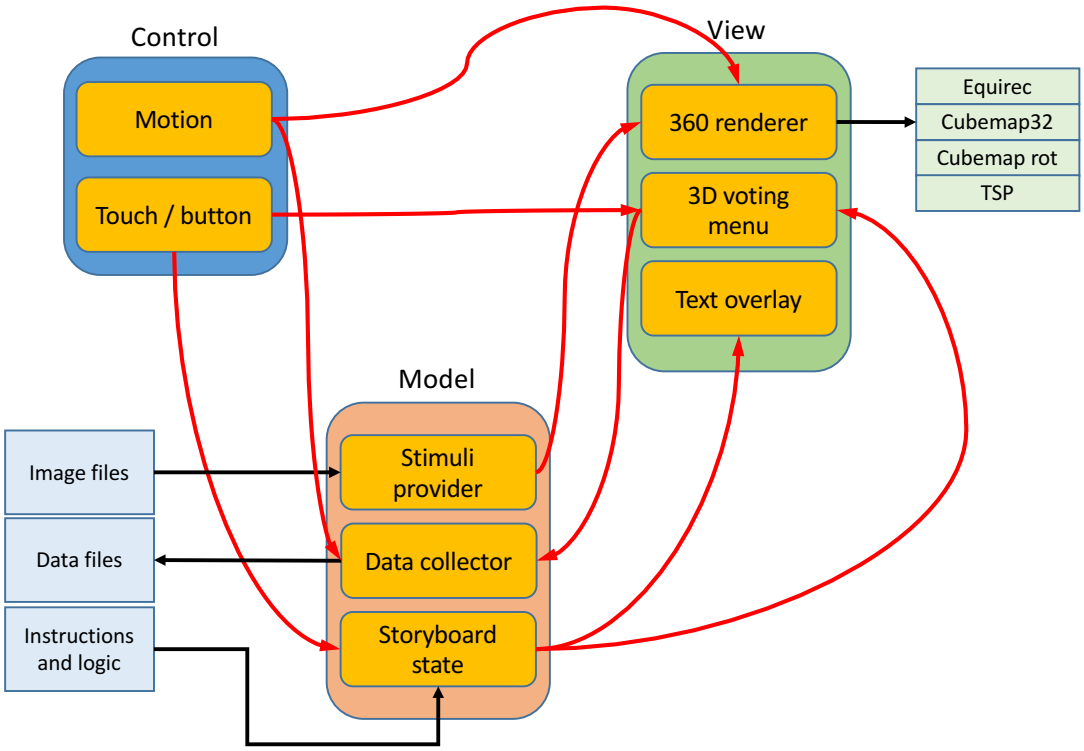


Figure 4.2 – Testbed application architechture.

representing user interactions from the Control module, and provides data and instructions to the View module. The Control module is responsible for capturing signals from motion sensors and input controls of a device. The View module performs the rendering of the omnidirectional visual content according the type of projection in which it is represented. Additionally, the View module controls an immersive voting menu and the overlay text.

The testbed application provides a special storyboard of screens to perform subjective quality evaluation of omnidirectional visual content. The storyboard contains customizable textual instructions displayed to subjects. This text can be easily updated and should contain information such as instructions to the subject on how to proceed during different parts of evaluation. For example, during the training, it can include information on how to score: what part of the content the subjects must pay attention to, what type of distortion and artifacts they should consider and what they should exclude from consideration.

After a welcome message and basic instructions, participants proceed to the training session. During the training session no special action is required from subjects as the training stimuli are shown with the corresponding votes already provided. Subsequently, during the evaluation session, subjects see the test stimuli and, when ready to give a quality score, they can activate a scoring menu. The scoring menu is displayed as an overlay on the screen and contains voting

items according to the selected evaluation methodology. Once the voting score is selected and recorded, the next stimulus is displayed immediately. The scoring menu as well as the entire storyboard can be customized for different subjective evaluation methodologies, such as single stimulus or double stimulus impairment scale. The data gathered from subjects are anonymized in the application by automatically assigning a unique identification number to each subject at the beginning of each session.

Prior to the experiment, a stimulus set must be prepared by an experimenter according to the special naming convention and uploaded to the device in order to perform tests. Additionally, the experimenter explicitly marks images for the training session and dummy stimuli. The developed testbed application automatically performs randomization of the test stimuli according to the ITU recommendations described in Section 2.7 of [2], in such a way that the same content is never shown consecutively.

For each evaluation session the proposed testbed application stores: subjective scores, tracks of direction of view for each stimulus, and time spent by each subject on each stimulus. The direction of subjects' sight during the evaluation of each stimulus is represented by two coordinates: *yaw* and *pitch*, where the former represents horizontal angle and the latter represents vertical angle. Attitude coordinates correspond to the center of a current viewport. A separate track is recorded for each test stimulus with a configurable sampling rate, e.g. 10 to 60 samples per second. Additionally, during the evaluation session, time spent by a subject on each stimulus is measured. It is also possible to restrict the subjects to spend limited time on each stimulus. All experimental data, including scores and viewing direction tracks are stored locally on the device in comma-separated-values (CSV) format. It can be transferred from the device to a server or a work station for further analysis.

The testbed software is implemented for iOS and Android platforms and can be found on Github by the following URLs, respectively: `https://github.com/mmspg/testbed360`, `https://github.com/mmspg/testbed360-android`.

Types of data which can be obtained

The types of data which can be obtained during subjective evaluation and consuming of omnidirectional visual content includes opinion scores and interaction information. Subjective ratings can be collected along with the track of head rotational and translational movements for each stimulus, which can help during further analysis.

Additionally, there exists an option to collect eye tracking data and obtain fixation information after analysis of this data. However, eye tracking is only possible in laboratory conditions and cannot be used, for example, in crowdsourcing experiments, nor by a typical consumer.

4.2 Impact of projections on perceived quality

4.2.1 Projections reducing spatial resolution

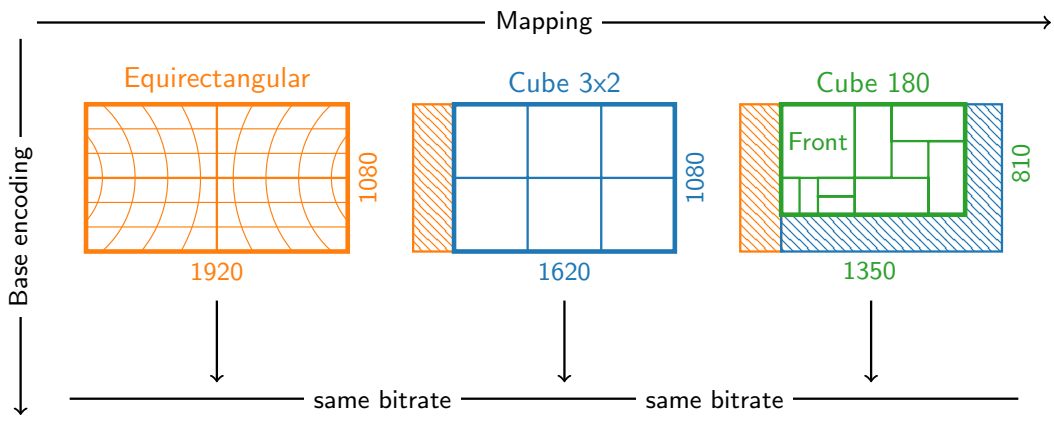


Figure 4.3 – Impact of remapping on the number of required pixels.

Reducing the spatial resolution of certain regions of an omnidirectional image is a practical approach to optimize encoding via representation. One can apply this set of methods to either reduce the bitrate or possibly increase the visual quality within a limited bandwidth transmission channel.

Figure 4.3 depicts an approach for remapping equirectangular projection to a cube based projection preserving effective angular resolution. However, the Cube 180 projection only preserves effective resolution of the front face and the front half of each side face. After transforming an omnidirectional image represented in equirectangular projection to a cube map, we save 16% of pixels required to store it. See the image area hatched in orange in the Cube 3x2 schematic drawing in Figure 4.3. Further remapping to a half-resolution-back cube projection releases 32% more pixels (blue hatched area) from the original equirectangular image making the picture only 52% of the input size. It preserves, however, the effective angular resolution only in the front part of the cube.

In contrast to the approach shown in Figure 4.3, one can keep the same height for the Cube 180 representation. (See Figure 4.4). In this case, the front facet of the Cube 180 has increased effective resolution comparing to the Cube 3x2, while the back facet decreases in quality.

4.2.2 Impact of projections on perceptual quality

A limited-scale subjective evaluation experiment was conducted in order to assess how different projections affect perceptual visual quality of omnidirectional visual content. The purpose of this experiment was only to estimate the impact of remapping and compression in a conventional sense was not considered.

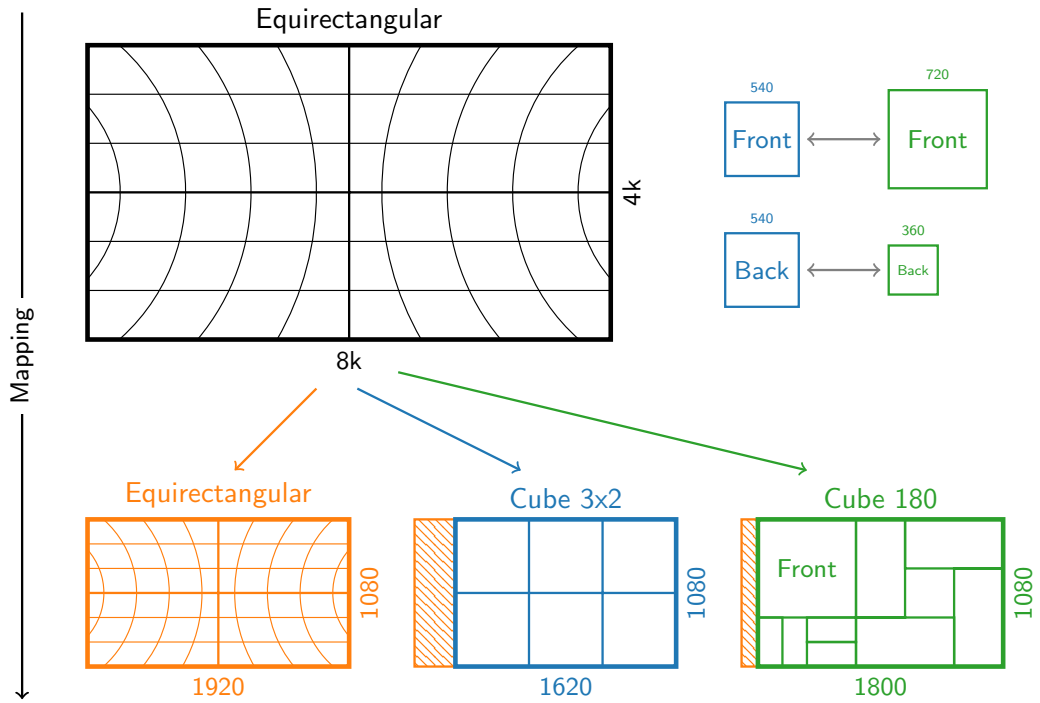


Figure 4.4 – Increasing effective pixel density of certain regions by remapping.

The stimuli set (Figure 4.5 (e-l)) was created from four stitched uncompressed omnidirectional video sequences (Figure 4.5 (a-d)) of 10 seconds duration each. The originals were remapped following the approach depicted in Figure 4.4 and compressed with visually transparent quality with AVC using libx264 library. Three different types of projections were used in the experiment: namely, equirectangular (EQR), cubic (C32), and cube map with half resolution in the back hemi-cube (C180). The compression step was necessary because of the limitations of the equipment unable to playback raw uncompressed video.

A special player for omnidirectional video supporting required projections was developed for the purpose of this experiment. A hand-held device was used to render the stimuli. The size of the device was 250 mm (9.7 inches) diagonally, the resolution was 2048x1536 pixels. The display is a color IPS LCD at (264 ppi) with a 4:3 aspect ratio. The tablet was held by subjects in front of the face at a distance of approximately 30 cm, while sitting in a revolving chair.

The pair comparison methodology was used to assess omnidirectional video. It consists in picking one out of two subsequently shown video stimuli by choosing an answer "*A is better than B*", "*B is better than A*", and "*A and B are equal*" in terms of visual quality. Subjects were asked to select the one which had a better visual quality, or in case no difference could be perceived, they were allowed to mark the pair as having the same visual quality. 18 subjects participated in the experiment. Each of them made 12 comparisons for 4 different test sequences. This resulted in total 216 answers, 72 for each pair of projections being compared.

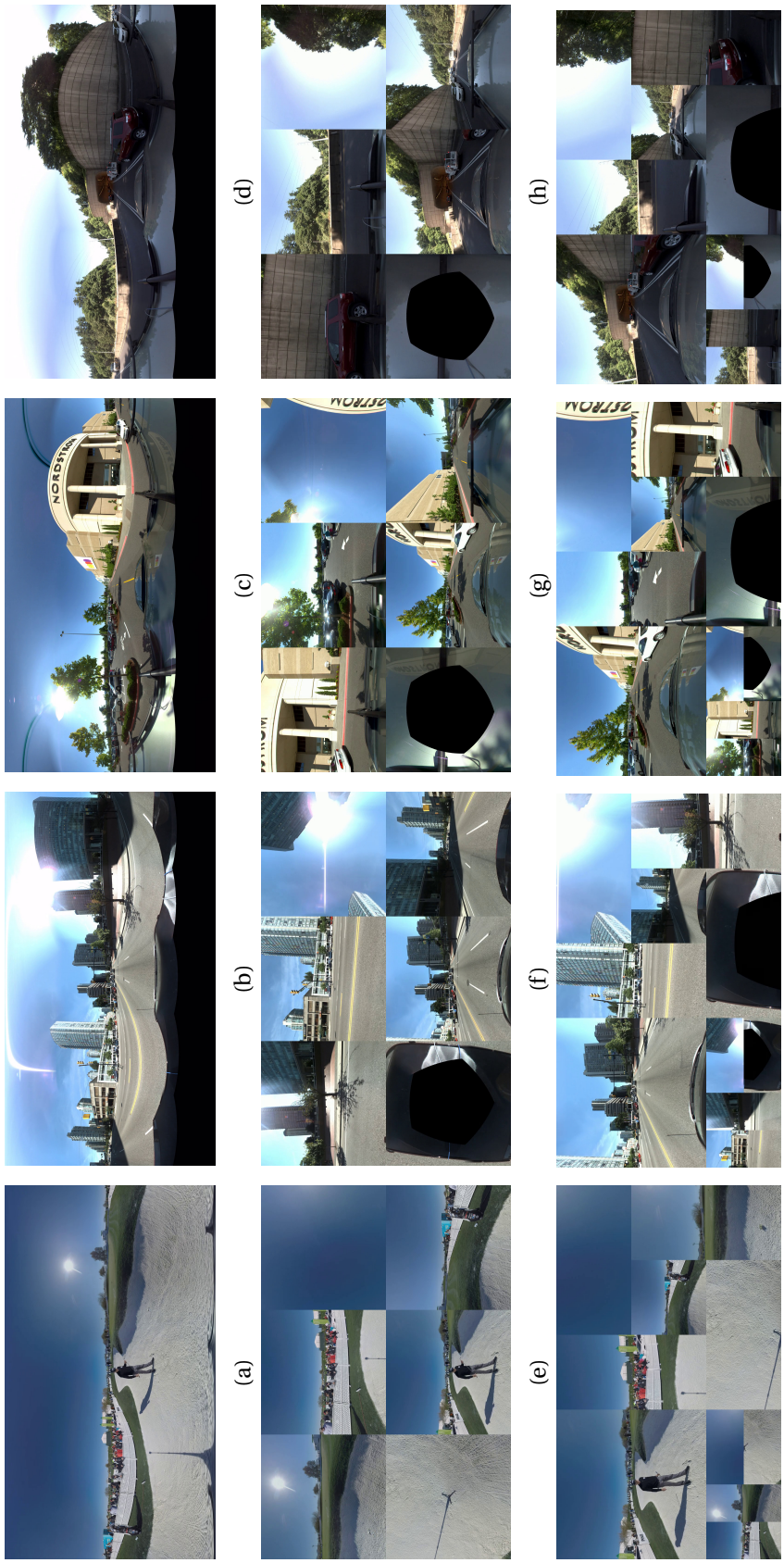
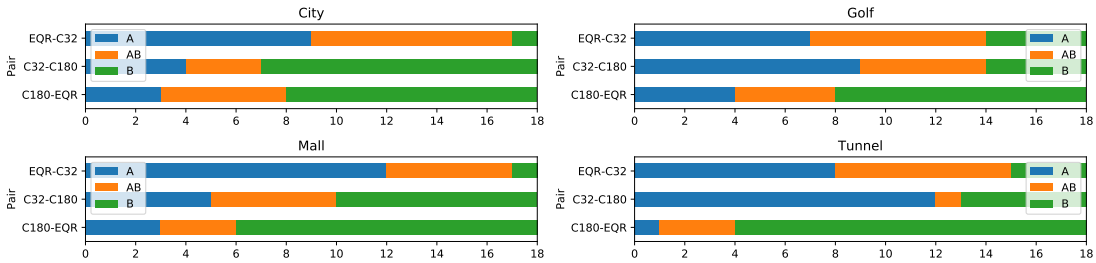
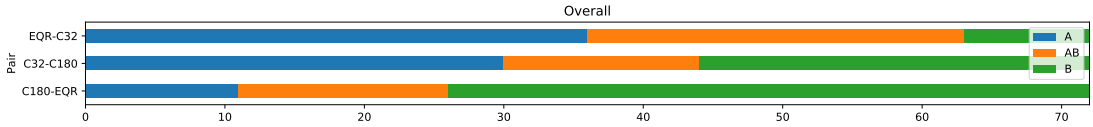


Figure 4.5 – Dataset of omnidirectional video used for remapping evaluation

Figure 4.6 shows charts of vote distribution for each video and overall. Figure 4.7a depicts normalized frequencies of choosing mapping A over B, e.g. coefficient 0.69 at the point (A=EQR, B=C32) means that 69 times out of 100 the subjects preferred EQR over C32, when comparing the pair. This matrix allowed us to estimate mean opinion scores using Bradley-Terry-Luce (BTL) model [16, 76] described in [41]. BTL scores are shown in Figure Figure 4.7b.

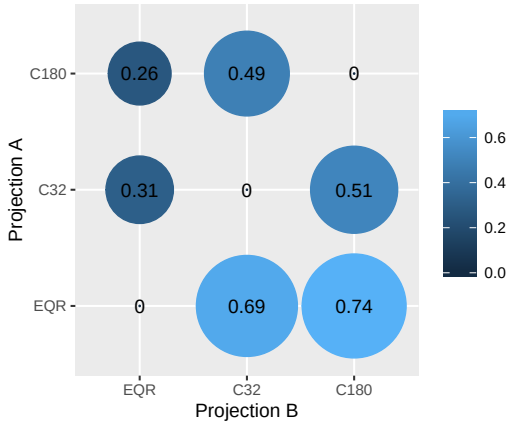


(a) Results per content

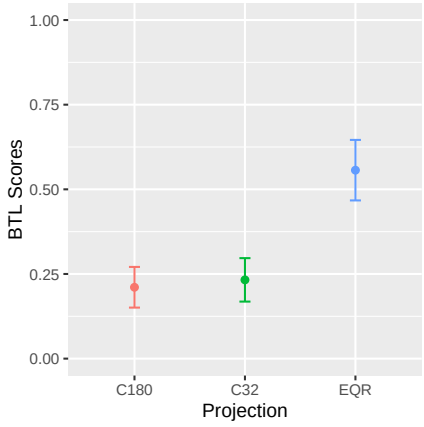


(b) Overall results

Figure 4.6 – Results of pair comparison. EQR: equirectangular projection, C32: cubic projection, C180: cubic projection with half resolution back part. The labels on the left of each plot denote the pairs A-B being compared.



(a) Normalized frequency of choosing projection A over B



(b) Scores for every projection calculated using Bradley-Terry-Luce model

Figure 4.7 – Subjective scores of remapping evaluation

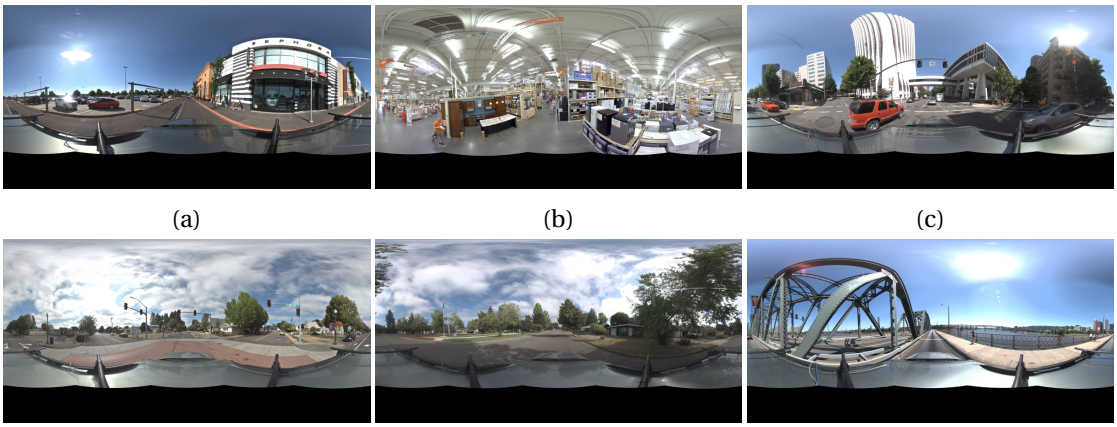
Results show that subjects find the quality of video samples in equirectangular representation higher than that of both cube maps. Moreover, cube map and half-back cube map scores lie within the margin of error with respect to each other.

4.3 Subjective quality evaluation experiments

4.3.1 Testbed pilot subjective evaluation

Six different 360-degree image contents were used in the experiment. Original unstitched and uncompressed raw samples from *Ladybug 5* omnidirectional camera were provided by Point Grey Research Inc. Images were processed and stitched to equirectangular projection using LadybugSDK software package. Figure 4.8 shows examples of equirectangular projections of the test samples used in the experiment.

Reference images in a prepared stimulus set were in lossless compressed PNG format, sRGB 8-bit color space, and a resolution of 3000x1500 pixels for equirectangular format and 2250x1500 for cube map. Test stimuli were prepared by compressing the reference images with a JPEG encoder using four different quality parameters for each image: 20, 45, 60, and 92. The latter was set to achieve transparent quality. The stimuli used in the experiment contained a training set of five samples obtained from the same content representing all the quality levels, including original and an evaluation set consisting of 25 images for each of two projections obtained from five different contents. To select the lower and upper quality bounds, an expert screening session was conducted for each content separately with the aim to cover the full quality scale for each content.



(a) (b) (c)

(d) (e) (f)

Figure 4.8 – Dataset for testbed pilot subjective evaluation

The Absolute Category Rating with Hidden Reference (ACR-HR) methodology [1] was selected for evaluations. A five-grade quality scale (1: Bad; 2: Poor; 3: Fair; 4: Good; 5: Excellent) was used. The subjects were asked to judge the overall quality of the evaluated omnidirectional images. To reduce contextual effects, the order of rendered stimuli was randomized so that the same content was never shown consecutively. The randomization was done automatically by the testbed software.

Half of 48 subjects participating in the subjective assessment study evaluated test stimuli rep-

resented in equirectangular format; another half evaluated the cube map projection. Subjects (36 males and 12 females) were between 19 and 36 years old with a corresponding age average and median of 25.1 and 24.7 years, respectively.

In order to minimize visual fatigue effects, each test session was designed to take no longer than 15 minutes. Prior to each experiment, short written training instructions were provided to subjects to explain their tasks. Subsequently, more detailed instructions were shown on the screen during the training session, where five training samples, representing all quality levels were displayed to familiarize subjects with the assessment procedure. The training instructions and samples as well as the entire test were presented to subjects using the evaluation testbed described in Section 4.1.3.

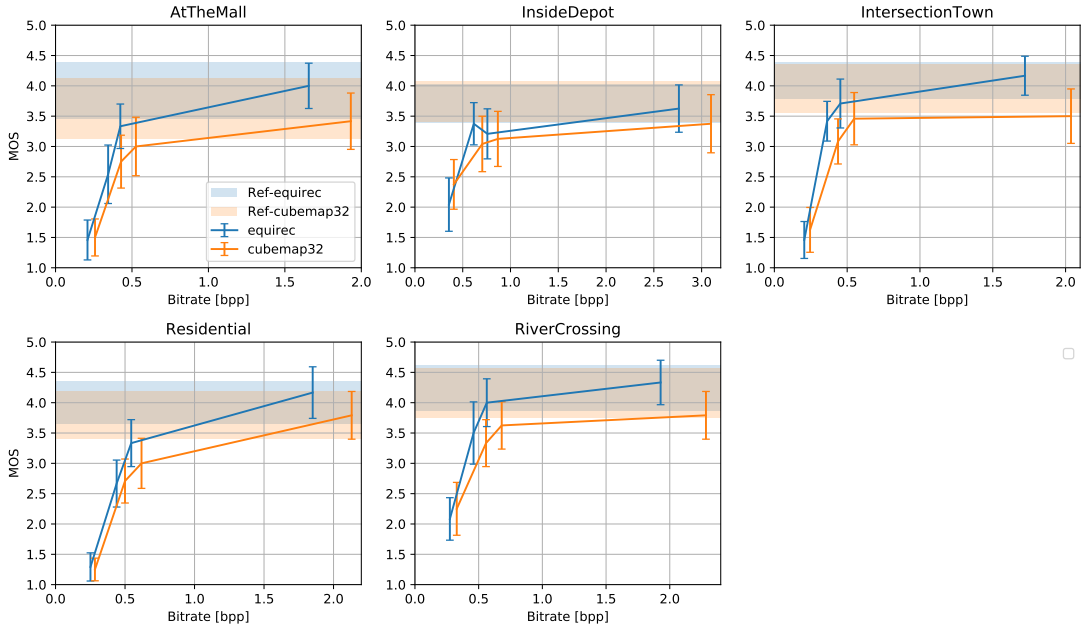


Figure 4.9 – MOS with CI obtained using ACR-HR method for JPEG-compressed omnidirectional images. Blue lines depict MOS for equirectangular projection, orange lines are for cube map. Collor-filled area corresponds to the confidence interval of the reference for each projection (blue for equirectangular, orange for cube map). JPEG quality parameters are: 20, 45, 60, 92.

Subjects were asked to observe the images using a particular HMD mount ("MergeVR"[1]) with a mobile device (iPhone 6S) placed inside as a screen. Resolution of the device screen was 1334x750 pixels overall, or 667x750 per eye. The HMD mount had two buttons on the top to select and scroll within a menu in order to rate the quality of presented stimuli. All subjects were asked to sit on a revolving chair with an ability to easily rotate left and right.

In order to evaluate perceived quality, standard statistical indicators describing distribution of scores across subjects for each test condition were computed. Firstly, outlier detection was

[1] https://mergevr.com/

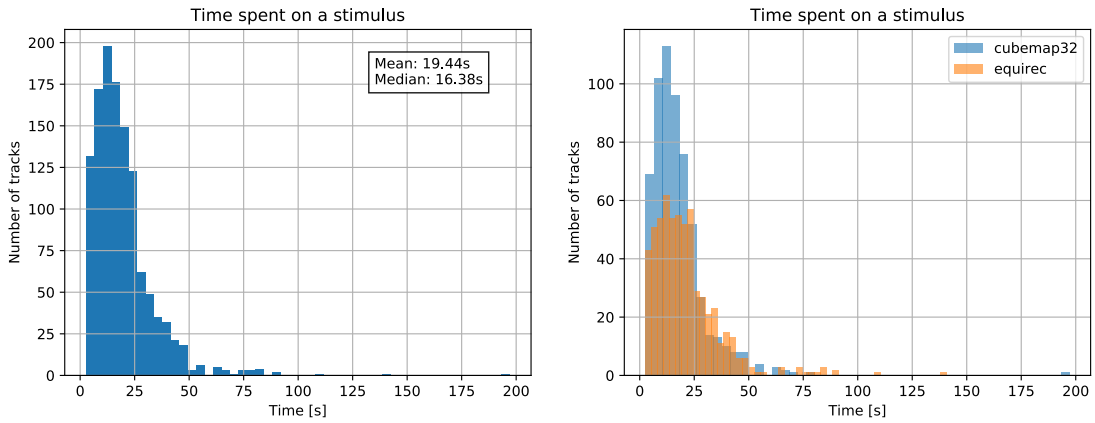


Figure 4.10 – Histogram of time spent on a stimulus by subjects during testbed pilot subjective evaluation experiment.

performed according to the guidelines described in Section 2.3.1 of Annex 2 in [2] to remove subjects whose scores deviated strongly from others. Two subjects were detected as outliers for all test sessions. Secondly, the mean opinion scores (MOS) and corresponding 95% confidence intervals (CI), assuming a Student's t-distribution of the scores, were computed for each test condition.

Figure 4.9 shows the resulting MOS and CI plots for equirectangular and cubic projections obtained from subjective evaluation experiments performed using ACR-HR method. Color-filled areas correspond to the confidence intervals of the reference stimuli for each projection. One can observe in the plots that, for certain contents, stimuli which were compressed with quality parameters 60 and 92 were perceived as having transparent quality compared to the reference, and, for other contents, stimuli which were compressed with the quality parameter 45 are also perceived as transparent. However, the MOS for the stimuli with the lowest quality parameter 20 had statistically significant differentiation from the neighbors for all the contents.

There exists a tendency among subjects to under-rate stimuli. It can be possibly explained by the fact that many of participants reported themselves evaluating existing stitching artifacts, despite the instructions. According to the recorded values of the time spent to evaluate each test stimulus by the participants of the experiment with a combined number of 48 volunteers, in average subjects looked at each omnidirectional image for a duration of 30 seconds (see Figure 4.10). These data can be used in designing future subjective quality evaluation tests for omnidirectional images and video.

4.3.2 Impact of compression and projections on perceived visual quality

Four high fidelity uncompressed omnidirectional images represented in equirectangular projection were used to compose an evaluation test-set; an additional image different from

Figure 4.11 – Dataset used in impact of compression and projections experiment

those four was selected for training. The dataset is based on omnidirectional video test-set proposed by Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG. Omnidirectional video sequences were examined, and a still frame from each of the selected four sequences was taken to compose the dataset. Figure 4.11 depicts the selected contents in equirectangular projection. In order to keep the balance in spatial complexity and colors, spatial index (SI) [1] and colorfulness (CF) [48] were taken into account when selecting the images. Table 4.1 shows SI and CF for the contents used in the experiment.

Original images are represented in YUV color-space format with 4:2:0 chroma sub-sampling. Initial high resolution images were down-sampled using bi-cubic interpolation to 3000 × 1500 pixels in order to correspond with the resolution of the HMD screen used in the experiments. Reference original images were then remapped to a cubic projection with rotated faces. This projection is a variation of a cube mapping introducing the least amount of additional non-continuities (face edges) to the picture. Figure 4.11 (e-h) shows examples of images represented in the cubic projection with rotated faces.

Both equirectangular and cubic images were compressed with three different codecs, namely JPEG, JPEG 2000, and HEVC intra. Afterwards, an expert screening was conducted to select bitrates representing the full scale of visual quality. As a result, four target bitrates, namely 0.25, 0.50, 0.75, and 1.00 bits per pixel, were selected. To compress original images with JPEG, JPEG 2000, and HEVC, the libjpeg[2], OpenJPEG[3], and FFmpeg with x265[4] software packages were used, respectively. The selected parameters and settings for all the codecs exploited in the subjective experiments are presented in Table 4.2. It also contains the details on the command-line arguments used to produce the compressed images. In order to perform subjective assessments, all the encoded images were decompressed using the same respective software packages to produce reconstructed impaired stimuli.

[2]https://github.com/thorfdbg/libjpeg. Commit: 0x0009dcc
[3]https://github.com/uclouvain/openjpeg. Ver.: 2.1.2, commit: 0x1f1e968
[4]https://ffmpeg.org/ Ver.: 3.2.2, http://x265.org/ Ver.: 1.9

Table 4.1 – Spatial index (SI) and colorfulness (CF) computed for the test contents used in the experiment.

	Harbor	KiteFlite	PoleVault	SkateboardTrick
Spatial Index (SI)	7.96	10.45	10.33	6.31
Colorfulness (CF)	15.64	14.88	42.81	25.62

Table 4.2 – Selected parameters and settings for all codecs exploited in the subjective experiments.

Codec	Software	Command line
JPEG	libjpeg	`jpeg -q quality referencefile.png compressedfile.jpg`
JPEG 2000	OpenJPEG	`opj_compress -q quality -i referencefile.ppm -o compressedfile.j2k`
HEVC	libx265	`ffmpeg -f rawvideo -r 1 -s size -pix_fmt yuv420p -i infile.yuv -c:v hevc -crf qlt outfile.hevc`

The experiment was conducted in the Multimedia Signal Processing Group (MMSPG) laboratory at EPFL where naïve subjects were invited to participate. It was performed according to Absolute Category Rating with Hidden Reference (ACR-HR) method described in [1]. ACR-HR is a single stimulus evaluation where the reference stimuli are randomly shown to observers among the impaired images. Stimuli were presented subsequently to subjects, and voting was performed after each viewing. Images were assessed using five-grade quality scale with the following levels: "*5 - Excellent*", "*4 - Good*", "*3 - Fair*", "*2 - Poor*", and "*1 - Bad*".

The observers were placed in an immersive environment where omnidirectional images were presented to them by means of an HMD. Immersive textual instructions were provided inside the VR along with a verbal guidance by the experimenter. Every test session started with an immersive training that consisted of three consequently shown images of a content not used in the evaluations. Subjects observed the examples of "Excellent", "Bad", and "Fair" quality levels shown in this particular order and were provided with the explanations and illustrations of impairment artifacts which can be found in the images. During the evaluation, subjects assessed the stimuli shown to them consequently without any time restrictions. When ready to rate an image subjects had to activate a 3D immersive voting menu by pressing a button and select the grade proceeding immediately to the next image. All stimuli were automatically randomized in each session.

All the steps described above in the current subsection including immersive training and evaluation were conducted using a testbed for subjective evaluation of omnidirectional visual content proposed in [111]. This software was developed for iOS and Android platforms. The source code is publicly available for download[5]. It renders omnidirectional images with OpenGL using perspective projection and bi-cubic interpolation. The testbed allows uploading test stimuli to a device and changing immersive textual instructions. Voting data is acquired by the software and stored on the device. It can be further transmitted to a server for processing.

[5] https://github.com/mmspg/testbed360

The following hardware equipment was used to perform immersive subjective quality evaluation of omnidirectional images along with the software testbed. During the experiment subjects were wearing an HMD composed of a VR head-mount with buttons[6] and a mobile device installed inside as a screen. An iPhone 6 was used to display the images. The overall resolution of the phone screen was 1334 × 750 pixels, which gives 667 × 750 pixels per eye. The vertical field of view provided by the hardware-software solution is 90 degrees and corresponds to 8.33 pixels per degree. All the subjects were sitting on a revolving chair during the assessment.

Prior to the experiment, a non-immersive training was provided to the subjects. The experimenter explained the purpose of the evaluation, showed examples of compression artifacts, and pointed to differences between coding and stitching artifacts. Subjects were instructed not to assess stitching artifacts. The test material, consisting of 104 test stimuli, was randomly distributed between two sessions. Each participant took part only in one session in order to shorten the time when the subject is exposed to VR immersive environment to a maximum of 25 minutes. Overall, 41 naïve subjects participated in the experiment. One subject was not able to complete the evaluations due to motion sickness. Subjects, 25 males and 15 females, were between 18 and 32 years old with an average and median of 24.9 and 24.8, respectively. All the participants were tested for correct color vision and visual acuity using Ishihara and Snellen charts respectively. An additional evaluation session was independently conducted with 5 expert viewers at the 74th JPEG meeting in Geneva.

Outlier detection was performed separately on the raw experimental data from each of two test sessions, since an individual subject had only assessed stimuli from one subset. A boxplot based method was used to remove outliers in the same way as in [29]. One subject was detected as an outlier in the first session. Therefore, to preserve the symmetry of the data, one randomly selected subject was removed from the second session. MOS values were computed for each stimulus in the entire dataset as mean values for the set of scores provided by different subjects. In order to estimate statistical significance, 95% confidence intervals (CI), assuming a Student's t-distribution of the scores, were computed alongside with MOS values.

Figure 4.12 shows MOS and CI plotted for different contents. MOS obtained from naïve subjects were shown to be highly correlated with expert subject results. More particularly, standard correlation indexes between naïve and expert scores are *PLCC = 0.95*, *SROCC = 0.87*, *RMSE = 0.40*. Certainly, this allows us to consider the subjective evaluation results being reliable and consistent.

The results of the subjective evaluation experiment and the data analysis show, as expected, higher performance of HEVC and JPEG 2000 when compared to JPEG at lower bitrates. Some of the contents, however, namely "Pole Vault" and "Skateboard Trick", are systematically underrated, which can be possibly explained by the lower perceptual quality of the original pictures. Other explanations for the former can be the following. There are many human faces

[6] https://mergevr.com

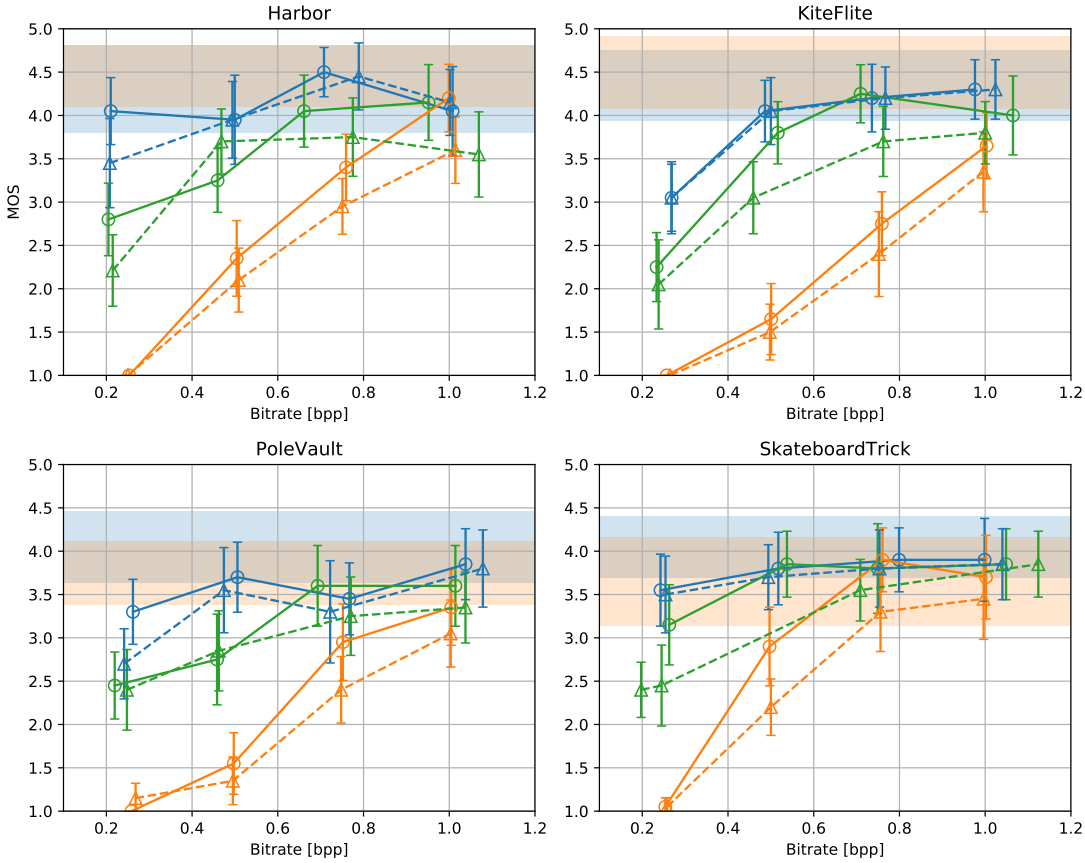


Figure 4.12 – Mean opinion scores of compression and projections experiment.

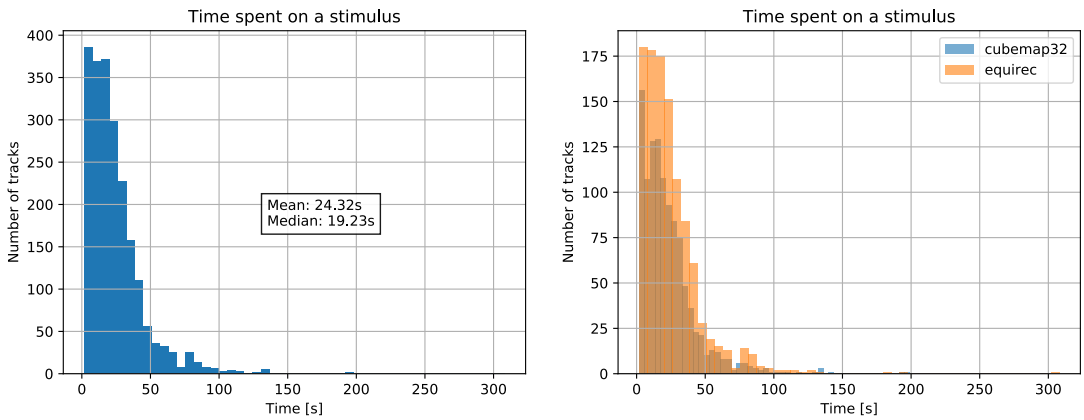


Figure 4.13 – Histogram of time spent on a stimulus by subjects during subjective evaluation experiment.

in the "Pole Vault" content, and thus, due to a relatively low resolution of the HMD screen, observers' expectations to distinguish facial features were not met. In the "Skateboard Trick"

content, there is an artificially blurred circle below the camera used to camouflage a tripod, which could influence the decision of naïve subjects.This hypothesis is supported by the fact that in the expert subjective results "Skateboard Trick" reference stimuli was not underrated, whilst "Pole Vault" was.

When compressing images represented in a cube map projection, edges of continuous parts of the frame are distorted non-uniformly with different intensity. This makes cube-face borders distinguishable for some stimuli in the rendered viewport when observed using an HMD. Experimental results, indeed, show lower scores for cubic mapping at medium bitrates and the same scores as for equirectangular mapping at high and low bitrates. This may occur for the reason that at high bitrates there are no impairments, and at low bitrates the entire image is distorted. Thus, the cube-facet borders are distinguishable only at medium bitrates, due to compression artifacts.

5 Objective Metrics for Perceptual Visual Quality

In this chapter, we review the state of the art in objective perceptual visual quality measurement for omnidirectional content. Afterwards, we present results of benchmarking of existing objective metrics against subjective mean opinion scores. In the second part of this chapter, we propose a novel method for objectively assessing perceptual visual quality based on visual attention, following with its validation and conclusive discussion.

5.1 Review of the state of the art

An important part of the development of future 360-degree image and video compression algorithms is related to the selection of objective metrics required to automate the process of visual quality assessment. Thus far, however, there is no agreement on which metrics should be used to predict perceived quality of omnidirectional content, as there is not enough evidence about their performance. Recently, new objective metrics for omnidirectional content have been introduced [124, 127]. To benchmark the available objective metrics, a ground-truth data is necessary. The most reliable way to obtain such ground-truth data is by means of subjective

quality evaluation. Further in this chapter, we use the results of the experiments described in Section 4.3 of Chapter 4.

State-of-the-art research on perceptual visual quality assessment of omnidirectional content mainly focuses on adaptation of conventional full-reference objective metrics in order to cope with geometrical distortions and spatial entropy redistribution introduced by different representations of such content. A review and benchmarking results of recently proposed objective quality metrics for omnidirectional visual content are provided by authors in [112, 111]. Among the proposed metrics methodology varies from applying forward-and-backward geometrical mappings as in [127] to different schemes of weighting during pixel-wise comparison as in [102, 124]. Croci et al. propose in [27] a framework for perceptual visual quality control in stereoscopic omnidirectional imaging. Their method considers empirical visual attention data to define the significance of regions.

5.2 Benchmarking of existing methods

We aim here to assess the performance of available objective metrics designed specifically for omnidirectional visual content against ground-truth subjective mean opinion scores (MOS). Additionally, a comparison to the performance of conventional 2D objective metrics has been carried out. In particular, the 96 mean opinion score (MOS) and corresponding confidence interval (CI) values to measure the correlation between objective and subjective scores is used. The objective metrics are evaluated, in terms of commonly used performance indexes, i.e. linearity, monotonicity, accuracy, and consistency, based on their correlation with the perceived visual quality. It is shown that the VIFp objective metric provides the best performance indexes. However, overall results indicate the need for new algorithms, which better predict perceived quality of omnidirectional content.

5.2.1 Objective evaluation

This section presents objective evaluation data for omnidirectional visual content obtained by calculating particular metrics. Performance of these metrics is then evaluated by comparison to the ground-truth subjective scores. Finally, the results are presented alongside with a discussion.

Omnidirectional visual content can be assessed with conventional 2D objective metrics as well as with metrics designed specifically for 360-degree images. Here we provide a list of objective evaluation methods used in this study. The following objective metrics were computed:

- Conventional 2D metrics

  1) *Peak signal-to-noise ratio (PSNR)*
  2) *Structural Similarity (SSIM)*

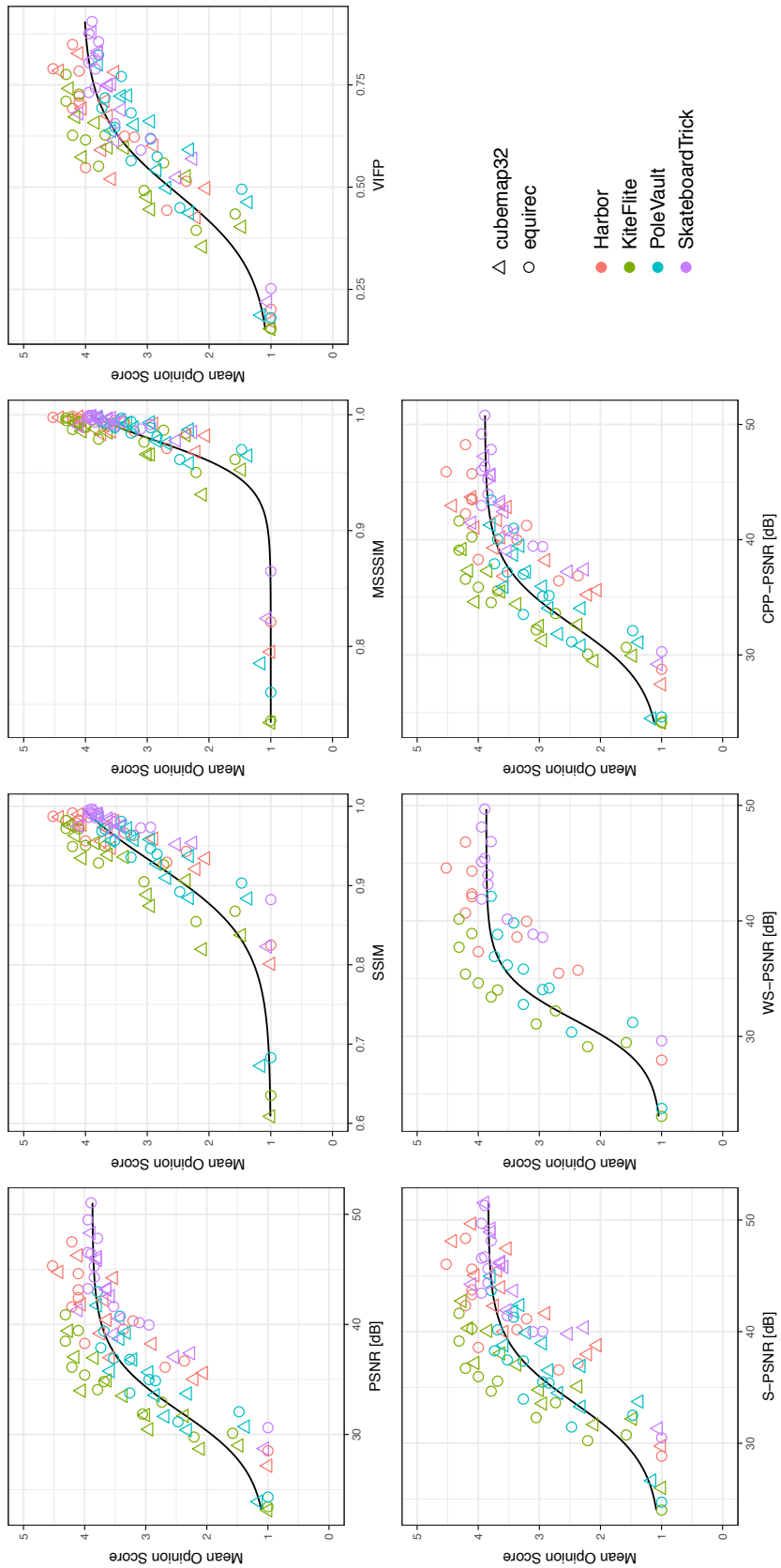


Figure 5.1 – Mapping of objective scores to subjective ratings. Triangles represent cubic projection, circles represent equirectangular projection. Different contents are marked with colors: "Harbor" - red, "Kite Flite" - green, "Pole Vault" - cyan, "Skateboard Trick" - magenta. Solid black line depicts a logistic fitting.

3) *Multi-Scale Structural Similarity (MSSSIM)*
4) *Visual Information Fidelity in pixel domain (VIFp)*

- Metrics designed for omnidirectional visual content

1) *Spherical PSNR (S-PSNR)* computes PSNR for the set of points uniformly distributed on a spherical surface, where corresponding pixels from a reference and an assessed image are reprojected to this set [124].
2) *Weighted Spherical PSNR (WS-PSNR)* computes PSNR in such a way that intermediate values for pixels in an equirectangular image of height h are weighted with a coefficient $w_{i,j} = cos((i - h/2)\pi/h)$ [103]. This weighting reduces the impact of the pixels with higher latitudes. It should be noted that WS-PSNR is only applicable for images in equirectangular projection.
3) *Craster Parabolic Projection PSNR (CPP-PSNR)*. Both an assessed image and a reference are re-mapped to a Craster parabolic projection, then PSNR is computed in that domain [127].

To compute conventional objective metrics, namely PSNR, SSIM, MSSSIM, and VIFP, a publicly available software package *VQMT*[1] was used. For metrics designed specifically for omnidirectional content, S-PSNR, WS-PSNR, and CPP-PSNR, publicly available *Samsung 360 Tools*[2] were used.

5.2.2 Performance evaluation

Standard performance indexes, namely, the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SROCC), the Root mean square error (RMSE), and the Outlier ratio (OR), were computed to compare objective results with the ground-truth subjective ratings. To calculate the above listed performance coefficients, the raw objective evaluation data was fitted to the MOS values. Logistic fitting was performed considering that the data were in different ranges and in order to compensate possible saturation of subjective scores. One can see the fitted curves in Figure 5.1.

Table 5.1 presents linearity, monotonicity, accuracy, and consistency indexes. These indexes were computed assuming different mapping schemes of test data:

A - for equirectangular projection, on all the contents,
B - for cubic projection, on all the contents,
C - for both projections, on all the contents,
D - for both projections, each content separately.

[1] http://mmspg.epfl.ch/vqmt
[2] https://github.com/Samsung/360tools. Commit: 0x54845f0

Table 5.1 – Standard performance indexes. Subcolumns A, B, and C, represent the results for equirectangular, cubic, and both projections computed over all the contents, respectively. Subcolumn D shows an average of coefficients computed for each content separately.

	PLCC				**SROCC**				**RMSE**			
Metric	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
PSNR	0.8714	0.8437	0.8553	0.9487	0.7176	0.7731	0.7567	**0.8909**	0.4804	0.5103	0.5008	0.2929
SSIM	0.8898	0.8632	0.8740	0.9459	0.7365	0.7927	0.7709	0.8821	0.4464	0.4790	0.4689	0.3050
MSSSIM	0.9059	0.8661	0.8860	0.9123	0.7539	0.7796	0.7814	0.8394	0.4143	0.4755	0.4483	0.3887
VIFp	**0.9116**	**0.8875**	**0.8994**	0.9319	**0.7608**	**0.8029**	**0.7953**	0.8538	**0.4025**	**0.4374**	**0.4221**	0.3395
S-PSNR	0.8766	0.8482	0.8392	0.9168	0.7376	0.7836	0.7307	0.8214	0.4715	0.5035	0.5257	0.3705
WS-PSNR	0.8748	-	-	**0.9583**	0.7297	-	-	0.8648	0.4746	-	-	**0.2544**
CPP-PSNR	0.8800	0.8521	0.8658	0.9467	0.7403	0.7745	0.7697	0.8843	0.4654	0.4975	0.4838	0.2966

More specifically, for A and B cases the fitting was performed only for the data points representing each individual projection, for C and D cases the fitting was performed on all the contents to compute the indexes. Moreover, for D case, an average of resulted indexes for each content was considered.

5.2.3 Results and discussion

Figure 5.1 shows scatter plots of MOS values against objective metrics. For the cases A, B, and C, the results of the objective metrics performance evaluation show only moderate correlations with the ground-truth subjective scores and do not significantly change for different projections. As can be seen from scatter plots in the Figure 5.1, points are sparse and not concentrated along the fitting curve. Moreover, objective metrics designed specifically for omnidirectional visual content do not show better performance when compared to common objective quality evaluation measures. For the case D, performance per content is significantly higher compared to cases A, B, and C. However, conventional metrics still outperform those designed for 360-degree content.

Since S-PSNR, WS-PSNR, and CPP-PSNR are all based on PSNR, it is reasonable to compare them. Looking at the scatter plots in the Figure 5.1, one can notice that the distribution patterns of the score points are of high similarity for all the PSNR based metrics showing strong content dependency.

Analysis of the obtained subjective and objective scores indicates moderate performance of the investigated metrics for omnidirectional visual content. Being PSNR based, these metrics do not outperform significantly their ancestor in predicting visual quality of omnidirectional content. All the evidence above suggests that the problem of better objective quality evaluation methods for omnidirectional visual content remains open. Future work should consider developing a more suitable objective metric for 360-degree content.

5.3 Saliency driven perceptual quality metric

In this section, we propose a novel objective perceptual visual quality metric which takes into account ground-truth viewer's visual attention information in order to make image quality assessment selective with respect to regions of interest.

5.3.1 Visual attention weighting

Weighting pixel-based objective quality metrics is a well known approach. However, to the best of our knowledge, very little has been done regarding visual attention in this topic.

As a base for our method we decided to use the Peak Signal to Noise Ratio (PSNR) metric because it is widely accepted, its implementation is simple, and its performance is satisfactory

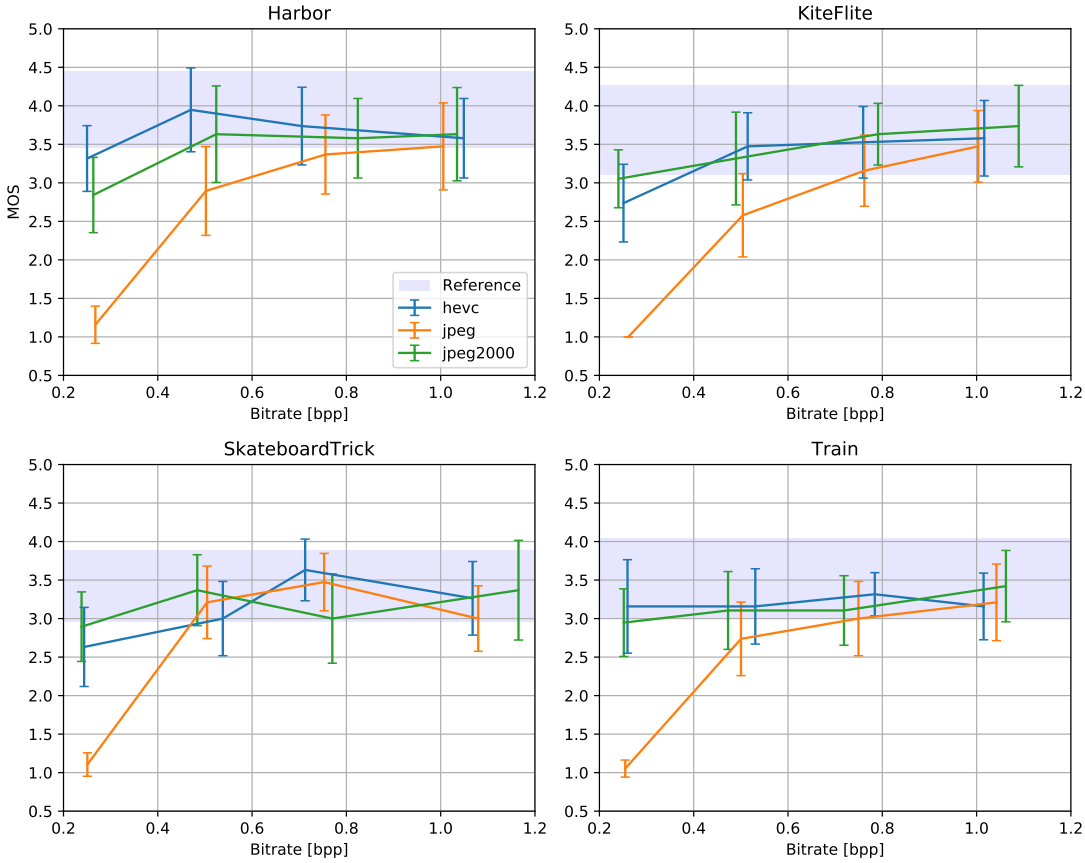


Figure 5.2 – Subjective mean opinion scores (MOS) with 95% confidence intervals. The area filled with transparent purple color depicts the 95% confidence interval of the hidden reference.

to test our hypothesis. We define a ground-truth image as $I(i,j)$, where $i = 0,1,...,H, j = 0,1,...,W$, with W and H dimensions of the image. The impaired image is defined as $\hat{I}(i,j)$. Thus, PSNR is described by the following equation:

$$PSNR = \frac{MAX_I^2}{MSE}$$

where

$$MSE = \frac{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left(I(i,j) - \hat{I}(i,j)\right)^2}{H * W}$$

and MAX_I is the maximum possible value of pixel intensity of the assessed image, e.g. for an 8-bit image it equals 255.

Given that sufficient amount of empirical data of head movements is available for an assessed omnidirectional image, one can obtain a continuous visual saliency map using the method described in [109].

The saliency map can be defined as:

$$h_{i,j} \in [0,1], i = 0,1,...,H, j = 0,1,...,W$$

where each pixel of $h_{i,j}$ provides a visual attention value for each corresponding pixel of $\hat{I}(i,j)$. The saliency map $h_{i,j}$ can be obtained independently for different degradation levels of impaired images. This issue is further addressed in Section 5.3.4.

Visual saliency map is used to compute a saliency-weighted mean square error MSE_{VA} which contributes to the PSNR equation as a denominator.

$$MSE_{VA} = \frac{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left(I(i,j) - \hat{I}(i,j)\right)^2 h_{i,j}}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} h_{i,j}}$$

Therefore, a Visual Attention PSNR (VA-PSNR) is defined as:

$$PSNR_{VA} = \frac{MAX_I^2}{MSE_{VA}}$$

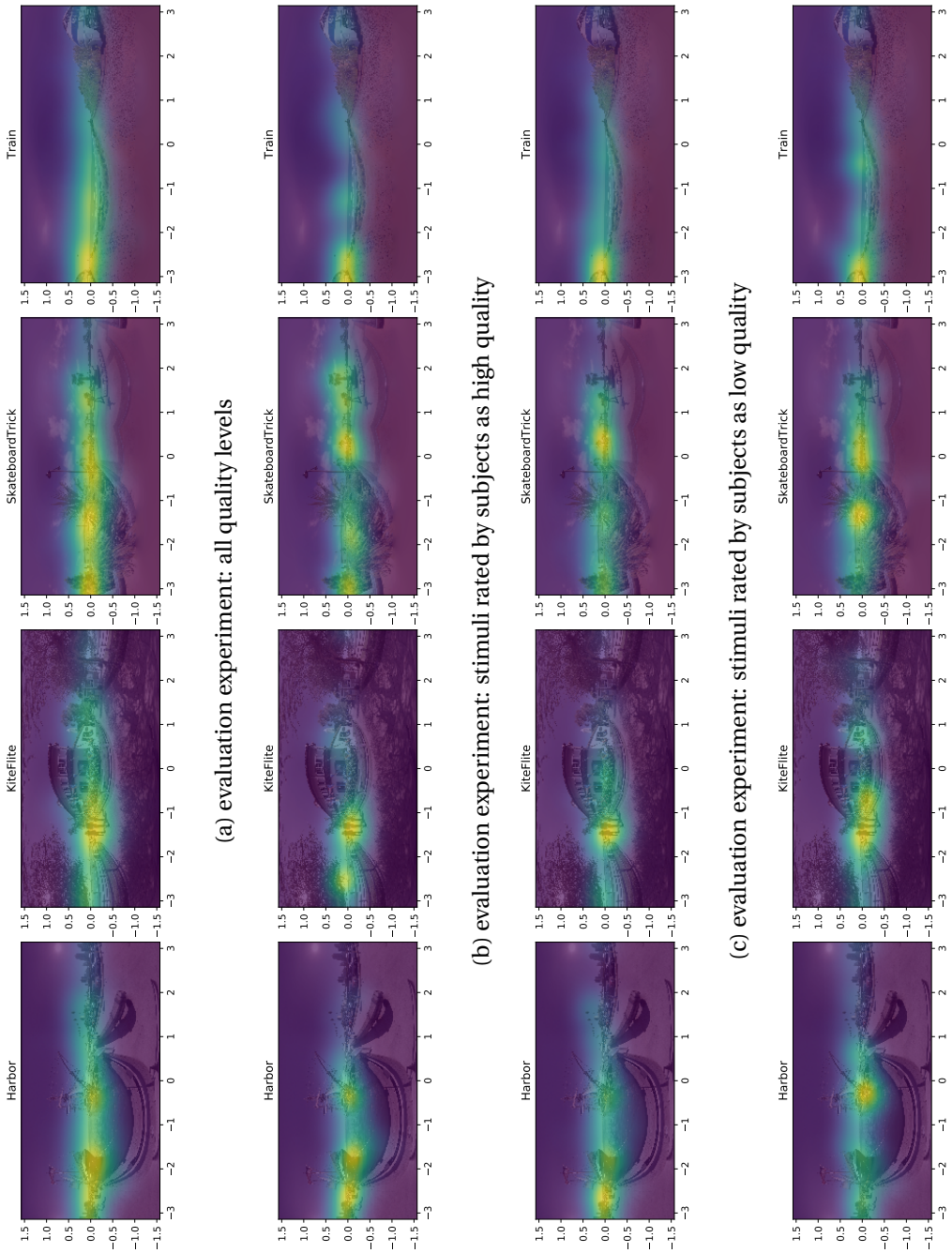
VA-PSNR allows comparison of two omnidirectional images regardless of the projection (equirectangular, cubic etc.) they are represented in, provided that both are represented in the same way.

The source code and data are publicly available on-line at:
`https://github.com/mmspg/saliencymetric360`

5.3.2 Validation with subjective experiments

This section describes experiments performed by the authors with the goal to validate and benchmark proposed objective perceptual visual quality evaluation method.

Two independent content viewing sessions were conducted. Participants were divided in two disjoint groups: one was asked to evaluate omnidirectional images according to visual quality, whilst another performed free exploration with a dummy task to assess the aesthetic value of the pictures.



(a) evaluation experiment: all quality levels

(b) evaluation experiment: stimuli rated by subjects as high quality

(c) evaluation experiment: stimuli rated by subjects as low quality

(d) exploration experiment

Figure 5.3 – Visual attention heatmaps obtained from experimental data.

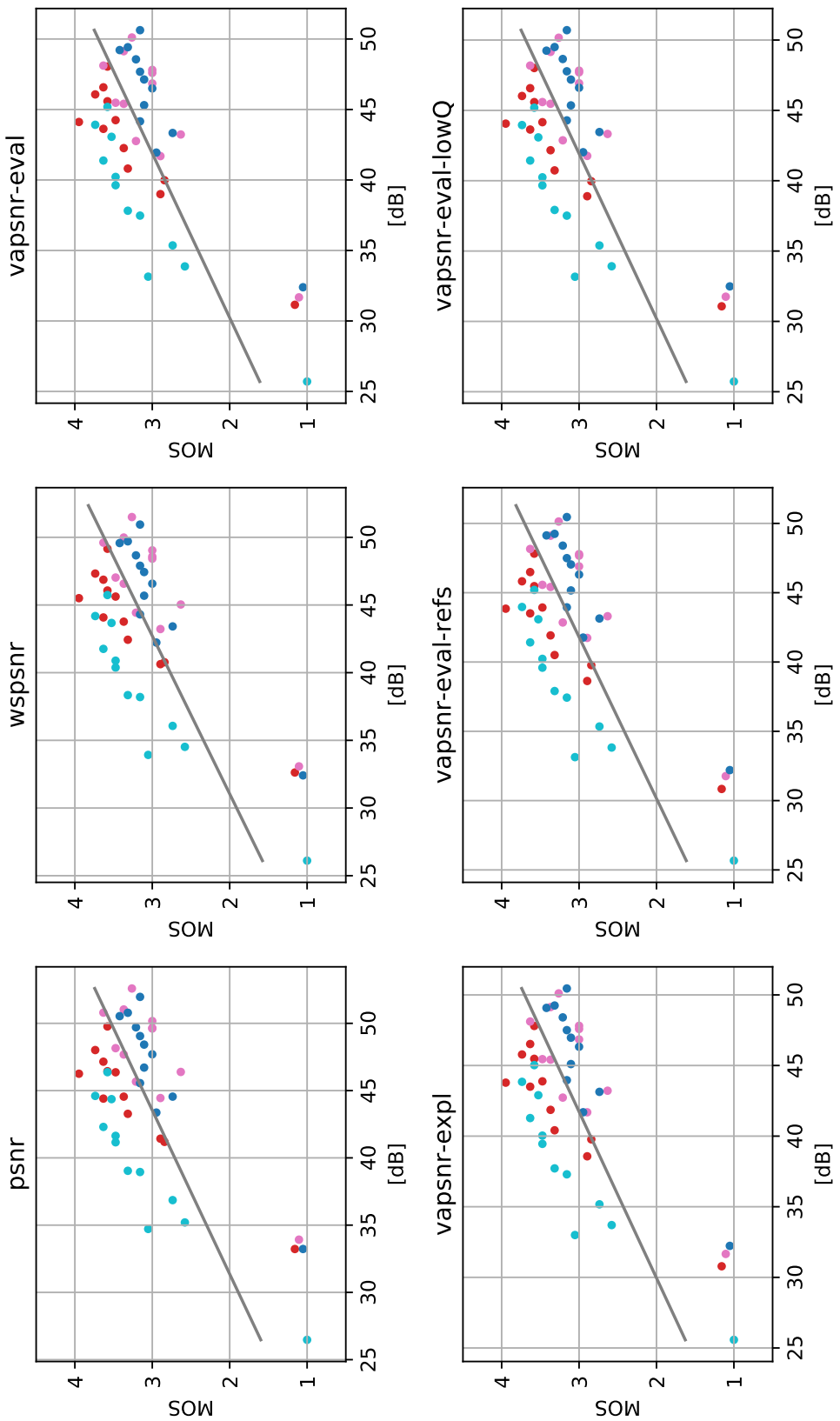


Figure 5.4 – Mapping of objective scores to subjective ratings. Grey line depicts linear fitting. Different colors represent different contents: blue - Train, red - Harbor, cyan - SkateboardTrick, magenta - KiteFlite.

	PSNR	WS-PSNR [102]	VA-PSNR Eval	VA-PSNR Expl	VA-PSNR Eval-Refs	VA-PSNR Eval-LowQ
PLCC	0.6959	0.7106	0.7107	0.7074	**0.7114**	0.7083
SRCC	0.3706	0.4131	0.4131	0.4075	**0.4163**	0.4080
KRCC	0.2706	0.2976	**0.3012**	0.2904	0.2976	0.2958

Table 5.2 – Standard performance indexes. Pearson linear correlation coefficient (PLCC), the Spearman rank correlation coefficient (SRCC), and Kendall rank correlation coefficient (KRCC). Bold text shows the best result per index, italic text shows the second best result for PLCC.

Dataset and equipment

Selected contents were compressed using three different codecs, namely JPEG, JPEG 2000, and HEVC Intra-frame. Images were compressed using the same software as in [112] with the quality parameters specified in Table 5.3. Original high-fidelity images were downscaled to 5760×2880 pixels before compression in order to comply with technical requirements of the display.

Codec	Harbor	KiteFlite	Skateboard..	Train
JPEG	9,53,79,87	4,23,54,73	8,71,87,93	8,65,85,92
JPEG 2000	41,44,46,47	35,39,42,44	44,47,49,51	43,46,48,50
HEVC-I	32,27,24,21	37,30,26,23	29,23,21,18	30,24,21,19

Table 5.3 – Quality "Q" parameters used to encode images. The software is the same as used in [112].

Experiments were conducted with the help of a testbed for subjective evaluation of omnidirectional content proposed in [111] which is publicly available for downloading[3]. Participants were observing stimuli using a head-mount[4] with a mobile device acting as a screen. Galaxy S7 Edge SM-G935F was used. The resolution of the device was 2560×1440 pixels. During the experiments, subjects were sitting on a rotating chair. All subjects passed color vision and visual acuity tests.

5.3.3 Evaluation and exploration

During the evaluation experiment subjects were assessing omnidirectional images following the methodology called Absolut Category Rating with Hidden Reference (ACR-HR). They were asked to rate stimuli on the five-level scale ”5 - Excellent”, ”4 - Good”, ”3 - Fair”, ”2 - Poor”, and ”1 - Bad”. 19 subjects participated in the evaluation session, among which 9 were females, with an overall median age of 24.5. Results of subjective assessment are presented in Figure 5.2.

Exactly the same set up as for evaluation was used in the exploration experiment. However, subjects were asked to evaluate the aesthetic value of the pictures and only uncompressed

[3]https://github.com/mmspg/testbed360-android
[4]https://mergevr.com

stimuli were used. Their subjective scores were discarded and only head direction tracks were collected. Exploration sessions had 17 participants, of which 10 were females, with an overall median age of 24.3.

5.3.4 Visual attention and quality

Head direction tracks were collected from both evaluation and exploration experiments. They were processed according to the method described in [109] in order to produce saliency maps. Additionally, raw visual attention data from evaluation sessions were grouped into three categories: all tracks, tracks from stimuli which have Mean Opinion Scores (MOS) lying withing the 95% confidence interval of hidden reference, and with MOS lower then 3.0. The resulting saliency maps are depicted in Figure 5.3.

5.3.5 Validation and discussion

VA-PSNR as well as other metrics were computed for all the stimuli using each set of saliency maps described in Section 5.3.4. Standard performance indexes were calculated (Table 5.2). Notably, VA-PSNR-Refs computed using saliency maps from high quality evaluation stimuli outperforms VA-PSNR-Expl, VA-PSNR-Eval, and VA-PSNR-lowQ computed using maps from exploration sessions, from all evaluation tracks, and from low quality evaluation stimuli tracks respectively.

The proposed method requires empirical visual saliency data and it can be applied in post-production of cloud services where, after a certain time from the moment of initial release, sufficient amount of data can be collected and used *a posteriori* to estimate quality during re-compression of the content which can be beneficial for saving bandwidth.

We proposed a novel method called VA-PSNR which estimates perceptual quality of omnidirectional content considering visual attention. We validated our method against subjective MOS and benchmarked it against state-of-the-art objective metrics. VA-PSNR shows better performance when compared to alternative approaches based on PSNR.

6 Coding of Omnidirectional Visual Content

In this chapter, we propose OmniJPEG, a JPEG backward-compatible solution to encode omnidirectional images. In order to ensure the JPEG backward compatibility, OmniJPEG extracts predefined regions of interest from omnidirectional images, as well as uses properties of equirectangular projection, while at the same time also keeps complete equirectangular information to preserve the capability of correctly rendering an omnidirectional image with appropriate devices and software.

6.1 Review of existing methods

The approaches to code omnidirectional visual content have been proposed in the past few years. The most notable of these approaches to compress omnidirectional images and video presented in the literature can be listed as follows: a) adaptive and partial content delivery methods, b) algorithms exploiting or adapting to 2D spherical surface geometry of the content to be compressed, and c) geometric representation or projection based methods as pre- and postprocessing prior to compression.

One of the very first attempts to introduce a solution to code an omnidirectional or a panoramic image data was developed by Apple Inc. The so-called Quicktime VR [22], which refers to both a file format and visualization software, allows for creation and display of panoramic images. More specifically, it proposes to store 360 degree cylindrical panoramic images divided in tiles. While displaying panoramic images with a specially designed viewer, only the tiles visible in the current viewport are decoded. However, the proposed panoramic image format did not

have backward compatibly with conventional image viewers, such as legacy JPEG decoders, and it did not consider any region of interest (RoI) estimation to provide users with a default viewport.

The idea of using RoI to code omnidirectional visual content recently appeared in a work considering adaptive coding and partial delivery methods [93]. This approach proposes to deliver only a part of the omnidirectional content which is being viewed. Each frame, after an equirectangular projection, is divided into tiles or regions which are then coded separately with a different quality according to an adaptive model. The tiles covered by the portion of the frame which is being viewed are encoded with the highest possible quality. The quality of other regions is determined considering their probability of being viewed next. However, authors did not investigate any statistical model to predict the next candidate tiles which will be viewed. Additionally, it did not consider that an omnidirectional image can be divided into tiles corresponding to a predicted RoI.

Coding by taking into account a specific 2D spherical surface has also been investigated in the past [106]. Assuming that a raw image from all different types of cameras after stitching can be mapped onto a sphere, authors proposed a generic compression method based on decomposition over a dictionary of geometric atoms. A redundant dictionary is built over two generating functions (low frequency and high frequency) extended with scaling and affine transformations on a 2D sphere. The coder performs matching pursuit to select atoms from the dictionary, sorts the atoms along the decreasing magnitude of their coefficients, and then applies adaptive quantization. The proposed codec outperforms JPEG 2000 at low bitrates, however, yields to it at high bitrates.

Representing omnidirectional visual content using geometric projections, which produce less amount of data, is another approach. Example of such an approach can be found in [38], where authors propose a rhombic dodecahedron (RD) mapping model. This convex polyhedron was chosen considering the restriction that the faces should be of quad-based nature, which allows constructing unfolded rectangular images for encoding. The model provides almost uniform pixel distribution without significant oversampling or undersampling, which allows applying DCT and wavelet based coding more efficiently when compared to alternatives, such as cubic mapping. However, this method has not been widely adopted, because of its complexity.

Regarding solutions with JPEG backward compatibility, several such approaches [47, 64] have been proposed in the past. However, none of them targets omnidirectional image compression.

6.2 OmniJPEG: JPEG backward compatible coding

We propose a format called OmniJPEG, as an extension of the most popular image format JPEG. The architecture of the corresponding proposed codec is designed with an emphasis on the backward compatibility with the legacy JPEG decoders. Additionally, we assess the

performance of the proposed approach.

As mentioned above, OmniJPEG exploits the RoI within an omnidirectional image, and relies on properties of the projection onto a plane, to visualize an omnidirectional image content from a representative viewport by a legacy JPEG decoder, while taking at the same time full advantage of the omnidirectional content by an extended decoder. Although, the proposed architecture for OmniJPEG can cope with various capture, representation, and projection alternatives. The prototype software implemented in the framework of this work only supports 360 omnidirectional images of a spherical model projected to an equirectangular representation.

6.2.1 Proposed coding architecture

The proposed scheme for JPEG backward compatible coding of omnidirectional images aims at providing the ability to visualize omnidirectional content in cases when advanced omnidirectional viewing is not available. Rendering of omnidirectional image content requires efficient implementation of sphere-to-plane projections, such as the widely used equirectangular or cube map projections. Therefore, to successfully and interactively visualize a desired viewport, a 3D graphics powered viewer is necessary. To display omnidirectional image information with a conventional 2D image viewer without a need for an advanced rendering algorithm, a predefined viewport representing a limited field of view of the original content has to be identified a priori. This viewport can then be encoded with a conventional 2D image compression algorithm and consequently displayed after decoding. However, if one can perform necessary geometrical transformations to a particular part of the omnidirectional image and embed it into the file, then the viewport can be displayed by any conventional image viewer without applying any rendering algorithm.

The block diagram of the proposed JPEG backward compatible coding scheme for omnidirectional image content, OmniJPEG, is presented in Figure 6.1. The input is either a compressed or an uncompressed omnidirectional image in one of the sphere-to-plane geometric projections (such as equirectangular or cube map). The input data is first read into a memory block and, if needed, decompressed. Afterwards, the RoI of an input image is extracted and its corresponding viewport is generated and geometrically corrected. Then, a viewport image is compressed with a conventional JPEG encoder and stored in the JPEG File Interchange Format (JFIF) [58], which can then be displayed by any legacy viewer. Subsequently, the entire omnidirectional image is stored in the same JFIF file as a metadata.

A block diagram of OmniJPEG decoder is presented in Figure 6.2. The process of decoding consists in extracting the original omnidirectional image from metadata of the JFIF file. The obtained image can be displayed by an omnidirectional image viewer supporting OmniJPEG coding, or stored in a file in order to be displayed later by an omnidirectional visualization software.

On the other hand, Figure 6.3 depicts how an omnidirectional image encoded by the proposed

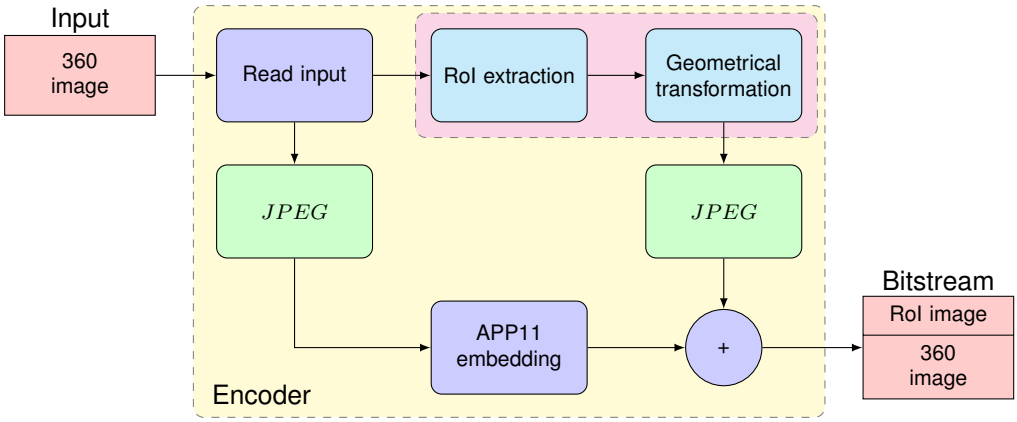


Figure 6.1 – Block diagram of the proposed OmniJPEG encoder.

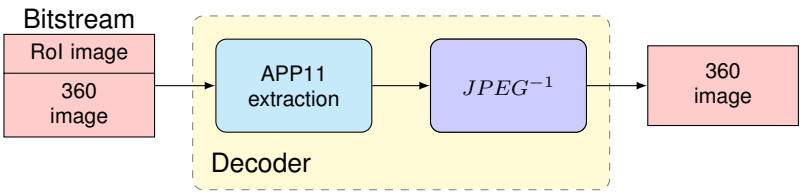


Figure 6.2 – Block diagram of the proposed OmniJPEG decoder.

OmniJPEG format can be decoded and displayed both by legacy JPEG viewers or by decoders supporting OmniJPEG decoding, respectively. Generally, an omnidirectional image can only be displayed by dynamically rendering the particular viewport with a special software allowing users to freely change the view direction. However, when an OmniJPEG encoded image file is opened with a conventional legacy viewer, the latter can only access the predefined viewport corresponding to a particular RoI specified during encoding. On the contrary, when an encoded image is opened by an omnidirectional viewer, which supports OnmiJPEG, the application has access to an entire omnidirectional image stored in OmniJPEG metadata.

6.2.2 Implementation

To test OmniJPEG encoding process, a prototype software was implemented. The implemented software allows to encode and decode image files with the proposed OmniJPEG algorithm. The assumed input is an omnidirectional image represented in equirectangular projection. The software reads the input image to a memory block and decompress it, if already compressed.Then, it extracts the predefined RoI and renders its corresponding viewport, performing geometric transformations necessary for the given position within the omnidirectional image. The produced viewport is subsequently compressed with JPEG and

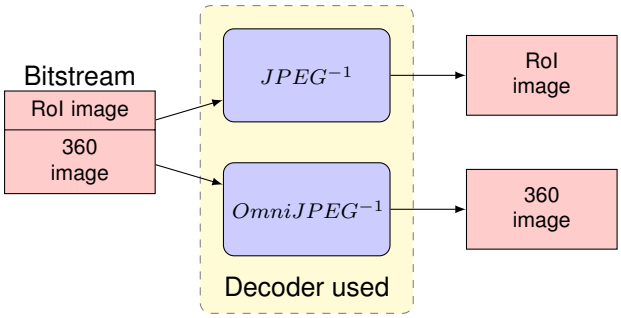


Figure 6.3 – Visualization of an omnidirectional content encoded in OmniJPEG format.

stored in a JFIF file. The original omnidirectional image is stored in the APP11 application marker in JPEG format [126].

The developed application requires the following parameters:

- yaw, as a horizontal spherical coordinate in degrees,
- pitch, as a vertical spherical coordinate in degrees,
- vertical field of view in degrees, and
- aspect ratio of the desired viewport.

The implemented prototype uses the open-source library LibJPEG to cope with JFIF file format and JPEG compression. Additionally, a graphical user interface has been developed to help in the process of OmniJPEG encoding.

To apply the geometrical transformation and to produce an undistorted viewport for the specified RoI of an omnidirectional image, we use the pinhole camera model. Omnidirectional image represented in equirectangular format is mapped to a sphere first. Then, perspective projection transformation, which removes the geometric distortions, is applied either to a manually defined or to an automatically extracted RoI. Examples of RoI areas in equirectangular projection and their viewport counterparts extracted with the developed prototype are shown in Figure 6.4. The desired RoI, highlighted in green (see Figure 6.4), is processed, all geometric distortion related to sphere-to-plane projection are removed, and the corresponding viewport representing the omnidirectional image is produced.

6.2.3 Performance evaluation

Table 6.1 presents results of the proposed backward compatible coding scheme for omnidirectional visual content. For five different contents in their original resolution , we calculate the

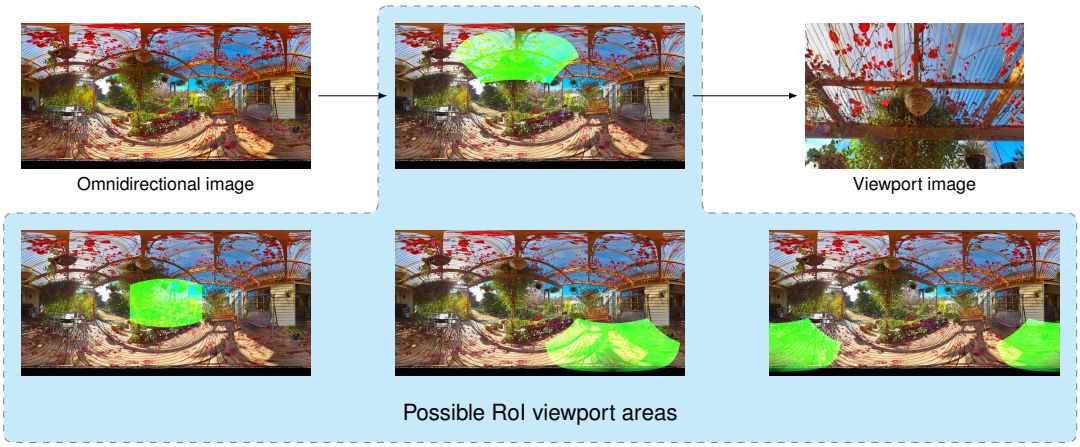


Figure 6.4 – Examples of visualization of regions of interest in equirectangular projection and its counterpart viewport.

file size overhead after encoding with OmniJPEG. Three different relative viewport sizes with a vertical and horizontal field of view, corresponding to 30×60, 60×90, and 90×120, are used. As expected, file size is increasing with increasing size of the desired viewport. Another factor affecting the file size overhead is the actual content in the viewport, because, for example, the amount of details in the picture can decrease JPEG compression efficiency. Results show an average overhead of approximately 8% for the viewport size of 60×90.

Table 6.1 – OmniJPEG performance for tested dataset. Overview of file size overhead for different sizes of extracted RoI.

Content		*C1*	*C2*	*C3*	*C4*	*C5*
file size [MB]	original file	6.9	5.6	4.4	4.5	0.3
	30x60	7.1	6.0	4.5	4.6	0.3
	60x90	7.4	6.6	4.6	4.7	0.3
	90x120	7.9	7.5	4.8	4.8	0.3
file size overhead [%]	30x60	3.2	5.7	1.6	1.5	3.4
	60x90	7.2	17.9	4.8	3.7	6.9
	90x120	14.2	33.6	9.8	6.6	10.3
resolution WxH [px]	original file	6000×3000	6512×3256	4000×2000	4096×2048	2048×1024
	30x60	1000×500	1085×543	667×333	683×341	341×171
	60x90	1500×1000	1628×1085	1000×667	1024×683	512×341
	90x120	2250×1500	2442×1628	1500×1000	1536×1024	768×512

7 Applications and Extensions of Omnidirectional Imaging

In this chapter, we explore certain related extensions and applications in omnidirectional imaging. At first, we look into privacy protection which is yet another field drawing more attention with the advances in image processing, visual and social media. We present a method for protecting user privacy in omnidirectional media, by removing parts of the content selected by the user, in a reversible manner.

In the second part of this chapter, we investigate a possible extension to 3+ degrees of freedom by considering an individual case of rendering narrow baseline light field images with limited translational interactions. We provide also results of an extensive analysis of those iterations, including: circular histograms of directions of user movements, average vectors for a next perspective view, and charts of time spent on a view.

7.1 Privacy and perceptual visual quality

Privacy protection is drawing more attention with the advances in image processing, visual and social media. Photo sharing is a substantive contemporary activity, which also brings

the concern of regulating permissions associated with shared content. We present here a method for protecting user privacy in omnidirectional media, by removing parts of the content selected by the user, in a reversible manner. Object removal is carried out using three different state-of-the-art inpainting methods, employed over the mask drawn in the viewport domain so that the geometric distortions are minimized. The perceived quality of the scene is assessed via subjective tests, comparing the proposed method against inpainting employed directly on the equirectangular image. Results on distinct contents indicate that our object removal methodology on the viewport enhances perceived quality, and thereby improves privacy protection as the user is able to hide objects with less distortion in the overall image.

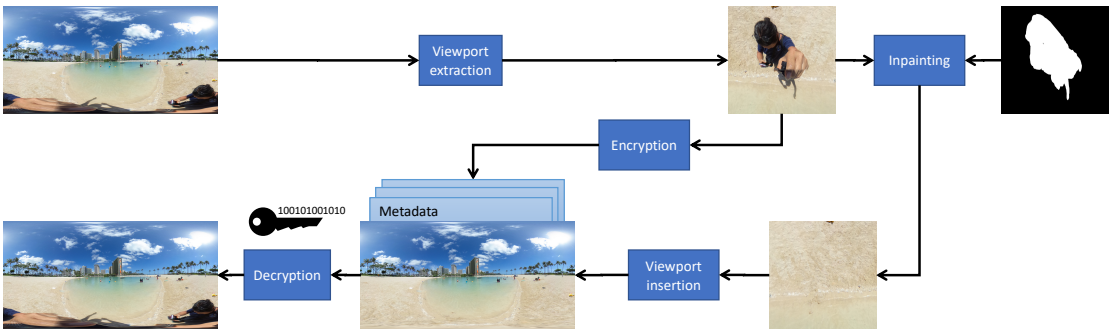


Figure 7.1 – Viewport extraction method for object removal using inpainting

7.1.1 Related work

With the advent of smart mobile devices and social networks, photo sharing has become an easy and widespread activity among users. The increasing distribution of images also raises issues on privacy protection and creates the need for adjusting permissions, as the shared content contains sensitive information concerning users. Access control over contents provide exclusive rights to only selected correspondents, thereby enhancing user security and privacy. A widely preferred form of privacy protection is to obfuscate parts of images, instead of encrypting or permuting the whole image [126]. This results in less visual distortions that are confined only to the specific area of interest. The obfuscation can be achieved using a variety of image processing techniques, such as blurring, mosaicking, censoring and object removal. Among these methods, the first three have to introduce a high amount of distortion to hinder the underlying content, whereas object removal provides more natural viewing conditions while still being able to protect the content. This process can also be made reversible, as in work [126]. Therefore, the access to the original data can be granted to selected users with permissions.

We present an object removal methodology via inpainting on omnidirectional images performed on the selected viewport instead of the equirectangular representation, which yields visually plausible results. Given an omnidirectional image, we extract the viewport and ap-

ply the mask defining the objects to be removed on the viewport. We remove objects using three distinct state-of-the-art inpainting algorithms [26, 63, 123]. Inpainting on the viewport rather than on the equirectangular image minimizes the geometric distortions and limits the source region to more relevant components within the content. After removing an object and inpainting the background, we project the viewport back onto the equirectangular image. We, furthermore, the assess quality of the protected content by performing subjective evaluations, where we compare the images inpainted using our method and in the equirectangular domain directly, using Absolute Category Rating (ACR) [1].

Image inpainting algorithms can be divided into four general classes: statistical methods, partial differential equation (PDE)-based methods, exemplar-based methods and deep generative models based on convolutional neural networks [123, 6]. Statistical methods make use of parametric models to describe the input textures. However, they fail in the presence of additional intensity gradients [69]. PDE-based methods propagate information from the known part of the image [10, 107, 40] using smoothness priors, which introduces blurring when large and high frequency regions needs to be inpainted. Exemplar-based methods and deep generative models are most widely used, where the former fill the holes in the image using exemplars from local or global search regions [26, 63, 67, 49, 19] and the latter exploit semantics learned from large scale datasets [119, 55, 121]. We have selected two robust methods for exemplar-based inpainting [26, 63] as well as one semantic learning-based state-of-the-art method [122] in order to reduce the bias of the preferred inpainting technique on our object removal strategy on the viewport.

While most works on object removal and inpainting focus on planar images, panoramic content is considered in [77]. A field-of-view expansion method using retargeting techniques combined with Graphcut Textures is proposed to remove objects near the equator, and extended to farther portions of the sphere by tripod rotations. Although objects can be removed regardless of their locations, rotation of the full equirectangular image is more costly than viewport extraction. Our approach minimizes the geometric distortions within a limited search region, thereby reducing the complexity of inpainting simultaneously.

7.1.2 Viewport extraction method

In this section, we present a method to perform object removal in omnidirectional images using inpainting.

Viewport extraction

A viewport is a part of an omnidirectional image which is observed by a user at one moment. During the process of rendering a viewport is extracted from an equirectangular representation and shown to the user. Unlike in the back-end equirectangular representation, the geometrical distortion in the viewport image is negligible.

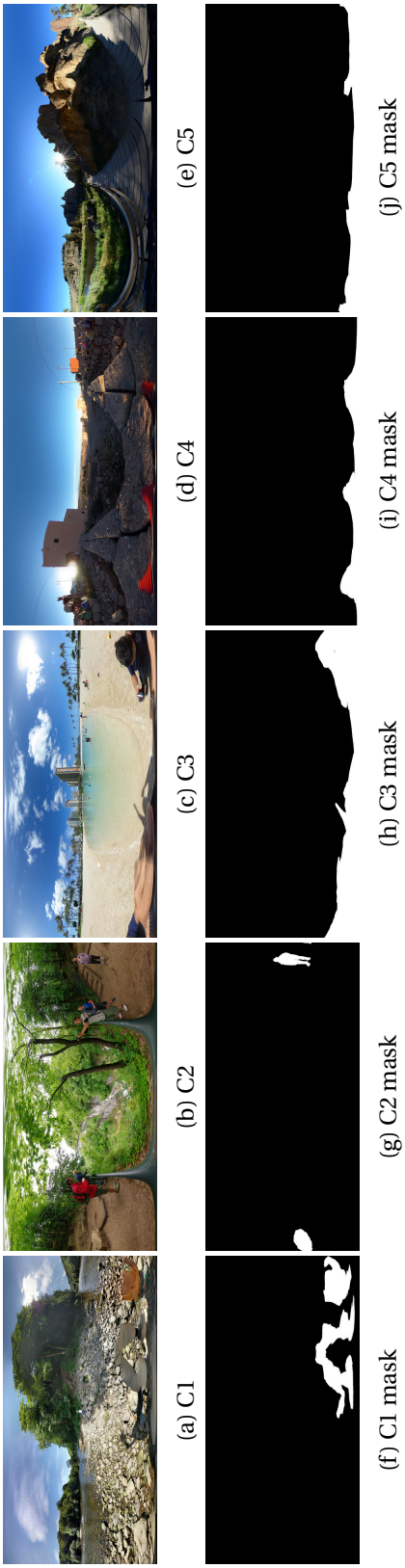


Figure 7.2 – Contents selected for experiments (top row) and masks for object removal (bottom). The objects to be removed are depicted with white color on masks.

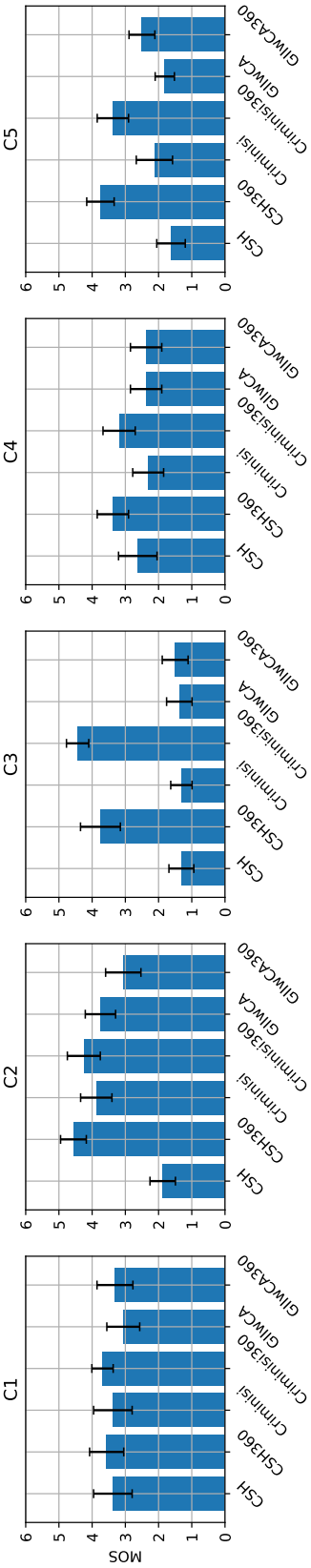


Figure 7.3 – Mean opinion scores with 95% confidence intervals.

The block-diagram in Figure 7.1 describes the method of viewport extraction for object removal in omnidirectional images. Here we apply inpainting algorithms in the viewport domain such as it would be performed on common planar images. Afterwards, the viewport with the inpainted area is inserted back to the equirectangular image. In order to make the process of object removal reversible, one can keep the original viewport and store it as metadata in the image file, similarly to what is done in [92]. The original viewport which contains information critical for privacy protection can be also encrypted.

7.1.3 Inpainting experiment

This section contains a step-by-step description of the removal of objects on images in order to further assess the performance of the proposed method.

Dataset

Five distinct omnidirectional contents were selected amongst publicly available photographic works licensed with Creative Commons[1]. Masks for object removal were created manually, as depicted in figure 7.2. Original contents and prepared masks can be retrieved from a GIT repository in [2]. Each content was downsampled to 2048x1024 resolution and the resolutions of masks were identical. All contents were natural images with outdoor views, where the location of objects to be removed varies from the equator to the south pole of the scene.

Inpainting algorithms

Abbreviation	Base algorithm name	Viewport	Source code
CSH	Coherency sensitive hashing	No	http://github.com/PetterS/patch-inpainting [Commit: 03cc575]
CSH360	Coherency sensitive hashing	Yes	- -
Criminisi	Exemplar-Based Image Inpainting	No	http://github.com/cheind/inpaint [Commit: 864128c]
Criminisi360	Exemplar-Based Image Inpainting	Yes	- -
GIIwCA	Generative Image Inpainting w/ Contextual Attention	No	http://github.com/JiahuiYu/generative_inpainting [Commit: 6bfaa20]
GIIwCA360	Generative Image Inpainting w/ Contextual Attention	Yes	- -

Table 7.1 – Inpainting algorithms used in the experiments

We used three state-of-the-art algorithms to perform inpainting in the viewport domain. Table 7.1 lists the algorithms selected to be enhanced with viewport extraction.

[1] https://creativecommons.org/
[2] https://github.com/mmspg/inpainting360

Content	Viewport position	FoV°	Size
C1	yaw: 30°, pitch: -80°	90×90	1024×1024 px
C2	yaw: 180°, pitch: 0°	90×90	1024×1024 px
C3	yaw: 180°, pitch: -80°	90×90	1024×1024 px
C4	yaw: 0°, pitch: -90°	90×90	1024×1024 px
C5	yaw: 0°, pitch: -90°	90×90	1024×1024 px

Table 7.2 – Viewport positions for inpainted contents

Inpainting procedure

Table 7.2 details the viewport parameters selected for each content from the dataset described in Section 7.1.3. All the viewports have the same size and field of view. The positions were selected in such a way that an object to be removed appears approximately in the canter of the viewport and the whole inpainted area fits inside.

For each omnidirectional image represented in equirectangular projection the viewport was extracted according to the parameters specified in Table 7.2. Then inpainting was applied in the viewport domain in the areas depicted in Figure 7.2 using the algorithms CSH, Criminisi, and GIIwCA (Table 7.1). The inpainted viewport was then inserted back in the equirectangular image. Further here, these methods are called CSH360, Criminisi360, and GIIwCA360, respectively.

In order to have a reference to assess viewport extraction enhancement, we also applied inpainting directly in the equirectangular domain for all the images from the dataset using exactly the same masks.

7.1.4 Subjective evaluation

Test methodology

The Absolute Category Rating (ACR) method described in [1] was chosen to assess the effect of the proposed viewport extraction method for object removal on the quality of selected contents. ACR is a single stimulus evaluation where the test stimuli are randomly presented to subjects and voting is performed after each viewing. Overall quality is assessed using a five-grade scale with the following levels: ”5 - Excellent”, ”4 - Good”, ”3 - Fair”, ”2 - Poor”, and ”1 - Bad”.

A total of 16 consenting subjects participated in the study, of which 7 were female, with an overall median age of 26.5. All subjects had passed color vision and visual acuity tests prior to experiments, using Ishihara and Snellen charts, respectively. The subjects were placed in an immersive environment; and stimuli were presented to them using a head mounted display (HMD) composed of a VR head-mount with buttons[3] and a mobile device installed inside as a screen. Samsung Galaxy S7 edge SM-G935F with a screen resolution of 2560x1440 pixels was

[3] https://mergevr.com

used to display the images. Subjects were sitting on a rotatable chair during the assessment. Immersive textual instructions were provided inside the VR along with a verbal guidance by the experimenter, as described in [112].

Each experiment started with an immersive training, where original contents were presented to the subjects with a red circle indicating the position of the objects to be removed later in the test session. This way, subjects were familiarized with the contents and surroundings of the objects to be removed. The circles were omitted during test session in order not to disturb the natural viewing of the stimuli. During evaluation, subjects assessed the stimuli shown to them consequently without any time restrictions. When ready to rate an image subjects had to activate a 3D immersive voting menu by pressing a button and select the grade proceeding immediately to the next image. All stimuli were automatically randomized in each session. The experiments were conducted using a testbed for subjective evaluation of omnidirectional visual content proposed in [111]. This software was developed for Android and is publicly available for downloading[4].

Data processing

Outlier detection was performed separately on the raw scores using a method described in [2]. None of the subjects were identified as outliers during our experiments. The mean opinion score (MOS) and 95% confidence intervals (CIs) assuming a Student's t-distribution of the scores were computed for each test condition [29].

7.1.5 Results and discussion

Figure 7.3 presents the results of the subjective evaluation described in Section 7.1.4. Bar plots with 95% confidence intervals show how different methods perform on different contents. The plots are grouped by content. As one can see, viewport extraction significantly enhances inpainting. CSH360 performs better than CSH on four contents, Criminisi360 is better on three contents, and GIIwCA360 is superior to GIIwCA on one content.

For the contents C3-C5 viewport extraction brings higher quality gain than for C1-C2. The fact that on C2 the quality does not improve in two cases out of three can be possibly explained by the position of inpainted area lying near equator where geometrical distortion is minimal hence its compensation is not needed.

We presented a method for reversible object removal in omnidirectional images, which is targeted for privacy protection in immersive media. We show by performing subjective quality evaluation involving 16 naive subjects that viewport extraction can enhance the performance of state-of-the-art inpainting algorithms in omnidirectional images.

[4] https://github.com/mmspg/testbed360-android

7.2 Towards 3+ degrees of freedom extension

Here, we explore a possibility of extending omnidirectional imaging to 3+ degrees of freedom by considering an individual case of rendering narrow baseline light field images with limited translational interactions. We provide, afterwards, results of extensive analysis of those iterations, including: circular histograms of directions of user movements, average vectors for a next perspective view, and charts of time spent on a perspective view.

Light field is yet another type of immersive multimedia content which describes the amount and the dicertion of light passing trough each point of picture. In practice, this allows refocusing, changes of depth of field and limited change of perspective during consumption of such content.

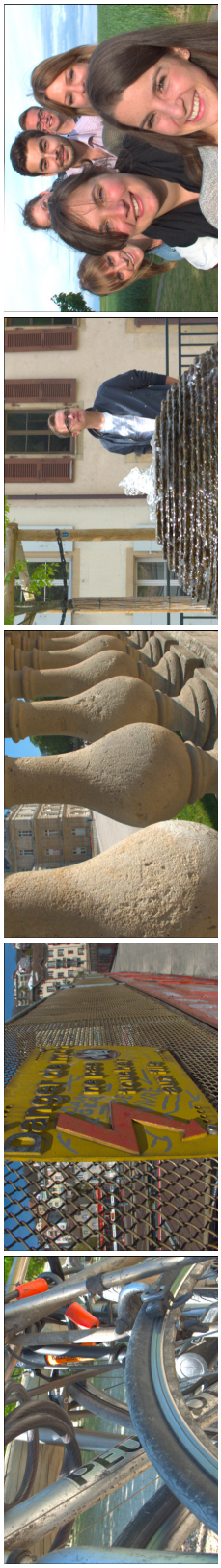
Light field imaging is associated with augmented reality (AR) more often than it is with virtual reality (VR). Nonetheless, it is becoming a very desirable type of content for VR, along with omnidirectional imaging and point clouds which are naturally designed for this type of media.

7.2.1 Related work

When compared to traditional omnidirectional content, light field rendering allows for more realistic visualization of 3D spaces in a virtual or augmented reality scenario, thanks to the full parallax environment that can provide depth and focus cues. In recent years, several wearable light field display prototypes have been designed and proposed by both academics [53, 79, 74] and industry [66, 54]. The development of commercial devices such as the Avegant light field display headset[5] or Magic Leap Digital Lightfield[6] promises near-eye light field head-mounted displays (HMD) to be available to consumers in the near future. However, as those devices are currently either in prototype state or too expensive to be widely accessible to the public, off-the-shelf solutions are needed to perform quality assessment of light field content in a virtual reality scenario. In particular, if already available devices are used to perform quality assessment, crowd-sourcing can be employed to collect a large number of scores with reduced costs in terms of time and expenses [95].

Several studies of quality assessment for light field images can be found in the literature. Paudyal et al. investigate the impact of watermarking on visual quality of light fields using Absolute Category Rating (ACR), and in particular on the relationship between watermark strength and visual quality [86]. Darukumalli et al. and Kara et al. examine the quality of experience using light field displays, and their relationship with angular resolution and zooming levels [28, 60]. Viola et al. evaluate a compression solution through subjective quality assessment using passive and interactive methodologies on conventional 2D displays [115]. They also perform a statistical comparison between the two methodologies to determine the impact of interaction on the results [116]. Konrad et al. evaluate the quality of experience

[5] https://www.avegant.com/
[6] https://www.magicleap.com/



(a) I01 (b) I02 (c) I04 (d) I09 (e) I10

Figure 7.4 – Stimuli used in the pilot experiment

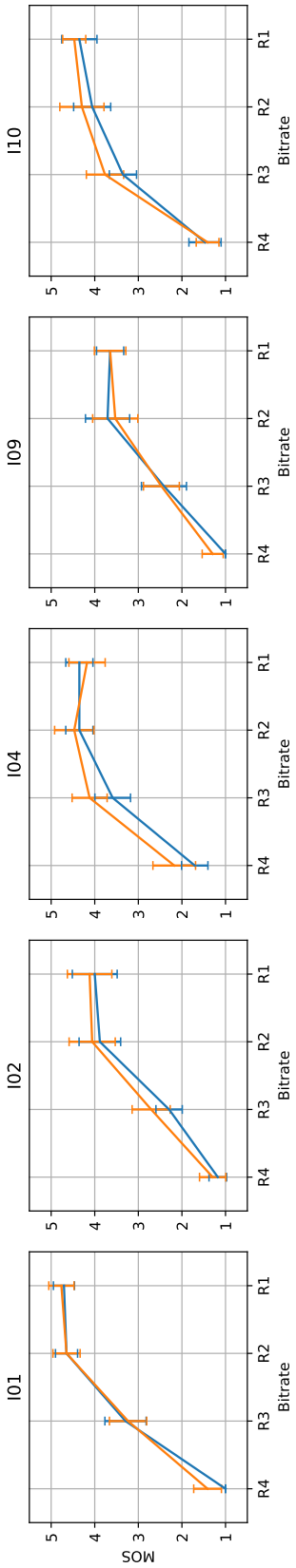


Figure 7.5 – Mean opinion scores (MOS) plotted against bit-rates and their 95% confidence intervals (CI). A blue solid line depicts HEVC and an orange solid line depicts VP9 encoded stimuli.

related to several focus-tunable near-eye display modes, as well as the effect of the display mode on the user performance [62]. However, to the best of the author's knowledge, no quality assessment of compression artifacts for light field images has been performed on HMD.

7.2.2 Rendering light field in VR

In this section, we propose a software solution to render narrow baseline light filed images in virtual reality implemented using WebGL. We describe its architecture, main components and features.

The proposed VR rendering solution allows visualizing light field images on mobile HMD platforms, such as Google Cardboard or Samsung GearVR, desktop computers, and head-mounted displays, such as HTC Vive and Oculus Rift. The portability is achieved by using a web-based platform. The types of implemented interactions include horizontal and vertical narrow baseline perspective changes. In the VR environment viewers interact with movements of a head, whilst on personal computer a mouse or a trackpad can be used.

The rendering can be performed in any web-browser which supports WebGL standard in its version 1.0, including all mobile devices supporting OpenGL ES 2.0. The source code is written in JavaScript language and requires Three.JS 3D graphics library. Light field images are rendered as separate perspective views which are changed in real time according to the data from motion sensors of a device in the mobile and HMD case, and according to mouse or trackpad movements in the desktop computer case. In order to display a light field image, all its perspective views are downloaded from the server as texture files. Then those textures are loaded into GPU memory. The application tracks user movements and renders the texture corresponding to the current perspective view image.

Besides the rendering *per se*, the application provides additional features allowing one to use the software to conduct subjective quality evaluation experiments for narrow baseline light field assessment. This includes a storyboard implementing training and evaluation scenarios, an ability to assign a score to a light field image in an immersive way within VR, and store resulting assessment data on a server.

The storyboard currently includes a training session followed by an evaluation session. An absolute category rating (ACR) methodology [1] is implemented to collect subjects' votes and send them to the server after evaluation is completed. During the evaluation process subjects use a 3D menu for voting without leaving the immersive VR environment.

An important direction of research in immersive imaging and in particular in light field and omnidirectional imaging is the analysis of user interactions. The proposed software tracks how users interact with narrow baseline light field content. Every time a subject moves from one perspective view to another, it is recorded and sent to the server.

In order to deploy the developed software, one needs an HTTP-server supporting PHP server-

side scripting. The latter is required for storing results on the server. Once the software is on-line it can be accessed by multiple users simultaneously over the Internet. Assuming the high availability of affordable consumer HMD, this can allow for large scale crowd-sourcing subjective quality evaluations with interaction analysis.

The source code of the developed software for rendering light field in VR and a demo are publicly available on-line at `https://mmspg.github.io/lightfieldvrtb/`.

7.2.3 Pilot experiment

This section describes a pilot experiment on subjective quality evaluation in VR environment conducted in Multimedia Signal Processing Group laboratory at EPFL with the purpose to validate the solution proposed in Section 7.2.2.

Population and environment

The experiment was performed with 17 subjects, of which 9 were males and 8 were females. The age of the subjects ranged from 18 to 37 years old, with the average equal 25.38 and the median equal 26.73. Prior to the experiment all the subjects were tested for their visual acuity and color vision.

Equipment

To render narrow baseline light field images in VR, experimenters used the software solution described in Section 7.2.2. Subjects were wearing a Google Cardboard compatible HMD-mount for mobile devices (MergeVR[7]) with a Samsung Galaxy S7 Edge smartphone installed inside. The resolution of the device was 2560 × 1440 pixels or 1280 × 1440 pixels per eye. The pixel density was 534 pixels per inch. The field of view of this HMD-mount was 96 degrees. It had 42 mm lenses and allowed for adjustment of interpupillary distance.

Stimulus set

The set of stimuli used in the pilot subjective quality evaluation experiment is based on the light field image data set created by Rerabek et al. [91]. Five light field images have been selected to represent different categories. Figure 7.4 shows central-view all-in-focus thumbnails of the unimpaired stimuli in 2D representations.

Light field images have been compressed using two different codecs adapted in such a way that they process the perspective views as a pseudo-temporal sequence in a serpentine order. Before being fed to encoders, all the perspective images were padded with black pixels, color-space was converted to YUV and re-sampled to 422 with 10-bit depth.

[7] https://mergevr.com/goggles

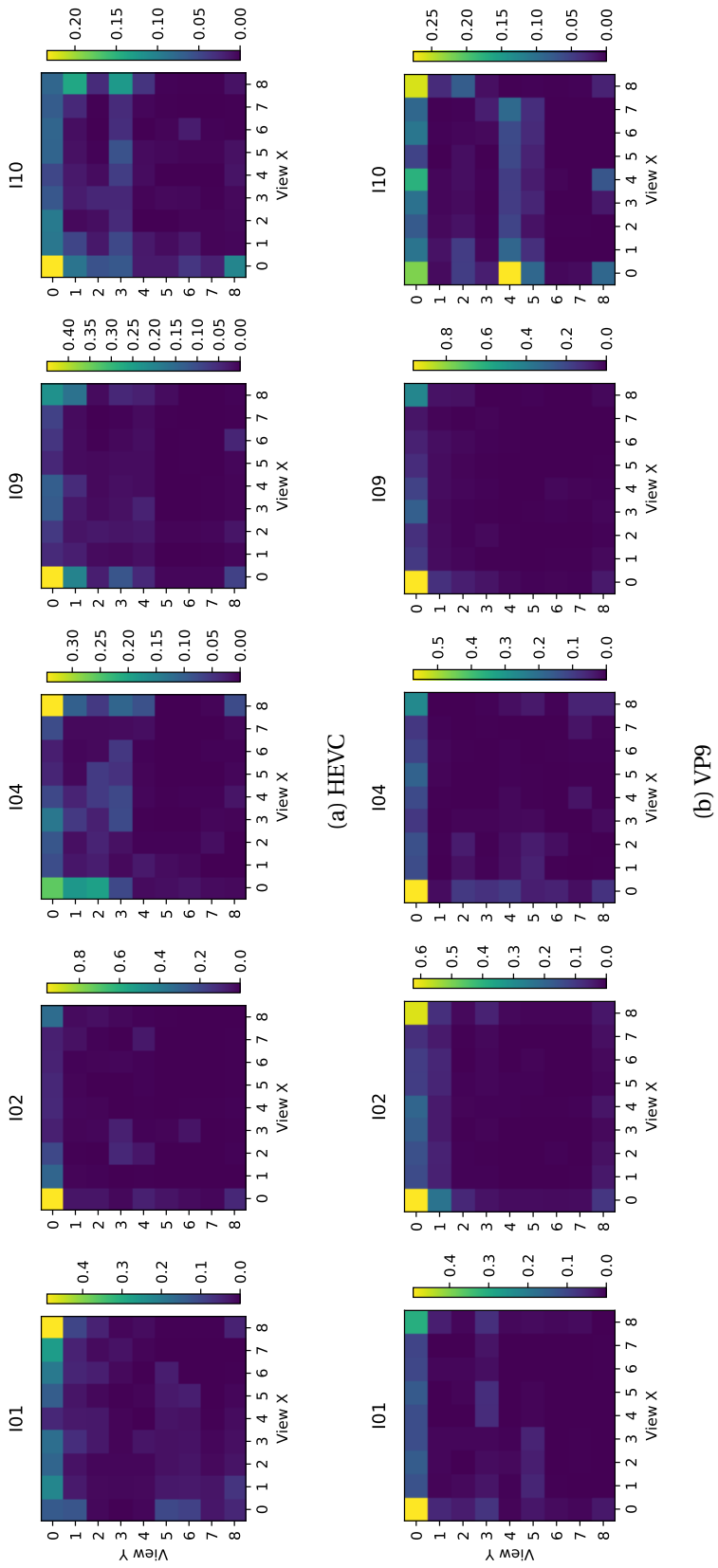


Figure 7.6 – Charts of time spent on a view in seconds

Table 7.3 – QP values selected to encode contents with HEVC.

Content	R1	R2	R3	R4
I01 (Bikes)	13	24	33	44
I02 (Danger de mort)	15	26	35	43
I04 (Stone pillars)	14	23	30	40
I09 (Fountain)	14	24	32	43
I10 (Friends)	12	21	29	40

Table 7.4 – Settings for VP9 coder.

--i422 --input-bit-depth=10 --profile=3 -w < *Width* > -h < *Height* >
--target-bitrate=< *bitrate*> --cq-level=0 --bit-depth=10 --codec=vp9
--fps=30000/1000 --best -o < *Output* > < *Input* >

The codec number one (P1) was the HEVC Main10 profile. The x265[8] library was used to perform the compression. The quantization parameters (QP) were set to match the preselected compression ratios. In the Table 7.3, one can find the exact values of different QP used in the test.

The codec number two (P2) was the VP9[9]. The full command line used to produce compressed stimuli can be found in Table 7.4. The target bitrate was chosen to match the corresponding compression ratios as defined below.

The codecs were evaluated on four bitrates, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, $R4 = 0.005$ bpp. The compression ratios were computed as ratios between the size of the uncompressed raw images in 10-bit precision ($5368 \times 7728 \times 10$ bits = 414839040 bits = 10 bpp) and the size of the compressed bitstream.

Methodology

The subjective quality evaluation experiment was designed to follow a single stimuli Absolute Category Rating (ACR) method [1]. ACR is a single stimulus evaluation where stimuli are presented subsequently to subjects, and voting is performed after each viewing. Images are assessed using five-grade quality scale with the following levels: ”5 - Excellent”, ”4 - Good”, ”3 - Fair”, ”2 - Poor”, and ”1 - Bad”.

Due to distortions naturally occurring in lenslet-based light field content, the border perspective views were deemed not suitable for visualization, since they would negatively bias subjects. Hence, only the central 9×9 perspective views out of the 15×15 views were selected

[8] https://www.videolan.org/developers/x265.html
[9] https://www.webmproject.org/vp9/

for the test. The contents were converted from *PPM* file format in 10 bits to *PNG* file format in 8 bits, due to limitations of the display.

7.2.4 Results and discussion

This section describes the analysis of experimental data and presents its results including mean opinion scores and interaction vectors, including mean opinion scores for every stimulus, average time spent on a perspective view per content per codec, circular histograms of interaction vectors per content per codec, and average interaction vectors for a perspective view per content per codec.

Raw experimental data obtained from the evaluations consists of ACR scores given by each subject for each stimulus and iteration records for each subject grouped by stimulus. The latter contains time stamps for every change of a perspective view initiated by a subject. Before proceeding to further analysis the data was screened for outlier subjects using the method described in [2]. Zero outliers were detected among the subjects.

Mean opinion scores

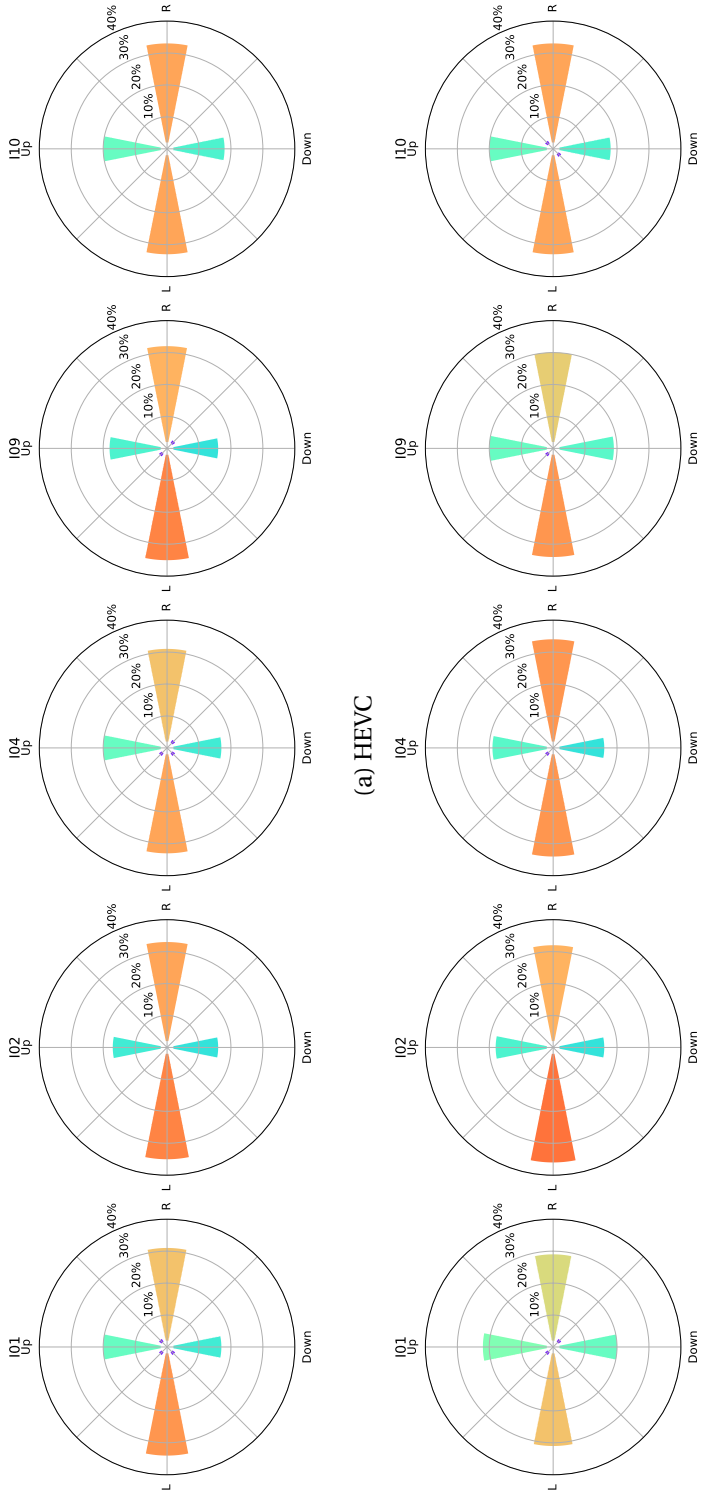
Figure 7.5 presents subjective mean opinion scores (MOS) for five contents compressed with two different codecs at four different bitrates selected as described in Section 7.2.3. 95% confidence intervals were computed for each MOS assuming T-Student's distribution of subjective scores for a stimulus. The MOS values are plotted against the bitrates R4-R1 where R4 is the lowest bitrate and R1 is the highest bitrate. The exact values can be found in Subsection 7.2.3.

Interaction analysis

Interaction analysis includes two main parts: temporal analysis and spatial vector analysis. The former concerns the time spent looking at each perspective view of a narrow baseline light field image. The latter computes interaction vectors of changes between the perspective views shown to a user.

In order to compute the time spent on each view, the difference between the time stamps of two subsequently shown perspective views was taken. All the time stamps were grouped by stimulus. The first and the last time stamps in each group were dropped. Then the average time spent on a view was computed. Stimuli compressed with all the bitrates were counted for average. Figure 7.6 depicts maps of average time spent on a view for each content for two codecs.

Interaction vectors were computed in the following way. The time stamps for every change of a perspective view initiated by a user were grouped by stimulus and subject. The first and



(a) HEVC

(b) VP9

Figure 7.7 – Circular histograms of the directions of user interactions

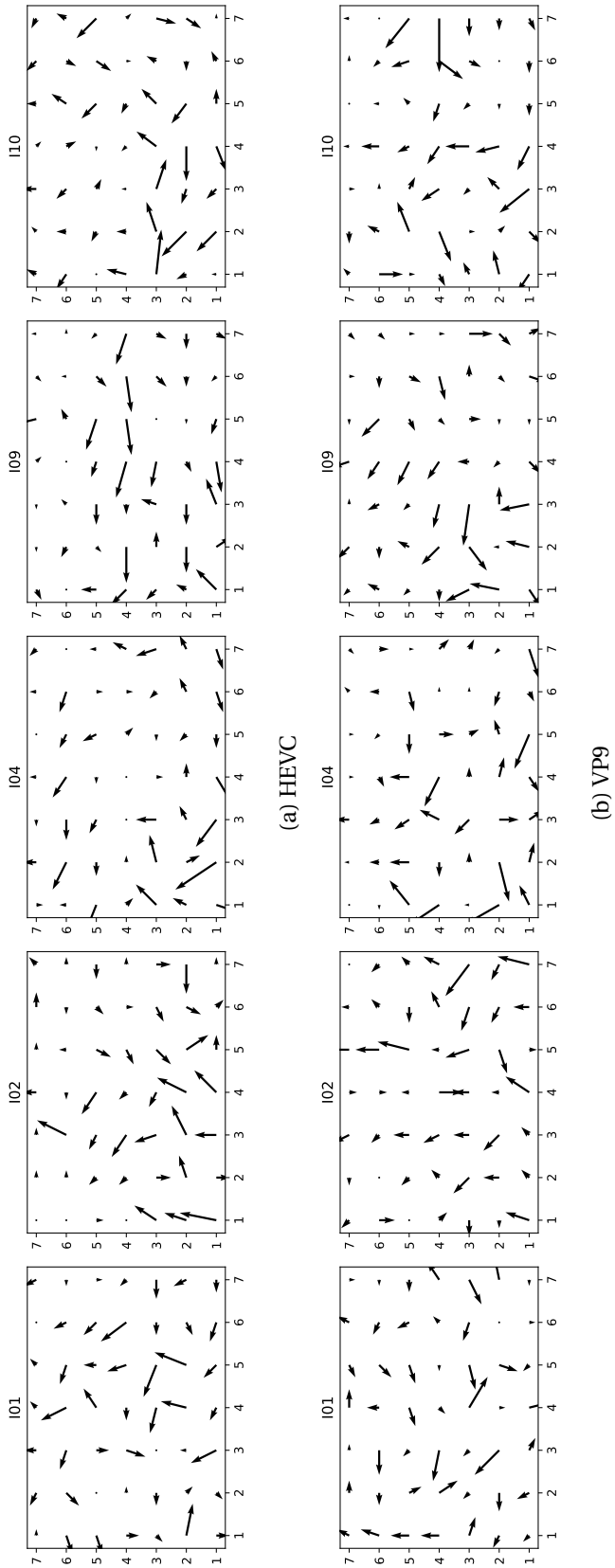


Figure 7.8 – Average interaction vectors for each perspective view per content per codec

the last time stamps in a group were dropped. An interaction vector was computed per each time stamp record as a difference between corresponding x and y coordinate pairs of the subsequent view and the current view compensating the reverse Y-axis direction (i.e. from top to down) in perspective view coordinate system. Figure 7.7 shows circular histograms of interaction vectors per content per codec.

In order to analyze a typical path from each perspective view to a subsequent one, we have computed an average interaction vector for each view. This average was taken among all subjects and all bitrates grouped by content and by codec. One can find the vector field plots depicting average interaction vectors for each view in Figure 7.8.

Discussion

Mean opinion scores in Figure 7.5 show no statistically significant difference in visual quality for the codecs except in bitrate R3 for content I04 where VP9 outperforms HEVC.

From the average time spent on a view in Figure 7.6 one can see that subjects were systematically biased towards spending more time on the top row. This can be possibly explained by a wrong vertical position of light field images in the VR space, which was exactly in front of the camera. In order to compensate this bias in future experiments one should consider placing images higher than the camera. Furthermore, the average time spent on a view can be used in the future to compute weighted subjective scores for each view.

Circular histograms of the interaction vectors in Figure 7.7 show clearly that subjects tend to interact more horizontally than vertically. One can also notice correlation between the histograms for the same content with respect to different codecs.

Average interaction vectors for each perspective view presented in Figure 7.8 can be used as ground truth data for estimation of the most probable subsequent view. This is required to develop efficient compression algorithms providing fast random access to perspective views of a light field image.

In conclusion, we propose a solution for rendering narrow baseline light filed images in a VR environment which allows interactions with their perspectives. The developed software includes features to perform subjective evaluations and tracks users interactions. A pilot subjective quality evaluation experiment for light field in VR was conducted with 17 subjects participating in the assessments. The results of the pilot experiment have been presented, including MOS and interaction analysis for 5 light field images compressed with two different codecs.

8 Conclusion

In this dissertation, we have studied visual attention and perceptual visual quality in omnidirectional imaging. We thoroughly investigated the following topics in the field: visual attention in head-mounted virtual reality, subjective assessment of perceptual visual quality, and objective measurement of perceptual visual quality. In addition to the above, we have also looked into other specific problems in omnidirectional imaging: namely, backwards compatible coding, privacy applications, and extension to 3+ degrees of freedom.

8.1 Accomplishments

First of all, we established a new direction in the topic of visual attention analysis by using in our method the data obtained from head-direction only trajectories. By computing angular speed of rotational movements we were able to estimate in which regions viewers fixate their attention. Afterwards, by applying particular filters to one- and two-dimensional signals, we obtained pixel-based saliency maps. This approach has many applications, thanks to the fact that every rendering device can record such head-direction trajectories, allowing massive data collection from consenting consumers.

We have made contributions to the methodology for subjective evaluation of perceptual visual quality of omnidirectional images in head-mounted virtual reality environment by proposing a testbed for subjective quality evaluation of omnidirectional visual content.The testbed allows researchers to perform experiments using different methods for subjective quality evaluations. Experimental data that can be obtained with this testbed includes subjective mean opinion scores, time spent on stimulus, and view direction tracks. The software implementation is publicly available as open source under the GNU license. Furthermore, with the proposed testbed, we conducted multiple subjective evaluation experiments and used the obtained data for benchmarking a number objective metrics specifically designed for omnidirectional visual content.

In objective measurement of perceptual visual quality, we designed a novel metric which

incorporates experimental visual attention information; and we showed that the performance improves comparing to the same base measurement without visual attention weighting. The proposed metric has a particularly important application in on-demand streaming of omnidirectional video, where the first critical number of viewers contribute to the statistics of visual attention the sources can be re-encoded for better quality or bandwidth performance.

As contributions to specific related problems, we achieved the following: We developed OmniJPEG, a JPEG backward compatible solution to encode the omnidirectional images. In order to ensure the JPEG backward compatibility, OmniJPEG extracts predefined regions of interest from omnidirectional images, as well as properties of equirectangular projection, while at the same time also keeping complete equirectangular information to preserve the capability of correctly rendering an omnidirectional image with appropriate devices and software. We presented a method for reversible object removal in omnidirectional images, which is targeted for privacy protection in immersive media and showed by performing subjective quality evaluation involving 16 naive subjects that viewport extraction can enhance the performance of state-of-the-art inpainting algorithms in omnidirectional images. We explored a possibility of extending omnidirectional imaging to 3+ degrees of freedom by considering an individual case of rendering narrow baseline light filed images with limited translational interactions and provided results of extensive analysis of those iterations, including: circular histograms of directions of user movements, average vectors for a next perspective view, and charts of time spent on a perspective view.

8.2 Future directions

In this section, we discuss possible future directions to continue the work presented in this dissertation. We focus mainly on the topics of visual attention and objective measurement of perceptual visual quality.

We approached the problem of visual attention in omnidirectional imaging by considering only head rotational movements of a viewer. The rationale behind this approach was the simplicity of obtaining these data in a large scale due to the fact that every rendering device can provide such information. Future work may consider also statistics of eye movements with respect to the viewport and its relation to the rotational position of the head. Research in this direction may improve the accuracy in estimation of visual attention.

Further incorporating of visual attention approaches in subjective and objective evaluation of perceptual visual quality of omnidirectional visual content may potentially bring additional improvements to the state of the art. One possible scenario would be to indirectly estimate perceptual visual quality based on user interactions with the content. This may lead also to integration of visual attention predictors into methods of objective quality measurement.

In the present dissertation, we did not explore deep learning approaches for compressing omnidirectional visual content. The variety of recent findings in the field of artificial neural

networks are remarkably promising; and applying such approaches, for example, autoencoders, to the specific domain of omnidirectional imaging may result in improvements of current compression methods for this type of multimedia.

Bibliography

[1] ITU-T Rec. P.910 Subjective video quality assessment methods for multimedia applications, April 2008.

[2] ITU-R Rec. BT.500-13 Methodology for the subjective assessment of the quality of television pictures, 2012.

[3] Next-generation video encoding techniques for 360 video and VR. Facebook Engineering. Web resource: https://engineering.fb.com/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/, January 2016.

[4] A. De Abreu, C. Ozcinar, and A. Smolic. Look around you: Saliency maps for omnidirectional images in VR applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017.

[5] SN Akula, A Singh, A Dsouza, RN Gadde, et al. Ahg8: Efficient frame packing for icosahedral projection. *JVET-E0029*, 2017.

[6] Pinar Akyazi and Pascal Frossard. Graph-based inpainting of disocclusion holes for zooming in 3d scenes. In *Proceedings of EUSIPCO*, number CONF, 2018.

[7] S. Baker and S. K. Nayar. A theory of catadioptric image formation. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 35–42, January 1998.

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, Lecture Notes in Computer Science, pages 404–417. Springer Berlin Heidelberg, 2006.

[9] Ryad Benosman and Sing Bing Kang. *Panoramic Vision: Sensors, Theory, and Applications*. Springer Science & Business Media, May 2001.

[10] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[11] I. Bogdanova, A. Bur, and H. Hugli. Visual Attention on the Sphere. *IEEE Transactions on Image Processing*, 17(11):2000–2014, November 2008.

[12] Iva Bogdanova, Alexandre Bur, Heinz Hügli, and Pierre-André Farine. Dynamic visual attention on the sphere. *Computer Vision and Image Understanding*, 114(1):100–110, January 2010.

[13] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient Object Detection: A Benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[14] A. Borji and L. Itti. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[15] Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv:1810.03716 [cs]*, 2018.

[16] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[17] A. M. Bruckstein and T. J. Richardson. Omniview cameras with curved surface mirrors. In *Proceedings IEEE Workshop on Omnidirectional Vision (Cat. No.PR00704)*, pages 79–84, June 2000.

[18] Peter J. Burt and Edward H. Adelson. A Multiresolution Spline with Application to Image Mosaics. *ACM Trans. Graph.*, 2(4):217–236, October 1983.

[19] Pierre Buyssens, Maxime Daisy, David Tschumperlé, and Olivier Lézoray. Exemplar-based inpainting: Technical review and new heuristics for better geometric reconstructions. *IEEE transactions on image processing*, 24(6):1809–1824, 2015.

[20] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *Computer Vision – ECCV 2016*, pages 809–824, 2016.

[21] Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-Preserving Half-Projective Warps for Image Stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3254–3261, 2014.

[22] Shenchang Eric Chen. QuickTime VR: An image-based approach to virtual environment navigation. In *Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, pages 29–38. ACM, 1995.

[23] Zhenzhong Chen, Yiming Li, and Yingxue Zhang. Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation. *Signal Processing*, 146:66–78, May 2018.

[24] Jaechoon Chon, Hyongsuk Kim, and Chun-Shin Lin. Seam-line determination for image mosaicking: A technique minimizing the maximum local mismatch and the global cost. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):86–92, January 2010.

[25] James J Clark and Nicola J Ferrier. Modal control of an attentive vision system. pages 514–523, 1988.

[26] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

[27] S. Croci, S. Knorr, L. Goldmann, and A. Smolic. A framework for quality control in cinematic VR based on Voronoi patches and saliency. In *2017 International Conference on 3D Immersion (IC3D)*, pages 1–8, December 2017.

[28] Subbareddy Darukumalli, Peter A Kara, Attila Barsi, Maria G Martini, and Tibor Balogh. Subjective quality assessment of zooming levels and image reconstructions based on region of interest for light field displays. In *2016 International Conference on 3D Imaging (IC3D)*, 2016.

[29] Francesca De Simone, Lutz Goldmann, Jong-Seok Lee, and Touradj Ebrahimi. Towards high efficiency video coding: Subjective evaluation of potential coding technologies. *Journal of Visual Communication and Image Representation*, 22(8):734–748, November 2011.

[30] G.V. der Auwera, M. Coban, Hendry, and M. Karczewicz. Ahg8: Tsp evaluation with viewport-aware quality metric for 360 video. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-E0070*, 2017.

[31] David P. Doane. Aesthetic Frequency Classifications. *The American Statistician*, 30(4):181–183, November 1976.

[32] A. Doshi and M. M. Trivedi. Head and gaze dynamics in visual attention and context learning. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 77–84, June 2009.

[33] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, and Y. Wang. A subjective study of viewer navigation behaviors when watching 360-degree videos on computers. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[34] Andrew Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer Science & Business Media, September 2007.

[35] Zeev Farbman, Raanan Fattal, and Dani Lischinski. Convolution Pyramids. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, pages 175:1–175:8, New York, NY, USA, 2011. ACM. event-place: Hong Kong, China.

[36] H. Sheikh Faridul, T. Pouli, C. Chamaret, J. Stauder, E. Reinhard, D. Kuzovkin, and A. Tremeau. Colour Mapping: A Review of Recent Methods, Extensions and Applications. *Computer Graphics Forum*, 35(1):59–88, 2016.

[37] A. Fatma, K. Khaled, and F. Zemzemi. Design, construction and calibration of an omnidirectional camera. In *2013 International Conference on Individual and Collective Behaviors in Robotics (ICBR)*, pages 49–55, December 2013.

[38] Chi-Wing Fu, Liang Wan, Tien-Tsin Wong, and Chi-Sing Leung. The Rhombic Dodecahedron Map: An Efficient Scheme for Encoding Panoramic Video. *IEEE Transactions on Multimedia*, 11(4):634–644, June 2009.

[39] L. Gaemperle, K. Seyid, V. Popovic, and Y. Leblebici. An Immersive Telepresence System Using a Real-Time Omnidirectional Camera and a Virtual Reality Head-Mounted Display. In *2014 IEEE International Symposium on Multimedia (ISM)*, pages 175–178, December 2014.

[40] Mahmoud Ghoniem, Youssef Chahir, and Abderrahim Elmoataz. Geometric and texture inpainting based on discrete regularization on graphs. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1349–1352. IEEE, 2009.

[41] Mark E. Glickman. Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, January 1999.

[42] J. Gluckman and S.K. Nayar. Ego-motion and omnidirectional cameras. In *Sixth International Conference on Computer Vision, 1998*, pages 999–1005, January 1998.

[43] David M. Goldberg and I. I. I. J. Richard Gott. Flexion and Skewness in Map Projections of the Earth. *Cartographica: The International Journal for Geographic Information and Geovisualization*, December 2007.

[44] N. Greene. Environment Mapping and Other Applications of World Projections. *IEEE Computer Graphics and Applications*, 6(11):21–29, November 1986.

[45] D. Guitton and M. Volle. Gaze Control in Humans: Eye-Head Coordination During Orienting Movements to Targets Within and Beyond the Oculomotor Range. *Journal of neurophysiology*, 58(3):427–459, 1987.

[46] F. Gustafsson. Determining the initial states in forward-backward filtering. *IEEE Transactions on Signal Processing*, 44(4):988–992, April 1996.

[47] Philippe Hanhart, Pavel Korshunov, Martin Rerabek, and Touradj Ebrahimi. JPEG backward compatible format for 3d content representation. pages 1–4. IEEE, 2013.

[48] David Hasler and Sabine E. Suesstrunk. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, volume 5007, pages 87–96. International Society for Optics and Photonics, June 2003.

[49] Kaiming He and Jian Sun. Statistics of patch offsets for image completion. In *Computer Vision–ECCV 2012*, pages 16–29. Springer, 2012.

[50] Stuart L. Henley. Seamless multi-camera panoramic imaging with distortion correction and selectable field of view, August 1997.

[51] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford, September 2011.

[52] V. Hosu, F. Hahn, O. Wiedemann, S. Jung, and D. Saupe. Saliency-driven image coding improves overall perceived JPEG quality. In *2016 Picture Coding Symposium (PCS)*, pages 1–5, December 2016.

[53] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Transactions on Graphics (TOG)*, 34(4):60, 2015.

[54] Fu-Chung Huang, Gordon Wetzstein, Brian A Barsky, and Ramesh Raskar. Eyeglasses-free display: towards correcting visual aberrations with computational light field displays. *ACM Transactions on Graphics (TOG)*, 33(4):59, 2014.

[55] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.

[56] S. Ikeda, T. Sato, and N. Yokoya. High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system. In *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI2003.*, pages 155–160, August 2003.

[57] Masaki Isshiki and Keiji Matsuki. Achromatic super wide-angle lens, August 1970.

[58] ITU-R T.871. Information technology - digital compression and coding of continuous-tone still images: JPEG file interchange format (JFIF), 2011.

[59] Tilke Judd, Frédo Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. January 2012.

[60] Peter A Kara, Maria G Martini, Peter Kovacs, Samdor Imre, Attila Barsi, Kristof Lackner, Tibor Balogh, et al. Perceived quality of angular resolution for light field displays and the validity of subjective assessment. In *2016 International Conference on 3D Imaging (IC3D)*, 2016.

[61] Christof Koch and Shimon Ullman. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In Lucia M. Vaina, editor, *Matters of Intelligence*, Synthese Library, pages 115–141. Springer Netherlands, 1987.

[62] Robert Konrad, Emily A Cooper, and Gordon Wetzstein. Novel optical configurations for virtual reality: evaluating user preference and performance with focus-tunable and monovision near-eye displays. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1211–1220. ACM, 2016.

[63] Simon Korman and Shai Avidan. Coherency sensitive hashing. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1099–1112, 2016.

[64] Pavel Korshunov and Touradj Ebrahimi. Context-dependent JPEG backward-compatible high-dynamic range image compression. 52(10):102006–102006, 2013.

[65] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut Textures: Image and Video Synthesis Using Graph Cuts. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 277–286, New York, NY, USA, 2003. ACM. event-place: San Diego, California.

[66] Douglas Lanman and David Luebke. Near-eye light field displays. *ACM Transactions on Graphics (TOG)*, 32(6):220, 2013.

[67] Olivier Le Meur, Josselin Gautier, and Christine Guillemot. Examplar-based inpainting based on local geometry. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3401–3404. IEEE, 2011.

[68] S. Leorin, L. Lucchese, and R. G. Cutler. Quality Assessment of Panorama Video for Videoconferencing Applications. In *2005 IEEE 7th Workshop on Multimedia Signal Processing*, pages 1–4, October 2005.

[69] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *null*, page 305. IEEE, 2003.

[70] 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)*, pages 370–374, September 2016.

[71] HC Lin, CY Li, JL Lin, SK Chang, and CC Ju. Ahg8: An efficient compact layout for octahedron format. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0142*, 2016.

[72] Hung-Chih Lin, Chao-Chih Huang, Chia-Ying Li, Ya-Hsuan Lee, Jian-Liang Lin, and Shen-Kai Chang. Ahg8: An improvement on the compact ohp layout. *JVET-E0056*, 2017.

[73] Jing Ling, Kao Zhang, Yingxue Zhang, Daiqin Yang, and Zhenzhong Chen. A saliency prediction model on 360 degree images using color dictionary based sparse representation. *Signal Processing: Image Communication*, 69:60–68, 2018.

[74] Mali Liu, Chihao Lu, Haifeng Li, and Xu Liu. Near eye light field display based on human visual features. *Opt. Express*, 25(9):9886–9900, May 2017.

[75] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[76] R. Duncan Luce. On the possible psychophysical laws. *Psychological Review*, 66(2):81–95, 1959.

[77] Andrew MacQuarrie and Anthony Steed. Object removal in panoramic media. In *Proceedings of the 12th European Conference on Visual Media Production*, page 2. ACM, 2015.

[78] P. C. Madhusudana and R. Soundararajan. Subjective and Objective Quality Assessment of Stitched Images for Virtual Reality. *IEEE Transactions on Image Processing*, 28(11):5620–5635, November 2019.

[79] Andrew Maimone and Henry Fuchs. Computational augmented reality eyeglasses. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, pages 29–38. IEEE, 2013.

[80] Vida Maliene, Vytautas Grigonis, Vytautas Palevičius, and Sam Griffiths. Geographic information system: Old principles with new capabilities. *URBAN DESIGN International*, 16(1):1–6, January 2011.

[81] Chadwick B. Martin. Design issues of a hyperfield fisheye lens. In *Novel Optical Systems Design and Optimization VII*, volume 5524, pages 84–92. International Society for Optics and Photonics, October 2004.

[82] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, March 2013.

[83] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. SalNet360: Saliency maps for omni-directional images with CNN. *Signal Processing: Image Communication*, 69:26–34, 2018.

[84] S.K. Nayar. Catadioptric omnidirectional camera. In *1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings*, pages 482–488, June 1997.

[85] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2018.

[86] Pradip Paudyal, Federica Battisti, Alessandro Neri, and Marco Carli. A study of the impact of light fields watermarking on the perceived quality of the refocused data. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*, pages 1–4. IEEE, 2015.

[87] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson Image Editing. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 313–318, New York, NY, USA, 2003. ACM. event-place: San Diego, California.

[88] Y. Rai, P. Le Callet, and P. Guillotel. Which saliency weighting for omni directional image quality assessment? In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017.

[89] Srikumar Ramalingam. Catadioptric Camera. In Katsushi Ikeuchi, editor, *Computer Vision*, pages 85–89. Springer US, 2014.

[90] Donald W Rees. Panoramic television viewing system. *United States Patent, No. 3, 505, 465*, 1970.

[91] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[92] Martin Rerabek, Evgeniy Upenik, and Touradj Ebrahimi. JPEG backward compatible coding of omnidirectional images. In *Applications of Digital Image Processing XXXIX*, volume 9971. SPIE, September 2016.

[93] Patrice Rondao Alface, Jean-Francois Macq, and Nico Verzijp. Interactive Omnidirectional Video Delivery: A Bandwidth-Effective Approach. *Bell Labs Technical Journal*, 16(4):135–147, March 2012.

[94] Dario D. Salvucci and Joseph H. Goldberg. Identifying Fixations and Saccades in Eye-tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78, New York, NY, USA, 2000. ACM.

[95] Dietmar Saupe, Franz Hahn, Vlad Hosu, Igor Zingman, Masud Rana, and Shujun Li. Crowd workers proven useful: A comparative study of subjective video quality assessment. In *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*, 2016.

[96] K. Seyid, V. Popovic, O. Cogal, A. Akin, H. Afshari, A. Schmid, and Y. Leblebici. A Real-Time Multiaperture Omnidirectional Visual Sensor Based on an Interconnected Network of Smart Cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(2):314–324, February 2015.

[97] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

[98] John P. Snyder. *Flattening the Earth: Two Thousand Years of Map Projections.* University Of Chicago Press, Chicago, 1993.

[99] Mikhail Startsev and Michael Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Processing: Image Communication*, 69:43–52, 2018.

[100] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5), May 2011.

[101] Herbert A. Sturges. The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153):65–66, March 1926.

[102] Y. Sun, A. Lu, and L. Yu. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Processing Letters*, 24(9):1408–1412, September 2017.

[103] Yule Sun, Ang Lu, and Lu Yu. WS-PSNR for 360 video quality evaluation. Proposal, MPEG2016/M38551, ISO/IEC JTC1/SC29/WG11, Geneva, May 2016.

[104] Richard Szeliski. Image Alignment and Stitching: A Tutorial. *Found. Trends. Comput. Graph. Vis.*, 2(1):1–104, January 2006.

[105] Masayuki Tanaka, Ryo Kamio, and Masatoshi Okutomi. Seamless Image Cloning by a Closed Form Solution of a Modified Poisson Problem. In *SIGGRAPH Asia 2012 Posters*, SA '12, pages 15:1–15:1, New York, NY, USA, 2012. ACM. event-place: Singapore, Singapore.

[106] I. Tosic and P. Frossard. Low bit-rate compression of omnidirectional images. In *Picture Coding Symposium, 2009. PCS 2009*, pages 1–4, May 2009.

[107] David Tschumperle and Rachid Deriche. Vector-valued image regularization with pdes: A common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):506–517, 2005.

[108] Evgeniy Upenik, Pinar Akyazi, Mehmet Tuzmen, and Touradj Ebrahimi. Inpainting in omnidirectional images for privacy protection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[109] Evgeniy Upenik and Touradj Ebrahimi. A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 73–78, July 2017.

[110] Evgeniy Upenik and Touradj Ebrahimi. Saliency Driven Perceptual Quality Metric for Omnidirectional Visual Content. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4335–4339, September 2019.

[111] Evgeniy Upenik, Martin Rerabek, and Touradj Ebrahimi. Testbed for subjective evaluation of omnidirectional visual content. In *2016 Picture Coding Symposium (PCS)*, pages 1–5, December 2016.

[112] Evgeniy Upenik, Martin Rerabek, and Touradj Ebrahimi. On the performance of objective metrics for omnidirectional visual content. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017.

[113] Evgeniy Upenik, Irene Viola, and Touradj Ebrahimi. A Rendering Solution to Display Light Field in Virtual Reality. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 246–250, September 2018.

[114] M. Uyttendaele, A. Eden, and R. Skeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II, December 2001.

[115] Irene Viola, Martin Řeřábek, and Touradj Ebrahimi. Comparison and evaluation of light field coding approaches. *IEEE Journal of selected topics in signal processing*, 2017.

[116] Irene Viola, Martin Rerabek, and Touradj Ebrahimi. Impact of interactivity on the assessment of quality of experience for light field content. In *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[117] Ruiqi Wang, Long Ye, Wei Zhong, Li Fang, and Qin Zhang. Novel pseudo-cylindrical projection based tile segmentation scheme for omnidirectional video. In *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, volume 10806, page 1080646. International Society for Optics and Photonics, August 2018.

[118] Eric W. Weisstein. Gnomonic Projection. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/GnomonicProjection.html.

[119] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.

[120] W. Xu and J. Mulligan. Performance evaluation of color correction approaches for automatic multi-view image and video stitching. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 263–270, June 2010.

[121] Raymond A Yeh, Chen Chen, Teck-Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *CVPR*, volume 2, page 4, 2017.

[122] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[123] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

[124] M. Yu, H. Lakshman, and B. Girod. A Framework to Evaluate Omnidirectional Video Coding Schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 31–36, September 2015.

[125] Matt Yu, Haricharan Lakshman, and Bernd Girod. Content Adaptive Representations of Omnidirectional Videos for Cinematic Virtual Reality. In *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, ImmersiveME '15, pages 1–6, New York, NY, USA, 2015. ACM.

[126] L. Yuan and T. Ebrahimi. Image transmorphing with jpeg. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3956–3960, Sept 2015.

[127] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park. Quality metric for spherical panoramic video. In *Proceedings Volume 9970, Optics and Photonics for Information Processing X*, volume 9970, 2016.

[128] Julio Zaragoza, Tat-Jun Chin, Michael S. Brown, and David Suter. As-Projective-As-Possible Image Stitching with Moving DLT. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2339–2346, 2013.

[129] C Zhang, Y Lu, J Li, and Z Wen. Ahg8: Segmented sphere projection for 360-degree video. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-E0025*, 2017.

[130] Yan Zhang, Lina Zhao, and Wanbao Hu. A Survey of Catadioptric Omnidirectional Camera Calibration. *International Journal of Information Technology and Computer Science*, 5(3):13–20, February 2013.

[131] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360° videos. pages 488–503, 2018.

[132] Minhua Zhou. Ahg8: A study on compression efficiency of cube projection. *Document JVET-D0022, Chengdu, CN*, 2016.

CURRICULUM VITAE

EVGENIY UPENIK

Location: Lausanne, Switzerland

Phone: +41 79 560 13 97
Email: evgeniy.upenik@gmail.com

EDUCATION

Docteur ès Sciences (PhD) in Electrical Engineering
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2015–2020 School of Electrical Engineering
Multimedia Signal Processing Group
Advisor: Prof. Dr. Touradj Ebrahimi

Engineer Diploma (MEng) in Laser Systems
Novosibirsk State Technical University, Novosibirsk, Russia
2008–2010 Laboratory of Laser Electronic Systems, Institute of Laser Physics Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia

Bachelor of Technology in Engineering Physics
Novosibirsk State Technical University, Novosibirsk, Russia
2007 Department of Laser Systems, Faculty of Physical Engineering

PUBLICATIONS

Peer Reviewed Conference Papers

1. E. Upenik, M. Rerabek and T. Ebrahimi. A Testbed for Subjective Evaluation of Omnidirectional Visual Content. 32nd Picture Coding Symposium, Nuremberg, Germany, December 4-7, 2016.
2. M. Rerabek, E. Upenik and T. Ebrahimi. JPEG backward compatible coding of omnidirectional images. SPIE Optics + Photonics, San Diego, California, USA, August 28 - September 1, 2016., Proceedings of SPIE.
3. E. Upenik, M. Rerabek, and T. Ebrahimi, On the performance of objective metrics for omnidirectional visual content, 9th International Conference on Quality of Multimedia Experience (QoMEX 2017), Erfurt, Germany, 2017.
4. E. Upenik and T. Ebrahimi, A Simple Method to Obtain Visual Attention Data in Head Mounted Virtual Reality, in IEEE International Conference on Multimedia and Expo 2017, Hong Kong, 2017.
5. E. Alexiou, E. Upenik, and T. Ebrahimi, Towards Subjective Quality Assessment of Point Cloud Imaging in Augmented Reality, 19th International Workshop on Multimedia Signal Processing (MMSP 2017), London-Luton, UK, 2017.
6. E. Upenik, I. Viola, and T. Ebrahimi, A Rendering Solution to Display Light Field in Virtual Reality, 26th European Signal Processing Conference (EUSIPCO 2018), Rome, Italy, 2018.
7. E. Upenik and T. Ebrahimi, "Saliency Driven Perceptual Quality Metric for Omnidirectional Visual Content," in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4335–4339.

8. E. Upenik, P. Akyazi, M. Tuzmen, and T. Ebrahimi, "Inpainting in Omnidirectional Images for Privacy Protection," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2487–2491.

Technical documents

1. ISO/IEC JTC 1/SC29/WG1 M72021, A test bed for the assessment of visual quality in 360 VR, E. Upenik, D. McNally and T. Ebrahimi, 72st JPEG Meeting, Geneva, Switzerland, June 2016
2. ISO/IEC JTC 1/SC29/WG1 M76089, Objective and Subjective Assessment of Omnidirectional 360 Visual Content, E. Upenik and T. Ebrahimi, 76th JPEG Meeting, Turin, Italy, July 2017
3. ISO/IEC JTC 1/SC29/WG1 M76090, Visual Attention for Omnidirectional 360 Visual Content, E. Upenik and T. Ebrahimi, 76th JPEG Meeting, Turin, Italy, July 2017

EMPLOYMENT

Doctoral Researcher at *École Polytechnique Fédérale de Lausanne,* Lausanne, Switzerland;
January 2015–Present

Multimedia Signal Processing Group headed by Prof. Touradj Ebrahimi

- Subjective and objective evaluation of 360 degree images and video
- Compression methods for 360 degree images and video
- Augmented and virtual reality
- Teaching assistance in "Media Security" and "Image and Video Processing" courses

Committee Member at *ISO/IEC JTC 001/SC 29/WG 01 "Coding of still pictures" (JPEG)*
March 2016–Present

Committee Member at *ISO/IEC JTC 001/SC 29/WG 11 "Coding of moving pictures and audio" (MPEG)*
March 2016–Present

DSP Software Engineer at *R&D, Streambox, Inc.*, Novosibirsk, Russia; Seattle, USA.
August 2008–December 2014

Video encoding, Massively Parallel Processor Arrays, DSP, ARM, Embedded Linux

- Software development for embedded Linux (custom ARM-based platform): boot loader, kernel customizing, drivers, and user's apps.
- Implementation of AVC based video encoding algorithms on a massively parallel processors array device (Ambric, 336 cores MPPA) including parts as DCT, motion estimation, entropy coding, image filtering etc. HD SDI (1080i, 720p) video streams software processing.
- Software and HDL development for a DSP+FPGA systems including inter-chip communication and entropy coding.

Assistant Lecturer at *Institute of Laser Physics, Novosibirsk State Technical University, Novosibirsk*, Russia
February 2010–June 2013

- Matlab practice class (spring semester)

Contract Research Engineer at *Institute of Laser Physics, Siberian Branch of Russian Academy of Sciences*, Novosibirsk, Russia
January 2013–May 2013

- Laser beam position stabilization system implemented and put in operation. System is based on analysis of the image signal from CMOS censor to define a beam position, and driving motorized mirrors.

Contract Firmware Engineer at *Center of Financial Technologies, Inc.*, Novosibirsk, Russia.
March 2008 – July 2008

Firmware design for a wireless transceiver

- Firmware development for a handheld wireless device (logistic tracking and data collection system) including radio protocol design, battery usage optimization, graphic LCD module operating, software ADC implementation etc. on a low-power MCU (MSP430).

Electronic Design Assistant Engineer at *Institute of Design and Technology of Computers, Siberian Branch of Russian Academy of Sciences*, Novosibirsk, Russia
October 2006–March 2008

Analog and digital electronics, PCB, Firmware, Manufacturing management

- Development of a time-domain reflectometer device for cable integrity checking, including system modeling, circuit design, and firmware design for the DSP.
- Development of data collection modules for the mine conveyor control system. Manufacturing management and test procedures elaboration.
- Assisting in development of an adaptive control system for the asynchronous three-phase motor.

Intern at *Institute of Design and Technology of Computers, Siberian Branch of Russian Academy of Sciences*, Novosibirsk, Russia
September 2005–June 2006

Assisting in analog and digital electronic design, Programming MCUs

- As a part of practical training a universal data-com network based on a multi-point wireless link had been developed.

Quality Assurance Engineer at *SWsoft Inc. (since Jan 2008 Parallels Inc.)*, Novosibirsk, Russia.
August 2004–September 2005

Quality Assurance

- Using Bugzilla, CVS, VMWare; advanced bash/sed/awk scripting, Perl, SQL; Testing servers maintaining (Linux, FreeBSD)

OTHER EDUCATION

Massively Parallel Processor Arrays Programming Training at *Ambric Inc. HQ*, Portland, Oregon, USA. July 2009.

High Performance Computing at the *Institute of Computational Technologies* of Russian Academy of Sciences (Novosibirsk, Russia) in cooperation with *High Performance Computing Center Stuttgart (HLRS)* of the University of Stuttgart by Prof. Thomas Bönisch**.** Certificate of accomplishment. October 2011.

Functional Programming Principles in Scala at *Coursera* by Prof. Martin Odersky (École Polytechnique Fédérale de Lausanne). Certificate signed by instructor. November 2012.

Machine Learning at *Coursera* by Prof. Andrew Ng (Stanford University). Certificate signed by instructor. November 2012.

Summer School on Domain Specific Programming Languages at *EPFL* by M. Odersky, Ph. Wadler, M. Flat, E. Visser, M. Püschel et al**.** Doctoral course. Lausanne, Switzerland. June 2015.

LANGUAGES

Russian – mother tongue;
English – full professional working proficiency;
French – professional working proficiency (B2)