Objective Quality Assessment of 3D Synthesized Views

Luís Miguel Domingos Nunes

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. Fernando Manuel Bernardo Pereira

Prof. João Miguel Duarte Ascenso

Prof. Catarina Isabel Carvalheiro Brites

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Supervisor: Prof. Fernando Manuel Bernardo Pereira

Members of the Committee: Prof. Tomás Gomes Silva Serpa Brandão

November 2017

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledments

In first place, I would like to thank to my father Eduardo, my mother Maria, my sister Ana, my brothers Eduardo and Rui, my brother-in-law Telmo, and my sisters-in-law Carolina and Alexandra for supporting me in every moment of this journey and for always doing everything to help me. Anything I say is not enough to express my sense of debt to you.

A special thanks to my supervisors, Prof. Fernando Pereira, Prof. João Ascenso and Prof. Catarina Brites, for all the availability, guidance, patience and support in every matter possible, and also for all shared valuable lessons that will be useful in the rest of my life; and to Alexandre, João, André, Miguel, Fabio, Guilherme, Pedro and Renam for the support and company, and also for sharing their friendship with me.

Finally, a special word of appreciation to all that in any way have shared their time and energy with me during this journey, I feel humbled and honoured.

Resumo

Com a explosão das tecnologias digitais surgem novos modelos de representação visual, do stereo ao vídeo multiview, e destinados a diversas aplicações como as de difusão 3D de desporto, pós-processamento de filmes, etc. Um formato de representação que se mostra interessante e promissor é o multiview plus depth, onde a distância da câmara aos objetos (referida como mapa de profundidade) pode ser obtida para além da textura, por várias vistas. Este formato permite sintetizar vistas no descodificador evitando a necessidade de adquirir, codificar e transmitir um número maior das mesmas. No descodificador, as técnicas de síntese das vistas criam outras novas vistas a partir de um conjunto de vistas vizinhas. Infelizmente, este processo pode criar diversos tipos de distorções geométricas na vista sintetizada. Portanto, para monitorizar a qualidade da experiência do utilizador final ou até conduzir à otimização do processo de codificação e transmissão (ex: decidir que vistas devem ser enviadas) devem ser procuradas métricas automáticas, objetivas e de qualidade.

No seguinte trabalho, é apresentada uma métrica objectiva full-reference, relativa à qualidade de vídeo. Esta funciona no domínio espácio-temporal e considera o flickering e as distorções 2D. De modo a avaliar o desempenho da métrica, foi utilizada uma base de dados relevante, referente à qualidade de vídeos sintetizados.

A métrica de avaliação de qualidade de vídeo proposta supera o estado da arte 2D e métricas objetivas de qualidade 3D. Além disso, verificou-se também que utilizar um modelo just-noticeable difference ao nível do pixel para a distorção causada pelo flickering melhorou a correlação linear da métrica, mas diminuiu a correlação monotónica.

Palavras-chave: Qualidade, Qualidade de Experiência, 3D, 3D-HEVC, Vistas Sintetizadas, Qualidade de Vistas Sintetizadas.

Abstract

With the explosion of digital technologies new models of visual representation have also emerged, from stereo to multiview video and for several applications such as 3D sports broadcasting, movie post-processing, etc. An interesting and promising representation format is the multiview plus depth where the distance from the camera to the objects can be acquired (referred as depth map) besides the texture for several views. This format allows to synthesize views at the decoder, avoiding the need to acquire, code and transmit a large number of views. At the decoder, the view synthesis techniques create novel views from a set of neighbouring views. Unfortunately, this process can create several types of geometric distortions in the synthesized view. Thus, to monitor the quality of experience for the end-user or even drive optimization of the encoding and transmission process (e.g. decide which views must be sent), automatic objective quality metrics need to be pursued.

In this document, a full-reference objective video quality metric is presented. This metric works on the spatiotemporal domain and address the flickering and 2D spatial distortions. In order to assess the metrics performance, a relevant synthesized video quality database was used.

The proposed video quality assessment metric outperforms relevant state-of-the-art 2D and 3D objective quality metrics. Also, it was shown that using a pixelwise just-noticeable difference model for flickering distortion has improved the metric linear correlation, but it lowered the monotonic correlation.

Keywords: Quality, Quality of Experience, 3D, 3D-HEVC, Synthesized Views, Synthesized Views Quality.

Table of Contents

Acknowledments v
Resumo vii
Abstract ix
Table of Contents xi
List of Figures xv
List of Tables xvii
Acronyms xix
1. Introduction 1
1.1. Context and Motivation 1
1.2. Objectives 3
1.3. Thesis Structure 3
2. Multiview Video: Basics, Coding, and View Synthesis 5
2.1. 3D Perception: Basic Concepts and Systems 5
2.2. High Efficiency Video Coding Standard: Brief Review 9
2.2.1. Architecture 9
2.2.2. Block Partitioning 10
2.2.3. Intra Coding Tools 11
2.2.4. Inter Coding Tools 12
2.2.5. Transform and Quantization 13
2.2.6. Entropy Coding 13
2.2.7. In-loop Filtering 13
2.2.8. Performance 13
2.3. Multiview Video Coding: Brief Evolution Review 15
2.3.1. Texture based Multiview Video Coding Formats 15
2.3.2. Texture plus Depth based Multiview Video Coding Formats 16
2.4. Multiview High Efficiency Video Coding Standard: Brief Overview 17
2.5. 3D-HEVC Coding Standard: Brief Review 18
2.5.1. Architecture 18
2.5.2. Texture Data Coding 18
2.5.3. Depth Data Coding 20
2.5.4. Performance 22
2.6. View Synthesis 23
2.6.1. Basics 24

2.6.2. Main Tools 25
3. Image and Video Quality Assessment 27
3.1. Image and Video Distortions (2D and 3D) 28
3.2. 2D Objective Quality Metrics 30
3.2.1. PSNR: Peak Signal-to-Noise Ratio 31
3.2.2. SSIM: Structural Similarity Index 32
3.2.3. VIF: Visual Information Fidelity 33
3.2.4. VQM: Video Quality Metric 34
3.2.5. MOVIE: Motion-based Video Integrity Evaluation 35
3.3. 3D Quality Metrics 36
3.3.1. 3DSwIM: 3D Synthesized view Image Quality Metric 37
3.3.2. SIQE: Synthesized Image Quality Evaluator 38
4. Spatio-Temporal Quality Assessment for Synthesized Views Metric 41
4.1. Architecture and Walkthrough 42
4.2. Main Modules Detailed Description 43
4.2.1. Quality Assessment Group of Pictures Splitting 43
4.2.2. Reference View Spatio-Temporal Tubes Creation 43
4.2.3. Synthesized View Spatio-Temporal Tube Creation 46
4.2.4. Spatial Gradient Computation 47
4.2.5. Temporal Gradient Computation 47
4.2.6. Spatio-Temporal Activity Distortion Computation 47
4.2.7. Flickering Distortion Computation 48
4.2.8. Overall Distortion Computation 53
5. 3D Synthesized Views Relevant Databases 55
5.1. Synthesized Image Quality Assessment Databases 55
5.1.1. Media Communications Lab 3D Database 56
5.1.2. IRCCyN-IVC DIBR Image Quality Assessment Database 58
5.2. Synthesized Video Quality Assessment Databases 59
5.2.1. IRCCyN-IVC DIBR Video Quality Assessment Database 59
5.2.2. SIAT Synthesized Video Quality Assessment Database 60
6. Quality Metrics Performance Assessment 63
6.1. VQA Metric Performance Assessment Workflow 63
6.2. VQA Metric Configuration Profiles 65
6.2.1. Configuration #1: Motion Vector Estimation Approach (MVEA) 65
6.2.2. Configuration #2: Flickering Distortion Perception Threshold Model (FDPTM) 65
6.2.3. Configuration #3: Edge Detection Threshold Approach (EDTA) 65
6.3. Performance Study of VQA Metric Configurations 65

6.3.1. Motion Vector Estimation Approach Analysis 67
6.3.2. Edge Detection Threshold Approach Analysis.....69
6.4. Performance Assessment of the proposed video quality metric.....70
6.4.1. Comparison to 2D Objective Quality Metrics 70
6.4.2. Comparison to 3D Objective Quality Metrics 73
7. Summary and Future Work 77
7.1. Summary.....77
7.2. Future Work 78
References.....81

List of Figures

Figure 1.1 – Stereo and motion parallax [2]. 2
Figure 2.1 – a) Lenticular imaging display; b) Integral imaging display [14]. 8
Figure 2.2 – a) Barrier-grid display; b) Moving-slit display (adapted from [14]). 8
Figure 2.3 – HEVC hybrid encoding architecture [32]. 10
Figure 2.4 – a) CU partitioning units; b) Hierarchical block coding structure [33]. 11
Figure 2.5 – a) Intra prediction modes; b) Intra prediction mode-dependent coefficient scanning order [33]. 12
Figure 2.6 – Interactive scenario HEVC performance results [40]. 14
Figure 2.7 – Entertainment scenario HEVC performance results [40]. 14
Figure 2.8 – Subjective test results: mean opinion score versus bitrate [40]. 15
Figure 2.9 – Subjective test results: bitrate savings versus mean opinion score [40]. 15
Figure 2.10 – a) Spatial Multiplexing: Side-by-side; b) Spatial Multiplexing: Top-Bottom; c) Temporal Multiplexing. 16
Figure 2.11 – a) MVC typical temporal plus inter-view prediction structure; b) Depth Enhancement Multiview Coding: Frame (top-left), Depth (top-right), occlusion layers (bottom) [42]. 16
Figure 2.12 – MV-HEVC encoder architecture [44]. 17
Figure 2.13 – 3D-HEVC encoder architecture [46]. 18
Figure 2.14 – a) Inter-view residual prediction; b) Temporal residual prediction [44]. 20
Figure 2.15 – Wedge (top) and Contour (bottom) Intra prediction modes [50]. 21
Figure 2.16 – Subjective performance results [50]. 23
Figure 2.17 – Average bitrate distributions: 2-SD (top), 3-ASD (bottom) [50]. 23
Figure 2.18 – Simplified view synthesis architecture [56]. 26
Figure 3.1 – Common 2D image-based distortions: a) blurring effect; b) blocking effect [62]. 29
Figure 3.2 – Common 3D image-based distortions: a) geometric distortion; b) ghosting distortion; c) cracks; d) occluded areas [62]. 30
Figure 3.3 – Visual Information Fidelity metric architecture. 33
Figure 3.4 – MOVIE architecture. 35
Figure 3.5 – 3DSwIM architecture [59]. 37
Figure 3.6 – SIQE architecture [75]. 39
Figure 4.1 – Processing architecture for the Spatio-Temporal Video Quality metric. 42
Figure 4.2 – QA-GOP and S-T tube structure [80]. 44
Figure 4.3 – S-T tubes creation architecture. 44

Figure 4.4 – Block-based Motion Estimation: DST and SRC block examples when the SRC is the central frame..... 45
Figure 4.5 – Spatial gradient kernels: a) horizontal; b) vertical. 47
Figure 4.6 – Edge emphasized JND model architecture. 49
Figure 4.7 – Image decomposition: a) Image; b) Structural; c) Textural. 50
Figure 4.8 – Directional high-pass filters for texture detection. 52
Figure 5.1 – MCL-3D database content and processes [89]. O, R and D refer to the original, reference and distorted data/views, respectively; lower script T and D refer to texture and depth data. 56
Figure 6.1 – Estimated motion vectors histogram: a) PoznanStreet; b) PoznanHall2; c) GT_Fly. 67
Figure 6.2 – Motion compensated blocks: a) Reference frame; b) BM_32; c) BM_64; d) BM_64_CF. 68
Figure 6.3 – BookArrival: a) Structural image; b) Edge map (static threshold); c) Edge map (adaptive threshold). 69
Figure 6.4 – BookArrival: a) Luminance samples; b) Edge map (static threshold); c) Edge map (adaptive threshold). 69
Figure 6.5 – Pearson correlation coefficient for each 2D image metric. 71
Figure 6.6 – DMOS versus $DMOS_p$ for different 2D objective quality metrics. 72
Figure 6.7 – 3DSwIM DMOS vs Score per: a) Sequence; b) Subset. 73
Figure 6.8 – DMOS versus $DMOS_p$ for different 3D objective quality metrics. 74

List of Tables

Table 2.1 – 3D-HEVC average bitrate savings (BD-Rate). 22
Table 5.1 – SIAT Synthesized Video Quality Database: content and coding characteristics. 61
Table 6.1 – Motion Vector Estimation: configuration characteristics. 65
Table 6.2 – Performance Assessment on Relevant Metric Configurations. 66
Table 6.3 – Performance Comparison of Objective Video Quality Assessment: 2D VQA. 70
Table 6.4 – Performance Comparison of Objective Video Quality Assessment: 3D VQA. 73
Table 6.5 – Performance Comparison by Subset: SIAT and Proposed. 74
Table 6.6 – RMSE of the proposed metric per sequence. 74

Acronyms

3D-HEVC	3D High Efficiency Video Coding
3DSwIM	3D Synthesized view Image quality Metric
ACR-HR	Absolute Category Reference with Hidden Reference
AMVP	Advanced Motion Vector Prediction
ARP	Advanced Residual Prediction
AMM	Affine Motion Model
ANSI	American National Standards Institute
ASD	Auto-Stereoscopic Display
BO	Band Offset
BS	Boundary Strength
BV	Bounded Variation
CB	Coding Block
CTB	Coding Tree Block
CTU	Coding Tree Unit
CU	Coding Unit
CABAC	Context-Based Adaptive Binary Arithmetic Coding
CM	Contrast Masking
DBF	Deblocking Filter
DIBR	Depth Image Based Rendering
DIBR	Depth Image-Based Rendering
DIS	Depth Intra Skip
DLT	Depth Lookup Table
DoNBDV	Depth Oriented Neighbouring Block Based Disparity Vector
DMOS	Differential Mean Opinion Score
DCT	Discrete Cosine Transform
DST	Discrete Sine Transform
DWT	Discrete Wavelet Transform
DCP	Disparity Compensated Prediction
DMV	Disparity Motion Vector
DA	Distortion Activity
DN	Divisive Normalization
EDTA	Edge Detection Threshold Approach
EM	Edge Masking

EO	Edge Offset
fps	Frames Per Second
GSM	Gaussian Scale Mixtures
HD	High Definition
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
HLS	High-Level Syntax
HSV	Human Visual System
IVC	Image and Video-Communication
IQA	Image Quality Assessment
IBMR	Image-Based Modelling and Rendering
ISF	Instance Scene Flow
IRCCyN-IVC	Institut de Recherche en Communications et Cybernétique de Nantes - Images and Video-communications
ITU	International Telecommunications Union
JND	Just Noticeable Difference
JNDe	Edge Emphasized Just Noticeable Difference
KS	Kolmogorov-Smirnov
LA	Luminance Adaptation
MAD	Mean of Absolute Differences
MOS	Mean Opinion Score
MSE	Mean Square Error
MCP	Motion Compensated Prediction
MV	Motion Vector
MVEA	Motion Vector Estimation Approach
MVP	Motion Vector Prediction
MOVIE	MOtion-based Video Integrity Evaluation
MV-HEVC	Multiview High Efficiency Video Coding
MVD	Multiview Plus Depth
MVC	Multiview Video Coding
NTIA	National Telecommunications and Information Administration
NSS	Natural Scene Statistics
NBDV	Neighbouring Block Disparity Vector
NAL	Network Abstraction Layer
NAMM	Nonlinear Additivity Model for Masking
PC	Pairwise Comparison
PSNR	Peak Signal-To-Noise Ratio
PTM	Perception Threshold Model
PPS	Picture Parameter Set
PRSM	Piecewise Rigid Scene Model

$DMOS_P$	Predicted Difference Mean Opinion Score
PDI	Predicted Disparity Information
PDV	Predicted Disparity Vector
PB	Prediction Block
PU	Prediction Unit
QTL	Quadtree Limitation
QA-GOP	Quality Assessment-Group of Pictures
QoE	Quality of Experience
Q_P	Quantization Parameter
Q_{STEP}	Quantization Step
RD	Rate-Distortion
RV	Reference View
RMSE	Root Mean Square Error
SAO	Sample Adaptive Offset Filter
SIAT	Shenzhen Institute of Advanced Technology
S-T	Spatio-Temporal
SD	Stereoscopic Display
SSIM	Structural Similarity Index
SBP	Sub-Block Partitions
SIQE	Synthesized Image Quality Evaluation
TMV	Temporal Motion Vector
TM	Texture Masking
TV	Total Variation
TB	Transform Block
TU	Transform Unit
2D	Two-Dimensional
3D	Three-Dimensional
UHD	Ultra-High-Definition
URQ	Uniform Reconstruction Quantization
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
VSRS	View Synthesis Reference Software
VR	Virtual Reality
VIF	Visual Information Fidelity

CHAPTER 1

Introduction

This chapter will introduce the topic addressed in this Thesis, which is the objective quality assessment of 3D synthesized views. Thus, it will start by presenting the context and motivation of this work followed by the presentation of the Thesis main objectives and structure.

1.1. Context and Motivation

The development of communication technologies has always played a key role in Human evolution due to its capability to exchange experiences amongst individuals, societies and cultures. As technology evolved, new methods have emerged to store and share information, from a simple paper coated with silver chloride able to preserve a black-and-white picture of a prosperous man in mid-later 1800s, up to yesterday's video posted online of a cat playing piano. The evolution of communication tools gained momentum during the nineteenth century, notably with the invention of the telephone. In less than 200 years, the state-of-the-art on imaging went from two-dimensional (2D) black-and-white photograph to High Definition (HD) 3D colour video. The major factor for the increased speed on the evolution of imaging technologies was the digitization of multimedia information, which began around mid-1900s with speech data. After a couple of decades, digital image, video and audio emerged as a new major communication technology to store and transmit moving pictures, giving credit again to the popular stating: "If a picture is worth a thousand words then a video is worth a million", meaning that although a still image is far more descriptive than words, moving pictures (videos) are far even more descriptive.
However, the real world is not 2D but rather 3D, and thus visual representation naturally evolved towards 3D data. Although 3D image and video representation has deep historical roots [1], only a few years ago this kind of representation techniques became more popular, bringing to our lives 3D cinema experiences, virtual reality headsets, and home stereo displays able to improve users' immersion through 3D experiences. Naturally, image stereo pairs are the simplest form of 3D imaging where the 3D experience is created by providing the left and right eyes with two different scene views, the so-called stereo parallax, see Figure 1.1. However, this solution does not offer motion parallax as the pair of views do not change when the viewer's position changes. To provide richer 3D experiences, autostereoscopic displays use a larger number of scene views allowing the viewer to see different views

and thus slightly different perspectives of the scene, as it takes different positions in front of the screen, thus offering also motion parallax. Representing the scene with a high number/density of views leads to acquisition, storage and transmission challenges as it requires a high number of cameras covering the visual scene, larger storage devices, and large bandwidth resources. To avoid the acquisition and transmission of all views while still offering smooth motion parallax, a new representation approach was developed, known as view synthesis. View synthesis opens the possibility to interpolate as many as wanted *virtual views* at the receiver from two (or more) decoded views. Typically, the view synthesis process uses the views located at the left and right of the view to be interpolated. This solution allows the display of many views, while only coding a few selected views, thus enriching the 3D experience without increasing too much the rate.

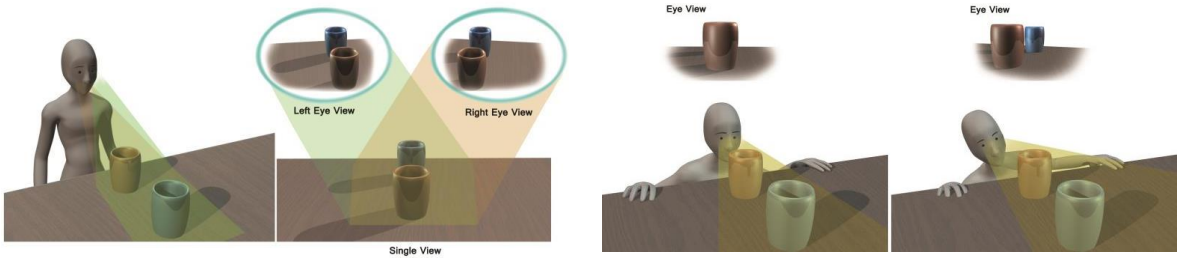


Figure 1.1 – Stereo and motion parallax [2].

The increased number of views and the increasing resolution of each view required the development of highly efficient compression and decompression tools, so-called codecs. Similarly to the 2D case, the 3D video codecs are lossy, meaning that they exploit the data perceptual irrelevancy, as well as the redundancy among the views and components in addition to spatial, temporal and statistical redundancies. Lossy codecs are preferable in video applications due to their higher compression capabilities, while still achieving high perceptual quality. The key to the high performance of 3D codecs is the exploitation of the temporal and inter-view correlations with efficient tools that are able to capture the similarities in time and space together. In this domain, there are several international standards, notably developed by the ISO/IEC MPEG standardization group, such as MVC, MV-HEVC and 3D-HEVC.

However, whatever the coding and synthesis solutions, at the end of the day, quality has to be assessed to validate the developed solutions.

In fact, quality assessment is paramount in evaluating the performance of video capture, compression and transmission steps. An accurate video quality assessment will provide information to optimize the overall system performance, and may lead to an optimization of the Quality of Experience (QoE) for the end users, which means enhancement of the perceivable quality for the same bitrate. While the best type of quality assessment is the subjective assessment, where human subjects are asked to rate the quality, many times this is not possible, as it represents a cumbersome and time consuming process. Thus, objective quality assessment methods are developed with subjective assessment estimation as the target. Moreover, while the quality assessment of the decoder views can be full reference by definition as the original frame to be coded is available, for the synthesized views this may not be true, as these frames are only synthesized and not coded. Moreover, the type of processing artefacts that

can be found in synthesized views are different, as the applied processing mechanism is a synthesizer and not an encoder.

The video quality assessment methods for the synthesized views have a fundamental role in the performance evaluation of several 3D video coding systems. In these systems, the synthesized view quality is considered very relevant in the overall QoE for the 3D video processing chain. Also, by designing an efficient quality assessment metric, it would become possible to optimize video codecs using this distortion metric, therefore improving the rate-quality efficiency of 3D video coding systems.

1.2. Objectives

Following the context above, the main goal of this MSc Thesis is to present an objective video quality assessment metric for 3D synthesized views. In order to achieve this objective, the developed work was organized according to the following steps:

- Revision of the state-of-the-art in multiview video coding and the main related tools;
- Revision of the most relevant 2D and 3D image and video objective quality assessment metrics;
- Revision of the most relevant 3D synthesized views quality assessment databases available in the literature;
- Implementation of an objective video quality assessment metric for synthesized views;
- Assessment of the developed metric and comparison with other 2D and 3D objective quality assessment metrics available in the literature.

1.3. Thesis Structure

In order to successfully achieve the proposed objectives presented in Section 1.2, this Thesis is organized as follows:

- Chapter 2 begins by describing the 3D basic perception concepts, and by reviewing the state-of-the-art on multiview video, notably: 3D video coding standards; and the main tools used in the view synthesis process.
- Chapter 3 presents the most common 2D and 3D distortions and describe their cause. It also reviews the most widely used image and video objective quality assessment metrics for 2D and 3D synthesized views, as well as the main subjective methods to evaluate image and video perceptual quality.
- Chapter 4 presents the developed solution architecture with a walkthrough characterization, and describes its main tools and methods to compute the objective score.
- Chapter 5 presents the four most relevant 3D synthesized views databases available in the literature, discussing their importance and content. Moreover, one of these databases will be used to evaluate the performance of different quality assessment metrics in the view synthesis context.
- Chapter 6 presents the assessment workflow, test conditions, and performance assessment metrics used to measure the performance of objective quality metrics. Also, reviews the performance of the implemented objective video quality assessment metric using different

configuration profiles. These configuration profiles allow the deeper understanding of the metrics behaviour, and it is carried out by evaluating the impact of the different configurations, notably: motion vector estimation; perceivable distortion models; and methods to extract the edge map. After this process, a single profile is proposed and its performance compared against several 2D and 3D objective quality metrics.

- Chapter 7 presents the main conclusions achieved during this process and some directions for the future work.

CHAPTER 2

Multiview Video: Basics, Coding, and View Synthesis

This chapter presents the basics, evolution and a number of relevant solutions related to multiview video coding. To achieve this objective, it starts by explaining several basic concepts about 3D perception and systems, then proceeds to a brief evolutional review of multiview video coding with special emphasis on the most recent standard, 3D-HEVC (High Efficiency Video Coding), with associated synthesis and depth extraction tools. Finally, it reviews some basic concepts on view synthesis and presents the main related tools used in the context of multiview coding standards.

2.1. 3D Perception: Basic Concepts and Systems

Humans sense the world through a series of sensors/receptors, which when stimulated send signals across the nervous system to the relevant sensory cortex area located in the brain. Since the scope of this Thesis is visual data and its perception, the *stimuli* associated to the non-visual receptors will not be considered, *e.g.* hearing. To begin with, it is important to highlight that 3D perception is largely based on the perception of depth. This perception is attained through the patterns *cognized* by the visual cortex, which provide information about points in space; these patterns are denominated **depth cues** [1]. These cues have different importance and impact depending on the relative distance between the observer and the object, *e.g.* the blur effect is an effective depth cue only for near distances [3]. Additionally, the perception of depth is enhanced when the information obtained from each cue is correlated [3] [4]. Although the human brain processes the cues *on the fly,* depth is not only perceived from stimulus acquired in a specific time instant as the human brain also performs correlations with the accumulated knowledge from prior experiences [5].

Most humans are born with a pair of visual receptors (eyes). Considering that both eyes are in front of the skull, in a stationary position, the binocular field of vision is more than 75%, thus providing stereopsis, *i.e.* the perception of depth and 3D structures based on binocular vision [6]. *Stereopsis*, or *stereo parallax***,** provides advantages in basic and also more demanding visual tasks, such as detecting

camouflaged objects, reading, and eye-hand coordination [7] [8]. This feature is associated to two exclusively binocular depth cues denominated binocular *disparity* and *vergence* [6]:

- ***Binocular disparity*:** Also known as binocular/stereo parallax, refers to the difference between the pair of views acquired with the two eyes, which through *binocular summation* provides a high-fidelity depth perception; this summation is a process combining the right and left eyes information to perform some tasks [3] [6];
- ***Binocular vergence*:** Refers to the act of *convergence* or *divergence* of both eyes towards a specific point in space so that the projection of the image is at the centre of the retina in both eyes, where the density of photoreceptor cells is higher, therefore obtaining a clearer vision of that point in space.

The only monocular cue that grants a high level of depth precision is **motion parallax**, which is associated to depth perception through motion. Motion parallax and binocular parallax are closely related, since the successive views (multiple views) obtained through motion and the interocular distance allow a similar perception of depth [9]. However, with the observer's motion, the objects within his/her field of vision will suffer a series of deformation disparities. These disparities can be decomposed into a weighted sum of four first-order differential transformation components: *Expansion,* or dilatation; *Curl*, or rotation; a couple of *Shear-deformation* components [10]. Combining each weighted component considering its own velocity gradient leads to five types of motion parallax [4] [11]:

- ***Linear parallax*:** When a pair of objects are moving at the same speed at different distances to the observer, the farther object will seem to be moving slower;
- ***Looming parallax*:** A moving object will *expand* faster when closer to the observer;
- ***Rotation parallax*:** A rotating object with a rotation axis perpendicular to the observer line of sight will create a strong impression of depth;
- ***Shear-deformation parallax*:** A moving object will display some deformities during its movement, depending on its geometry;
- ***Compression- or expansion-deformation parallax*:** When an object is getting closer or farther from the observer, it will produce a deformation effect where the closer part of the object will seem to be bigger than the farther part.

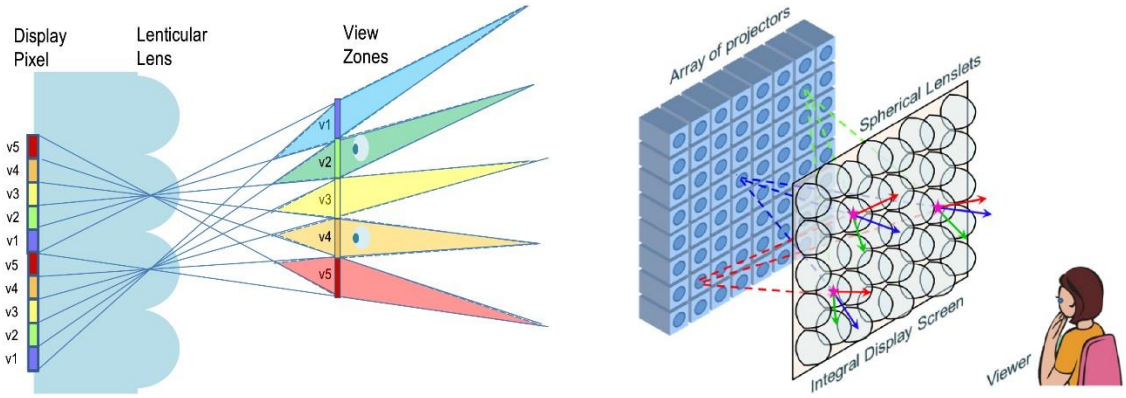
Throughout the years, the Human need to represent reality in the most authentic way has kept pushing further all types of 3D technologies. Although some studies were conducted in past centuries [1], only in the past couple of decades this evolution accelerated, notably due to the emerging computer and digital systems technologies. Since then, techniques to display 3D pictures and movies had evolved from passive stereoscopic displays to active stereoscopic displays and auto-stereoscopic and more recently to light field displays [1].

Stereoscopic displays involve two views and use multiplexing techniques to enable the displaying of both views in a particular area within a specific time period (simultaneously or not) with each view processed by the appropriate eye. Stereoscopic display techniques typically involve the use of a pair of glasses and can be grouped into four classes which differ on the way they control the exposure of the appropriate view to the right eye, notably by using different wavelengths, polarizations, times and

spaces. In *wavelength multiplexing,* each view is modulated in different wavelengths; after, to enable stereopsis, a pair of filters is used in the glasses, one for each eye, so that the appropriate view can be delivered to the appropriate receptor. *Polarization multiplexing* is similar to wavelength multiplexing, but instead of using wavelength modulation for the two views, the views are orthogonally polarized and filtered, eventually on the same wavelength. *Time multiplexing* refers to active stereoscopic systems that alternatively display each view in succeeding short periods of time; these periods must be really short to avoid a flickering effect, which may occur when the observer can perceive the refreshing periods. Finally, *spatial multiplexing* refers to stereoscopic systems that display each view in different areas of the display. Nowadays, this technique is used in virtual reality (VR) headsets and in YouTube VR 3D videos, where each view is relayed to the appropriate eye through separate sets of lenses and mirrors. Moreover, this technique is also used in movie theatres and TV sets in combination with wavelength, polarization and time (active shutter) multiplexing systems [12]. All these systems require the observer to use a pair of glasses or another sort of apparatus, which is largely considered a negative factor in terms of QoE. Stereoscopic displays are known to cause viewer discomfort and fatigue, if the period of exposure is long; this is due to the **vergence-accommodation conflict** associated to the dissociation between the converging (in the space) and focal (on the screen) distances [13].

Auto-stereoscopic: These displays are often described as *glass-free* 3D display systems and can be divided into four classes, notably multiview, volumetric, holographic and light field displays [14]:

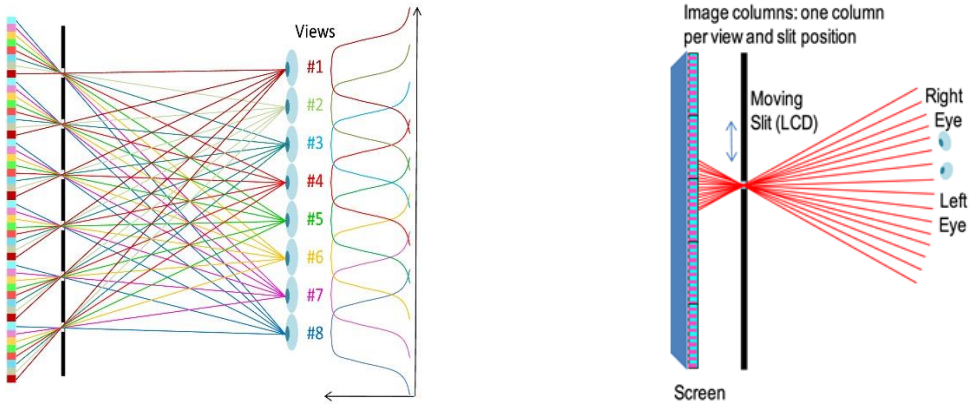
- ***Multiview 3D displays*:** This type of displays may explore different optical effects such as diffraction, refraction, reflection and occlusion to build *parallax panoramagrams*. Parallax panoramagrams are associated to an autostereoscopic viewing method where each view is spatially multiplexed, e.g. the left-eye and right-eye views are divided into narrow juxtaposed strips and viewed through a superimposed ruled or lenticular screen in such a way that each of the observer's eyes is able to see only the correct views. The different ways the spatial multiplexing is performed distinguish different types of autostereoscopic displays; an example is shown in Figure 2.1 a), where different views are displayed spatially multiplexed (view 1 up to view 5). Within multiview 3D displays, it is also possible to refer to *super multiview*, multiview with eye tracking and directional backlight display systems. Nowadays, the most commonly used systems are those designed based on the refraction and occlusion approaches [12] [14] which are briefly described in the following:
  - **Refraction-based:** This approach uses different techniques to produce the panoramagrams, notably the most commons are integral and lenticular imaging displays. The *lenticular imaging* solution, shown in Figure 2.1 a), uses a cylindrical lenslet to grant horizontal stereo and motion parallax [14]. On the other hand, the *integral imaging* solution, illustrated in Figure 2.1 b), uses a spherical lenslet to provide both vertical and horizontal spatial parallax.



a) **b)**

Figure 2.1 – a) Lenticular imaging display; b) Integral imaging display [14].

- **Occlusion-based:** This approach uses different techniques to grant the 3D multiview effect, notably the most common are barrier-grid and moving-slit solutions. The *barrier-grid* solution, illustrated in Figure 2.2 a), uses an aperture mask in front of a raster display to mask singular screen sections that are not meant to be displayed for a certain point of view. On the other hand, the *moving-slit* solution, shown in Figure 2.2 b), uses a moving slit in front of a high-speed display where different views are displayed depending on the slit position [14].
- ***Super Multiview*:** This 3D displays use a very high number of views (*e.g.* 256 views) to smooth or even supress the discontinuity between views as well as the vergence-accommodation conflict [15]. For smooth parallax, the interval between the displayed views should be smaller than a human pupil diameter, to harvest simultaneously the light from (at least) a pair of views. These effects help building smoother stereo and motion parallaxes and may also reduce the vergence-accommodation conflict [16]. This type of displays explores the same optical effects as those described above, therefore they are just a particular case of multiview displays with a higher density of views.



a) **b)**

Figure 2.2 – a) Barrier-grid display; b) Moving-slit display (adapted from [14]).

- ***Volumetric Displays*:** The basic principle behind this type of displays is to represent a *voxel* within a well-defined space volume. For example, this volume can be filled with a fluorescent gas that when excited produces light (*Static-Volume display*), or with a rotating or linear panel which is able to produce a similar effect (*Swept-Volume display*) [17].

- ***Holographic Displays*:** In principle, these displays should reproduce the light wavefront [18] of the original object/scene. To produce the desired effect, the light is diffracted from the microscopic interference fringes stored in the hologram onto the holographic surface [19] [20] [21] [22] [23] [24] [25] [26] [27]. The holographic data itself is not a conventional image where luminance and chrominances are measured but rather an interference pattern between two wave fields: the *reference wave field* and the *object wave field* corresponding to the diffraction of the reference wave field by an object.
- ***Light field Displays*:** Similar to holographic displays, light field displays should reproduce the scene light field previously captured although using different physical methods. Instead of using holographic data and methods, this emerging technology uses different techniques such as tomographic, multi-layer and directional backlighting [28] [29] [30] [31].

After reviewing the basic concepts on 3D perception and systems, the next sections will specifically address multiview video coding.

2.2. High Efficiency Video Coding Standard: Brief Review

The HEVC standard, also known as ITU-T H.265 or MPEG-H part 2(ISO/IEC 23008-2), is a video coding standard that offers 50% bitrate reduction for the same perceptual video quality, with respect to its predecessor, the join ITU-T and MPEG, H.264/AVC standard. This major improvement opens two distinct evolution paths: i) lower bitrates for same target quality, *e.g.* for mobile services; ii) higher quality/resolution with lower rates, *e.g.* for ultra-high-definition content (UHD), namely UHD 4K and 8K, and high dynamic range (HDR) with more than 8-bit per sample (HDR is addresses in one of the HEVC standard amendments) [32]. The approach followed in the standard design and development was to propose a new set of rather efficient tools while still maintaining the usual predictive block-based architecture. As previous standards, HEVC only defines normatively the bitstream syntax and semantics, and the decoding process; the encoding process is non-normative, meaning that the decision of whether using or not specific tools and with which parameters it is the encoders' entire responsibility, as well the final achieved quality-rate performance. In addition, new parallel processing tools were developed to increase the codec throughput, such as wavefront parallel processing, overlapped wavefront and tiles [33] [34].

2.2.1. Architecture

In terms of architectural design, the HEVC architecture is largely based in the previous hybrid coding standards, as the encoder architecture shown in Figure 2.3 illustrates. Although the HEVC encoder architecture has no substantial changes with respect to previous standards, each block features new, more efficient tools to more efficiently exploit the temporal, spatial and statistical redundancies. The following subsections present a brief description of the new coding tools introduced by HEVC, while grouping them in a meaningful way.

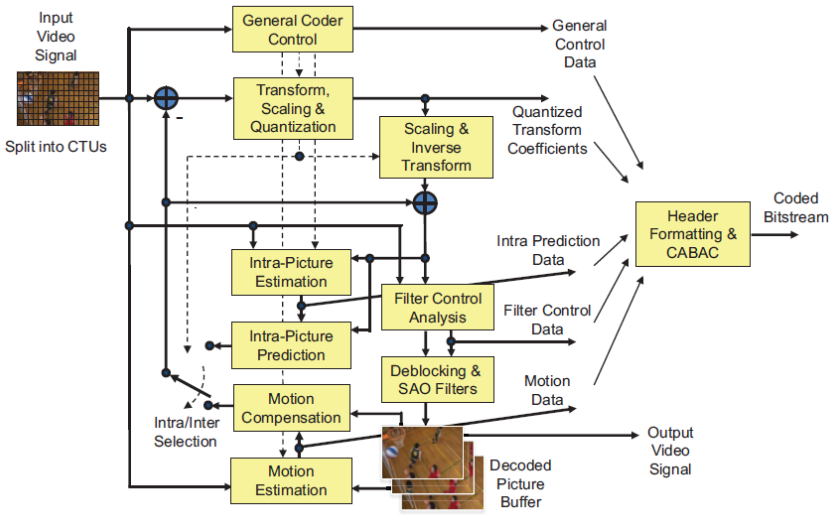


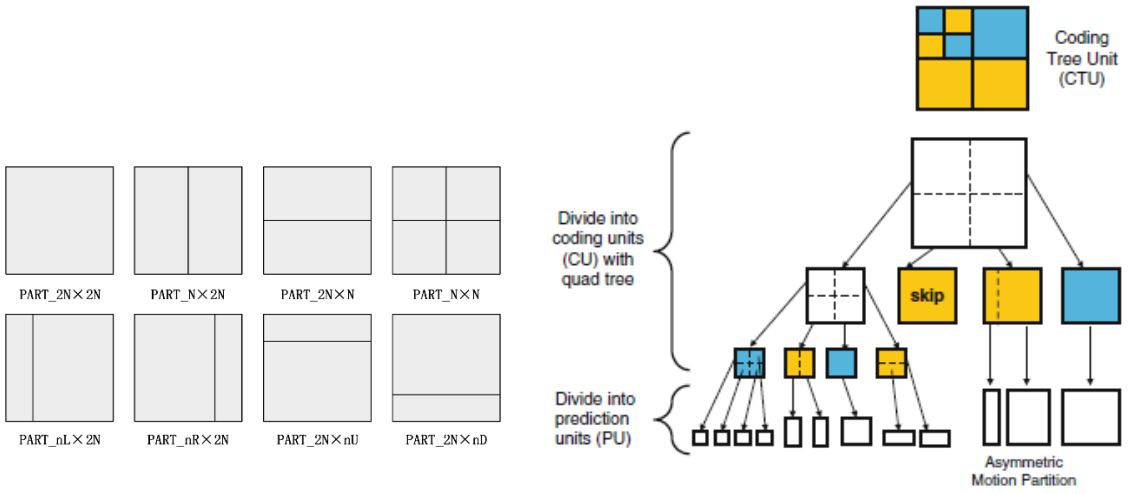
Figure 2.3 – HEVC hybrid encoding architecture [32].

2.2.2. Block Partitioning

As usual, the coding structure will partition a frame into slices. Each slice is composed by a sequence of coding tree units (CTUs), which can be independently decoded from other slices within the same frame, enhancing resilience by enabling re-synchronization due to data loses. The CTUs inside each slice are processed in a raster scan order. For the first time in a video coding standard, the concept of macroblock is not used.

The HEVC standard introduces a new block partitioning structure, notably coding trees, coding units and coding blocks. As represented in Figure 2.4 b), the hierarchical partitioning structure has the CTU as the basic processing unit, which may contain one coding unit (CU) or be partitioned to form multiple CUs. Each CU has one or several partitioned prediction units (PUs) and also a tree of transform units (TUs). While the PU is a data structure associated to the prediction modes (either Inter or Intra prediction) and motion compensation, the TU is the data structure associated to the type and size of the used integer basis function transform and the quantized coefficients.

A CTU has an $L_{CTU} \times L_{CTU}$ resolution, where $L_{CTU} \in \{16, 32, 64\}$ and can be partitioned into N CUs, with size of $L_{CU} \times L_{CU}$, where $L_{CU} \in \{8, 16, 32, 64\} \cap L_{CU} \leq L_{CTU}$. PUs may use different partition modes as illustrated in Figure 2.4 a), depending on the size and selected coding mode. This dependency is described later in this section. TUs may use two different integer basis function transforms, namely the discrete cosine transform (DCT) and the discrete sine transform (DST); TUs support four transform sizes, notably 4×4, 8×8, 16×16 and 32×32. As will be mentioned later, the DST is only used for Intra predictions with a 4×4 TU size [33]. Each set of these logical units (CTU, CU, PU, TU) have their own block elements, respectively, coding tree block (CTB), coding block (CB), prediction block (PB) and transform block (TB) for each of the (typically 3) colour components. For example, a CTU aggregates three CTBs, one for the luminance samples, and a pair for the chrominance samples, and associated syntax elements. This relationship is reproduced in the remaining logical units and their blocks.



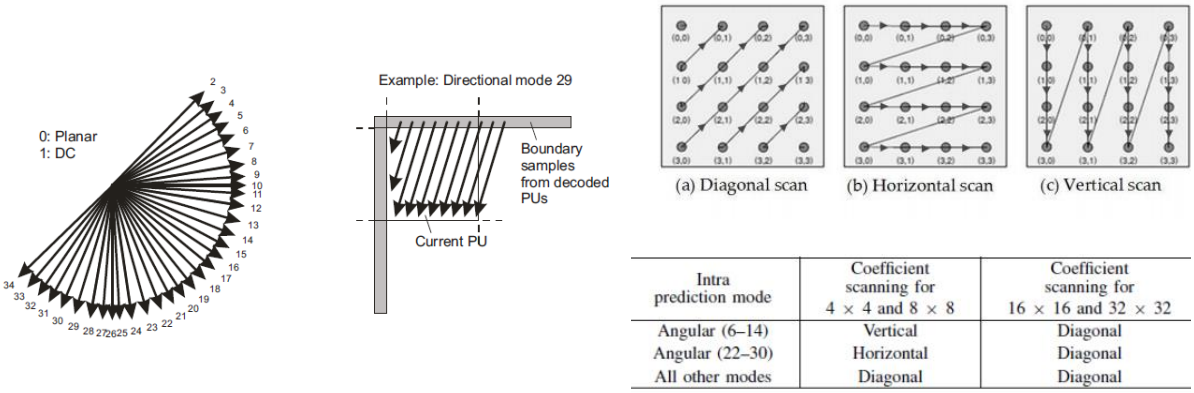
a) b)

Figure 2.4 – a) CU partitioning units; b) Hierarchical block coding structure [33].

2.2.3. Intra Coding Tools

HEVC Intra coding and the associated Intra prediction tools are used to exploit the spatial redundancy within a frame, more precisely at the CU level. HEVC includes some new Intra coding tools, notably regarding the previous H.264/AVC standard, presented below:

- ***Additional Intra prediction modes*:** With the Intra prediction modes, spatial prediction is performed, *i.e.* the decoded boundary samples of adjacent blocks are used as reference data for the prediction of the block under coding. In HEVC, 35 prediction modes are defined (from up to 9 in H.264/AVC): planar, DC and 33 angular modes, as illustrated in Figure 2.5 a). The planar mode (Mode 0) provides good predictions in areas of smoothly-varying structures, the DC mode (Mode 1) offers fine predictions for large areas of nearly constant or slowly varying smooth regions, and the angular modes (Mode 2 to 34) provide high-fidelity predictions for areas with directional structures. Additionally, reference samples might be filtered by a 3-tap $[1\ 2\ 1]/4$ smoothing filter while boundary values are filtered by a 2-tap $[3\ 1]/4$ filter [32].
- ***Intra prediction mode-dependent coefficient scanning*:** Contrary to the past, the transform coefficients scanning order depends on two factors, the Intra prediction mode and the TU block size, as shown in Figure 2.5 b). This feature was adopted in the standard since these coefficients are even more residual/noise data (and not *samples*) than in the past because of the more efficient prediction modes.



Intra prediction mode	Coefficient scanning for 4 × 4 and 8 × 8	Coefficient scanning for 16 × 16 and 32 × 32
Angular (6–14)	Vertical	Diagonal
Angular (22–30)	Horizontal	Diagonal
All other modes	Diagonal	Diagonal

a) b)

Figure 2.5 – a) Intra prediction modes; b) Intra prediction mode-dependent coefficient scanning order [33].

2.2.4. Inter Coding Tools

HEVC Inter coding and the associated prediction tools exploit the temporal frame correlation, often using motion compensated prediction (MCP). MCP encodes the motion vector difference, which is based on the difference between the estimated MVs and the motion vectors predictors obtained by motion vector prediction (MVP). The HEVC technical novelties regarding Inter coding are:

- ***New partition modes*:** As illustrated in Figure 2.4 a), HEVC allows a new set of Inter partitioning modes associated to asymmetric mode partitioning (AMP). This tool offers four new asymmetrical partitioning modes to be applied to the PUs, providing some efficiency improvements if properly selected by the encoder [35].
- ***Fractional sample interpolation*:** Motion vectors (MVs) can be defined with half-pel and quarter-pel accuracy for luminance samples and one-eighth-pel accuracy for chrominance samples. No intermediate rounding operations are used in the motion compensation interpolation process which allows improved precision while also simplifying the interpolation process which is also performed at the decoder [36].
- ***Motion vector prediction*:** Motion vector predictors derive all motion data for a PU block from neighbouring blocks. This process is based on a candidate list in which the predictors depend on the selected mode. This list is filled with five spatial candidates and two temporal collocated candidates temporally and spatially collocated blocks. MVP is performed once for each MV, thus once or twice for uni- or bidirectional PU, respectively. Motion vector predictors might be obtained by one of the following modes:
  - **Normal mode:** This mode is also referred as advanced motion vector prediction (AMVP) and uses up to two spatial candidates and one temporal candidate. The selected candidate and the residue obtained are coded in the bitstream.
  - **Merge mode:** This mode uses up to four spatial candidates and one temporal to predict the motion data; the MVs are inferred so that only residue data is coded in the bitstream.
  - **Skip mode:** This mode uses also up to four spatial and one temporal candidate per candidate list as the merge mode but now no residual data is sent.
- ***Weighted prediction*** – This tool is very useful to deal with fading sequences by adding a multiplicative and an additive offset factor on the computation of the motion compensated prediction.

The reason for using a specific tool to deal with global illumination changes results from the poor efficiency obtained when using only motion compensation in such patterns [37].

2.2.5. Transform and Quantization

As mentioned earlier, HEVC transforms include a DCT and a DST. The DST transform is only used for 4×4 TUs with Intra coding. This choice was made because the DST is able to model better the residual data in this specific case and it is not more computationally demanding. For all the other cases, a DCT is applied to the residual data block, *i.e.* with size $L \times L$ where $L \in \{4, 8, 16, 32\}$. Experimental data has shown that the DST provides a bitrate reduction of approximately 1%, only for 4×4 Intra coded blocks, while it only achieves marginal gains for other sizes and modes.

The HEVC quantization process is similar to H.264/AVC, *i.e.* it uses a uniform reconstruction quantization (URQ) scheme controlled by a quantization parameter (Q_P). Q_P is used to determine the quantization step (Q_{Step}); for 8-bit video sequences, the relationship between Q_P and Q_{Step} may be adjusted differentially at CU level [33].

2.2.6. Entropy Coding

As H.264/AVC, HEVC uses an entropy coding tool based on context-based adaptive binary arithmetic coding (CABAC). For HEVC, some improvements were made in terms of throughput, complexity and context memory requirements [33] [38].

2.2.7. In-loop Filtering

HEVC defines two in-loop filters, namely the deblocking filter (DBF) and the sample adaptive offset filter (SAO). Both filters are used in the encoding and decoding loop, before the entropy encoder and after the inverse quantization, respectively.

- ***Deblocking filter*** – This is the first filter used in the prediction loop and is able to reduce the block artefacts by attenuating discontinuities at the PB and TB boundaries. The DBF classifies the boundaries in three levels of boundary strength (Bs). Only boundaries with a Bs value of '1' and '2' are filtered to avoid oversmoothing.
- ***Sample adaptive offset*** – This filter improves the subjective quality of the reconstructed frames by attenuating the ringing and banding artefacts while increasing edge sharpness. The SAO filter classifies the samples as: i) minimum; ii) maximum; iii) edge with the sample having the lower value; iv) edge with the sample having the higher value; and v) monotonic [39]. Edge offset (EO) classification is based on the neighbouring samples and band offset (BO) for a given sample value.

2.2.8. Performance

Experimental tests have proven that HEVC can achieve approximately an average of 50% bitrate savings for similar subjective quality in comparison with the previous H.264/AVC standard [40]. These tests were conducted for two types of application scenarios, namely interactive and entertainment application scenarios. Figure 2.6 shows the performance results for the interactive scenarios, notably video conferencing, in terms of rate-distortion (RD) performance and bitrate savings regarding previous

relevant standards for two test sequences. On average, the results show a 40,3% bitrate saving when comparing the HEVC Main profile with the H.264/AVC High profile.

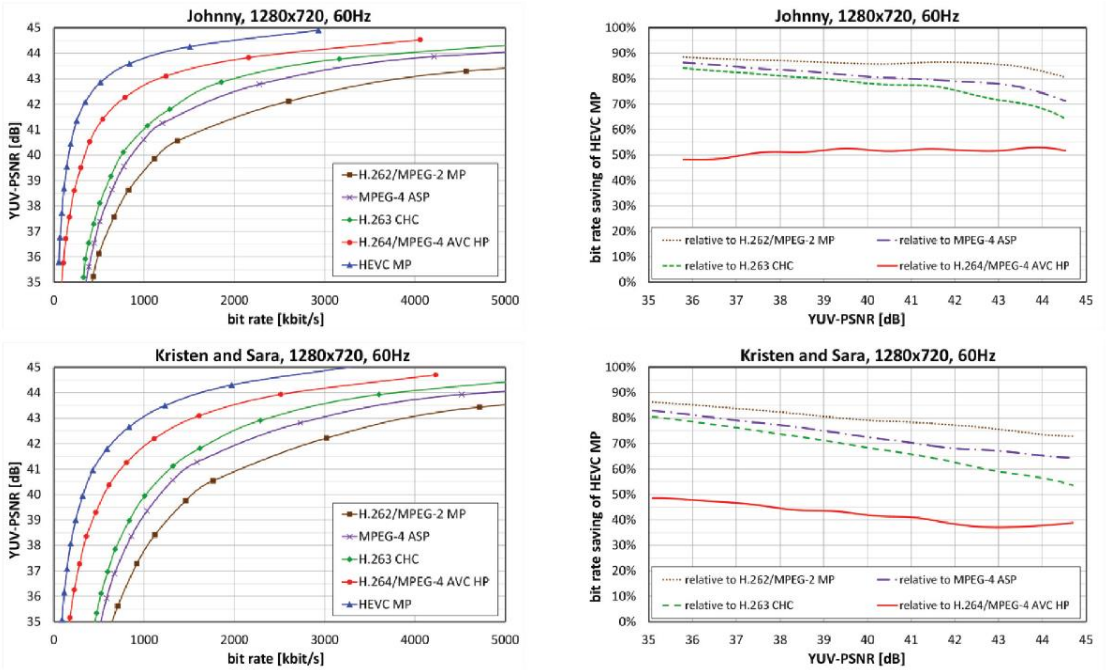


Figure 2.6 – Interactive scenario HEVC performance results [40].

For the media entertainment scenario, the results show an average 35,4% bitrate saving again when comparing the HEVC Main profile with the H.264/AVC High profile.

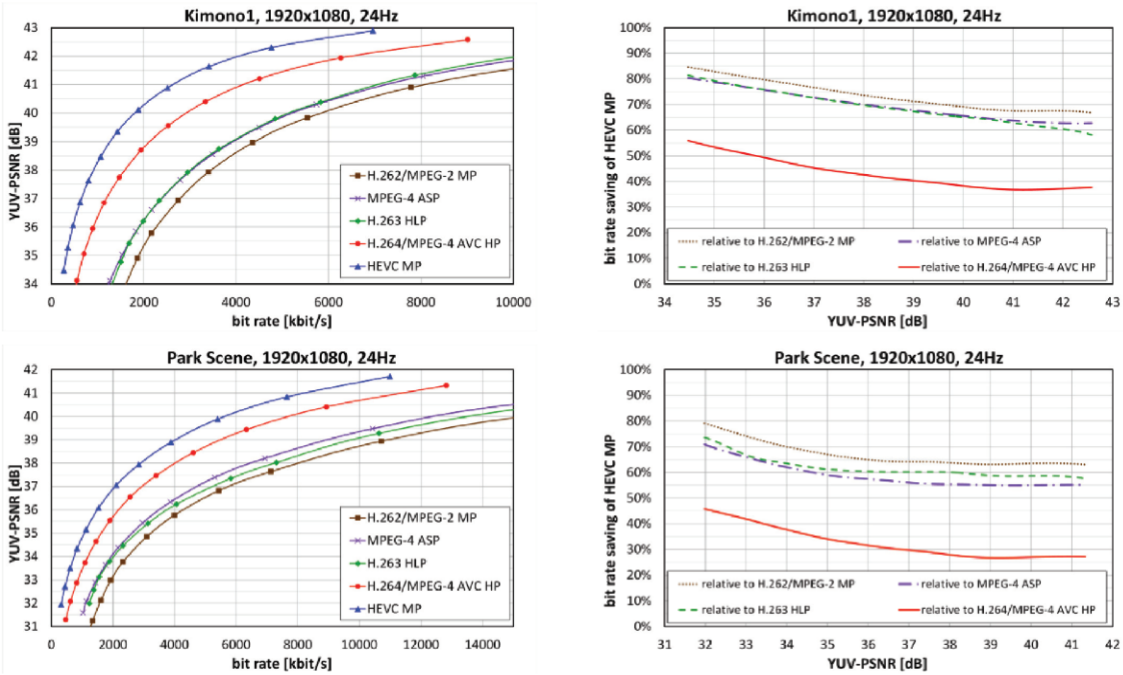


Figure 2.7 – Entertainment scenario HEVC performance results [40].

Additionally, subjective quality assessment tests were conducted using the test setup referenced in [40]. The obtained mean opinion score results comparing the HEVC Main profile with the H.264/AVC High profile are shown in Figure 2.8 and Figure 2.9 and allow concluding that there are significant bitrate savings for the same perceptual quality, notably around 50%.

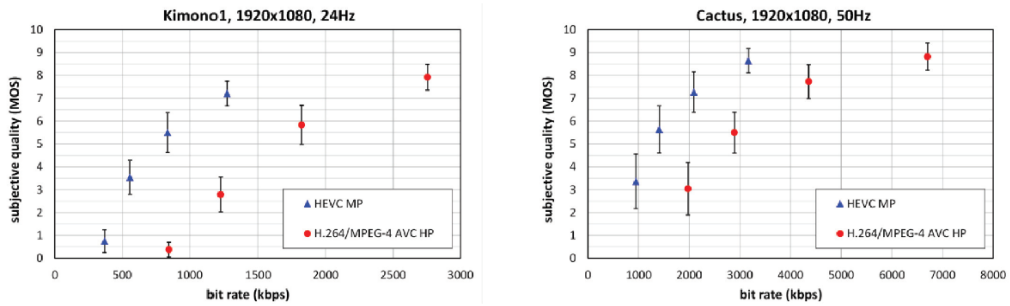


Figure 2.8 – Subjective test results: mean opinion score versus bitrate [40].

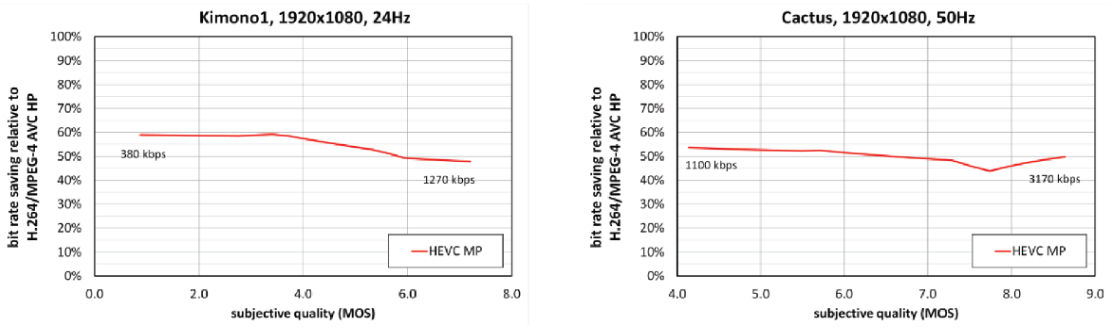


Figure 2.9 – Subjective test results: bitrate savings versus mean opinion score [40].

2.3. Multiview Video Coding: Brief Evolution Review

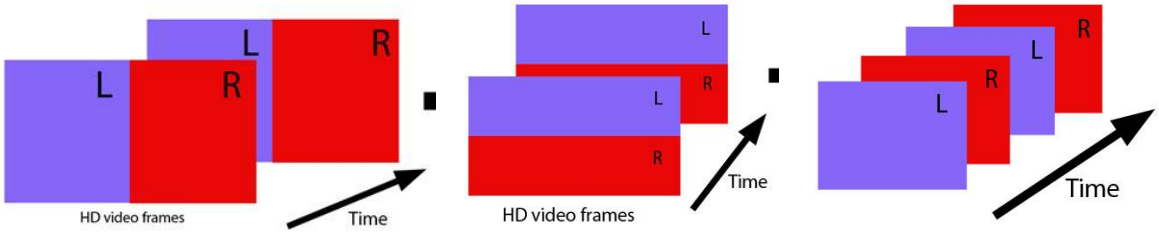
Multiview video coding regards the coding process of multiple views from the same scene, thus typically exhibiting high inter-view correlation. In principle, multiview video representation formats may be divided into two major classes depending on the type of data considered: only texture, this means only RGB or YUV components, or texture plus depth where depth data is used together with the texture data. Within each of these classes, different types of coding solutions have been defined and eventually deployed in practice as listed below.

2.3.1. Texture based Multiview Video Coding Formats

This type of formats uses only texture data to create the final 3D experiences; several coding solutions may be adopted depending on the critical application requirements, notably:

- ***Multiview Simulcasting*:** This approach is conceptually very simple as each view is independently coded using any regular video codec, *e.g.* the MPEG-2 Video, H.264/AVC, and High Efficiency Video Coding (HEVC) standards; therefore, this format does not exploit the redundancy between views and thus this is not a very rate efficient solution.
- ***Frame Compatible Stereo*:** This approach considers only a pair of views, which left and right frames are combined together into a regular single view video signal to be coded with any regular video codec. The combination/multiplexing of the frames from the two views into a single frame may be performed in many ways, notably:
  - **Spatial Multiplexing:** As illustrated in Figure 2.10 a) and b), where the frames of the two stereo views are spatially combined into a single frame either side by side or top and bottom to be encoded after with a regular video codec; in this case, the vertical or horizontal spatial resolutions of the stereo frames need to be halved to fit into a single frame;

- **Time Multiplexing**, as illustrated in Figure 2.10 c), where the frames of the two stereo views are temporally multiplexed leading to a single video signal with doubled frame rate (assuming that half the rate for each view is not acceptable for many applications).



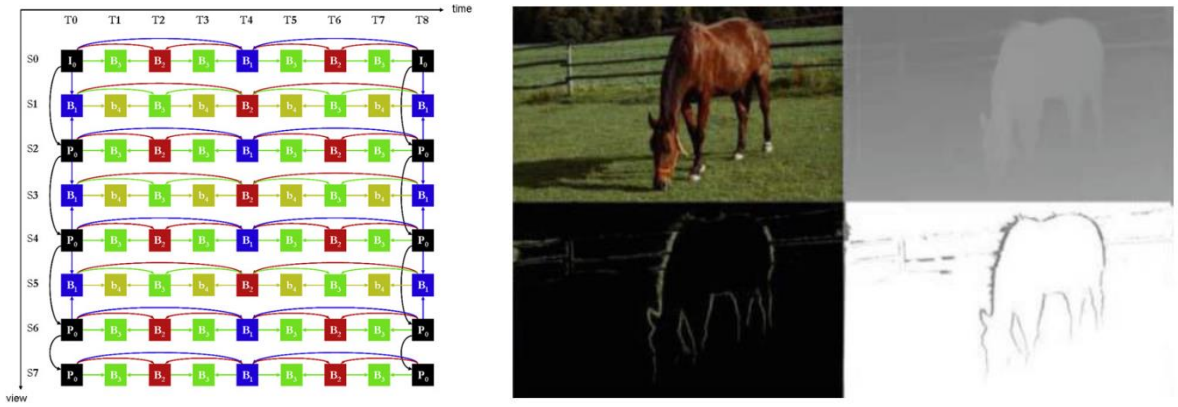
a) **b)** **c)**

Figure 2.10 – a) Spatial Multiplexing: Side-by-side; b) Spatial Multiplexing: Top-Bottom; c) Temporal Multiplexing.

- ***Multiview Video Coding*:** This coding approach exploits the inter-view correlation to reduce the total bitrate; it may be applied to only a stereo pair or tens of views. In terms of standards, this approach was first adopted in the MPEG-2 Video standard, around 1996, which created a specific profile for multiview video coding. Later, this same approach was adopted in the Multiview Video Coding (MVC) standard, a backward compatible extension of the H.264/AVC standard (November 2009) and the MV-HEVC standard (Multiview Video – High Efficiency Video Coding), a backward compatible extension of the HEVC standard (October 2014) [41]; a typical temporal plus inter-view prediction structure is shown in Figure 2.11 a).

2.3.2. Texture plus Depth based Multiview Video Coding Formats

The texture plus depth format differs from the texture-only format by using one more data component, this means a depth map representing the distance from the camera plane to each object in the scene for each view. This depth map can be produced in different ways, e.g. 3D scanners, disparity processing of the texture views, etc. As shown in Figure 2.11 b), the depth data has different characteristics from the texture data, typically large smooth areas separated by sharp edges. There are several ways to encode the texture plus depth data format, notably:



a) **b)**

Figure 2.11 – a) MVC typical temporal plus inter-view prediction structure; b) Depth Enhancement Multiview Coding: Frame (top-left), Depth (top-right), occlusion layers (bottom) [42].

- ***Multiview Video plus Depth* with same coding solution:** This approach codes independently the sequence of texture and depth frames using the same coding solution, e.g. the MVC or MV-HEVC standards. Thus, this type of solution does not typically consider the specific characteristics of the depth data as the depth is coded as one more 'texture component'; naturally, also the texture-depth inter-component correlation is not exploited.
- ***Multiview Video plus Depth* with different coding solutions:** This approach adopts a different coding solution for the texture and depth data to consider the specific characteristics of each data type. The 3D-HEVC standard has adopted this approach by adding to the texture coding modes a set of depth specific coding modes. Moreover, 3D-HEVC also exploits the texture-depth inter-component correlation to reduce the overall bitrate.

2.4. Multiview High Efficiency Video Coding Standard: Brief Overview

The Multiview High Efficiency Video Coding (MV-HEVC) standard is an extension of the HEVC standard approved in October 2014 with the target to efficiently code several views of the same scene by exploiting the inter-view redundancy. MV-HEVC is a rather short specification as it only contains changes to HEVC at the high-level syntax (HLS) since it fully reuses the complete set of HEVC tools at CU level, thus preserving all HEVC block-level coding processes. The changes regard a multilayer coding design where each view is represented in the bitstream as a multiplexed layer. To achieve higher compression, MV-HEVC follows the same inter-view prediction design as for MVC as shown in the architecture illustrated in Figure 2.12. In practice, each view is dependently coded to some previous view(s) using the Inter-view prediction tools (denoted by the red arrows) which are basically the same as the Inter-frame prediction tools; to guarantee backward compatibility with HEVC, there is a HEVC coded base view. Inter-view prediction is, in fact, the same as Inter-frame prediction tampered in a way that instead of including frame references in time it also includes frame references from neighbouring views, as shown in Figure 2.12, thus generating the so-called disparity motion vectors (DMVs) [43]. While MV-HEVC is focused on texture coding only, the second HEVC 3D extension to be reviewed in the following has adopted a different approach where both texture and depth are coded.

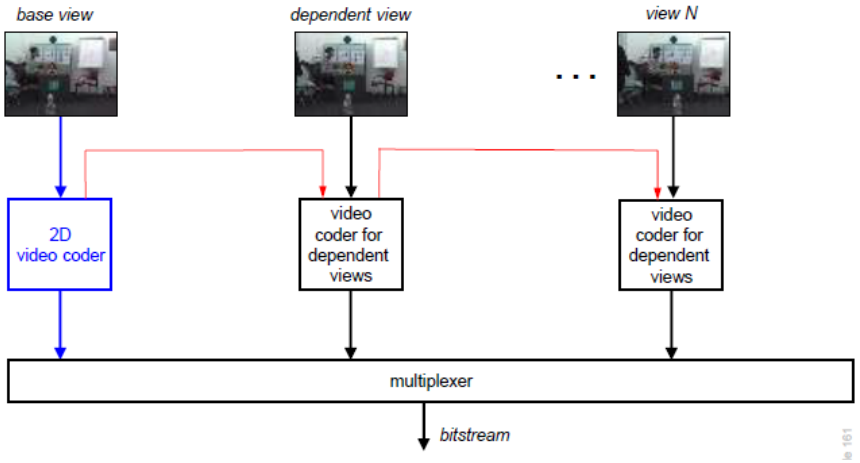


Figure 2.12 – MV-HEVC encoder architecture [44].

2.5. 3D-HEVC Coding Standard: Brief Review

The 3D High Efficiency Video Coding (3D-HEVC) standard [45] was approved in 2015 as a HEVC backward compatible extension where, differently from the MV-HEVC standard, both texture and depth are coded for each available view. While the same multilayer coding design as for MV-HEVC has been adopted, each layer includes now not only a texture component per view but also a depth component. The depth provides geometry information about the scene in order non-coded views may be synthesized at the receiver side based on the decoded texture and depth views using a so-called Depth Image based Rendering (DIBR) techniques. Since 3D-HEVC considers a depth component which has rather distinct signal characteristics in comparison with texture, notably smooth areas separated by sharp edges, a new set of coding tools has been specified to efficiently address these characteristics; moreover, also the inter-component dependencies between both components may be exploited for the first time.

2.5.1. Architecture

Figure 2.13 shows the 3D-HEVC encoder architecture where the main novelties are the coding of depth maps and the exploitation of the inter-component redundancies. To achieve a higher compression factor, new coding tools have been specific for the depth component to exploit depth properties that are not present in the texture components. Additionally, since depth is naturally highly correlated to its corresponding texture components, depth coding may be performed exploiting its redundancy with the texture components as indicated by the red solid arrows in Figure 2.13. Moreover, there are also some dashed red arrows in Figure 2.13 corresponding to the possibility to use DIBR predictions, this means view synthesis, to more efficiently code some views based on adjacent decoded views. It is important to stress that HEVC backward compatibility is provided by coding the first view with pure HEVC, this means independently from the other views.

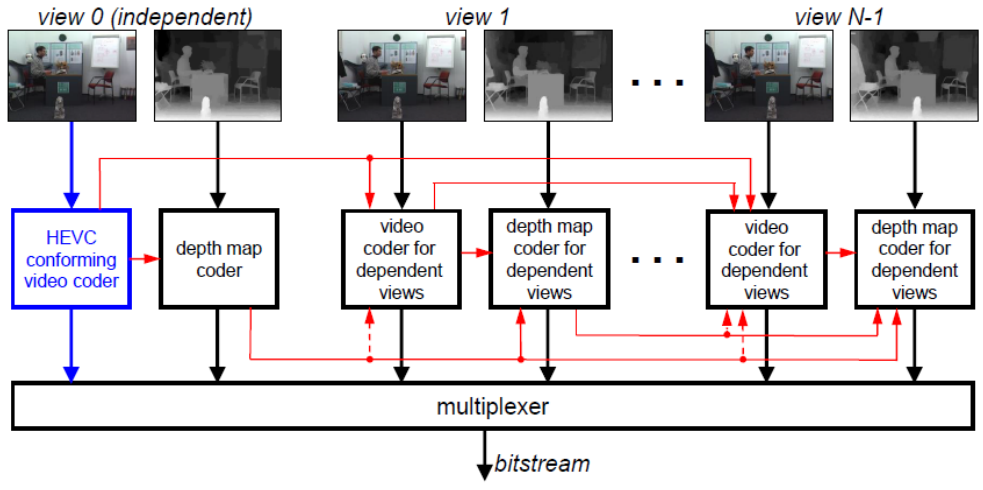


Figure 2.13 – 3D-HEVC encoder architecture [46].

2.5.2. Texture Data Coding

The same concepts and coding tools used to code independent views, also referred as base view, are used to code dependent views; notably, independent views are coded using an unmodified HEVC codec. On the other hand, to increase the coding efficiency of the dependent views which do not have

to be compliant to any previous standard, notably HEVC, new coding tools have been additionally specified, particularly to better exploit the redundancy with previously coded neighbouring views as well as the new added component, notably depth maps. This subsection describes the additional coding tools for the dependent views, particularly disparity compensated prediction and how disparity vectors are derived, Inter-view motion prediction and Inter-view residual prediction.

I. Disparity Compensated Prediction

As discussed previously, there is evident redundancy between views, thus disparity estimation is a key technique to exploit this redundancy, resulting into disparity vectors. Disparity vectors between views, also named as DMVs may be predicted through disparity compensated prediction (DCP) to improve their coding efficiency. DCP works the same way as MVP (see Section 2.2-2.2.4), where instead of using the same view at different access units (this means different time instants) to derive temporal MVs (TMVs), other views within the same access unit are used to derive DMVs.

II. Disparity Vectors Derivation

The Inter-view coding tools rely on Inter-view prediction to efficiently code a dependent view in terms of sample values, motion information, predicted residue and partitioning scheme. Therefore, a couple new techniques were standardized to enable the predicted disparity information (PDI) to be standardly derived at the decoder instead of sending it directly thru the bitstream. The PDI, referring to a predicted disparity vector (PDV) and a reference view (RV), is used to identify the spatial displacement and the view wherefrom the PDV was derived, respectively.

The first technique, named as **neighbouring block disparity vector** (NBDV), relies only in texture components to derive the disparity vectors of a dependent view. Notably, NBDV operates at the texture component CU level to derive the PDI from motion information of spatially and temporally neighbouring blocks [47]. The second technique, named as **depth oriented neighbouring block based disparity vector** (DoNBDV) or in some literatures referred as depth refinement tool, use the reference view decoded depth map (if it is available) to enhance the disparity vectors accuracy. The refined disparity vector is derived from the maximum of four corner sample values of the depth block [43] [48].

III. Motion Prediction

Considering that a scene is captured from different viewpoints with slightly different positions, it is possible to understand that when an object moves through a scene, the movement captured from a specific viewpoint shows a similar behaviour for other viewpoints. This implies that the MVs for the various different views are highly correlated and thus this correlation may be exploited to increase the motion coding efficiency. To exploit the Inter-view correlation, 3D-HEVC has specified an extended candidate list for the Merge mode, namely: i) texture; ii) Inter-view; iii) disparity information; and iv) view synthesis prediction [43]:

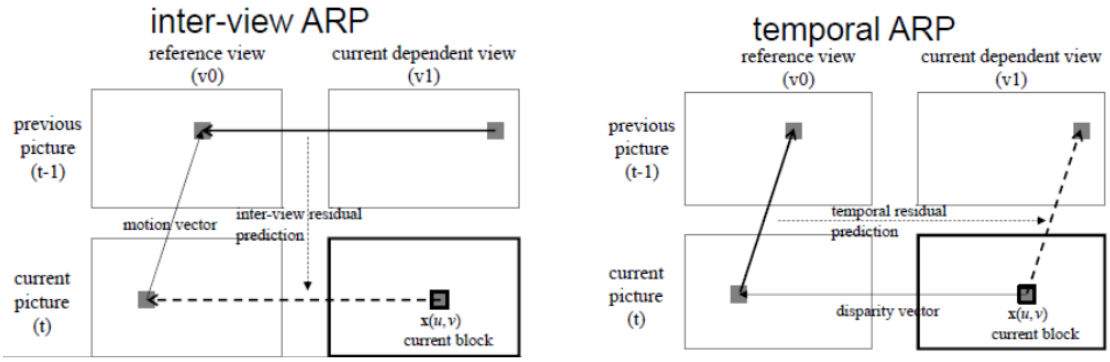
- ***Texture candidate*:** MV derived from the texture components of the same view and same access unit.
- ***Inter-view candidate*:** Inherited MVs from a frame within the same access unit but different view.
- ***Disparity information candidate*:** MVs predicted from a disparity block through PDI.

- ***View synthesis prediction candidate*:** Using the PDI, the corresponding depth block on another view is fetched to estimate the depth block on the current view, from which the MVs are derived accordingly to the camera parameters.

IV. Inter-view Residual Prediction

Using the same logic as above, for similar positions in various viewpoints of a scene, similar residues may be expected. This motivated the development of tools to predict such residues based on relevant neighbouring residues to further improve their compression performance, notably:

- ***Advanced residual prediction*:** As Figure 2.14 illustrates, the advanced residual prediction (ARP) uses the MV of the current PB to be coded to predict its residue from a coded residue block in one of two ways [43]:
  - **Inter-view ARP:** Motion compensated residue is predicted from different views;
  - **Temporal ARP:** Disparity compensated residue is predicted from different access units, this means different time instants.
- ***Illumination compensation*:** Illumination compensation is performed by using a weighted Inter-view prediction computed by means of a scale factor and an offset, similarly to the weighed prediction tool in HEVC Inter coding [43].
- ***Depth-based block partitioning*:** The texture partitioning is predicted based on the corresponding depth map partitioning, thus improving the compression of dependent texture [43] [49].



a) b)

Figure 2.14 – a) Inter-view residual prediction; b) Temporal residual prediction [44].

2.5.3. Depth Data Coding

For depth coding, the same concepts and tools as used for the texture components are applied. However, due to the depth characteristics, new specific depth coding tools have been included in 3D-HEVC. In terms of quadtree structure, depth coding units cannot be further split regarding its collocated texture block corresponding to a new coding tool named quadtree limitation (QTL).

I. Intra Coding Tools

As mentioned above, depth data presents some distinct features when compared with texture, thus justifying the specification of additional depth data coding tools, notably:

- ***Additional Intra prediction modes*:** HEVC Intra prediction tools lead to coding artefacts at sharp edges, causing undesirable strongly visible artefacts in synthesized views [50]. Therefore, to mitigate these problems, new Intra coding prediction modes have been specified, notably:
  - **Intra_Wedge:** (Mode 35) This new prediction mode provides a good prediction solution for linear sharp edged blocks. As illustrated in Figure 2.15, a PB is partitioned in two sub-block partitions (SBPs) by explicitly signalling an index value, wherein this value refers to a set of binary patterns named *wedgelets*. Subsequently, the predicted depth value for each SBP is derived through DC type prediction from a subset of decoded neighbouring sample values [43].
  - **Intra_Contour:** (Mode 36) This mode is a worthy prediction solution for little depth variation elements where depth is not so largely constant; this behaviour is commonly found in natural scenes, *e.g.* a garden. This mode divides the PB into two SBPs which are predicted based on the texture component (same view) and a threshold value, as shown in Figure 2.15. Then, these SBPs have their depth values predicted through the same process as for the Intra_Wedge mode, notably DC type prediction [43].
  - **Intra_Single:** (Mode 37) When using this mode, the Intra prediction is computed using a single neighbouring boundary sample value at position $\left({}^{N}/_{2}, 0\right)$ or $\left(0, {}^{N}/_{2}\right)$ of the PU, where N stands for PU size. This prediction mode is highly efficient for large homogeneous areas [43].

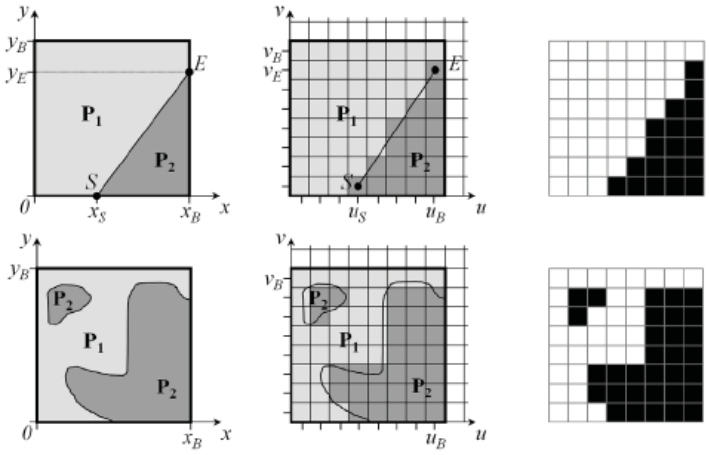


Figure 2.15 **–** Wedge (top) and Contour (bottom) Intra prediction modes [50]**.**

- ***New Skip mode*:** This tool is applied at CU-level, and may only be used together with the Intra_Single prediction mode. This Depth Intra Skip (DIS) mode is used to skip redundant information between two PBs predicted with the Intra_Single mode, thus resulting only in CU level syntax elements.

II. Disparity Compensated Prediction

As the depth and texture components express similar disparity behaviours, the same DCP may naturally be applied for all components. Still, the predicted disparity vector might be enhanced by using the so-called **depth refinement** tool, which through NBDV identifies the appropriate depth block from a disparity view to derive a refined disparity vector. The refined disparity vector is derived from the maximum of four corner sample values of the depth block [43] [48].

III. Inter-view Residual Prediction

Inter-view residual prediction for depth maps is trickier as fractional sample interpolation at sharp edges may cause ringing artefacts thus leading to strong and unpleasant visible artefacts, and thus poor QoE, when synthesizing a view. As a consequence, 3D-HEVC only supports full sample motion accuracy for the depth component Inter-view residual prediction [43] [50].

IV. Residual Coding Tools

Experience has shown that the depth maps high frequency components can be irrelevant when compared with the DC component for view synthesis. This fact justified the specification of a DC-only mode which explicitly signals a DC offset in addition to the quantized transform coefficients.
Also, a so-called **depth lookup table** (DLT) technique was added, wherein residual values are mapped to a lookup table; the residual values may be coded by signalling only the index for the DLT. The values stored in the DLT are using a picture parameter set (PPS) NAL (Network Abstraction Layer) unit. This enables the encoder to predict depth values adaptively, notably by refreshing the DLT and send only some indexes. This is useful because depth maps exhibit a highly constant DC variation in time, *i.e.* the depth value range is often only sparsely used [43] [51] [52].

2.5.4. Performance

To assess the 3D-HEVC coding performance gains regarding relevant alternative coding solutions, some tests were conducted for eight video sequences [50]. Table 2.1 shows that the average 3D-HEVC bitrate reductions over MV-HEVC and HEVC Simulcasting measured with BD-Rate are around 21% and 45%, respectively.

Table 2.1 – 3D-HEVC average bitrate savings (BD-Rate).

Sequence	Overall PSNR vs total bitrate, (original and synthesized positions)		
	MV-HEVC vs Simulcast	3D-HEVC vs Simulcast	3D-HEVC vs MV-HEVC
Balloons	−24%	−39%	−20%
Kendo	−21%	−38%	−22%
Newspaper	−26%	−39%	−19%
GT_Fly	−40%	−54%	−24%
Poznan_Hall	−24%	−38%	−19%
Poznan_Street	−33%	−41%	−13%
Undo_Dancer	−36%	−50%	−21%
Shark	−40%	−58%	−30%
Average	**−40%**	**−45%**	**−21%**

Additionally, subjective quality assessment tests were conducted using the same test setup described in [50]. The results shown in Figure 2.16 relate the obtained mean opinion scores (MOS) for the eight test sequences at four different bitrates R1-R4 which evaluate a 2-view scenario on a stereoscopic display (SD) and a 3-view scenario on an auto-stereoscopic 28-view display (ASD). The results show that multiview coding solutions improve greatly the coding efficiency, consequently improving MOS for

a fixed bitrate. Comparing the two 3DV codecs, the video-only performed better at lower bitrates while the texture plus depth performed slightly better in higher bitrates.

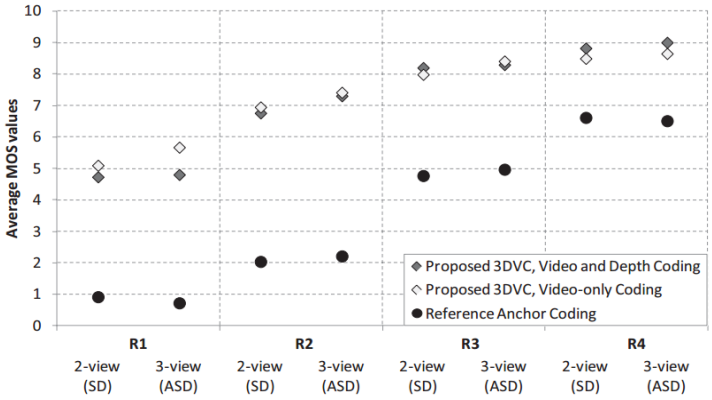


Figure 2.16 **–** Subjective performance results [50]**.**

Moreover, when considering the average bitrate distribution illustrated in Figure 2.17 depth maps prove to be high efficiently coded. Notably, as referred in [50], merging the depth maps into the multiview video sequences only represent an 8% decrease of the overall bitrate for textured components, considering R4 bitrate and 3-ASD. Also, for an 8% bitrate tradeoff depth maps will be available at the decoder, thus enhancing the quality of synthesized views [50].

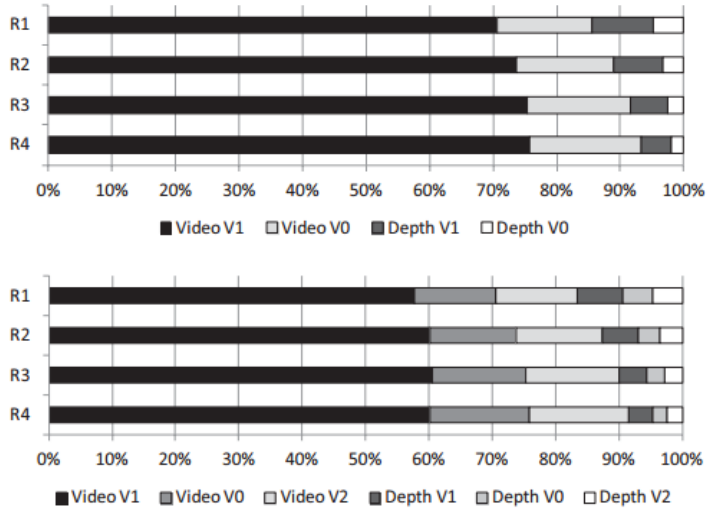


Figure 2.17 – Average bitrate distributions: 2-SD (top), 3-ASD (bottom) [50].

2.6. View Synthesis

Image-based modelling and rendering (IBMR) is a field of study in both computer graphics and computer vision. In recent years, IBMR gained considerable attention from the scientific community, as emerging techniques increasingly allowed to generate very realistic synthesized images. View synthesis is a main IBMR topic and involves the synthesis of additional views by only exploiting data from distinct, available views of the same scene.

View synthesis can be performed in a variety of ways depending on the available dataset; notably, if acquired views have or do not have any implicit or explicit geometric information. If no explicit geometric information is available, then the view is usually synthesized using view interpolation techniques. View interpolation, or view morphing, generates in-between views along the line of two

original camera centres, based on point correspondences obtained through optical flow obtained by means of feature matching instead of geometric information. Also, there are techniques which involve obtaining geometric information from the views, e.g. depth estimation, to later perform the view synthesis using a 3D synthesis model. Frequently, these techniques do not offer good performance as geometric information is hard to obtain from real images even using state-of-the-art computer vision algorithms. View synthesis techniques which explicitly involve geometric information are called depth image-based rendering (DIBR), and involve a depth map which describes the geometry of the scene. Some of these techniques have been adopted in the context of 3D-HEVC standard development, targeting further reducing the overall bitrate for the target QoE.

2.6.1. Basics

Synthesized views can be generated from available views by a process denoted as 3D warping. If the available adjacent views are not in the same plane (parallel) to the desired viewpoint to be synthesized, then a three-step approach such as proposed in [53] may be used; this process begins by transforming adjacent non-parallel views to parallel ones, which are then used in the following two-step approach. At first, available views are projected into a 3D space by making use of their corresponding depth map; then, 3D space points are re-projected into the 2D plane of the synthesized view. To achieve this result, there are two sets of parameters that need to be known:

- ***Intrinsic camera parameters:*** These parameters are essential to establish the transformation matrix (A) which enables the conversion of the 3D camera coordinates to its 2D image coordinates, notably:

$$A = \begin{bmatrix} f_u & 0 & d_u \\ 0 & f_v & d_v \\ 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

These parameters depend on the image sensor format, focal length and principal point; notably, d_u and d_v designate the position of intersection of the optical axis with the image plane in pixel coordinates (*i.e.* coordinates of the image centre), and, respectively, f_u and f_v denote the horizontal and vertical focal lengths (in samples) and may be obtained as in equation (2.2) where f denotes the focal length in millimetres and k_u and k_v express the number of pixels per millimetre along the u and v axes, respectively.

$$\begin{aligned} f_u &= -f \times k_u \\ f_v &= -f \times k_v \end{aligned} \tag{2.2}$$

- ***Extrinsic camera parameters*:** These parameters express the coordinate system transformations from 3D space to 3D camera coordinates (position and orientation), as denoted by equation (2.3), where E, R and t are, respectively, the extrinsic parameters, the rotation matrix and the translation vector. In practice, the extrinsic parameters define the position of the camera center and the camera's heading in world coordinates.

$$E = \langle R|t \rangle \tag{2.3}$$

2.6.2. Main Tools

Depth image-based rendering techniques have significantly evolved over the past decade and although multiple solutions have been proposed, this Thesis will only focus on the two tools included by MPEG in the 3D-HEVC standard Test Model 16. Although these tools are non-normative, MPEG has adopted a depth estimation tool and a view synthesis tool as they are essential to accurately assess the performance of the coding solutions; these tools are briefly described in the following.

I. Depth Estimation

Most state-of-the-art depth estimation techniques use 2D Markov random fields as a solution to derive depth maps from texture data, where each node is defined by all possible disparities and conforming probabilities [54]. These probabilities are represented as log-probabilities and are scored by a sum of two functions, namely similar cost and smooth cost. In this way, the disparities are computed and consequently a depth map is generated, notably an estimated (not acquired) depth map.

MPEG adopted a depth estimation solution based on stereo matching and graph cuts algorithms [55], which estimates the depth maps based on a pair of stereo images regularly selected from the views in a linear and horizontal direction. The resulting depth maps are represented by an 8-bit gray value map, where level 0 denotes the farthest point and the level 255 the nearest point [55].

II. View Synthesis

This type of view synthesis method is often called DIBR since it uses texture and depth maps to synthesize additional texture views. The synthesized views are derived out of correlation between the different perspectives by means of interpolation tools. As illustrated in Figure 2.18 [56], the synthesis process includes the following steps:

1) **Depth mapping to target viewpoint:** First, depth maps are mapped into the desired viewpoint using appropriate camera information;
2) **Depth filtering by median filter:** The mapped depth maps are processed through an arithmetic mean filter to fill the smaller holes caused by rounding operations;
3) **Texture mapping:** Textures from left and right views are warped into corresponding left and right target viewpoints using the remapped depth maps;
4) **Hole filling:** Any holes in a new right mapped view are filled with information from the left mapped view, and vice-versa;
5) **Blending:** The desired texture view is then obtained by blending the two previously mapped texture views, depending on a factor related to their spatial distances to the desired viewpoint position;
6) **Inpainting:** Finally, the holes within the synthesized frame are filled by an inpainting algorithm.

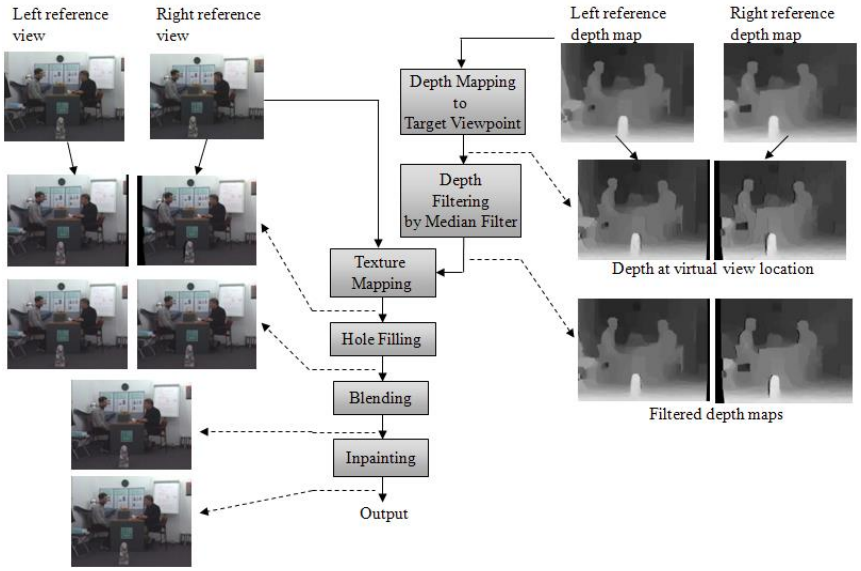


Figure 2.18 – Simplified view synthesis architecture [56].

CHAPTER 3

Image and Video Quality Assessment

The main goal of any multimedia signal processing architecture is to provide the best experience to the final user, in this case a human being. This means that, at the end of the multimedia processing flow, there is a person that assesses the displayed multimedia data, after decoding and synthesis, using the human visual system. This assessment is also based on his/hers psychophysical and physiological background, specifically the context, personal experience, expectations, *etc*. Since the scope of this Thesis is limited to visual quality assessment, any other dimension related to multimedia quality of experience (QoE) will not be considered.

There are two main types of image and video quality assessment methodologies, notably subjective protocols and objective metrics. Subjective protocols consist in experimental processes that follow some recommended procedure, e.g. as defined in ITU-R Rec.BT.500 [57] and ITU-T Rec.P.910 [58]. These experimental processes are performed by asking human subjects to score the quality of some (mostly) processed/distorted image and video content after its visualization. Subjective experiments are more reliable than objective metrics as they correspond to direct quality assessment based on Human quality perception. However, subjective experiments are far more resource demanding and slow, as a number of detailed procedures and requirements must be accurately met to produce trustworthy and statistically relevant results. Thus, reliable objective quality assessment metrics are of paramount importance to reduce the complexity and speed of the quality assessment process; naturally, to be useful, objective quality scores must correlate as much as possible with the corresponding subjective scores. However, most available quality metrics do not present very reliable results for synthesized views when compared to the corresponding subjective scores [59]. In general, objective video quality assessment metrics may be classified into three major types, namely:

- ***Full Reference*:** This type of metrics directly compares the processed/decoded/synthesized view, also known as test sequence, with the corresponding original view which is taken as reference.
- ***Reduced Reference*:** This type of metrics compares features from the test sequence with features from the corresponding original view which are transmitted and taken as reference.
- ***No Reference*:** This type of metrics directly assesses the test sequence without a reference, as no reference is available to make any comparisons.

The reliable assessment of video quality plays a key role in improving the end users QoE, especially for synthesized views since geometric distortions influence directly the users' QoE. In this context, some new quality assessment metrics have been specifically designed to categorically achieve a reliable quality assessment for synthesized views. Before considering any multimedia communication system, there is the visual scene which is composed by objects, gist (general context) and a layout (relationships among previous two items) [60]. When a scene is captured by a digital camera, either as an image or a video sequence, there are numerous types of errors that may be introduced by several processes such as the acquisition, processing, compression, storage and transmission. These errors may exhibit distinct behaviours that result in some particular subjective effects, some as Gaussian-like white noise and blurring; these image artefacts may affect the overall quality of an image or video sequence in peculiar ways. The image artefacts may have different subjective impact depending on:

- ***Input type*:** Depending on the type of stimuli, *i.e.* image or video sequence, quality assessment guidelines change. As the amount of information processed by the brain in a given period of time is the same in both situations, and motion from the video sequence will flood the brain with a lot of information, some errors will be concealed by the motion in the video itself and may not affect the overall QoE [61]. However, if the video is paused and scrutinized frame by frame, the perceivable quality will suffer a high degradation, offering a poor visual experience to the user, since the user will be able to perceive all its visual 'static' content.
- ***Input content*:** The input content has also an effect on the impact of an error when considering image quality assessment. Depending if it is natural or computer-generated content, it may display a different energy distribution, notably between the higher and lower frequencies; moreover, also the amount of temporal variation and motion across the captured scene may change. All these features affect how a use perceives a given sequence, either psychophysical and physiologically.

Image quality assessment metrics should consider all these features as a common ground, despite the fact that each and every single metric has its own paradigm and approach to the problem. This chapter begins by addressing image and video typical distortions, emphasising the most common 2D and 3D distortions and artefacts. Then, it will describe the most widely used image and video quality assessment metrics for 2D and 3D (synthesized views), as well as the main subjective methods to assess image and video perceptual quality. In this context, each metric will be presented with its mathematical equations, principles and limitations.

3.1. Image and Video Distortions (2D and 3D)

Every communication system is prone to errors, and multimedia communication systems are no exception to the matter. Regarding multimedia systems, errors related to transmission and coding tools may result in some visible distortions. Depending on the type and strength of the distortions, as well as on the content, the user's degree of immersion may be significantly reduced, thus seriously affecting the overall QoE. 2D image-based distortions have been extensively studied throughout the years, notably:

- ***Blurring Effect*:** Refers to the blurry look that an image may display, see Figure 3.1 a); this look may result from over-filtering the texture samples and may be found in several applications; for example, it is rather common to find the blurring effect in JPEG 2000 overly compressed images.
- ***Blocking Effect*:** Refers to the blocking aspect that an image may exhibit, see Figure 3.1 b); it typically results from a highly lossy compression with block-based coding algorithms. The blocking effect is a broadly perceivable effect that disturbs several image/video-based applications, in particular those using the JPEG image coding standard and most video coding standards.

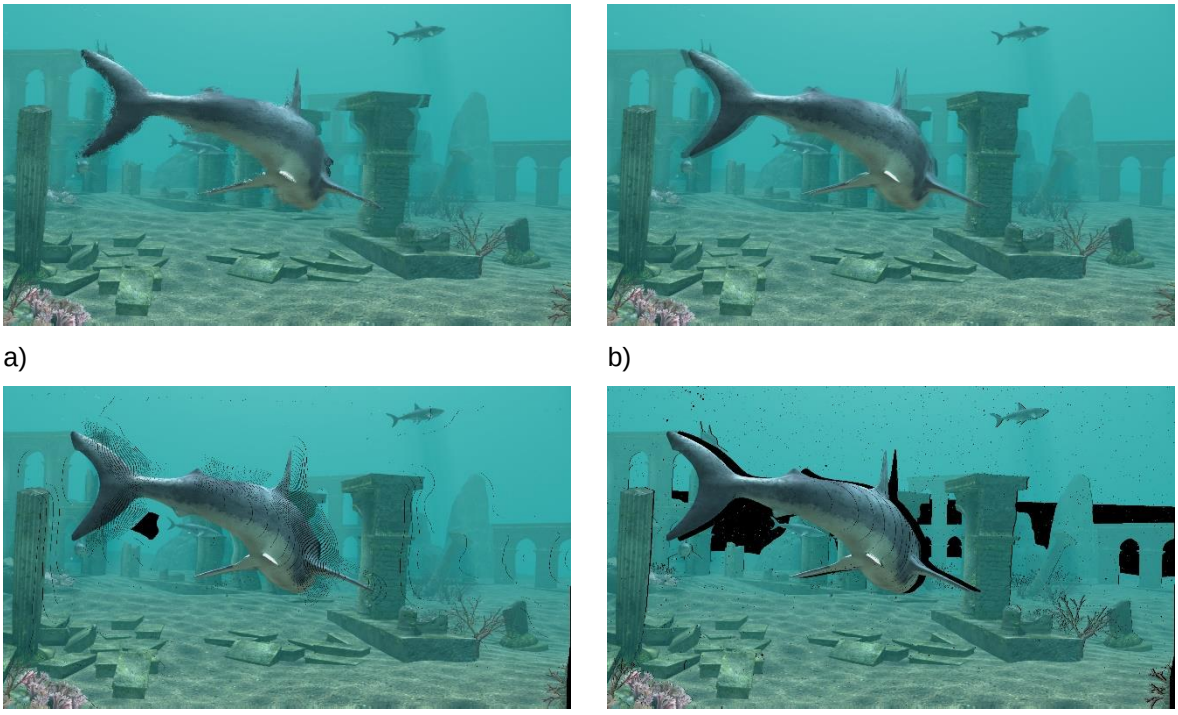
a) b)

Figure 3.1 – Common 2D image-based distortions: a) blurring effect; b) blocking effect [62].

As described in Section 2.5, the view synthesis process renders a synthesized view by exploiting its correlation with its neighbouring perspectives by using appropriate interpolation tools; depending on the specific DIBR technique involved, irregularities and errors are produced, thus resulting into new distortion types. DIBR distortions might be originated from different sources, from the synthesis process itself to the texture and depth compression. As views synthesis involves new distortion types, the available 2D image-based objective quality metrics may not perform well to evaluate the quality of the synthesized views. In addition to the typical 2D distortions, synthesized views include a brand-new variety of artefacts, notably:

- ***Geometric Distortions*:** Refers to non-uniform distortions that deform the objects in the synthesized view. As shown in Figure 3.2 a), geometric distortions are usually found in the object edges and are usually caused by depth map incongruences, e.g. resulting from high compression.
- ***Ghosting Distortions*:** As illustrated in Figure 3.2 b), ghosting is a major visual handicap for the QoE which occurs after the blending process, usually due to inaccurate camera parameters or errors in the edge matching process; it is typically caused by overfiltering and high texture compression.
- ***Cracks*:** As shown in Figure 3.2 c), they refer to perceivable image cracks, meaning missing texture samples. Cracks normally appear as a result of rounding operations and neighbourhood interpolations when reference pixels are mapped to non-integer coordinates in the synthesized view.
- ***Occluded Areas*:** When the reference views to be used for the synthesis process have occluded areas that are not anymore occluded in the synthesized viewpoint, large holes that need to be filled may appear. This problem may be dealt with using inpainting techniques as otherwise seriously degrading visual artefacts as those shown in Figure 3.2 d) may result.

- ***Flicker Distortions*:** Refer to the flickering effect that may only take place in video sequences. This distortion may be caused by depth map inaccuracies as those described above for the geometric distortions, which project pixels to wrong positions during short periods of time, thus creating some flickering effects.



a) b)

c) d)

Figure 3.2 – Common 3D image-based distortions: a) geometric distortion; b) ghosting distortion; c) cracks; d) occluded areas [62].

3.2. 2D Objective Quality Metrics

Image-based quality assessment (IQA) metrics provide an objective, mathematical way to compute the perceivable quality without directly using human observers. Each IQA metric offers its own approach to the problem, as well as tools and processes that follow four broadly related although distinct assessment paradigms [61]:

- ***Signal-based approach*:** The quality assessment of a sequence is based on the signal fidelity, meaning that by measuring the signal fidelity of a test sequence against the reference signal, a score will be created; while this approach is used in every single metric, in its purest form, this is the paradigm used by the peak signal-to-noise ratio (PSNR).
- ***Perceptual-based approach*:** The quality assessment based on a perceptual-based approach considering that perceptually-like quality estimators have a better understanding of human visual system perception; this approach includes criterion such as luminance and contrast distortion measurements; this is the approach used in video quality metric (VQM) [69].
- ***Structural-based approach*:** Structural-based assessment methods are also included in the perceptual-based approach, as these methods have their theory based on the human visual

structural perception as human visual perception is based on the extraction of structural information; this is the approach used in the structural similarity index (SSIM) [64].

- ***Human visual system (HVS)-based approach*:** This approach is also related to the perceptual-based approach, as it is a more detailed application of the perceptually-like quality estimators. However, this HVS-based approach relies on HVS modelling from psychophysical experiments. These models can be grouped into two categories: neurobiological and properties of psychophysical human vision. This is the main paradigm behind visual information fidelity (VIF) [68].

3.2.1. PSNR: Peak Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio (PSNR) and the related Mean Square Error (MSE) are the most broadly used metrics to assess lossy compressed signals. The PSNR importance is related to its ubiquitous use as a quality metric for visual communication systems up to date. Although this subsection addresses the PSNR, it starts by presenting the MSE, as MSE is used in the PSNR computation.

MSE is a mathematical error function used to assess the quality of an estimation; in the context of FR image quality assessment, it essentially measures the difference between a distorted visual signal and its reference (ground truth) in a pixelwise way as follows:

$$MSE = \frac{1}{I_L I_W} \sum_{j=1}^{I_L} \sum_{i=1}^{I_W} [I_R(i,j) - I_D(i,j)]^2 \tag{3.1}$$

where I_L and I_W represent the number of pixels in the image horizontal and vertical dimensions, I_R and I_D are the reference and distorted image, and i and j are the indexes used as pixel coordinates in the MSE computation process.

The PSNR is a quality metric, and not anymore a distortion metric as the MSE, and it is computed as the ratio between the image maximum dynamic range and the MSE expressed in base 10 logarithmic scale, as shown in equation (3.2), where b represents the image bit depth, typically 8 bits:

$$PSNR = 10 \log_{10} \left(\frac{(2^b - 1)^2}{MSE} \right) \tag{3.2}$$

PSNR inherits some of the MSE features; therefore, it usually presents a fairly good assessment performance considering several strict testing conditions (*e.g.* no image displacement). However, when it has to deal with distortion types that may be present in synthesized views, such as distortion around objects edges, shear-deformations and image-shifting, it typically fails. This is due to the high correlation between the reference and estimated sample of an image as these types of distortions do not follow a uniform spatial distribution, penalizing greatly the image subjective score without penalizing in the same way the objective score, or the other way around. This means that two differently distorted images with the same objective score may have remarkably distinct types of deformations, some of which are much more visually penalizing than others.

3.2.2. SSIM: Structural Similarity Index

The Structural Similarity (SSIM) Index was first introduced by Wang and Bovik in 2004 [64], and brought a new image quality assessment paradigm to the scientific community. Instead of relying on an error-sensitivity-based approach, as the more conservative solutions like the PNSR, it is based on the assumption that the HVS perception is highly depend on the image structural information. Considering this assumption, the SSIM organizes the similarity measurement into three comparisons between the reference (R) and the distorted (D) images, namely: luminance ($l(\cdot)$), contrast ($c(\cdot)$) and structure ($s(\cdot)$), as shown in equation (3.3):

$$SSIM(R,D) = \left[l(R,D)^{\alpha} \cdot c(R,D)^{\beta} \cdot s(R,D)^{\gamma}\right] \tag{3.3}$$

Each comparison contemplates an exponential weighting factor, hereby referred as α, β and γ, which are related factors for the luminance, contrast and structure, respectively. These weighting factors are used to adjust the relative importance of each comparison, and the sum of all factors must be equal to 1.

As shown in equation (3.4), the luminance comparison is a function of the luminance mean values for both the reference (R) and distorted (D) images and also involves a constant (c_1); c_1 is a really small value that was added to avoid instability in the case the denominator value is too close to zero.

$$l(R,D) = \frac{(2\mu_R\mu_D + c_1)}{(\mu_R^2 + \mu_D^2 + c_1)} \tag{3.4}$$

The contrast comparison, on the other hand, uses the standard deviation values for both the reference and distorted images, while it includes a second really small constant (c_2) for the same reason as mentioned for the luminance comparison; the comparison is expressed as in equation (3.5):

$$c(R,D) = \frac{(2\sigma_R\sigma_D + c_2)}{(\sigma_R^2 + \sigma_D^2 + c_2)} \tag{3.5}$$

Additionally, the luminance and contrast comparisons are proven to be consistent with the Weber's law, which states that the just noticeable difference (JND) threshold varies in a nearly linear proportional way with the background intensity [64]. This effect is often called as the *luminance or contrast masking effect* and is contemplated in the SSIM equation [65].

The structure comparison is performed by computing the Pearson correlation coefficient between the samples in the reference (R) and distorted (D) images, to seek for the structural correlation among them. As shown in equation (3.6), the structure comparison is defined as the statistical covariance for the discrete variables associated to the two image signals divided by their multiplied standard deviations.

$$s(R,D) = \frac{(\sigma_{RD} + c_3)}{(\sigma_R\sigma_D + c_3)} \tag{3.6}$$

SSIM has shown better performance when applied locally, and typically using an 8x8 square window; this window is associated to a pixel by pixel displacement through the whole image, thus leading to a series of partial scores that are used to derive a final score via an arithmetical mean. SSIM typically shows better performance when compared with PSNR in terms of correlation with the subjective assessment; however, its performance worsens in cases of translated, scaled or rotated images, as a result of the pixelwise comparison across the entire image.

3.2.3. VIF: Visual Information Fidelity

In 2015, Sheikh and Bovik proposed a new quality metric that is based on Shannon's information theory featuring statistical models named Visual Information Fidelity (VIF) [68]. This solution assumes that human beings developed their visual system over the years to best perceive natural scenes, which correspond to a small subset of all possible signals/images. To apply this hypothesis, VIF adopts wavelet trees to model the Natural Scene Statistics (NSS); these wavelet trees are used as a class of non-Gaussian multiscale stochastic processes defined by random cascades on trees of multiresolution coefficients, which produce Gaussian Scale Mixtures (GSM) in the wavelet domain. As shown in Figure 3.3, VIF is defined by a set of three different statistical models: Natural Image Source Model, Distortion Model and Human Visual System Model.

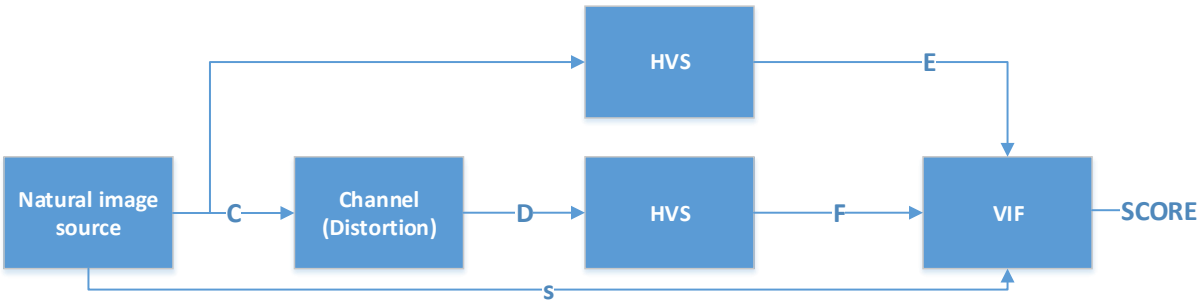


Figure 3.3 – Visual Information Fidelity metric architecture.

I. Natural Image Source Model

The Natural Image Source Model is conveniently included in the VIF design to extract NSS from the image, resulting into a set of GSM models in the wavelet domain, where each GSM model describes a single wavelet subband. This decomposition works best for natural scene statistics because it captures key statistical features of natural images due to their kurtotic behaviour (shape of their probability distribution), notably high peaks and heavy-tailed distribution. Such behaviour is highly attractive in a number of ways, notably the image properties are easily captured (enabling the use of very general and flexible models) and favours the assessment of introduced distortions (by applying distortion classification models) [67] [68].

II. Distortion Model

The Distortion Model, or Channel Distortion, used in the VIF design is defined as a wavelet subband GSM model similarly to the Source Model. This model has the single purpose of assessing how the image statistics is altered when generic distortions are applied (*e.g.* blur, chromatic distortions); however, it is designed to identify the perceptual annoyance of a distortion and not to model the image artefacts.

III. Human Visual System Model

The Human Visual System (HVS) model is also represented in the wavelet domain. Since the VIF metric already makes use of the NSS to model the source signal, many aspects of the HVS are already modelled. This model assumes that human beings are only capable of perceiving a certain amount of visual information in a given time period; this implies that, depending on the source of information (image), that are features that stand-out more than others, thus masking some effects in a given stimuli.

According to this assumption, the HVS model is used to extract the *image information*; some empirical analysis has shown that just by modelling the internal neural noise, the assessment performance may be significantly boosted.

As expressed by equation (3.7), the VIF score is computed as the ratio between the mutual information of the reference ($I(\vec{C}^{N,j};\vec{E}^{N,j}|s^{N,j})$) and distorted ($I(\vec{C}^{N,j};\vec{F}^{N,j}|s^{N,j})$) images, in all wavelet bands; $\vec{C}^{N,j}$, $\vec{E}^{N,j}$, $\vec{F}^{N,j}$ and $s^{N,j}$ stand, respectively, for the reference image random fields, reference image information content random fields, distorted image information content random fields and natural scene model parameters.

$$VIF = \frac{\sum_{j\in subbands} I(\vec{C}^{N,j};\vec{F}^{N,j}|s^{N,j})}{\sum_{j\in subbands} I(\vec{C}^{N,j};\vec{E}^{N,j}|s^{N,j})} \quad (3.7)$$

The VIF unconventional way to assess the image quality has resulted into a good overall performance across different distortion types, unlike other state-of-the-art FR quality assessment metrics that usually only standout for one or two types of distortion. Although VIF uses only the image luminance component it has higher computational complexity than SSIM due to the wavelet decomposition and distortion model.

3.2.4. VQM: Video Quality Metric

The Video Quality Metric (VQM) adopts a general-purpose model (*i.e.* works well for a wide span of bit depth ranges and quality range (QP)), and was proposed in late 2003 by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA) as a model to estimate video quality [69]. The VQM was adopted in several standards by various organizations, notably the International Telecommunications Union (ITU) Rec. J.144 and the American National Standards Institute (ANSI) Rec. T1.801.03. The VQM consists in a four-staged process, notably: i) Calibration; ii) Extraction and Perception-based Features Processing; iii) Video Quality Parameters Computation; iv) Final Score.

I. Calibration

The Calibration process includes spatial and temporal misalignment handling, identifying valid regions and correcting the gain and level values. All these processes tackle the issues created when non-perceivable distortions in videos quality may needlessly penalize the VQM final score; for example, if a video has a spatial misalignment of 2 pixels on both axis, and no calibration procedure is performed, VQM will score it as really poor, while the user will not be able to tell the difference between the reference and distorted video sequences.

II. Extraction and Perception-based Features Processing

The features extraction process applies a perceptual filter to enhance some image-based properties (*e.g.* edge information) in the spatial, temporal and chrominance domains, in order to extract after these quality features. This process is held in spatio-temporal sub-regions to exploit the varying video sequence behaviour.

III. Video Quality Parameters Computation

The video quality parameters are derived from a set perceptual changes, which are computed by comparing the extracted perceptual features from the reference image with the extracted perceptual features from the distorted image. The following parameters are considered: si_loss detects the loss of spatial information (*e.g.* blur distortion); hv_loss and hv_gain detect edge shifts by blurring or blocking effects; $chroma\_spread$ detects the spreading of two colour sample distributions in 2D; si_gain detects image enhancements (*e.g.* edge sharpening); ct_ati_gain detects the weight of perceivable impairments given the amount of motion involved in a spatio-temporal region; and $chroma\_extreme$ detects severe loci colour impairments.

IV. Final Score

The VQM final score computation involves the linear combination of the parameters that were previously derived, as expressed in equation (3.8):

$$VQM = -0.2097 * si\_loss + 0.5969 * hv\_loss + 0.2483 * hv\_gain + 0.0192 * chroma\_spread - 2.3416 * si\_gain + 0.0431 * ct\_ati\_gain + 0.0076 * chroma\_extreme \quad (3.8)$$

3.2.5. MOVIE: Motion-based Video Integrity Evaluation

In early 2009, Seshadrinathan and Bovik released a new FR video quality assessment metric that integrates both the spatial and temporal aspects of distortion assessment, names as MOtion-based Video Integrity Evaluation (MOVIE) [70]. As illustrated in Figure 3.4, MOVIE uses a 3D spatio-temporal Gabor filter bank to extract multiscale spatial and temporal coefficients from the reference and test video sequences; these coefficients are selected according to the motion estimated from the reference video sequence in the form of an optical flow field. After the multiscale decomposition, the algorithm computes spatial and temporal assessment scores using two different components: Spatial MOVIE and Temporal MOVIE.

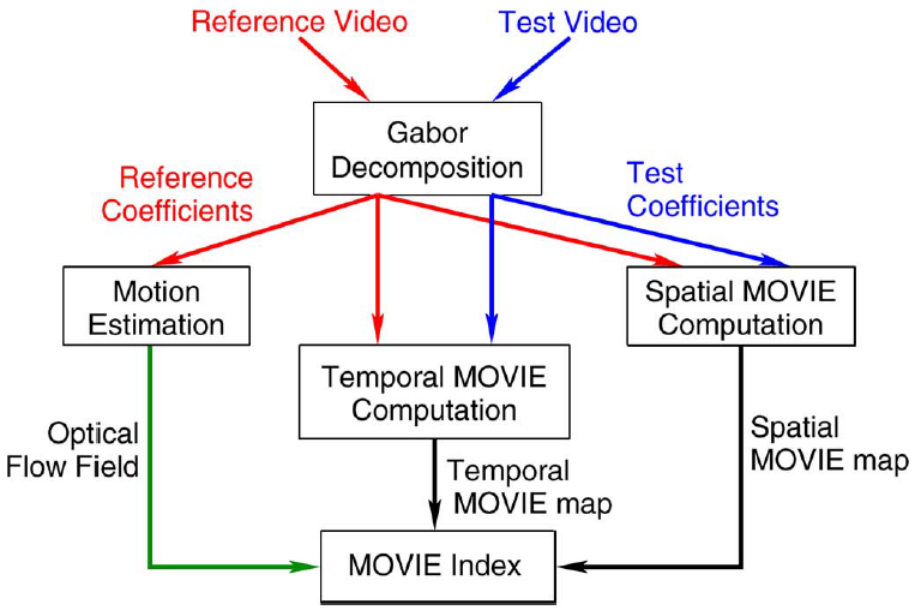


Figure 3.4 – MOVIE architecture.

I. Spatial MOVIE

Spatial MOVIE captures spatial distortions in a similar way as the SSIM, but does it by analysing the output coefficients from the Gabor filter bank. Spatial distortions, such as blur, ringing artefacts and false contouring, are captured using errors computed between the reference and test multiscale coefficients; this involves a MSE computation and a normalized masking function as defined in [71]. Using the captured spatial distortion results, a spatial frame quality score ($FQ_S(\cdot)$) is computed; this results in a spatial quality score for a single frame, which will then be used to estimate the video quality. The spatial index score is computed as the average score over τ frames of the spatial frame quality computed for a given time instant (t_j) as:

$$Spatial\ MOVIE = \frac{1}{\tau}\sum_{j=1}^{\tau} FQ_S(t_j) \tag{3.9}$$

II. Temporal MOVIE

Temporal MOVIE captures temporal degradations in the video sequence that result from either the motion in a scene, and/or motion between viewpoints. The various temporal distortion types are captured by the output coefficients of the Gabor filter bank and the optical flow field that is obtained by the motion estimation process, notably by computing the video quality along the motion trajectories. The temporal index score is computed as the square root of the average score over τ frames of the temporal frame quality ($FQ_T(\cdot)$) computed for a given time instant (t_j) as:

$$Temporal\ MOVIE = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau} FQ_T(t_j)} \tag{3.10}$$

The overall MOVIE score is computed as the multiplication of the two computed indexes, as expressed in equation (3.11). MOVIE is an intuitively designed framework that bases its features on a Gabor filter bank and an optical flow estimation algorithm to properly capture the distortions in spatial and temporal domain.

$$MOVIE = Spatial\ MOVIE \times Temporal\ MOVIE \tag{3.11}$$

3.3. 3D Quality Metrics

As mentioned at the beginning of this chapter, view synthesis techniques use new processing tools that generate new forms of distortions. This means that the previously available image-based quality assessment metrics do not typically perform well in the assessment of synthesized views and thus new, appropriate quality assessment metrics have to be developed. The new 3D quality assessment metrics typically follow a perceptual-based approach, either structural or HVS-based.

3.3.1. 3DSwIM: 3D Synthesized view Image Quality Metric

I. Objective

The 3D Synthesized view Image quality Metric (3DSwIM) is a FR video quality assessment metric for 3D synthesized views [59] that relies on two main assumptions: i) the visual quality of synthesized images is not greatly affected by displacement differences, which is relevant as synthesized views will typically display shifted objects regarding other views; and ii) distortion effects around human subjects are far more visually impacting.

II. Technical Approach

3DSwIM performs quality assessment in a block-based structure, meaning that the frame is partitioned into $B_n \times B_m$ sized blocks (B). Each block is then processed by analysing the statistics between the reference and synthesized views, while adding a weighting factor, in this case related to skin detection, to exploit the second assumption presented above. The 3DSwIM architecture is presented in Figure 3.5 [59]. Considering the two inputs as the reference and synthesized views, the initial processing for both views is rather similar, only differing on the Skin detection module only applied to the synthesized view. Finally, both pre-processed views go through a statistical analysis algorithm which will derive the quality score for the synthesized view.

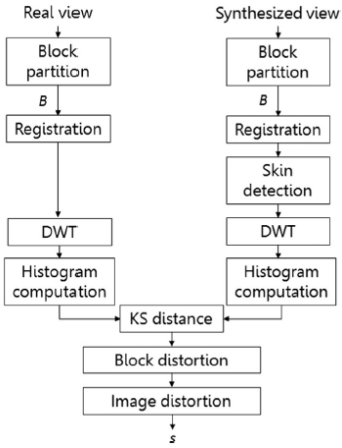


Figure 3.5 **–** 3DSwIM architecture [59].

To better understand the proposed quality metric, the methods and motivation of each architectural module are now reviewed by order of entrance:

1) **Block partition:** An image is partitioned into non-overlapping blocks (B) with a height and width determined by the metric parameters B_n and B_m, respectively.
2) **Registration:** This registration procedure aligns the best matching blocks from the reference and synthesized views, *i.e.* registers both blocks together for further evaluation. To do so, an Exhaustive Motion Estimation Search-like [72] algorithm is used here with a search window size (W) parameter to be initially defined. This step is associated to the first assumption mentioned above, implying that translational displacements do not have a major quality impact in terms of synthesized views visual quality [59].
3) **Skin detection:** This procedure is only performed for the synthesized view. Based on HSV colour space segmentation, each block is perceptually weighted depending on skin detection using the H

component [73]. After, a weighting factor is computed to penalize the distorted blocks with "human skin" pixels, nominally W_{skin}. This step is associated to the second assumption above stating that humans are more sensitive to distortion in skin areas.

4) **Discrete Wavelet Transform:** A discrete wavelet transform (DWT) is applied to each block from both views to obtain the horizontal details [74]. This derives from the fact that some authors [59] believe that high frequency coefficients in the horizontal direction are the main source of distortion for synthesized views, as the pair of views used to synthesize an image are typically horizontally displaced.
5) **Histogram computation:** After the DWT, a histogram is computed to obtain a statistical model related to the frequency components associated to a given block for each view, namely h_o for the reference view and h_s for the synthesized view.
6) **Kolmogorov-Smirnov distance:** For both histograms above, the statistical non-parametric Kolmogorov-Smirnov (KS) metric measures the distance between the two probability distributions represented by the two histograms. Notably, the two distributions are KS compared as in (3.12), where F_{o_b} and F_{s_b} refer to the probability distribution function for a block from the reference and synthesized views, respectively.

$$KS_b(x) = \left|F_{o_b}(x) - F_{s_b}(x)\right| \tag{3.12}$$

7) **Block distortion:** This module computes the maximum difference distance (supremum) for the KS distance for a given histogram. Hence, as expressed by Equation (3.13), a block distortion (d_b) value is assigned to the derived supremum value for that block.

$$d_b = \max\left(KS_b(x)\right) \tag{3.13}$$

8) **Image distortion:** The image distortion (D) is computed as the normalized sum of the block distortions (d_b) measured throughout the full set of image blocks as expressed by Equation (3.14); the normalization factor (D_0) depends on the number of samples evaluated and the skin detection weighting factor (W_{skin}) has a default, initially defined setup value of 15.

$$D = \frac{1}{D_0}\sum_{b=1}^{B}(W_{skin} \cdot d_b) \tag{3.14}$$

Finally, the image quality score (S) is computed using the normalized image distortion (D) as in Equation (3.15). The quality score values are normalized in a way such that $S \in [0; 1]$. The final quality score is inversely proportional to the derived image distortion as follows:

$$S = \frac{1}{1 + D} \tag{3.15}$$

3.3.2. SIQE: Synthesized Image Quality Evaluator

I. Objective

The Synthesized Image Quality Evaluation (SIQE) metric is a RR video quality assessment metric [75] that is based upon the so-called *cyclopean eye theory*, which relates to the central reference midway point between the two eyes [76], also designated as *mental image* [75]. The process to estimate a reference view begins by applying the divisive normalization transform to both views and then fuse them

together. Regarding the synthesized view, the 3D video quality is then objectively scored through a statistical evaluation that compares the similarities between the synthesized and estimated reference cyclopean views, as described in the following. This RR video quality assessment is held in the Divisive Normalization transform, which has been endorsed as an effective nonlinear coding transform for natural sensory signals [77].

II. Technical Approach

The SIQE architecture is illustrated in Figure 3.6 [75]. It adopts the multiview plus depth data format as input regarding a stereoscopic view pair. Although a texture plus depth data format is used to synthesize a view employing DIBR techniques, the cyclopean view is estimated only from the texture components.

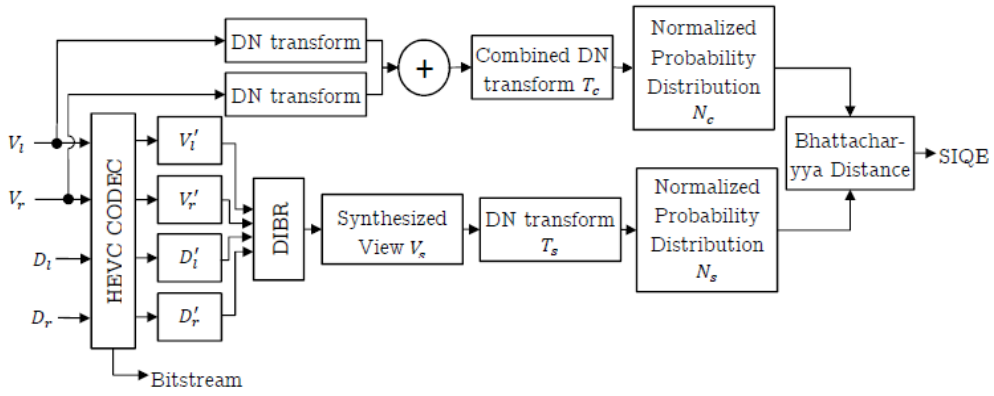


Figure 3.6 – SIQE architecture [75].

After HEVC coding and decoding of the texture and depth views, the DIBR created synthesized view as well as both the original texture views go through the following process:

1) **Divisive Normalization transform:** A Divisive Normalization (DN) transform is applied to the original texture views with the target to estimate the statistical characteristics of both uncompressed views in order to fuse them together to build a statistical model for the cyclopean view. This process is performed to derive the statistical features of the synthesized view (s). The DN transform was originally proposed in [78], and is based on the standard psychophysical and physiological model that describes the visual process from the initial stage until the V1 cortex [79]. The DN transform process is mathematically described in equation (3.16), considering $k = \{s, l, r\}$, u and v the geometric centre of a given $m \times n$ block, l and r the left and right views, respectively, V_k the image for the k view, μ_k the local average as computed by equation (3.17), σ_k the standard deviation as computed by equation (3.18), and ϵ a small constant value to avoid divisions by zero.

$$T_k(u,v) = \frac{V_k(u,v) - \mu_k(u,v)}{\sigma_k(u,v) + \epsilon} \tag{3.16}$$

$$\mu_k(u,v) = \sum_{i=-m}^{m} \sum_{j=-n}^{n} w(i,j) V_k(u+i, v+j) \tag{3.17}$$

$$\sigma_k(u,v) = \sqrt{\sum_{i=-m}^{m}\sum_{j=-n}^{n} w(i,j)[V_k(u+i,v+j) - \mu_k(u,v)]^2} \quad (3.18)$$

The parameter w refers to the 2D circularly symmetric Gaussian weight function shown in (3.19) where σ^2 is the variance, and i and j are the indexes used to derive the aperture of the weighting function that is used to filter the importance of the features regarding their location.

$$w(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}} \quad (3.19)$$

2) **Combined DN Transform**: This step is only performed for the transformed left (T_l) and right (T_r) views. The transformed views are used to estimate the cyclopean transformed view (T_c), which is obtained by simply concatenating both transformed views as $T_c = [T_l|T_r]$.
3) **Normalized Probability Distribution:** After obtaining the statistical characteristics for the synthesized (s) and cyclopean (c) views, a normalized probability distribution is calculated, independently for both views.
4) **Bhattacharyya Distance:** The similarities between the two normalized probability distributions are computed as a Bhattacharyya coefficient (ρ), notably by determining the statistical overlapping between the two groups of samples, *i.e*. the likelihood between the two images. The authors consider that this method has proven more reliable when compared to other alternative metrics [75].

$$\rho(\mathcal{N}_c, \mathcal{N}_s) = \sum_{x\in\kappa} \sqrt{\mathcal{N}_c(x)\mathcal{N}_s(x)} \quad (3.20)$$

The final SIQE score is given by the Hellinger distance, regarding the Bhattacharyya coefficient, as expressed in equation (3.21). SIQE measures the distortion in the synthesized image, meaning that the higher the correlation, the lower the distortion associated and the overall score.

$$SIQE = \sqrt{1 - \rho(\mathcal{N}_c, \mathcal{N}_s)} \quad (3.21)$$

CHAPTER 4

Spatio-Temporal Quality Assessment for Synthesized Views Metric

In recent decades, video processing systems evolved from a single 2D perspective to multiple perspectives, starting with stereo, eventually also with a higher spatial resolution, higher frame rate, and higher dynamic range. Driven by this evolution, available video quality assessment (VQA) methods became less appropriate as the quality-related artefacts in 2D and 3D are really not the same, and thus poor-quality assessment performance results may be achieved by somehow using 2D metrics for 3D information. Reliable VQA methods play a fundamental role in the evaluation of video processing, coding and synthesis systems as they enable, in a more tangible way, the effective quality assessment of the processed video content.

Recently, 3D video systems became more popular. These 3D systems are commonly based in the Multiview plus Depth (MVD) format, which supports the synthesis of intermediate views based on the available (acquired) views, the so-called *synthesized views*. View synthesis allows making available at the receiver as many as desired perspectives of a scene at no bitrate cost, thus allowing a very smooth navigation of the scene (or alternatively bitrate reduction, if a target number of perspectives is required).

The view synthesis process opens up a broad range of applications. However, to perform a reliable quality assessment of the synthesized views, new VQA solutions are need as the conventional 2D metrics do not perform well due to the new types of distortion associated to the synthesis process, which are naturally not considered by the conventional 2D metrics.

The VQA solution presented in this chapter is largely inspired on the solution proposed in [80] and involves a full-reference VQA metric which follows a perceptive-based approach (see Chapter 3), notably for two types of distortions: i) conventional 2D spatial distortions; and ii) flickering distortions. According to this approach, the video sequence quality is assessed in the spatio-temporal domain along block-shaped, motion-coherent temporal tubes which are first detected in the reference view and after projected to the synthesized view.

4.1. Architecture and Walkthrough

This section presents the Spatio-Temporal Video Quality metric processing architecture and its walkthrough. As this is a full reference metric, this process involves both the synthesized view and the so-called reference view, this means the original view for the same perspective position of the synthesized view. As illustrated in Figure 4.1, this VQA metric includes four main parts (red dashed boxes): i) Quality assessment structure definition; ii) Flickering distortion measurement; iii) Spatio-Temporal activity distortion measurement; and iv) Overall distortion measurement.

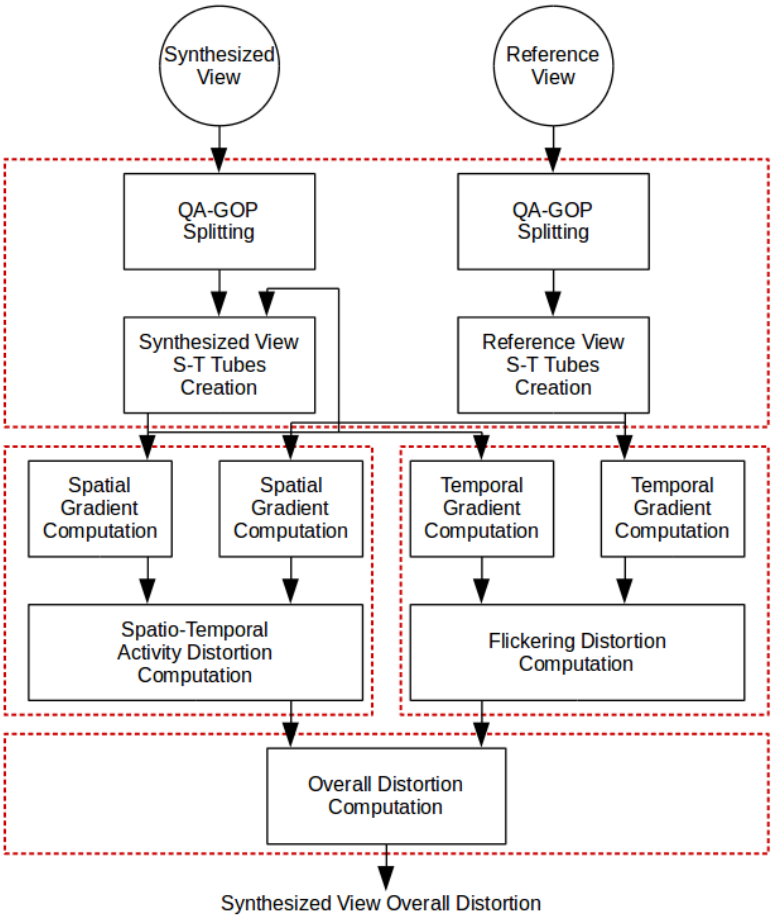


Figure 4.1 – Processing architecture for the Spatio-Temporal Video Quality metric.

The processing walkthrough for the Spatio-Temporal Video Quality metric involves the following main steps:

1) **QA-GOP Splitting:** Splits a sequence into groups of 2N+1 pictures here called a *Quality Assessment-Group of Pictures* (QA-GOP);
2) **Reference View Spatio-Temporal Tubes Creation:** Creates spatio-temporal (S-T) tubes for the reference view based on the motion estimated for each QA-GOP; the motion vectors are estimated using a block-based motion estimation process, notably full-search block-matching in a window size of 32 for blocks with size 8;
3) **Synthesized View Spatio-Temporal Creation:** Uses the spatio-temporal tubes definition data created in Step 2) for the reference view to generate equivalent spatio-temporal tubes for the synthesized view video sequence;

4) **Spatial Gradient Computation:** Computes the spatial gradient for each block within each S-T tube;
5) **Temporal Gradient Computation**: Computes the temporal gradient for the blocks within each S-T tube;
6) **Spatio-Temporal Activity Distortion Computation:** Using the spatial gradient computed in Step 4), it computes the spatial-related distortion along time;
7) **Flickering Distortion Computation:** Using the temporal gradient computed in Step 5), it computes the so-called flickering distortion;
8) **Overall Distortion Computation:** Combines the activity distortion and flickering distortion scores to compute the final Spatio-Temporal Video Quality score.

4.2. Main Modules Detailed Description

This section will now present the details for each module in the Spatio-Temporal Video Quality metric processing architecture. As this algorithm only processes the luminance component, the term 'sample' will always refer to 'luminance sample' in the following.

4.2.1. Quality Assessment Group of Pictures Splitting

The Quality Assessment Group of Pictures Splitting module is responsible to divide the full set of pictures in a video sequence into groups of pictures with a limited number of pictures, the so-called *quality assessment-group of pictures* (QA-GOP), each group composed by 2N+1 pictures. The quality assessment pipeline is first applied to each of these QA-GOPs and, once all groups have been assessed, the partial quality assessment results are integrated for the full sequence quality assessment. The QA-GOP defining parameter *N* is derived following the basic assumption that, on average, the human fixation duration is around 200-400 ms [80], *i.e.* the QA-GOP size should not be larger than the human fixation duration. Considering the maximum value for the fixation time, it comes:

$$2N + 1 \leq int\,(0.4 * Framerate) \tag{4.1}$$

For 25 Hz video, equation (4.1) returns 2N+1 ≤ 9.

4.2.2. Reference View Spatio-Temporal Tubes Creation

The main objective of this Reference View Spatio-Temporal Tubes Creation module is to extract and define a temporal data structure called *spatio-temporal (S-T) tubes* for the reference view sequence by exploiting the motion in the scene, as locally represented in a QA-GOP. As illustrated in Figure 4.2, the S-T tubes are defined as the temporal aggregation of (spatial, frame located) blocks, corresponding to specific motion-tracked blocks along time. This method enables to perform spatial-temporal distortion measurements along the motion trajectories; this is especially relevant because view synthesis distortions tend to appear around the edges of the moving objects in a scene.

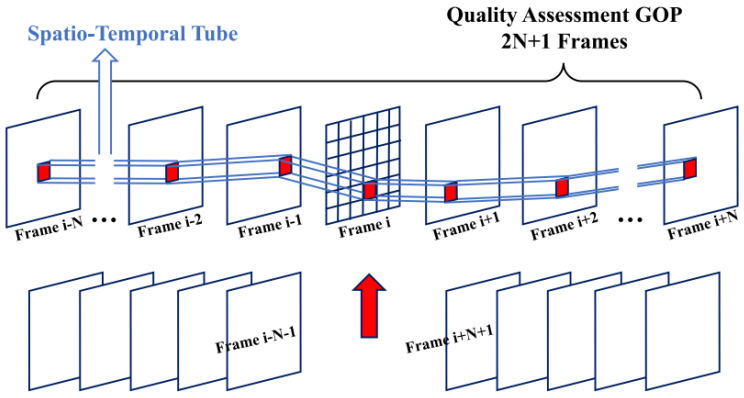


Figure 4.2 – QA-GOP and S-T tube structure [80].

Following this motivation, the reference view S-T tubes creation architecture is represented in Figure 4.3. This module implements an iterative process that runs for all pictures in each QA-GOP, starting from its center and moving to both ends of the QA-GOP. The reason for the central frame to be used as the starting point is to minimize the error propagation along the successive motion estimation processes performed in this module.

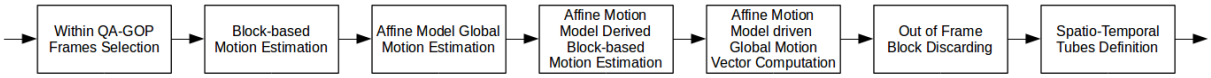


Figure 4.3 – S-T tubes creation architecture.

The various steps in this module are described in the following (this process is repeated for all the QA-GOPs in the sequence):

1) **Within QA-GOP Frames Selection:** This step is responsible to define the pairs of pictures to feed to the processing chain within this module, notably for motion estimation. It is an iterative process that selects only a couple of pictures within each QA-GOP at a time, the so-called source (SRC) and destination (DST) frames, each will be processed as described below. This step is subdivided into two sub-steps, which split a QA-GOP into two sub-sets of pictures, the so-called backward and forward sub-sets, located to the left and to the right of the central picture within the QA-GOP. Considering the indexes presented in Figure 4.2, the backward sub-set is composed by the pictures from *i* to *i-N*, while the forward sub-set is composed by the pictures from *i* to *i+N.* The processing starts with the backward sub-set, until it reaches its end at the QA-GOP far left side, and moves after to the forward sub-set until it finishes at the QA-GOP far right side. The SRC and DST frames are always two adjacent pictures which slide to the left or to right, depending if the backward or forward sub-sets are being processed. The first SRC frame for each sub-set processing is always the QA-GOP central picture and the SRC frame in the next pair of frames is the previous DST frame. In practice, this creates sliding pairs of frames moving to the left and to right, starting from the central frame. For each pair of SRC-DST pictures, the next steps are applied:

2) **Block-based Motion Estimation:** This motion estimation step consists in applying a full-search block matching algorithm to estimate the motion vectors for the SRC blocks regarding the DST frames. For the first pair of SRC-DST frames, the SRC frame is the central frame of the QA-GOP being assessed; in this case, the SRC frame is divided into a grid of *b* sized, square, non-overlapping

blocks where b is typically 8×8. For the subsequent frame pairs, the SRC blocks are the blocks previously selected for motion compensation in the previous DST frame (which is now the SRC frame); naturally, these blocks are not anymore located in the frame block grid. This block matching algorithm uses a *w* square sized samples search window in the DST frame to find the best motion vector for a given SRC block. The best prediction block is found by adopting the mean of absolute differences (MAD) as the error metric to assess all the block estimations available within the search window. This step performs this block-based motion estimation for all the blocks in the SRC frame defined as mentioned above, thus producing at the end a motion vector field connecting all SRC blocks to a block in the DST frame; in practice, the S-T tubes are extended by one frame, step by step. In detail, the block matching algorithm proceeds as follows:

a) Take a (typically 8×8) block of samples in the SRC frame; this block will be taken from a regular grid when the SRC is the central frame like in Figure 4.4);
b) Define a search window in the target DST frame around the collocated position of the SRC block;
c) Perform the mean of absolute differences (MAD) for the luminance samples between the SRC block under processing and all the blocks within the target search window, thus obtaining a MAD map with a size defined by the search window size;
d) The motion vector selected is the one corresponding to the minimum MAD value within the MAD map;
e) Go to a) until motion vectors for all blocks in the SRC frame are estimated from the DST frame.

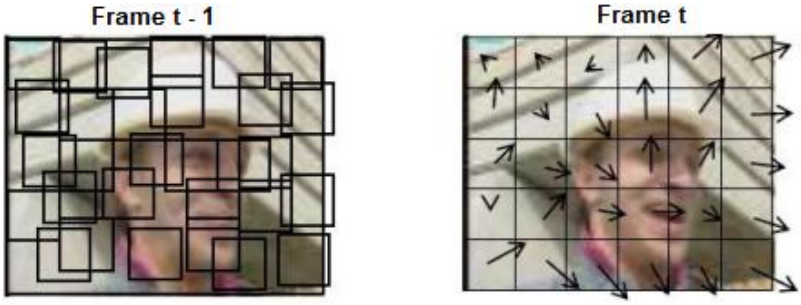


Figure 4.4 – Block-based Motion Estimation: DST and SRC block examples when the SRC is the central frame.

3) Affine Model Global Motion Estimation: Global motion is defined as the motion that applies to all parts of the frame, e.g. as generated by a camera horizontal displacement. To capture global motion, an estimation algorithm has to be performed. In this case, the global motion is modeled by a six-parameter affine motion model, which parameters have to be estimated from the motion vectors field computed in the previous step. The affine motion model (AMM) is characterized by a matrix (Θ) with 6 degrees of freedom and is able to describe rotations, sheers, scaling and translation transformations, notably occurring between the SRC and DST frames as shown in equation (4.2):

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \Theta_1 x + \Theta_2 y + \Theta_3 \\ \Theta_4 x + \Theta_5 y + \Theta_6 \end{bmatrix} \tag{4.2}$$

where (x, y) represents a sample position in the SRC frame and (x', y') is the position of the corresponding warped point in the DST frame. The AMM parameters are estimated using the least

squares estimation method, which is a regression analysis approach minimizing the sum of the squared matching error between two motion vector fields, notably the previously block-based estimated MVs (input MVs) and the MVs estimated from the current global AMM which parameters are being estimated. Each input block-based estimated MV (MV_{x_i}, MV_{y_i}) and the corresponding globally estimated MV ($\widehat{MV}_{x_i}$, $\widehat{MV}_{y_i}$) are associated the spatial coordinates (x_i, y_i), which correspond to the center of block (i). The motion vector error is defined as:

$$E = \sum_i \left(\left(MV_{x_i} - \widehat{MV}_{x_i} \right)^2 + \left(MV_{y_i} - \widehat{MV}_{y_i} \right)^2 \right) \quad (4.3)$$

4) **Affine Motion Model Derived Block-based Motion Estimation:** In this step, a motion vector is derived for each block in the SRC frame by computing the displacement between the SRC block under processing and the affine model warping of the same block. This will provide block-based motion information but now derived from the global motion information, this means the previously derived affine motion model, which characterizes the motion at the frame level.

5) **Affine Motion Model driven Global Motion Vector Computation -** The global motion vector (GMV) is here computed as the average of the block-based motion vectors derived at the previous step. The GMV is computed as in equation (4.4), where N_{blocks} is the number of blocks in the SRC frame and MV_i^{global} is the block motion vector estimated in the previous step:

$$GMV = \frac{1}{N_{blocks}} \sum_{i=1}^{N_{blocks}} \left(MV_i^{global} \right) \quad (4.4)$$

6) **Out of Frame Blocks Discarding:** This step is responsible to identify and discard the SRC blocks that might be leaving the visual scene, meaning that a motion compensated block in the DST frame falls out of it. This implies that the corresponding S-T tube is leaving the scene/frame and thus should be discarded, as it is not extending along the full QA-GOP. This module performs the following two steps:
   a) Perform motion compensation for all SRC blocks using the GMV;
   b) Check if the motion compensated block in the DST frame is outside the frame boundaries, notably if the top left or the bottom right corners of the compensated block are outside the frame, *i.e.* if any point of a motion compensated SRC block is positioned outside the DST frame;

7) **Spatio-Temporal Tube Aggregation:** This final step is responsible to aggregate the set of motion compensated blocks into a specific S-T tube structure defined within a QA-GOP. While the S-T tube creation process always starts with the motion estimation for the central frame, the motion compensation process extends to both ends of the QA-GOP, thus creating a 'snake' of motion connected blocks. The final result are several S-T tubes for each QA-GOP as the one illustrated in Figure 4.2.

4.2.3. Synthesized View Spatio-Temporal Tube Creation

The main objective of this module is to create the spatio-temporal (S-T) tubes for the sequence of frames in the synthesized view, notably using the S-T tubes already defined for the reference view. As mentioned before, the reference view is here the original view for the same position of the synthesized

view. The S-T tubes along the synthesized sequence are created using the S-T tubes computed for the reference view, obtained in the previous reference view S-T tubes creation module. In practice, this module replicates the S-T tubes structure extracted from the reference view sequence into the synthesized view sequence. This process is independent from the reference view S-T tubes creation process as the first S-T tubes are originally created only using the reference view sequence, and here replicated for the synthesized view sequence using the previously obtained S-T tubes defining information.

4.2.4. Spatial Gradient Computation

The main objective of this module is to compute the spatial gradient for each sample in the blocks contained in the previously defined S-T tubes, independently for both the reference and synthesized views. The spatial gradient ($\nabla I^{spatial}$) at a given position (x, y) in a frame t is defined by the magnitude of the spatial gradient vector, which is the sum of the horizontal ($\nabla \vec{I}^{horizontal}$) and vertical ($\nabla \vec{I}^{vertical}$) spatial gradient vectors, as defined in equation (4.5):

$$\nabla I_{x,y,t}^{spatial} = \sqrt{\left|\vec{\nabla} I_{x,y,t}^{horizontal}\right|^2 + \left|\vec{\nabla} I_{x,y,t}^{vertical}\right|^2} \tag{4.5}$$

The horizontal and vertical gradient vectors are computed by a convolution operation using the defined horizontal and vertical kernels shown in Figure 4.5.

1	1	0	-1	-1
3	3	0	-3	-3
8	8	0	-8	-8
3	3	0	-3	-3
1	1	0	-1	-1

a)

1	3	8	3	1
1	3	8	3	1
0	0	0	0	0
-1	-3	-8	-3	-1
-1	-3	-8	-3	-1

b)

Figure 4.5 – Spatial gradient kernels: a) horizontal; b) vertical.

4.2.5. Temporal Gradient Computation

The main objective of this module is to compute the temporal gradient for each block in the S-T tubes, both for the reference and synthesized views, based on the (sample) intensity changes along the motion trajectory. As expressed in equation (4.6), the temporal luminance gradient is computed as the difference between the luminance intensities (I) at a given position (x, y) in frame t and the corresponding motion compensated position (x', y'), this means the position along the motion trajectory, at frame *t-1*:

$$\vec{\nabla} I_{x,y,t}^{temporal} = I(x, y, t) - I(x', y', t-1) \tag{4.6}$$

4.2.6. Spatio-Temporal Activity Distortion Computation

The spatio-temporal activity distortion computation module is responsible to measure rather traditional 2D distortions, notably blurring and blocking artifacts. This is achieved by calculating the standard

deviation of the spatial gradient for each S-T tube. This requires to compute first the arithmetical mean of the spatial gradient values along a S-T tube, thus involving time, as expressed by equation (4.7), where y_n and x_n represent the top-left corner position for the blocks in frame t, while b_h and b_w are, respectively, the blocks height and width:

$$\overline{\nabla I^{spatial}_{tube}} = \frac{\sum_{t=i-N}^{i+N} \sum_{y=y_n}^{y_n+b_h} \sum_{x=x_n}^{x_n+b_w} \nabla I^{spatial}_{x,y,t}}{b_h \times b_w \times (2N+1)} \tag{4.7}$$

Using the computed arithmetical mean, the standard deviation for the spatial gradient in the S-T tubes is computed with equation (4.8):

$$\sigma_{tube} = \sqrt{\frac{\sum_{t=i-N}^{i+N} \sum_{y=y_n}^{y_n+b_h} \sum_{x=x_n}^{x_n+b_w} \left(\nabla I^{spatial}_{x,y,t} - \overline{\nabla I^{spatial}_{tube}} \right)^2}{b_h \times b_w \times (2N+1)}} \tag{4.8}$$

To avoid selecting imperceptible sample gradients, the computed standard deviation values are thresholded with a perceptual threshold (ξ) appropriately selected, thus obtaining a spatio-temporal activity (Γ) per tube.

The Distortion Activity (DA) for each S-T tube is computed as the absolute value of the ratio between the spatio-temporal activity in the synthesized sequence ($\tilde{\Gamma}$) and the corresponding spatio-temporal activity in the reference tube, using a base 10 logarithmic scale, as follows:

$$DA^{tube} = \left| log_{10} \left(\frac{\tilde{\Gamma}_{tube}}{\Gamma_{tube}} \right) \right| \tag{4.9}$$

After computing the distortion activity for each S-T tube, a distortion activity per QA-GOP is computed. As expressed in equation (4.10), the distortion activity per QA-GOP is computed as the arithmetical mean of the *W%* worst cases in terms of tube distortion activity for that specific QA-GOP, *i.e.* the tubes with higher distortion activity values in the QA-GOP, thus obtaining:

$$DA^{GOP} = \frac{1}{N_W} \sum_{k \in W} DA^{tube}_k \tag{4.10}$$

where N_W is the number of selected S-T tubes in the specific QA-GOP.

The distortion activity for the whole synthesized view video sequence is computed by averaging the distortion activities for all QA-GOPs as presented in equation (4.11):

$$DA^{seq} = \frac{1}{N_{QA-GOP}} \sum_{m=1}^{N_{QA-GOP}} DA^{GOP}_m \tag{4.11}$$

where N_{QA-GOP} is the number of QA-GOPs in the sequence.

4.2.7. Flickering Distortion Computation

The main objective of this module is to measure the flickering distortion in the synthesized video sequence. By studying the temporal gradient fluctuation for the samples along a specific S-T tube, the flickering distortion can be measured as:

$$DF(x_t, y_t) = \sqrt{\frac{\sum_{t=i-N+1}^{i+N} \Phi(x_t, y_t, t) \cdot \Delta(x_t, y_t, t)}{2N}} \tag{4.12}$$

where Φ express the potential sensibility function and Δ the flickering distortion intensity.

The flickering distortions are perceivable if: i) the directions of the temporal gradients for the reference ($\vec{\nabla} I$) and synthesized ($\vec{\nabla} \tilde{I}$) sequences are different and non-null (first conditions in 4.13); ii) the synthesized sequence has temporal variations and; iii) these last changes are perceivable, *i.e.* the temporal gradient magnitude is over a perceivable threshold (μ) (last condition in 4.13).

$$\Phi(x, y, t) = \begin{cases} & \vec{\nabla} I_{x,y,t}^{temporal} \times \vec{\nabla} \tilde{I}_{x,y,t}^{temporal} \leq 0 \\ 1, & \wedge \ \vec{\nabla} \tilde{I}_{x,y,t}^{temporal} \neq 0 \\ & \wedge \ \left| I(x, y, t) - \tilde{I}(x, y, t) \right| > \mu(x, y, t) \\ 0, & otherwise \end{cases} \tag{4.13}$$

A. Potential Sensibility Function Computation

The perceivable threshold, μ, denotes the limit above which the HVS is capable of perceiving the absolute luminance difference as typically depends on the content, in this case in space and time.

To compute the perceivable threshold, a method so-called *edge emphasized Just Noticeable Difference (JND) model*, has been adopted [81]. As illustrated in Figure 4.6, the computation of the JND values following this model can be split into two main parts (red dashed boxes): i) JND model computation; and ii) Edge emphasizing post-processing.

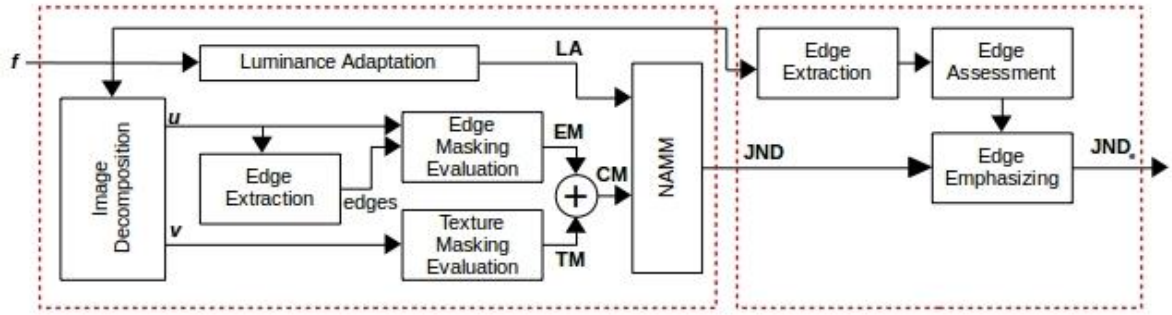


Figure 4.6 – Edge emphasized JND model architecture.

A.1 JND Model Computation

The JND model computation relies upon two main factors: luminance adaptation (LA) and contrast masking (CM). The CM is expressed as a sum of two masking effects, notably edge masking (EM) and texture masking (TM).

As shown in Figure 4.6, the JND model computation proceeds as follows:

- ***Luminance Adaptation*:** Luminance adaptation, as expressed in equation (4.14) [82], attempts to model the HVS retina adaptation by reproducing relative changes described by the Webers' law [64], and ultimately addressing the desired luminance masking effect.

$$LA(x,y) = \begin{cases} 17 \times \left(1 + \sqrt{\frac{f(x,y)}{127}}\right) + 3 & \text{if } f(x,y) \leq 127 \\ 3 \times \frac{(f(x,y) - 127)}{128} + 3 & otherwise \end{cases} \tag{4.14}$$

- ***Image decomposition*:** Contrast masking is an effect associates to the fact that the HVS is less sensitive in areas with high spatial variation and more sensitive in large smooth areas. To address these distinct characteristics, the input frame is decomposed into two sub-images, the structural (u) and textural (v) sub-images, in a way that the sum of both sub-images constitutes the original image, *i.e.* $f = u + v$, as proposed in [81]. The structural image, also called cartoon image (see Figure 4.7 b)), gained this name due to its coarse granular description of the image f, meaning that is piecewise smooth, while containing at the same time sharp edges like in a cartoon. On the other hand, the textural image contains only the fine-scale details, usually with some oscillatory nature (see Figure 4.7 c)).

a) b) c)

Figure 4.7 – Image decomposition: a) Image; b) Structural; c) Textural.

The image decomposition is performed using the TV-L1 model [83], which offers a solution to the standard Total Variation (TV) denoising problem defined as:

$$\min\{\|s(u)\|_A \mid \|t(u,f)_B\| \leq \sigma\} \tag{4.15}$$

where the first and second terms represent the regularization and fidelity terms, respectively; notably, $\|s(u)\|_A$ is small for regular signals (u) but large for irregular signals (v), and

$\|t(u,f)_B\| \leq \sigma$ forces u to be very close to f. Euclidean norms (or semi-norms) are chosen for each term accordingly to the desired application; hence, for the first and second terms, the L_1 norm is chosen as ideally it allows the regularization term to preserve the geometric features [84], and the fidelity term to favor exclusively the texture (sharp variations) [83]. Typically, $\|s(u)\|_A$ is the integral of the absolute generalized derivative of *u*, where *u* is contained in a space of functions of bounded variation (BV) denoted as G space [83]. For discrete signals, the L1 norm of the gradient is used as the fidelity term.

Considering the past definitions, in addition to both terms being convex in *u*, the Chambolle-Pock primal-dual algorithm [85] solves the optimization problem by applying in each step/iteration two proximal operators: $(I + \sigma\partial F^*)^{-1}$ reduces to pointwise Euclidean projectors onto L2 balls; $(I + \tau\partial G)^{-1}$ is the shrinkage operation, notably:

$$p = (I + \sigma\partial F^*)^{-1}(\tilde{p}) \Leftrightarrow p_{i,j} = \frac{\tilde{p}_{i,j}}{\max\left(1, \left|\tilde{p}_{i,j}\right|\right)} \tag{4.16}$$

$$u = (I + \tau\partial G)^{-1}(\tilde{u}) \Leftrightarrow u_{i,j} = \begin{cases} \tilde{u}_{i,j} - \tau\lambda & \text{if} \quad \tilde{u}_{i,j} - g_{i,j} > \tau\lambda \\ \tilde{u}_{i,j} + \tau\lambda & \text{if} \quad \tilde{u}_{i,j} - g_{i,j} < -\tau\lambda \\ g_{i,j} & \text{if} \quad \left|\tilde{u}_{i,j} - g_{i,j}\right| \leq \tau\lambda \end{cases} \tag{4.17}$$

where, τ and λ controls the tradeoff between both fidelity and regularization terms (closeness to f).

- ***Edge extraction***: From the extracted structural image, an edge map is computed. This map is obtained by using the Canny edge detection algorithm with Otsu's method to improve its outcome. Otsu's method performs an automatic threshold selection for image segmentation, meaning that this method is used to obtain the optimal threshold to be used in the Canny edge detection algorithm. Canny edge detection algorithm is a five-step process, notably: i) smoothing (Gaussian filter); ii) image gradients computation; iii) non-maximum suppression (edge thinning) [86]; iv) double thresholding; and v) edge tracking by hysteresis [87].
- ***Contrast Masking*****:** Contrast masking is defined as the sum of two similar masking effects which occur in high textured areas and edges. However, textured areas present a more intense masking effect than edge areas.
- ***Edge Masking Evaluation*****:** Edge masking evaluation, as expressed in equation (4.18), is the product of content variation (C_s) times a β constraint, that is set to 0.117 as in [82], times the edge masking weight value (w_e) of 1.

$$EM^u(x,y) = C_s^u(x,y) \cdot \beta \cdot w_e \tag{4.18}$$

C_s denotes the maximal weighted average of gradients, also known as mean filter, around a pixel. Gradients are computed independently for each sub-image, using four directional high-pass filters for texture detection (g_k , $k \in [1,2,3,4]$) as illustrated in Figure 4.8.

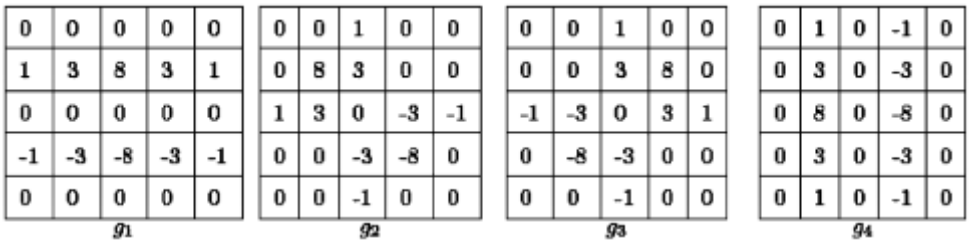


Figure 4.8 – Directional high-pass filters for texture detection.

- ***Texture Masking Evaluation*:** The texture masking evaluation follows the same approach as the edge masking evaluation, as it can be seen in equation (4.19). However, due to the characteristics of the HVS being less sensitive to artefacts in high textured areas, the texture masking weight parameter (w_t) is assigned 3.

$$TM^v(x,y) = C_s^v(x,y) \cdot \beta \cdot w_t \tag{4.19}$$

- ***Nonlinear Additivity Model for Masking (NAMM)*:** To obtain the JND model, both LA and CM are joint using the nonlinear additivity model for masking (NAMM) [82]. NAMM, as expressed in equation (4.20), use a gain reduction factor (C^{lc}) which is related to overlapping effect in masking, and is also related to the viewing conditions (ambience lighting, display device, viewing distance, etc.). $C^{lc}$is set to 0.3 to adopt the same value as in [82].

$$JND = LA + CM - C^{lc} \times \min\{LA, CM\} \tag{4.20}$$

A.2 Edge-based JND Emphasizing Post-Processing

The resulting JND map is finally post-processed for edge emphasizing as a way to increase the sensibility around the object edges; this is expected to deliver better results as the JND map has not been used in the flickering distortion context. This method enhances the JND sensibility at pixel level for the edges in a given frame. This is achieved by performing an **edge extraction** process for the frame and processing the resulting edges using a 8x8 block-based approach. At this stage, **edge assessment** is performed where any block containing more than 48 edge pixels is classified as a texture block and the associated pixels are not emphasized. On the contrary, the remaining edges are used for **edge emphasizing** where the position corresponding previously computed JND values is emphasized to increase the sensitivity of the metric to these areas. Thus, expressing the fact that the user sensitivity is also higher in these areas. This emphasis is performed by multiplying the JND values by 0.1.

B. Flickering Distortion Intensity Computation

To measure the contribution of each sample in terms of flickering distortion intensity, equation (4.21) adopts a squared function that considers both the magnitude of the temporal gradient distortion (numerator), and the temporal masking effect (denominator). A C constant term is added to avoid divisions by zero. The flickering distortion intensity at position (x, y) is computed along the S-T tubes which capture the motion trajectory as follows:

$$\Delta(x,y,t) = \left(\frac{\vec{\nabla}\tilde{I}^{temporal}_{x,y,t} - \vec{\nabla}I^{temporal}_{x,y,t}}{\left|\vec{\nabla}I^{temporal}_{x,y,t}\right| + C} \right)^2 \quad (4.21)$$

C. Flickering Distortion Computation

The flickering distortion per S-T tube is given by the average of all the flickering distortions measured at sample level within the S-T tube blocks as:

$$DF^{tube} = \frac{\sum_{x=1}^{b_w}\sum_{y=1}^{b_h} DF(x,y)}{b_w \times b_h} \quad (4.22)$$

The flickering distortions at QA-GOP level and, subsequently, at video sequence level, follow the DA approach, meaning that the flickering distortion is computed as the arithmetical mean of the *W%* worst S-T tubes for each QA-GOP (see equation (4.23)), and the average of all QA-GOPs flickering distortions, see equation (4.24).

$$DF^{GOP} = \frac{1}{N_W}\sum_{k\in W} DF_k^{tube} \quad (4.23)$$

$$DF^{seq} = \frac{1}{N_{QA-GOP}}\sum_{m=1}^{N_{QA-GOP}} DF_m^{GOP} \quad (4.24)$$

4.2.8. Overall Distortion Computation

The final module computes the overall distortion for the synthesized sequence, as shown in equation (4.25), by combining both the spatio-temporal activity and the flickering distortions computed in the previous modules. According to the authors [80], the relationship/correlation between the flickering distortion and the scores resulting from subjective assessment may be made more linear (as typically desired), by using a base 10 logarithmic scale for DF before combination with DA as follows [80]:

$$D = DA \times log_{10}(1 + DF) \quad (4.25)$$

CHAPTER 5

3D Synthesized Views Relevant Databases

As mentioned before, subjective experiments are the most reliable way to perform the assessment of image and video quality. The scores collected in these subjective experiments may be stored in image and video quality assessment databases as they provide key data to develop effective objective quality metrics, this means metrics with high correlation with the subjective scores.

In fact, the development of image and video quality objective assessment metrics requires image and video databases including the processed data and the corresponding subjective scores, notably for testing and validating novel objective quality assessment metrics. Often, these metrics are developed under some specific constraints based on assumptions and modelling the signals to exploit some desired features. This results in objective metrics that perform well for some specific types of degradation and artefacts, while not performing so effectively for other types of distortion and artefacts.

This chapter presents four databases with image and video synthesized views and associated subjective scores, which may be used in this Thesis to assess 3D Image Quality Assessment (IQA) metrics. The databases will be individually identified by their origin, image/video sequences, artificially introduced artefacts, view synthesis algorithms, subjective assessment methodology, and finally any other particular features. All sequences are summarized in a table at the end of each section.

5.1. Synthesized Image Quality Assessment Databases

Traditional 2D image quality assessment databases consist on a collection of original and processed/distorted images with subjective scores obtained through experimental assessment protocols. These experiments collect scores from a population of subjects, resulting in a mean opinion score (MOS) for each displayed image or pair of images. Synthesized image quality assessment databases are built similarly, differing only on how the processed images were obtained. Unlike 2D images, synthesized images are not captured directly using an optical-based device (*e.g.* camera); instead, the images correspond to synthesized views generated by some synthesis algorithm based on some adjacent captured views. Notably, it is a common practice to use Depth Image-Based Rendering

(DIBR) algorithms for Multiview plus Depth (MVD) video sequences to create the synthesized frames which should build the synthesized image quality assessment databases.

5.1.1. Media Communications Lab 3D Database

This database, denoted as MCL-3D, has been developed by the Media Communications Lab at the University of Southern California, USA [88]. Originally designed to assess stereoscopic image quality metrics, its content results from the stereoscopic image pair synthesis system shown in Figure 5.1. The distortion types addressed in this database are related to compression, transmission and imperfect rendering artefacts.

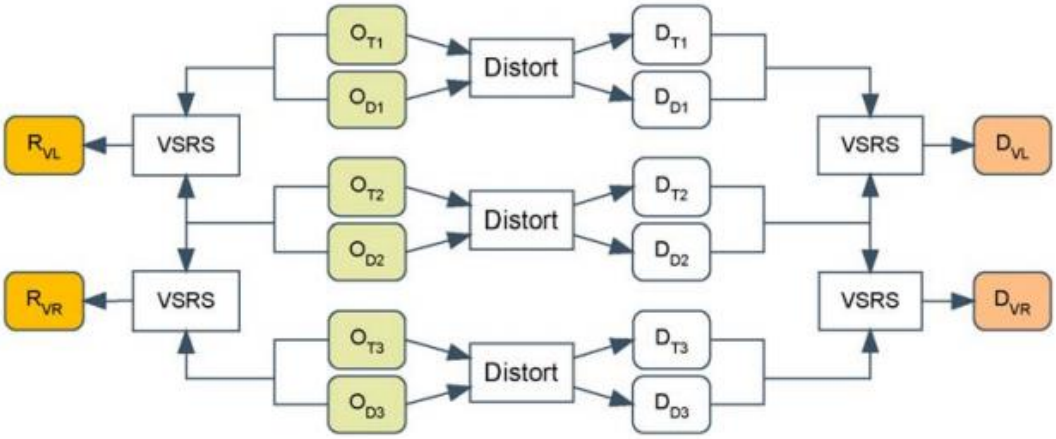


Figure 5.1 – MCL-3D database content and processes [89]. O, R and D refer to the original, reference and distorted data/views, respectively; lower script T and D refer to texture and depth data.

This system contains four main types of image data:

- ***Original data/views*:** The MCL-3D database includes nine original multiview plus depth images, where two thirds have a 1920×1080 resolution and the last third has a 1024×768 resolution. Each image was captured from three adjacent viewpoints and is accompanied by its depth data.
- ***Distorted data/views*:** Besides the original views, this database includes artificially distorted views, where the original views are distorted according to some relevant type of distortion, denoted as *Distort* in Figure 5.1. This distortion function allows studying the quality impact of different types of distortions; for example, these distortions allow to simulate the effect of some typical compression and transmission related distortions on the synthesized views. The considered six distortion types are Gaussian Blur, Sampling Blur, JPEG coding, JPEG 2000 coding, Additive White Gaussian Noise and Transmission Losses, each one with four different levels of distortion intensity; these distortions are applied to both the texture and depth data.
- ***Reference synthesized data/views*:** Reference views correspond to the synthesized views using the original neighbouring data/views and some selected synthesis algorithm. Song *et al* [89] set forth that the View Synthesis Reference Software (VSRS) algorithm [90] offers a near perfect stereoscopic image synthesis; for this reason, the MCL-3D database includes the pair of VSRS synthesized views, using the original neighbouring data, and takes them as reference views for the synthesized views created based on distorted neighbouring views; hence, each pair represents the reference view left (R_VL) and view right (R_VR) images. The reference synthesized views are here created by using the MPEG VSRS 3.0/3.5 synthesis algorithm [90].

- ***Synthesized data/views*:** For this database, a couple of approaches were considered regarding the synthesized data/views: i) compression and transmission related artefacts; and ii) imperfect rendering artefacts. While for the first approach the synthesized data/views are created using one synthesis algorithm (VSRS) applied to a pair of distorted neighbouring views, the second approach uses original data/views to synthesize the stereoscopic pair of views. For this database, five different DIBR synthesis algorithms were considered:
  - **MPEG VSRS 3.0/3.5:** The first synthesis algorithm is the DIBR MPEG VSRS [90]. This algorithm is the only one that uses distorted data to synthesize views; in particular, it uses texture distorted data with its depth original data, texture original data with its correspondent depth distorted data, and both texture and depth distorted data, to create the synthesizes views. This algorithm includes an inpainting method to fill any holes in the synthesized views.
  - **DIBR without hole-filling:** In this synthesis solution, the views are rendered using MPEG VSRS with the inpainting method disabled [91], which means the synthesized views may have holes and line scratches.
  - **DIBR with depth map smoothing:** In this case, the depth map is pre-processed to filter out insignificant depth discontinuities, crop the borders and perform the necessary interpolation to reach the required original size [92]; this algorithm may introduce shifting artefacts.
  - **DIBR with horizontal inpainting:** This solution is based on the *DIBR with depth map smoothing* solution but now using also an inpainting technique named *fast marching inpainting* method [93]; this algorithm may introduce shifting artefacts and blurring discontinuities around the scene objects.
  - **DIBR with hierarchical hole-filling:** This solution uses a lower resolution estimate of the synthesized textured view in a pyramid like structure to fill the missing texture samples associated to disocclusions in the synthesized view [94] [95]; this algorithm may introduce geometric distortions [96].
- ***Subjective protocol and conditions*:** A pairwise comparison (PC) subjective assessment method (double stimulus and discrete scores) was adopted to assess the subjective quality of the synthesized stereoscopic image pairs. In this context, the two stereo pairs are observed simultaneously and the assessor is asked to select the stereo pair exhibiting the best perceivable quality. A stereo pair is scored against the other, thus accumulating comparison scores across multiple assessment rounds. Subjective testing was performed with 270 subjects, 32 experts and 236 naïve, in a ITU-R Rec.BT.500 environment, using a 46.9" LG 47LW5600 display. Each view-pair was assessed by 30 subjects, where the highest and lowest 10% for each stereo pair score were discarded and treated as outliers. Finally, a MOS score was computed as the arithmetical mean of all subjective scores obtained from the multiple subjects.
- ***Database elements*:** In summary, the MCL-3D database includes with the following elements:
  a) Set of original data/views in a folder named as "reference_texture_and_depth". The original data refers to the images captured for three different views, including both texture and depth data for each of the given sequences.

b) Set of distorted data/views in a folder named as “distorted_texture_and_depth”. Distorted data image files include both the texture and depth data using several types and levels of distortion.
c) Set of synthesized views using the various synthesis algorithms in a folder named as “rendered_left_and_right”.
d) MOS scores in a couple of .xlsx files in a folder named as “scores”. The first file is named as “MOS_fr” and includes the subjective scores for the synthesized views generated with the distorted data; the second file, denoted as “MOS_nr”, contains the MOS for all distorted views.

This database is publicly available at the University of Southern California Media Communications Lab website [88] and can be used for research and development on objective image quality metrics.

5.1.2. IRCCyN-IVC DIBR Image Quality Assessment Database

The IRCCyN-IVC DIBR_Images is a DIBR image database designed by the Images and Video-communications research group at the *Institut de Recherche en Communications et Cybernétique de Nantes* (IRCCyN-IVC) [97]. This database was designed to allow the performance assessment of DIBR algorithms, particularly for stereoscopic-related IQA metrics. The IRCCyN-IVC DIBR_Images database is characterized by:

- ***Original data*:** It includes three different 1024×768 resolution multiview plus depth sequences, with diverse extrinsic and intrinsic camera parameters, *i.e.* number of cameras in the array, and different spacing.
- ***Reference data/views*:** Three reference images, for each original data sequence, were extracted from distinct viewpoints in two separated time instants (*t1* and *t2*); notice that the meaning of reference views here is totally different from the previous database.
- ***Synthesized data/views*:** This database only uses a single view to synthesize other views, what significantly increases the probability of having hole artefacts around object edges. For *t1*, the collected reference images use the left view to render the centre and right views while, for *t2*, the right view is used to synthesize the centre and left views. These views are synthesized considering seven different DIBR algorithms, notably:
  - **Fehn cropped:** Pre-processes image depth map to filter out any insignificant depth discontinuities, crop the borders, and interpolate to reach the original size [92]; this algorithm may introduce shifting artefacts.
  - **Fehn interpolated:** Same as the Fehn cropped synthesis algorithm but substituting its inpainting method by the fast marching inpainting method [93]; this algorithm may introduce blurring discontinuities around scene objects.
  - **MPEG VSRS:** Off-the-shelf MPEG VSRS algorithm [98]; this algorithm may introduce blurry regions in synthesized views.
  - **Müeller:** Post-processes depth discontinuities in rendered views (smoothing filter) and includes a new hole filling method that is helped by depth information [99].
  - **ICME:** Holes in the synthesized view are patched using textured synthesis patches with patches computed by estimation derived from spatially adjacent original textures [100].

- **ICIP TMM:** Exploits temporal information from depth components to improve the view synthesis in areas with disocclusions [101].
- **Holes:** Off-the-shelf MPEG VSRS with inpainting method disabled [98]; this algorithm may produce holes and line scratches in the synthesized view.

- ***Subjective protocol and conditions*:** Image quality subjective tests were conducted in an ITU conforming test environment, notably following ITU-R BT.500 recommendations. Stimuli were displayed in a TVLogic LVM401W. A total of 43 subjective tests (all naïve subjects) were carried out using two testing methodologies, pair comparison (PC) and ITU-R absolute category reference with hidden reference (ACR-HR) (single stimulus and discrete scores). The observers were at a 4H distance and the 96 testing sequences were assessed in a 1 to 5 discrete scale. Using all subjective scores, the arithmetical mean was computed to obtain the final MOS score for each assessed asset in the database.
- ***Database elements*:** In summary, the IRCCyN-IVC DIBR_Images database includes:
  a) Set of images related to the reference and synthesized views.
  b) MOS scores in .xls format files, one for each adopted subjective assessment methodology.

This database considers several view synthesis algorithms, thus offering rich data about the impact of inpainting methods on the observers QoE. However, it lacks comprehensive information about other typical distortions that usually result from lossy coding compression techniques (e.g. JPEG) or transmission errors. This is an openly available database that can be found at the Image and video-communication (IVC) research group website at [97].

5.2. Synthesized Video Quality Assessment Databases

Synthesized video quality assessment databases are built with original and processed/synthesized video sequences with the corresponding subjective scores obtained through subjective assessment protocols. These databases are used to study the reliability of objective video quality metrics when evaluating virtual synthesized views in a multiview plus depth video context. As mentioned in Chapter 2, DIBR algorithms involve geometric transformations, thus resulting in new types of distortions and impairment artefacts that are not typically considered by 2D video quality assessment metrics, and also not present in 2D video quality assessment databases, such as flickering distortions. Therefore, the availability of DIBR video quality assessment databases is critical to study video quality assessment metrics for synthesized views.

5.2.1. IRCCyN-IVC DIBR Video Quality Assessment Database

The IRCCyN-IVC DIBR video quality assessment database, denoted as DIBR_Videos, considers several DIBR algorithms to study the performance of video quality metrics for synthesized views [102]. This DIBR videos database is characterized by:

- ***Original data/views*:** Includes three 1024x768 resolution multiview plus depth video sequences with 6 seconds each, with a frame rate between 15 and 30 frames per second (fps), and three different viewpoints (left, centre and right).

- ***Reference data/views*:** The three viewpoints in the original video sequences are taken as reference views in this database.
- ***Distorted reference data/views*:** The left view of each video sequence (4:2:2 format) is intentionally distorted by coding it with a H.264/AVC encoder and applying three different quantization parameters, notably 26, 34 and 44. The resulting distorted views enable the DIBR_Videos database to assess video quality assessment metrics considering video coding-related artefacts.
- ***Synthesized views*:** Seven different DIBR algorithms are used to synthesize the additional viewpoints. As already mentioned in the previous subsection, here the left original view is used to synthesize the central and right view, and the right original view is used to the central and left views. The selected DIBR algorithms for the view synthesis process are mainly related to different hole filling strategies in the synthesized views; the synthesis algorithms mentioned above for the IRCCyN-IVC DIBR Image Quality Assessment Database are also used here.
- ***Subjective protocol and conditions*:** An absolute category reference with hidden reference (ACR-HR) subjective assessment method (single stimulus and discrete scores) was adopted as the subjective assessment methodology. The ACR-HR methodology involves observers scoring test sequences using a discrete scale and one sequence at a time; in this particular case, each sequence was assessed by 32 subjects (all naïve) in a 1 to 5 scale. In addition to the synthesized views, also the reference views are scored to enhance the database reliability by computing the differential mean opinion score (DMOS) over the scored MOS, for each testing sequence. The subjective tests were performed in an ITU-R BT.500-11 recommendation environment, where the stimuli were displayed at a 4H distance on a TVLogic LMV401W monitor.
- ***Database elements*:** In summary, the IRCCyN-IVC DIBR Video Quality Assessment database includes:
  a) Set of videos with the reference, distorted and synthesized views.
  b) Subjective scores in a .xls format file containing the MOS and DMOS for all tested sequences.

The DIBR_Videos database in [102] is indeed a good source of information regarding the synthesis performance of different hole filling algorithms; however, DIBR algorithms usually use a pair of adjacent views (and not only a single one as here) to synthesize the desired viewpoint, what greatly reduces the distortion at the hole filling stages. Additionally, distortions concerning compression of texture and/or depth maps are not considered in this database, thus resulting in a poor synthesized video quality assessment database regarding video coding effects.

5.2.2. SIAT Synthesized Video Quality Assessment Database

The Shenzhen Institute of Advanced Technology (SIAT) in China developed a synthesized video quality database [103] which is characterized by:

- ***Original data*:** It includes ten different multiview plus depth video sequences that were captured from three adjacent viewpoints. As described in Table 5.1, these video sequences have 1024×768 and 1920×1088 spatial resolution.
- ***Reference views*:** The central views of each video sequence are used as reference views.

- ***Distorted data*:** Motivated by the non-existence of DIBR video databases considering the impact of the quantization process happening while coding, this database includes for each sequence 14 different texture/depth quantization combinations. Video sequences were coded using the H.264/AVC coding standard in a 4:2:0 format, in particular using the reference software 3DV-ATM v10.0, which is the H.264/AVC Test Model available at the Nokia Research Center website [104]. As recommended by the Video Quality Experts Group (VQEG), the selected quantization parameters include a wide range of values; at the same time, the quality gap between each decoded test sequence should not be too small to make difficult assessing the differences. This database has divided the synthesized videos into four categories: i) Uncompressed Texture and Uncompressed Depth ($U_T U_D$); ii) Uncompressed Texture and Compressed Depth ($U_T C_D$); iii) Compressed Texture and Uncompressed Depth ($C_T U_D$); and iv) Compressed Texture and Compressed Depth ($C_T C_D$). Naturally, the coding of the various components was performed before the view synthesis. As illustrated in Table 5.1, the SIAT database has been built considering VQEG recommendations and using appropriately different QPs for the texture and depth data.
- ***Synthesized views*:** Views were synthesized using the VSRS-1D-Fast mode included in the 3D-HEVC reference software 3D-HTM v8.0 provided by the Fraunhofer Heinrich-Hertz-Institut [105].

Table 5.1 – SIAT Synthesized Video Quality Database: content and coding characteristics.

Sequence	Spatial Resolution	Input Views	Output View	Depth QP ($\boldsymbol{U_T C_D}$)	Texture QP ($\boldsymbol{C_T U_D}$)	QP Pair ($\boldsymbol{C_T C_D}$)
BookArrival	1024×768	$6 - 10$	8	28,36,40,44	28,34,38,42	(22,26),(28,32),(34,36),(38,40),(42,44)
Balloons	1024×768	$1 - 5$	3	32,36,40,46	24,32,38,42	(24,32),(28,36),(32,40),(40,42),(42,46)
Kendo	1024×768	$1 - 5$	3	32,38,44,48	24,28,32,40	(24,32),(32,34),(36,38),(40,44),(44,46)
Lovebird1	1024×768	$4 - 6$	5	36,38,40,48	28,30,34,38	(28,36),(30,40),(34,44),(38,48),(42,50)
Newspaper	1024×768	$2 - 4$	3	28,36,44,50	24,30,34,38	(28,32),(32,40),(38,44),(36,50),(42,48)
Dancer	1920×1088	$1 - 9$	5	24,28,40,45	28,32,40,44	(24,20),(30,24),(32,28),(32,40),(44,35)
PoznanHall2	1920×1088	$5 - 7$	6	28,32,40,46	24,28,32,38	(24,28),(26,32),(34,36),(32,40),(40,42)
PoznanStreet	1920×1088	$3 - 5$	4	32,38,44,48	26,30,38,42	(22,28),(26,40),(30,44),(34,48),(42,35)
GT Fly	1920×1088	$1 - 9$	5	28,36,44,48	24,36,40,44	(24,28),(32,36),(34,38),(40,44),(44,48)
Shark	1920×1088	$1 - 9$	5	28,36,40,44	24,32,36,40	(24,28),(32,36),(36,40),(40,36),(42,48)

- ***Subjective protocol and conditions*:** Subjective experiments were split into a two-session experiment, with 84 and 56 videos assessed in the first and second sessions, respectively. Using 56 naïve subjects, each test video was evaluated 40 times using a 0 to 100 continuous scale while

adopting the ITU-R ACR-HR assessment methodology. Both experimental sessions were conducted as recommended in Rec. ITU-R BT.500, and stimuli were displayed on a LG 42LH30FR monitor with FHD (1920x1080) resolution. For each subjective score, a differential score was derived by subtracting the scores associated to the reference sequence for the same session. Subsequently, differential scores were normalized by the statistical standard score computation denoted as *z-score*. Regarding the SIAT database, it is specified that the reference, mean and standard deviation values are respectively described as the differential score, 0.5, and $1/6$. Subsequently, DMOS values were computed for each test sequence as the arithmetical mean of all normalized differential scores. The SIAT database is mainly focused on video coding as the source of distortions regarding the video quality assessment of synthesized views. This is an important and realistic context when applying this database for the study and design of objective quality assessment metrics for 3D synthesized views.

- ***Database elements*:** In summary, the SIAT database includes:
  a) Set of original, distorted views and including both the texture and depth maps.
  b) Synthesized views for each study case as presented in Table 5.1.
  c) MOS and DMOS values in a .xlsx format file.

This database is publicly available at the Center for Digital Media Computing, Shenzhen Institute of Advanced Technology website [103] and can be used for research and development of objective video quality metrics. It is important to highlight that this is the only database with sequences illustrating temporal distortions, such as flickering, which are a remarkable source of subjective annoyance in 3D synthesized views.

Openly available synthesized video quality databases are a scarce resource in the QoE scientific community, as there are only two available databases addressing video content: the SIAT and the DIBR_Videos databases. Despite this limitation, the databases address different aspects: While the first database considers the effects of several different inpainting methods, and thus hole filling solutions, used in the DIBR algorithm, the second database considers abundant variations in terms of coding distortions.

CHAPTER 6

Quality Metrics Performance Assessment

There are several algorithms to perform image quality assessment, exploring different domains, techniques and methodologies as well as diverse psychophysiological and psychological approaches and assumptions to this challenge. The idea is to find a method that assigns objective scores to degraded video sequences, while considering the human perception of that specific video sequence. Note that, subjective and objective quality scores may have rather distinct values for exactly the same test parameters and sequences, as these depend entirely on the used subjective protocol and objective metric scale. For this reason, to evaluate the performance of an objective metric, objective scores need to be modelled to fit the same scale (and to the same statistic) as the collected subjective scores. The scope of this Chapter is the performance assessment of the designed and implemented metric described in the Chapter 4, and it is split into four Sections, notably: performance assessment workflow; VQA metric configuration profiles; study of the different VQA metric configuration profiles; performance assessment comparison of the objective VQA metric implemented.

6.1. VQA Metric Performance Assessment Workflow

The performance assessment workflow has three main part: i) 3D Synthesized Views Video Quality Assessment Database; ii) Objective Quality Assessment Metric; iii) Quality Metrics Performance Assessment. The 3D synthesized views video quality assessment database used for the performance assessment procedure was the SIAT synthesized video quality database described in Section 5.2.2 [103], which is a 10-sequence database with 14 different synthesized views per sequence.

The objective quality metric assessed in this Chapter follows the algorithmic description presented in Chapter 4. Also, several configurations of this metric were evaluated; to study the impact of different methods and parameters on the overall performance and on its components.

The objective quality metrics performance assessment makes use of a non-linear regression process to study the correlation between subjective and objective results. Following the ITU-T recommendation [106], the VQA metric scores (Q) are transformed (fitted) into predicted difference mean opinion score

($DMOS_p$) by applying the non-linear least squares regression analysis to the model defined in (6.1); notably, the β parameters are computed using the second-order optimization model [107] (Newton-Raphson) for a confidence interval greater than *95%*.

$$DMOS_p = \frac{\beta_1}{1 + e^{-\beta_2 \times (Q - \beta_3)}} \tag{6.1}$$

This step is necessary because the subjective and objective scores are scaled differently; thus, objective scores need to be fitted into the subjective scale. As recommended by ITU-T [106], the performance assessment of a VQA metric is based in the following three criteria:

- ***Pearson Linear Correlation Coefficient*:** Pearson correlation coefficient is the most commonly used metric in the image-based QA field to measure the correlation between the subjective and objective scores. This bivariate analysis measures the linear correlation in an *n* number of bi-sampled variables (*X* and *Y*), resulting in a correlation coefficient (ρ) that is computed according to (6.2). Regarding the image-based QA performance assessment, the *X* and *Y* variables represent the DMOS and $DMOS_p$ values, and the resulting score is used to quantify the linear relationship between both variables, *i.e.* subjective and objective scores, to measure the prediction accuracy.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{6.2}$$

- ***Spearman Rank Correlation Coefficient*:** Spearman correlation coefficient is a non-parametric test that makes no assumptions about the form of the relationship (linear, polynomial, etc.) used to measure the prediction monotonicity. Spearman correlation coefficient (r_s) is given by (6.3) where d_i is the difference between the converted ranks derived from the raw scores of two variables (X_i and Y_i), for a n number of samples. In image-based QA, this correlation coefficient quantifies the monotonicity between the objective and the subjective scores, thus enabling the quantification of the monotonic variation behaviour of an objective metric, *i.e.* if the objective scores follow the same crescent or descend behaviour as the subjective assessment.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{6.3}$$

- ***Root Mean Square Error*:** Root Mean Square Error (RMSE) is a widely-used statistical metric that quantifies the error between two variables (*X* and *Y*), by aggregating the magnitude of errors between each variable-sample pair (X_i and Y_i). RMSE is computed as expressed in (6.4), where n represents the number of samples for both the *X* and *Y* variables. Considering image-based QA performance assessment, *X* and *Y* represent the $DMOS_p$ and the DMOS score values and expresses numerically the consistency of an objective QA metric, *i.e.* on average how much error is accumulated between (objective) estimations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{n}} \tag{6.4}$$

6.2. VQA Metric Configuration Profiles

To study how the different approaches and techniques impact the metrics performance, the performance assessment of the objective VQA (see Chapter 4) for synthesized views was evaluated using several configurations. The metric performance depends on three key aspects, notably: i) Motion vector estimation approach; ii) Flickering distortion perception threshold model; iii) Edge detection threshold approach. All configurations were evaluated on the SIAT synthesized video quality database (see Section 5.2.2).

6.2.1. Configuration #1: Motion Vector Estimation Approach (MVEA)

Motion vector estimation has a key role in the objective VQA developed, as it is used to compute this metrics' basic analysis structure, called S-T tubes (Section 4.2.2). As illustrated in Table 5.1, three different approaches were tested with: window size (*w*) and motion vector amplitude penalty function (*PF*) for each configuration.

Table 6.1 – Motion Vector Estimation: configuration characteristics.

Algorithm	Block Matching		
Label	**BM_32**	**BM_64**	**BM_64_CF**
Details	***w*** 32p **PF** MAD	***w*** 64p **PF** MAD	***w*** 64p **PF** MAD + $K \times \left\|\|mv\right\|\|_2$ where $K = 0.05$

6.2.2. Configuration #2: Flickering Distortion Perception Threshold Model (FDPTM)

The flickering distortion perception threshold is computed as described in Section 4.2.7. However, two different hypotheses were considered as models of the perception threshold, notably: i) the JND model as described in the Subsection A.1; ii) the edge emphasized JND model, so-called JNDe, as described in Subsection A.2.

6.2.3. Configuration #3: Edge Detection Threshold Approach (EDTA)

The edge detection is performed using the Canny edge detection algorithm as described in Section 4.2.7. Nevertheless, the following two approaches were considered, one static and another adaptive. The static approach fixes the high and low thresholds used in the Canny edge detection algorithm as 200 and 100 respectively. The adaptive approach is based on the Otsu method, which is used to compute the high threshold value based on the content of the image, see Section 4.2.7 [108]; then, the low threshold is found by multiplying the high threshold by 0.5 as used in [109].

6.3. Performance Study of VQA Metric Configurations

As a baseline, all different configurations described in the previous Section are studied considering the three different score results, notably: i) Distortion Activity; ii) Flickering Distortion; iii) Overall Distortion. This approach promotes an analysis of the influence that different options have in each intermediate distortion score, and how it impacts the overall distortion score. Table 6.2 presents the performance

assessment correlation results of each distortion score for each configuration; hence, Motion Vector Estimation Approach (MVEA), Flickering Distortion Perception Threshold Model (FDPTM) and Edge Detection Threshold Approach (EDTA).

Table 6.2 – Performance Assessment on Relevant Metric Configurations.

Configurations			**Distortion Activity**			**Flickering Distortion**			**Overall Distortion**		
MVEA	***FDPTM***	***EDTA***	*ρ*	*r*	*RMSE*	*ρ*	*r*	*RMSE*	*ρ*	*r*	*RMSE*
BM_32	JND	Static	**0.826**	**0.827**	**0.072**	0.703	0.689	0.091	0.815	0.802	0.074
		Adaptive	**0.826**	**0.827**	**0.072**	0.697	0.685	0.092	0.811	0.800	0.075
	JNDe	Static	**0.826**	**0.827**	**0.072**	0.663	0.651	0.096	0.805	0.797	0.076
		Adaptive	**0.826**	**0.827**	**0.072**	0.665	0.655	0.096	0.807	0.799	0.076
BM_64	JND	Static	0.819	0.815	0.074	**0.711**	0.696	**0.090**	0.819	0.805	0.074
		Adaptive	0.819	0.815	0.074	0.706	0.691	0.091	0.816	0.802	0.074
	JNDe	Static	0.819	0.815	0.074	0.674	0.662	0.095	0.812	0.799	0.075
		Adaptive	0.819	0.815	0.074	0.674	0.664	0.095	0.814	0.801	0.075
BM_64_CF	JND	Static	0.825	0.823	0.072	0.705	**0.701**	0.091	**0.820**	**0.809**	**0.074**
		Adaptive	0.825	0.823	0.072	0.698	0.688	0.092	0.814	0.804	0.075
	JNDe	Static	0.825	0.823	0.072	0.661	0.650	0.096	0.808	0.800	0.077
		Adaptive	0.825	0.823	0.072	0.660	0.647	0.096	0.807	0.798	0.076

The motion vector estimation technique has a direct impact over the S-T tubes creation module, which in turn has repercussion on the metrics performance. As it can be seen in Table 6.2, changing the technical solution of the motion estimation has a considerable level of impact. The MVE has a higher impact over the flickering distortion, which is expected as it represents temporal analysis, meaning that it is highly sensible to the estimations quality.

The second configuration, so-called flickering distortion perception threshold model, refers to the selection of one of the two available JND models. These models express a pixel-wise map of perceivable thresholds ($\mu(x, y, t)$) that explicitly define the level from which the computed flickering effect is perceived by a human observer. The use of different models has impact over the flickering distortion sensibility function, that consequently impacts the flickering distortion and the overall distortion metric. As shown in Table 6.2, the model selection is a factor that has a great impact on the flickering distortions' performance, in the order of 4-5% for the Pearson and Spearman correlation coefficients for all cases.

The edge detection is only performed in the flickering distortion computation module, namely at the JND model computation and the edge emphasizing post-processing. As presented in Table 6.2, the major difference of Pearson value is found for the JND flickering distortion perception threshold model; this means that the edge detection threshold approach have some impact (0.7%) over the flickering distortion performance when the JND is used as the FDPTM. For the Spearman correlation value, the difference between both EDTA in the JND flickering distortion perception threshold model is in the order of 1.3%, which indicate that the adaptive method for this case produce some jittery results.

To understand the impact that each profile has in the overall performance of the objective quality metric, a deeper analysis is performed for the motion vector estimation and edge detection threshold approaches.

6.3.1. Motion Vector Estimation Approach Analysis

To perform a deeper analysis over the results obtained, three different MV histograms were computed for different types of camera motion: i) static camera; ii) camera translation; iii) camera zooming. Thus, different sequences were used: 1) PoznanStreet is a sequence which has a static viewpoint that only captures motion from the objects in scene; 2) PoznanHall2 is a sequence that have a translational (panning) movement in the x axis; 3) GT_Fly is a sequence which has camera zoom. All sequences were used to identify differences of the estimated motion vectors using the three different configurations of the MVEA, as shown in Figure 6.1.

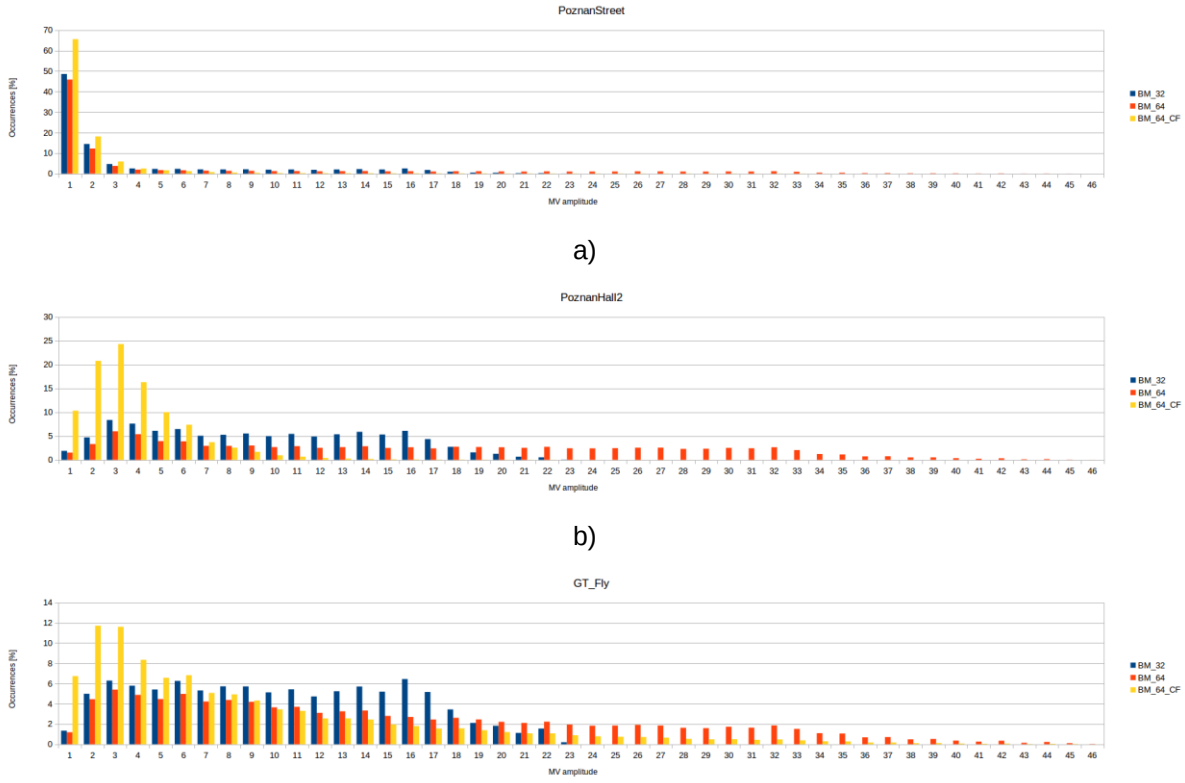


Figure 6.1 – Estimated motion vectors histogram: a) PoznanStreet; b) PoznanHall2; c) GT_Fly.

From Figure 6.1 it is possible to conclude that: 1) both BM_32 and BM_64 approaches are substantially similar even though the BM_64 has a slightly more spread histogram; 2) the size of the search window used in the BM_32 is too small; 3) using the proposed motion vector amplitude penalty function (BM_64_CF) reduce the number of higher amplitude motion vectors, especially for the sequences with a higher amount of motion; 4) BM_64_CF has higher concentrated estimations, not being evenly spread across the histogram.

To examine the quality of the estimations of each approach, the computed S-T tubes had their motion compensated; this means that if estimations are optimal, the video content remains static for the whole QA-GOP. This procedure allows to visually inspect the quality of the motion vectors computed by

each approach. Figure 6.2 shows four images of the first frame of the sequence PoznanHall2, where: a) reference frame; b), c) and d) are the MC frames using the MVs computed by the respective BM_32, BM_64 and BM_64_CF approaches.

a) b) c) d)

Figure 6.2 – Motion compensated blocks: a) Reference frame; b) BM_32; c) BM_64; d) BM_64_CF.

The worst quality was obtained for the BM_64 approach, which presents blocking artefacts as shown in Figure 6.2 c). This effect is present when the block for which the motion is estimated is in a smooth area, but gets worse when areas are bigger. Additionally, BM_32 shown the same behaviour but in a smaller area due to the smaller window size. BM_64_CF was the best one, producing really consistent estimations. Therefore, it can be concluded that simply using the MAD criterion for motion vector estimation can lead to lower quality in comparison with better motion vector modelling criteria such as MAD plus the addition of a cost function that increase the penalization over great distances, or using an adaptive support-weight window penalizing function [110]; notably, MAD can create jittery S-T tubes and poor object tracking, which may also lead to inaccurate S-T tubes exclusion. Independently to the considered approach, some circumstances were found to drive the lack of estimation quality, notably:

- ***Smooth areas*:** Smooth areas are highly constant areas where gaussian noise has a great impact in the estimation process. The addition of the penalizing term ($K \times ||mv||_2$) only mitigates this issue.
- ***Overexposed/Underexposed areas*:** Over/underexposed areas are areas in the frame that have a really high/low luminance samples which make them highly sensitive areas, leading to poor motion vector estimation quality.
- ***Blurred frames*:** Some frames were found to be blurry. This effect constitutes a source of error to estimate the motion, producing jittery S-T tubes.
- ***Shadows*:** Shadows created by scene objects also affected by the quality of the motion vectors, as these lead the MVE to follow their behaviour and not following the objects motion where these shadows are cast upon. This leads to inconsistent tracking of motion across the scene and faulty S-T tubes.
- ***Reflective surfaces*:** Reflective surfaces are seen as luminance and chrominance changes over time even when the object remains static, providing poor motion estimation quality that in consequence produces faulty S-T tubes.
- ***Occlusion/Disocclusion*:** The block matching algorithm cannot deal with occlusions or disocclusions, as it tries to find the most similar block in a given neighbourhood.

6.3.2. Edge Detection Threshold Approach Analysis

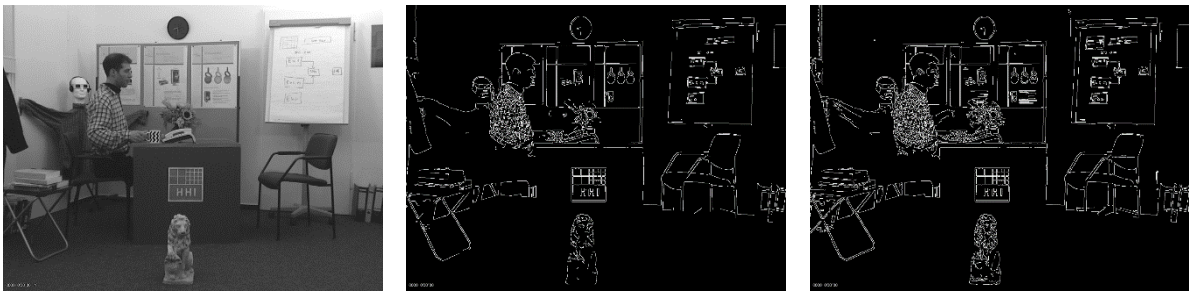
We will now examine where this edge detection threshold has influence, in the edge masking evaluation and in the edge emphasizing post-processing technique. Each one has different input frames, one is the structural image and the other is the luminance samples of the frame being analysed. The structural image, as described in Section 4.2.7, is the image obtained of the decomposition process, described as piecewise smooth, while containing at the same time sharp edges, as shown in Figure 6.3 a). To illustrate the differences between the two threshold approaches, the edge detection for the first frame of the BookArrival sequence was performed, and the results in Figure 6.3 b) and c) that respectively represent the static and adaptive thresholds computed edge maps. This sequence was chosen because it has a lot of edges and textured areas.

a) b) c)

Figure 6.3 – BookArrival: a) Structural image; b) Edge map (static threshold); c) Edge map (adaptive threshold).

As shown in Figure 6.3, there are differences between both approaches, notably the use of the adaptive threshold shows more edges than the edges obtained with the static threshold. These cases have repercussion in the edge masking evaluation, increasing the threshold sensibility in some areas of the frame.

Edge emphasizing post-processing technique is used to lower the sensibility threshold around objects boundaries, which is where the flickering distortions usually appear. The way that this procedure is performed, is to detect the edges in a frame and multiplying their JND value to 0.1, lowering them to a tenth of their original value. Therefore, it is expected that the edge detection is a key part of the quality evaluation metric. To examine the difference between each threshold computation solution, the first frame of the BookArrival sequence was chosen based in the same reason described above (Figure 6.4 a), and both static (b) and adaptive (c) methods were applied.



a) b) c)

Figure 6.4 – BookArrival: a) Luminance samples; b) Edge map (static threshold); c) Edge map (adaptive threshold).

The difference between both solutions is perceivable, adaptive threshold classifies more pixels as edges than the static approach. Considering these results with the performance results presented in Table 6.2, it is possible to conclude that the number of edge marked pixels is inversely proportional to the performance results.

Considering all possible configurations, the best performed configuration is found for: BM_64_CF using the JND model with a Canny threshold operator fixed, which will be used on the next Sections and called *proposed VQA*.

6.4. Performance Assessment of the proposed video quality metric

The performance assessment of the proposed objective video quality assessment metric developed for synthesized views was compared with other 2D and 3D objective quality metrics on the SIAT synthesized video quality database (Section 5.2.2).

6.4.1. Comparison to 2D Objective Quality Metrics

This section presents the performance assessment of the 2D objective quality metrics referred in Section 3.2, but also adds the following: MSSIM [111]; UQI [112]; IFC [113]; NQM [114]; WSNR [115]; and SNR. The overall performance assessment results for the SIAT synthesized video quality database are shown in Table 6.3.

Table 6.3 – Performance Comparison of Objective Video Quality Assessment: 2D VQA.

VQA	ALL DATA		
	ρ	*r*	*RMSE*
MSE	0.653	0.631	0.097
PSNR	0.650	0.627	0.098
SSIM	0.581	0.546	0.104
MSSIM	0.748	0.736	0.085
VSNR	0.678	0.667	0.094
VIF	0.631	0.629	0.100
VIFP	0.658	0.630	0.097
UQI	0.477	0.459	0.113
IFC	0.477	0.459	0.113
NQM	0.554	0.515	0.107
WSNR	0.620	0.588	0.101
SNR	0.759	0.718	0.084
VQM	0.674	0.665	0.095
S-MOVIE	0.705	0.696	0.091
T-MOVIE	0.518	0.461	0.110
MOVIE	0.679	0.660	0.094
Proposed	**0.820**	**0.809**	**0.074**

As shown in the Table 6.3 quality metrics which are a mixture of signal and perceptual principles, such as MSSIM and SNR, have shown to perform quite well. Perceptual-based 2D objective quality metrics have proven to perform worse, as some assumptions are made based on some specific 2D cases and

do not consider the temporal-based distortions generated by the view synthesis process, *e.g.* flickering distortions.

Figure 6.5 illustrates the performance assessment of each 2D objective quality metric, for the subsets of the dataset: UU: uncompressed texture and uncompressed depth; UC: uncompressed texture and compressed depth; CU: compressed texture and uncompressed depth; CC: compressed texture and compressed depth.

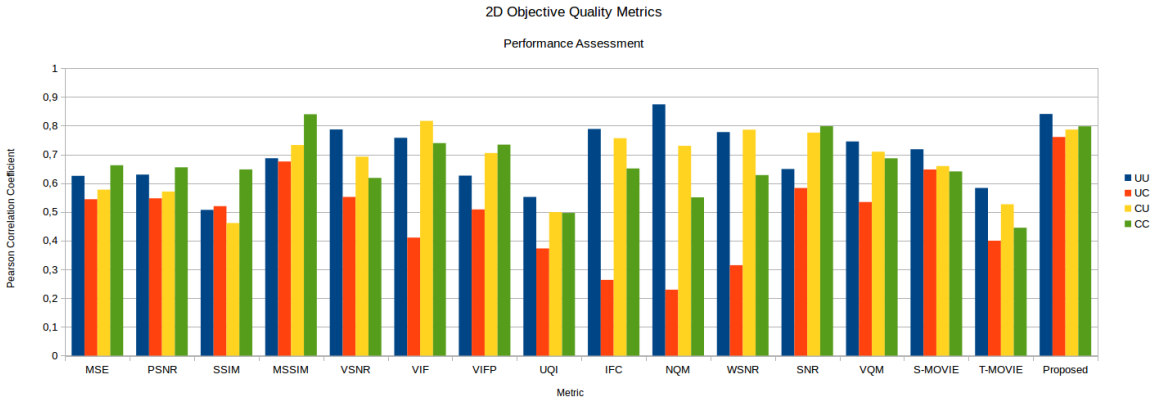


Figure 6.5 – Pearson correlation coefficient for each 2D image metric.

As illustrated in Figure 6.5, it is noticeable that the subset for which the 2D metrics has lower performance is the case of synthesized videos where the texture is uncompressed and the depth is compressed, notably where the flickering distortions have a greater impact over the video quality. This effect is smaller for the pure signal-based approaches, such as MSE and SNR, because they essentially rely upon the signals' difference between the reference and the test sequences.

There are metrics which show a great difference in different subsets, for example, NQM shows a good performance when the depth and texture is uncompressed, much higher than the MSE; however, NQM performance is poor when depth and texture are both compressed, resulting in the lowest performance than the consistent MSE for the whole dataset.

The scatter plots of DMOS versus $DMOS_p$ for each 2D objective quality metrics are shown in Figure 6.6, marking the performance of each subset as the previous figure.

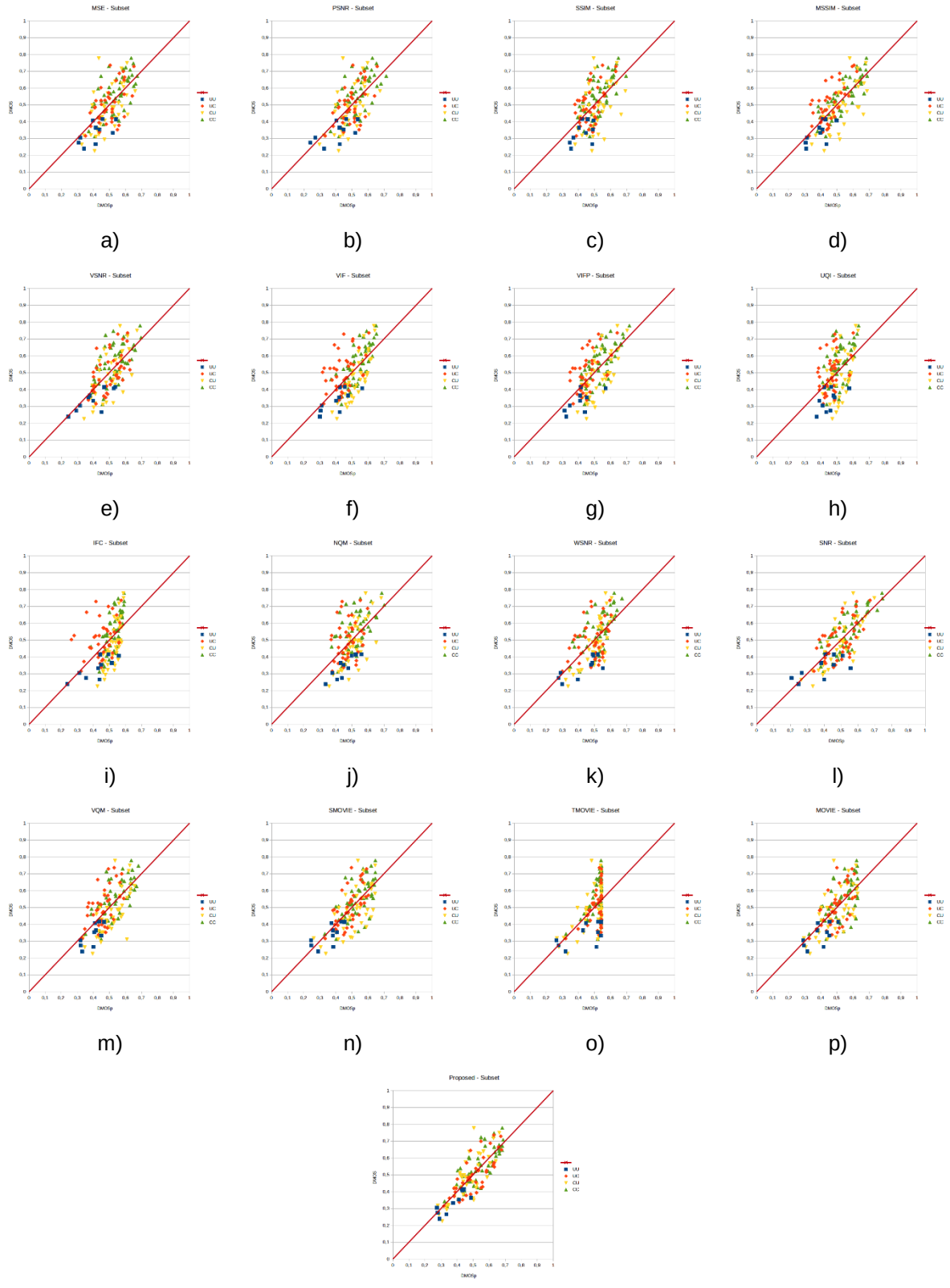


Figure 6.6 – DMOS versus $DMOS_p$ for different 2D objective quality metrics.

Among all the compared 2D QA metrics, the SNR has the highest performance across all subsets and for the entire dataset. This reveals that although new artefacts emerge during view synthesis process, such as geometric distortions, these spatial distortions can be measured somewhat efficiently by 2D QA

metrics. A special note for MSSIM and Spatial MOVIE that correlated well with DMOS on the entire dataset, and the proposed metric had the highest linear correlation.

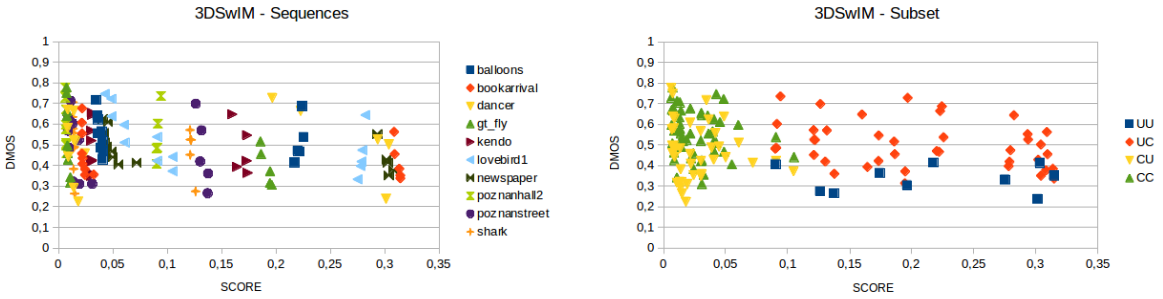
6.4.2. Comparison to 3D Objective Quality Metrics

This section presents the performance assessment comparison of the 3D objective quality metrics described in Section 3.3, and the SIAT database proposed metric. The overall performance assessment results for the SIAT synthesized video quality database are presented in Table 6.4.

Table 6.4 – Performance Comparison of Objective Video Quality Assessment: 3D VQA.

VQA	ALL DATA		
	ρ	*r*	*RMSE*
SIQE	0.140	0.139	0.127
3DSwIM	0.238	0.260	0.125
SIAT	0.815	**0.869**	**0.074**
Proposed	**0.820**	0.809	**0.074**

The performance assessment results show that 3D quality metrics (Section 3.2), like SIQE and 3DSwIM, exhibit a poor performance when evaluated by the SIAT synthesized video quality database. The lack of good performance for these metrics are consequence of being image oriented, and following a perceptual-based approach which do not consider in anyway the temporal artefacts created by the synthesis process. Hence, as shown in Figure 6.7, which illustrates the 3DSwIM score versus DMOS by sequence (a) and by subset (b), it can be seen that the 3DSwIM for each sequence evaluates the uncompressed texture and depth ($U_T U_D$) subset as being similar to the uncompressed texture and compressed depth ($U_T C_D$) subset, giving to both subsets an equivalent score, disregarding the flickering distortions.



a) b)

Figure 6.7 – 3DSwIM DMOS vs Score per: a) Sequence; b) Subset.

The proposed metric follows a similar structure as the SIAT 3D video quality metric. However, the performance results are quite different; notably, showing an improvement of the linear correlation (Pearson) but a lower monotonic correlation (Spearman). This indicates that the proposed metric gives better but noisier scores, in terms of monotonic behaviour. To identify the source of the disparity performance, the performance comparison to each subset is computed, and shown in Table 6.5. Note that results of the subset uncompressed texture and uncompressed depth are unavailable, because no information was present in the SIAT paper.

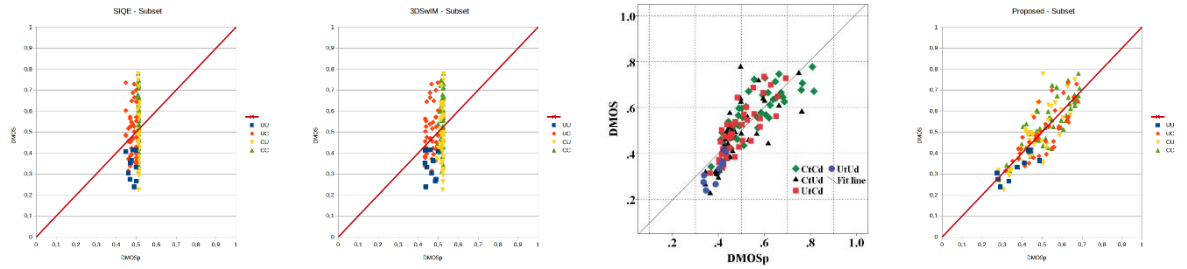
Table 6.5 – Performance Comparison by Subset: SIAT and Proposed.

VQA	$U_T C_D$			$C_T U_D$			$C_T C_D$			ALL DATA		
	ρ	*r*	*RMSE*	*ρ*	*r*	*RMSE*	*ρ*	*r*	*RMSE*	*ρ*	*r*	*RMSE*
SIAT	**0.815**	**0.824**	**0.065**	0.732	**0.838**	0.090	**0.827**	**0.863**	**0.067**	0.815	**0.869**	0.074
Proposed	0.761	0.761	0.073	**0.786**	0.765	**0.083**	0.798	0.789	0.069	**0.820**	0.809	**0.073**

The performance results for the $C_T U_D$ subset is lower than the other subsets, as shown in Table 6.5. This suggests that flickering distortion scores damage the SIAT performance when flickering artefacts are not present.

The proposed metric falls behind the SIAT metric with respect to the Spearman coefficient, expressing that the proposed metric has less outliers but their behaviour is noisier, neglecting some specific distortion effects.

The scatter plots of DMOS versus $DMOS_p$ for each quality assessment metric is shown in Figure 6.8. The quality assessment metrics that follow the signal-based approach do not have a good correlation regarding the perceptual scores.



a) b) c) d)

Figure 6.8 – DMOS versus $DMOS_p$ for different 3D objective quality metrics.

The proposed method has a good performance as suggested in Figure 6.8, with a better Pearson linear correlation since the points are closer to the ideal line DMOSp=DMOS line.

The proposed method presents higher RMSE for the kendo and lovebird1 sequences, as shown in Table 6.6.

Table 6.6 – RMSE of the proposed metric per sequence.

Sequence	RMSE
Balloons	0.045
BookArrival	0.062
Dancer	0.044
GT Fly	0.054
Kendo	0.101
Lovebird1	0.101
Newspaper	0.063
PoznanHall2	0.097
PoznanStreet	0.087
Shark	0.045

The kendo sequence has large smooth areas and a considerable percentage of overexposed area, that constituted a major issue for the quality of the MVE. The lovebird1 sequence, is characterized by a large area which has little or no motion at all, thus the observer pays little attention to that areas and much of the attention is devoted to the two human actors in scene. As mentioned by several objective metrics,

the HVS is more sensitive to distortions that might occur around humans, thus the proposed method gives a lower distortion score than the subjective one, increasing the RMSE for this sequence.

As seen in Figure 6.8 d), there is an obvious outlier that belongs to the $C_T U_D$ subset. This outlier is from the PoznanHall2 sequence with the QP 38 and 0 for the texture and depth, respectively.

CHAPTER 7

Summary and Future Work

Objective video quality assessment metrics have always been important in many scenarios, from assessing the overall performance of a video compression algorithm or codec to the computation of a fidelity measure to monitor video transmission, acquisition or synthesis process or even a complete multimedia system. However, video quality assessment methods still lack techniques that accurately mimic the visual perceptual model of the human visual system and thus, the research in this field is still needed. In the past, accurate models for 2D image and video quality assessment were proposed with a high level of efficiency, especially for some types of degradations. Nonetheless, with new ways to represent the world, namely 3D visual representation models, new and more reliable models are needed. The state-of-the-art have not yet reached this stage of development for many 3D visual representation models, although several techniques described in the literature use features that attempt to replicate some psychophysiological traits. This Chapter presents a brief summary and suggestions for future work.

7.1. Summary

In this MSc Thesis, some relevant quality assessment metrics have been reviewed, which include both image and video objective quality assessment metrics for 2D and 3D synthesized images. The synthesized views quality assessment metrics address new types of distortions introduced by the rendering process. Various 2D and 3D objective quality assessment metrics were studied in this MSc Thesis, and heavily based on [80] a metric was proposed and implemented. This full reference objective video quality assessment metric relies on the spatio-temporal domain along block-shaped motion-coherent temporal tubes which are first detected in the reference view, and later projected to the synthesized view. The proposed VQA metric evaluates two types of distortions, one for the conventional 2D distortions in a spatio-temporal domain, and another one to evaluate flickering distortions.

The most relevant 3D synthesized views databases available in the literature were reviewed, as one had to be selected to perform the quality assessment of the described metrics. The criterion for the database selection was to ensure that it is discriminative enough, has a wide range of different quality levels, and emulates several distortion artefacts that occur in view synthesis process. Considering this

criterion, the SIAT synthesized video quality assessment database (Section 5.2.2) was selected, which exhibits these specific characteristics according to [80]. The proposed metric was assessed using the SIAT database, and compared with other objective quality assessment metrics. The experimental results shown that the proposed metric has a good performance compared with the state-of-the-art objective quality assessment metrics on the entire database, and is particularly prominent in the subset that has significant temporal flickering distortions caused by depth compression and by the view synthesis technique.

7.2. Future Work

Despite the good overall performance results, showing that the proposed metric is reliably consistent, some issues need to be addressed to further improve the performance, notably:

1) Spatio-temporal tubes are the basic data structured from which the proposed VQA metric is computed, meaning that they play a crucial role in the metric performance. Therefore, the estimation accuracy of the S-T tubes is critical. These tubes are created using a block-based MVE algorithm to compute the motion vectors from frame to frame. However, during the development and evaluation of the metric, it had been found that the block matching algorithm produced some inconsistent jittery tubes, mostly when some effects took place, notably: blurred central frames, shadows, reflections, occlusion/disocclusion and burned areas. Some of these issues cannot be easily deal with (*e.g.* reflections), but others can be mitigated using several approaches, such as: block matching with adaptive support/weight window penalizing function [110], piecewise rigid scene model [116] and instance scene motion flows [117].
2) The performance assessment of the proposed objective video quality metric depends on the JND model computed as described in the Chapter 4 and its accuracy is rather important. Although the proposed method performed well, some imprecisions occurred since this model was designed to work for 2D objective quality assessment metrics. For this reason, an improvement over the JND model is desirable. One approach suggested is the use of a visibility prediction model of flicker distortions on natural videos [118]. In a nutshell, this model predicts target-related activation levels in the excitatory layer of neural networks for displayed video frames by following frames via spatiotemporal backward masking. It then scales the flickering intensity, and performs an accumulation or adaptation process depending on the impact of the scenes motion.
3) The proposed objective quality assessment method uses the 10% worse S-T tube cases of both distortion scenarios, *i.e.* activity distortion and flickering distortion. Considering that the HVS can focus in small areas of the frame, certain portions of a frame may have a greater impact over the others, depending in the spatial and temporal information. To tackle this challenge, a new model can be used to identify the different regions of interest. One suggestion is by using the fast region-based convolutional networks for object detection [119]. Additionally, quantify the importance of each region of interest can be performed with a scalable visual sensitivity profile estimation [120].
4) The proposed method follows a perceptual-based approach, notably for two types of distortions: i) conventional 2D spatial distortions; and ii) flickering distortions. One way to improve the objective video quality assessment metrics performance is by adding other perceptually-based distortion

score, *e.g.* geometric distortion score (Section 3.1). Geometric distortions occur driven by the overly compressed depth map, and are a major handicap in terms of QoE. It is known that human perception uses its prior geometrical knowledge of the objects in a visual scene to quickly evaluate the grade of distortion associated with them. One suggestion to tackle this issue is the use of a deep neural network that was trained to explore temporal [121], or adapting the DeepQA [122] to the temporal domain maybe considering using long short-term memory network adaptation.

References

[1] I. P. Howard and B. J. Rogers, Perceiving in Depth, Basic Mechanisms, vol. 1, New York: Oxford University Press, Inc., 2012.

[2] [Online]. Available: http://developer.zspace.com/docs/ui-guidelines/Content/intro.php.

[3] H. Pashler and S. Yantis, Stevens' Handbook Of Experimental Psychology, Sensation and Perception, 3rd ed., vol. 1, New York, NY, USA: John Wiley & Sons, Inc., 2002.

[4] I. P. Howard and B. J. Rogers, Perceiving in Depth, Other Mechanisms of Depth Perception, vol. 3, New York: Oxford University Press, Inc, 2012.

[5] J. E. Cutting, "How the eye measures reality and virtual reality," *Behavior Research Methods, Instruments and Computers,* vol. 29, no. 1, pp. 27-36, March 1997.

[6] I. P. Howard and B. J. Rogers, Binocular Vision and Stereopsis, New York: Oxford University Press, Inc., 1995.

[7] R. K. Jones and D. N. Lee, "Why two eyes are better than one: The two views of binocular vision.," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 7, no. 1, pp. 30-40, February 1981.

[8] J. E. Sheedy, I. L. Bailey, M. Buri and E. Bass, "Binocular vs. monocular performance," *American Journal of Optometry and Physiological Optics,* vol. 63, no. 10, pp. 839-846, October 1986.

[9] R. Klein, "Stereopsis and the representation of space," *Perception,* vol. 6, no. 3, pp. 327-332, 1977.

[10] I. P. Howard and B. J. Rogers, Perceiving in Depth, Depth Perception, vol. 2, New York: Oxford University Press, Inc., 2012.

[11] M. L. Braunstein, E. C. Carterette and M. P. Friedman, Depth Perception Through Motion, Irvine, Los Angeles, California: Academic Press, Inc., 1976.

[12] O. Schreer, P. Kauff and T. Sikora, 3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication, John Wiley & Sons, Ltd, 2005.

[13] D. M. Hoffman, A. R. Girshick, K. Akeley and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision,* vol. 8, no. 3, pp. 1-30, March 2008.

[14] J. Geng, "Three-dimensional display technologies," *Advances in Optics and Photonics,* vol. 5, no. 4, pp. 456-535, November 2013.

[15] Y. Takaki and N. Nago, “Multi-projection of lenticular displays to construct a 256-view super multi-view display,” *Optics Express,* vol. 18, no. 9, pp. 8824-8835, April 2010.

[16] Y. Takaki, K. Tanaka and J. Nakamura, “Super multi-view display with a lower resolution flat-panel display,” *Optic Express,* vol. 19, no. 5, pp. 4129-4139, February 2011.

[17] K. Langhans, C. Guil, E. Rieper, K. Oltmann and B. Detlef, “Solid Felix: a static volume 3D-laser display,” in *SPIE*, Santa Clara, California, USA, May 2003.

[18] J. M. Geary, Introduction to Wavefront Sensors, Bellingham, Washington, USA: SPIE - The International Society for Optical Engineering, 1995.

[19] M. Halle, “Autostereoscopic displays and computer graphics,” *ACM SIGGRAPH Computer Graphics,* vol. 31, no. 2, pp. 58-62, May 1997.

[20] M. Oikawa, T. Shimobaba, T. Yoda, H. Nakayama, A. Shiraki, N. Masuda and T. Ito, “Time-division color electroholography using one-chip RGB LED and synchronizing controller,” *Optics Express,* vol. 19, no. 13, pp. 12008-12013, June 2011.

[21] M. Makowski, M. Sypek and Kolodziejczyk, “Colorful reconstructions from a thin multi-plane phase hologram,” *Optics Express,* vol. 16, no. 15, pp. 11618-11623, July 2008.

[22] M. Makowski, M. Sypek, I. Ducin, A. Fajst, A. Siemion, J. Suszek and A. Kolodziejczyk, “Experimental evaluation of a full-color compact lensless holographic display,” *Optics Express,* vol. 17, no. 23, pp. 20840-20846, October 2009.

[23] T. Ito and K. Okano, “Color electroholography by three colored reference lights simultaneously incident upon one hologram panel,” *Optics Express,* vol. 12, no. 18, pp. 4320-4325, September 2004.

[24] T. Shimobaba and T. Ito, “A Color Holographic Reconstruction System by Time Division Multiplexing with Reference Lights of Laser,” *Optical Review,* vol. 10, no. 5, pp. 339-341, September 2003.

[25] T. Shimobaba, A. Shiraki, N. Masuda and T. Ito, “An electroholographic colour reconstruction by time division switching of reference lights,” *Journal of Optics: Pure and Applied Optics,* vol. 9, no. 7, pp. 757-760, July 2007.

[26] T. Shimobaba, A. Shiraki, Y. Ichihashi, N. Masuda and T. Ito, “Interactive color electroholography using FPGA technology and time division switching method,” *IEICE Electronics Express,* vol. 5, no. 8, pp. 271-277, April 2008.

[27] R. Häussler, S. Reichelt, N. Leister, E. Zschau, R. Missbach and A. Schwerdtner, “Larger real-time holographic displays: from prototypes to a consumer product,” in *Stereoscopic Displays and Applications XX*, San Jose, CA, USA, 2009.

[28] G. Wetzstein, D. Lanman, M. Hirsch and R. Raskar, “Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting,” *ACM Transactions on Graphics,* vol. 31, no. 4, pp. 1-11, 1 July 2012.

[29] G. Wetzstein, D. Lanman, W. Heidrich and R. Raskar, “Layered 3D: Tomographic Image Synthesis for Attenuation-based Light Field and High Dynamic Range Displays,” *ACM Transactions on Graphics,* vol. 30, no. 4, 1 July 2011.

[30] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich and R. Raskar, “Polarization Fields: Dynamic Light Field Display using Multi-Layer LCDs,” *ACM Transactions and Graphics,* vol. 30, no. 6, December 2011.

[31] D. Lanman, M. Hirsch, Y. Kim and R. Raskar, “Content-Adaptive Parallax Barriers: Optimizing Dual-Layer 3D Displays using Low-Rank Light Field Factorization,” *ACM Transactions and Graphics,* vol. 29, no. 6, pp. 1-10, December 2010.

[32] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1649-1668, September 2012.

[33] V. Sze, M. Budagavi and G. J. Sullivan, High Efficiency Video Coding (HEVC): Algorithms and Architectures, Switzerland: Springer International Publishing, 2014.

[34] C. C. Chi, M. Alvarez-Mesa, B. Juurlink, G. Clare, F. Henry, S. Pateux and T. Schierl, “Parallel Scalability and Efficiency of HEVC Parallelization Approaches,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1827 - 1838, December 2013.

[35] I.-K. Kim, S. Lee, M.-S. Cheon, T. Lee and J. Park, “Coding efficiency improvement of HEVC using asymmetric motion partitioning,” in *IEEE international Symposium on Broadband Multimedia Systems and Broadcasting*, Seoul, South Korea, June 2012.

[36] K. Ugur, A. Alshin, E. Alshina, F. Bossen, W.-J. Han, J.-H. Park and J. Lainema, “Motion Compensated Prediction and Interpolation Filter Design in H.265/HEVC,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 7, no. 6, pp. 946-956, July 2013.

[37] A. Leontaris and A. M. Tourapis, “Weighted Prediction Methods for Improved Motion Compensation,” in *IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, 2009.

[38] V. Sze and M. Budagavi, “High Throughput CABAC Entropy Coding in HEVC,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1778 - 1791, December 2012.

[39] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S.-M. Lei, J.-H. Park and W.-J. Han, “Sample Adptive Offset in the HEVC Standard,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1755 - 1764, October 2012.

[40] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan and T. Wiegand, “Comparison of the Coding Efficiency of Video Coding Standards — Including High Efficiency Video Coding (HEVC),” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1669 - 1684, December 2012.

[41] A. Smolic, K. Mueller, P. Merkle, P. Kauff and T. Wiegand, “An Overview of Available and Emerging 3D Video Formats and Depth Enhanced Stereo as Efficient Generic Solution,” in *Picture Coding Symposium*, Chicago, IL, USA, May, 2009.

[42] A. Kondoz and T. Dagiuklas, 3D Future Internet Media, 1 ed., London, UK: Springer-Verlag New York Inc., 2014.

[43] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro and Y.-K. Wang, “Overview of the Multiview and 3D Extensions of High Efficiency Video Coding,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 26, no. 1, pp. 35 - 49, January 2016.

[44] K. Müller, “3D Coding Tools for High-Efficiency Video Coding,” in *IEEE Visual Communications and Image Processing (VCIP’2014)*, Valetta, Malta, December 2014.

[45] ITU-T, *High efficiency video coding,* Geneva, Switzerland: ITU-T, 2015.

[46] Y. Chen, G. Tech, K. Wegner and S. Yea, “Test Model 11 of 3D-HEVC and MV-HEVC,” *ISO/IEC JTC1/SC29/WG11, N15141,* February 2015.

[47] J.-W. Kang, Y. Chen, L. Zhang and M. Karczewicz, “Low complexity Neighboring Block based Disparity Vector Derivation in 3D-HEVC,” in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, Melbourne, Australia, 2014.

[48] Y.-L. Chang, C.-L. Wu, Y.-P. Tsai and S. Lei, “CE1.h: Depth-oriented Neighboring Block Disparity Vector (DoNBDV) with virtual depth retrival,” Geneva, Switzerland, January, 2013, document JCT3V-C0131.

[49] F. Jäger, “Depth-based block partitioning for 3D video coding,” in *Picture Coding Symposium*, San Jose, CA, USA, December 2013.

[50] K. Müller, H. Schwartz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. H. Rhee, G. Tech, M. Winken and T. Wiegand, “3D High-Efficiency Video Coding for Multi-View Video and Depth Data,” *IEEE Transactions on Image Processing,* vol. 22, no. 9, pp. 3366 - 3378, September 2013.

[51] F. Jäger, M. Wien and P. Kosse, “Model-based intra coding for depth maps in 3D video using a depth lookup table,” in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, Zurich, Switzerland, October, 2012.

[52] F. Jäger, “Simplified depth map intra coding with an optional depth lookup table,” in *2012 International Conference on 3D Imaging (IC3D)*, Liège, Belgium, December, 2012.

[53] S. M. Seitz and C. R. Dyer, “View Morphing,” in *International Conference on Computer Graphics and Interactive Techniques*, New Orleans, Louisiana, USA, August, 1996.

[54] I. Tosic, B. A. Olshausen and B. J. Culpepper, “Learning Sparse Representations of Depth,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 5, pp. 941 - 952, May, 2011.

[55] M. Tanimoto, T. Fujii and K. Suzuki, “Improvement of Depth Map Estimation and View Synthesis,” *ISO/IEC JTC1/SC29/WG11, M15090,* January 2008.

[56] M. Tanimoto, T. Fujii and M. Suzuki, “View Synthesis Algorithm in View Synthesis Reference Software 2.0 (VSRS2.0),” *ISO/IEC JTC1/SC29/WG11,* February 2009.

[57] ITU-R, *Rec. BT.500 Methodology for the subjective assessment of the quality of television pictures.,* Geneva, Switzerland: ITU-R, January 2012.

[58] ITU-T, *Rec. P910 Subjective video quality assessment methods for multimedia applications,* Geneva, Switzerland: ITU-T, April 2008.

[59] F. Battisti, E. Bosc, M. Carli, P. Callet and S. Perugia, “Objective image quality assessment of 3D synthesized views,” *Signal Processing Image Communication,* vol. 30, no. C, pp. 78-88, January 2015.

[60] R. A. Rensink, “Scene perception,” in *Encyclopedia of Psychology*, vol. 7, New York, Oxford University Press, 2000, pp. 151-155.

[61] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing,* vol. 15, no. 2, pp. 430-444, February 2006.

[62] R. Song, H. Ko and C. C. Kuo, “MCL-3D: a database for stereoscopic image quality assessment using 2D-image-plus-depth source,” *Journal of Information Science and Engineering,* vol. 31, no. 5, pp. 1593-1611, March 2014.

[63] Z. Wang, A. C. Bovik and L. Lu, “Why is image quality assessment so difficult?,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002*, Orlando, FL, USA, May 2002.

[64] D. J. Swift and R. A. Smith, “Spatial frequency masking and Weber's Law,” *Vision Research,* vol. 23, no. 5, pp. 495-505, September 1982.

[65] A. Liu, W. Lin and M. Narwaria, “Image quality assessment based on gradient similarity,” *IEEE Transactions on Image Processing,* vol. 21, no. 4, pp. 1500-1512, November 2011.

[66] A. Srivastava, A. Lee, E. P. Simoncelli and S.-C. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision,* vol. 18, no. 1, pp. 17-33, January 2003.

[67] Z. Wang and A. C. Bovik, “Reduced- and no-reference image quality assessment: the natural scene statistic model approach,” *IEEE Signal Processing Magazine,* pp. 29-40, November 2011.

[68] A. K. Moorthy and A. C. Bovik, “Statistics of natural image distortions,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, March 2010.

[69] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting,* vol. 50, no. 3, pp. 312-322, September 2004.

[70] D. J. Fleet and Y. Weiss, “Optical flow estimation,” in *Mathematical models for Computer Vision: The Handbook*, Springer, 2005, pp. 239-257.

[71] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Transactions on Image Processing,* vol. 19, no. 2, pp. 335-350, February 2010.

[72] F. Dufaux and F. Moscheni, “Motion estimation techniques for digital TV: a review and a new contribution,” *Proceedings of the IEEE,* vol. 83, no. 6, pp. 858 - 876, June 1995.

[73] K. B. Shaika, P. Ganesana, V. Kalista, B. Sathisha and J. M. M. Jenithab, “Comparative study of skin color detection and segmentation in HSV and YCbCr color space,” in *3rd International*

Conference on Recent Trends in Computing 2015 (ICRTC-2015), Ghaziabad, India, August 2015.

[74] P. Porwik and A. Lisowska, “The Haar-wavelet transform in digital image processing: its status and achievements,” *Machine Graphics and Vision,* vol. 13, no. 1/2, pp. 79--98, November 2004.

[75] M. S. Farid, M. Lucenteforte and M. Grangetto, “Objective quality metric for 3D virtual views,” in *IEEE International Conference on Image Processing (ICIP)*, Quebec, Canada, December 2015.

[76] I. P. Howard and B. J. Rogers, Perceiving in depth, depth perception, vol. 2, New York: Oxford University Press, Inc., 2012.

[77] S. Lyu, “Divisive normalization: justification and effectiveness as efficient coding transform,” in *Advances in Neural Information Processing Systems 23*, Vancouver, British Columbia, Canada, December 2010.

[78] P. Teo and D. Heeger, “Perceptual image distortion,” in *IEEE International Conference Image Processing*, Austin, TX, USA, November 1994.

[79] D. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual Neuroscience,* vol. 9, no. 2, pp. 181-197, August 1992.

[80] X. Liu, Y. Zhang, S. Hu, S. Kwong, C. -C. J. Kuo and Q. Peng, “Subjective and Objective Video Quality Assessment of 3D Synthesized Views With Texture/Depth Compression Distortion,” *IEEE Transactions on Image Processing,* vol. 24, no. 12, pp. 4847 - 4861, 2015.

[81] A. Liu, W. Lin, M. Paul, C. Deng and F. Zhang, “Just Noticeable Difference for Images With Decomposition Model for Separating Edge and Textured Regions,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, no. 11, pp. 1648 - 1652, November, 2010.

[82] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong and S. Yao, “Just noticeable distortion model and its applications in video coding,” *Signal Processing: Image Communication,* vol. 20, no. 7, pp. 662-680, August, 2005.

[83] W. Yin, D. Goldfarb and S. Osher, “Image Cartoon-Texture Decomposition and Feature Selection Using the Total Variation Regularized L1 Functional,” in *Variational, Geometric, and Level Set Methods in Computer Vision*, Beijing, China, October, 2005.

[84] V. Duval, J.-F. Aujol and Y. Gousseau, “The TVL1 Model: A Geometric Point of View,” *Multiscale Model and Simulation,* vol. 8, no. 1, p. 154–189, November, 2009.

[85] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *Journal of Mathematical Imaging and Vision,* vol. 40, no. 1, pp. 120-145, May, 2011.

[86] S. Mallat and S. Zhong, “Characterization of signals from multiscale edges,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 14, no. 7, pp. 710 - 732, July, 1992.

[87] A. Al-kubati, J. Saif and M. A. Taher, “Evaluation of Canny and Otsu Image Segmentation,” in *International Conference on Emerging Trends in Computer and Electronics Engineering*, Dubai, Emirates, March, 2012.

[88] University of Southern California, “USC Media Coomunications Lab - MCL-3D Database,” Media Communications Lab, January 2014. [Online]. Available: http://mcl.usc.edu/mcl-3d-database/. [Accessed July 2016].

[89] R. Song, H. Ko and C. C. Kuo, “MCL-3D: a database for stereoscopic image quality assessment using 2D-image-plus-depth source,” March 2014. [Online]. Available: https://arxiv.org/pdf/1405.1403v1.pdf. [Accessed January 2017].

[90] M. Tanimoto, T. Fujii and K. Suzuki, *View synthesis algorithm in view synthesis reference software 3.5 (VSRS3.5) Document M16090,* May 2009.

[91] M. Tanimoto, T. Fujii and M. Suzuki, *View synthesis algorithm in view synthesis reference software 2.0 (VSRS2.0)Document M16090,* February 2009.

[92] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV,” in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, May, 2004.

[93] A. Telea, “An image inpainting technique based on the fast marching method,” *Journal of Graphics Tools,* vol. 9, no. 1, pp. 23-34, April 2012.

[94] M. Solh and G. AlRegib, “Hierarchical hole-filling for depth-based view synthesis in FTV and 3D video,” *IEEE Journal on Selected Topics in Signal Processing,* vol. 6, no. 5, pp. 495-504, June 2012.

[95] M. Solh and G. AlRegib, “Depth adaptative hierarchical hole filling for DIBR-based 3D videos,” in *Proceedings of SPIE - The Internationa Society for Optical Engineering*, Burlingame, CA, USA, June 2012.

[96] M. Solh and G. AlRegib, “Hierarchical hole-filling (HHF): depth image based rendering without depth map filtering for 3D-TV,” in *IEEE International Workshop on Multimedia Signal Processing*, Saint Malo, France, November 2010.

[97] IRCCyN - IVC, “DIBR_Images - Databases IVC - IRCCyN lab,” IVC - IRCCyN, November 2011. [Online]. Available: http://ivc.univ-nantes.fr/en/databases/DIBR_Images/. [Accessed March 2017].

[98] Y. Mori, N. Fukushima, T. Yendo, T. Fujii and M. Tanimoto, “View generation with 3D warping using depth information for FTV,” *Signal Processing: Image Communication,* vol. 24, no. 1-2, p. 65–72, January 2009.

[99] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff and T. Wiegand, “View synthesis for advanced 3D video systems,” *EURASIP Journal on Image and Video Processing,* vol. 2, no. 1, pp. 1-11, February 2009.

[100] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller and T. Wiegand, “Depth image based rendering with advanced texture synthesis,” in *IEEE International Conference on Multimedia and Expo (ICME)*, Singapore, July 2010.

[101] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller and T. Wiegand, “Temporally consistent handling of disocclusions with texture synthesis for depth-image-based

rendering," in *2010 IEEE International Conference on Image Processing*, Hong Kong, China, September 2010.

[102] IRCCyN - IVC, "DIBR_Videos - Databases IVC - IRCCyN lab," DIBR_Videos Videos quality assessment (using ACR-HR), November 2012. [Online]. Available: http://ivc.univ-nantes.fr/en/databases/DIBR_Videos/. [Accessed March 2017].

[103] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. Jay Kuo and Q. Peng, "SIAT Synthesized Video Quality Database," SIAT Synthesized Video Quality Database, December 2015. [Online]. Available: http://codec.siat.ac.cn/SIAT_Synthesized_Video_Quality_Database/. [Accessed January 2017].

[104] "Reference Software for 3D-AVC: 3DV-ATM V10.0," [Online]. Available: http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/3DV-ATMv10.0/. [Accessed 24 April 2017].

[105] "VSRS-1D-Fast," [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-8.0/. [Accessed 24 April 2017].

[106] ITU-T, "Objective perceptual assessment of video quality: Full reference television," ITU-T, Geneva, Switzerland, 2004.

[107] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge, United Kingdom: Cambridge University Press, 2009.

[108] Y.-K. Huo, G. Wei, Y.-D. Zhang and L.-N. Wu, "An adaptive threshold for the Canny Operator of edge detection," in *2010 International Conference on Image Analysis and Signal Processing*, Zhejiang, China, April 2010.

[109] Y.-K. Huo, G. Wei, Y.-D. Zhang and L.-N. Wu, "An adaptive threshold for the Canny Operator of edge detection," in *International Conference on Image Analysis and Signal Processing (IASP)*, Zhejiang, China, April 2010.

[110] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 4, pp. 650 - 656, April 2006.

[111] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, USA, November, 2003.

[112] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters,* vol. 9, no. 3, pp. 81-84, March 2002.

[113] H. R. Sheikh, A. C. Bovik and G. Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Transactions on Image Processing,* vol. 14, no. 12, pp. 2117-2128, December 2015.

[114] N. Damera-Venkata, T. D. Kite and W. Geisler, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing,* vol. 9, no. 4, pp. 636 - 650, April 2000.

[115] T. Mitsa and K. L. Varkur, “Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms,” in *ICASSP-93., 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, USA, USA, April 1993.

[116] C. Vogel, K. Schindler and S. Roth, “3D Scene Flow Estimation with a Piecewise Rigid Scene Model,” *International Journal of Computer Vision,* vol. 115, no. 1, p. 1–28, October 2015.

[117] A. Behl, O. H. Jafari, S. K. Mustikovela, H. A. Alhaija, C. Rother and A. Geiger, “Bounding Boxes, Segmentations and Object Coordinates: How Important is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?,” in *International Conference on Computer Vision*, Venice, Italy, 2017.

[118] L. K. Choi, L. K. Cormack and A. C. Bovik, “Visibility prediction of flicker distortions on naturalistic videos,” in *48th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, November, 2014.

[119] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in *Neural Information Processing Systems (NIPS)*, Palais des Congrès de Montréal, Montréal, Canada, December, 2015.

[120] G. Zhai, Q. Chen, X. Yang and W. Zhang, “Scalable visual sensitivity profile estimation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, May, 2008.

[121] H. Mobahi, R. Collobert and J. Weston, “Deep learning from temporal coherence in video,” in *26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, June, 2009.

[122] J. Kim and S. Lee, “Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii Convention Center, HI, USA, July, 2017.

[123] D. M. Hoffman, A. R. Girshick, K. Akeley and M. S. Banks, “Vergence-accommodation conflicts hinder visual performance and cause visual fatigue,” *Journal of Vision,* vol. 8, no. 3, pp. 1-30, 28 March 2008.

[124] J. Geng, “Three-dimensional display technologies,” *Advances in Optics and Photonics,* vol. 5, no. 4, pp. 456-535, 22 November 2013.

[125] Y. Takaki, K. Tanaka and J. Nakamura, “Super multi-view display with a lower resolution flat-panel display,” *Optic Express,* vol. 19, no. 5, pp. 4129-4139, 16 February 2011.

[126] Y. Takaki and N. Nago, “Multi-projection of lenticular displays to construct a 256-view super multi-view display,” *Optics Express,* vol. 18, no. 9, pp. 8824-8835, 13 April 2010.

[127] K. Langhans, C. Guil, E. Rieper, K. Oltmann and B. Detlef, “Solid Felix: a static volume 3D-laser display,” in *SPIE*, Santa Clara, California, USA, 30 May 2003.

[128] M. Oikawa, T. Shimobaba, T. Yoda, H. Nakayama, A. Shiraki, N. Masuda and T. Ito, “Time-division color electroholography using one-chip RGB LED and synchronizing controller,” *Optics Express,* vol. 19, no. 13, pp. 12008-12013, 6 June 2011.

[129] M. Makowski, M. Sypek and Kolodziejczyk, “Colorful reconstructions from a thin multi-plane phase hologram,” *Optics Express,* vol. 16, no. 15, pp. 11618-11623, 18 July 2008.

[130] M. Makowski, M. Sypek, I. Ducin, A. Fajst, A. Siemion, J. Suszek and A. Kolodziejczyk, “Experimental evaluation of a full-color compact lensless holographic display,” *Optics Express,* vol. 17, no. 23, pp. 20840-20846, 29 October 2009.

[131] T. Ito and K. Okano, “Color electroholography by three colored reference lights simultaneously incident upon one hologram panel,” *Optics Express,* vol. 12, no. 18, pp. 4320-4325, 6 September 2004.

[132] T. Shimobaba, A. Shiraki, N. Masuda and T. Ito, “An electroholographic colour reconstruction by time division switching of reference lights,” *Journal of Optics: Pure and Applied Optics,* vol. 9, no. 7, pp. 757-760, 3 July 2007.

[133] T. Shimobaba, A. Shiraki, Y. Ichihashi, N. Masuda and T. Ito, “Interactive color electroholography using FPGA technology and time division switching method,” *IEICE Electronics Express,* vol. 5, no. 8, pp. 271-277, 25 April 2008.

[134] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich and R. Raskar, “Polarization Fields: Dynamic Light Field Display using Multi-Layer LCDs,” *ACM Transactions and Graphics,* vol. 30, no. 6, 1 December 2011.

[135] D. Lanman, M. Hirsch, Y. Kim and R. Raskar, “Content-Adaptive Parallax Barriers: Optimizing Dual-Layer 3D Displays using Low-Rank Light Field Factorization,” *ACM Transactions and Graphics,* vol. 29, no. 6, pp. 1-10, 1 December 2010.

[136] V. Sze and M. Budagavi, “High Throughput CABAC Entropy Coding in HEVC,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1778 - 1791, 1 December 2012.

[137] R. Blake and R. Fox, “The psychophysical inquiry into binocular Summation,” *Perception and Psychophysics,* vol. 14, no. 1, pp. 161-185, February 1973.

[138] M. E. Ono, J. Rivest and H. Ono, “Depth perception as a function of motion parallax and absolute distance information,” *Journal of Experimental Psychology: Human Perception and Performance,* vol. 12, no. 3, pp. 331-337, August 1986.

[139] P. Lebreton, A. Raake, M. Bakowsky and P. L. Callet, “Perceptual Depth Indicator for S-3D Content Based on Binocular and Monocular Cues,” in *Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, November, 2012.

[140] F. Dufaux, B. Pesquet-Popescu and M. Cagnazzo, Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering, Chichester, West Sussex: John Wiley & Sons, Ltd., 10 June 2013.

[141] V. Jantet, *Layered Depth Images for Multi-View Coding,* 2013.

[142] M. M. Hannuksela, Y. Yan, X. Huang and H. Li, “Overview of the multiview high efficiency video coding (MV-HEVC) standard,” in *International Conference on Image Processing*, Quebec City, QC, September 2015.

[143] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1649-1668, 28 September 2012.

[144] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S.-M. Lei, J.-H. Park and W.-J. Han, “Sample Adptive Offset in the HEVC Standard,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, no. 12, pp. 1755 - 1764, 5 October 2012.

[145] F. Jäger, “Depth-based block partitioning for 3D video coding,” in *Picture Coding Symposium*, San Jose, CA, USA, 2013.

[146] L. Zhang, Y. Chen, X. Li and S. Xue, “Low-complexity advanced residual prediction design in 3D-HEVC,” in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, Melbourne, Victoria; Australia, June 2014.

[147] F. Battisti, E. Bosc, M. Carli, P. Callet and S. Perugia, “Objective image quality assessment of 3D synthesized views,” *Signal Processing Image Communication,* vol. 30, no. C, pp. 78-88, 2015.

[148] K. B. Shaika, P. Ganesana, V. Kalista, B. Sathisha and J. M. M. Jenithab, “Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space,” in *3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*, Ghaziabad, India, 2015.

[149] M. Solh, A. Ghassan and J. M. Bauza, “3VQM: A vision-based quality measure for DIBR-based 3D videos,” in *2011 IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, July, 2011.

[150] P.-H. Conze, P. Robert and L. Morin, “Objective view synthesis quality assessment,” in *Society of Photo-Optical Instrumentation Engineers (SPIE)*, Burlingame, California, United States, February, 2012.

[151] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando and A. M. Kondoz, “Perceptual Video Quality Metric for 3D video quality assessment,” in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Tampere, Finland, June, 2010.

[152] M. S. Farid, M. Lucenteforte and M. Grangetto, “Objective Quality Metric for 3D Virtual Views,” in *IEEE International Conference on Image Processing (ICIP)*, Quebec, Canada, 2015.

[153] P. Porwik and A. Lisowska, “The Haar-wavelet transform in digital image processing: its status and achievements,” *Machine Graphics and Vision,* vol. 13, no. 1/2, pp. 79--98, 2004.

[154] F. Dufaux and F. Moscheni, “Motion estimation techniques for digital TV: a review and a new contribution,” *Proceedings of the IEEE,* vol. 83, no. 6, pp. 858 - 876, June, 1995.

[155] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV,” in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, May 21, 2004.

[156] A. Telea, “An Image Inpainting Technique Based on the Fast Marching Method,” *Journal of Graphics Tools,* vol. 9, no. 1, pp. 23-34, April 6, 2012.

[157] Y. Mori, N. Fukushima, T. Yendo, T. Fujii and M. Tanimoto, “View generation with 3D warping using depth information for FTV,” *Signal Processing: Image Communication,* vol. 24, no. 1-2, p. 65–72, January 2009.

[158] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff and T. Wiegand, “View Synthesis for Advanced 3D Video Systems,” *EURASIP Journal on Image and Video Processing,* vol. 2, no. 1, pp. 1-11, February 15, 2009.

[159] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller and T. Wiegand, “Depth image based rendering with advanced texture synthesis,” in *Multimedia and Expo (ICME), 2010 IEEE International Conference*, Suntec City, Singapore, July 2010.

[160] ITU-T, *Subjective video quality assessment methods for multimedia applications,* Geneva, Switzerland: ITU-T, 2008.

[161] P. Teo and D. Heeger, “Perceptual Image Distortion,” in *IEEE International Conference Image Processing*, Austin, TX, USA, 1994.

[162] D. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual Neuroscience,* vol. 9, no. 2, pp. 181-197, 1992.

[163] Institut de Recherche en Communications et Cybernétique de Nantes - Images and Video-communications, [Online]. Available: http://ivc.univ-nantes.fr/en/databases/DIBR_Images/.

[164] MediaCommLab - University of Southern California, [Online]. Available: http://mcl.usc.edu/mcl-3d-database/.

[165] Shenzhen Institutes of Advanced Technology, [Online]. Available: http://codec.siat.ac.cn/SIAT_Synthesized_Video_Quality_Database/index.html.

[166] K. Müller, “3D Coding Tools for High-Efficiency Video Coding,” in *IEEE Visual Communications and Image Processing (VCIP’2014)*, Valetta, Malta, 2014.

[167] R. Song, H. Ko and C. C. Kuo, “MCL-3D: a database for stereoscopic image quality assessment using 2D-image-plus-depth source,” arXiv:1405.1403, arXiv preprint, 2014.

[168] ITU-T, “Recommendation ITU-T P.910, Subjective video quality assessment methods for multimedia applications,” ITU, Geneva, Switzerland, 1996.

[169] J. Geng, “Three-dimensional display technologies,” *Advances in Optics and Photonics,* vol. 5, no. 4, pp. 456-535, November 2013.

[170] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez and N. García, “NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences,” in *Proceedings of the 4th International Workshop on Quality of Multimedia Experience*, Melbourne, Australia, July 2012.

[171] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV,” in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, May 2004.

[172] A. Telea, “An Image Inpainting Technique Based on the Fast Marching Method,” *Journal of Graphics Tools,* vol. 9, no. 1, pp. 23-34, April 2012.

[173] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff and T. Wiegand, “View Synthesis for Advanced 3D Video Systems,” *EURASIP Journal on Image and Video Processing,* vol. 2, no. 1, pp. 1-11, February 2009.

[174] P. Teo and D. Heeger, “Perceptual Image Distortion,” in *IEEE International Conference Image Processing*, Austin, TX, USA, November 1994.

[175] X. Liu, Y. Zhang, S. Hu, S. Kwong, C. -C. J. Kuo and Q. Peng, “Subjective and Objective Video Quality Assessment of 3D Synthesized Views With Texture/Depth Compression Distortion,” *IEEE Transactions on Image Processing,* vol. 24, no. 12, pp. 4847 - 4861, August 2015.

[176] A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, “Considering temporal variations of spatial visual distortions in video quality assessment,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 3, no. 2, pp. 253-265, April 2009.

[177] S. L. Cloherty, M. J. Mustari, M. G. Rosa and M. R. Ibbotson, “Effects of saccades on visual processing in primate MSTd,” *Vision Search,* vol. 50, no. 24, pp. 2683-2691, December 2010.