



République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Djilali Liabes de Sidi Bel Abbés

Faculté de Technologies
Département d'Informatique

THÈSE

pour obtenir le titre de
Docteur en Sciences

Spécialité: INFORMATIQUE
Option: INTELLIGENCE ARTIFICIELLE

Présenté et soutenue par
Sid Ahmed FEZZA

Stereo and Multi-view Video Coding Extensions of the H.264/AVC Standard: from Coding to Quality of Experience

Thèse dirigée par **Kamel Mohamed FARAOUN**

Membres du jury:

<i>Président :</i>	Zakaria ELBERRICHI	Professeur	UDL - Sidi Bel Abbés
<i>Directeur de thèse :</i>	Kamel Mohamed FARAOUN	Professeur	UDL - Sidi Bel Abbés
<i>Examineurs :</i>	Sofiane BOUKLI HACENE	MCA	UDL - Sidi Bel Abbés
	Benmammar BADR	MCA	Université de Tlemcen
	Fatima DEBBAT	MCA	Université de Mascara
	Reda Mohamed HAMOU	MCA	Université de Saïda
<i>Invité :</i>	Mohamed-Chaker LARABI	MCA	Université de Poitiers

2014 - 2015

Abstract

Over the last few years, 3D video has been gaining interest with a fast pace. It can be regarded as a type of visual media that expand the capabilities of a 2D video by providing depth sensation of the observed scene. 3D video applications require Multi-View content (MVV) with a very large number of views, captured by several synchronized cameras from different viewpoints. As this huge amount of data is making its way to consumers, its transmission over existing bandwidth-limited infrastructure is a great challenge. Consequently, efficient compression is essential to store or transmit MVV streams. Based on these considerations, Multi-view Video Coding (MVC), an extension of the H.264/MPEG-4 AVC standard, has been developed to address the encoding of MVV content. MVC takes advantage of the redundancies existing between views to achieve high coding performance. However, one major problem with MVV systems is linked to illumination and color variations between views. These color mismatches may impair both coding performance and rendering quality. To fix this issue, we propose a preprocessing method based on an improvement of the histogram matching for correcting these color discrepancies between views. Furthermore, this thesis tackles the problem of stereoscopic video (*i.e.*, 2 views only) coding by relying on an asymmetric coding method that exploits the binocular suppression theory. For this point, based on recent findings on binocular perception, a fully automated model that dynamically determines the best level of asymmetry is proposed, thus offering optimal 3D visual experience. Finally, in 3D imaging systems, it is desirable to assess degradations and to predict the perceived quality of 3D content in line with the human visual system (HVS). Therefore, reliable quality evaluation metrics are really needed, in order to guarantee the final 3D viewing experience. To this end, this thesis presents three novel quality assessment metrics for stereoscopic images. Thanks to the integration of the perceptual binocular characteristics, these metrics handle effectively the asymmetric distortions of stereoscopic images.

In light of these developments, relying on the perceptual characteristics of the HVS, several methods have been presented in this thesis, thus covering the end-to-end 3D video system, from preprocessing to quality assessment, through coding, with the aim to provide an optimal quality of experience.

Keywords : 3D Video, Multi-view Video, Stereoscopic video, Multi-view Video Coding, H.264/MPEG-4 AVC, Asymmetric Coding, Color Correction, View Rendering, Depth, Binocular Just Noticeable Difference, Quality of Experience.

Résumé

L'avènement de la 3D stéréoscopique a suscité un grand intérêt à la fois chez les professionnels et le grand public. Nul doute que cet intérêt est lié à l'effet immersif et l'impression de profondeur que procure la 3D. Au-delà des données stéréoscopiques, les vidéos multi-vues (*MVV : Multi-View Video*) permettent d'étendre les capacités de la 3D avec un visionnage sans lunette et un accès à différents points de vue de la scène. Ces séquences MVV sont constituées de plusieurs vidéos captées avec un ensemble de caméras. Cependant, cela se fait au prix d'une taille de données considérable endiguant leur transmission sur des infrastructures avec une bande passante limitée. Par conséquent, l'usage d'une solution de compression efficace devient crucial pour stocker ou transmettre les flux MVV. C'est pourquoi la norme H.264/MPEG-4 AVC a été étendue avec le profil Multi-view Video Coding (MVC). Ce dernier exploite la forte corrélation existant entre les vues en plus de la corrélation temporelle. Cependant, les systèmes d'acquisitions multi-vues peuvent générer une non-uniformité de luminance et/ou de chrominance entre les vues. Cette variation de couleur peut affecter négativement les performances de codage et la qualité du rendu visuel. Pour pallier à ce problème, nous proposons une méthode efficace de prétraitement basée sur une amélioration significative de la méthode *histogram matching*. En outre, cette thèse s'attèle au problème de la compression des vidéos stéréoscopiques en proposant une méthode de codage asymétrique exploitant la théorie de suppression binoculaire. En nous basant sur les récents résultats sur la perception binoculaire, nous proposons un modèle entièrement automatisé qui détermine dynamiquement le meilleur niveau d'asymétrie entre les vues, offrant ainsi une expérience visuelle 3D optimale. Enfin, dans les systèmes d'imagerie 3D, l'évaluation de la qualité perçue est une étape primordiale pour la validation des solutions développées. Même si l'évaluation subjective reste l'approche la plus fiable, il est indispensable de disposer de métriques de qualité se rapprochant du jugement humain. C'est ainsi que nous avons proposé trois nouvelles métriques de qualité dédiées aux images stéréoscopiques. Grâce à l'intégration des caractéristiques de la perception binoculaire, ces métriques évaluent efficacement les distorsions asymétriques dans les images stéréoscopiques.

À la lumière de ces développements et en s'appuyant sur les caractéristiques perceptuelles du SVH, plusieurs méthodes ont été présentées dans cette thèse, couvrant de bout en bout la chaîne d'exploitation des vidéos 3D, allant du prétraitement à l'évaluation de la qualité, en passant par le codage, et tout cela dans le but de fournir une qualité d'expérience optimale.

Mots clés : Vidéo 3D, Vidéo Multi-Vues, Vidéo Stéréoscopique, Codage de Vidéos Multi-Vues, H.264/MPEG-4 AVC, Codage Asymétrique, Correction de Couleur, Rendu de Vue, Profondeur, JND-Binoculaire, Qualité de l'Expérience.

Contents

Abstract	i
Résumé	ii
Remerciements	iii
List of Figures	ix
List of Tables	x
List of Abbreviations	xi
Introduction	1
1 Overview of 3D Perception	5
1.1 Introduction	5
1.2 Basic concepts of depth perception	6
1.2.1 Anatomy and organisation	6
1.2.2 Depth cues	7
1.3 Some perceptual computational models	13
1.3.1 Saliency models	14
1.3.2 Just Noticeable Difference	17
1.3.3 Binocular Just Noticeable Difference	19
1.4 Conclusion	21
2 3D Video Representation and Coding	22
2.1 Introduction	22
2.2 3D video application scenarios	23
2.2.1 Three-dimensional television	24
2.2.2 Free view point video	24
2.2.3 Immersive teleconference	25
2.3 3D video data representation	26

2.3.1	Stereoscopic 3D Video	26
2.3.2	Multi-View Video	27
2.3.3	Video-plus-Depth	28
2.3.4	Multi-view Video-plus-Depth	29
2.3.5	Layered Depth Video	30
2.4	3D video coding	30
2.4.1	Principle	31
2.4.2	H.264/MPEG-4 AVC Video Codec	34
2.4.3	Stereo/Multi-view Video Coding	37
2.4.4	Coding performance	40
2.5	Disparity estimation	41
2.6	Conclusion	44
3	Color Correction of Multi-View Video	45
3.1	Introduction	45
3.2	Color compensation methods for multi-view images/videos	47
3.2.1	Calibration of multi-camera setup	48
3.2.2	Embedded illumination/color compensation	50
3.2.3	Preprocessed color correction	53
3.3	Proposed color correction method	58
3.3.1	Feature correspondence	58
3.3.2	Reference view selection	60
3.3.3	Cumulative histogram calculation	62
3.3.4	Color mapping function	64
3.4	Experimental results	65
3.4.1	Effect on the visual color consistency	65
3.4.2	Effect on coding performance	69
3.4.3	Effect on rendering performance	73
3.5	Conclusion	77
4	Asymmetric Coding for Stereoscopic Video	79
4.1	Introduction	79
4.2	Asymmetric stereoscopic video coding methods	81
4.2.1	Temporal resolution reduction	81
4.2.2	Asymmetric quantization	83
4.2.3	Mixed resolution coding	84
4.2.4	Asymmetric luminance/chrominance	88
4.3	Perceptual limits of asymmetry	88
4.4	Proposed adaptive asymmetric stereoscopic coding	89
4.5	Experimental results	95
4.5.1	Experience 1	95
4.5.2	Experience 2	100
4.6	Conclusion	102

5	Stereoscopic 3D Quality Assessment	103
5.1	Introduction	103
5.2	Visual quality assessment	105
5.2.1	Subjective quality assessment	105
5.2.2	Objective quality assessment	107
5.2.3	Performance evaluation of objective quality metrics	111
5.3	Stereoscopic image quality assessment (SIQA): State-of-the-art . . .	112
5.3.1	SIQA based on 2D metrics	112
5.3.2	SIQA including depth/disparity information	114
5.3.3	SIQA considering binocular perception	114
5.4	Proposed quality assessment metric for stereoscopic images	116
5.4.1	Binocular perception model	116
5.4.2	Proposed FR IQA metric for stereoscopic images	118
5.4.3	Proposed cyclopean image quality approach for stereoscopic images	120
5.4.4	Proposed NR perceptual blur metric for stereoscopic images .	121
5.5	Experimental results	124
5.6	Conclusion	127
6	Conclusion and Future Work	128
	Bibliography	131
	Publications	155

List of Figures

1.1	Vertical section of the human eye and an overview of the HVS	7
1.2	Categories of depth cues.	8
1.3	Basic geometry of horizontal binocular disparities	8
1.4	Illustration of binocular disparity perception	10
1.5	Depth information from Pictorial cues.	11
1.6	Example of motion depth cues	12
1.7	Accommodation and vergence processes at near and far distance. . .	12
1.8	Just-discriminable depth thresholds versus the log-distance from the observer.	13
1.9	Computational modelling of the visual attention.	14
1.10	Flowchart of the model proposed by <i>Itti et al.</i>	16
1.11	Contrast sensitivity function	18
2.1	An examples of typical multi-view capture setup.	24
2.2	Multi-view video system for different application scenarios	25
2.3	Stereoscopic 3D video representation	26
2.4	Stereoscopic multiplexing pictures	27
2.5	Example of a multi-view video representation	27
2.6	Example of a video-plus-depth representation.	28
2.7	Rendering of stereo video from V+D.	28
2.8	Example of MVD format with 5 views and corresponding depth maps. .	29
2.9	Virtual views by DIBR for efficient support of multi-view auto- stereoscopic displays	30
2.10	Example of LDI.	31
2.11	An example of spatial and temporal correlation in a 2D video sequence. .	32
2.12	Video encoder block diagram.	32
2.13	DPCM/DCT video encoder.	33
2.14	Inter prediction	33
2.15	Intra prediction	34
2.16	Example of GOP structure.	35
2.17	Macroblocks and slices organization.	35

2.18	H.264/MPEG-4 AVC video encoder scheme.	36
2.19	Hierarchical reference picture structure for temporal prediction.	36
2.20	Simulcast coding approach.	37
2.21	Disparity-based and motion-based prediction in a stereo pair.	37
2.22	A basic inter-view prediction structure for stereoscopic video.	38
2.23	Overall structure of a MVC system with stereoscopic video	38
2.24	Typical MVC prediction structure.	39
2.25	Sample coding results for <i>Ballroom</i> , <i>Race1</i> , and <i>Rena</i> sequences.	40
2.26	Disparity principle.	42
2.27	Cones stereo pair and the corresponding disparity map	42
2.28	An example of features correspondence.	43
3.1	An example of color inconsistencies between views.	46
3.2	An illustration of binocular luster.	47
3.3	An example of Macbeth color checker.	48
3.4	Modified search loop for the current block	50
3.5	Block diagram of the MB-based adaptive illumination change compensation method	52
3.6	An example for the resulting histogram of the corrected image compared with the reference image and the distorted image.	54
3.7	Flowchart of the proposed method	58
3.8	Overview of the proposed approach	59
3.9	Common SIFT keypoints across five views of the <i>Race1</i> sequence	60
3.10	An example of some SIFT regions	63
3.11	Visual effect of the proposed method on the <i>Flamenco2</i> test sequence	66
3.12	Visual effect of the proposed method on the <i>Race1</i> test sequence	66
3.13	Visual effect of the proposed method on the <i>Ballroom</i> test sequence	67
3.14	Visual effect of the proposed method on the <i>Exit</i> test sequence	67
3.15	Illustration of the automatic selection of the reference view for the four used test sequences.	68
3.16	Impact of impairments on the reference view selection for the <i>Race1</i> sequence	69
3.17	Enlarged parts of the <i>Flamenco2</i> and <i>Race1</i> sequences	70
3.18	Objective color difference comparison for the <i>Flamenco2</i> , <i>Race1</i> , <i>Ballroom</i> , and <i>Exit</i> sequences	71
3.19	Rate distortion curves for the <i>Flamenco2</i> , <i>Race1</i> , <i>Ballroom</i> , and <i>Exit</i> test sequences	72
3.20	The experiment scenario to evaluate the effect on the DIBR-based application	73
3.21	Rendering results for <i>Balloons</i> sequence	74
3.22	Rendering results for <i>Kendo</i> sequence	74
3.23	Rendering results for <i>Newspaper</i> sequence	75
3.24	Rendering results for <i>Lovebird1</i> sequence	75

3.25	Enlarged parts of the rendering results for <i>Balloons</i> and <i>Newspaper</i> sequences	76
3.26	CIEDE2000 color-difference after rendering by DIBR for the <i>Balloons</i> , <i>Kendo</i> , <i>Newspaper</i> , and <i>Lovebird1</i> sequences	77
4.1	Illustration of dropping-and-repeating frames procedure adopted in temporal asymmetry.	81
4.2	The structure of the encoder and decoder including temporal asymmetry	82
4.3	An example of asymmetric quantization.	83
4.4	Stereo image pair with low-pass filtered right view.	84
4.5	Block diagram of MR coding.	85
4.6	MVC with two additional down-sampling blocks	86
4.7	Implementing MR coding method by blur filter or down/up sampling approaches	87
4.8	Relation between objects' distance, depth and sharpness	90
4.9	Disparity map before and after the recursive bilateral filtering.	93
4.10	The obtained visual results for the left image of <i>Teddy</i> stereo pair.	96
4.11	The obtained visual results for the left image of <i>Art</i> stereo pair.	97
4.12	Rate/distortion curves of the <i>Cones</i> , <i>Midd1</i> , <i>Moebius</i> , and <i>Reindeer</i> stereo pairs.	99
4.13	DMOS values for each image of the test-set	100
4.14	Rate/distortion curves of the coded stereoscopic videos.	101
5.1	3D video processing chain artifacts	104
5.2	The three main classes of objective quality metrics.	107
5.3	Example of the upper limit of the PSNR	108
5.4	3D perceived quality based on stereoscopic image-pair	113
5.5	Examples of local information content maps	118
5.6	The flowchart of the FR metric based cyclopean image.	120
5.7	Flowchart of the proposed NR SIQA metric.	122
5.8	Examples of local perceptual blurriness maps	123
5.9	Example of images from the LIVE 3D IQA database (phase 2)	124
5.10	Scatter distributions of subjective DMOS versus predicted scores	126

List of Tables

3.1	Comparison of PSNR averages using Bjøntegaard measure	71
4.1	Bit-rate saving and average PSNR comparisons measured against SSC method using Bjøntegaard measurement	98
4.2	Bit-rate saving and average PSNR comparisons between the proposed asymmetric stereoscopic video coding approach and SSC method using Bjøntegaard measurement	100
5.1	Specification of subjective quality assessment in ITU-R BT.500. . . .	106
5.2	Recommended rating scales to quantify the perceived quality	106
5.3	Performance comparison of the proposed FR SIQA metrics using state-of-the-art tools	125
5.4	Performance comparison of the proposed NR SIQA metric using state-of-the-art tools	125
5.5	Performance comparison for each distortion	125

List of Abbreviations

2D-QA	2D Quality Assessment
3D	Three-Dimensional
3D-QA	3D Quality Assessment
3DTV	Three-Dimensional TeleVision
ACR	Absolute Category Rating
AVC	Advanced Video Coding
BJND	Binocular Just-Noticeable Difference
CABAC	Context-Adaptive Binary Arithmetic Coding
CAVLC	Context-Adaptive Variable-Length Coding
CIE	International Commission on Illumination
CSF	Contrast Sensitivity Function
DCP	Disparity-Compensated Prediction
DCT	Discrete Cosine Transform
DIBR	Depth-Image-Based Rendering
DMOS	Difference Mean Opinion Score
DPB	Decoded Picture Buffer
DPCM	Differential Pulse Code Modulation
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DVIC	Difference Illumination Change Value
DWT	Discrete Wavelet Transform
FR	Full-Resolution
FTV	Free Viewpoint Television
FVV	Free Viewpoint Video
GBVS	Graph Based Visual Saliency
GMM	Gaussian Mixture Model
GOP	Group Of Pictures
HEVC	High Efficiency Video Coding
HM	Histogram Matching
HVS	Human Visual System
IC	Illumination Compensation

IEC	International Electrotechnical Commission
IQA	Image Quality Assessment
ISO	International Standardization for Organization
ITU	International Telecommunication Union
JMVM	Joint Multiview Video Model
JND	Just-Noticeable Difference
JPEG	Joint Photographic Experts Group
JVT	Joint Video Team
LCC	Linear Correlation Coefficient
LDI	Layered Depth Image
LDV	Layered Depth Video
LGN	Lateral Geniculate Nucleus
MB	MacroBlock
MC	Motion Compensation
MCP	Motion-Compensated Prediction
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MR	Mixed Resolution
MSE	Mean Square Error
MS-SSIM	Multi-Scale Structural Similarity Index
MV	Motion Vector
MVC	Multi-view Video Coding
MVD	Multi-view Video plus Depth
MVV	Multi-view Video
NAL	Network Abstraction Layer
NR	No-Reference
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
QP	Quantization Parameter
RANSAC	Random Sample Consensus
R-D	Rate-Distortion
RDO	Rate-Distortion Optimization
RMSE	Root Mean Square Error
RR	Reduced-Reference
S-3D	Stereoscopic 3D
SAD	Sum of Absolute Difference
SIFT	Scale Invariant Feature Transform
SIQA	Stereoscopic Image Quality Assessment
SNR	Signal-to-Noise Ratio
SROCC	Spearman Rank Order Correlation Coefficient
SSC	Symmetric Stereoscopic Coding
SSIM	Structural Similarity Index
UMR	Uniform MR coding
V+D	Video-plus-Depth
VCEG	Video Coding Experts Group
VSRS	View Synthesis Reference Software
ZNCC	Zero mean Normalized Cross Correlation

General Introduction

In the last few years, visual media have changed significantly and evolved considerably. The two main significant and remarkable changes are linked to the increase in resolution and the extension in spatial dimension. Figure 1 shows the historical development of visual media.

For instance, we can see the gradual evolution of video from black/white to color, and then an increase in frame rate and spatial resolution. However, 2D monoscopic video only offers a plane perception, which is quite different to what the human experiences in a natural environment.

As a consequence, with the aim of increasing quality of experience (QoE), re-

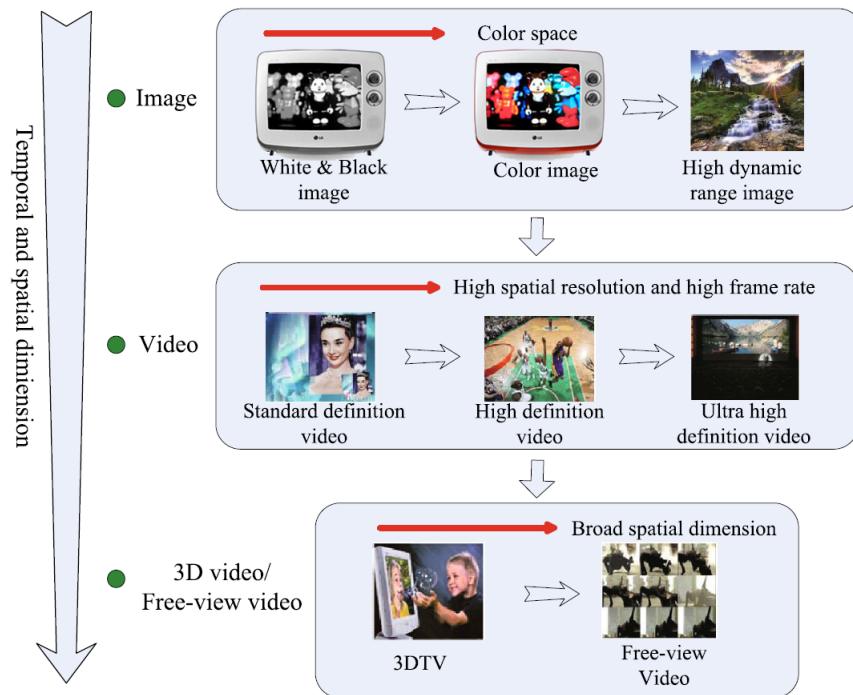


Figure 1: The historical development of visual media.

search and development in 3D video are attracting sustained interest. In recent years, great efforts have been made to bring a more realistic 3D visual experience to consumers. 3D image/video is introduced in our daily life through various multimedia applications including cinema, TV, Blu-ray Disc, PCs, smart phones, broadcasting, etc. As an evidence for this consumer interest is the significant increase in the production of 3D contents. For example, the world's leading manufacturers of consumer electronics expect that, in 2017, 3D-enabled TV will account for 58% of all TVs sold across the globe, rising from 18% in 2013 [1]. This great consumer acceptance is partly due to the recent advances in capturing, coding, and display technology, but also thanks to the 3D feeling of immersion.

Two major 3D video applications may be cited, the 3D television (3DTV) and free viewpoint video (FVV). 3DTV delivers to viewers a feeling of immersion, by bringing an additional dimension (depth) for the perception of the scene, while FVV allows the user to change his viewpoint and viewing direction within a scene in an interactive manner. However, to enable such applications, the acquisition of multiple video sequences (N views) is required. When only two views ($N = 2$) are used, we talk about stereoscopic video. The set of N video sequences is acquired from the same scene with slightly different viewpoints, usually referred to as Multi-View Video (MVV). In sight of the huge amount of data created by MVV, depending linearly on the number of views, storage or transmission becomes not plausible in media delivery systems, such as broadcasting, multimedia streaming services, and other commercial applications. Therefore, efficient compression methods are of paramount importance for streaming MVVs while preserving the QoE.

In this direction, the topic of Multi-view Video Coding (MVC) has been and still an active research area. Consequently, the 2D video coding standard H.264/MPEG-4 AVC has been extended to support MVC. The adopted MVC schemes have been designed under the assumption that each camera view, in the MVV system, captures the same scene from slightly different perspectives, and are therefore highly correlated. MVC combines temporal and inter-view predictions, where frames are not only predicted from their temporal neighbors but also from corresponding ones in adjacent views. Compression efficiency is significantly increased compared to the simulcast approach encoding each view independently.

Scope of the thesis

One major problem with multi-camera systems is the color balance between individual cameras. Since MVV is captured from different viewpoints with multiple cameras, significant illumination and color inconsistencies can be observed between views. These color mismatches reduce the correlation existing between views, and therefore can significantly decrease the compression efficiency. Furthermore, these inconsistencies can negatively affect rendering of new virtual views that will appear unnatural. This aspect may be annoying when viewers switch between the different views, thus affecting their QoE.

Compression of MVV remains a great challenge towards the design of advanced 3D video systems. Most of the previous developed approaches have mainly focused on bit-rate gain while neglecting the perceptual aspect. In parallel, we have noticed a huge increase in the use of perceptual approaches to improve the performance of image processing techniques. Among these promising techniques, we can quote those used for coding stereoscopic 3D content, known as *asymmetric coding*, *i.e.*, one of the views is provided with lower quality than the other. Thanks to the binocular suppression theory, the fused perceptual 3D image is dominated by the higher quality view. However, despite abundant works devoted to this concept, there are still unresolved issues, such as the problem of determining the best level of asymmetry between the views.

Indeed, the performance of 3D imaging systems must be determined and predicted in order to guaranty the optimal 3D visual experience for the end user. As human observer is the ultimate receiver, subjective tests remain the best way to assess the perceived 3D image quality. However, such experiments are time consuming, not feasible for real-time applications, expensive, and depend strongly on the observer context (personal experience, emotional state, etc.). Therefore, objective measurements are required to assess and predict automatically the perceived quality of 3D content in line with the HVS.

To this end, this thesis is focused on the aforementioned issues and has led to the following scientific contributions:

- An effective color correction method of MVV is proposed, with the aim to enhance coding performance and view rendering quality.
- Based on recent findings in binocular perception, a fully automated model is proposed. The latter dynamically determines the best level of asymmetry offering optimal 3D visual experience.
- Three quality assessment metrics are proposed for stereoscopic images. Thanks to the incorporation of perceptual binocular characteristics, these metrics handle effectively the asymmetric distortions of stereoscopic images.

Thesis outline

This thesis is divided in six chapters. Two of them are dedicated for the review of existing literature related to the work presented in this thesis. In this review, the fundamental concepts are described providing a basis for the methodological and experimental developments in the rest of the thesis. Three chapters corresponding to each of the contributions and a final one draws conclusions and future work:

Chapter 1

This chapter outlines the theory relevant for understanding the 3D perception. Therefore, we give an introduction to the HVS, and its main features related to

depth perception. Next, some important HVS models are described, which are exploited in the different proposed methods.

Chapter 2

This chapter, first, reviews the major 3D video applications. Next, a state-of-the-art of 3D data representations enabling 3D video services is presented. Following that, the principle of video coding algorithms is introduced, and finally coding algorithms for MVV are described.

Chapter 3

In this chapter, the issue of color inconsistency in MVV is addressed. This chapter presents a preprocessing color correction method to enhance coding performance and rendering quality, and thus increasing the quality of experience.

Chapter 4

This chapter is devoted to the issue of asymmetric coding for stereoscopic 3D. In this chapter, the focus is put on identifying the just noticeable threshold of asymmetry between views of the stereo pair, for which the stereoscopic 3D effect is not affected.

Chapter 5

This chapter is dedicated to 3D image quality assessment. It presents three image quality assessment metrics for stereoscopic images designed to handle effectively the asymmetric distortions, by incorporating HVS characteristics.

Chapter 6

The last chapter provides a summary of the contributions of this thesis and gives future directions.

Chapter 1

Overview of 3D Perception

Contents

1.1	Introduction	5
1.2	Basic concepts of depth perception	6
1.2.1	Anatomy and organisation	6
1.2.2	Depth cues	7
1.3	Some perceptual computational models	13
1.3.1	Saliency models	14
1.3.2	Just Noticeable Difference	17
1.3.3	Binocular Just Noticeable Difference	19
1.4	Conclusion	21

1.1 Introduction

It is obvious that multimedia systems have improved the human's quality of life. For many years, researchers in industry and academia have been working on providing realistic feelings of immersion to viewers. Accordingly, multimedia systems have evolved exponentially during the past century to provide richer content to the users. Recently, there has been a lot of interest on the part of technology suppliers, broadcasters, and content providers to bring 3D video to the consumer. To this end, a new set of 3D video applications have been introduced to the user through various consumer devices. The impact of these technologies are almost comparable to the introduction of color television in the 60s.

3D video is commonly known as a type of visual media that provides the illusion of depth perception. In order to provide this sensation of depth, the 3D video technologies are based on exploiting the properties of the human visual system

(HVS). Therefore, it is essential to understand the vision's mechanisms behind depth perception, in order to provide a satisfactory quality of experience (QoE).

HVS is based on complex mechanism requiring the coordinated work of many components including the eyes, neuronal pathway to the brain and the brain. This complicated structure gives the ability to process the visual informations. HVS perceives visual stimuli from natural environments via the two horizontally located frontal eyes, and form a three-dimensional percept using the two corresponding responses generated in the primary visual cortex of the brain. Inspired by this, the different 3D video delivery systems try to simulate this biological mechanism such that they reproduce a 3D viewing experience as close as possible to what humans naturally view in real life. The main followed direction adopted in 3D video technologies is to fool the brain by providing two disparate views of the same scene to each eye in order to generate a positive or negative disparity that will be interpreted as depth.

The rest of the chapter is organized as follows. Section 1.2 addresses the human visual system and the basic concepts of depth perception. Following that, Section 1.3 describes some perceptual computational models that are relevant to the proposed methods introduced in the following chapters. Finally, 1.4 summarizes and concludes the chapter.

1.2 Basic concepts of depth perception

Human vision is a very complex process that is still not fully understood. In this section, we explore some key concepts playing a role in 3D viewing.

First, a review of human vision is given. Afterwards, depth perception process and depth cues allowing the human to perceive its three-dimensional environment are discussed.

1.2.1 Anatomy and organisation

Vision is the most important senses, since 80–90% of all neurons in the human brain are involved in visual perception [9]. Therefore, human vision is a very complex neural process that enables the perception of our surrounding environment. The HVS can be subdivided into two major components: the eyes, which capture light and convert it into signals that can be exploited by the nervous system, and the visual pathways in the brain, along which these signals are transmitted and processed.

The eye is the first organ identified as responsible for vision. In the first stage, a visual image in the form of light passes through the conjunctiva, a transparent area covering the front of the eye, and then passes through the following optical components: the cornea, the aqueous humour, the pupil, the lens, and lastly, the vitreous humour, as illustrated in Figure 1.1. After crossing the optics of the eye, the light is projected onto the retina, the neural tissue at the back of the eye. The result is referred to as a *retinal image*. When light hits the retina, two types of photoreceptor cells (rods and cones) get excited depending on light intensity and

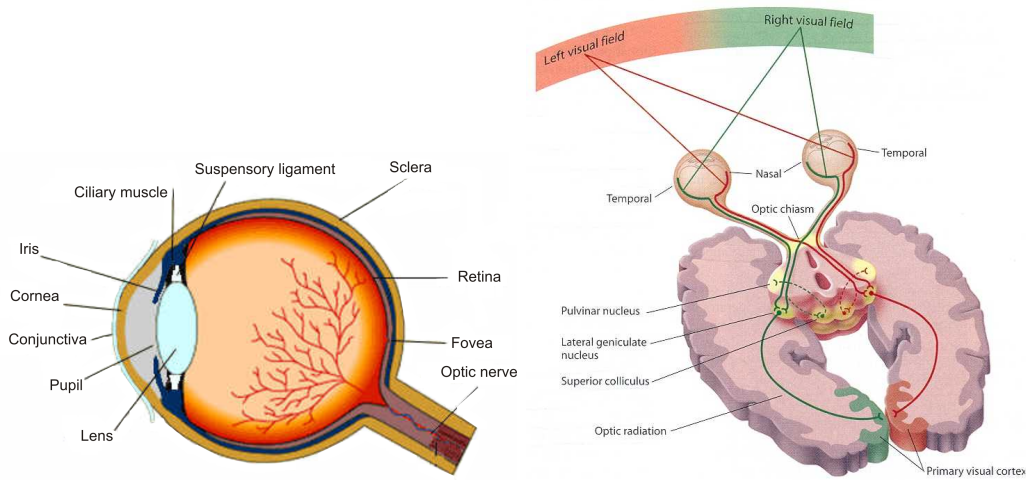


Figure 1.1: Vertical section of the human eye (left), and an overview scheme of the HVS (right) [13].

frequency. The cones are responsible for vision in photopic conditions and high resolution color vision, while the rods are responsible for vision in scotopic conditions and peripheral vision. The cones are highly concentrated at the center of the eye, in a small area called *fovea*, and are categorized into three types, according to their sensitivity to the different wavelengths: the L-cones, M-cones, and S-cones (L for large, M for medium, and S for short). Due to this three subgroups, humans are able to distinguish three colors (commonly names red, green and blue).

The information encoded in the retina is transmitted through the optic nerve to the LGN (lateral geniculate nucleus), before being relayed to the visual cortex, at the back of the brain. The latter is responsible for visual perception. The LGN is also the place where the information from the left and right eyes is merged. While images travel through optic tracks, the image is processed in simple ways by neurons so extra information to the visual cortex is provided. This low level processing is then used to provide higher level information such as disparity (depth) estimation or pattern recognition. For more details on human anatomy, please refer to [2, 16].

1.2.2 Depth cues

Depth perception is the ability to see the environment in three-dimensions and be able to evaluate distance of the surface of objects from the observer. The HVS uses several different cues to comprehend the depth in sensed images. Some cues require both eyes to be open (binocular), others are available also when looking with only one open eye (monocular). The depth cues can be classified mainly into two categories, namely, *oculomotor cues* and *visual cues*, as illustrated in Figure 1.2. In the real world, the HVS automatically uses all the available depth cues to perceive the depth [66]. These depth cues are discussed in the following.

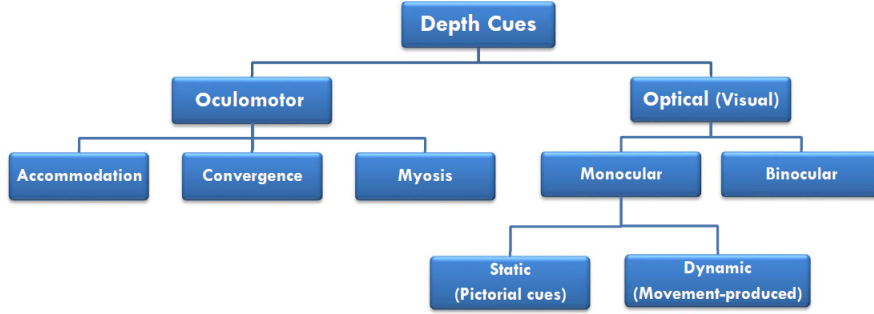


Figure 1.2: Categories of depth cues.

1.2.2.1 Binocular depth cues

Binocular or stereoscopic depth cues are based on the fact that human eyes are horizontally separated by an inter-ocular distance equal to 6.3 cm on average [15]. Due to this lateral displacement, we get two slightly different images of the same scene from each eye, *i.e.*, two different retinal images. The difference in image location of observed 3D points in the scene is called *binocular disparity*. These disparities provide the HVS with information about the relative distance of objects (depth). Thus, this perception of depth obtained from the relative displacement (binocular disparity) between the two retinal images is a binocular depth cue and is referred to as *stereopsis*. Furthermore, binocular stereopsis is most often the only additional depth cue that current stereoscopic 3D displays provide.

Figure 1.3 illustrates how the geometry of the binocular vision projects slightly different images on the retinas of the two eyes. If both eyes fixate on the point P ,

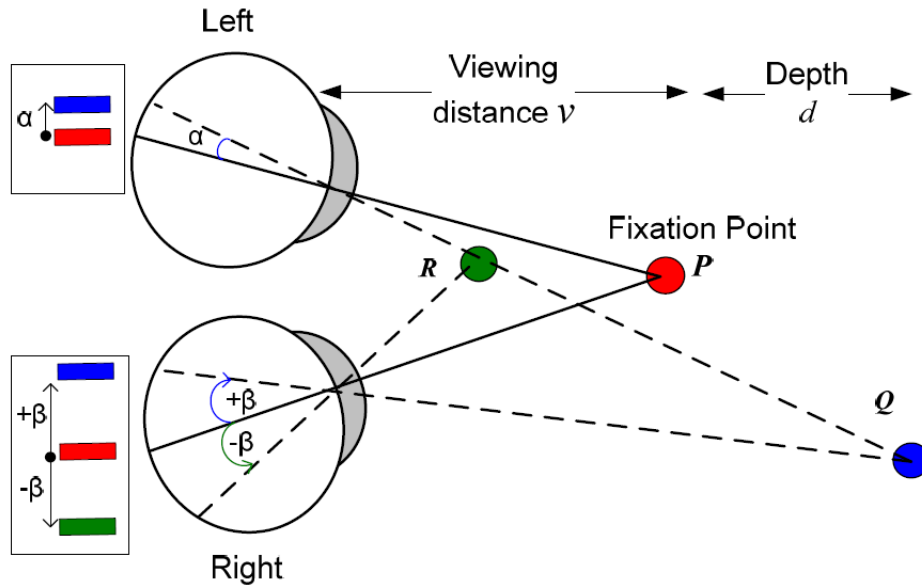


Figure 1.3: Basic geometry of horizontal binocular disparities [66].

then the image cast by that point will fall on the center of the fovea of each eye. Assume that the point Q casts the image α degrees away from one eye's fovea and β degrees away from the other eye's fovea. The binocular disparity η in this case is given as $(\beta - \alpha)$, measured in degrees of visual angle [67]. Binocular disparity acts as the stimulation that enables depth perception with binocular stereopsis. It has both a magnitude and a direction. Let the angular disparity stimulated by point Q be denoted by η_Q ,

$$\eta_Q = (\beta - \alpha) \quad (1.1)$$

Geometrically, it can be proven that the disparity is directly proportional to the depth (d) of the point Q with relatively to P , and inversely proportional to the square of the viewing distance v

$$\eta \approx \frac{e \cdot d}{v^2} \quad (1.2)$$

where e is inter-ocular distance. Similarly, the angular disparity caused by the point R is given as in equation (1.3):

$$\eta_R = (-\beta) - \alpha = -(\beta + \alpha) \quad (1.3)$$

According to equation (1.1) and equation (1.3), $\eta_Q > 0$ and $\eta_R < 0$. The brain interprets this sign difference as the relative positioning of points Q and R relative to point P . When, Binocular disparity is greater than zero, the brain interprets that the point is behind the fixation point and vice versa. Thus, binocular disparity provides stimulation to the brain to perceive the depth of objects, relative to a fixation point.

Furthermore, the amount of displacement of the points relative to the fixation point (in outward or inward direction) indicates how much farther or closer these points are, in depth, compared to the fixation point. Basically, when looking at the 3D world, the eyes rotate until their optical axes converge (intersect at a single point). If the *point of convergence* is projected at the same location on the respective retinas, this does not produce any retinal disparity. The same holds for all points on the *horopter* which is defined by the fixation point and the nodal points of both eyes. All other points, however, will produce retinal disparities whose magnitude becomes larger, the further away the 3D points are from the horopter. Disparities that are caused by points in front of the horopter are said to be *crossed* (closer point), while disparities that result from points behind the horopter are called *uncrossed* (farther point), as illustrated in Figure 1.4. As long as the crossed- or uncrossed-disparities do not exceed a certain magnitude, the two separate viewpoints can be merged by the human brain into a single 3D percept. The small region around the horopter within which disparities are fused is known in the literature as *Panum's fusional zone*. Points outside this area are not fused and double images will be seen, a phenomenon that is called *diplopia* [68].

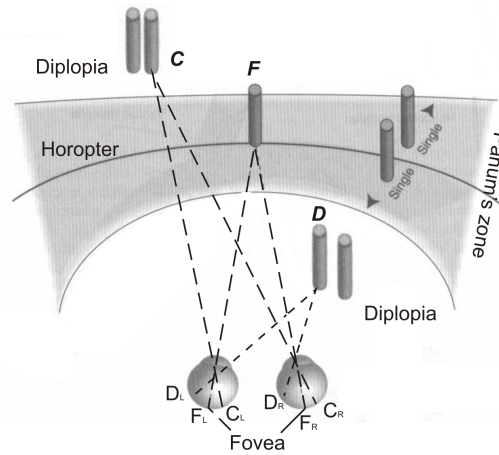


Figure 1.4: Illustration of binocular disparity perception. F is a fixation point on the horopter, D is the point with crossed disparity and C is the point of uncrossed disparity.

1.2.2.2 Pictorial depth cues

The pictorial depth cues are the set of cues responsible for depth inferred from static 2D picture captured from one viewpoint. For instance, when closing one eye, one can still perceive the depth of his surroundings. This ability is possible thanks to the learning process and previous experience. To infer depth using only one image, there are several monocular pictorial depth cues that can be used; some of them are presented below:

Light shading and shadows give knowledge on the shape of the object thanks to the analysis of the reflection of objects and the casted shadows [7, 8] (Figure 1.5(f)).

Occlusion (overlapping, interposition) occurs when a part of the observed object is masked by another. In this case, the HVS considers the occluding object closer than the occluded one (Figure 1.5(b)) [3–5].

Familiar size refers to the use of our knowledge about objects of familiar size to assist our perception of the size and depth location of other objects. If for example a person and the Eiffel tower were to appear in your field of view and they were of a similar size, you would perceive the person as being much closer than the Eiffel tower (Figure 1.5(e)) [6].

Aerial perspective occurs when objects are viewed from very long distances. This phenomenon relates to the way light scatters in the air, causing objects on the horizon to appear in faded colors or washed in lower light luminance and contrast, while objects that are near appear to have vibrant or intense colors with strong contrast (Figure 1.5(a)) [11, 12].

Texture gradient refers to the fact that the texture pattern, its density or orientation change with distance. The closer we are to an object, more details we can see of its surface texture. Therefore, objects with smooth textures are usually

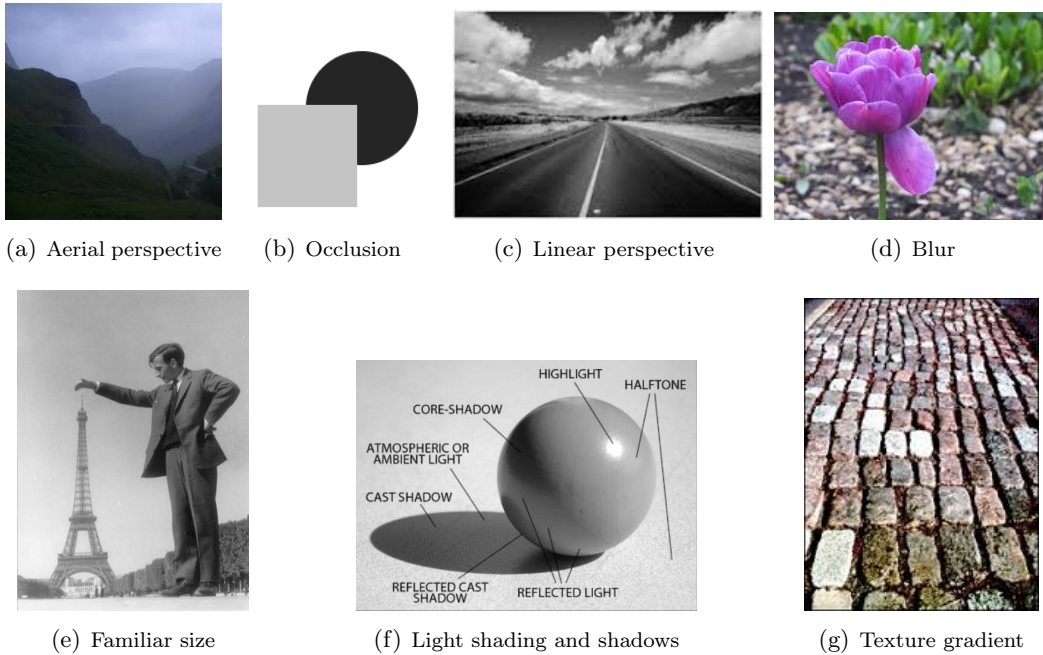


Figure 1.5: Depth information from Pictorial cues.

interpreted as being farther away (Figure 1.5(g)).

Linear perspective occurs with the observation of parallel lines. Such lines appear to recede and converge in the image into a vanishing point. The distance between parallel lines becomes increasingly small with the increasing distance to the observer (Figure 1.5(c)).

Blur/Sharp The amount of blur in the image is lawfully related to its distance from fixation, and can provide a cue to depth. As the distance from fixation increases, areas closer than fixation blur more rapidly than those farther away (Figure 1.5(d)).

1.2.2.3 Dynamic depth cues

The dynamic cues refer to depth information from motion. Displacement of object on the retinas over time provides dynamic visual information to the HVS named *optic flow*. When the observer moves, the relative movement of objects in the scene with respect to the point of view may give hints about their relative or absolute distance. A typical example of this situation is the view from a moving vehicle window such as a car or a train. If one would look through a lateral window, far objects will appear to be much more fixed than objects near the vehicle. An illustration can be seen in Figure 1.6, where the trees near the moving car appear to move faster than far away trees, the cloud and the sun (which almost does not move). This difference in the relative movement of near and far objects is referred to as *motion parallax*, and the latter provides the relative information about the

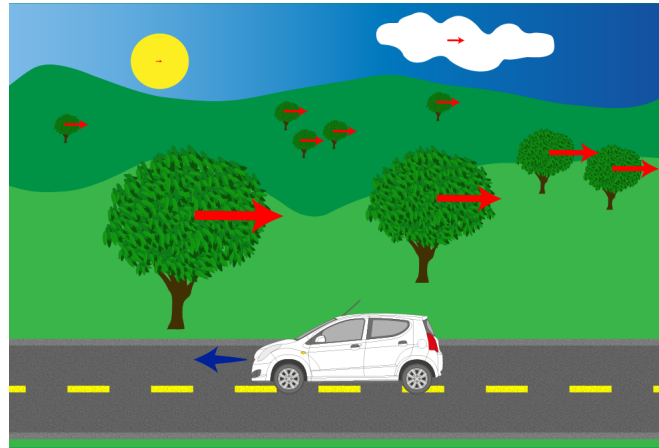


Figure 1.6: Example of motion depth cues; motion vectors are represented by red arrow; close objects move faster than far away ones.

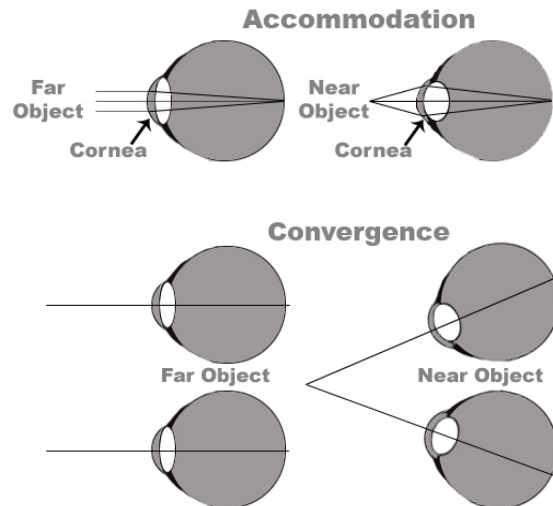


Figure 1.7: Accommodation and vergence processes at near and far distance.

distance to an object.

1.2.2.4 Ocular depth cues

When our eyes move to fixate on objects at different distances, feedback to the brain from eye muscles can provide the brain with information about their location. These oculomotor cues are of two types: *accommodation* and *convergence*. Both of these phenomena are related to the state of the eye (Figure 1.7), and provide depth information only for objects that are less than 1 meter or so away.

Accommodation is the tension of the muscle that changes the focal length of the lens of eye. Thus, it brings into focus objects at different distances. The muscular activity necessary for this accommodation acts as a signal for the brain to

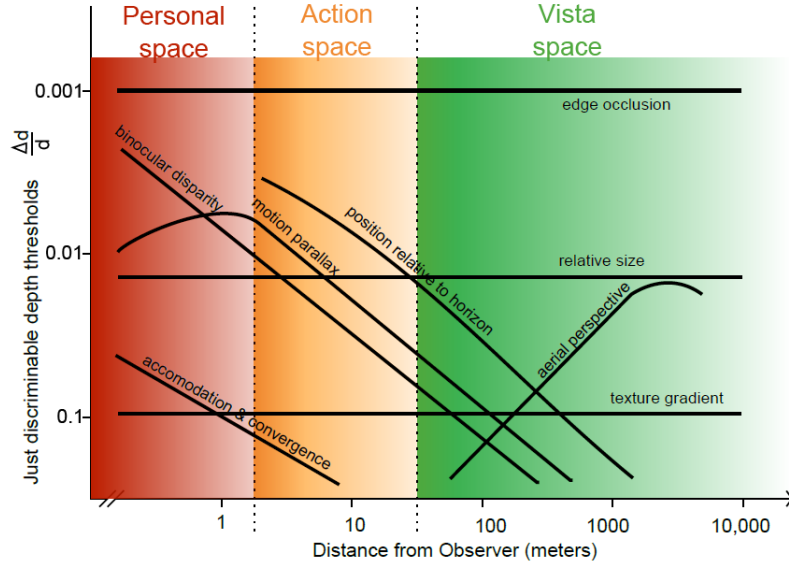


Figure 1.8: Just-discriminable depth thresholds versus the log-distance from the observer.

generate perception of depth and distance. This depth cue is quite weak, and it is effective only at short viewing distances (less than 2 metres) and with other cues.

Vergence is defined as the movement of the two eyes in opposite direction to enable the fixation to a point or region of interest. When objects are very near, the eyes converge to fixate on them; as the object moves further away, the eyes diverge to maintain fixation. The brain uses the information it receives about the degree of rotation as a cue to interpret the distance of the perceived objects. This depth cue is effective only for short distances (less than 10 metres).

1.2.2.5 Combination of depth sources

As mentioned previously, a huge number of depth information sources can be perceived and used by the HVS. If a real scene is observed with the two eyes, stereopsis is available. If, instead, a 2D image/video is observed, only static/dynamic cues are available. Depth perception is the result of the use of many cues with various degrees, as demonstrated by Cutting *et al.* [11] and illustrated in Figure 1.8. Many studies investigate how these cues are combined [17, 18], but there is a general agreement in the community that the most reliable cues are the ones obtained when both eyes operate (stereopsis), followed by dynamic cues and static cues.

1.3 Some perceptual computational models

In the last decade, we have witnessed a huge increase in the use of perceptual approaches to improve the performance of multimedia processing techniques [79]. Based on the characteristics of the HVS, different models have been developed in

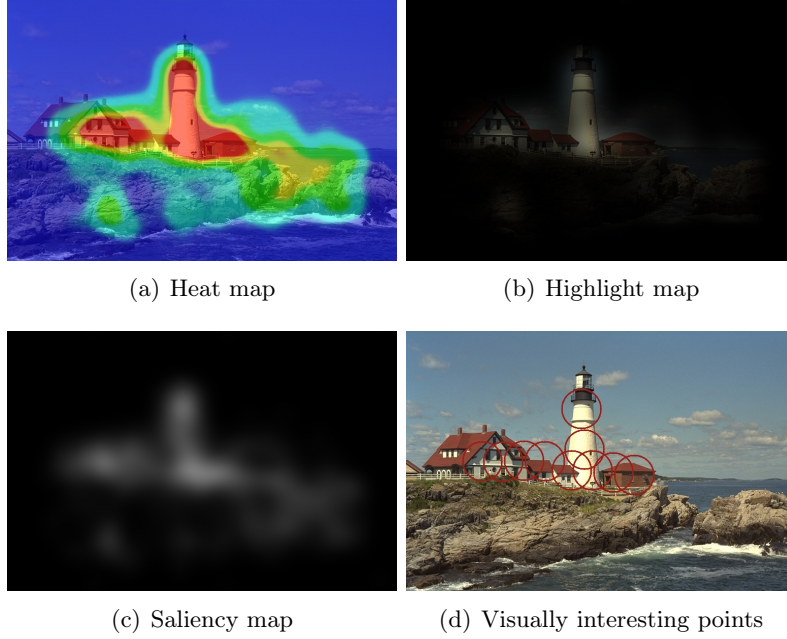


Figure 1.9: Computational modelling of the visual attention.

many challenging areas of image processing such as image coding, enhancement and watermarking, image quality assessment and so on. In the following we try to describe some of them.

1.3.1 Saliency models

A very important HVS characteristic to be considered is the visual attention mechanism. When we observe our visual environment, we do not perceive all its components with equal interest. Some objects draw our attention more than others, either by their saliency or by our a priori intention. This attention is facilitated by a retina that contains a high-resolution central fovea and a low-resolution periphery. The visual attention guides this anatomical structure to important parts of the scene to gather more detailed information [19]. This ability allows us to remove redundant visual information and allocate limited perceptual and cognitive resources for the most relevant parts of our visual field.

Thus, the aim of the computational attention models is to predict the areas within an image/video attracting human attention. The result of such algorithms is often an attention map or a saliency map providing higher intensities for the most important areas (Figure 1.9).

Before introducing the two main saliency computational models, it is important to define some key terms used in this research field. While the terms attention, saliency, and gaze are often used interchangeably, each has a more subtle definition that allows their delineation [19].

Attention is a general concept covering all factors that influence selection mechanisms.

Saliency intuitively characterizes some parts of a scene, which could be objects or regions, that appear to an observer to stand out relative to their neighboring parts.

Gaze is a coordinated motion of the eyes and head, has often been used as a proxy for attention in natural behavior. For instance, a human or a robot has to interact with surrounding objects and control the gaze to perform a task while moving in the environment. In this sense, gaze control engages vision, action, and attention simultaneously to perform sensorimotor coordination necessary for the required behavior.

Following the psychological experiments, some emerging research focused on computational models of visual attention. Saliency estimation methods can broadly be classified, according to which mechanism the cortex is driven, as *bottom-up* or *top-down* models.

1.3.1.1 Bottom-up saliency model

This mechanism describes our focus of attention on salient changes in the environment in an unconscious fashion; it is fast and most likely feed-forward. Bottom-up cues are mainly based on characteristics of a visual scene (stimulus-driven), due to its low level features transmitted to the HVS. A prototypical example of bottom-up attention is looking at a scene with only one horizontal bar among several vertical bars where attention is immediately drawn to the horizontal bar.

One of the earliest and the most referenced model is the framework proposed by Koch *et al.* [20], and Itti *et al.* [21]. Figure 1.10 illustrates the general architecture of this model. Visual input is first decomposed into three independent channels that represent the most important early visual features such as intensity, color, and orientation. For each channel, a set of feature maps is then computed by a center-surround operation which is implemented as the difference between fine and coarse scales of a Gaussian pyramid. Finally, the overall saliency map is given by normalizing and linearly combining all these feature maps. Apart from Itti *et al.*'s model which is biologically inspired, other pure mathematics-based approaches have been also proposed in the literature. In [23], Gao *et al.* proposed a simple definition for saliency using the idea that a pixel is salient if it is distinguished from its surround. The saliency of a point is measured as the Kullback-Leibler divergence between the histogram at the point and its surrounding region computed on the Gaussian (or Gabor) filtered image. In [24], Zhang *et al.* define the visual saliency as the conditional probability of finding a target at every location given the visual features observed. A Bayesian probabilistic framework is then applied to solve this problem and provide the final saliency map.

While many models fall into this category, they can only explain a small fraction of eye movements since the majority of fixations are driven by task.

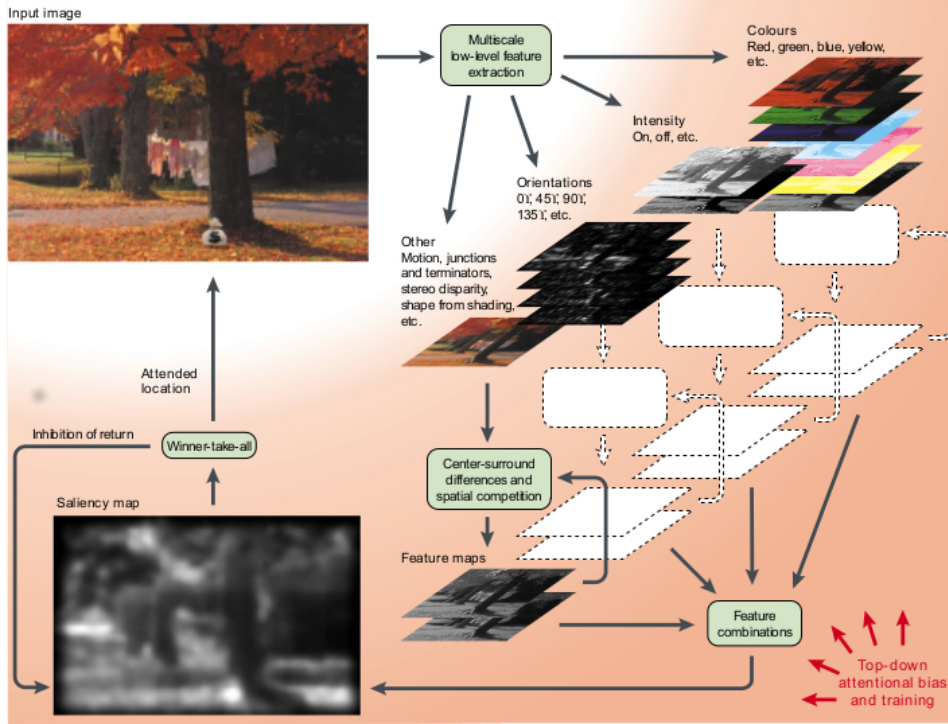


Figure 1.10: Flowchart of the model proposed by *Itti et al.* [22].

1.3.1.2 Top-down saliency model

This is a high level feature of the visual cortex. Top-down cues (task-driven) are determined by cognitive phenomena like knowledge, expectations, reward, and current goals. One of the most famous examples of top-down attention guidance is from Yarbus in [27], who showed that eye movements depend on the current task with the following experiment: Subjects were asked to watch the same scene under different conditions (questions) such as “estimate the material circumstances of the family,” “what are the ages of the people?”, or simply to freely examine the scene. Eye movements differed considerably for each of these cases.

Some top-down factors in free-viewing are already known although active investigation still continues to discover more semantic factors. For instance, Einhauser *et al.* proposed that objects are better predictors of fixations than bottom-up saliency [28]. Elazary *et al.* showed that interesting objects are more salient [29]. Cerf *et al.* showed that the meaning objects such as faces and text attract human attention [30].

According to the results of these methods, the basic idea is that a weighted combination of features, where weights are learned from a large repository of eye movements over natural images, can enhance saliency detection compared with unadjusted combination of feature maps.

Current computational models of visual attention mainly focus on bottom-up much more than top-down mechanism. However, visual attention is assumed to

be guided by a combination of at least these two factors. Some studies about the top-down visual attention as well as models incorporating these two mechanisms can be found in [25, 26]. To explore in more detail the saliency models, the reader is referred to [19].

1.3.2 Just Noticeable Difference

Typically, the various image processing tasks process the visual signal with the aim to improve its visual quality, reduce its representation or to protect its content. It is well known that the HVS cannot detect all changes in an image due to its underlying physiological and psychological mechanisms. Therefore, since the HVS is the ultimate receiver of the visual signal, it is advantageous to integrate the HVS visibility thresholds into image processing algorithms. This allows to avoid altering the sensitive regions and to be more aligned with human perception [31].

Just-noticeable difference (JND) refers to the visibility threshold below which any change cannot be detected by a standard observer [32]. In other words, JND can be defined as the minimum amount by which a stimulus intensity must be changed relative to a background intensity in order to produce a noticeable variation in sensory experience [260]. The determination of the JND is complex and is related to the HVS characteristics. Therefore, before describing some main JND models, it is important to introduce the HVS features on which the JND is based, namely, contrast sensitivity, luminance adaptation, and contrast masking.

Contrast masking

Masking is a general concept referring to the reduction of visibility of a target signal (the masked) due to the presence of another (masker). The contrast masking effect is the minimum contrast value (threshold contrast) necessary for an observer to detect a change in intensity. It is measured by superposing target stimulus onto the masker stimuli, and varying the masker contrast until to detect a stimulus masked. Generally higher masking contrast and larger similarity between the masker and the target in their spatial frequencies, orientations, and phases will lead to higher masking effect [33, 34].

Luminance adaptation

Also called *luminance masking* refers to the variation of the HVS sensitivity to a wide range of background luminance. Outside of this range, our intensity discrimination ability deteriorates. This implies that background mean luminance masks the luminance variation, *i.e.*, higher the background mean luminance, the stronger the masking effect is.

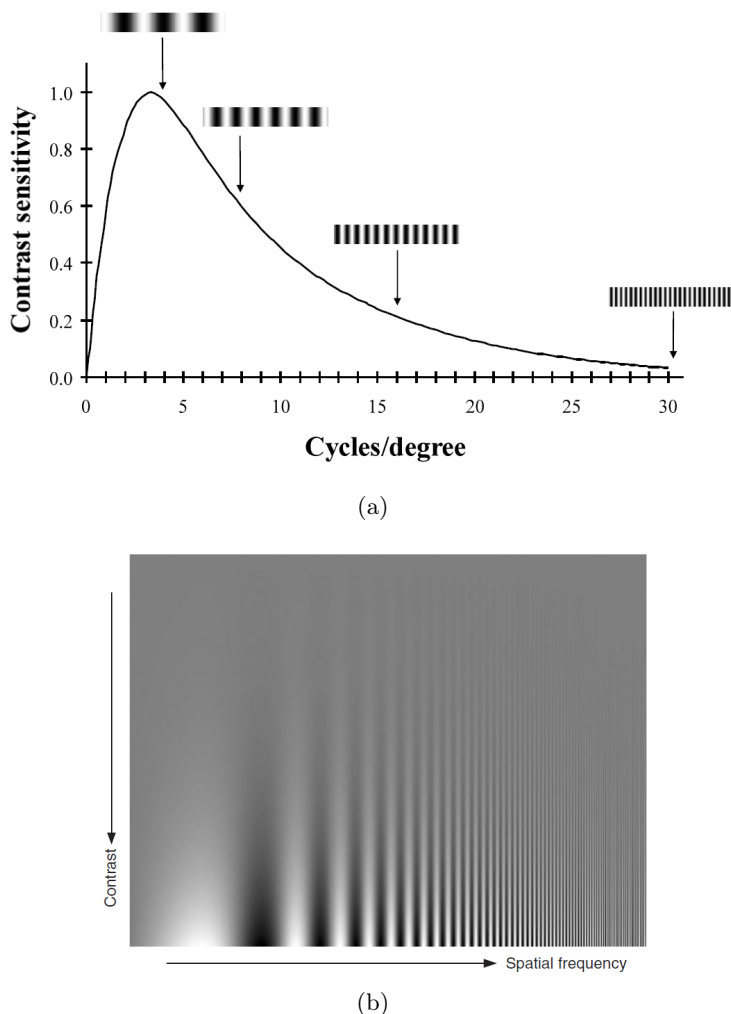


Figure 1.11: Contrast sensitivity function. (a) normalized visual sensitivity as a function of spatial frequency, (b) Campbell-Robson CSF chart [35].

Contrast Sensitivity Functions

Contrast is a key concept in vision science because the information in the visual system is represented in terms of contrast and not in terms of the absolute level of light. Thus, relative changes in luminance are important rather than absolute ones [260].

The contrast sensitivity function (CSF) models the sensitivity of the HVS as a function of the spatial frequency content in visual stimuli. CSF is the inverse of the contrast threshold and it derived from psychophysical experiments. A typical CSF is shown in Figure 1.11(a). In general, the CSF has a band-pass nature. The peak sensitivity is around four cycles per degree of visual angle and drops significantly with increasing frequencies [207]. This effect is demonstrated in Figure 1.11(b), which is widely known as the Campbell-Robson CSF chart [35]. In this figure, the

minimum and maximum amplitude along a horizontal line through the image remain constant. However, one will have the impression that the intensity is smaller on the left and right. This perceptual intensity represents the notion of contrast sensitivity.

1.3.2.1 JND models

In the literature, several JND models have been proposed. Typically, they can be determined in the spatial domain or the transform domain, such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), or the combination of the two schemes [36].

JND models in the spatial domain [37, 257], named as the *pixel-based JND*, mainly focus on the background luminance adaptation and the spatial contrast masking. For instance, the model proposed by Chou *et al.* takes into account the luminance adaptation and the so-called *texture masking* [37]. The latter is determined according to local spatial activities of the region (*e.g.*, characterized by computing gradients around the pixel). Finally, JND is decided as the dominant effect between texture masking and luminance masking. This model has been improved by Yang *et al.* in [38].

However, pixel-based JND models do not consider the human vision sensitivities such as CSF. Therefore, it cannot describe the HVS properties accurately. To overcome this problem, some authors proposed to design JND models in the frequency domain.

JND models generated in the transform domain, namely the *subband-based JND*, usually incorporate all the major affecting factors, such as CSF, luminance adaptation and contrast masking. One of the earliest and the most known DCT based models is the one developed by Watson *et al.* [39]. Since most of image and video compression standards are based on the DCT domain, this model has been adopted for many works. In this model, the JND threshold for each DCT subband is computed by incorporating the most important mechanisms of the HVS described previously. More recently, Wei *et al.* [40] incorporate new formulas of luminance adaptation, contrast masking and Gamma correction to estimate the JND threshold in the DCT domain. Zhang *et al.* [36] propose to estimate the JND profile by summing the effects in DCT and spatial domain together.

Concerning the JND models in the wavelet domain, Liu *et al.* proposed a JND model for the discrete wavelet transform [41]. The luminance adaptation and contrast masking functions used in this model are the same as in the Watson's model (with the same parameters). For CSF modeling, they make use of the wavelet detection thresholds measured in the experiments of Watson *et al.* [42].

1.3.3 Binocular Just Noticeable Difference

In order to build a Binocular-JND (BJND) model, Zhao *et al.* conducted psychophysical experiments [83]. Based on these experiments, they proposed a BJND model to measure the joint visibility of a pair of distortions in two views. The

BJND model determines the minimum distortions in one view that evoke binocularly visible differences, given the background information and the distortions in the corresponding area of the other view.

In their proposed model, they considered two HVS characteristics namely luminance and contrast masking effects, and they modeled them to the case of binocular vision as described in the following.

Given the left and right views, as well as the disparity map (see Section 2.5) of the left image. The BJND of the left view (BJND_l) is defined as:

$$\begin{aligned} \text{BJND}_l(i, j) &= \text{BJND}_l(bg_r(i - d_l, j), eh_r(i - d_l, j), n_r(i - d_l, j)) \\ &= A_C(bg_r(i - d_l, j), eh_r(i - d_l, j)) \\ &\quad \times \left(1 - \left(\frac{n_r(i - d_l, j)}{A_C(bg_r(i - d_l, j), eh_r(i - d_l, j))} \right)^\lambda \right)^{\frac{1}{\lambda}} \end{aligned} \quad (1.4)$$

where i and j are the pixel coordinates, d_l is the horizontal disparity value at pixel (i, j) . The parameter λ controls the influence of the noise in the right image, and it was suggested in [114] that $\lambda = 1.25$. It should be noted that BJND_l is dependent on the background luminance bg_r , the edge height eh_r , and the noise amplitude n_r of the corresponding pixel in the right image (found by stereo matching). Where no noise in the right image, *i.e.*, $n_r(i - d, j) = 0$, the BJND_l is reduced to A_C , which is defined as

$$A_C(bg, eh) = A_{limit}(bg) + K(bg) \cdot eh \quad (1.5)$$

Thanks to psychophysical experiments, authors defined $A_{limit}(bg)$ and $K(bg)$, respectively by

$$A_{limit}(bg) = \begin{cases} 0.0027 \cdot (bg^2 - 96 \cdot bg) + 8, & \text{if } 0 \leq bg < 48 \\ 0.0001 \cdot (bg^2 - 32 \cdot bg) + 1.7, & \text{if } 48 \leq bg \leq 255 \end{cases} \quad (1.6)$$

$$K(bg) = -10^{-6} \cdot (0.7 \cdot bg^2 + 32 \cdot bg) + 0.07 \quad (1.7)$$

where bg is the mean of the luminance values of a block of 5×5 centered on the corresponding pixel position, and the edge height eh is computed by the 5×5 Sobel operators as follows:

$$eh(i, j) = \sqrt{E_H^2(i, j) + E_V^2(i, j)}, \quad (1.8)$$

$$E_k(i, j) = \frac{1}{24} \sum_{h=1}^5 \sum_{v=1}^5 I(i - 3 + h, j - 3 + v) \cdot G_k(h, v), \quad (1.9)$$

with $k = H, V$.

where $I(i, j)$ represents the luminance value of pixel (i, j) , and the detailed repre-

sensation of the Sobel operator $G_k(h, v)$ is as follows

$$G_H = \begin{pmatrix} -1 & -2 & 0 & 2 & 1 \\ -2 & -3 & 0 & 3 & 2 \\ -3 & -5 & 0 & 5 & 3 \\ -2 & -3 & 0 & 3 & 2 \\ -1 & -2 & 0 & 2 & 1 \end{pmatrix}, \quad G_V = \begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 3 & 5 & 3 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ -2 & -3 & -5 & -3 & -2 \\ -1 & -2 & -3 & -2 & -1 \end{pmatrix} \quad (1.10)$$

To the best of our knowledge, this is the first model to measure the perceptible distortion threshold of stereoscopic images for binocular vision, and it has been used in the proposed approaches described in Chapters 4 and 5.

1.4 Conclusion

In this chapter, first, the physiology of the human vision has been overviewed along with the most important high level mechanisms of the human visual system. Furthermore, the different depth cues have been introduced and the sensitivity of humans for these various sources of depth information has been discussed. Even though numerous depth cues exist, the main exploited sources of depth for 3D technology are the binocular cues. Accordingly, the most existing 3D display systems aim to project two slightly different images into each eye.

Given that 3D video is a technology that exploits the properties of HVS, understanding of binocular vision is essential for reliably developing and testing any 3D video system. To do so, the chapter described some human factors of 3D viewing, which is a very important topic.

Finally, the last sections are dedicated to the presentation of some perceptual computational models. These models have been developed based on the characteristics of the HVS, with the aim to optimize and enhance the different image processing tasks. Specifically, visual attention models with their two main approaches, bottom-up and top-down, were presented. Next, we have reported the definition of the JND models, followed by an extension to the case of binocular vision, known as BJND. The latter has been adopted in the different approaches described in the following chapters.

Chapter 2

3D Video Representation and Coding

Contents

2.1	Introduction	22
2.2	3D video application scenarios	23
2.2.1	Three-dimensional television	24
2.2.2	Free view point video	24
2.2.3	Immersive teleconference	25
2.3	3D video data representation	26
2.3.1	Stereoscopic 3D Video	26
2.3.2	Multi-View Video	27
2.3.3	Video-plus-Depth	28
2.3.4	Multi-view Video-plus-Depth	29
2.3.5	Layered Depth Video	30
2.4	3D video coding	30
2.4.1	Principle	31
2.4.2	H.264/MPEG-4 AVC Video Codec	34
2.4.3	Stereo/Multi-view Video Coding	37
2.4.4	Coding performance	40
2.5	Disparity estimation	41
2.6	Conclusion	44

2.1 Introduction

The consumers' thirst for more immersive multimedia technologies, together with the industry interest to provide a leap forward in QoE as close as possible to the

reality, this has led to the fast popularization of 3D technologies, creation of 3D video products, and deployments of new 3D services. Two major 3D video applications may be cited, the 3-Dimensional TeleVision (3DTV) [77] and Free Viewpoint Video (FVV) [82]. 3DTV delivers to viewers a feeling of immersion, by bringing an additional dimension of depth for the perception of the scene, while FVV allows the user to change his viewpoint and viewing direction within a scene in an interactive manner.

In order to deploy the 3D technology on real consumer electronic platforms, several challenges need to be addressed, from acquisition, representation, coding, transmission, and to display. Furthermore, it turns out that every single step of the processing chain is linked to the others. In this respect, it is obvious that standardized media formats for representation, coding and transmission are needed, because interoperability and compatibility are crucial for the success of 3D video applications.

Three types of formats need to be considered separately for each step of the 3D video processing chain: 1) production format for the acquisition step, 2) transport format exploited for coding and transmission, and 3) adopted format for display [53]. Consequently, data representation is a central issue in 3D video system and has a significant impact on the overall performance. For instance, different factors must be taken into account when selecting the 3D video format, including available bandwidth, minimum acceptable visual quality, display type as well as backward compatibility with existing equipment and infrastructure [77]. In addition, several attributes are highly desirable, in particular decoupling of content creation from display technology has to be achieved.

Furthermore, to enable 3D video applications a very large number of views is required. As this huge amount of data is making its way to consumers, their transmission over existing bandwidth-limited infrastructures poses a great challenge for the realization of these applications. Consequently, efficient compression and transmission is essential for the deployment of 3D video.

The rest of the chapter is organized as follows. Multi-view video systems and their different applications are presented in Section 2.2. Section 2.3 describes different representation formats of 3D video and discusses their respective strengths and shortcomings. Section 2.4 reviews state-of-the-art techniques to compress 3D video. Next, the disparity estimation problem is introduced in Section 2.5. Finally, Section 2.6 summarizes and concludes the chapter.

2.2 3D video application scenarios

The denomination *3D video* is associated with the new applications providing the ability to the viewer to perceive an illusion of depth, or the illusion to navigate in a 3D space, in order to improve its immersion experience. 3D video applications can be grouped under three categories, free-viewpoint video/television (FVV/FTV), three-dimensional television (3DTV), and immersive teleconferencing. Each of them

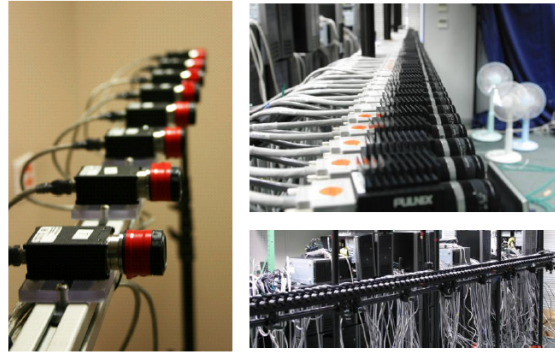


Figure 2.1: An examples of typical multi-view capture setup.

will be described below.

These applications share the same concept of capturing multiple view points of the same scene by a set of synchronized cameras. Two examples of practical multi-view capture setups are illustrated in Figure 2.1, one from Mitsubishi Electric Research Labs in US [44] and another from Nagoya University in Japan [45].

Figure 2.2 shows the end-to-end architecture of different multi-view video (MVV) application scenarios. In this figure, a MVV is first captured and then encoded by a multi-view video coding (MVC) encoder. A server transmits the coded bitstream(s) to different clients with different capabilities, possibly through media gateways. At the final stage, coded video is decoded and rendered with different means according to the application scenario and capabilities of the receiver [97].

2.2.1 Three-dimensional television (3DTV)

After the black-and-white and high-quality color television, 3DTV is believed to be the next major revolution in the history of television providing to the viewer a feeling of immersion. 3DTV refers to the extension of traditional 2DTV with the depth perception capability. In this application, more than one view is decoded and displayed simultaneously, where each viewer sees two views, one for the right eye, and other for the left eye. 3DTV can be simply achieved by using stereoscopic video. Stereoscopic displays show two simultaneous views which requires the use of special glasses (polarized or active shutter glasses). However, it is desirable to experience 3D viewing directly through a 3D display without the need to wear special glasses. This can be realized by autostereoscopic displays. The latter are able to display higher number of views at the same time increasing the observer freedom by supporting head parallax.

2.2.2 Free viewpoint video (FVV)

In FVV, the users have the ability to select their viewing directions and viewpoints in an interactive way, which provides realistic impressions, *i.e.*, the viewer can navigate

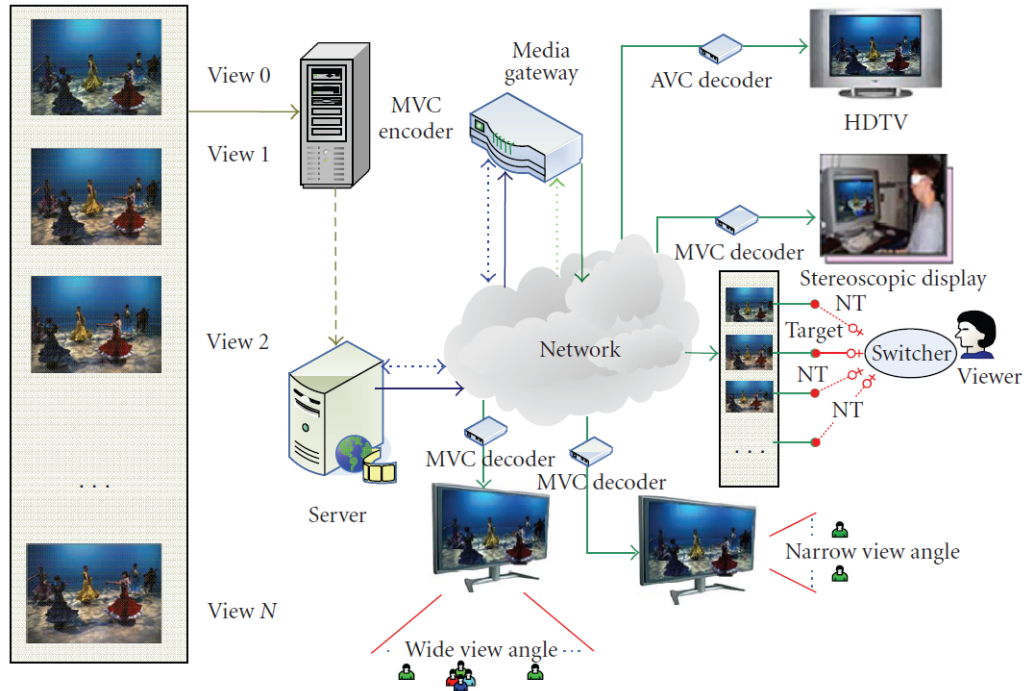


Figure 2.2: Multi-view video system for different application scenarios [97].

freely in the scene within a certain range, and view the 3D scene from different viewing angles. FVV finds potential usage in the following applications [47]:

- As a mode of entertainment – sports events, multi-user games, movies, drama, and news coverages
- Development of education videos – archives of cultural activities, instructions of sports playing, illustrating medical surgery
- For sightseeing websites and videos – zoos and aquariums, botanical gardens, museums, historical and archaeological sites
- For surveillance applications – traffic monitoring, monitoring of parking bays, place that require security such as banks
- As a tool for the creation of new types of media art and digital content

2.2.3 Immersive teleconference

A number of companies have already produced high-end video-conferencing facilities, which attempt to provide greater sense of the remote participants being in the same room. Therefore, the introduction of 3D to such communication systems would seem like a natural step forward in terms of immersion. In this family of applications, the users are at different geographical locations and see each other in either free view point or 3DTV mode. This brings a feeling of immersiveness and more naturalness



Figure 2.3: Stereoscopic 3D video representation: example of left and right views.

to their teleconference. Notable examples of these applications are teleconference, remote training and telesurgery.

2.3 3D video data representation

This section introduces different 3D video data representations currently in use, or currently being considered by the research community, including stereoscopic 3D (S-3D) video, multi-view video (MVV), video-plus-depth (V+D), multi-view video-plus-depth (MVD), and layered depth video (LDV) [78]. The common requirement is that all of them must provide means for stereoscopic viewing, since this is the underlying principle behind depth perception from visual information [54]. These various representations have different advantages and drawbacks with regard to complexity, efficiency and functionality.

2.3.1 Stereoscopic 3D (S-3D) Video

This format is considered as the most widespread data format and the most simple type of acquisition technique for 3D video technology [101, 102]. S-3D consists of capturing two images from two slightly different positions and provides each of them to each viewer's eye. The distance separating cameras is called *stereo baseline*. As a result, a pair of 2D videos is acquired, one for the left eye, and the other for the right eye, as illustrated in Figure 2.3.

To facilitate the introduction of stereoscopic services through the existing infrastructure and equipment, a simple way to represent stereoscopic video data is to multiplex them into a single frame or sequence of frames. The different multiplexing approaches are shown in Figure 2.4, which include time and spatial multiplexing. For instance, with spatial multiplexing, left and right views are filtered and decimated horizontally or vertically to fit within the size of the original image, at the cost of losing the spatial resolution, and stored either side-by-side or over/under. Temporal multiplexing is also possible, and is achieved by interleaving left and right views as alternating frames or fields.

Nevertheless, this way of capturing and projecting has shown its limitations. A principal drawback lies in the necessity of wearing special 3D glasses. This often

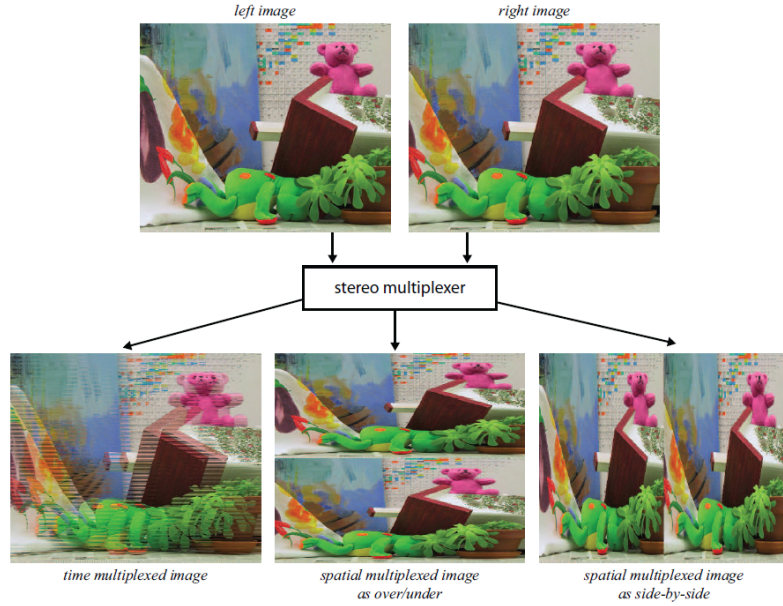


Figure 2.4: Stereoscopic multiplexing pictures [55].

produces discomfort to the user and reduces the opportunity of 3D deployment at home. In addition to that, since the stereo baseline is fixed at the capture stage, depth cannot be adjusted according to the viewing conditions and display characteristics (types, sizes, etc.). Also, head motion parallax is not supported with such data format (different perspective, occlusions and dis-occlusion when moving the viewpoint).



Figure 2.5: Example of a multi-view video representation (*e.g.*, 5 views).

2.3.2 Multi-View Video (MVV)

MVV is considered as an extension of the S-3D representation to a higher number of views, where the latter are captured by several synchronized cameras from different viewpoints. An illustration of the MVV representation is given in Figure 2.5. Since multiple views are available, a wide range of autostereoscopic displays can be supported with this representation. Therefore, this allows head motion parallax viewing. However, the amount of data to be processed and transmitted increases significantly (depends linearly on the number of views) compared to the conventional stereo data or monoscopic 2D video.

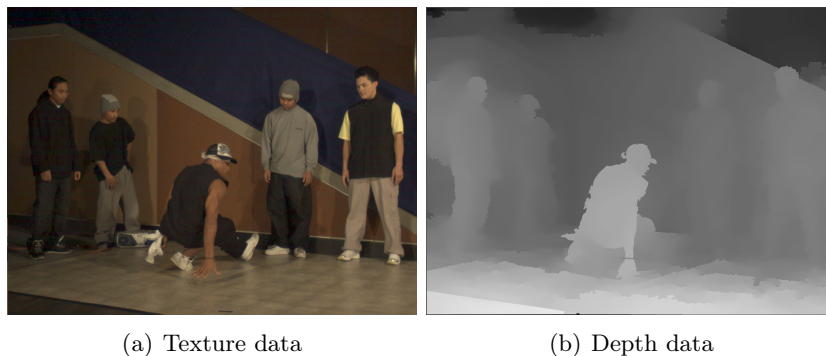


Figure 2.6: Example of a video-plus-depth representation.

2.3.3 Video-plus-Depth (V+D)

Another important class of 3D representation format is the video-plus-depth (V+D). This format consists, for one given viewpoint, of a 2D color video (namely texture data) and its associated depth map (namely depth data), as shown in Figure 2.6. This depth-based format allows the generation of virtual views at the decoder through depth-based image rendering (DBIR) techniques [47, 155] (see Figure 2.7). Moreover, depth information provides the opportunity to adjust the stereo baseline and therefore depth perception, in addition to the adaptation to the screening environment and user preference, thus reducing visual discomfort if necessary. The depth information may be obtained from a stereo pair by estimating disparity information (see Section 2.5) or obtained directly through special range cameras [56].

The depth map can be regarded as a grayscale video (luminance-only), where each pixel is assigned a depth value representing the distance between the cor-



Figure 2.7: Rendering of stereo video from V+D.



Figure 2.8: Example of MVD format with 5 views and corresponding depth maps.

responding 3D point in the scene and the camera. Typically, the depth data is quantized with 8 bits, *i.e.*, the closest point is associated with the value 255 and the farthest point is associated with the value 0. The depth data range is restricted to a range in between two extremes Z_{near} and Z_{far} , corresponding to the minimum and maximum depth-distance.

The concept of V+D is attractive due to the backward compatibility and extended functionalities. Add to that, depth information leads to a display-independent solution supporting both stereo and multi-view displays. However, the shortcomings of this approach lies the increase in complexity for both transmitter and receiver sides. In addition, it is only capable of rendering a limited depth range making difficult to handle occlusions and disocclusions.

2.3.4 Multi-view Video-plus-Depth (MVD)

In order to overcome the aforementioned drawbacks, more advanced 3D video systems providing more flexible 3D visualization are required. These types of displays require a 3D video format that allows rendering a very large number of views (28-52), and display tens of views simultaneously, but ensure through lens technology that the viewers only perceive a stereo pair from a given viewpoint without the need to wear any special glasses.

The previously described 3D video formats are not sufficient to support such requirements and are not feasible to provide this large number of views. For instance, V+D supports only a very limited continuum around the available original view since view synthesis artifact increase dramatically with the distance of the virtual viewpoint. In order to address this issue, the Multi-view Video plus Depth (MVD) format has been proposed [48, 49]. MVD consists of multiple color videos captured from different viewing angles, with a depth value associated with each texture pixel, as illustrated in Figure 2.8.

By adopting the MVD format, the advanced 3D video systems, such as multi-view auto-stereoscopic display and FTV applications, can be effectively supported with only transmitting a limited subset of views plus their associated depth maps to the receiver. For instance in Figure 2.9, the 3D display projects 9 views (view

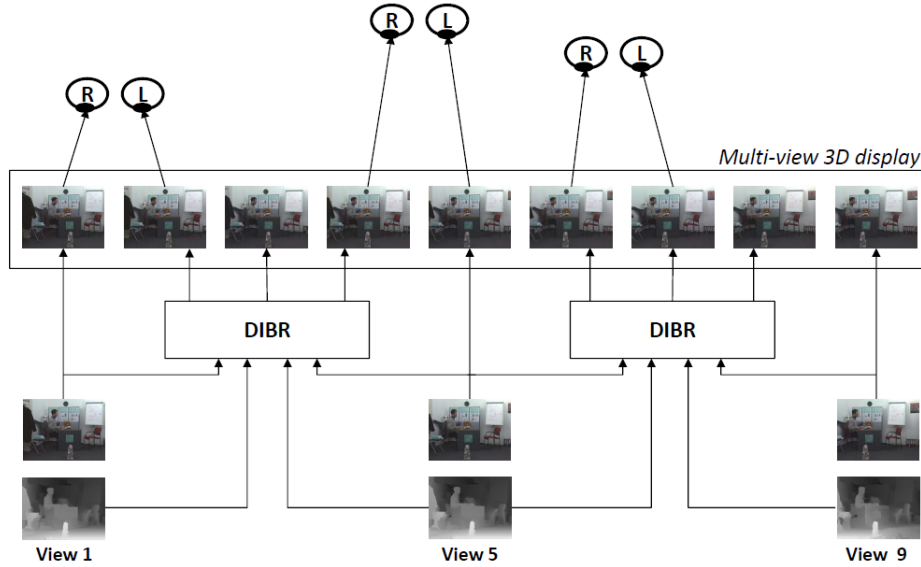


Figure 2.9: Virtual views by depth-image-based rendering (DIBR) for efficient support of multi-view auto-stereoscopic displays.

1 to view 9) at the same time. However, only 3 views (view 1, view 5 and view 9) plus their associated depth maps are transmitted. After decoding these data (texture and depth), the remaining views are generated as virtual views by DIBR techniques, in between existing camera positions.

2.3.5 Layered Depth Video (LDV)

Layered Depth Video (LDV) is a derivative and an alternative to MVD representation. LDV is the temporally extension of the so-called *layered depth image* (LDI). LDI includes one texture image with associated depth map as the main layer, together with an enhancement layer including a residual texture image and a residual depth map (see Figure 2.10). The residual layers include data from other viewing directions, not covered by the main view. LDV supports rendering of virtual views and therefore multi-view autostereoscopic displays similarly to the concept illustrated in Figure 2.9 [78].

LDV can be generated from MVD by warping the main layer image onto other contributing input images (*e.g.*, an additional left and right views). Thus avoiding MVD limitation of storing redundant identical textures obtained from different views, the common textures are all fused and expressed in one common view.

2.4 3D video coding

As mentioned previously, in order to maximize the 3D viewing experience multiple views are needed. This set of views can be provided by using different 3D video representations introduced above. Since representation formats consist of at least

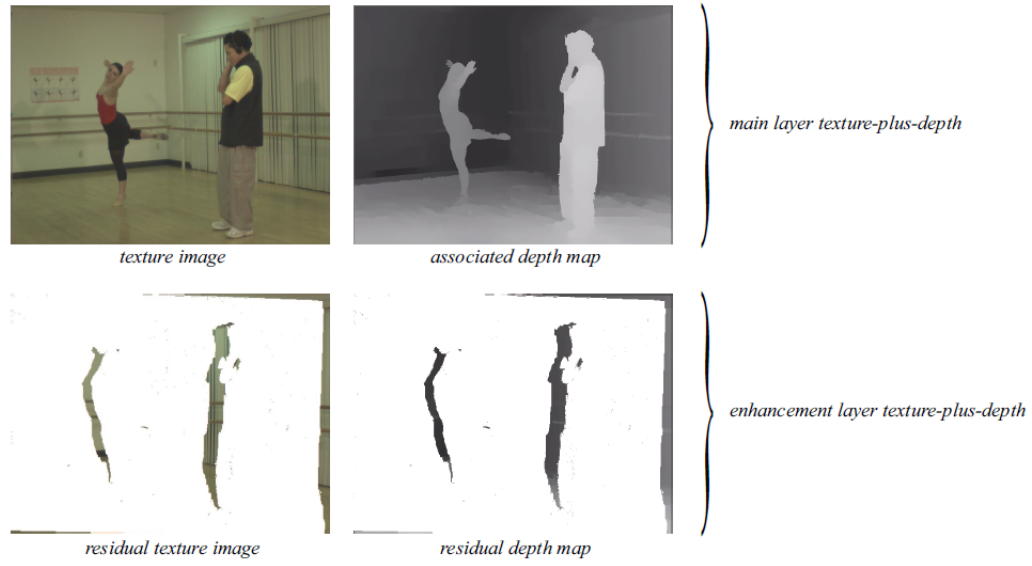


Figure 2.10: Example of LDI.

two video sequences and possibly additional depth data, this makes the task of compression essential for 3D video delivery, especially over a channel with a limited bandwidth. High efficiency corresponds to a high video quality with a fixed bandwidth, or a lower bandwidth with the same visual quality.

In this section, we introduce the existing advanced 3D video coding standards, which respond to the problem of efficiently encoding the huge amount of data generated by MVV. We will first present the fundamental principles of a video coding, since most of the standards for encoding 3D video data took inspiration from it. Then, the multi-view video coding (MVC) standard will be further discussed.

2.4.1 Principle

A video encoder and decoder taken together are known as a video codec. Most of the conventional video codecs [50] reduce the amount of data to be stored or transmitted by exploiting as much redundancies as possible. Typically, two types of redundancies or correlations are present in conventional 2D video, specifically, the spatial and temporal correlation (see Figure 2.11). The spatial correlation is the similarity within regions in the same frame, between pixels that are close to each other, *i.e.*, the values of neighboring pixels are often very similar especially in homogeneous areas. While, the temporal correlation represents the similarities between successive frames in time order.

A video encoder consists of three main functional units: a prediction model, a transform or spatial model and an entropy encoder [57]. This sequence of steps is illustrated in Figure 2.12. The major video coding standards (MPEG-1, MPEG-2, MPEG-4 Visual, H.261, H.263 and H.264) since 1990's have been based on this same

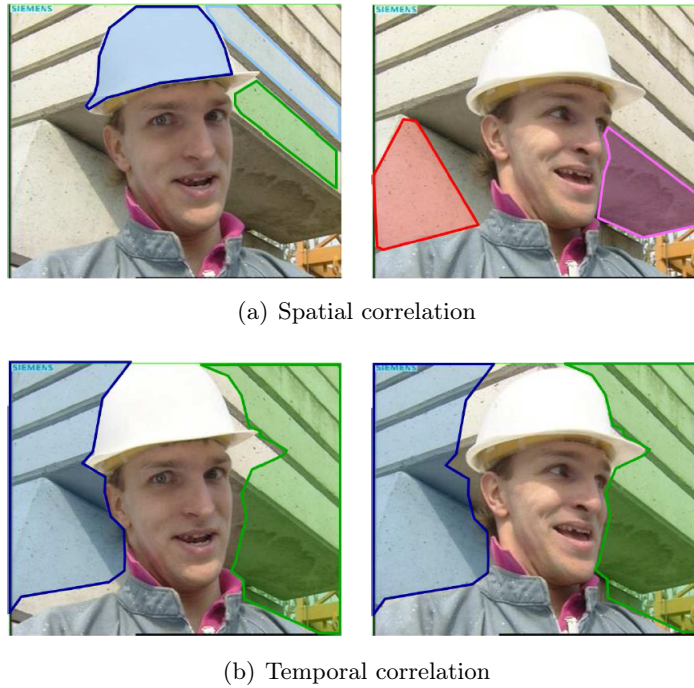


Figure 2.11: An example of spatial and temporal correlation in a 2D video sequence.

generic design, including a temporal prediction, sometimes referred to as DPCM (Differential Pulse Code Modulation), a transform stage, where the ever-popular discrete cosine transform (DCT) is used, and finally an entropy coder. The model is often described as a hybrid DPCM/DCT codec or block-based hybrid video coding approach. Figure 2.13 shows the basic DPCM/DCT hybrid encoder.

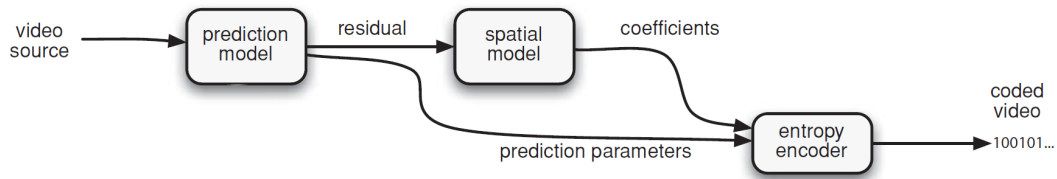


Figure 2.12: Video encoder block diagram.

The prediction model attempts to reduce redundancy by exploiting the similarities between neighboring image pixels and/or neighboring video frames (spatial and/or temporal correlations). Typically, this is achieved by a combination of temporal and spatial predictions. The prediction model aims at sending a reference image sample and a distance to the reference instead of sending two similar samples.

The temporal prediction (inter prediction) is carried out by forming a predicted frame, and subtracting it from the current frame. The output of this process is the residual frame. The accuracy of the prediction can be improved by estimating the

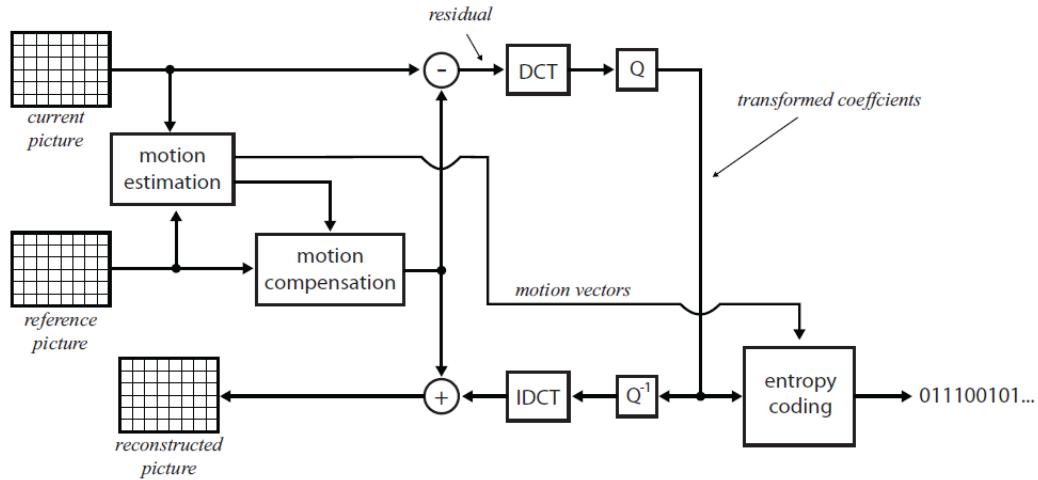


Figure 2.13: DPCM/DCT video encoder.

motion between adjacent frames, and then generating a motion vector field, which minimizes the residual, as illustrated in Figure 2.14. This prediction type is called *motion compensation prediction* (MCP).

In the spatial prediction (intra prediction), the current block is predicted from previously coded blocks in the same frame. Assuming that the blocks of image are coded in raster-scan order, the upper/left blocks are available for intra prediction (see Figure 2.15). These blocks have already been coded and placed in the output bit stream. Once the prediction has been generated, it is subtracted from the current block to form a residual in a similar way to temporal prediction.

Thus, the output of the prediction model is a residual frame, and a set of model parameters indicating the spatial prediction type or describing how the motion was

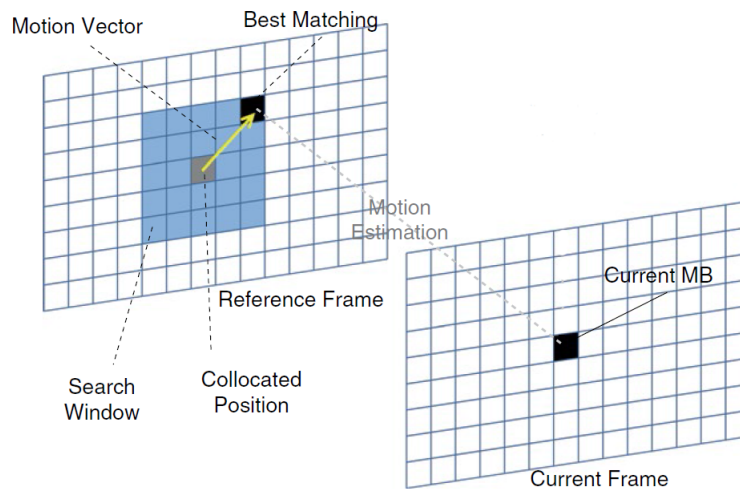


Figure 2.14: Inter prediction (temporal prediction).

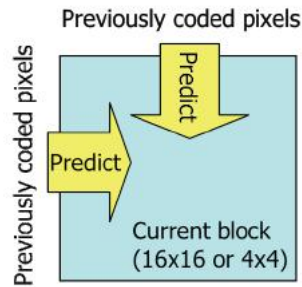


Figure 2.15: Intra prediction (spatial prediction).

compensated. The residual frame forms the input to the spatial model. The latter applies a transform to the residual samples and quantizes the results. The transform converts the samples into another domain in which they are represented by transform coefficients. The coefficients are quantized in the lossy mode to remove insignificant values, leaving a small number of significant coefficients that provide a more compact representation of the residual frame. The output of the spatial model is a set of quantized transform coefficients. Finally, after a quantization step, the parameters of the prediction model and coefficients are compressed by the entropy encoder. The entropy encoder produces a compressed bit stream that may be transmitted and/or stored.

2.4.2 H.264/MPEG-4 AVC Video Codec

The H.264/MPEG-4 AVC video coding standard is developed by the Joint Video Team (JVT) of the Moving Picture Experts Group (MPEG) of the International Standardization for Organization (ISO)/International Electrotechnical Commission (IEC) and Video Coding Experts Group (VCEG) of the International Telecommunication Union (ITU-T). H.264/MPEG-4 AVC is published by MPEG as MPEG-4 part 10 Advanced Video Coding (AVC) and by ITU-T as ITU-T Recommendation H.264.

In the following, a brief description of H.264/MPEG-4 AVC is limited to the key features that are relevant for understanding the concepts of extending H.264/MPEG-4 AVC towards multi-view video coding (MVC). For more detailed information about H.264/MPEG-4 AVC, the reader is referred to the standard description [73] and the various dedicated overview papers [51, 52].

As illustrated in Figures 2.16 and 2.17, the video sequence is first divided into groups of pictures (GOP). A picture consists of slices, and a slice consists of a contiguous sequence of macroblocks (MB). Next, a MB consists of a 16×16 block of pixels. A MB thus consists of four 8×8 blocks (sub-MB). Depending on the degree of freedom for generating the prediction signal, each GOP may include three types of pictures, as shown in Figure 2.16:

- **Intra pictures** or **I-pictures**, are coded using only information present in the picture itself, and provide potential random access points and error robustness

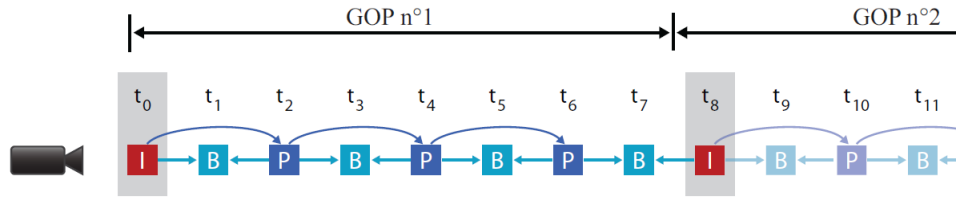


Figure 2.16: Example of GOP structure.

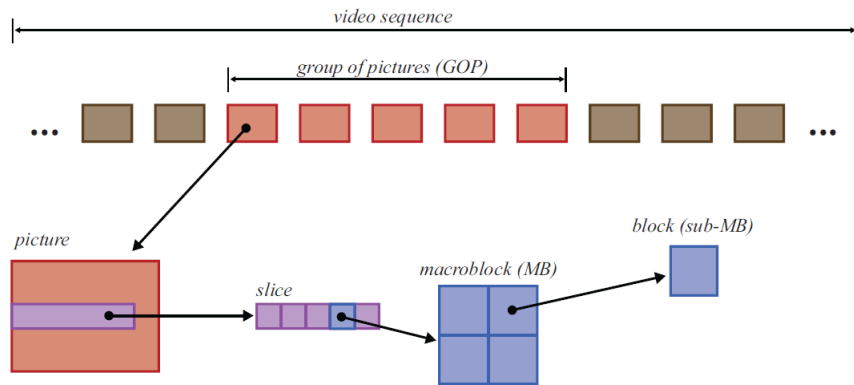


Figure 2.17: Macroblocks and slices organization.

inside the video sequence. They do not use motion prediction, and use only spatial transform coding.

- **Predicted pictures** or **P**-pictures, are coded with respect to some previous I-picture or P-picture. This technique is known as forward prediction. Similarly to I-pictures, P-pictures can also serve as prediction reference for B-pictures and future P-pictures. Moreover, P-pictures use motion compensation to provide more compression efficiency than I-pictures.
- **Bi-directional** or **B**-pictures, uses both past and future pictures as a reference. This technique is called *bi-directional prediction*. B-pictures provide the best compression since they use the past and future pictures as reference. However, the computational time and delay efficiency is the largest.

The H.264/MPEG-4 AVC follows the so-called block-based hybrid video coding approach described in the previous section. A basic H.264/MPEG-4 AVC video encoder is illustrated as in Figure 2.18. After splitting the input sequence in units of MB. A prediction MB is generated and subtracted from the current MB to form a residual. The latter is transformed, quantized and encoded. In parallel, the quantized data are re-scaled, inverse transformed and added to the prediction MB to reconstruct a coded version of the frame, which is stored for later predictions [51].

The H.264/MPEG-4 AVC includes new features that enable it to achieve a signif-

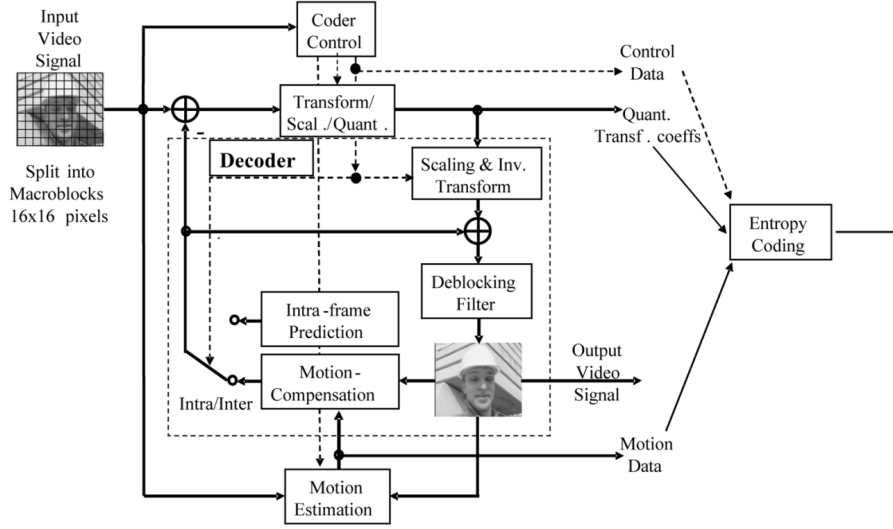


Figure 2.18: H.264/MPEG-4 AVC video encoder scheme.

icant improvement in compression efficiency relative to any prior video coding standard [51, 52]. This is mainly due to the consideration of variable block sizes for the temporal prediction, multiple reference frames, various intra prediction modes, integer transform (integer DCT), context-based adaptive entropy coding, and adaptive deblocking filter. Moreover, also due to better exploitation of the spatio-temporal correlation that may exist between adjacent MBs, with the SKIP mode in predictive P-slices and the DIRECT mode in bi-predictive B-slices.

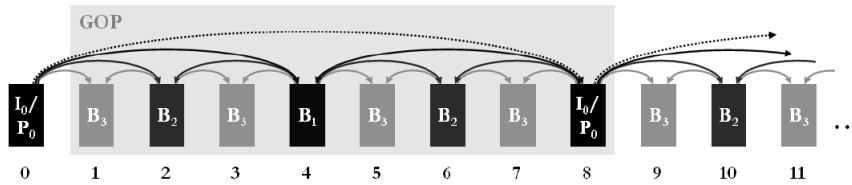


Figure 2.19: Hierarchical reference picture structure for temporal prediction.

In addition, the main difference relative to previous standards is the greatly increased flexibility and adaptability of the H.264/MPEG-4 AVC design. In contrast to older video coding standards, the coding and display order of pictures is completely decoupled. Furthermore, any picture can be marked as reference picture and used for motion-compensated prediction of following pictures independently of the coding types of the corresponding slices. These features of H.264/MPEG-4 AVC allow the selection of arbitrary coding/prediction structures, which are not possible with older video coding standards. As a result, more efficient temporal prediction structures can be supported by this standard. For instance, hierarchical prediction structure (see [58] for a detailed description) with three stages of a dyadic hierarchy is depicted in Figure 2.19. This type of prediction structure demonstrated an

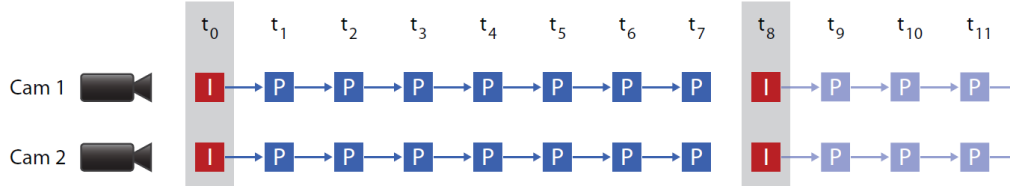


Figure 2.20: Simulcast coding approach.

improved compression performance [58].

2.4.3 Stereo/Multi-view Video Coding

The obvious solution to compress stereo or multi-view video is to encode each view independently of the other, using any video coding standard such as H.264/MPEG-4 AVC or High Efficiency Video Coding (HEVC) [75]. This solution is usually referred to as simulcast coding approach. In this case, the redundancies between views are not exploited, as shown in Figure 2.20.

This approach keeps computation and processing delay to a minimum since dependencies between views are not exploited. In addition, backward compatibility with 2D systems is maintained by decoding only one view. However, the main drawback of a simulcast solution is that coding efficiency is not maximized since redundancy between views, *i.e.*, inter-view redundancy, is not considered.

Consequently, in order to improve coding efficiency of MVV, both temporal and inter-view redundancies should be exploited. Because MVV shows the same scene from slightly different perspectives and therefore contains a large amount of inter-view statistical dependencies. These can be exploited by combining temporal prediction with inter-view prediction.

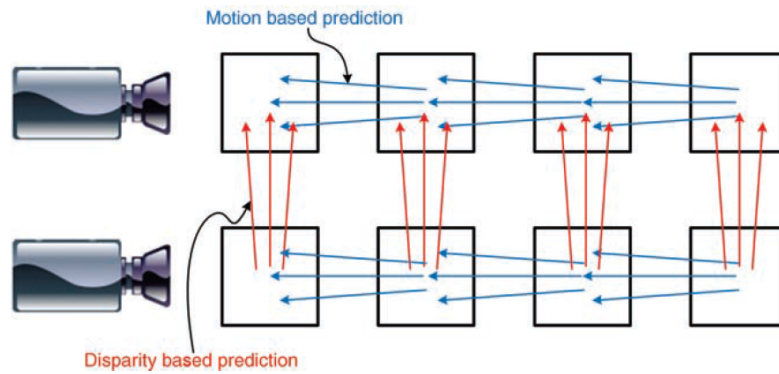


Figure 2.21: Disparity-based and motion-based prediction in a stereo pair.

As mentioned in Section 2.4.1, hybrid video coding schemes perform temporal motion-compensated prediction (MCP). This mechanism can be straightforwardly extended to inter-view prediction from neighboring views [74]. More specifically, as illustrated in Figure 2.21 for two views, temporal predictions inside each view

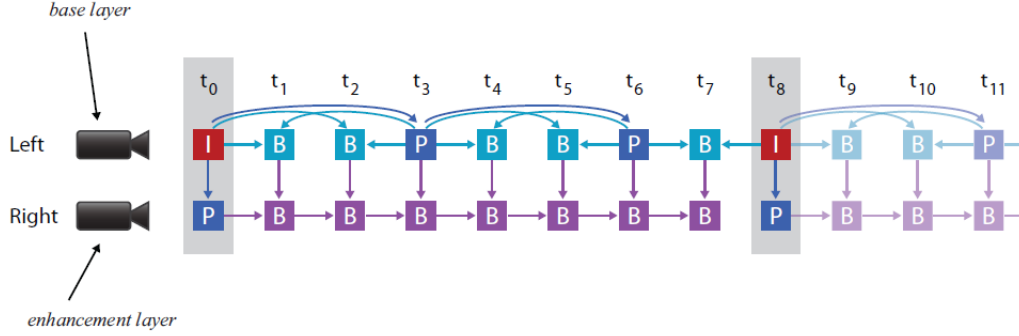


Figure 2.22: A basic inter-view prediction structure for stereoscopic video.

are combined with disparity-compensated predictions (DCP) between frames at the same temporal instant, but in different views. This approach is referred to as multi-view video coding (MVC) [74]. A typical prediction structure considering both temporal and inter-view predictions is shown in Figure 2.22.

As it can be seen, temporal prediction only is used on the base layer, while temporal prediction and inter-view prediction are simultaneously performed on the enhancement layer. As a consequence, backward compatibility with legacy 2D decoders is achieved, since the base layer represents a conventional 2D video sequence. Moreover, compression efficiency is significantly increased compared to the simulcast approach, however, at the expense of high encoding complexity.

The concept of inter-view prediction, or disparity compensated prediction, was first developed in the 1980s [59] and subsequently supported in amendments of the MPEG-2 standard [60–63]. Most recently, the H.264/MPEG-4 AVC standard has been amended to support MVC [73].

The MVC extension of the H.264/MPEG-4 AVC standard was first introduced in 2008 to support S-3D and MVV representations [73, 96], described respectively in Sections 2.3.1 and 2.3.2. Basically, as shown in Figure 2.23, the multi-view encoder receives N temporally synchronized video streams and generates one bit-stream. The multi-view decoder receives the bit-stream, decodes and outputs the N video signals.

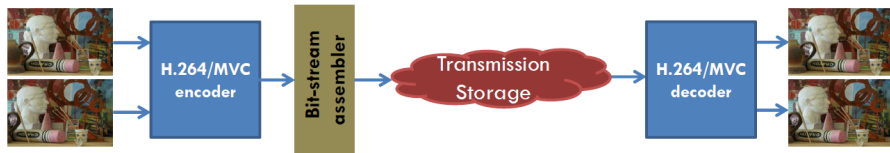


Figure 2.23: Overall structure of a MVC system with stereoscopic video ($N = 2$).

The MVC extension allows to encode multiple view videos while keeping as much as possible the same structure of H.264/MPEG-4 AVC. Inter-view prediction is a key feature of the MVC design. It is enabled through flexible reference picture man-

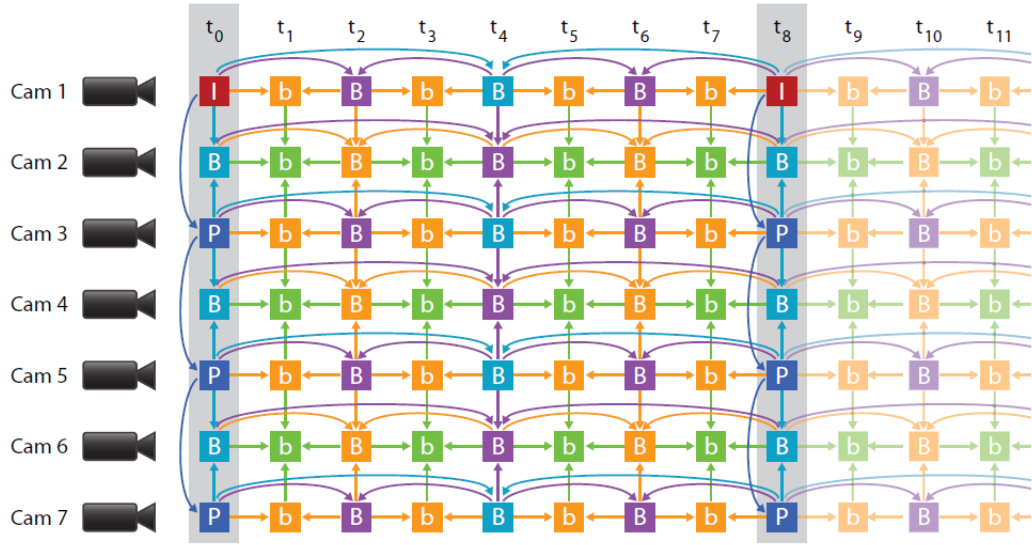


Figure 2.24: Typical MVC prediction structure.

agement that had already been designed into H.264/MPEG-4 AVC, where decoded pictures from other views are essentially made available in the reference picture lists. According to the MVC specification, only pictures corresponding to the same capture (or display) time can be added to the reference lists. Using the terminology of the standard, inter-view reference pictures must belong to the same access unit as the current picture. Block-level coding decisions are adaptive, so a block in a particular view may be predicted by a temporal reference, while another block in the same view can be predicted by an inter-view reference. With this design, decoding modules are not necessarily aware of whether a reference picture is a temporal reference or an inter-view reference picture. This distinction is managed at a higher level of the decoding process.

Hence, supporting MVC only involves small changes to high-level syntax, *e.g.*, an indication of the prediction dependency. A major consequence of not requiring changes to lower levels of the syntax (at the macroblock level and below it) is that MVC is compatible with existing hardware for decoding single-view video with H.264/MPEG-4 AVC. In other words, supporting MVC as part of an existing H.264/MPEG-4 AVC decoder should not require substantial design changes.

In MVC, different prediction structures can be envisaged. For example, a typical MVC prediction structure, utilizing both inter-view prediction and temporal prediction, is shown in Figure 2.24, wherein 7 views are coded and the GOP size is 8. The first view sequence provided by camera 1 corresponds to the base view, and other views are non-base views. In each view, a hierarchical B temporal prediction structure is used, thus the pictures in one view are with different temporal levels. This prediction structure is called *fully hierarchical encoding*, since bidirectional predictions are allowed along both time and view dimensions. The inter-view prediction is applied to every second view. This configuration aims at better exploiting all the

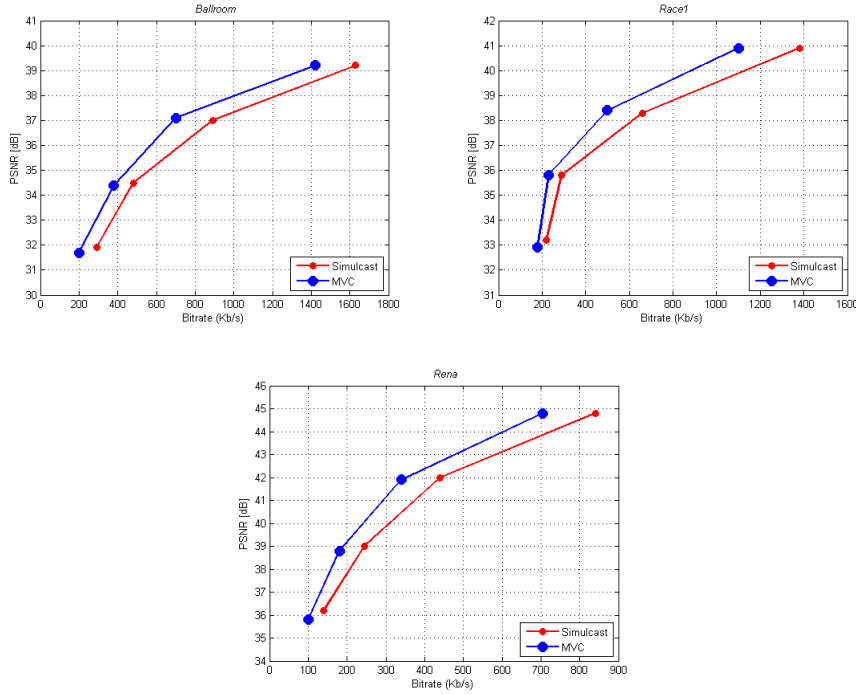


Figure 2.25: Sample coding results for *Ballroom*, *Race1*, and *Rena* sequences.

existing dependencies in time and view.

Thus, coding efficiency is improved at the cost of increased coding complexity. Furthermore synchronization and random access features are still provided by coding the key pictures of the base view in intra mode (I-pictures). Further backward compatibility is provided by this prediction structure, as the base view can be extracted and the resulting bit stream is conforming to the H.264/MPEG-4 AVC standard.

There are many possible variations on the prediction structure considering both temporal and inter-view dependencies. The structure not only affects coding performance, but has notable impact on delay, memory requirements and random access.

2.4.4 Coding performance

Several studies have been carried out in order to compute the compression performance improvement of MVC with respect to simulcast coding [64, 65, 74].

A synthetic common way to report compression performance is to use the Bjøntegaard measurement [197], which estimates the rate reduction of a given codec with respect to the reference one for the same quality over a given rate interval. This rate reduction is usually referred to as Bjøntegaard delta rate. Likewise, it is possible to define a Bjøntegaard delta peak signal-to-noise ratio (PSNR) as the average PSNR increase for a same rate in a given PSNR interval.

It has been shown that coding MVV with inter-view prediction can give signifi-

cantly better results compared to independent coding [74]. A comprehensive set of results for MVC over a broad range of test material was also presented in [64]. This study used the common test conditions and test sequences specified in [149], which were used throughout the MVC development. For MVV with up to 8 views, an average of 20% reduction in bit-rate relative to the total simulcast bit-rate was reported with equal quality for each view. Also, in some specific cases, bit-rate savings of up to 50% can be obtained [74]. Figure 2.25 shows sample rate-distortion curves comparing the performance of simulcast coding with the performance of MVC that employs hierarchical predictions in both temporal and view dimensions, according to common test conditions [149].

Subjective quality assessment results are presented in [77] for HD stereo content. It is shown that, on the whole, the subjective quality obtained by MVC is equally good to that obtained using H.264/MPEG-4 AVC simulcast, as long as the dependent-view bit rate is larger or equal to 25% of the base-view bit rate.

2.5 Disparity estimation

As mentioned in the previous section, in order to improve coding efficiency, in addition to classical temporal prediction, an efficient inter-view prediction should be performed. The inter-view correlation between adjacent views is removed via the so-called disparity estimation and compensation. Two main approaches, block- and pixel-based, have been used to estimate the disparity vectors.

The disparity consists of the spatial displacement between two images generated by recording two slightly different view angles of the same scene (see Figure 2.26). The computation of such displacement across all the pixels leads to the so-called *disparity map*. The disparity estimation problem, also known as the stereo matching problem, has been extensively studied in computer vision, and surveys of the different techniques proposed in the literature can be found in [85, 86].

Given stereo pair composed from the left I_l and right I_r images. One image is considered as a reference image while the other is the target or current one. In order to estimate the disparity values for all the pixels in the reference image, a stereo matching process is performed by searching for each point p_l in the reference image its homologous point p_r in the other image (as illustrated in Figure 2.26).

Thus, the stereo matching problem can be mathematically formulated as the search for a function d that associates a disparity vector to each pixel $p_l(x_l, y_l)$ in the reference image:

$$\begin{aligned} d : \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ (x_l, y_l) &\mapsto d(x_l, y_l) = (x_l - x_r, y_l - y_r) \end{aligned} \quad (2.1)$$

When the two images I_l and I_r are rectified, the displacement between the

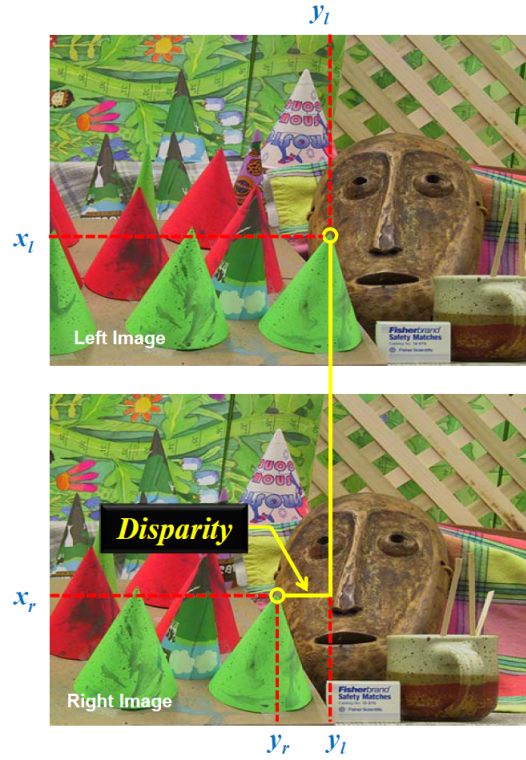


Figure 2.26: Disparity principle.

corresponding points is purely horizontal. In this case, equation (2.1) becomes

$$\begin{aligned}
 d : \mathbb{R}^2 &\rightarrow \mathbb{R} \\
 (x_l, y_l) &\mapsto d(x_l, y_l) = y_l - y_r
 \end{aligned} \tag{2.2}$$

Figure 2.27 shows an example of the *cones* stereo image pair and the corresponding disparity map.

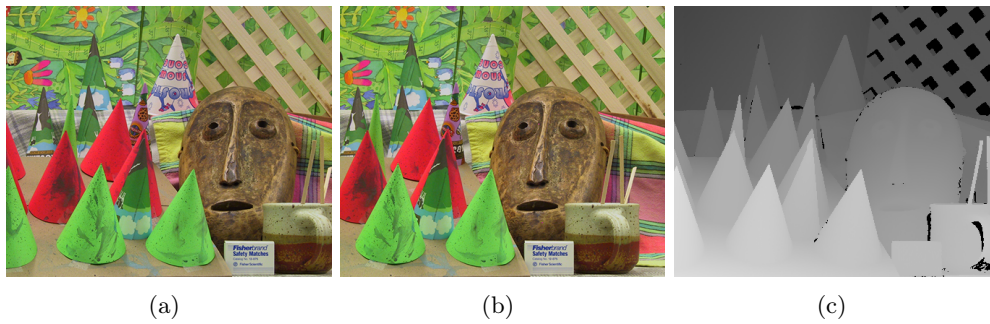


Figure 2.27: Cones stereo pair: (a) left image, (b) right image, and (c) disparity map.

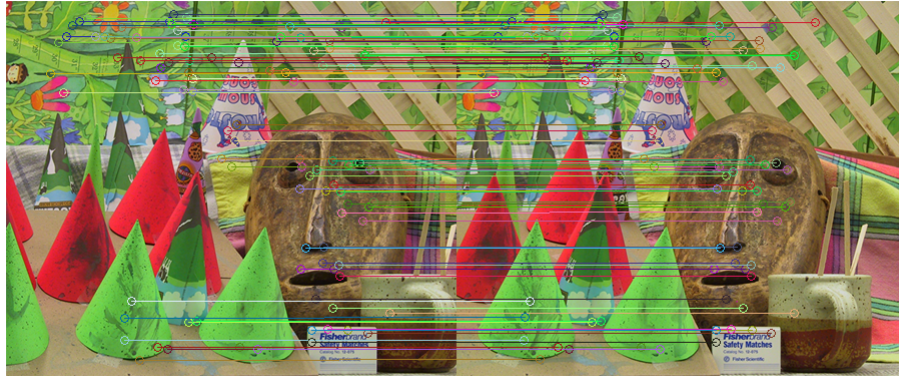


Figure 2.28: An example of features correspondence.

Traditionally, two approaches, block- and pixel-based, can be used to estimate the disparity field. In the former class, it is assumed that the disparity is block-wise constant. However, this assumption does not always hold, especially around object discontinuities. Furthermore, using the block-matching technique yields inaccurate results in textureless regions. Pixel-based approaches attempt to overcome this drawback by assigning a disparity value to each pixel, thus producing a dense disparity map [84].

In turn, pixel-based algorithms are classified into two categories: local and global methods. In local methods, the disparity at each pixel only depends on the intensity values within a local window. However, global methods find the disparity map by minimizing a global energy function over the whole image.

Local methods are based on comparing the similarity between two sets of pixels. According to the type of the matching sets, these methods can be mainly classified into two categories: feature- and area-based approaches. For feature correspondence methods, before matching, a set of features is detected and their invariant descriptors are extracted from both views. Then, a correspondence step between the extracted features is established. Over the past years, a number of local invariant features have been proposed in the literature, such as Scale Invariant Feature Transform (SIFT) [87], Speeded Up Robust Features (SURF) [88], Maximally Stable Extremal Regions (MSER) [89], Harris-Laplace [90], Principal Component Analysis (PCA)-SIFT [91] and SIFT-Flow [92], etc. These feature detection methods have been evaluated and compared [93, 94]. They have the advantage of being less sensitive to noise and photometric distortions, rotation, scaling, and so on. Moreover, compared to pixel intensities-based description, these methods do not provide a dense correspondence map but extracts accurate and sparse correspondences. Figure 2.28 illustrates an example of features correspondence between two images.

In contrast to feature-based methods, area-based ones generate a dense disparity map, but they are more sensitive to locally ambiguous regions like the occlusion and

the untextured ones. This class of methods, where pixels or regions are matched to find the corresponding pixels, is also known as *correlation-based methods*. The basic idea of this method consists firstly of defining a window around the pixel to be matched, and then comparing it with the window surrounding each possible candidate corresponding pixel located in the other image.

A major difficulty of the area-based methods are that they fail at object edges and boundaries. Furthermore, using small windows in untextured regions do not capture enough intensity variations to make reliable the correspondence task, whereas the use of large windows tends to blur the boundaries of disparity map.

2.6 Conclusion

3D contents have to be created and store. This brings the issue of 3D scene representations. We have reviewed various types of 3D video representation techniques that are envisioned to enable advanced 3D video applications requiring multiple viewpoints of a scene. The choice for the appropriate 3D scene representation widely depends on the expected case of use.

Finally, since advanced 3D video applications require multiple viewpoints simultaneously, efficient compression methods are of paramount importance for streaming this amount of data, while preserving the QoE. In response, H.264/MPEG-4 AVC has been amended to support multi-view video coding. To achieve coding gain, the latter is extended by exploiting, in addition to the temporal and spatial prediction, the inter-view prediction. This is achieved by taking advantage of the redundancies between the views.

To achieve this inter-view prediction, the disparity information is needed. At the end of this chapter, we introduced the basic concept of stereo matching or disparity estimation used in the following chapters.

Chapter 3

Color Correction of Multi-View Video

Contents

3.1	Introduction	45
3.2	Color compensation methods for multi-view images/videos	47
3.2.1	Calibration of multi-camera setup	48
3.2.2	Embedded illumination/color compensation	50
3.2.3	Preprocessed color correction	53
3.3	Proposed color correction method	58
3.3.1	Feature correspondence	58
3.3.2	Reference view selection	60
3.3.3	Cumulative histogram calculation	62
3.3.4	Color mapping function	64
3.4	Experimental results	65
3.4.1	Effect on the visual color consistency	65
3.4.2	Effect on coding performance	69
3.4.3	Effect on rendering performance	73
3.5	Conclusion	77

3.1 Introduction

Advanced 3D video applications such as multi-view autostereoscopic displays and FVV, requires a very large number of views. This set of views is obtained by capturing the same scene with multiple synchronized cameras from different view-points. However, these multi-view acquisition systems have an inherent problem of



Figure 3.1: An example of color inconsistencies between views.

color/illumination difference between the different views. There are several potential sources that contribute to these color discrepancies between views. For example, camera orientation, focal length, and manufacturing may be inconsistent, even between cameras of the same model with the same settings. Also, non-uniform camera parameters or editing errors could create color/luminance mismatches. Another important factor is linked to calibration differences between cameras. Moreover, even if all cameras are perfectly calibrated, it is often difficult to capture an object under perfectly constant lighting conditions. The reflectance properties over the object surface such as: scene geometry, angles between cameras and objects, and light incidence, should also be considered. All these factors negatively impact color consistency, leading to a possible difference of appearance of the same object on the various views, as illustrated by Figure 3.1.

These color inconsistencies negatively affect the multi-view video (MVV) processing framework. For instance, they reduce the correlation existing between views, and provoke a significant decrease in compression efficiency. Consequently, it causes performance reduction in multi-view video coding (MVC) that exploits the inter-view statistical dependencies. Furthermore, when rendering new virtual views and switching between the different views, observer may perceive color variations across them. This will appear unnatural and will affect his quality of experience (QoE) [100]. Add to that, if these color differences exceed a given threshold, the fusion difficulty and binocular rivalry can occur inducing visual fatigue and discomfort [101–104].

In addition, when the fused regions are bounded by contours of opposite contrast, this can be perceived with variable brightness, as *binocular luster* effect described in the psychology literature [158]. An illustration demonstrating this effect is shown in Figure 3.2. In this example, the smaller square is lighter than the background in both of the views and no rivalry occurs. However, the larger squares have different contrast polarities in each view, which causes it to appear as if it is shining [158]. These mismatches are immediately noticeable by the viewer, making the luster artifact extremely salient. Therefore, an important problem is how to compensate



Figure 3.2: An illustration of binocular luster.

color differences between different views of MVV to enhance the performance of 3D systems [105].

To cope with such important problems, various approaches were proposed to compensate for the illumination and color changes between views. To achieve this, some of them proposed to fine-tune shooting conditions, such as camera arrangements and parameters, and other methods proposed to customize the codecs in order to compensate the color mismatches. In a completely different way, other approaches performed color correction by preprocessing the captured views to make them color consistent and harmonized.

The rest of the chapter is organized as follows. Section 3.2 describes the state-of-the-art covering the most important strategies for color compensation applied to stereo and/or multi-view images/videos. We decomposed the existing methods in three categories depending on their objectives and behaviors. The first category addresses the methods applied on the capturing stage aiming at significantly reducing color discrepancies between the calibration of the used cameras. The second category describes methods embedded in the core coding and proposing strategies to apply illumination/color compensation. Finally, the third category can be seen as a preprocessing applied to captured scene before being fed to coding or rendering steps. Following this review, the detailed steps of the proposed approach are described in Section 3.3. Section 3.4 presents some simulation results with the proposed technique for different test sequences, including visual comparisons, coding performance, and rendering quality. Finally, Section 3.5 summarizes the chapter with some thoughts for future work.

3.2 Color compensation methods for multi-view images/videos

As mentioned previously, several methods have been proposed in the literature to resolve the problem of color inconsistency in MVV. This section is dedicated to their description with a classification into three categories: **1)** calibration of multi-camera setup, **2)** color compensation in the core coding process, and **3)** correction applied as a preprocessing stage before compression.



Figure 3.3: An example of Macbeth color checker.

3.2.1 Calibration of multi-camera setup

Over the last few decades, the use of multi-camera systems has grown exponentially. Different applications resort to such acquisition systems, for example 3DTV, FVV, and surveillance, to cite a few examples. However, inherent problems with this type of acquisition, are geometric errors and inconsistent color responses. Many methods have been developed to geometrically calibrate camera arrays [106–109], but much less attention is paid to color calibration of cameras, relying only on manual adjustment.

To calibrate color responses of cameras, Nanda *et al.* [110] proposed a method using scene statistics. In this method any camera can be chosen as the reference. The latter is color calibrated by adjusting its brightness and white balance. The brightness adjustment is performed by acquiring an absolutely black scene at zero exposure, and then by making the mean intensity of the acquired image reach a desired black value. In turn, the white balance is calibrated by adjusting two parameters, specifically red and blue gain values, to make the amount of green, blue and red equal in the scene. Once the reference camera has been color calibrated, to ensure color uniformity between cameras, the settings of the neighboring cameras are adjusted to make the colors of the overlapping regions across cameras similar. This method produces varying results depending on the part of the scene acquired by the camera.

Another color calibration way is to use a Macbeth color checker (see Figure 3.3). For example, in [111, 112], under static lighting conditions an acquisition of a Macbeth color checker is made with the whole set of cameras. Then, color patches of the Macbeth color checker are automatically detected in the captured images of all cameras. Next, the gain and offset settings of each camera response are iteratively adjusted on each channel to fit a line through the RGB values recorded from the six gray patches of the Macbeth color checker. Following this step, some errors in the gain and offset settings, non-linearity in the sensors, and color distortions remain. To address that, the authors proposed a three-stage post-processing. The first one is to correct the residual errors in gain and offset by computing them once again in a floating-point domain, allowing more precision and accuracy. The next stage

addresses the non-linearity in the sensor. A look-up table that re-maps the response curve of each channel in each camera to the target color is calculated. Finally, the third stage is to correct global color errors by computing a 3×3 RGB to RGB color transform that minimizes the error between each recorded color values from the patches and the average values of the patches recorded by the set of cameras. Inspired by this work, Ilie *et al.* [113] proposed a color calibration method for a multi-camera system. This method is implemented in two stages. Similar to [112], once all the cameras have taken an image of the Macbeth color checker, one of these acquired images is considered as the reference image. The first stage consists of adjusting each camera's parameter settings by minimizing the color difference between the reference image and images of the Macbeth color checker acquired by all cameras. This minimization is iteratively performed using a modified Powell's method. Starting from the obtained initial cameras' parameters, the same optimization process for all cameras is repeated, but by comparing with a new reference image calculated as the average of all cameras images computed in the previous step. In the second step of the method, a post-processing refinement is performed. The refinement is applied to enhance the cameras' parameters obtained from the first stage. However, this method requires a constant lighting condition, which cannot be guaranteed in a real practical system. Also, they are focused only on planar acquisition systems.

Li *et al.* [114] formulated the multi-camera color calibration problem as an overdetermined linear system, in which each camera collaborates with other cameras on color calibration. The authors begin by establishing global correspondences between all views. Based on these global correspondences, the camera sensor response can be approximated by a linear model with a multiplicative gain and an additive offset. The next step is to find proper gain and offset values so that the color values of the global correspondences will be the same in all camera images, and consequently a consistent color response of the cameras will be obtained. Add to that, in order to ensure proper dynamic range, the authors incorporated a dynamic range shaping process into the described linear system. Slavkovikj in [115] first detects and captures a white target to generate calibration data. This is done by capturing a white sphere of diffuse reflectance placed in the scene and by using the cameras' response model described in [116]. Thanks to this step, the gain parameters of the color channels are instantiated for each of the cameras. Then, the author acquired images of a scene containing colored objects. To take into account the non-uniformity of the cameras' sensors, the acquired images are corrected as a matrix containing per-pixel scaling factors. Following that, the cameras' images are white balanced, and the gain and exposure are adjusted according to one reference camera. After applying this initial correction, correspondence points are calculated between adjacent camera pair images. The latter are used to form corresponding color pairs and to derive color correction matrices by performing nonlinear multiple regression. Finally, the correction is performed by using the obtained color correction matrices.

Zhong *et al.* [117] proposed a method to control each camera in a multi-camera system with the aim to enhance video quality. They proposed to address both inter-

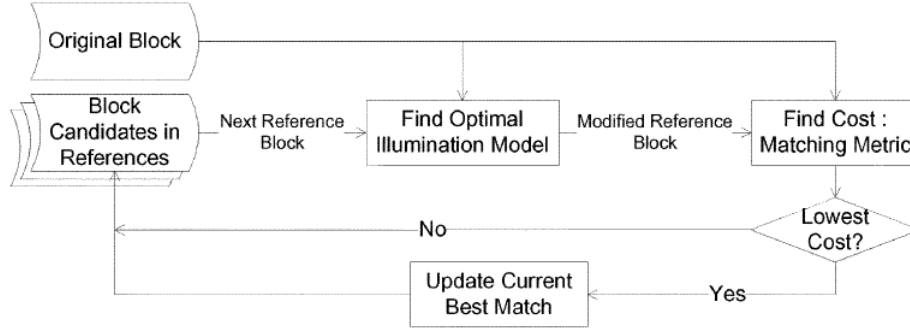


Figure 3.4: Modified search loop for the current block [118].

camera color inconsistency and object location difference problems. To achieve this objective, each camera is individually and automatically controlled by adjusting its pan angle and focal length. They formulated the problem as a nonlinear optimization problem, and then find the optimal camera control of pan angle and focal length. The authors first estimate initial values for the camera control, and then perform an iterative optimization by the bivariate Newton-Raphson method to get the optimal ones. Unlike the previous described methods, this one has been evaluated for coding purposes. To test the performance of their method, the authors have taken three sequences with and without their camera control method. The obtained MVV sequences have been compressed using MVC. The achieved coding gains range from 0.5 to 1.9 dB.

Despite the improved color consistency among camera views, this class of methods does not come with performance guarantees, because of the impossibility to control the reflectance properties over the objects surface, which is dependent on the capturing angle of each camera, lighting conditions, and so on. As a result, color camera responses are not well balanced. Consequently, several color correction approaches based on image processing techniques were proposed to compensate for illumination and color changes between views. This is the subject of the next sections.

3.2.2 Embedded illumination/color compensation

When illumination/color changes occur between adjacent views, disparity estimation exploited by disparity-compensated prediction (DCP) in the MVC cannot be accurately performed. Therefore, the residual energy will be higher and, consequently, the coding efficiency will decrease. One way to address this problem is to tackle it during the coding process, and specifically, as part of the inter-view prediction process on a block level. To achieve that, several approaches used a modified block matching measure.

For instance, block-based illumination compensation (IC) technique has been introduced in [118, 119]. Based on an idea outlined by Kamikura *et al.* in [120], the authors proposed a modified matching metric that incorporates an IC model (see

Figure 3.4). They modeled illumination mismatches as an additive offset term and a multiplicative scale factor. Before matching, the candidate reference block and the current block to be matched are decomposed into a mean and a mean-removed signal. Next, the IC parameters, namely the scale and the offset, are applied to the reference block. Thus, for each candidate block, optimal IC parameters are computed. This process is repeated until finding the best matching block in the reference frame.

In order to transmit IC parameters to the decoder, the authors predicted IC values from neighboring blocks, and encoded them using the context adaptive binary arithmetic coder (CABAC) [98]. Finally, decision on whether or not to use the IC model for each block is based on rate-distortion (R-D) criteria. According to simulation results presented in [118], a coding improvement up to 0.8 dB is obtained with this approach.

In the same vein, an adaptive local illumination change compensation method has been presented in [121, 122]. Figure 3.5 shows the encoding and decoding structure of this method. The compensation is performed at the macroblock (MB) level. The authors assume that the discrepancy caused by the illumination change is equal to the discrepancy of the DC component between the current and the reference blocks. Accordingly, the difference of DC values of the current coded MB and the reference one is subtracted during the motion and disparity estimation/compensation. This difference, referred to as the illumination change value (DIVC), which is in fact an offset factor is used to cancel out the illumination change between MBs. As the method described above, DIVC needs to be encoded and signaled to the decoder for each MB in the bitstream. Each DIVC value is predicted from the neighbors of the current MB, and entropy-encoded using CABAC. Also, for each MB, the R-D control is used to decide whether or not the IC process is needed. It is reported by the authors that gains of up to 0.6 dB can be achieved. This approach has been adopted by the JVT standardization committee in the joint multi-view video model (JMVM) software [99], but not in the standard itself [96].

In [123], the authors proposed to extend the H.264/MPEG-4 AVC weighted prediction [124] for illumination compensation in MVC. They proposed to calculate the offset and weight parameters of the weighed prediction for each MB. Firstly, the slice weight value is determined. Thereafter, the best MB partition mode and motion vectors are set based on the RD-optimization using the normalized sum of absolute differences (NSAD). In the next step, they built the predicted MB with the motion compensation. Then, weight and offset parameters are calculated by minimizing a distance between the current MB and the predicted one. Consequently, the slice header contains slice weight values and each MB header contains weight and offset to be mandatorily sent to the decoder.

Sohn *et al.* [125] proposed an inter-view balanced disparity estimation (IBDE) method. As a first step, the authors calculate the global balancing parameters for a whole frame with mean and standard deviation. Using these initial parameters and their probabilities, they derive the local balancing parameters (scale and offset) for each MB. This is achieved by using the 2D Lloyd-max algorithm. The reported cod-

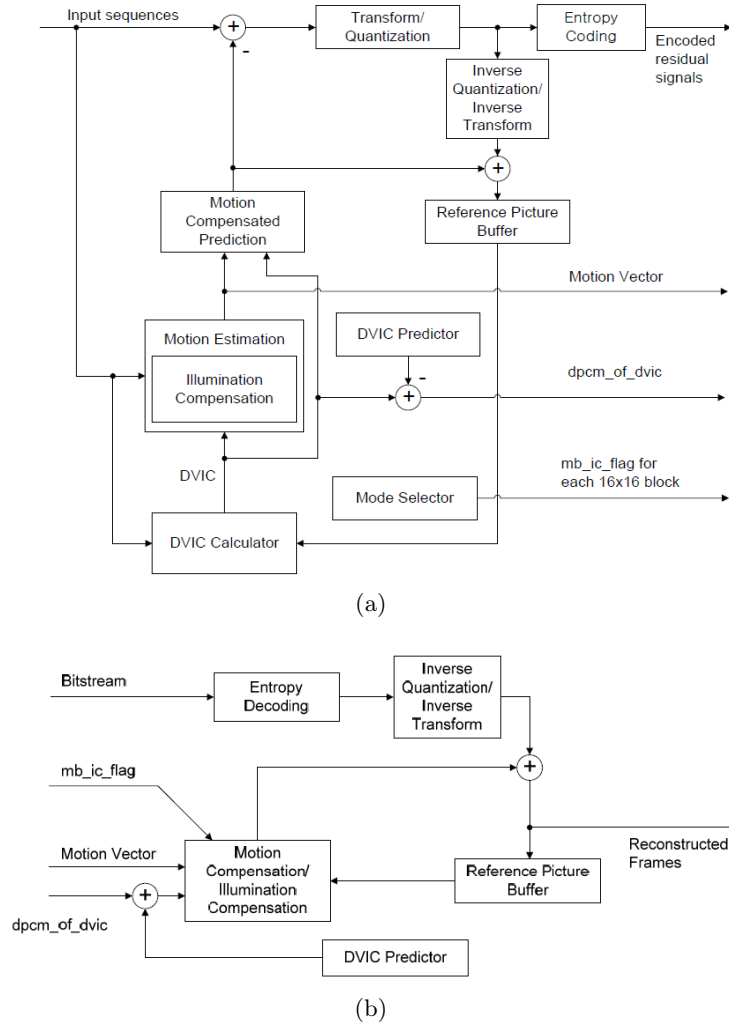


Figure 3.5: Block diagram of the MB-based adaptive illumination change compensation method [121]: (a) Encoder, (b) Decoder.

ing gain is about 0.1~0.3 dB. Li *et al.* [126] proposed an illumination compensation algorithm based on what they call a template, *i.e.*, the reconstructed neighboring region (left and above) of the current block. Their idea lies in the fact that the illumination of the current block has a strong correlation with its adjacent template. Similarly, the reference block and its corresponding template also have this correlation. Therefore, they derived the illumination change between the current block and the reference one from the change between their templates. The key advantage of this approach is that the local illumination change is calculated by the template and compensated without transmitting additional bits.

Another method proposed by Yamamoto *et al.* [127] applies color compensation in the decoded picture buffer (DPB). They performed color correction of each frame being used for prediction, *i.e.*, decoded pictures in DPB. Firstly, each frame is con-

verted from YUV to the RGB color space. After that, they calculate look-up tables of each color channel from RGB. The latter are created first by finding correspondence points between pairs of views. Subsequently, the found correspondence points are used to create corresponding-colors. However, to be robust to possible noise in the frames, the authors produced some Gaussian-filtered frames from input frames and detected corresponding-colors from them. Once the corresponding-colors are defined, they formulated the color-mapping relationships as an energy-optimization problem and resolve it by dynamic programming to obtain look-up tables. Color correction is applied on both encoder and decoder sides, which implies sending look-up tables in the codestream or as side information.

The previous methods are integrated into the H.264/MPEG-4 AVC, for evaluation purposes, and the coding gains were notable. However, compensating the illumination/color changes during the coding process has some disadvantages. It involves a new syntax and requires design changes affecting the low levels of the encoding and decoding processes. In addition, since the compensation is forced to be carried out in both encoder and decoder, the complexity is increased. Also, an overhead is needed to send the IC parameters to the decoder. All the above described methods, except for [127], only deal with variations in brightness, and they do not compensate for color variations. Because they are only applied on the luminance component. Finally, these methods do not perform color correction of the input sequence, but they only enhance coding performance with the aim of R-D optimization. Thus, the problem of inconsistencies in luminance and color still persists after decoding, which will unfortunately introduce errors on DIBR for the creation of new virtual views.

3.2.3 Preprocessed color correction

As an alternative to embedded illumination/color compensation, appropriate preprocessing consisting in color correction methods can be applied in an independent manner before feeding the view sequences to the encoder. In essence, the color correction operation on MVV can be seen as a process transferring the color palette of a source view (or reference view) to a target view. The source view is chosen by the user from the set of views, and all the remaining views are corrected to acquire the color palette of that reference. Following this idea, several tentative approaches have been proposed [128–131, 133–139]. The key advantage of such approaches is that no changes are necessary to either the encoder/decoder nor the bitstream syntax. In the following, we try to present a comprehensive overview of these methods with the emphasis on those that are applied to MVC.

A straightforward solution for correcting color in MVV as a preprocessing step is the histogram matching (HM) method. HM has already demonstrated its efficiency to perform image color correction [140]. Fecker *et al.* [128] proposed an extension of the HM method to MVV. In their method, all camera views of a multi-view sequence are corrected toward a common reference view in the center of the camera setup. To achieve this, the authors calculate look-up tables using cumulative histograms of

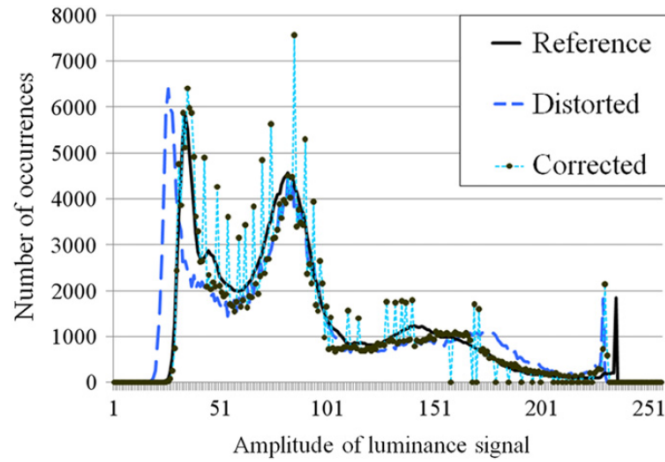


Figure 3.6: An example for the resulting histogram of the corrected image compared with the reference image and the distorted image.

the view being corrected and the reference. They adapt the cumulative histogram of each view to match that of the reference. To maintain the temporal correlation, and avoid the introduction of color variation between successive frames within a single view, the authors proposed a time-constant mapping function. They compute the histogram for the whole sequence; it is summed up over all frames of the sequence. So, the same correction is used for every temporal frame in the sequence. An example of the result from luminance compensation by the HM is shown in Figure 3.6. It can be seen that the histogram of the corrected image is closer to the reference histogram than to the histogram of the distorted image.

However, for the cases when the camera setup is not fixed and in real-time applications, this solution is not sound. The obtained gain in coding efficiency is up to 0.7 dB for luminance and 1.9 dB for chrominance. Similarly, Chen *et al.* [129] proposed an approach based on HM, where two parameters, namely scaling and offset, are derived optimally for each YUV channel based on the histograms of the chosen reference view and the view being corrected.

Both approaches have common drawbacks linked to HM. Since the MVV sequences do not show exactly the same content, especially in a sparse camera system (*i.e.*, widely separated views), frames from different viewpoints contain non-overlapping regions (occluded regions). Therefore, applying HM to the entire frame will produce inaccurate and inconsistent results.

Another approach is presented by Pitié *et al.* [130, 131], where they consider the colors of an image as an N -dimensional probability density function (PDF), and repeatedly project these N -dimensional distribution into a series of random 1-dimensional marginal PDFs. In order to map the PDF of the reference image, the color distribution of the target image is iteratively manipulated until matching that of the reference. To avoid the increase in graininess over the corrected image, the authors proposed in [131] a post-processing algorithm to preserve its gradient

field. This method has been extended to MVV by using only the first frame of each view to estimate the color transformation. Then, this fixed transformation is applied to the remaining frames of the view [132]. This method provides reasonable performance, but if the reference and the target images have different color statistics in the non-overlapping regions, it can produce poor results.

In order to avoid occlusion problems, several recent approaches used correspondences between views. For instance, Lee *et al.* [133] proposed to segment the image according to depth, and perform HM for each sub-image. The image segmentation is performed by using an expectation maximization clustering method based on a Gaussian mixture model (GMM). However, this method depends heavily on the quality of depth/disparity information, for which unfortunately the performance of disparity estimation methods are still error prone. Also, disparity estimation requires knowledge about the range of disparities for all views, which is a difficult task as it depends on the scene and camera geometry, and will have a significant effect on the performance of color correction. In [134], a histogram-offset based color correction method is proposed. As a first step, the authors identify maximum matching regions between the reference view and the target one. To achieve that, disparity estimation is performed on the rank-transformed domain, thus allowing to handle occlusion regions. Then, the histograms of the reference view and the target view are calculated for each YUV channel only within the identified matching regions. According to an iterative threshold, the calculated histograms are divided into high and low occurrences bin classes, and only the histogram bins having high occurrences are calibrated. Finally, the generated histogram offset is exploited to correct the target view. However, this approach considers only the global color discrepancies between views, since it computes only a rough color mapping between views, which does not produce a locally-refined result.

Another method is presented in [135], where the authors proposed to find the average color of the set of views and modify all views to match this average. Hence, in this approach, the average color is considered as the color reference, and was calculated by finding matching points between all views using block-based disparity estimation. The authors expressed each of the corrected YUV channels as a weighted linear sum of the three original ones with an offset. Then, they performed a least-squares regression on the matching points to find optimal weighting parameters leading to a good fit between the average YUV values and those of the view being corrected. With this method, the compression efficiency is increased by up to 1 dB in luminance component, and up to 2.1 dB in chrominance component. Unfortunately, choosing the average color as the color reference cannot guarantee a good quality, because, if some views are distorted or have a significant color shift, this will strongly influence the average color. In addition, it is difficult to compute the average, especially when we have large distances between the different views. Another limitation lies in the fact that only ten successive frames are used from the sequence in order to calculate these correction parameters. The extracted set of parameters is used to correct every frame in a given view, under the assumption that correction parameters are time-constant during the whole sequence. However, this assumption

does not always hold, especially when the acquiring system of MVV is not fixed. In addition, assigning a single disparity vector for 8×8 pixels blocks, as described in their approach, can be the cause of a number of bad matches. Finally, disparity estimation and least-squares regression can lead to a significant increase of computational cost, which cannot be neglected, even though the process is applied offline. A less complex preprocessing method was proposed in [136]. In order to achieve fast color correction, instead of performing correction in a frame-by-frame basis, the authors divide the set of frames of each view into keyframes and non-keyframes, and adopt different color correction strategies for each of them. The keyframes are corrected directly by the same color correction method proposed by [135]. While for non-keyframes, the color corrected values are derived from previous frames by using a temporal variation model. Basically, the latter consists of two steps, time-invariant detection and temporal variation modeling. The purpose of the time-invariant detection step is to extract the background regions that are unchanged over time. These regions are used for extracting stable temporal variation information from a single view. Once these regions are defined, the color discrepancy relationship between adjacent frames is derived under the assumption that color distributions in background regions are Gaussians. The interval between keyframes is fixed and equals the size of the group of pictures (GOP) in MVC. This approach has two main advantages. Firstly, by exploiting the temporal variation model, computational complexity is significantly decreased. Secondly, as the color corrected values are derived from previous frames, color variations between successive frames can be avoided.

Jung *et al.* [137] proposed to improve the color consistency by modeling the characteristics of the cameras and estimating their coefficients. To consider occlusion regions, first, the authors extract correspondences between pairs of views. After that, based on the calculated correspondences, they estimate the coefficients (gain, gamma, and offset) of the camera characteristic curve via an outlier-removed nonlinear regression based on the Levenberg-Marquardt algorithm. With these estimated coefficients, look-up tables are generated and followed by color correction. The coding improvement made by this method is of up to 0.9 db and 0.8 db for luminance and chrominance components, respectively. The main disadvantage of this approach is linked to the fact that it may be used only when the setup is fixed, because the authors consider fixed coefficients for each camera. The same authors proposed a color correction algorithm using adaptive weights of correspondences [138]. They exploited two types of correspondences: global and local. Global correspondences are common features on all views. While, local correspondences are matched features for a certain stereo pair. In order to increase correspondences, an additional matching on Gaussian-filtered images was applied. The authors consider that the change between the reference and target views is nonlinear for the Y component and, linear for U and V components. Accordingly, to compensate for relative color differences between views, they used Gain-Offset-Gamma (GOG) and linear models for luminance and chrominance components, respectively. The coefficients of these linear and nonlinear models were estimated from correspondences. In addition, an

adaptive weighting factor is included into these two models to weight correspondences according to their locations.

Tehrani *et al.* [139] proposed a similar method to the one developed by Yamamoto *et al.* [127], with some important differences. Instead of performing the correction in the core coding process, it is achieved as preprocessing step. Also, there is no unique reference view for which all target views are corrected. The algorithm can start from any camera view from the set of views and continues by shifting the pair sequentially on the multi-camera system for several iterations, until the average color change between all views becomes small enough. This allows its use for any camera configuration, from sparse to dense multi-camera systems. Another difference lies in the increase of correction accuracy by enhancing the matching step. This includes appending corresponding intensities from several time steps, suppressing the outliers, and by nonlinearly weighting the model.

The main advantage of these approaches, is that no changes are necessary neither to the encoder/decoder, nor to bitstream syntax. Nonetheless, most of these methods are not appropriate when the camera setup is not fixed or in real-time applications, since they use only a few frames to calculate the correction parameters, or in some approaches consider all frames but the same correction parameters are used for every successive frame in the sequence. Moreover, the selection of the reference view (*i.e.*, the one considered as the color reference and used to correct mismatched views) is often assumed in the center. However, in case the center view is itself distorted, this may lead to the impairment of the whole sequences of MVV. From the previously discussed approaches, only a couple [135, 139] does not consider the use of the center view.

With the aim of overcoming the aforementioned drawbacks, we proposed a preprocessing method based on histogram matching [141, 142]. However, in contrast to methods described above, we define a customized version of the HM algorithm. The proposed method relies on taking into account only the common features areas among all views when calculating the histogram, thereby avoiding the problem of occlusions and allowing a more precise color correction. To effectively define this set of common areas, we use the Scale Invariant Feature Transform (SIFT) [87], followed by the Random Sample Consensus (RANSAC) [95] to remove matching outliers. Thereafter, we calculate the cumulative histogram of each view based only on corresponding areas already detected. Through our color correction method, we try to improve the inter-view prediction, while maintaining the temporal correlation that exists between successive frames within a single view. In order to keep this temporal correlation, the HM method is performed in a sliding-window, where each color mapping function is defined using a group of pictures, and this may also ensure an adaptation to time-varying acquisition system and real-time broadcasting. Another aspect tackled by the proposed approach corresponds to reference view selection. Indeed, this step ignored by many works, may highly influence the correction results. We propose to perform it in a fully automatic manner, by defining a robust criterion relying on views statistics and structural content. Details about the proposed approach are given in the following section.

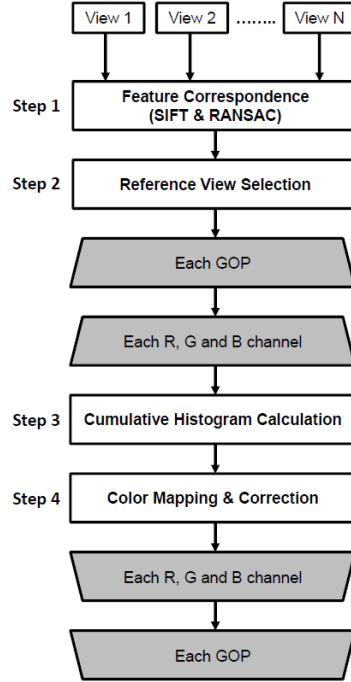


Figure 3.7: Flowchart of the proposed method. The shaded pairs of trapezoids represents loops.

3.3 Proposed color correction method

This section is dedicated to the description of the proposed color correction method for MVV [141, 142]. Firstly, a particular view is chosen to be the reference. All remaining views are corrected to make them similar to this reference as closely as possible. To accomplish this goal, our proposed method is performed in four steps, summarized in Figure 3.7. We operate the correction in the RGB color space, where the red, green and blue color channels are corrected individually and non-linearly; this choice will be justified later.

3.3.1 Feature correspondence

Before selecting the reference view and performing the correction, an important step in the proposed method consists in the extraction and matching the features that are present in all views. Here, the challenge is, to detect these shared features in the different views despite the variation of colors and geometric deformations usually due to the scene and cameras arrangement. For this purpose, there exist two main approaches: the first is based on disparity estimation [85, 86], which in our case is not applicable because it usually relies on color intensity as the matching criterion and may produce unreliable matches that will negatively affect the rest of the process. The second approach resolving reliably the correspondence problem is a feature

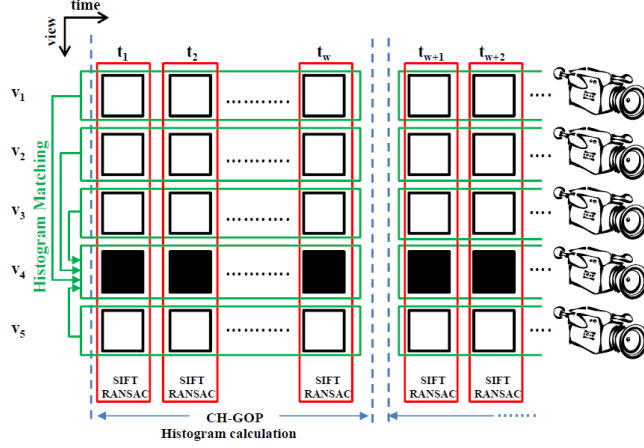


Figure 3.8: Overview of the proposed approach. The black squares represent the defined reference view.

matching algorithm. These features are robust to different changes (illumination, image noise, blur, etc.), and to a wide range of 3D transformations.

Over the past years, a number of local invariant features have been proposed in the literature, such as Scale Invariant Feature Transform (SIFT) [87], Speeded Up Robust Features (SURF) [88], Maximally Stable Extremal Regions (MSER) [89], Harris-Laplace [90], Principal Component Analysis (PCA)-SIFT [91] and SIFT-Flow [92], etc. These feature detection methods have been evaluated and compared [93,94]. Each method has its own advantages and drawbacks, and the choice mainly depends on the targeted application. However, it is important to note that SIFT and SURF outperform the other methods in terms of robustness and stability in most situations. Therefore, we chose the SIFT method for the detection and matching of feature points between the different views. SIFT detects the feature points, and extracts the invariant feature descriptors allowing to generate large numbers of matching features points corresponding to the same parts of the scene, while being robust to different changes of illumination, image noise, rotation, scaling, and so on [87].

In our method, SIFT is applied at each time step, between views' frames from the same time step but from different camera views, as illustrated by red rectangles in Figure 3.8.

Following the SIFT algorithm, the robust RANSAC algorithm is applied for outliers removal. Based on the epipolar constraint and the inliers of feature points, RANSAC algorithm estimates the fundamental matrix and makes use of it to eliminate the outliers in correspondences [95].

It is known that RANSAC does not perform well in case of a high number of outliers. In our case, this process is simplified by an implicit removal during the correspondence stage. To achieve this, the first set of SIFT keypoints obtained between frames of the two first matched views are used as the initial set of all possible points in the remaining frames. For each frame of a new view, the former

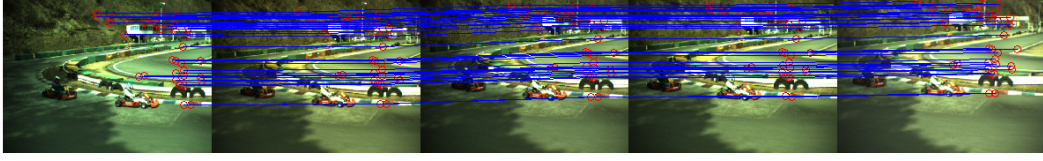


Figure 3.9: Common SIFT keypoints across five views of the *Race1* sequence. Only 15% of correspondences are shown for clarity.

set is regenerated by removing all points unreachable on it. At the end of the process, each SIFT keypoint will have correspondents in all views, and the matches are very reliable. Finally, the set of common SIFT keypoints is used as input of the RANSAC algorithm. After performing extensive tests, we noticed that the rate of outliers removed by RANSAC ranges from 2% to 10%.

Once the SIFT-RANSAC algorithms are applied, each frame f_{ij} will have a set of SIFT keypoints $K_{ij} = \{k_{j1}, k_{j2}, \dots, k_{jP}\}$ determined and matched, where i and j denote the view number and the temporal order, respectively, and P is the cardinality of K_{ij} . An example of common SIFT keypoints among five views of the *Race1* sequence is shown in Figure 3.9. It represents only part of them (15%) as well as only five views among eight for the sake of clarity.

The defined SIFT keypoints are used by the following steps to carry out the process only on the common regions among all views.

3.3.2 Reference view selection

The effective selection of the color reference is a key element of the color correction process. Since this choice influences the final correction and thus the coding performance and quality of rendered views. As stated before, all previous works, except [135, 139], choose the central view of the camera arrangement as the reference view. While this choice may appear natural and consistent, because the view shares the most parts of content with other views, the central view may be distorted thus leading to the impairment of the whole sequences of MVV. Additionally, in some cases there is no explicit central view, as for example with an even number of views. Furthermore, when using camera arrangements such as dome type, circular, and array, it is not obvious to choose a central view. In order to overcome that, in [135] they propose to find the average color of the set of views and modify all views to match the average. This solution cannot guarantee a good quality, as if some views are distorted or have a significant color shift, this will strongly influence the average. In addition, it is difficult to compute the average, especially when we have large distances between the different views. Another solution has been proposed in [139], which consists to choose any view from the set and continues by shifting the pair sequentially on the multi-camera system for several iterations, until the average color change between all views becomes small enough. However, this arbitrary choice of color reference depends heavily on the first corrected stereo pair.

Consequently, in order to select an appropriate reference view, we consider two

factors, the content of the view (S) and its visual quality (Q). We compare each view with the other views based on these two factors. For the first factor, the selected view is the one having the most similar content with the others; this can be evaluated by using the histogram differences across views. The similarity between histograms can be assessed with any technique from literature. In our case, we opted for the most widely used Bhattacharyya coefficient [143, 144], defined by

$$SH(H, H') = \sum_{b=1}^B p_n(n_b, n'_b),$$

$$\text{with } p_n(n_b, n'_b) = \frac{\sqrt{n_b n'_b}}{\sqrt{(\sum_{j=1}^B n_j)(\sum_{j=1}^B n'_j)}} \quad (3.1)$$

where H and H' are two histograms, and n_b is the number of pixels whose value is that of the b^{th} bin. The number B is the total number of bins. In our case, we set B to 256, which is the total number of levels for representing each color component.

For the second factor, the visual quality of the view, we assume that the view being compared is the reference view and the remaining views are the distorted ones. Traditional image quality/fidelity metrics as Peak Signal-to-Noise Ratio (PSNR) or Structural SIMilarity (SSIM) [145] fail in this case, because they are based on pixel-wise errors, and are not adapted to compare images with a different content even with a shift of one pixel. To overcome this problem, we propose to measure the structural fidelity only on common features by focusing on interest points to avoid comparing irrelevant content. This makes the quality measurement more reliable.

To compare two views (V_j and V'_j), we use the set of SIFT keypoints already calculated, and apply a customized SSIM metric as follows:

$$QS(V_j, V'_j) = \frac{1}{M} \sum_{i=1}^M SSIM(W_i, W'_i) \quad (3.2)$$

where W_j and W'_j are 11×11 windows from both being compared views (V_j and V'_j) centered on the SIFT keypoint i , and M being the number of detected and matched SIFT keypoints. SSIM metric is commonly used 2D image quality metric providing a quality score ranging between 0 and 1 (please refer to the definition of the metric in Section 5.2.2).

Finally, for each view V_j , a score C_{V_j} is calculated as follows:

$$C_{V_j} = \alpha \times S_{V_j} + (1 - \alpha) \times Q_{V_j} \quad (3.3)$$

where S_{V_j} and Q_{V_j} represent respectively the mean similarity histogram and the mean SSIM value of the view V_j over $N-1$ views

$$S_{V_j} = \frac{1}{N-1} \sum_{i=1}^{N-1} SH(V_j, V_i) \quad (3.4)$$

$$Q_{V_j} = \frac{1}{N-1} \sum_{i=1}^{N-1} QS(V_j, V_i) \quad (3.5)$$

with SH and QS are defined in (3.1) and (3.2), respectively.

The parameter α controls the relative importance of the two terms (S_{V_j} and Q_{V_j}). In our experiments, it has been set to 0.5, but if the camera views are widely separated, it would be more consistent to assign a higher value. Thus, the view V_j having the maximum score C_{V_j} is chosen as the reference view V_{ref}

$$V_{ref} = \{V_j / \arg \max_j (C_{V_j})\}. \quad (3.6)$$

This step can be seen as optional, and can be deactivated without changing the whole structure of the method.

3.3.3 Cumulative histogram calculation

In the proposed approach, one important difference with similar literature is the fact of selecting only common regions in the different views in order to perform color correction in the most accurate way. Therefore, to define these common regions we use the SIFT keypoints already detected. We have chosen the SIFT method not only for its robustness but also for its correlation/similarity with gaze points. In recent works [146–148], it has been shown that SIFT keypoints are often located in the most visually salient regions of an image. So, the SIFT keypoint and their neighboring pixels can be used as predictors of salient regions, and contribute in mimicking the human visual system. Therefore, by performing our color correction based on regions defined by the SIFT algorithm, we assume that the user will not perceive significant color variation between views, and thereby it will contribute to improve his QoE.

As stated by Lowe in [87], SIFT keypoints are generally located in textured regions allowing to cover a large color range, and hence providing sufficient color statistics for the HM method. In order to have sufficient and robust statistics that span over the color space, we start from the set of sparse SIFT keypoints already determined, and then compute a set of dense correspondences around those points to obtain reliable corresponding regions, thus enabling to cover a large color range. To achieve that, we defined circular regions centered at SIFT keypoint position, and its scale parameter is used as the radius of the region.

Finally, the pixels in circular regions are added to their corresponding K_{ij} set. In the rest of this chapter, we use the name of *SIFT regions* to refer to the corresponding circular regions. Figure 3.10 shows an example of some SIFT regions defined in *Book Arrival* sequence.

A possible extension of the HM method to MVV can be achieved by applying the algorithm to the entire sequence on a frame-by-frame basis. This means that, a separate histogram is calculated for each time step and each view sequence. However, as each time step is corrected separately, a temporal color variation may be created

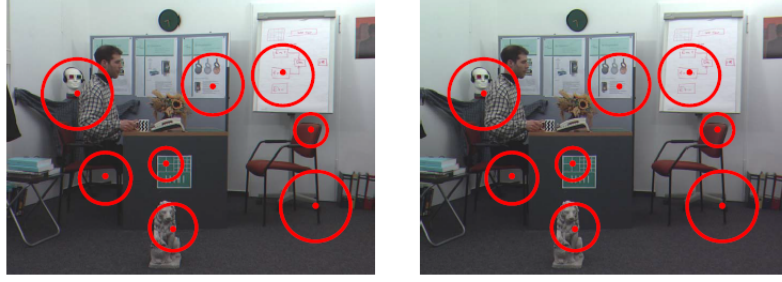


Figure 3.10: An example of some SIFT regions defined in *Book Arrival* sequence. The size of the circular regions is proportional to the scale parameter of its SIFT keypoint.

between frames, which adversely affects motion compensation and therefore coding efficiency thus leading to a perceptible flicker, decreasing consequently the QoE.

To overcome these problems, Fecker *et al.* proposed an extension of the HM method by introducing a time-constant mapping function [128]. They compute the histogram for the whole sequence; it is summed up over all frames of the sequence. So, the same correction is used for every temporal frame in the sequence. However, for the cases when the camera setup is not fixed and in real-time applications, this solution is not sound. Therefore, in our approach, we propose to perform the HM method in a temporal sliding-window. Unlike to [128], we calculate the histogram using only a group of picture (called CH-GOP), as shown in Figure 3.8. It is important to emphasize that the term “CH-GOP” here refers to the set of frames involved in the calculation of the cumulative histogram, and not the GOP during the coding process in MVC. The size of the CH-GOP (denoted w) can be fixed or variable within the same MVV sequence depending on the targeted application.

Besides that, we calculate the histogram based only on a union of all matched SIFT regions already detected, thus avoiding the problems of sensitivity to image content difference. To do so, the union of SIFT regions for each $CH-GOP_i^{(g)}$ in each view sequence is calculated as follows:

$$K_i^{(g)} = \bigcup_{j=t}^{w+t-1} K_{ij}. \quad (3.7)$$

where g is the CH-GOP number in the view i , and t the temporal order of the first frame of the $CH-GOP_i^{(g)}$.

Next, the histogram of each $CH-GOP_i^{(g)}$ is calculated as follows:

$$H_i^{(g)}(v) = \frac{1}{N} \sum_{k \in K_i^{(g)}} \delta(v, f(k)), \quad 0 \leq v \leq 255$$

$$\text{with } N = \text{Card}(K_i^{(g)}), \quad \text{and } \delta(x, y) = \begin{cases} 1 & \text{if } x=y \\ 0 & \text{else.} \end{cases} \quad (3.8)$$

f gives the intensity of the SIFT keypoint k in one of the color channels (R, G, B). Finally, the cumulative histogram of each $CH-GOP_i^{(g)}$ is derived

$$CH_i^{(g)}(v) = \sum_{j=0}^v H_i^{(g)}(j). \quad (3.9)$$

At this stage, the cumulative histogram of each $CH-GOP_i^{(g)}$ in each view i , is ready to be given as input of the color mapping function described in the next section.

3.3.4 Color mapping function

The purpose of this step is to minimize the difference between histograms for each CH-GOP of the reference view and those of all remaining views. This is done by means of the color mapping function that adjusts the cumulative histogram of each view to the common reference view. This correction is operated on a pairwise basis, by matching the cumulative histogram of each CH-GOP in the view being corrected ($CH_i^{(g)}$) to the cumulative histogram of correspondent CH-GOP in the reference view ($CH_{ref}^{(g)}$). Thus, this CH-GOP-based correction is performed as described in the following:

$$\begin{aligned} \forall x_1 \in [0, 255] \quad & M(x_1) = x_2, \\ \text{with } CH_i^{(g)}(x_1) &= CH_{ref}^{(g)}(x_2). \end{aligned} \quad (3.10)$$

where M is the color mapping function that maps each value of the view being corrected to a value in the reference view. In color correction, this function is also called *color transfer function*, *look-up tables* or, when only the grayscale is involved, *brightness transfer function* (BTF).

Consequently, one mapping function is derived for each CH-GOP. This means that, each CH-GOP is corrected individually. The derived mapping function is then applied to the set of frames belonging to the CH-GOP. By keeping the same reference view, this process is repeated for all remaining views.

The steps described so far are performed in the RGB color space instead of YUV. This is due to SIFT method that generates a negligible number of keypoints on the chromatic channels, since this components (U and V) have low texture, not allowing to have sufficient color statistics for the HM method. Although, SIFT parameters could be tuned to increase the number of keypoints, we made the decision to keep the same tuning in order to extract robust keypoints. We have also attempted to apply the SIFT algorithm only on the Y channel and have taken the same points for the chromatic channels (U and V). However, we obtained a lower performance than in RGB. In addition, usually most cameras used in recording video sequences operate in RGB, which makes it appropriate to perform the correction in this color space. Starting from these observations, RGB color space has been chosen for our

approach.

At the last stage, since the common video codecs operate in YUV, corrected RGB components are converted to this color space before feeding the sequence to MVC encoder.

3.4 Experimental results

This section is devoted to the presentation of experimental results related to the proposed color correction method for MVV. The assessment process is organized into three experiments reported in the following. The first investigates the impact on visual color consistency, the second evaluates its impact on coding performance, and finally, the third explores the effect on image-based rendering application. For the two first experiments, we used four standard MVV data test composed of: *Flamenco2* (5 views), *Race1* (8 views), *Ballroom* (8 views), and *Exit* (8 views) [149]. All these sequences have a resolution of 640×480 . We ran the third experience with four other standard test sequences namely *Balloons*, *Kendo*, *Newspaper*, and *Lovebird1*. The latter are in the multi-view video plus depth (MVD) format, with a resolution of 1024×768 .

It is worth pointing out that the size of CH-GOP (w), when calculating the cumulative histogram, plays an important role on the final results. After performing extensive tests, we noticed that when the CH-GOP is set to four frames ($w = 4$), this leads to introduce a slight color variation between CH-GOPs, which may appear unnatural to users and therefore reduce their QoE. To address this, we tried several sizes of CH-GOP on different test sequences, and thanks to visual inspection, we achieved to the conclusion that $w \geq 8$ produces improved results, and consequently no significant variation/impairment was noticed between successive CH-GOPs. So, transitions between CH-GOPs of corrected images are smooth and imperceptible to users. The results presented in the rest of this section are obtained by setting the CH-GOP to eight frames ($w = 8$).

3.4.1 Effect on the visual color consistency

Firstly, to prove the visual improvement produced by the proposed method in terms of color consistency, we show examples of frames before and after correction. Therefore, one frame is taken from each view at the same time step, as well as five views among eight are presented. Figures 3.11 to 3.14 illustrate the color correction results of *Flamenco2*, *Race1*, *Ballroom*, and *Exit* test sequences, respectively. We can notice that, in all these figures, before correction (original views) there are apparent color differences between the different views. It is thus clear that color consistency among these original views is poor. By following the procedures outlined in Section 3.3.2, from the left to the right on Figures 3.11 to 3.14, the chosen reference view (with red border) for each sequence is as follows: *Flamenco2* (view 2), *Race1* (view 3), *Ballroom* (view 5), and *Exit* (view 4). For example in *Flamenco2* sequence (Figure. 3.11), the center view (*i.e.*, view 3) in the camera arrangement is slightly blurred

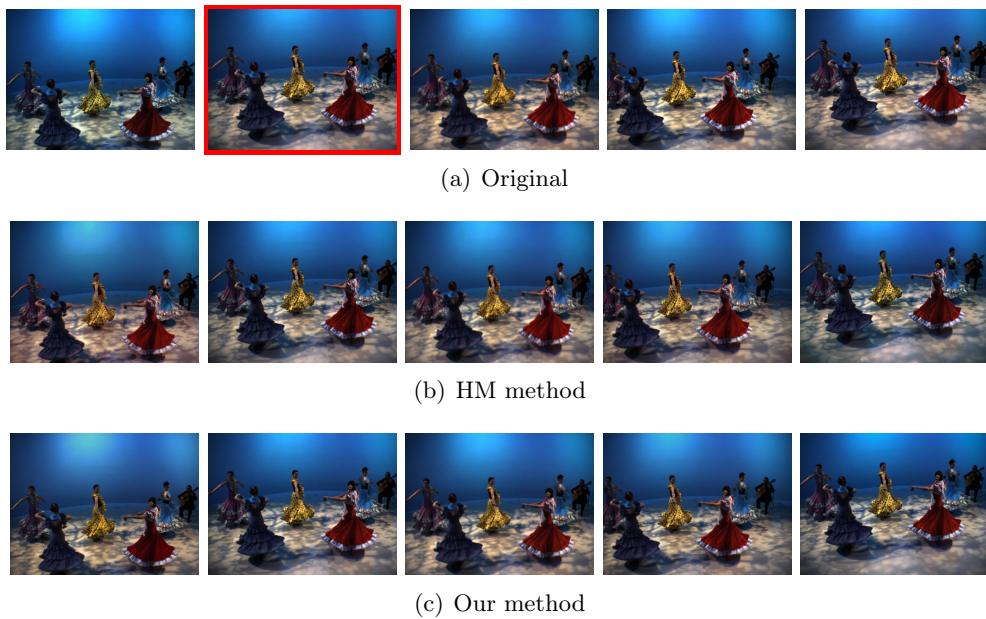


Figure 3.11: Visual effect of the proposed method on the *Flamenco2* test sequence. Comparison between: (a) without correction, (b) HM method, and (c) Our method. The image framed by a red border represents the reference image.

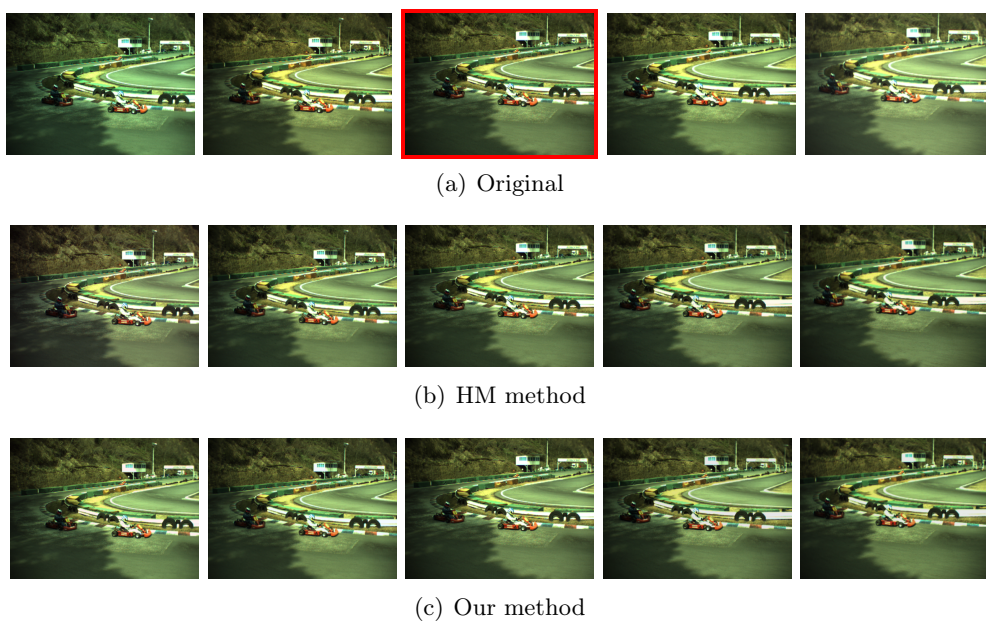


Figure 3.12: Visual effect of the proposed method on the *Race1* test sequence. Comparison between: (a) without correction, (b) HM method, and (c) Our method. The image framed by a red border represents the reference image.

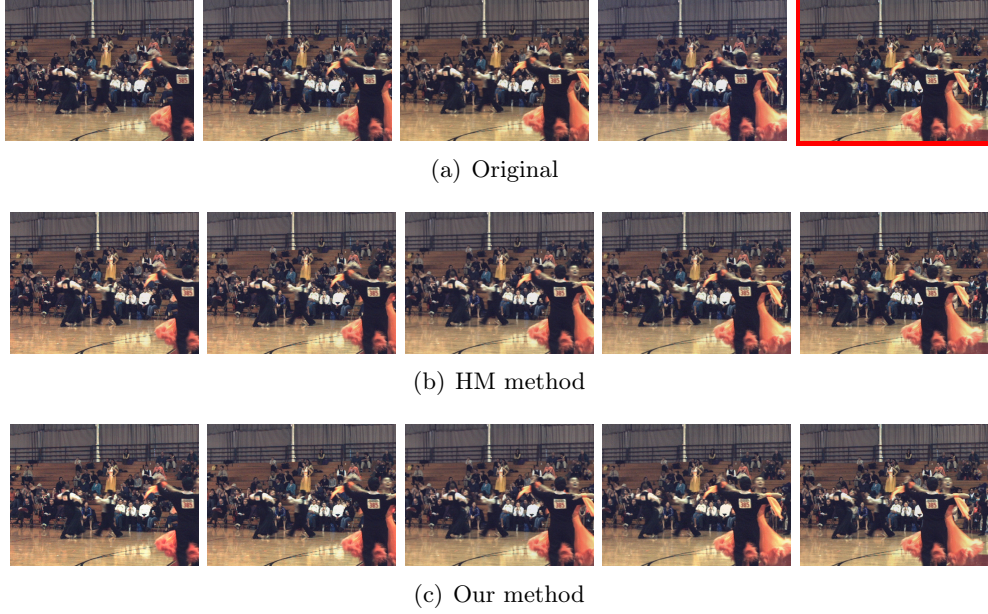


Figure 3.13: Visual effect of the proposed method on the *Ballroom* test sequence. Comparison between: (a) without correction, (b) HM method, and (c) Our method. The image framed by a red border represents the reference image.

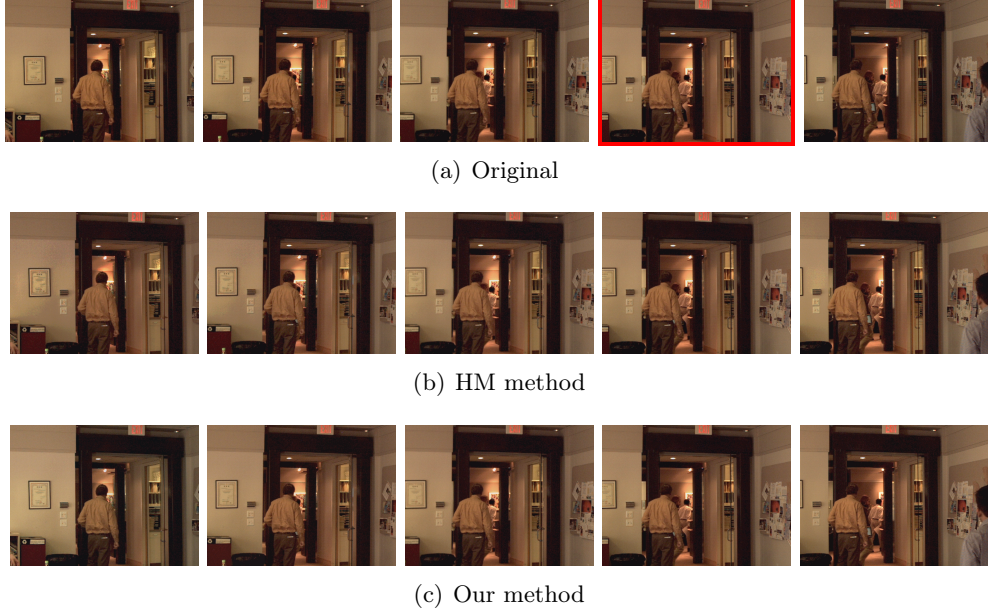


Figure 3.14: Visual effect of the proposed method on the *Exit* test sequence. Comparison between: (a) without correction, (b) HM method, and (c) Our method. The image framed by a red border represents the reference image.

	V1	V2	V3	V4	V5	C_{V_j}
V1	1	0,901	0,815	0,787	0,856	0,872
V2	0,901	1	0,82	0,913	0,944	0,916
V3	0,815	0,82	1	0,793	0,843	0,854
V4	0,787	0,913	0,793	1	0,881	0,875
V5	0,856	0,944	0,843	0,881	1	0,905

(a) *Flamenco2*

	V1	V2	V3	V4	V5	C_{V_j}
V1	1	0,836	0,852	0,806	0,723	0,843
V2	0,836	1	0,948	0,929	0,75	0,893
V3	0,852	0,948	1	0,978	0,744	0,904
V4	0,806	0,929	0,978	1	0,78	0,899
V5	0,723	0,75	0,744	0,78	1	0,799

(b) *Race1*

	V1	V2	V3	V4	V5	C_{V_j}
V1	1	0,862	0,848	0,852	0,854	0,883
V2	0,862	1	0,831	0,856	0,872	0,884
V3	0,848	0,831	1	0,853	0,878	0,882
V4	0,852	0,856	0,853	1	0,869	0,886
V5	0,854	0,872	0,878	0,869	1	0,895

(c) *Ballroom*

	V1	V2	V3	V4	V5	C_{V_j}
V1	1	0,856	0,811	0,845	0,827	0,868
V2	0,856	1	0,818	0,836	0,817	0,865
V3	0,811	0,818	1	0,824	0,801	0,851
V4	0,845	0,836	0,824	1	0,855	0,872
V5	0,827	0,817	0,801	0,855	1	0,860

(d) *Exit*

Figure 3.15: Illustration of the automatic selection of the reference view for the four used test sequences.

and is less contrasted. If this view was chosen as the reference (color reference), it would have impaired the other views. In this case, view 2 offers the best visual quality and at the same time has similar content with the other views. It is for those reasons that the process defined in Section 3.3.2 automatically chose this view as the reference. In contrast, on *Race1* (Figure 3.12), the center view (*i.e.*, view 3) is chosen as the reference, because it shares the most content with the other views and also offers good visual quality. This proves again that the proposed process of reference view selection allows selecting the appropriate one automatically and can adapt to different cases.

Figure 3.15 shows C_{V_j} calculation for the four used sequences. It can be noticed that for some sequences (*Flamenco2* and *Race1*) the difference is more important than the others (*Ballroom* and *Exit*). It is clear that if the central view is impaired (*e.g.*, in Figure 3.16(b)), the application of the proposed automated selection will help in avoiding the impairment of the whole views. Indeed, by applying a Gaussian blur to the central view, the proposed algorithm selects another view as a potential color reference, as illustrated in Figure 3.16(c).

Besides reference view selection, Figures 3.11 to 3.14 allow to compare the proposed color correction approach with the histogram matching (HM) described in [128]. One can notice that results obtained with our approach offer a better color consistency among views, especially for the Figures 3.11 and 3.12. Hence, resulting frames are visually closer to the reference image compared to HM method. In addition, HM introduced impairments in some cases, as for the *Flamenco2* test sequence. This is usually due to occluded regions. For a better analysis of the color correction results, some sensitive parts from *Flamenco2* and *Race1* sequences have been enlarged and depicted in Figure 3.17. This helps in concluding that our

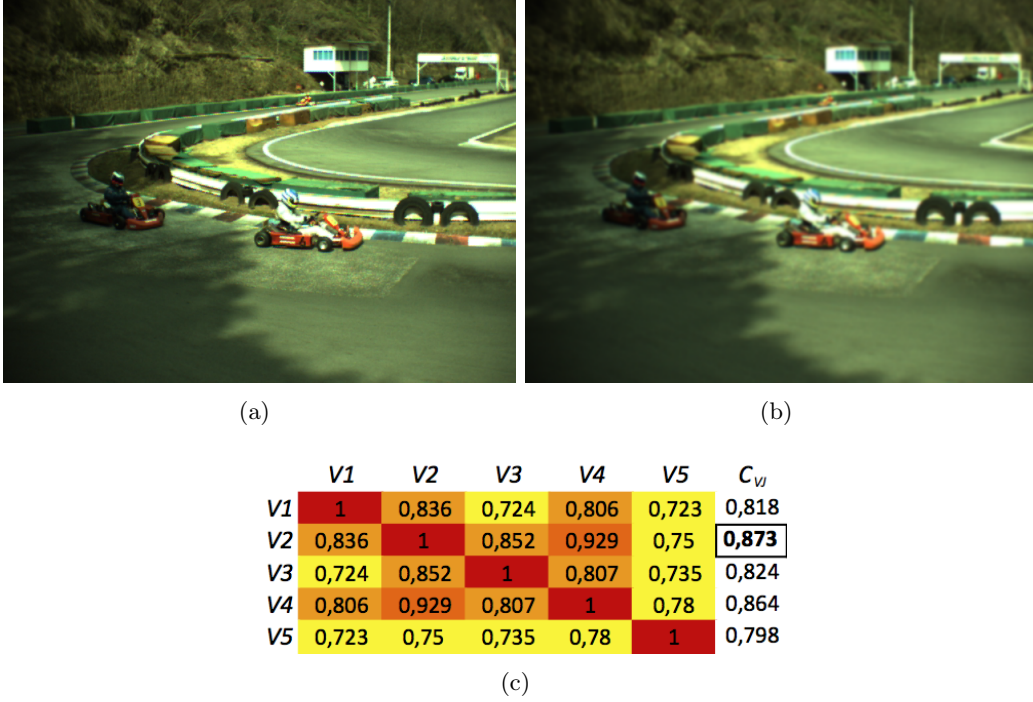


Figure 3.16: Impact of impairments on the reference view selection for the *Race1* sequence: (a) original frame, (b) distorted frame using Gaussian blur, (c) selection of the adapted reference.

method outperforms the HM method.

In addition, to objectively measure the color consistency across views, we use the well-known CIEDE2000 color-difference [151] defined and standardized by the CIE (International Commission on Illumination). The latter is based on the CIELAB color space, which has been designed to represent colors in a perceptual manner. The detailed definition and implementation of the CIEDE2000 equation can be found in [151, 152]. Figure 3.18 shows the color difference comparison for each sequence, before and after correction using HM and our proposal, between the reference view and all remaining views. The curves highlights that the color difference generated by the proposed method is significantly lower in comparison with HM, thus demonstrating the consistency of the proposed color correction.

3.4.2 Effect on coding performance

In order to assess the effect of the proposed color correction on coding performance, we used it as a preprocessing before feeding the sequences to MVC encoder (JMVM version 8.0) [99]. The experiments were performed using typical settings for MVC, for details regarding the test conditions the reader can refer to [149]. We compared the proposed method with the multi-view encoder that uses original sequences without any correction (denoted as *no correction* or *NC*), with the

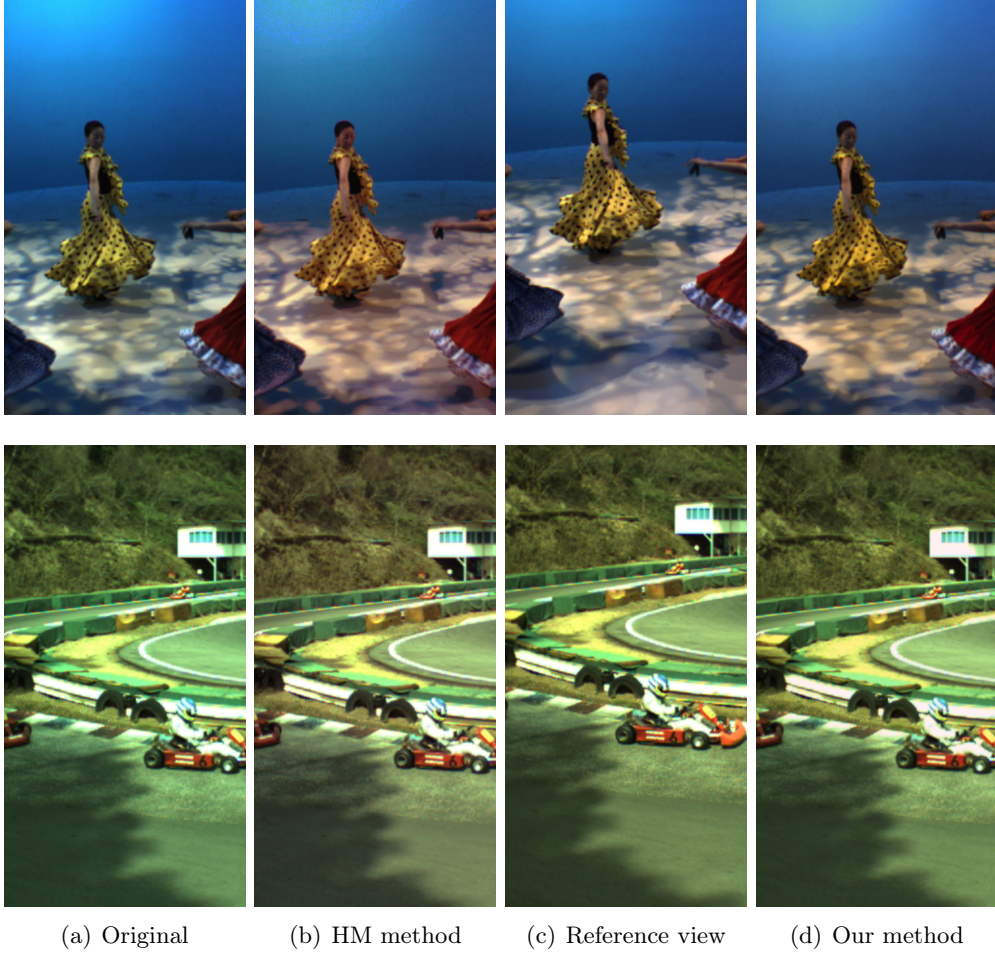


Figure 3.17: Enlarged parts of the *Flamenco2* (top) and *Race1* (bottom) sequences.

illumination compensation (denoted as *IC*) method implemented in the JMVM reference software [99, 153], as well as the HM method [128]. Figure 3.19 illustrates the PSNR *versus* bit-rate averaged over all views of the test sequence for both luma and chroma. To measure the relative gain, the average PSNR/bitrate differences between rate-distortion curves is calculated using the Bjøntegaard measure [154] and given in Table 3.1.

As it can be seen from the plots, the coding performance comparison clearly indicates the benefits of using the proposed color correction method. The quality improvement are reported for the luminance as well as the chrominance components, where the PSNR gains for the luma are about 0.5–1.1 dB compared to no correction. For chroma, the gains are even higher ranging from between 0.7 and 2.2 dB again in comparison with no correction. These results demonstrate that the color consistency created by our color correction helps in reducing the inter-view difference during the prediction stage. Besides, the comparison between our color correction and IC, shows also an improvement that is slightly lower than the one with regards to no

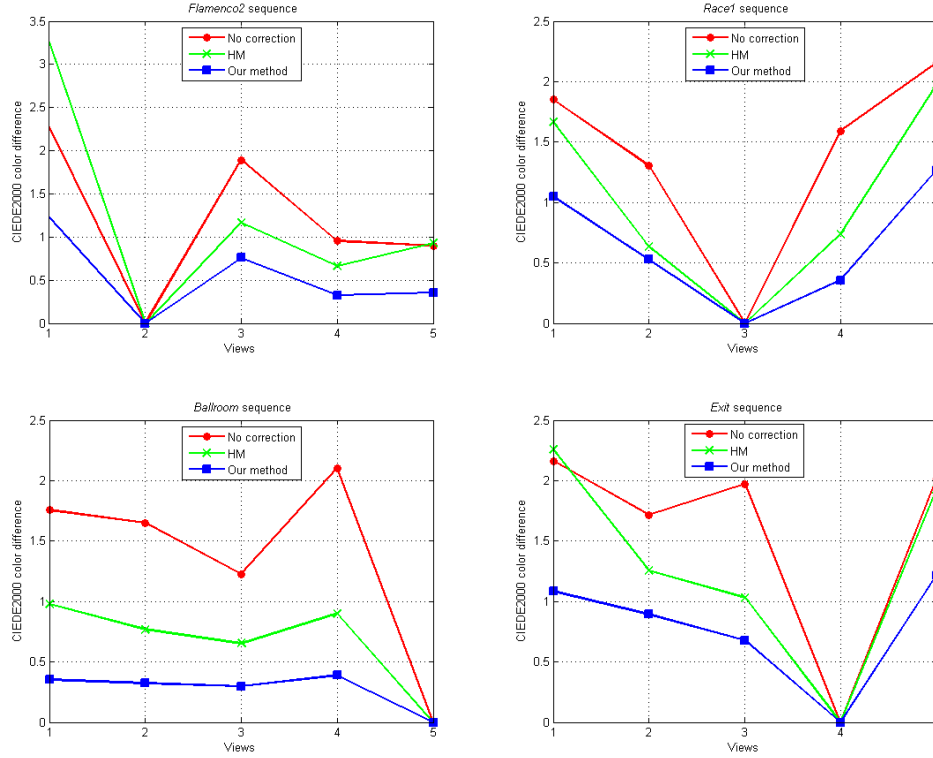


Figure 3.18: Objective color difference comparison for the sequences: *Flamenco2*, *Race1*, *Ballroom*, and *Exit*. CIEDE2000 color-difference equal to 0 correspond to the reference view (Lower values denote better correction).

Table 3.1: Comparison of PSNR averages (dB) between IC method, HM method and proposed method using Bjøntegaard measure.

Sequence	Y			UV		
	IC	HM	Proposed	IC	HM	Proposed
<i>Flamenco2</i>	0.34	0.56	0.87	0.29	0.45	1.04
<i>Race1</i>	0.49	0.52	1.01	0.37	0.40	1.76
<i>Ballroom</i>	0.08	0.32	0.74	-0.01	0.79	1.12
<i>Exit</i>	0.21	0.33	1.03	0.42	0.48	1.58

correction. Therefore, the gains are about 0.3–0.7 dB for luma, and about 0.6–1.8 dB for chroma. In the same vein, an improvement is also observed compared to using HM, which is about 0.1–0.6 dB for luma, and about 0.4–1.4 dB for chroma.

Finally, the PSNR curves and Bjøntegaard relative gain show that the proposed method improves the coding efficiency particularly the chrominance components, and confirmed its suitability as a color correction for multi-view video in comparison to the widely used HM.

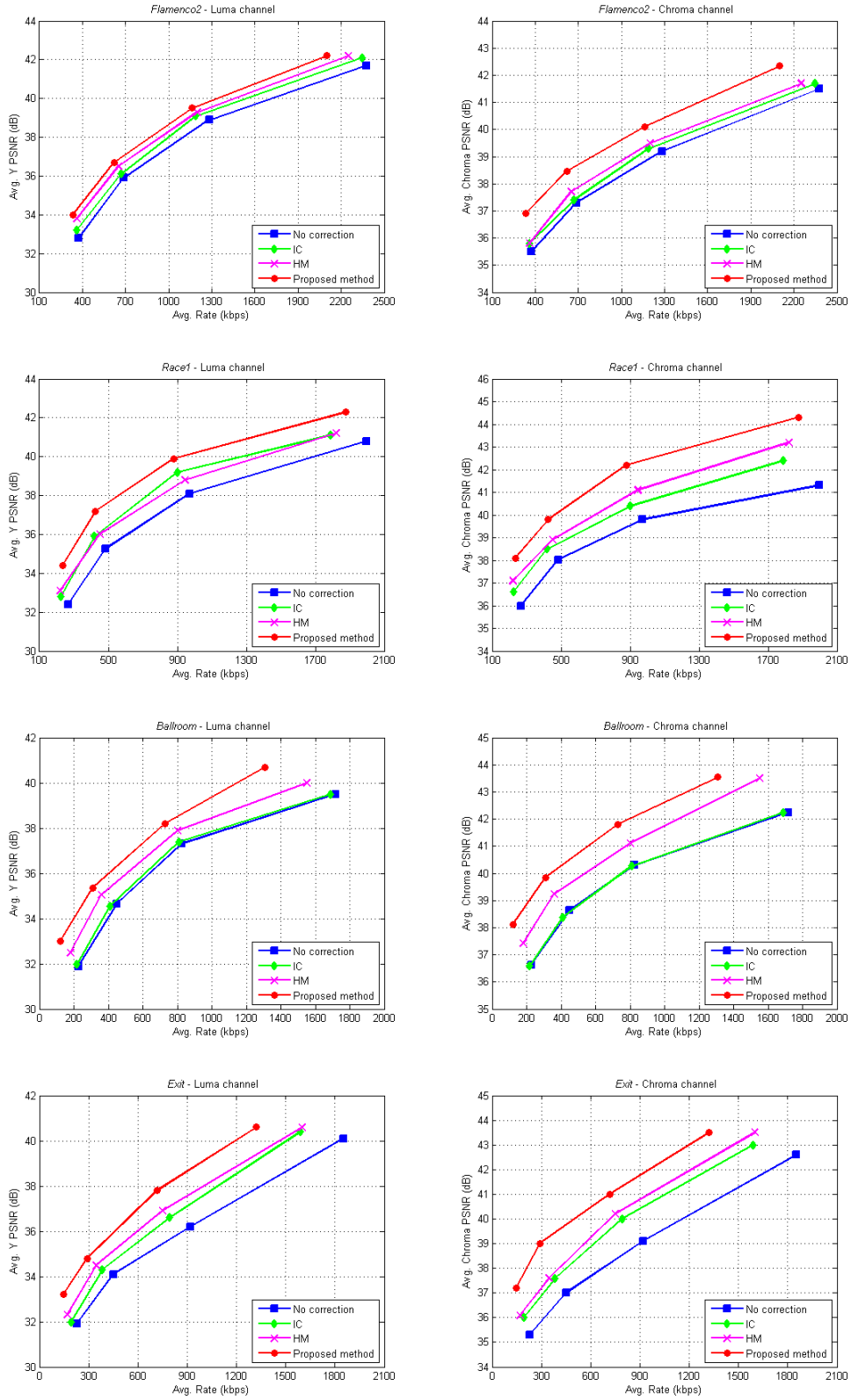
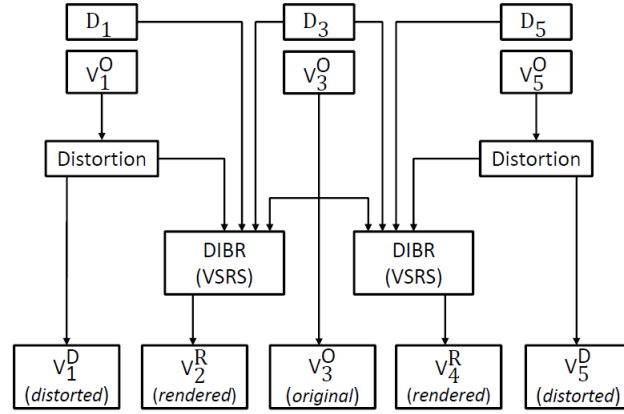
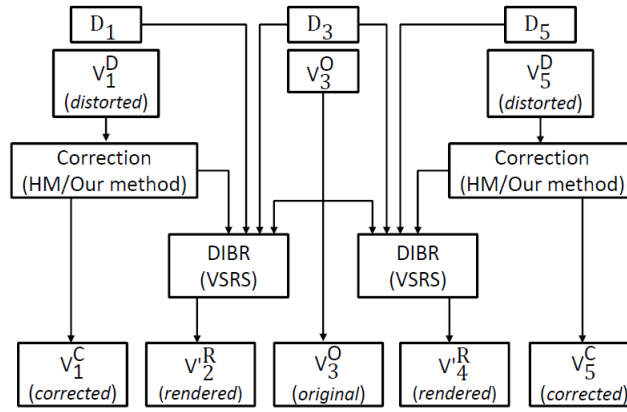


Figure 3.19: Rate distortion curves for the *Flamenco2*, *Race1*, *Ballroom*, and *Exit* test sequences. The chroma PSNR is the average PSNR of the U and V channels.



(a) No correction



(b) With correction

Figure 3.20: The experiment scenario to evaluate the effect on the DIBR-based application. V_i^* : View i (texture), D_i : Depth i . V_i^O : Original view, V_i^D : Distorted view, V_i^C : Corrected view, V_i^R : Rendered view before correction, $V_i'^R$: Rendered view after correction.

3.4.3 Effect on rendering performance

None of the previously cited works has directly addressed the color correction for DIBR-based applications. Indeed, most of them mentioned that their methods are helpful for rendering new virtual views. However, none of them has proved it by experimental results. Therefore, to analyze the impact of the proposed method on DIBR-based applications, it has been applied on the four MVD sequences (*Balloons*, *Kendo*, *Newspaper*, and *Lovebird1*). These sequences consist of multiple color videos with associated multiple depth data. To be in line with the test scenarios defined in [150], we selected 3 views (view 1, view 3, and view 5) from each sequence and performed the experiment scenario depicted in Figure 3.20. As illustrated in this figure, to highlight our contribution, we manually distorted two views (view 1 and

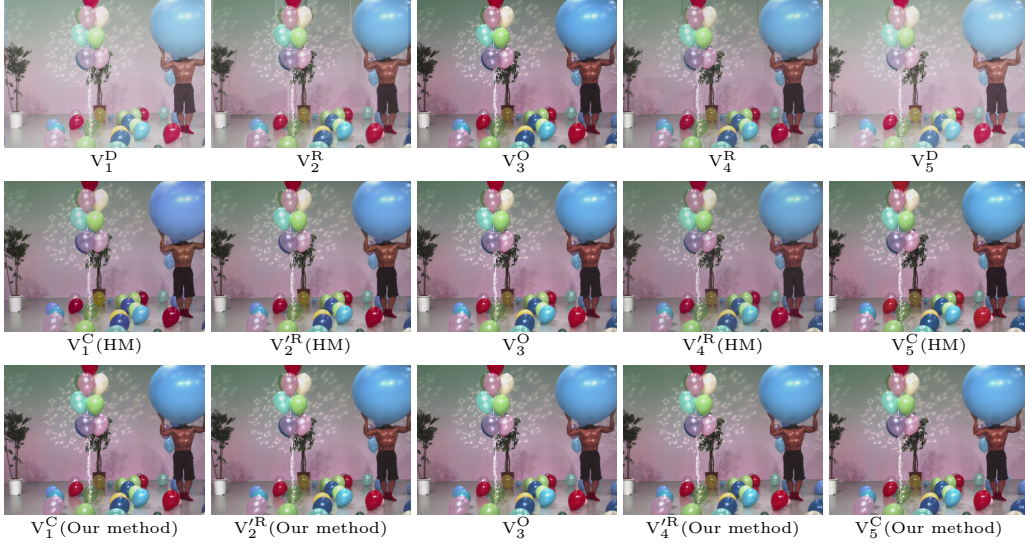


Figure 3.21: Rendering results for *Balloons* sequence (top: No correction, middle: HM method, and bottom: Our method). V_i^* : View i , V_*^O : Original view, V_*^D : Distorted view, V_*^C : Corrected view, V_*^R : Rendered view before correction, $V_*'^R$: Rendered view after correction.

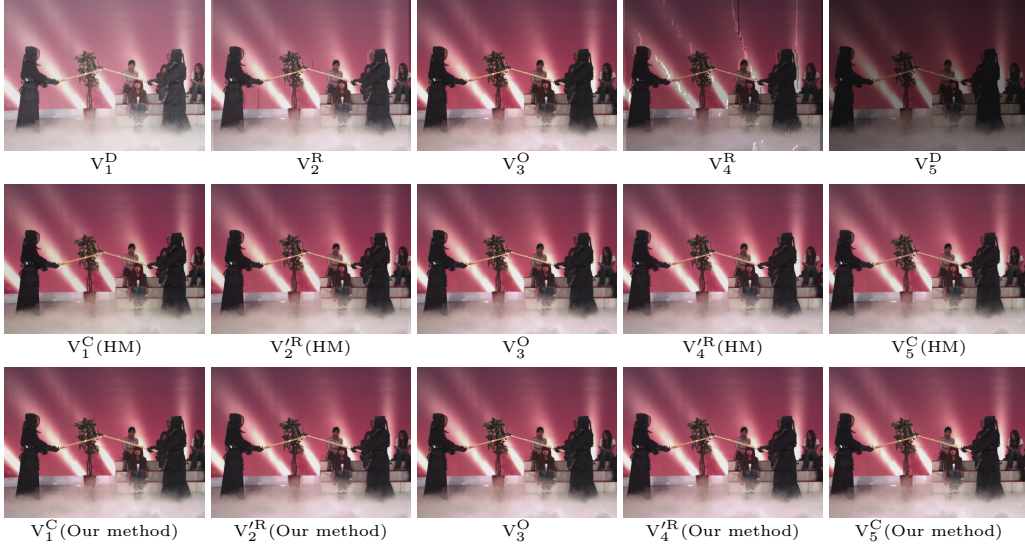


Figure 3.22: Rendering results for *Kendo* sequence (top: No correction, middle: HM method, and bottom: Our method). V_i^* : View i , V_*^O : Original view, V_*^D : Distorted view, V_*^C : Corrected view, V_*^R : Rendered view before correction, $V_*'^R$: Rendered view after correction.

view 5) by an increase/decrease in brightness and gamma correction. By following the procedures outlined in the Section 3.3.2, the remaining undistorted view 3 (V_3^O) is chosen as the reference view. Then, we performed the correction of the distorted views (V_1^D and V_5^D) by the proposed method and the HM method separately. Finally,



Figure 3.23: Rendering results for *Newspaper* sequence (top: No correction, middle: HM method, and bottom: Our method). V_i^* : View i , V_*^O : Original view, V_*^D : Distorted view, V_*^C : Corrected view, V_*^R : Rendered view before correction, $V_*'^R$: Rendered view after correction.

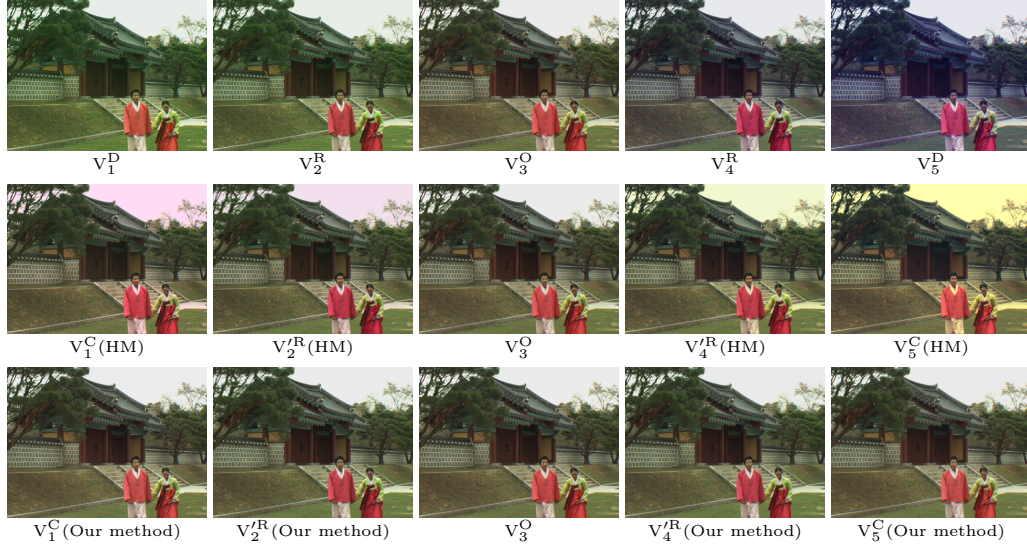


Figure 3.24: Rendering results for *Lovebird1* sequence (top: No correction, middle: HM method, and bottom: Our method). V_i^* : View i , V_*^O : Original view, V_*^D : Distorted view, V_*^C : Corrected view, V_*^R : Rendered view before correction, $V_*'^R$: Rendered view after correction.

we generated two virtual views before and after correction, by DIBR technique [155] based on the View Synthesis Reference Software (VSRS version 3.5) [156], between the reference view and the distorted (resp. corrected) ones. The results for the four sequences are depicted in Figures 3.21 to 3.24.



Figure 3.25: Enlarged parts of the rendering results for *Balloons* (top) and *Newspaper* (bottom) sequences.

We can notice from the Figure 3.25 (in addition to Figures 3.21 to 3.24) that when we used the distorted views (uncorrected), we have obtained virtual views (V_2^R and V_4^R) containing visible distortions. For instance, a ghosting effect can be seen, in addition to some darker/lighter regions in the foreground and background of the scene, which is annoying for the viewer. The distortions can be explained by the fact that the color inconsistency between views induces important errors in the rendered views. However, once the correction is performed, we obtained significant

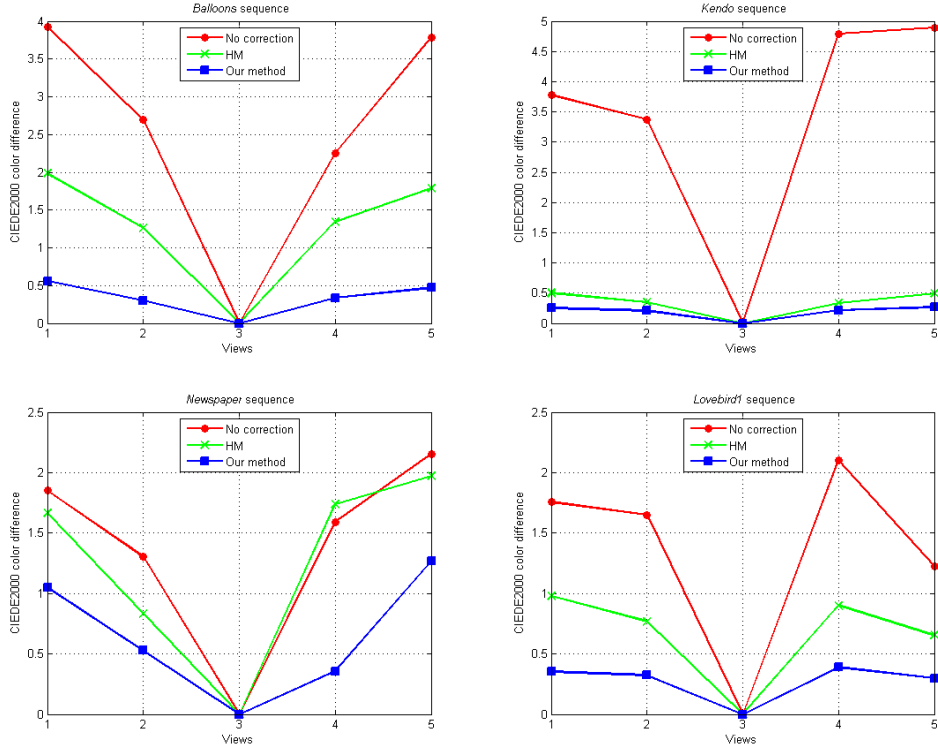


Figure 3.26: CIEDE2000 color-difference after rendering by DIBR for the *Balloons*, *Newspaper*, *Lovebird1*, and *Kendo* sequences. CIEDE2000 color-difference is computed between the reference view (view 3) and all remaining views (real and rendered).

improvements in the rendering quality (V_2^R and V_4^R). Similar to Section 3.4.1, to properly and objectively assess these improvements achieved by the two used color correction techniques, we used the CIEDE2000 color-difference between the reference view (view 3) and all remaining views (real and rendered), as shown in Figure 3.26. From these curves and the visual comparison, one can confirm that the proposed method gives the best performance and outperforms the HM method. Moreover, these results clearly demonstrate the benefits of using our method in DIBR-based applications such as FVV, as it allows to smooth the switch between the different views, leading to a more natural and enhanced 3D viewing experience.

3.5 Conclusion

In this chapter, we proposed a preprocessing method for color correction of MVV. Firstly, instead of choosing manually the center view as the reference view, we proposed to perform its selection automatically and effectively in a more adaptive manner, based on histogram difference and structural content. This allows to choose the view representing the best match of the whole set of views, and can be highly

adapted for some specific capture configuration. Then, the correction is performed using an improved version of the widely used histogram matching (HM) algorithm. To avoid its drawback, HM method is used only on overlapping regions over views, making our method robust to occlusion. The combination of the SIFT algorithm and RANSAC method is used to effectively define these common regions. In order to perform the correction while maintaining as much as possible the temporal correlation, we applied the method using temporal sliding-window, bringing the advantage of adaptation to time-varying acquiring system and real-time broadcasting.

Experimental results showed that, once the correction is performed with our method, the color of views is harmonized and consistent. This results in significant coding gains up to 1.1 dB for the luminance component and up to 2.2 dB for the chrominance components. Regarding rendering performance, the presented results clearly indicate the benefits of using our method before applying the DIBR technique. It contributes significantly in smoothing the transition between views, which leads to enhance the 3D viewing experience.

Among the future work that deserves to be addressed, is to seek the way to re-use the informations produced by feature correspondence step within the coding process itself. For example, to reduce the search area when calculating the disparity vectors, or to extract the parameters of the 3D geometric model [157], thus achieving the complete end-to-end 3D video system. In addition, maintaining the temporal coherence between successive frames within a single view needs to be improved. The proposed approach tackled this problem by defining a group of picture and perform the correction group by group. However, the size of group is fixed. Future studies should focus on proposing more advanced strategies allowing to tune adaptively and automatically the size of each group. Also, the reduction of computational complexity remains an important issue, especially, with the emergence of new real-time 3D applications. One solution to address this problem is to parallelize the color correction process by using a graphics processing pipeline or a GPU implementation. Another point is the fact that there is no ground truth allowing efficient evaluation. To address this, we plan to construct a benchmark in order to be able to evaluate and to compare each method of the stat-of-the-art. Finally, the impact of color variation is not fully investigated, the relation between color inconsistencies and binocular rivalry should be explored. As the latter can induce visual fatigue, the maximum tolerance of color difference between views need to be quantified.

Chapter 4

Asymmetric Coding for Stereoscopic Video

Contents

4.1	Introduction	79
4.2	Asymmetric stereoscopic video coding methods	81
4.2.1	Temporal resolution reduction	81
4.2.2	Asymmetric quantization	83
4.2.3	Mixed resolution coding	84
4.2.4	Asymmetric luminance/chrominance	88
4.3	Perceptual limits of asymmetry	88
4.4	Proposed adaptive asymmetric stereoscopic coding	89
4.5	Experimental results	95
4.5.1	Experience 1	95
4.5.2	Experience 2	100
4.6	Conclusion	102

4.1 Introduction

3D video is widely considered as one of the main advances in digital video technology of the last decade, and has already made its way to becoming a remarkable worldwide success story. As mentioned in Section 2.3, there exist different techniques to represent 3D content. However, the most widespread and widely enjoyed 3D video format is currently the stereoscopic one [77, 78]. Based on the underlying principles of stereopsis, most of the 3D content is captured and prepared in stereoscopic format and numerous consumer legacy devices are backward compatible with this format.

Since, stereoscopic 3D (S-3D) consists of capturing two views from two slightly different positions. Nominally, representing stereoscopic video requires at least twice the amount of data compared to monoscopic video. This increase in data rate poses one of the most imposing challenges that has received a commensurate amount of research attention. Thus, a central issue in the storage, media transmission and consumption of such S-3D content is the compression technology that is utilized. State-of-the-art coding standards or compression methods used for S-3D video include the Annexes G and H of H.264/MPEG-4 Part 10 AVC [73], simulcast approach, MVC [74, 96], the recently established High Efficiency Video Coding (HEVC) [75], and the depth enhanced extension for MVC (MVC+D) [76] (see Section 2.4).

In the last decade, we have witnessed a huge increase in the use of perceptual approaches to improve the performance of multimedia processing techniques [79]. Based on the characteristics of the HVS, different methods have been developed in many challenging areas of image processing such as image coding [80, 81], enhancement and watermarking, image quality assessment and so on. Among these promising techniques, we can quote those used for coding S-3D content known as asymmetric coding methods. Asymmetric coding techniques are based on the so-called *binocular suppression theory* [69–72], which specifies that if the eyes are provided with two views having a different quality, the fused perceptual 3D image is dominated by the higher quality view [162–166, 170–172].

Therefore, according to this theory, in order to reduce the required bit-rate for stereoscopic video delivery, one view of the stereo pair can be encoded at a sufficiently high quality, and the other view is encoded at a slightly lower quality (with a lower bit-budget), with transparent 3D viewing for the user. This approach has focused considerable research efforts, and has been exploited in different ways. Depending on how the quality (information) reduction of one of the views is performed, the different methods belonging to this concept can be classified into four categories: (a) spatial resolution reduction (spatial filtering), (b) asymmetric quantization (unequal QP), (c) temporal resolution reduction, and (d) asymmetric luminance/chrominance. With respect to the literature, spatial filtering is by far the most explored and is usually referred to as mixed resolution (MR) coding.

One large obstacle in asymmetric stereoscopic video coding methods is linked to the level of asymmetry between views, *i.e.*, the just noticeable threshold of asymmetry, so that 3D viewing experience is not altered. With the aim to address this issue, several works tried to identify this asymmetry threshold by means of subjective experiments. Even though these studies were successful in defining or identifying thresholds, the latter depend heavily on the experiment design and do not allow any adaptation to quality and content variations in the image. Consequently, in this chapter, we propose a new non-uniform asymmetric coding method of S-3D [159, 160] adjusting in an adaptive way the level of asymmetry for each image region to ensure unaltered binocular perception. This is achieved by exploiting several HVS-inspired models.

The rest of the chapter is organized as follows: Section 4.2 describes the different types of asymmetry that have been proposed in the literature for stereoscopic video

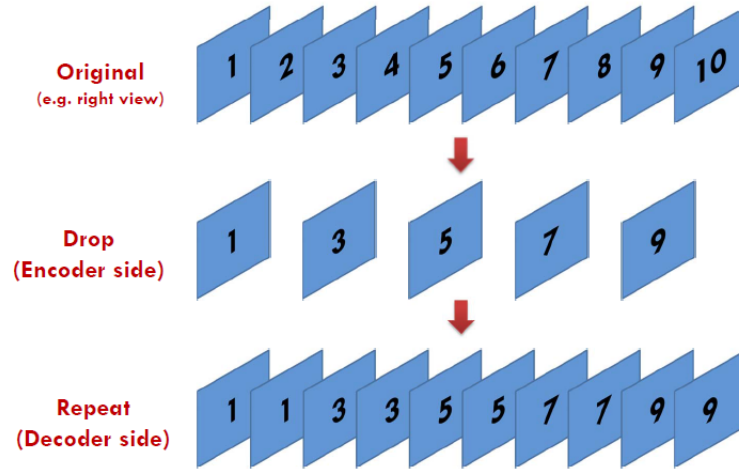


Figure 4.1: Illustration of dropping-and-repeating frames procedure adopted in temporal asymmetry.

coding. A review of the most important attempts for the determination of perceptual limits of asymmetry in the context of stereoscopic video coding is performed in Section 4.3. Section 4.4 describes the proposed approach and the various assumptions behind it. The experimental results, objective and subjective in addition to coding performance are given in Section 4.5. Section 4.6 summarizes and concludes this chapter.

4.2 Asymmetric stereoscopic video coding methods

Based on the suppression theory of binocular vision, the HVS fuses the retinal images coming from both eyes having different levels of quality such that the perceived visual quality is close to the view with higher quality. This property is effectively used to reduce the required bit-rate for stereoscopic video delivery, where one view of the stereo pair is encoded with higher strength than the other without inducing visible artifacts in the stereoscopic perception.

In the light of this standpoint, many asymmetric coding methods have appeared in the literatures. Depending on how the asymmetry in quality between the two views can be achieved, the different asymmetric stereoscopic video coding methods can be classified into four main categories detailed in the following.

4.2.1 Temporal resolution reduction

Generally, the temporal resolution reduction (reduction of frame rate) method corresponds to the decimation of some frames of one view, *i.e.*, frame dropping, while the other view is kept unaltered. Depending on the adopted 3D representation format, Stelmach *et al.* proposed two strategies to perform the temporal filtering [164]. In the first one, each pixel of the current field is averaged with the corresponding pixel

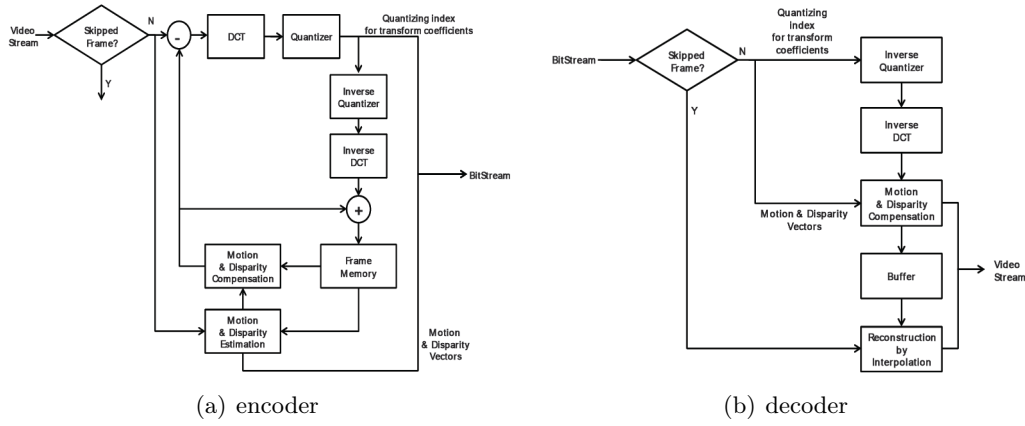


Figure 4.2: The structure of the encoder and decoder including temporal asymmetry [167].

in the next field of the same polarity. This method can be exploited with frame compatible format, when the views are interlaced into a signal sequence of frames (see Section 2.3.1). The second strategy consists of dropping either even or odd frames of one view of the stereo pair. For example, in Figure 4.1 only the odd frames are encoded, while all the even frames are dropped. At the decoder side, each frame is repeated to replace the dropped frames. Aksay *et al.* in [173] exploited this second method to perform temporal asymmetry. They implemented it by sending all the macroblocks of the dropped frame using the skipped mode of the H.264/MPEG-4 AVC standard. This has been done by a post-processing procedure in which NAL units of the dropped frames in the altered view are replaced with special NAL units signaling skip mode for each macroblock. When such a frame is decoded, each macroblock is copied from the reference frame resulting in frame repetition. More advanced strategies have been proposed in [167], where the authors developed two skipping modes. In the first skipping mode, all frames in the left view are encoded, while only odd frames in the right view are encoded. In the second mode, only even frames in the left view and only odd frames in the right view are encoded. On the encoder side, they signal the skipped frames in the bitstream, and they reconstruct them using the advanced bilateral interpolation on the decoder side, as illustrated in Figure 4.2. However, this approach requires syntax and design changes affecting low levels of the encoding and decoding processes.

The performance of temporal filtering has been previously studied in [164, 173]. In [164], the subjective results showed that the temporal filtering produces unacceptable results, with poor perceived 3D quality and sharpness, even if the perceived depth was relatively unaffected. Furthermore, there was noticeable double contouring with the 2-field averaged images approach, while the drop-and-repeat alternate frames approach had a jerky appearance. Aksay *et al.* concluded, based on subjective tests, that using temporal asymmetry leads to a relatively low subjective 3D

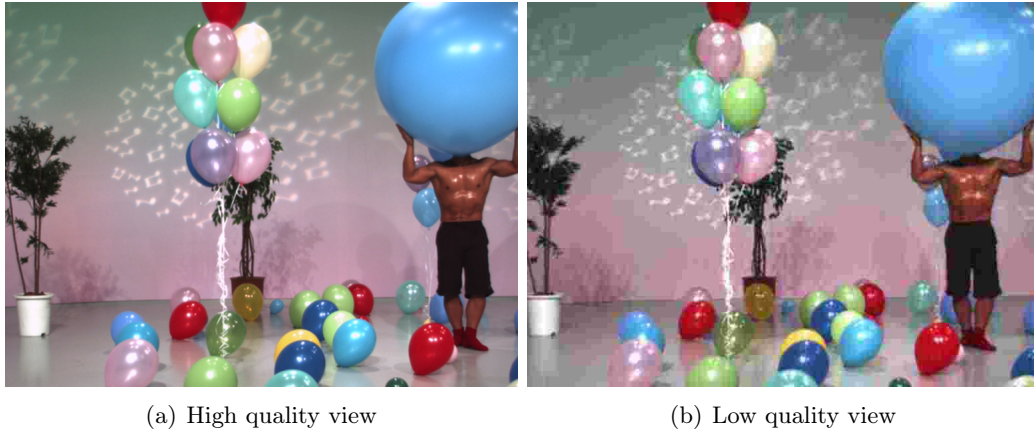


Figure 4.3: An example of asymmetric quantization.

video quality, and the performance is content dependent [173]. Moreover, this type of asymmetry is noticeable by viewers in case of fast motion content.

Finally, based on the previous studies it is not recommended to perform bandwidth reduction using this type of asymmetry.

4.2.2 Asymmetric quantization

Another way to reduce transmission bandwidth for stereoscopic video is to encode both views of the stereo pair at unequal PSNR or bit-rates, this method is referred to as PSNR (quality) reduction, SNR scaling, or asymmetric quantization. Implementing asymmetry by PSNR reduction is straightforward, and does not require any modification to the encoder/decoder. Since the encoding quality of a view depends on the quantization parameter, using different quantization parameters for the left and right views leads to PSNR asymmetry. Therefore, one view is encoded with a larger value of quantization than the other. As a result, one of the views has lower fidelity and may be subject to a greater amount of visible coding artifacts, such as blurring, blocking and ringing. An example of the effect of asymmetric quantization is depicted on the stereo pair of Figure 4.3.

The effect of asymmetric quantization method has been investigated in [162, 165, 166]. For instance, Stelmach *et al.* showed by fixing the bit-rate of the left view and reducing gradually that of the right view using MPEG-2, the binocular percept of the asymmetric MPEG-2 impaired images fell approximately midway between the 2D image quality scores of the left and right images [165]. Seuntiens *et al.* applied a different compression ratio on the left and right views of a stereo pair using JPEG [166]. The results showed that an increase in asymmetric JPEG coding of stereoscopic images had a negative effect on image quality, sharpness and eye-strain but had no effect on perceived depth. Furthermore, in contrast to the conclusion provided in [165], the results on asymmetric and symmetric coding showed that the relationship between perceived image quality and average bit-rate



Figure 4.4: Stereo image pair with low-pass filtered right view.

is not straightforward. In some cases, image quality ratings of a symmetric coded pair can be higher than for an asymmetric coded pair, even if the averaged bit-rate for the symmetric pair is lower than for the asymmetric pair. This conclusion has been further supported by the previous findings in [162], which demonstrated that the fused 3D image quality is mainly determined by the view of low level quality.

Saygili *et al.* [168] proposed to encode one view of the stereo pair using H.264/MPEG-4 AVC at sufficiently high quality (PSNR=40 dB), while the other view is encoded with reduced quality but kept above a certain PSNR threshold (this point will be detailed in Section 4.3). They noticed that the degradation in 3D viewing is unnoticeable. The results indicated that at high bitrates, asymmetric coding using PSNR reduction method achieves a good perceived quality. However, at low bitrates, symmetric coding achieves the best perceived quality, which is in conformance with the conclusion of Seuntiens *et al.* in [166].

Finally, to provide an acceptable 3D perceived quality when exploiting this type of asymmetry, a tradeoff between the subjective quality degradation and the bit-rate reduction should be considered. This point will be discussed in the section 4.3.

4.2.3 Mixed resolution (MR) coding

According to the perceptual phenomenon of binocular suppression, the HVS can tolerate the absence of high frequency information in one view. It is possible to reduce the spatial resolution of one view of the stereo pair while keeping the other view at its original high resolution, and finally without affecting the overall binocular perception. This approach is usually known as the spatial mixed resolution (MR) method.

Therefore, in the MR concept the left and right images are presented to the viewer with different levels of sharpness, *i.e.*, one image of the stereo pair is blurred, while the second is kept as is (sharp) (see Figure 4.4). Subjective tests have demonstrated that the perceived 3D fusion of these two images is relatively close to the sharper image [164]. This can be explained by the fact that the sharp image masks the blur effect of the lower quality one.

Perkins was the first to introduce the notion of MR coding [179]. In order to reduce the bit-rate for S-3D delivery, before the encoding stage, Perkins proposed to perform a low-pass filtering to one view, while the second view is unchanged. Due

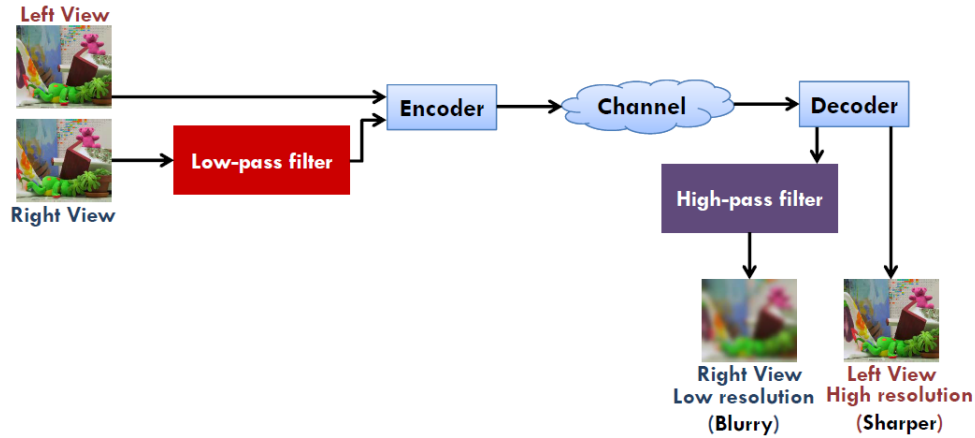


Figure 4.5: Block diagram of MR coding.

to the smaller number of pixels to be encoded compared to the full resolution (FR) case, this operation leads to reduce the transmission bandwidth. At the decoding stage, the view with lowest resolution is high-pass filtered to fit the resolution of the second view, as illustrated by Figure 4.5. Several works have followed this concept since then [170–178]. With respect to the literature, MR coding is by far the most explored and commonly used among the asymmetric coding approaches.

In [164], based on subjective tests, the authors studied the impact of uncompressed MR stereoscopic sequences at downsampling ratios of 1/2 and 1/4 applied on both horizontal and vertical axes. It was found that spatial filtering produced acceptable results, and had no effect on the overall sensation of depth. Add to that, when using half resolution the perceived sharpness and quality were nearly transparent, but quarter ratio provides a slight quality drop. Finally, the authors concluded that MR coding is an effective way to reduce transmission bandwidth, while maintaining acceptable image quality, sharpness, and depth. This conclusion is further supported by Seuntjens *et al.* in [166].

The performance of asymmetric quantization and spatial MR methods has been investigated in [162, 165, 171, 172, 177, 178]. According to their subjective results, they favor spatial MR over asymmetric quantization. They have found that viewers could identify relatively easily asymmetry achieved by quantization, than the one achieved by spatial resolution reduction. This is because the blocking artifacts appear as high spatial frequency noise, making them difficult to suppress by the other view, whereas the effect of blurring is masked. Finally, they conclude that in case of asymmetric blurring, the perceptual quality of a stereoscopic image pair is dominated by the view with higher resolution. However, in case of asymmetric quantization, the perceived 3D image quality is mainly determined by the low-quality component.

It was found in [178] that the subjective impairment caused by asymmetric blurring, is reduced by increasing resolution and viewing distance, and decreasing display size. In addition, the authors have studied the impact of bit-rate allocation between the left and right views in case of MR coding. The results showed that the

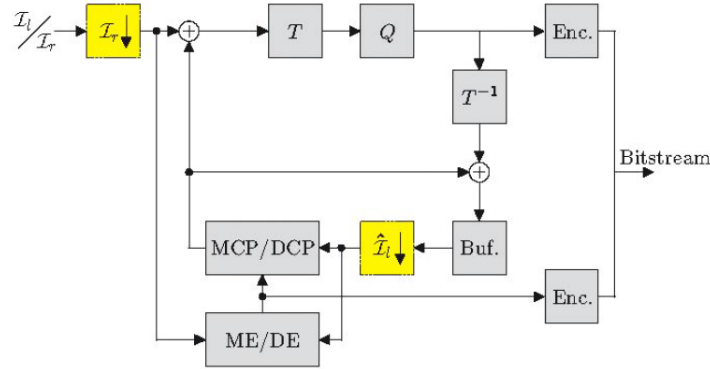


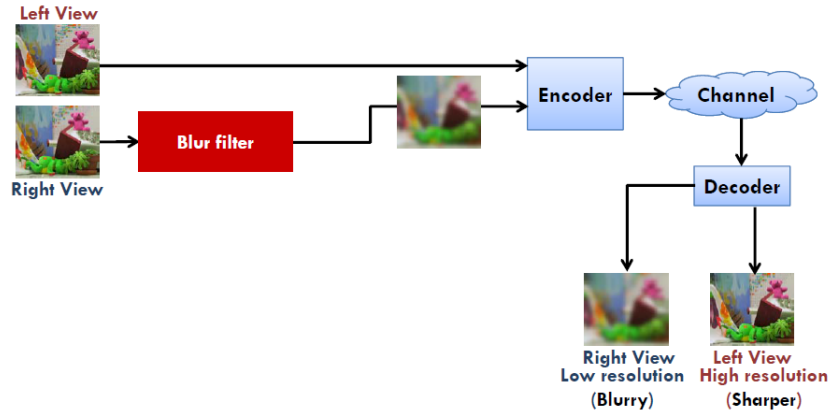
Figure 4.6: Asymmetric stereoscopic coding scheme based on MVC with two additional down-sampling blocks (in yellow) [174].

maximum PSNR is obtained at a distribution of 30% to 35% of the total bit-rate for the lower resolution view, when applying sub-sampling by a factor of two in both directions.

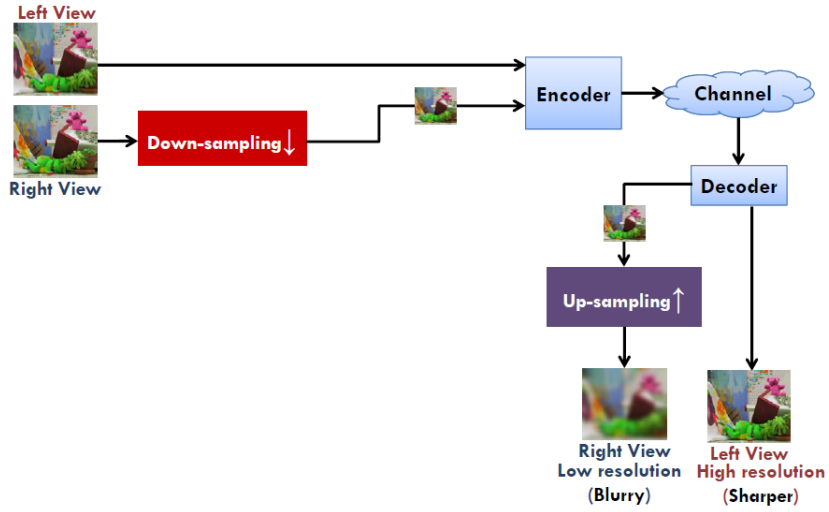
As mentioned previously (see section 2.4.3), using inter-view prediction produces additional coding gains. In order to increase the coding performance, some works have focused on how to combine the inter-view prediction with MR coding approach. Since, in the case of MR, the two views have different spatial resolutions not allowing to apply directly the inter-view prediction. One solution proposed by Fehn *et al.* [174] is to integrate an additional downsampling block attached to the reference picture buffer (as illustrated by Figure 4.6). This block is required to generate from the high-resolution view an inter-view reference picture with the same size as the low-resolution view, so that H.264/MPEG-4 AVC motion compensation can be applied for inter-view prediction. The results in [179] showed that the combination of MR and inter-view prediction yields the highest coding gains. Moreover, in [179] the coding performance has been evaluated according to the direction of the inter-view prediction. The authors have compared two directions of inter-view prediction: Prediction from the high-resolution view to the low-resolution view and vice versa. It was shown that for low bit-rates both prediction directions perform equally, while for medium and high bit-rates the prediction structure from high-resolution to low-resolution performs better.

More advanced and less complex strategies have been proposed by Chen *et al.* to combine MR approach with inter-view prediction [175, 180, 181]. The authors proposed an adaptive filter to generate a better-predicted signal in inter-view prediction. In their approaches, different filters are applied for different picture regions based on their distortions as well as the associated temporal and depth levels. Simulation results showed that these methods provide a bit-rate saving of 27% at maximum and 8% on average.

Usually, there are two ways to achieve MR coding. The first one is downsampling an image to a smaller size before encoding, and the second one is applying a blur filter (*e.g.*, disk filter, Gaussian filter, etc.), as illustrated in Figure 4.7. Spatially



(a)



(b)

Figure 4.7: Implementing MR coding method by (a) blur filter or (b) down/up sampling approaches.

downsampling an image reduces its size, since only specific pixels are kept depending on the used ratio. The MPEG-4 downsampling filter [182], with coefficients $[2, 0, -4, -3, 5, 19, 26, 19, 5, -3, -4, 0, 2]/64$, is often used. However, the blur filter removes the high frequency components, while the size of the image is unchanged. This allows to encode the image with less bits, because only few details (high frequency components) need to be encoded.

From the visual quality standpoint, both procedures have the same visual effect of blurring. However, downsampling and then upsampling back for 3D visualization, introduces more visual artifacts than using blur filters [176, 177]. From a compression point of view, as mentioned previously, combining inter-view prediction with downsampling concept requires an extra buffer size, in addition to design changes affecting low levels of encoding and decoding processes [174, 175, 179]. In contrast,

the blur filter is applied as a preprocessing step keeping the coding and decoding unchanged.

In general, it can be concluded that MR coding is a promising approach to decrease the bit-rate and yet provides close 3D visual quality to FR coding. However, in order to increase visual experience, it is important to effectively choose the type of low-pass filter and to define its appropriate strength.

4.2.4 Asymmetric luminance/chrominance

Inspired by prior works on visual perception, relating that HVS is far more sensitive to luminance than to chrominance and that, visual acuity is less for chrominance relative to luminance. Some works proposed to represent the chroma information of one view of the stereo pair by fewer samples than its respective in the other view [183]. It has been shown in [183], that applying a downsampling to chroma components of only one view of the stereo pair does not affect the perceived quality compared to the symmetric approach, provided that the downsampling ratio is properly selected.

Other works proposed to discard all the chrominance information of one view of the stereo pair before encoding [184, 185]. After decoding, the luminance-only frames are colored using chrominance information from the other view by applying color transfer techniques based on disparity information [184, 185]. The differences between the reconstructed and the original color images are insignificant, and HVS can ignore these differences due to binocular suppression. Thus, the bandwidth requirements can be relaxed while not affecting the final stereoscopic effect.

Finally, the previously described asymmetric stereoscopic coding methods can be combined, with the condition to keep one view completely unaltered. For instance, a combination of spatial resolution reduction, with chroma downsampling, and/or increasing the QP may lead to a better efficiency.

4.3 Perceptual limits of asymmetry

While subjective experiments showed that asymmetric stereoscopic video provides an acceptable visual experience, the amount of this asymmetry is still an open problem. Measuring the perceptual limits of asymmetric coding or quantifying the just noticeable threshold of asymmetry between views of the stereo pair, for which the S-3D effect is not affected, is an important issue. To the best of our knowledge, only few works have been dedicated to this topic [168, 170, 171, 184]. For instance, authors of [168] found that the just noticeable level of asymmetry depends on display technology and expressed this gap between both views in term of PSNR, *i.e.*, about 9 dB for a parallax barrier display and 7 dB for a full resolution projection display. In the same vein, Shao *et al.* [184] showed that 2 dB of difference between the left and right views, is the maximum tolerance level for which the viewer does not notice visual artifacts. Concerning the limit of asymmetric coding by spatial resolution

reduction, Aflaki *et al.* [170] made several trials with different downsampling ratios ($1/2$, $3/8$, and $1/4$) along both horizontal and vertical axes, and concluded that the range of downsampling ratios between $1/2$ and $3/8$ provides satisfactory 3D viewing experience.

The latest studies described above, use the aforementioned limit of asymmetry to perform effective asymmetric coding in such a way that the viewer cannot perceive any degradation. However, all the previous works are based on subjective experiments and the determination of the just noticeable level of asymmetry highly relies on them. Unfortunately, it is difficult to generalize the derived threshold to any other stereo pair, because the latter depends on the sequence content, the display type, the screen size and so on. A demonstration of this variability can be seen between [104] and [112], where, while both studies focused on asymmetric quantization following the same procedure, results gave very different threshold values.

Another issue lies in the fact of assigning a single threshold value to the entire image, which does not allow any adaptation to neither the content nor the original quality. Add to that, the techniques discussed earlier apply a uniform asymmetry by considering that all regions of an image have the same perceptual relevance, which is not in compliance with the characteristics of the HVS. Finally, none of the attempts have taken into account the characteristics of the view with higher quality, such as brightness or contrast. For example, a high contrast in one view may mask impairments in the other.

4.4 Proposed adaptive asymmetric stereoscopic coding

In this chapter, we propose a fully automated model that dynamically determines the best limits of asymmetry offering optimal 3D visual experience. The proposed method consists of adaptively selecting for each region of the image, the level of quality reduction. Our model is based on recent findings in visual perception, specifically: 1) the Binocular Just Noticeable Difference (BJND) [83] model (see Section 1.3.3), 2) the relationship between blur/sharpness and depth level, and 3) visual attention (see Section 1.3.1). We use the BJND model to determine the minimum distortions in one view that generate 3D perceptual difference, and combine depth information with visual saliency maps to adjust the resolution.

As stated before, when the left and right images are presented to the viewer with different levels of sharpness, *i.e.*, one image of the stereo pair is more blurry, the resulting 3D perception is close to the sharper image. On the other hand, the HVS tolerates a certain limited level of blur in one view that does not impair the 3D viewing experience. The principle of the proposed asymmetric stereoscopic coding approach relies on the fact that, the blur introduced in one view of a stereo pair should lead to a change below the binocular visibility threshold (BJND), thus avoiding visible artifacts. Moreover, as shown by subjective experiments described in [186], the sharpness visibility at different depth levels is different. The perceived blur of the object is highly dependent on the objects' distance (position in the

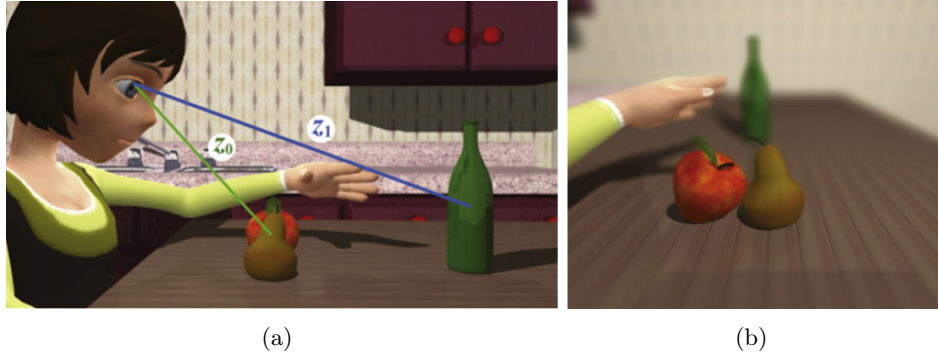


Figure 4.8: Relation between objects' distance, depth and sharpness [187].

scene), *i.e.*, the closer object appears sharper than deeper objects (see Figure 4.8). Accordingly, we use the depth information to control the strength of the introduced blur. In traditional methods of MR coding, one view is fully blurred at a fixed level of blur. However, the binocular visual properties are not spread equally across the stereoscopic images, because the response of the HVS to distortion in different regions is not uniform [244]. For example, the occluded regions cannot be binocularly compensated (masked), because they have no corresponding regions in the other view. Thus, the blur effect in such regions may be noticed. Therefore, we propose a novel MR coding method that enhances the concept of asymmetric coding relying on several aspects of the HVS.

In our model, in order to reduce the spatial resolution of one view of the stereo pair, we opted for the blur filter instead of downsampling an image to a smaller size. Because, using a blur filter allows assigning different degrees of blur to different regions of the image, thus providing more flexibility. More specifically, we use a circular kernel filter as blur kernel instead of the well-known Gaussian filter. Since to tune the strength of the blur filter, the disk filter depends on a single parameter, which is the diameter of the filter (kernel width), in contrast to Gaussian filter that depends on the standard deviation (σ) as well as the kernel width. Consequently, with a disk filter, we have fewer variables to optimize. Add to that, disk filter offers a radial symmetry as well as a flat response [176].

In the following, our three proposals are described, where the second and third are respectively a refinement and a temporal extension of the core proposal [159, 160].

– Proposition 1:

Given a stereo pair of images, denoted by I_l for the left view and I_r for the right. In the following, we consider the right image as the sharper image and it is not altered, while the left image is the image to be blurred. One may relate this choice to the notion of eye dominance. Because, approximately 70% of the world population is right eye dominant [188], while the rest are left eye dominant. This assures that if eye dominance has an effect on the perceived quality of stereoscopic asymmetric

image/video, it is better to provide the left view with a lower quality, since the majority of viewers are right eye dominant. However, according to [166] and [172], there is no significant difference between the left-eye and right-eye dominant subjects on binocular quality perception of asymmetric stereo images. Thus, the dominant eye is not considered in our model and the choice of the blurred view is arbitrary.

The aim is to reduce the spatial resolution of the left image, in order to decrease the bitrate without causing visible artifacts in 3D viewing. To achieve this, we blur I_l by a disk filter F defined as follows:

$$\tilde{I}_l^{(R_{i,j})}(i, j) = I_l(i, j) * F_{R_{i,j}} \quad (4.1)$$

$$\begin{aligned} \text{with } R_{i,j} &= \arg \max_{k>0} g_k(i, j) \\ &= \{k \mid |I_l(i, j) - \tilde{I}_l^{(k)}(i, j)| < \delta_{i,j} \cdot VT_{i,j}\} \end{aligned}$$

where $*$ is a convolution product, and $I_l(i, j)$ denotes the luminance value of the pixel at position (i, j) . $R_{i,j}$ is the diameter of the disk filter $F_{R_{i,j}}$ that determines the strength of blur for each pixel (i, j) , and $\tilde{I}_l^{(k)}$ is the output of the filtered left image with a k diameter disk filter. $\delta_{i,j}$ is the weighting factor related to the stereo matching result (disparity map), the definition of this factor will be discussed later. The change introduced by the blur shall not exceed the visual threshold $VT_{i,j}$ defined by equation (4.2):

$$VT_{i,j} = \frac{d_{max}}{d_l(i, j)} \cdot \text{BJND}_l(i, j) \quad (4.2)$$

where $d_l(i, j)$ is the disparity value of pixel (i, j) , and d_{max} is the maximum value of the disparity map. $VT_{i,j}$ depends on two factors: the $\text{BJND}_l(i, j)$ which represents the Binocular-JND of the left image at pixel position (i, j) , and the depth information.

It is known that the disparity is inversely proportional to depth (see Section 1.2.2.1 for more details), meaning that objects with greater disparity are closer, and vice versa. Accordingly, $VT_{i,j}$ of closer pixels ($d_l(i, j) \cong d_{max}$) are mainly determined by BJND_l , unlike to deeper pixels which are more blurred, because $VT_{i,j}$ is higher. This process is applied only to the luminance component, to all pixels of the left view, except those with zero disparity (*e.g.*, occluded pixels or bad matches). For the latter, they are kept unaltered. In our implementation, to estimate the disparity map, we use the stereo matching algorithm described in [189].

Since $VT_{i,j}$ is dependent on two factors which in turn are based on the disparity map, and as in practice the stereo matching algorithms can cause some bad matches, we introduce a weighting factor $\delta_{i,j}$ to control the influence of $VT_{i,j}$ according to the reliability of the stereo matching result. To achieve this, we use a dissimilarity function that assesses the stereo matching result in each pixel as follows:

$$\delta_{i,j} = a_1 \exp(-a_2 \cdot |ZNCC(i, j, d_l)|) \quad (4.3)$$

where a_1 and a_2 are two constants determined experimentally, and $ZNCC$ (Zero mean Normalized Cross Correlation) is window-based dissimilarity measures defined by equation (4.4). In this latter equation, W denotes a 5×5 window centered at pixel (i, j) , and $\bar{I}_*(i, j)$ is the mean value computed over all pixels inside W . The $ZNCC$ ranges from $[-1, 1]$, and it is used in equation (4.3) to exponentially decrease the influence of $VT_{i,j}$ according to the quality of the disparity map. On the other hand, if the disparity value is not reliable, then $ZNCC$ is high, and consequently, the value of the pixel is not significantly changed by the proposed process.

$$ZNCC(i, j, d_l) = \frac{\sum_{(i,j) \in W} (I_l(i, j) - \bar{I}_l(i, j)) \cdot (I_r(i - d_l, j) - \bar{I}_r(i - d_l, j))}{\sqrt{\sum_{(i,j) \in W} (I_l(i, j) - \bar{I}_l(i, j))^2 \cdot \sum_{(i,j) \in W} (I_r(i - d_l, j) - \bar{I}_r(i - d_l, j))^2}} \quad (4.4)$$

– Proposition 2:

In the previous proposition, the process is applied only to the non-occluded pixels, and the remaining occluded pixels are kept unaltered. One of the main problems in stereo matching is the occluded and unknown regions. After matching the stereo pair, the resulting disparity map identifies the missing information by zero values.

In order to enhance the previous described model, we start first by preprocessing the generated disparity map using a recursive bilateral filtering [190], in order to assign valid values to pixels with zero disparity (*e.g.*, occluded pixels or bad matches). Figure 4.12 shows an example of disparity map before and after the application of the recursive bilateral filtering.

Then, in order to assign properly the strength of blur to each region, we consider additional perceptual features such as visual saliency (see Section 1.3.1). Thus, we combine depth information with visual saliency to control more efficiently the blur level. This point has been integrated in equation (4.2) as follows:

$$VT_{i,j} = BJND_l(i, j) \cdot \left(\frac{d_{max}}{d_l(i, j)} \right)^{(1-S_l(i,j))} \quad (4.5)$$

where $S_l(i, j)$ represents the saliency value of pixel (i, j) . For the generation of the saliency map S_l , we used the Graph Based Visual Saliency (GBVS) model described in [191] due to its efficiency and good results [192]. Each $S_l(i, j)$ value ranges from 0 (un-salient pixel) to 1 (most salient pixel).

Therefore, in this second proposition we applied a preprocessing to disparity map and integrated saliency information for tuning the blur level. Thus, all the remaining process described by equations 4.1 to 4.4 (except equation (4.2)) remains the same as explained previously (Proposition 1).

Consequently, following this concept, deeper pixels are more blurred, provided that they are not salient, because $VT_{i,j}$ is higher. In contrast, if the region has a

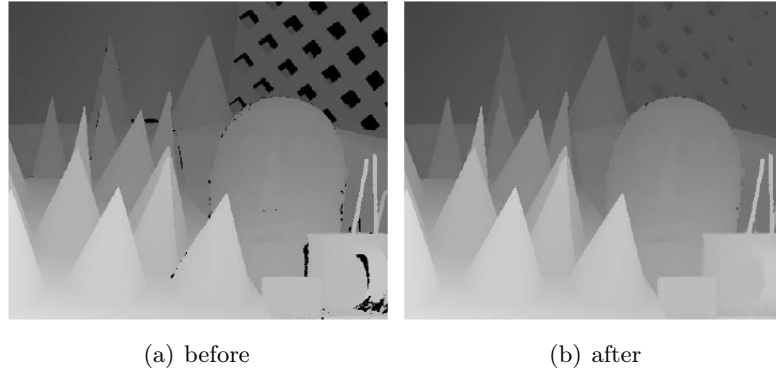


Figure 4.9: Disparity map before and after the recursive bilateral filtering.

high depth level and represents a perceptual importance (most salient), the effect of blur is somewhat mitigated by S_l factor.

– Temporal extension:

After performing a deep experimentation described in Section 4.5, including subjective and coding performance testing. We noticed that proposition 2 outperforms the first one in terms of both coding and subjective results. Based on these results, we tried to extend proposition 2 to stereoscopic video in order to tackle the problem of asymmetric stereoscopic video coding.

An obvious solution to extend the original algorithm to stereoscopic video is to apply it to the whole sequence on a frame-by-frame basis. However, this can lead to variations between subsequent time steps, which may deteriorate the coding performance and produces a noticeable flicker distortion affecting the QoE. In addition, since for each time step the whole process is repeated, this will increase the complexity making difficult to apply this approach in real time applications.

Therefore, in order to fix this issue, the input video is firstly segmented temporally into shots to reduce operations on similar frames. Then, the proposed adaptive asymmetric stereoscopic coding method is applied on selected key frames (the first frame of each shot). Next, within each video shot, the obtained key frames results are propagated from key frames to non-key frames.

The goal of temporal video segmentation is to divide the video stream into a set of meaningful and manageable segments (shots) [193]. Each shot is then represented by selecting key frames. There are two basic types of shot transitions: *abrupt* and *gradual*. Abrupt transitions (cuts) are simpler, they occur in a single frame when stopping and restarting the camera. Gradual transitions are more difficult to detect than cuts. They must be distinguished from camera operations and object movement that exhibit temporal variation of the same order and cause false positives [193]. Algorithms for shot boundaries detection were already discussed in several review papers [193–195]. Based on the metrics used to detect the difference between

successive frames, the algorithms can be divided broadly into three categories: pixel-, block- and histogram-based comparisons.

In order to reduce sensitivity to camera and object movements, in our method we used histogram comparison between successive frames. The idea behind histogram-based approaches is that two frames with unchanging background and objects will have little difference in their histograms. In addition, histograms are invariant to image rotation and change slowly under the variations of viewing angle and scale. More specifically, we use local histogram comparison instead of the global one, because the latter ignores the spatial information and, therefore, fails when two different images have similar histograms.

The adopted temporal video segmentation algorithm is performed in three steps as follows: (1) histogram calculation, (2) dissimilarity computation between frames and (3) shot boundary detection.

For the first step, color histogram in the RGB color space is computed from each frame of the sequence. In order to reduce the bin number (3 color channels $\times$ 8 bits create histograms with 2^{24} bins), only the upper two bits of each color intensity are used to compose the color code. We considered only 64 bins, which has given a sufficient accuracy.

A color histogram of a frame i is an n -dimensional vector $H_i(j)$, $j = 1, 2, \dots, n$, where j denotes a code value derived from the three color intensities of a pixel, and n is the number of color levels. Finally, $H(j)$ represents the number of pixels from the frame i with color level j .

Then, in the second step, the Euclidean distance is used as dissimilarity metric to detect the difference between successive frames. Before computing this difference between color histograms, each frame i is divided into b blocks that are compared with their corresponding blocks in following frame $i + 1$ as follows:

$$D(i, i + 1) = \sum_{k=1}^b DP(i, i + 1, K),$$

$$\text{with } DP(i, i + 1, k) = \sum_{j=1}^n \sqrt{(H_i(j, k) - H_{i+1}(j, k))^2}, \quad (4.6)$$

where k represents a block, and b is the total number of blocks.

Finally, in the third step, the shot boundary detection is determined by comparing the histogram distance ($D(i, i + 1)$) against a threshold T . An adaptive threshold at frame i is computed as follows:

$$T(i) = n \times \frac{\sum_{t=2}^i D(t, t - 1)}{i - 1}, \quad (4.7)$$

where t is the index of the t^{th} frame of one shot. It is an averaged dissimilarity up to frame i multiplying with a factor n , which is introduced to avoid under-segmentation (in our implementation $n=3$). A shot boundary is marked at i , when $D(i, i - 1) > T(i)$.

Once all the shots have been defined, this implies that the key- and non-key-frames have been determined. Then, we apply the process described in Proposition 2 on the key-frames. Following that, the same computed parameters for key-frame are applied to non-key frames of the same shot.

4.5 Experimental results

In this section, simulation results are provided to evaluate the performance of the proposed asymmetric stereoscopic coding method. The evaluation is organized in two parts. The first one is related to the first two propositions, and the second one is devoted to assess the proposed asymmetric stereoscopic video coding (temporal extension) method.

4.5.1 Experience 1

In order to evaluate Proposition 1 and 2, simulations have been carried out on the *Middleburry* stereo dataset [196]. The used test images were: *Art*, *Book*, *Cloth3*, *Cones*, *Midd1*, *Midd2*, *Moebius*, *Reindeer* and *Teddy*. The objective and subjective performance of the two proposed adaptive asymmetric stereoscopic coding methods and symmetric stereoscopic coding method (denoted as SSC) as well as traditional uniform MR coding (UMR) methods were compared.

Specifically, we considered two traditional uniform asymmetric stereoscopic coding methods. The first one achieves the MR coding by applying a downsampling ratio of $1/2$ along both coordinate axes (denoted as UMR-D), and the second consists in applying disk filter with a fixed diameter of 2 (denoted as UMR-B). The values of $1/2$ and 2 have been demonstrated experimentally in [170] and [176], respectively, so that they provide satisfactory 3D quality.

First, in Figures 4.10 and 4.11 we show examples of blurred images obtained from the different compared methods and also the computed blur maps. The blur maps illustrated in these figures represent the level of blur of each region of the image. We can see that the proposed methods affect adaptively the degree of blur according to the perceptual importance of each region, in contrast to the UMR methods providing uniform blur on the whole image.

4.5.1.1 Effect on coding performance

In order to assess the proposed methods in terms of coding performance, we used the JMVM software [99] with different QP values ($QP = 22, 27, 32, 37$). Usually, JMVM is used for coding MVV but in our case it is used to enable the inter-view prediction.

The right image was coded as I frame (reference image) and the left image (target image) was predicted from the reference one, usually this type of image is referred to as P frame. To allow the inter-view prediction in UMR-D method, we performed the downsampling (by $1/2$) of the right image at the reference picture buffer level.

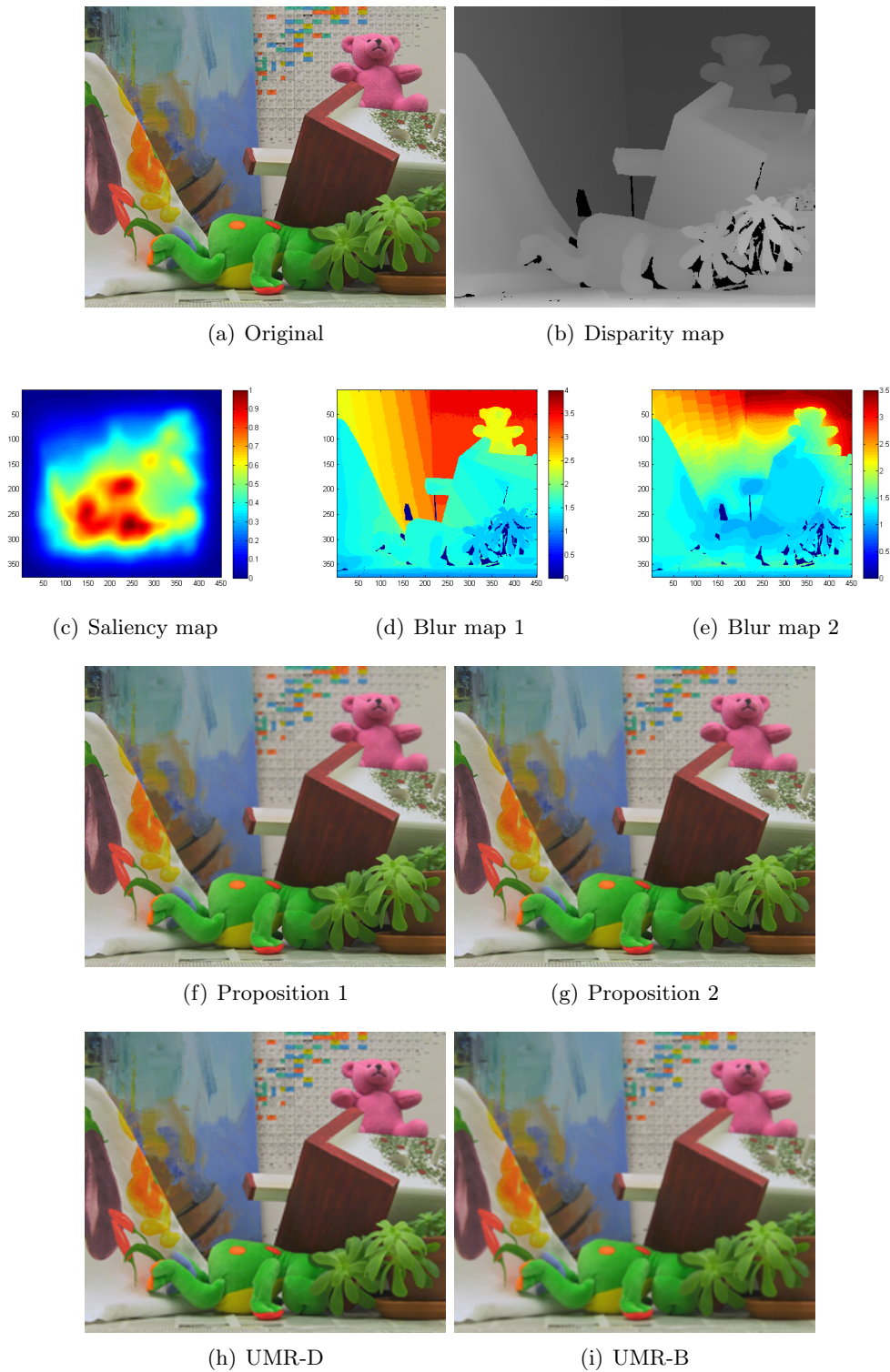
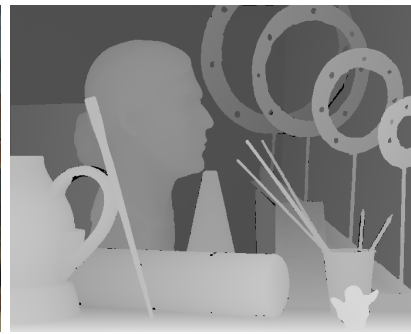


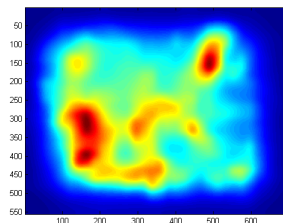
Figure 4.10: The obtained visual results for the left image of *Teddy* stereo pair.



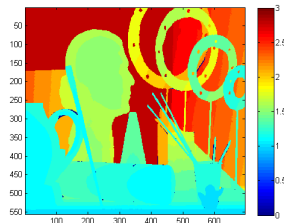
(a) Original



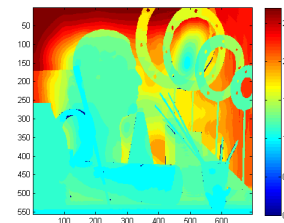
(b) Disparity map



(c) Saliency map



(d) Blur map



(e) Blur map 2



(f) Proposition 1



(g) Proposition 2



(h) UMR-D



(i) UMR-B

Figure 4.11: The obtained visual results for the left image of *Art* stereo pair.

Table 4.1: Bit-rate saving and average PSNR (dB) comparisons measured against symmetric coding method using Bjøntegaard measurement.

Method	UMR-B		UMR-D		Proposition 1		Proposition 2	
Image	Δ Bitrate saving	Δ PSNR	Δ Bitrate saving	Δ PSNR	Δ Bitrate saving	Δ PSNR	Δ Bitrate saving	Δ PSNR
<i>Art</i>	-35.69%	2.95	-33.90%	2.75	-36.75%	3.04	-37.07%	3.11
<i>Books</i>	-38.70%	2.79	-36.37%	2.50	-38.04%	2.93	-38.69%	2.98
<i>Cloth3</i>	-40.87%	3.94	-40.78%	3.52	-40.92%	4.22	-41.04%	4.81
<i>Cones</i>	-39.34%	2.27	-38.85%	2.12	-39.88%	2.32	-40.04%	2.42
<i>Midd1</i>	-32.27%	2.04	-31.69%	2.15	-34.51%	2.29	-34.12%	2.36
<i>Midd2</i>	-31.06%	2.18	-31.15%	2.23	-34.12	2.40	-35.03%	2.45
<i>Moebius</i>	-34.61%	2.01	-23.38%	1.25	-34.62%	2.03	-34.88%	2.10
<i>Reindeer</i>	-30.09%	2.48	-29.93%	2.15	-32.11%	2.62	-33.13%	2.41
<i>Teddy</i>	-38.46%	2.75	-38.90%	2.67	-39.97	2.77	-41.12%	2.82
<i>Average</i>	-35.67%	2.60	-33.88%	2.37	-36.76%	2.73	-37.23	2.82

Finally, after decoding, the downsampled left image is upsampled for PSNR calculation. Table 4.1 lists the bit-rate saving and average PSNR gain generated using the Bjøntegaard measurement [197] for the tested stereoscopic images. It also provides the average values for the whole test-set. Note that a bit-rate saving compared with a reference method corresponds to negative values.

From the obtained results, the bit-rate saving achieved by the Proposition 1 ranges from about 32% to 40% and from 33% to 41% for the Proposition 2, compared with SSC method. Also, UMR-B and UMR-D with lower performance than the two proposed methods, provide up to 39.34% and 40.78% bitrate saving, respectively. In terms of average PSNR, Table 4.1 shows that the gains are consistent and that both proposed methods outperform the others for almost all images. In addition to the Bjøntegaard measurement, Figure 4.12 shows the rate/distortion curves for some coded stereo images pairs. It clearly confirms the previous conclusions since our proposals provide the best results. However, it is important to notice that Proposition 2 has a relative advantage compared to Proposition 1.

4.5.1.2 Effect on perceived quality

In order to confirm the good results obtained with our approaches, we performed subjective test so as to quantify the 3D visual experience of the viewers. Therefore, the test was performed in the XLIM lab test-room, calibrated following the ITU-R BT.500 recommendations [198], using a Hyundai S456D stereoscopic 3DTV. The display was placed 0.5 m from the back wall and 4H from the viewer (H being the height of the image). Each viewer adjusted the height of their chair so that the position of his/her eyes were at about the same as the height of the center of the display. The room lighting was controlled and the display has been calibrated using the ildisplay2 spectrometer from Gretag Macbeth®.

Thirteen naive observers with age raging from 23 to 37 participated to the test. They have been pre-screened for visual acuity using fr@ct, color blindness using

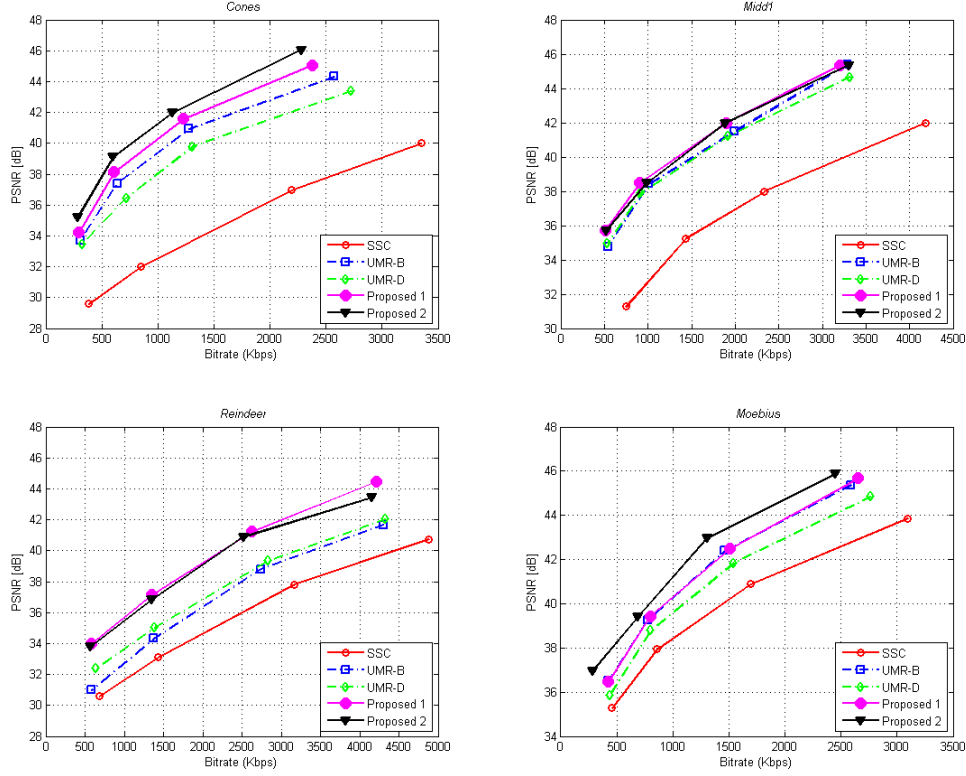


Figure 4.12: Rate/distortion curves of the *Cones*, *Midd1*, *Moebius*, and *Reindeer* stereo pairs.

Ishihara test and depth blindness using Randot Stereo test.

A set of 36 stereoscopic images, corresponding to 9 original pairs with associated results obtained using UMR-B, UMR-D and our proposals, was randomized and presented sequentially. We opted for an ACR-5 (Absolute Category Rating) method with hidden reference [125] for these tests. The aim being to be close to real condition when someone use his own display to view a stereo-pair without any cue about the original. Observers were asked to rate each S-3D image on a quality scale composed of 5 categories (Bad, Poor, Fair, Good, Excellent). Before the experiment starts, observers were provided with instructions explaining the task before going to the training session.

In order to account for the influence of the hidden reference, the final scores DMOS (Difference Mean Opinion Score) are calculated using the following equation:

$$DMOS_{test} = MOS_{test} - MOS_{Original} + 5 \quad (4.8)$$

where MOS_{test} and $MOS_{Original}$ are respectively the score of a given stereo pair for one of the used methods and the score of the original image.

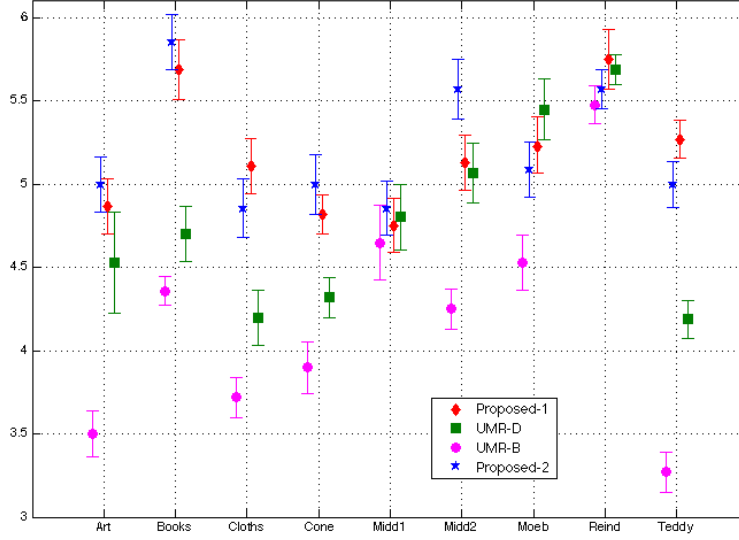


Figure 4.13: DMOS values for each image of the test-set using UMR-B, UMR-D and proposed approaches.

Table 4.2: Bit-rate saving and average PSNR (dB) comparisons between the proposed asymmetric stereoscopic video coding approach and symmetric coding method using Bjøntegaard measurement.

Sequences	UMR-B		UMR-D		Proposed	
	Bitrate saving	PSNR	Bitrate saving	PSNR	Bitrate saving	PSNR
<i>Balloons</i>	-10.13%	0.52	-12.09%	0.64	-14.17%	0.75
<i>Kendo</i>	-12.02%	0.9	-9.36%	0.51	-11.83%	0.89
<i>Undo Dancer</i>	-17.27%	0.83	-8.08%	0.39	-21.02%	1.04
<i>GT Fly</i>	-6.91%	0.34	-11.38%	0.45	-18.38%	0.9
<i>Poznan Hall2</i>	-14.09%	0.75	-6.74%	0.42	-17.3%	0.95
<i>Poznan Street</i>	-1.36%	0.28	-5.27%	0.12	-8.88%	0.32
<i>Pantomime</i>	-12.84%	0.78	-11.02%	0.66	-20.64%	1.28
<i>Champagne Tower</i>	-12.89%	0.80	-16.78%	1.06	-26.11%	1.74

Figure 4.13 gives the DMOS for all images from the test-set. Values above 5 indicate that observers scored the image better than the original one. One can notice that the proposed approaches obtain very good results almost for all images with a gap close to 1 MOS value for some images. Exception is made for *Moebius* and *Midd1*, the difference is very slight or negligible for the latter. This confirms the results obtained from the objective study.

4.5.2 Experience 2

To evaluate the rate-distortion performance of the proposed asymmetric stereoscopic video coding method, eight MPEG standard MVV sequences are chosen, namely *Balloons*, *Kendo*, *Undo Dancer*, *GT Fly*, *Poznan Hall2*, *Poznan Street*, *Pantomime*, and *Champagne Tower*.

Similarly to the previous coding evaluation described in Section 4.5.1.1, the

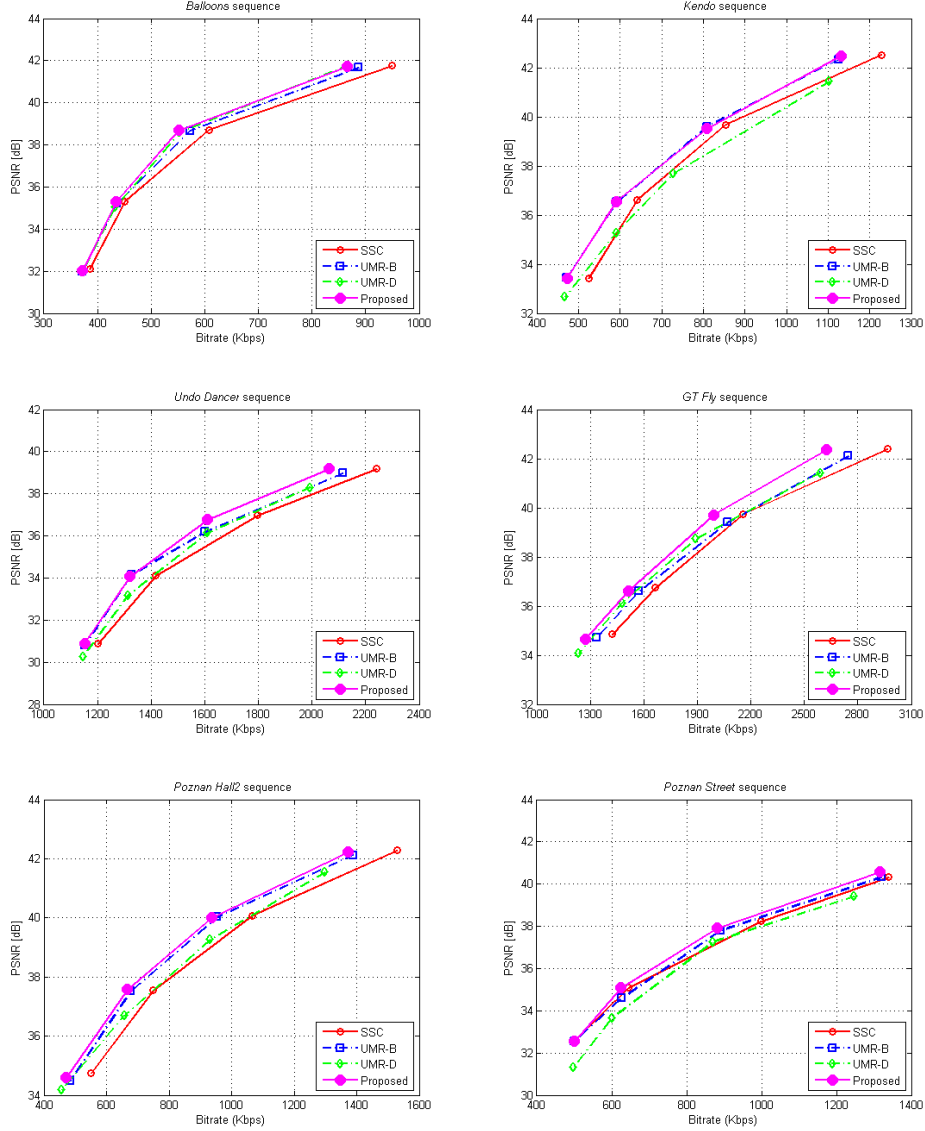


Figure 4.14: Rate/distortion curves of the coded stereoscopic videos.

experiments were run using the JMVM software with typical settings for MVC, for details regarding the test conditions the reader can refer to [149]. The right view was coded as reference view and the left view was predicted from the reference one. The same methods (SSC, UMR-B and UMR-D) have been evaluated and compared with the proposal.

To allow the inter-view prediction in UMR-D method, we performed the down-sampling (by 1/2) of the right view along both coordinates axes at the reference picture buffer level. This downsampling has been achieved using the MPEG-4 down-sampling filter [182], with coefficients $[2, 0, -4, -3, 5, 19, 26, 19, 5, -3, -4, 0, 2]/64$. Finally, after decoding, the downsampled left view is upsampled for PSNR calcula-

tion. Table 4.2 lists the bit-rate saving and average PSNR gain generated using the Bjøntegaard measurement [197] for the coded stereoscopic video.

In addition to the Bjøntegaard measurement, Figure 4.14 shows the results in terms of rate-distortion performances. The curves consist of 4 QP points which are 22, 27, 32, 37. According to these curves, it is clear that the proposed method provides significantly better rate-distortion performance than the other methods, which confirms the obtained results with Proposition 1 and 2.

4.6 Conclusion

In this chapter, novel asymmetric stereoscopic coding methods are proposed. The proposed approaches enhance the MR coding concept by introducing a non-uniform reduction of spatial resolution allowing decreasing the bandwidth required for S-3D delivery. Based on three HVS-inspired models, the proposed approaches determine the best level of blur to be introduced for each region of the image, making the blur effect almost transparent to the viewer, thus offering optimal 3D visual experience. Another contribution described in this chapter, is the extension of the proposed asymmetric coding method to stereoscopic video. Thanks to the adopted temporal video segmentation, we have been able to apply the proposed method to stereoscopic video, while maintaining the temporal correlation between successive frames and keeping the complexity of the proposed process very low.

Objective and subjective evaluation results demonstrated that the proposed methods can provide significant bit-rate saving without noticeable visual quality losses. From a visual quality standpoint, the produced 3D visual quality shows a gap, in favor of our approaches, compared to the state-of-the-art methods. Also, the proposed approaches can be considered as a preprocessing so as they do not require any overhead bits nor architecture changes at the encoder and decoder sides.

Even though the proposed asymmetric stereoscopic video coding method is providing satisfactory results, we believe that including perceptual temporal features is a promising step towards the improvement of the proposed method and consequently the quality of experience.

Chapter 5

Stereoscopic 3D Quality Assessment

Contents

5.1	Introduction	103
5.2	Visual quality assessment	105
5.2.1	Subjective quality assessment	105
5.2.2	Objective quality assessment	107
5.2.3	Performance evaluation of objective quality metrics	111
5.3	Stereoscopic image quality assessment (SIQA): State-of-the-art	112
5.3.1	SIQA based on 2D metrics	112
5.3.2	SIQA including depth/disparity information	114
5.3.3	SIQA considering binocular perception	114
5.4	Proposed quality assessment metric for stereoscopic images	116
5.4.1	Binocular perception model	116
5.4.2	Proposed FR IQA metric for stereoscopic images	118
5.4.3	Proposed cyclopean image quality approach for stereoscopic images	120
5.4.4	Proposed NR perceptual blur metric for stereoscopic images	121
5.5	Experimental results	124
5.6	Conclusion	127

5.1 Introduction

The introduction of 3D imaging technologies in different areas such as education, medicine, cinema, communication, gaming, has exponentially increased over the recent years. An evidence for this consumer interest is the significant increase in the

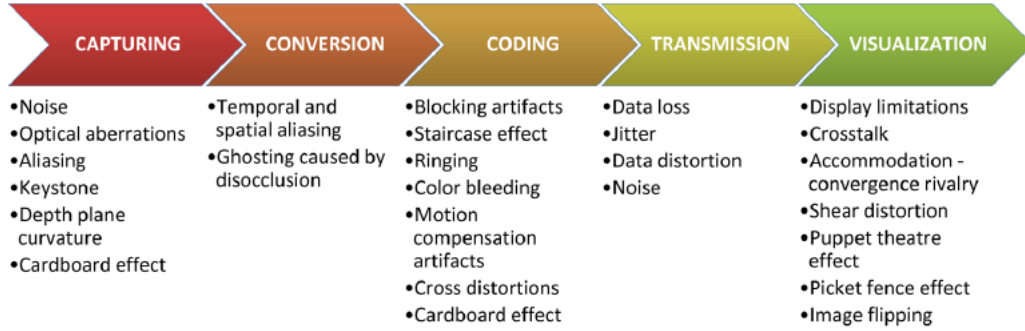


Figure 5.1: 3D video processing chain artifacts [200].

production of 3D contents. According to the latest theatrical market statistics gathered by the Motion Picture Association of America (MPAA) [199], the proportion of cinema screens that are 3D has reached 35% worldwide, and approximately half of all cinema-goers viewed at least one 3D movie in 2012. However, as these huge amounts of 3D content must be handled through the typical multimedia processing chain (*e.g.*, acquisition, editing, compression, transmission, reconstruction, restoration, etc.), distortions are inevitably introduced, which will contribute to the overall 3D perceptual quality. At each stage of this processing chain, different artifacts can be identified, as listed non-exhaustively on Figure 5.1. Thus, at every processing stage, it is desirable to identify and quantify the decrease of image quality, and finally to guarantee the final 3D viewing experience.

Moreover, to avoid reproducing the previous failed attempts to accept the 3D imaging systems, high-quality 3D visual content must be achieved. In this context, 3D quality assessment methods measuring the 3D visual experience are of a paramount importance. The latter are a central issue in the success of 3D technologies, because, it ensures the measurement and monitoring of 3D quality of experience (3D QoE) as perceived by the end-user.

As the human eyes are the ultimate receiver judging the quality, the human subjective opinion remains the most reliable measure reflecting the perceived 3D image quality. This user-perceived quality is obtained by means of subjective tests, where a large number of viewers provide their personal opinion on perceived quality. However, such experiments are time consuming and costly, not feasible for real-time applications, and depend strongly on the observer context (personal experience, emotional state, etc.) [201]. Therefore, reliable automated methods or algorithms are really needed to measure the perceived quality of 3D content in line with the HVS. Automated evaluation of visual quality using an algorithm is referred to as objective quality assessment.

Compared to 2D quality assessment (2D-QA), which has been studied extensively [202], 3D quality assessment (3D-QA) is still an open field and there is no commonly accepted metric that ensures reliable 3D quality evaluation *i.e.*, that correlates with human judgment. The reason for this difficulty lies mainly in the

lack of understanding of 3D human perception. The latter is very complex and requires a multidisciplinary approach. Furthermore, the nature of 3D images is also complex and contains several depth cues and perceptual attributes. All these problems represent a real challenge to the quality research community. Consequently, in response, this chapter presents three different quality assessment metrics for stereoscopic images based on perceptual binocular characteristics. In order to be as close as possible to the HVS behavior, the proposed metrics are designed so that they take into account the binocular visual phenomena.

The remainder of this chapter is organized as follows. Section 5.2 addresses the two major approaches of visual quality assessment, *i.e.*, subjective and objective. Next, a state-of-the-art of stereoscopic image and video quality metrics covering the major contributions to 3D objective quality assessment is presented in Section 5.3. Thereafter, the proposed approaches are described in Section 5.4, starting by defining the proposed binocular model, then detailing the three proposed stereoscopic image quality metrics. Section 5.5 is dedicated to experimental evaluation and performance analysis of the proposed metrics. Finally, in Section 5.6, the chapter is concluded with some remarks and future works.

5.2 Visual quality assessment

Globally, there are two main ways to measure image/video quality. The first is subjective quality assessment. This approach is a psychologically based method requiring structured experimental designs and involves human participants to evaluate the visual quality. The second is objective quality assessment, which corresponds to computational models that can predict perceived image quality accurately and automatically. In the following, these two approaches are consistently described.

5.2.1 Subjective quality assessment

Typically, in visual multimedia applications, the human is the ultimate receiver. Accordingly, subjective evaluation is the best way for assessing the quality of an image. In a subjective test, a large number of viewers (referred to as subjects) are asked to judge the perceived quality of the visual signal (referred to as stimuli) according to some predefined scales. Subsequently, these obtained subjective scores are processed across human viewers, providing a quality score, to referred as the mean opinion score (MOS) or differential mean opinion score (DMOS).

In order to ensure reliable and statistically significant results, the subjective tests should be run following the standards, which formalize the test set-up, the scoring procedure and so on. For instance, formal subjective testing has been described in the ITU recommendations [204] and ISO standards [203]. Table 5.1 specifies the common features and methods for subjective quality assessment described in ITU-R BT.500 [204]. Common features specify the conditions for conducting the subjective test. While, the assessment method refers to the protocol used to perform the evaluation task.

Table 5.1: Specification of subjective quality assessment in ITU-R BT.500.

Common features	General viewing condition
	Source signals
	Selection of test materials
	Range of conditions and anchoring
	Observers
	Instruction for the assessment
	The test session
	Presentation of the results
Assessment method	Particular method should be used to address particular assessment problems

For S-3DTV, ITU published the ITU-R BT.1438 recommendation [205], which is guidelines for subjective assessment of stereoscopic television pictures. The latter are mainly based on the ITU-R BT.500-11 for the assessment of two-dimensional television pictures. In this standard, three main categories of stimulus presentation and rating procedure are defined: (i) *double stimulus* (DS) methods (sequential presentation of each pair of a reference and a test stimulus), (ii) *single stimulus* (SS) methods (presentation and rating of the test stimulus only), and (iii) *stimulus comparison* (SC) methods (presentation of pairs of stimulus and rating of relative quality). The rating scale can be either continuous or discrete, and either numerical or categorical. Finally, the rating can be performed after each stimulus presentation for assessing the overall quality, or continuously during presentation for assessing temporal quality variations [200]. The recommended rating scales for these different methods are shown in Table 5.2.

Table 5.2: Recommended rating scales to quantify the perceived quality [206].

DSCQS Continuous Quality Scale			Comparison Scale of Stimulus–Comparison			
	Excellent	A	B	–3	Much worse	
					–2	Worse
	Good				–1	Slightly worse
					0	The same
	Fair				+1	Slightly better
					+2	Better
	Poor				+3	Much better
Bad						

Single Stimulus Quality Scale		DSIS and Single Stimulus Impairment Scale	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

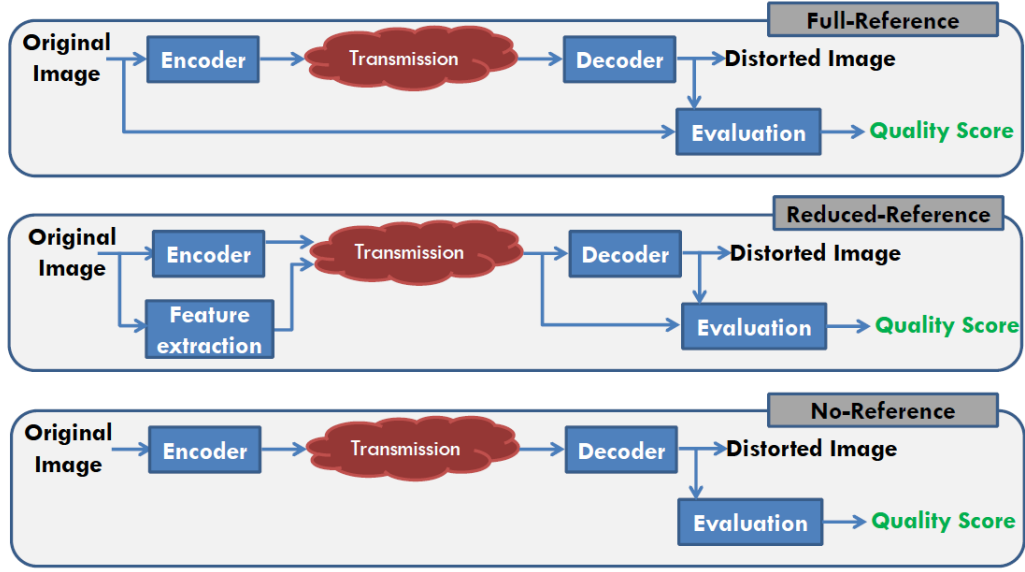


Figure 5.2: The three main classes of objective quality metrics.

The subjective evaluation remains the best and ideal way to assess perceived visual quality. However, this procedure is expensive, time-consuming, tedious, and finally does not allow any integration to real time applications. As a result, the ambition is to develop an automated method predicting the perceived quality of visual stimuli in a consistent and objective manner.

5.2.2 Objective quality assessment

The aim of objective quality assessment metrics is to predict a quality score of an image/video as close as possible to an average human observer will report [207]. In general, objective quality metrics can be divided into three categories based on the amount of reference information required by the metric to compute the quality score (see Figure 5.2):

- **Full-reference (FR)**

The FR metric consists in comparing a distorted image to the reference, thus needing the availability of the complete original/reference image. Among the various metrics, the most simple and broadly used FR metrics are the mean square error (MSE) and the related peak signal-to-noise ratio (PSNR). They are based on a pixel-wise comparison of two images I and $\tilde{I}$, as follows:

$$\begin{aligned}
 MSE &= \frac{1}{XY} \sum_i \sum_j \left[I(i, j) - \tilde{I}(i, j) \right]^2 \\
 PSNR &= 10 \log_{10} \frac{m^2}{MSE}
 \end{aligned} \tag{5.1}$$

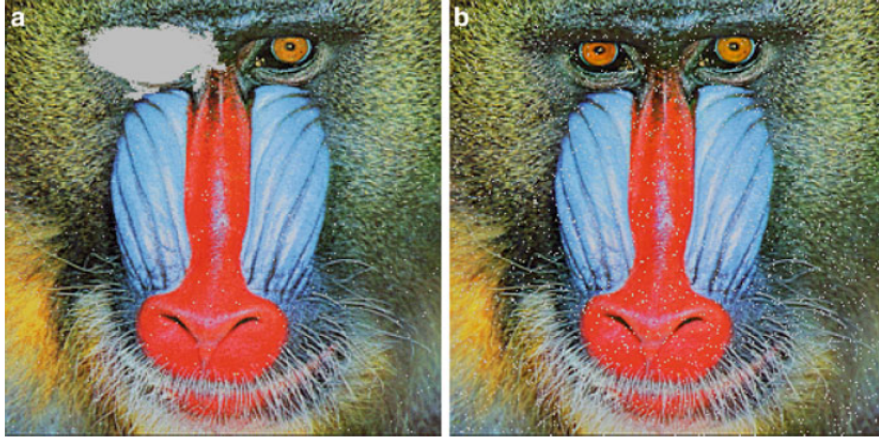


Figure 5.3: Example of the upper limit of the PSNR (a) Addition of a 1,700 pixels region and a Gaussian noise on 300 pixels, (b) Addition of a Gaussian noise on 2,000 pixels.

where $X \times Y$ is the size of the image, and m is the maximum value that a pixel can take (*e.g.*, 255 for 8-bit images). These measurements are popular and often used for image quality assessment because of their low complexity as well as their simple of implementation. However, they provide poor performance for the prediction of perceptual quality and do not correlate with the subjective ratings [208–210]. Because they are based on a pixel-based error measure, by equally weighting all the image areas, without taking into account the properties of HVS, such as masking effect. Figure 5.3 illustrates an example of the upper limits of PSNR interpretation. Both images have the same PSNR = 11.06 dB, while, the visual quality is obviously not the same.

To overcome these problems, some initiatives have been attempted to integrate the HVS features into these metrics [211, 214]. Therefore, to develop reliable quality metrics, it is important to take advantage of the HVS behavior.

Because of the abundance of quality metrics proposed in the literature, in the following, only the most relevant to the proposed work are described.

SSIM Structural Similarity Index [145]. This metric is one of the most popular objective quality metrics for 2D images. It is considered as an extension of UQI (universal image quality index) [212]. Under the assumption that human visual perception is highly adapted for extracting structural information, SSIM measures structural information degradation between original and distorted images. To achieve this, the task of similarity measurement is separated into three comparisons: luminance (l), contrast (c) and structure (s). They are computed as follows:

$$l(I, \tilde{I}) = \frac{2\mu_I\mu_{\tilde{I}} + C_1}{\mu_I^2 + \mu_{\tilde{I}}^2 + C_1}, \quad (5.2)$$

$$c(I, \tilde{I}) = \frac{2\sigma_I\sigma_{\tilde{I}} + C_2}{\sigma_I^2 + \sigma_{\tilde{I}}^2 + C_2}, \quad (5.3)$$

$$s(I, \tilde{I}) = \frac{\sigma_{I\tilde{I}} + C_3}{\sigma_I + \sigma_{\tilde{I}} + C_3}. \quad (5.4)$$

where μ and σ represent the mean and the standard deviation, respectively. The constants C_* are included to avoid instability.

Finally, the three comparisons are combined and the final SSIM index is computed as follows:

$$SSIM(I, \tilde{I}) = [l(I, \tilde{I})]^\alpha \cdot [c(I, \tilde{I})]^\beta \cdot [s(I, \tilde{I})]^\gamma \quad (5.5)$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components.

MS-SSIM Multi-Scale Structural Similarity Index [258]. In order to consider the variations of viewing conditions (viewer's distance), the authors in [258] extend the SSIM metric by adopting multi-scale structural similarity approach, which supplies more flexibility than the single-scale method.

The MS-SSIM score is obtained by combining the aforementioned features, at different scales j , as described by equation (5.6).

$$MS-SSIM_{I, \tilde{I}} = [l_M(I, \tilde{I})]^{\alpha_M} \prod_{j=1}^M [c_j(I, \tilde{I})]^{\beta_j} [s_j(I, \tilde{I})]^{\gamma_j} \quad (5.6)$$

VSNR Visual Signal-to-Noise Ratio [211]. This metric is also a FR perception-oriented metric. It exploits the just noticeable difference (JND) principle based on contrast in order to determine whether the distortions in the impaired image are visible. This metric operates via a two-stage approach. In the first stage, the metric verifies if the difference between the reference and processed images is below contrast thresholds (JND visibility threshold). If this is the case, the processed image is dimmed to have perfect visual fidelity (VSNR is set to plus infinity) and no further analysis is required. Otherwise, if the distortions are higher than the threshold, a second stage, in which the metric models some visual property to obtain the perceptual difference between the reference and the processed image, is applied. Finally, the VSNR score is computed based on a simple linear sum of these distances.

PSNR-HVS PSNR Human Visual System [214]. This metric is based on PSNR and UQI modified to take into account the HVS properties such as its sensitivity to contrast change and noise, and to low frequency distortions. The modification considers the removal of mean shifting and contrast stretching using a scanning window according to the method described in [212].

VQM Video Quality Metric [213]. VQM was proposed by Pinson *et al.* in [213].

It is a FR video quality metric measuring the perceptual effects of video impairments including blurring, jerky/unnatural motion, global noise, block distortion and color distortion, and combines them into a single metric. VQM is carried out in several steps: 1) calibration of the sampled video, 2) extraction of spatial and temporal features, 3) computation of video quality parameters, 4) linear combination of the calculated parameters to derive the final quality score. This metric is complex but well correlates with subjective scores [216].

MOVIE MOTion-based Video Integrity Evaluation [215]. This metric measures both spatial and temporal video distortions over multiple scales, and along motion trajectories. First, the reference and test videos are decomposed into spatio-temporal band pass channels using a Gabor filter family. Spatial quality measurement is accomplished by a method loosely inspired by the SSIM index. Temporal quality is measured using optical flow computed from the reference video sequence and the processed one, and is used to check the motion integrity between the two videos. Finally, the spatial and temporal quality scores are pooled to obtain an overall video integrity score known as the MOVIE index.

The majority of objective metrics proposed in the literature fall into this category. However, it is impossible to implement such metrics in practical situations where the original video is not available.

• No-reference (NR)

NR metric does not require any information about the reference image, instead, a quality score is computed based only on the tested image. Since, in practical applications, the reference/original image is rarely available, NR image/video quality assessment algorithms are highly desirable to develop measurement approaches that can blindly evaluate the image quality. However, the main limitation of existing NR algorithms is related to their focus to specific artifacts, *i.e.*, they assume that the distortion affecting the image is known. For instance, there exist NR IQA metrics that are developed to measure particular distortions, mostly for compression artifacts, such as blockiness [217], blurriness [259–261], and ringing [218, 219]. NR IQA algorithms predicting quality without any knowledge of the distortion affecting the image are rare [220, 221]. Shahid *et al.* provide a good survey of recent NR image and video QA approaches [252].

• Reduced-Reference (RR)

In the RR metric, only representative features of the reference image are used. Thus, this category lies in between the two extremes. The extracted features describing the reference image are transmitted to the receiver as side information to help evaluate the quality of the distorted image. Many RR IQA methods have been developed by considering the distortions behaviors, HVS properties, and the statistics modeling of visual signals [223–226].

5.2.3 Performance evaluation of objective quality metrics

In order to evaluate the performance of a visual quality metric, visual subjective quality databases are employed. These databases consist of a set of reference and distorted images or videos, with assigned subjective scores MOS or DMOS. The subjective scores are obtained through subjective tests following the standardized procedures as introduced in Section 5.2.1.

Consequently, the performances of objective metrics are evaluated by comparing their outputs to these subjective scores considered as the ground truth. The quality metric is considered as providing good prediction performance, if the correlation between both scores (subjective and objective) is high.

Before the correlation measurement, first, the objective scores are fitted to the set of subjective scores (MOS or DMOS). The fitting can be linear or nonlinear and is intended to convert the quality scores obtained with different objective metrics to the common range of the subjective scores. In addition, it removes any offset that may be present between the two compared sets of scores. Finally, in case of non linear relationship between the predictive and subjective scores, it is compensated by a non linear fitting.

The next step is to characterize prediction performance in terms of three attributes: *accuracy*, *monotonicity*, and *consistency*, with respect to subjective scores. To achieve this, several performance measurements can be used:

Prediction Accuracy: *Linear Correlation Coefficient (LCC)* and *Root Mean Square Error (RMSE)*

These two measures analyze the ability of the objective metric to predict subjective ratings with a minimum error on average, and are defined as:

$$LCC = \frac{\sum_{i=1}^N (MOS_i - \overline{MOS})(MOS_{p_i} - \overline{MOS_p})}{\sqrt{\sum_{i=1}^N (MOS_i - \overline{MOS})^2 \sum_{i=1}^N (MOS_{p_i} - \overline{MOS_p})^2}} \quad (5.7)$$

and

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (MOS_i - MOS_{p_i})^2} \quad (5.8)$$

where N is the number of test sequences or images. MOS_i and MOS_{p_i} are the subjective score and the predicted one, respectively. While $\overline{MOS}$ and $\overline{MOS_p}$ are the corresponding mean values.

The values of LCC vary between $[-1, 1]$, when they are close to the extremes, this indicates good correlation, while values around 0 signify no correlation. In contrast, RMSE values close to 0 mean better performance.

Prediction Monotonicity: *Spearman Rank Order Correlation Coefficient (SROCC)*

This measure quantifies the degree of monotonicity of the objective scores ac-

cording to the subjective ones. Ideally, differences of a metric's rating between two sequences should always have the same sign as the differences between the corresponding subjective ratings. SROCC measure is calculated as follows:

$$SROCC = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N \times (N^2 - 1)} \quad (5.9)$$

where d_i is the difference between ranks of the predicted score MOS_{p_i} and the reference one MOS_i . Similarly to LCC, the best monotonic case corresponds to -1 and 1 , while the worst one is close to 0 .

Prediction Consistency: *Outlier Ratio* (OR)

This measure checks if the objective metric provides consistent predictions for all images/videos and does not fail excessively for a subset of them:

$$OR = \frac{n_{outlier}}{N} \quad (5.10)$$

where $n_{outlier}$ is the number of predictions outside two standard deviations of the subjective scores.

More details on these and other measures can be found in [262, 263].

5.3 Stereoscopic image quality assessment (SIQA): State-of-the-art

Even though simulating 3D perception is closely related to 2D, it is much more complicated. The main difficulty lies into the impossibility of an algorithm to access 3D content as it is constructed and perceived, and more generally as it is experienced by the human observer. Because, we can only access the left and right views of the scene and possibly a depth/disparity map (see Figure 5.4). Consequently, the challenge is to find an approach to measure the perceptual quality of stereoscopic image-pair as performed by the HVS, and only by exploiting this stereoscopic content.

In the recent past, many 3D image quality assessment (3D-IQA) metrics have been proposed to quantify objectively the perceived 3D image quality [200]. Similar to 2D-QA, 3D-QA can be classified into the three previously described categories (FR, RR, and NR). In 3D-QA, each of these categories may in turn be divided into three families of methods, depending on whether or not they exploit depth (or disparity) information, and also according to the exploitation of binocular perception characteristics.

5.3.1 SIQA based on 2D metrics

In order to objectively assess the 3D image quality, a straightforward solution is to apply the state-of-the-art 2D-IQA methods to each view of the stereo pair, and

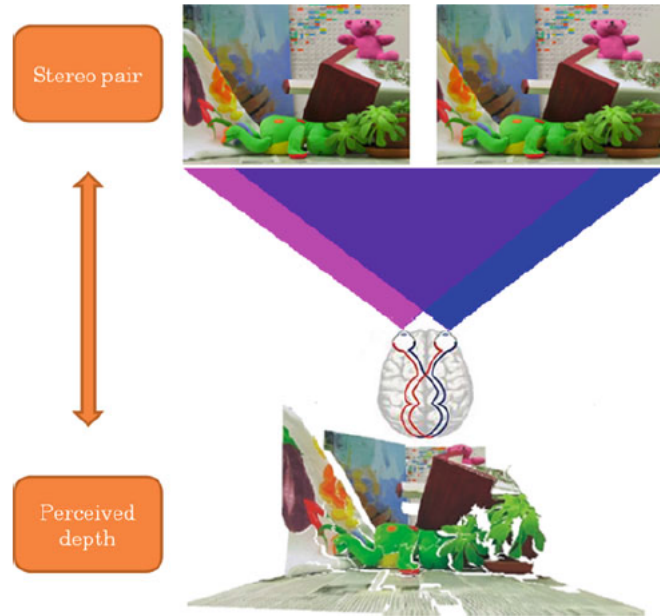


Figure 5.4: 3D perceived quality based on stereoscopic image-pair [239].

combine them into an overall score. This approach has been followed by several authors [227–230].

For instance, Campisi *et al.* [227] evaluated the appropriateness of four 2D-IQA metrics for stereoscopic images. Instead to always use average to combine the 2D IQ scores from the left and right views of a stereo pair. The authors have compared the average approach against the main eye approach and the visual acuity approach. For the main eye approach, only the objective score of the view corresponding to the dominant eye has been taken into account for each subject. For the visual acuity approach, the left and right scores have been weighted with the visual acuity of the left and right eyes, respectively. Results have shown no performance improvement over the average approach. However, in this study, only symmetric distortions were applied to the left and right images. Yasakethu *et al.* studied the relationship between subjective scores and several 2D image quality measures to predict the overall 3D image quality [229]. They found that some 2D image quality metrics yield fairly good correlation with overall image quality and perceived depth quality.

The obtained results of this class of methods can be regarded as moderate for the case of symmetric distortions, but are not satisfactory when the stereoscopic image-pair is asymmetrically distorted. Several studies demonstrated that simply predicting 3D image quality as the combination of left and right quality scores does not well correlate with the human quality judgment [85, 166]. This can be explained by the fact that 3D perception involves several binocular visual phenomena, such as binocular fusion and binocular rivalry/suppression, and the latter have a great impact on the perceived quality of 3D content.

5.3.2 SIQA including depth/disparity information

Since the HVS uses the so-called *binocular disparity* to perceive depth, other methods [231–238] have attempted to integrate the disparity (depth) map into the computation of SIQA metric, under the assumption that depth information has an important impact on 3D viewing experience. For example, Benoit *et al.* proposed to basically use a 2D quality metric to estimate the quality scores between the reference and distorted stereo pairs and combine them with the estimated scores of disparity distortion [231]. For these purposes, the distortion measure of disparity is obtained by the correlation between original and distorted disparity maps, while the image distortion is evaluated by SSIM [145] and C4 [222] metrics.

In the same vein, You *et al.* [232] applied ten well-known 2D quality metrics on both stereo pairs and disparity maps, combined the two respective scores in different ways, and concluded that the best results have been obtained using the SSIM metric as 2D image quality metric. Sarikan *et al.* proposed to assess the effect of different depth levels on the overall 3D quality. They presented in [233] a quality evaluation model for stereoscopic videos using depth map segmentation.

A NR 3D-IQA method was proposed by Sazzad *et al.* [236]. In their method, local features based distortions and relative disparity measures are estimated. Then, a logistic regression model is used to combine the different features, and finally to predict the quality score. However, the computational cost of this metric is high, since it requires many parameters and local features. The same authors have attempted to reduce the complexity in [237]. In [238], the authors present an evaluation of twelve 2D image quality metrics for assessing the quality of an image synthesized by DIBR. The results indicate that the performance of 2D-IQA metrics is moderate for measuring the quality of stereoscopic images, but not satisfactory when measuring the quality of synthesized images.

Despite the enhancement obtained by this second class of methods, these methods dependent heavily on the performance of disparity estimation methods, since these latter are still error prone. Moreover, disparity maps have different properties when compared to natural images. Assessing the depth/disparity map using 2D-IQA metrics is not efficient, because these 2D quality metrics are optimized to evaluate images targeted for visualization, in contrast to disparity map which is considered as one of the binocular cues.

5.3.3 SIQA considering binocular perception

Recently, the third class attracting a significant research effort considers the integration of different characteristics of the binocular perception in the 3D-QA algorithm [239–246]. For instance, Bensalma *et al.* [239] proposed modeling the phenomenon of binocular fusion, by estimating the difference of binocular energy between the reference and impaired stereo pairs. This is achieved by modeling the behavior of simple and complex cells of the HVS responsible of the described fusion. This bio-inspired method achieved good performance, however, it is relatively

complex for a use in real time application. In order to consider the binocular vision proprieties, Wang *et al.* [240] proposed a binocular spatial sensitivity weighted metric, the weighting is based on the JND model [140]. Despite the good results obtained by these methods, our understanding of 3D human perception remains limited. Therefore, study the different factors influencing the 3D QoE remain an interesting avenue for research.

Knowing that, when the stereoscopic images are presented to the viewer, the human brain combines the left and right images to create a synthesized image called *cyclopean image* [247]. Accordingly, to perform more reliable and accurate SIQA, some approaches performed the assessment process on this cyclopean view, such as [241–243]. Maalouf *et al.* [241] proposed a 3D image quality metric based on the computation of the cyclopean images of both reference and distorted stereo pairs. The authors computed the 3D score as an average of the difference between both corresponding cyclopean images and coherence of the disparity maps. Recently, Chen *et al.* [242] proposed a FR metric that performs SIQA on a cyclopean image. To simulate the binocular rivalry process, they used weighting coefficients for the left and right eyes computed from the normalized Gabor filter magnitude responses.

In the previous methods relying on the cyclopean image [241–243], it is important to highlight that there is no consideration of the inherent problem of occluded regions. To cope with this issue, authors of [244] proposed to first categorize the distorted stereoscopic images into three regions: non-corresponding region, binocular fusion region and binocular suppression region. Each region is evaluated separately based on its binocular perception property, and finally a pooling stage is performed to produce the overall 3D image quality score. Unfortunately, in this approach the score of binocular suppression/rivalry region is computed as the maximum score from the left and right quality scores, which is not in compliance with the literature. Indeed, as mentioned in [247–251], the view of high contrast or rich contours, and generally that contains more information, tends to dominate the 3D final perception in the case of binocular rivalry. This assumption proved its effectiveness and has been exploited and validated in the following recent works [240, 242, 246].

All the methods described above belong to FR 3D-IQA category. Although many NR IQA methods have been proposed for conventional 2D image/video during the last decade, only a few works [235, 236, 240, 243, 245] have developed RR or NR metrics for SIQA problem. The NR algorithms are considered as the most practical and desirable approaches since the reference image is rarely available in many multimedia systems, such as in computer networks, wireless communications, and real-time streaming services.

Based on an idea outlined in [242], the same authors proposed in [243] NR 3D-IQA algorithm that exploits 2D and 3D natural statistics features extracted from the tested stereo pair. Firstly, the extracted features are used to determine if the stereo pair is symmetrically or asymmetrically distorted. Next, the same features are used to compute the quality score. Ryu *et al.* [245] proposed NR SIQA metric based on the binocular quality perception model. In their model, they only consider the high and low quality components of the stereoscopic image-pair, making it unsuitable for

the case when the stereo pair is non-uniformly distorted.

5.4 Proposed quality assessment metric for stereoscopic images

In this section, we describe three proposed objective quality metrics for stereoscopic images taking into account the binocular visual phenomena. Two of them belong to the FR quality metric class (Sections 5.4.2 and 5.4.3) [161], while the third is a NR quality metric (Section 5.4.4). These three metrics are based on the proposed binocular perception model described in Section 5.4.1. Through the inclusion of the perceptual binocular characteristic of HVS, this model handles effectively the asymmetric distortions of stereoscopic images. Details of the proposed metrics are presented in the following sections.

5.4.1 Binocular perception model

As mentioned previously, depth perception is made by presenting two slightly different images to left and right retinas. Then, human brain analysis and fuses these two stimuli and constructs depth information. According to the degree of similarity between views of the stereo pair, different binocular phenomena may occur. When the left and right images are similar, the viewer perceives a unique scene resulting from the fusion process between the two retinal projections. This phenomenon made by the HVS is known as the *binocular fusion* [98, 99]. On the other hand, when dissimilar (incompatible) stimuli are presented, causing match failure, instead of being merged, the two images enter in a kind of competition. The latter known as the *binocular rivalry* phenomenon, where only one view dominates the fusion process while the other image incurs the *binocular suppression* [247–250]. In the psychophysical studies, the binocular suppression is considered as a special case of binocular rivalry. This suppression theory is exploited as a basis for validating the asymmetric stereoscopic coding methods (see Chapter 4) [96].

Regarding the quality of asymmetrically distorted stereoscopic images, results presented in [85, 166] are somewhat contradictory. For instance, concerning the blur distortion it has been demonstrated in [166] that 3D quality scores are mainly determined by the high-quality view of the stereo pair. On the other hand, for blockiness distortion, in [85] it is reported that the perceptual 3D quality is more affected by the view with lower quality, while in [166] it is specified that the quality can be calculated as the average between left and right views.

As mentioned in [166], 3D human perception is dominated by the view that contains more information, which is consistent with psychophysical findings described in [248, 250]. For example, if one view of the stereo pair is blurred and the other view unaltered, the resulting perceptual quality of the stereo-pair follows the high quality view. This is because the latter view (unaltered) contains more information than the former view (blurred image). On the other hand, for the case of blockiness

or noise distortion, the binocular perception does not follow the high quality view, because the latter has a lower amount of information than the blocky or noisy image. In other words, noise and blockiness introduce new information, thereby preventing them from being suppressed by the other view.

The conclusions drawn above suggest to concentrate on the local information contained in images for predicting quality. Hence, entropy used for estimating image information changes, has been already exploited for the development of quality metrics and proved its effectiveness [253–256].

Inspired by these findings, we integrated in our 3D-QA model the local information content as a weighing factor in order to simulate the binocular phenomena as follows:

$$Q3D(i, j) = e_l(i, j) \times Q2D_l(i, j) + e_r(i, j - d_l(i, j)) \times Q2D_r(i, j - d_l(i, j)) \quad (5.11)$$

where $Q2D_l$ and $Q2D_r$ are 2D image quality maps of the left and right images, respectively. $d_l(i, j)$ is the disparity value of the pixel (i, j) in the left view. In our implementation, to estimate the disparity map, we use the stereo matching algorithm described in [189]. Whereas $e_l(i, j)$ and $e_r(i, j - d_l(i, j))$ are weighting coefficients defined by:

$$\begin{aligned} e_l(i, j) &= \frac{EN_l(i, j)}{EN_l(i, j) + EN_r(i, j - d_l(i, j))} \\ e_r(i, j - d_l(i, j)) &= \frac{EN_r(i, j - d_l(i, j))}{EN_l(i, j) + EN_r(i, j - d_l(i, j))} \end{aligned} \quad (5.12)$$

where $EN_*(i, j)$ denotes the local entropy at each spatial location (i, j) . This entropy is calculated within a circular window of size 11×11 moving pixel-by-pixel over the entire image as follows:

$$EN(i, j) = - \sum_{k=1}^n p(x_k) \cdot \log_2(p(x_k)) \quad (5.13)$$

where n and $p(x)$ are the number of bins and the probability of a bin in the local histogram, respectively. An example of local entropy maps is depicted in Figure 5.5.

With such a modeling, our model can take into consideration the binocular perception properties. For instance, if we consider the case of two similar stimuli having equal or close local entropies, they will obviously contribute equally to the predicted quality as suggested by the binocular fusion process. On the other hand, if dissimilar stimuli are presented, accordingly the local entropies will be different, and the contribution to the predicted quality will be dominated by the view having the higher local entropies as suggested by the binocular rivalry/suppression.

To demonstrate that the local information content well correlates with distortion types, we show in Figure 5.5 the local information content maps of the original

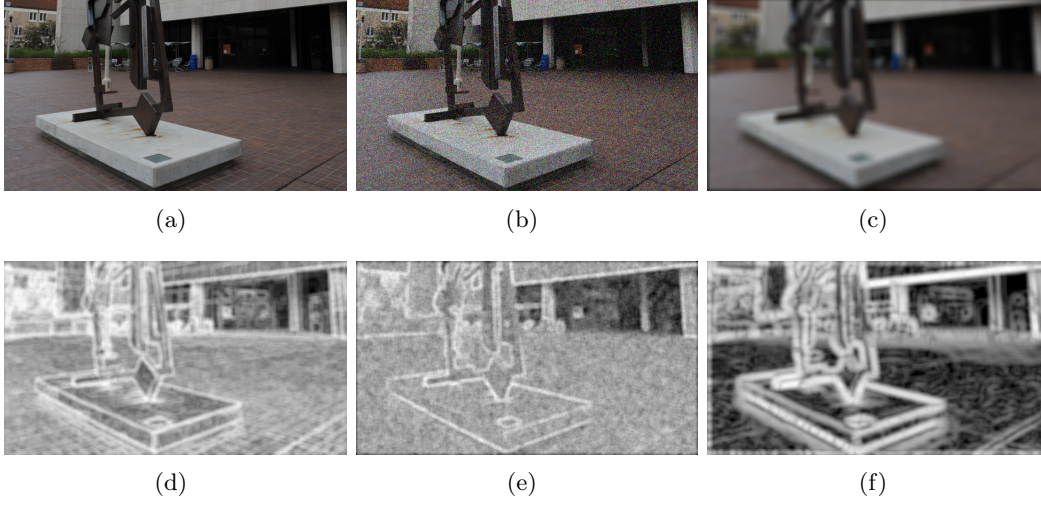


Figure 5.5: Examples of local information content maps. (a) Original image. Distortion types are: (b) Gaussian noise, (c) Gaussian blur. (d),(e),(f) Corresponding information content maps. Brighter indicates stronger information.

image and corresponding distortion images obtained by introducing a blur and a Gaussian noise. From this figure, it is easy to notice that when the Gaussian noise is introduced, the amount of information increases and in contrast, blur reduces the information. The following three described metrics are based on this model.

5.4.2 Proposed FR IQA metric for stereoscopic images

Let $(I_{r,l}, I_{r,r})$ and $(I_{d,l}, I_{d,r})$ be the reference and distorted stereo pairs, respectively, each contains left $I_{*,l}$ and right $I_{*,r}$ images. The first step aims at computing perceptual 2D image quality scores between both stereo pairs. Even though several metrics can be used, we opted for the SSIM metric for its simplicity and availability, to be used between the reference and distorted left (resp. right) images. As a result, two SSIM index maps are obtained (in the spatial domain) for the left and right images, noted as $SSIM_k$, $k \in \{l, r\}$.

Since the occlusion is an inherent problem in stereo vision, before moving to the next step, an essential classification shall be carried out. Each pixel in the image should be characterized as belonging to occluded or non-occluded class. To achieve this, we apply left-right consistency check between the corresponding pixels of two views. The check exploits the two disparity maps: the left-to-right disparity map $d_l(i, j)$ and the right-to-left disparity map $d_r(i, j)$. The matched pixels are considered as inconsistent if the sum of their disparity values is larger than a threshold T_{LR} (we use $T_{LR} = 1$):

$$|d_l(i, j) + d_r(i, j - d_l(i, j))| \geq T_{LR} \quad (5.14)$$

For occluded regions, the monocular vision occurs, whereas for the regions

that have a match in the other view, the binocular phenomena (fusion or rivalry/suppression) occur. In the following steps, the non-corresponding and corresponding regions are processed separately.

To mimic the sensitivity of the HVS, we modulate the quality score of each pixel according to its perceptual importance. For this purpose, we use the Just Noticeable Difference (JND) model [257] (see Section 1.3.2) for occluded pixels, and the Binocular-JND (BJND) [83] (see Section 1.3.3) for the non-occluded ones. A significant value of the JND/BJND means that the pixel tolerates a large distortion, and therefore has a low importance. Accordingly, we adjust the quality score of the occluded pixels in each view as follows:

$$Q_k^{oc} = \frac{\sum_{i,j} \frac{1}{JND_k(i,j)} \times SSIM_k(i,j)}{\sum_{i,j} \frac{1}{JND_k(i,j)}}, \quad k \in \{l, r\} \quad (5.15)$$

where i and j are the pixel coordinates. $JND(i, j)_k$ and $SSIM_k(i, j)$ denote the JND value and the SSIM score of the pixel (i, j) in the view k , respectively.

Finally, the quality score of the occluded regions is computed as the average between the left and right occluded regions as follows:

$$Q3D^{oc} = \frac{Q_l^{oc} + Q_r^{oc}}{2} \quad (5.16)$$

In the same vein, the non-occluded pixels are weighted as follows:

$$Q_k^{noc}(i, j) = \frac{\frac{1}{BJND_k(i,j)} \times SSIM_k(i, j)}{\max(\frac{1}{BJND_k})}, \quad k \in \{l, r\} \quad (5.17)$$

where $BJND_k(i, j)$ denote the BJND value of the pixel at position (i, j) .

Once the above score ($Q_k^{noc}(i, j)$) has been calculated for each non-occluded pixel of the left and right images, in order to consider the binocular phenomena, we apply the proposed binocular perception model (described in Section 5.4.1) within the 3D-QA framework as follows:

$$Q3D^{noc}(i, j) = e_l(i, j) \times Q_l^{noc}(i, j) + e_r(i, j - d_l(i, j)) \times Q_r^{noc}(i, j - d_l(i, j)) \quad (5.18)$$

where $e_l(i, j)$ and $e_r(i, j - d_l(i, j))$ are defined by equation (5.12).

Then, the perceptual quality score of the non-occluded regions is obtained by calculating the average of the $Q3D^{noc}$ index map as follows:

$$Q3D^{noc} = \frac{\sum_{i,j} Q3D^{noc}(i, j)}{N} \quad (5.19)$$

where N is the number of non-occluded pixels in the left image.

Finally, the overall 3D perceptual quality score $Q3D$ is calculated using a linear

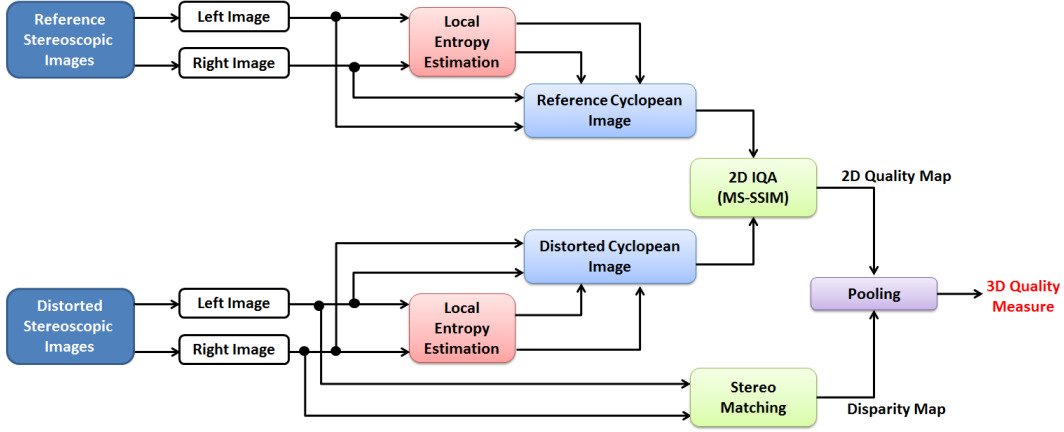


Figure 5.6: The flowchart of the FR metric based cyclopean image.

summation as follows:

$$Q3D = \alpha \cdot Q3D^{oc} + (1 - \alpha) \cdot Q3D^{noc} \quad (5.20)$$

where α is the weight used to adjust the relative importance of each region. In order to determine the parameter α , only a few training images (15 stereo pairs) are used. As a result, we obtained $\alpha = 0.142$, this clearly demonstrates that the occluded regions have less influence on the binocular vision.

5.4.3 Proposed cyclopean image quality approach for stereoscopic images

In order to ensure that the predicted 3D quality of experience is as reliable and close as possible to 3D human perception, we present in this section a new SIQA method performed on cyclopean image. In addition to the integration of the proposed binocular model (described in Section 5.4.1), we consider the disparity/depth masking effect by using the disparity map in the proposed 3D-IQA framework, thus reflecting more accurately the depth perception.

The flowchart of the proposed metric is shown in Figure 5.6. As the first step, we create two cyclopean images corresponding to the reference and distorted stereo pairs. By considering the left view as reference view, we synthesize the cyclopean view by fusing the left image with the right disparity-compensated image as follows:

$$I_{cyc}(i, j) = I_l(i, j) + I_r(i, j - d(i, j)) \quad (5.21)$$

where i and j are the pixel coordinates. I_{cyc} is the generated cyclopean image, I_l and I_r are the left and right images, respectively. $d(i, j)$ is the disparity value of the pixel (i, j) representing the horizontal shift of the pixel from left to right views.

With such a modeling, only the binocular fusion case is considered, *i.e.*, when the two stimuli are similar. However, in case of asymmetric distortions (dissimilar

stimuli), instead of being fused, the two images are in rivalry (*i.e.*, binocular rivalry). Usually, the human brain attempts to resolve this rivalry between two dissimilar stimuli by suppressing (masking) one of them. According to the proposed binocular model, to simulate this rivalry when computing the cyclopean image, we define weighting factors for each view. Therefore, equation (5.21) can be rewritten as follows:

$$I_{cyc}(i, j) = e_l(i, j) \times I_l(i, j) + e_r(i, j - d(i, j)) \times I_r(i, j - d(i, j)) \quad (5.22)$$

where $e_l(i, j)$ and $e_r(i, j - d(i, j))$ are defined by equation (5.12).

Once we obtained the cyclopean views of the reference (I_{cyc}^{ref}) and distorted (I_{cyc}^{dis}) stereo pairs, we then perform 2D-IQA metric between both obtained cyclopean images. Even though several metrics can be used, we opted for the Multi-Scale Structural SIMilarity (MS-SSIM) metric [258]. Because, the latter has provided the best results compared to other tested IQA metrics:

$$S = MS-SSIM(I_{cyc}^{ref}, I_{cyc}^{dis}) \quad (5.23)$$

where S denoted the obtained MS-SSIM index map (in the spatial domain).

Next, we modulate the quality score of each pixel according to its location in the scene. We claim that the visibility of different distortions at different depth levels is different. Accordingly, the HVS is more sensitive to closer objects than deeper ones. Also, it is known that disparity is inversely proportional to depth (see Section 1.2.2.1 for more details), meaning that objects with greater disparity are the closest, and vice versa.

Finally, the overall 3D perceptual quality score Q3D is calculated as follows:

$$Q3D = \frac{\sum_{i,j} d(i, j) \times S(i, j)}{\sum_{i,j} d(i, j)} \quad (5.24)$$

5.4.4 Proposed NR perceptual blur metric for stereoscopic images

To the best of our knowledge, except the work described in [245], there is no NR 3D-IQA metric focused on blur distortion. Blurriness of an image is commonly due to the attenuation of the high spatial frequencies, which could be caused by acquisition settings, filtering or image compression algorithms,..., etc. However, for the case of stereoscopic 3D, if one image of the stereo pair is blurred while the other is sharp, the resulting 3D perception is close to the sharper image. On the other hand, the HVS tolerates a certain limited level of blur in one view that does not impair the 3D viewing experience. Accordingly, it is very hard to assess the perceived blurriness in case of asymmetric blur distortion. Therefore, we propose a new NR SIQA method based on perceptual binocular characteristics assessing blurred stereo pair.

The flowchart of the proposed metric is shown in Figure 5.7. The first step aims at computing local perceptual blurriness maps for both left and right images of the stereo pair. To achieve this step, we have tested several NR sharpness metrics,

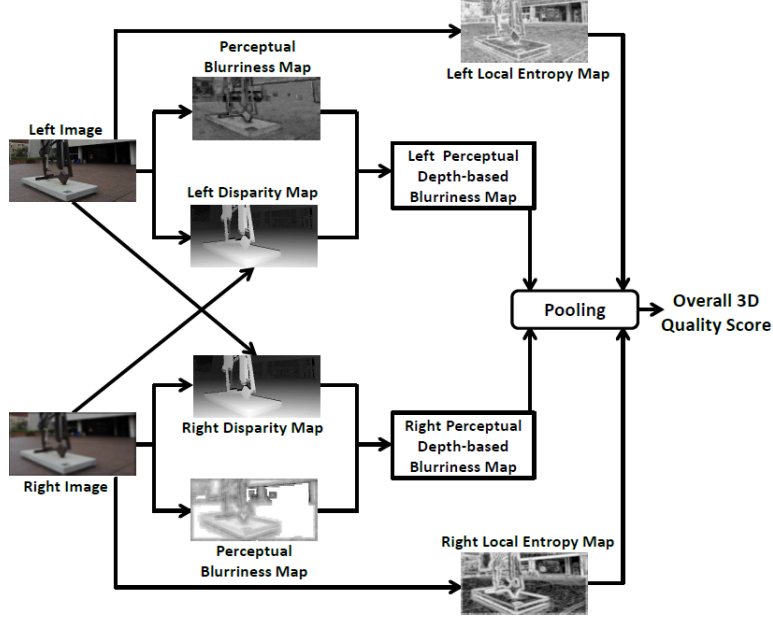


Figure 5.7: Flowchart of the proposed NR SIQA metric.

since blurriness and sharpness are inversely proportional. Among them, the one introduced by Ryu *et al.* [259], Narvekar *et al.* [260], and Vu *et al.* [261]. After performing a deep experimentation, we opted for the one proposed by Vu *et al.* (Referred to as S_3) [261], which provided a local perceived sharpness measure highly correlated with the HVS, and obtained the best performance. As a result, two local blurriness maps are obtained for the left and right images, noted as Bm_k , $k \in \{left, right\}$. Examples of local blurriness maps are presented in Figure 5.8.

As shown by subjective experiments described in [186], the sharpness visibility at various depth levels is different. The perceived blur of the object is highly dependent on the objects' distance (position in the scene), *i.e.*, the closer object appears sharper than deeper objects. Accordingly, to consider the relationship between blur/sharpness and depth level, we use depth information to pool the two local blurriness maps computed in the previous step as follows:

$$Bm_{k_d}(i, j) = Bm_k(i, j) \times \frac{d_k(i, j)}{d_{max}} \quad (5.25)$$

where $Bm_k(i, j)$ is the blurriness score of the pixel at position (i, j) , and $d_k(i, j)$ is the disparity value. Finally, d_{max} is the maximum value of the disparity map. The formulation of the pooling depth factor is justified by the fact that the disparity is inversely proportional to depth.

Next, in the following steps, the non-corresponding and corresponding regions are processed separately. Each pixel in the image should be characterized as belonging to occluded or non-occluded class. To achieve this, we apply left-right con-

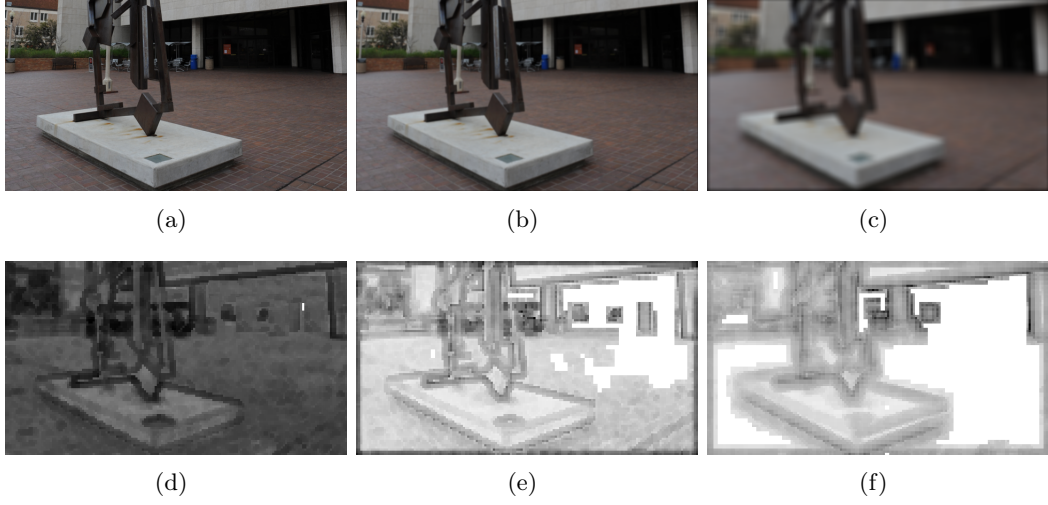


Figure 5.8: Examples of local perceptual blurriness maps. (a) Original image. (b),(c) Blurred images with different blur levels. (d)-(f) Corresponding local blurriness maps. Brighter indicates higher blurriness.

sistency check described by the equation (5.14), between the corresponding pixels of two views.

For occluded regions, the perceptual blurriness scores (Q_k^{oc}) for left and right images are calculated separately, and finally the average is computed to obtain the blurriness score ($Q3D_{blur}^{oc}$) of the occluded regions as follows:

$$Q_k^{oc} = \frac{\sum_{i,j} Bm_{k_d}(i,j)}{N}, \quad k \in \{l, r\} \quad (5.26)$$

$$Q3D_{blur}^{oc} = \frac{Q_l^{oc} + Q_r^{oc}}{2} \quad (5.27)$$

where N is the total number of occluded pixels.

Next, to take into account the binocular phenomena, we perform a weighting based on the local information content, as described in Section 5.4.1:

$$Q3D^{noc}(i,j) = e_l(i,j) \times Bm_{l_d}(i,j) + e_r(i,j - d_l(i,j)) \times Bm_{r_d}(i,j - d_l(i,j)) \quad (5.28)$$

where $e_l(i,j)$ and $e_r(i,j - d_l(i,j))$ are defined by equation (5.12).

The perceptual blurriness score of the non-occluded regions is obtained by calculating the average of the $Q3D^{noc}$ index map as follows:

$$Q3D_{blur}^{noc} = \frac{\sum_{i,j} Q3D^{noc}(i,j)}{M} \quad (5.29)$$

where M is the number of non-occluded pixels.

Finally, the overall perceptual blurriness score of the stereoscopic image-pair

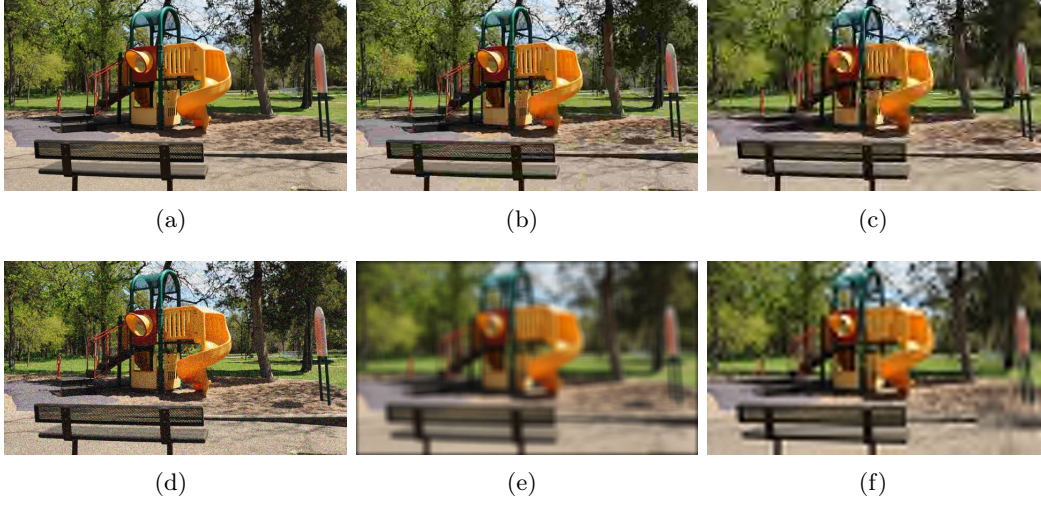


Figure 5.9: Example of images from the LIVE 3D IQA database (phase 2) [242, 242]: (a) Reference, (b) JPEG, (c) JPEG 2000, (d) Gaussian noise, (e) Gaussian blur and (f) Fast-fading.

$Q3D_{blur}$ is calculated using a linear summation as follows:

$$Q3D_{blur} = \alpha \cdot Q3D_{blur}^{oc} + (1 - \alpha) \cdot Q3D_{blur}^{noc} \quad (5.30)$$

where α is the weight used to adjust the relative importance of each region. After deep experimentation, we have found $\alpha = 0.135$.

5.5 Experimental results

In order to evaluate the performance of the proposed metrics, we used the LIVE 3D IQA database (phase 2) [242, 242]. This database, for which examples are given in Figure 5.9, contains both symmetric and asymmetric distortions. The distortions that were simulated are compression using the JPEG and JPEG 2000 standards, additive white Gaussian noise, Gaussian blur and a fast-fading model based on the Rayleigh fading channel. 360 distorted stereo pairs were generated, and the corresponding DMOS were obtained.

The evaluation of our metrics has been done using the three previously defined performance measures, namely LCC, SROCC, and RMSE. As explained in Section 5.2.3, we can say that a metric obtains good performance, if the values of PCC and SROCC are high (close to ± 1), and the value of RMSE is low (close to 0). All the measures are computed between DMOS and the scores provided by our metrics ($DMOS_p$) after a non-linear regression. This regression is performed using a 5-parameter logistic function as recommended in [262] and defined as follows:

$$DMOS_p(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp^{\beta_2(x - \beta_3)}} \right) + \beta_4 x + \beta_5. \quad (5.31)$$

Table 5.3: Performance comparison of the proposed FR SIQA metrics using state-of-the-art tools (Sym: Symmetric, Asy: Asymmetric).

Methods	LCC			SROCC			RMSE		
	All	Sym	Asy	All	Sym	Asy	All	Sym	Asy
PSNR	0.694	0.793	0.604	0.685	0.783	0.521	17.55	16.53	17.78
SSIM [145]	0.822	0.856	0.775	0.802	0.815	0.732	14.40	14.62	13.96
MS-SSIM [258]	0.835	0.893	0.723	0.793	0.898	0.582	15.11	13.45	16.01
Benoit [231]	0.756	0.804	0.74	0.732	0.811	0.706	16.1	15.08	16.72
You [232]	0.796	0.817	0.786	0.783	0.835	0.724	14.39	14.15	15.08
Gorley [230]	0.563	0.521	0.598	0.387	0.451	0.321	19.65	19.16	20.41
Hewage [235]	0.597	0.608	0.556	0.507	0.571	0.483	19.86	19.9	19.53
Chen1 [242]	0.901	0.937	0.879	0.893	0.929	0.854	10.58	9.37	10.29
Our FR1	0.925	0.956	0.903	0.928	0.937	0.919	9.05	8.04	10.03
Our FR2	0.938	0.961	0.915	0.930	0.941	0.920	5.55	5.06	5.95

Table 5.4: Performance comparison of the proposed NR metric using state-of-the-art tools (Sym: Symmetric, Asy: Asymmetric). Italicized algorithms are NR SIQA.

Methods	LCC			SROCC			RMSE		
	All	Sym	Asy	All	Sym	Asy	All	Sym	Asy
PSNR	0.663	0.695	0.610	0.675	0.779	0.581	8.235	8.023	8.551
SSIM [145]	0.807	0.856	0.775	0.8	0.865	0.741	6.853	6.634	7.064
MS-SSIM [258]	0.775	0.893	0.723	0.781	0.917	0.688	7.120	7.004	7.2
Benoit [231]	0.746	0.804	0.74	0.722	0.871	0.696	7.562	7.079	8.02
You [232]	0.796	0.817	0.786	0.781	0.897	0.704	6.881	6.642	7.128
Gorley [230]	0.523	0.521	0.598	0.387	0.412	0.271	9.650	9.551	9.803
Hewage [235]	0.515	0.608	0.556	0.502	0.591	0.491	9.464	9.218	9.58
Chen1 [242]	0.924	0.937	0.876	0.886	0.935	0.857	4.991	4.882	5.159
<i>Chen2 [243]</i>	0.895	0.910	0.879	0.881	0.919	0.854	5.203	5.012	5.591
<i>Our NR</i>	0.913	0.927	0.904	0.88	0.936	0.853	5.435	4.94	5.162

Table 5.5: Performance comparison for each distortion. Italicized algorithms are NR SIQA.

Methods	LCC		SROCC		RMSE	
	JPEG2000	Gaussian blur	JPEG2000	Gaussian blur	JPEG2000	Gaussian blur
PSNR	0.597	0.690	0.612	0.751	7.674	9.865
SSIM [145]	0.704	0.838	0.721	0.856	6.783	7.370
MS-SSIM [258]	0.798	0.801	0.820	0.841	5.621	8.397
Benoit [231]	0.784	0.535	0.751	0.455	6.096	11.763
You [232]	0.905	0.784	0.894	0.813	4.186	8.649
Gorley [230]	0.372	0.934	0.110	0.770	9.113	4.988
Hewage [235]	0.664	0.450	0.598	0.280	7.343	12.436
Chen1 [242]	0.834	0.963	0.814	0.908	5.562	3.747
<i>Chen2 [243]</i>	0.899	0.941	0.867	0.9	4.298	4.725
<i>Our NR</i>	0.904	0.958	0.859	0.913	4.930	3.671

Table 5.3 shows the results of the two proposed FR SIQA metrics, while the results of our NR 3D-IQA metric are depicted in Table 5.4, compared to state-of-the-art metrics. The first three lines in both tables are obtained by averaging the left and right 2D-IQA scores, obtained with PSNR, SSIM [145], MS-SSIM [258]. While, Benoit *et al.* [231], You *et al.* [232], Gorley *et al.* [230], Hewage *et al.* [235], and Chen1 *et al.* [242] are FR 3D-IQA metrics requiring the reference image.

On the other hand, in the Table 5.4, *Chen2 et al.* [243] and our proposed method are NR SIQA metrics. The distortions that were considered for the assessment of the proposed NR SIQA metric are Gaussian blur and compression using the JPEG 2000 standards. The performance comparison for each of these distortions is presented in Table 5.5.

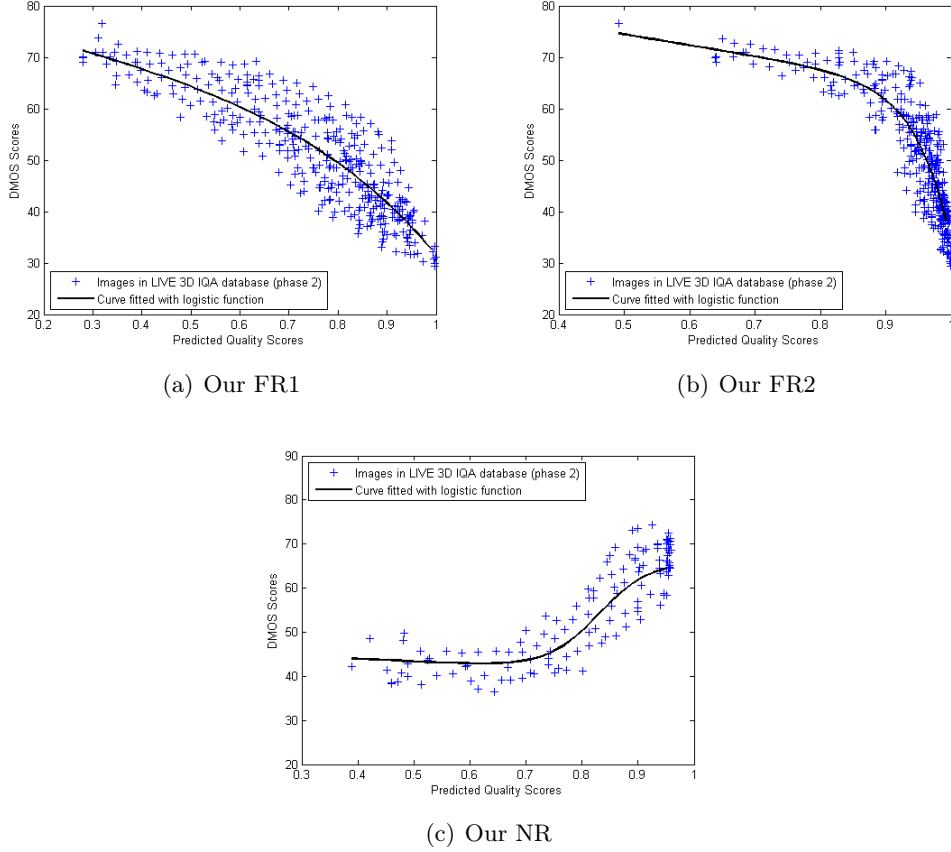


Figure 5.10: Scatter distributions of subjective DMOS versus predicted scores obtained by the proposed metrics. The continuous curve represent the non-linear fitting.

Based on the results shown in both tables, we can see that 2D image quality metrics provide acceptable results for the case of symmetric distortions. However, for the case of asymmetric distortions except the metrics proposed by Chen et al. [242, 243], all the metrics provide poor results.

However, according to the performance comparison, the two proposed FR SIQA metrics outperform the other metrics for both symmetric and asymmetric distortions, and this is particularly noticeable on the asymmetric distortions. Moreover, the proposed NR 3D-IQA metric outperforms most of the FR 3D-IQA metrics for both symmetric and asymmetric distortions, with a clear gap on asymmetric distortions. Also, from the result showed in Table 5.5, we can consider our method as an effective and reliable perceptual blurriness metric for stereoscopic 3D images. Thanks to the introduction of local information content weighing, through the proposed binocular model, our metrics handle the binocular phenomena more efficiently, without any prior detection of the distortion type or any additional information about the distortion.

Finally, besides the evaluation metrics used above, we provide in Figure 5.10 the plots showing the scatter distributions of subjective MOS versus the predicted scores obtained by the three proposed metrics, in addition to the non-linear fitted curve, on the LIVE 3D database (phase 2).

5.6 Conclusion

In this chapter, we proposed three novel quality assessment metrics for stereoscopic 3D images. These metrics are designed based on the assumption that 3D human perception places more emphasis on the view that contains more information. We integrated this psychophysical finding in the proposed 3D-QA framework under the form of a weighting based on the amount of available information. Consequently, the binocular visual phenomena, especially the binocular rivalry/suppression, have been taken into account when: combining the 2D image quality scores, calculating the cyclopean image, combining local blurriness maps. Add to that, the role of depth masking effect is also considered, we modulated the importance of each region (pixel) according its location in the scene. Furthermore, the proposed metrics have the advantage of locally adjust the quality score of 3D image, according to the HVS characteristics without any cue about the distortion type or distribution (symmetric or asymmetric).

The experimental results showed that the proposed metrics well correlate with subjective scores and outperform state-of-the-art metrics.

In the future, we will try to extend the metric to S-3D video by including temporal properties of the HVS. Also, the implementation of the metric in the frequency domain is planned, to take advantage of the representation. Finally, the asymmetric stereoscopic coding implemented through spatial filtering can cause visual fatigue. Consequently, we expect to include this factor in the proposed metric, which will lead to a more reliable 3D QoE prediction.

Conclusion and Future Work

Contributions of the thesis

In this thesis several contributions have been proposed, thus covering end-to-end 3D video system, from preprocessing to the assessment of the quality, through coding. The conducted works have been done with the aim to enhance coding performance while providing optimal quality of experience. This is achieved by incorporating perceptual features at in each step of the 3D video processing chain. Below we summarize the contributions of the thesis, and then propose some directions for future research.

- **Color correction of multi-view video for coding and rendering enhancement**

We showed that color inconsistency among different views of MVV reduces coding performance and introduces noticeable artifacts in rendered views. To fix this issue, we proposed a preprocessing method for color correction of MVV. First, instead of choosing manually the center view as the reference view, we proposed its automatic and effective selection in an adaptive manner, based on histogram difference and structural content. This allows to choose the view representing the best match with the whole set of views, and can be highly adapted for some specific capture configuration. Then, the correction is performed using an improved version of the widely used HM algorithm. To avoid its drawback, HM method is used only on overlapping regions over views, making our method robust to occlusion. The combination of the SIFT algorithm and RANSAC method is used to effectively define these common regions. In order to perform the correction while maintaining as much as possible the temporal correlation, we applied the method using a temporal sliding window, bringing the advantage of adaptation to time-varying acquiring system, camera moving capture, and real-time broadcasting. Experimental results illustrated the efficiency of the proposed method to calibrate and harmonize

the color of views, which led to coding gains and enhance the visual quality of the rendered views.

- **Asymmetric coding for stereoscopic 3D**

We proposed a novel asymmetric stereoscopic image/video coding method. The proposed approach enhances the mixed resolution coding concept by introducing a non-uniform reduction of spatial resolution allowing to decrease the bandwidth required for S3D delivery. The proposed method determines automatically and adaptively the maximum tolerance level of blur for each region (pixel) based on three HVS inspired models, making the blur effect almost transparent to the viewer. Objective and subjective performance evaluation results have demonstrated that the proposed approach can provide significant bit-rate saving without noticeable visual quality losses.

- **Stereoscopic image quality metrics based on binocular characteristics**

After a broad overview of 3D image quality assessment methods, we focused on new quality assessment metrics for stereo pairs containing asymmetric distortions, which is a much more challenging problem. Based on the assumption that 3D human perception places more emphasis on the view that contains more information, a new binocular model has been defined according to the local information content. Three stereoscopic image quality assessment metrics have been presented. The first one, is a full-reference quality assessment metric taking into account the binocular visual phenomena, especially the binocular rivalry/suppression that occurs when the stereo pair is impaired by asymmetric distortions. In the second one, in order to be as close as possible to the HVS behavior, we proposed to simulate the construction process of the cyclopean view as done by the human brain. To achieve that, the binocular visual phenomena have been taken into account when calculating the cyclopean image. Finally, the third metric is a no-reference perceptual blur metric for stereoscopic images. The latter, first, computes local blurriness for each image of the stereo pair. To consider the role of depth masking effect, we modulated the importance of each region (pixel) according to its location in the scene. Moreover, we used the proposed binocular model when combining local blurriness maps, thus allowing an adaptation to spatial variant blurriness. The experimental results pointed out that the proposed metrics well correlate with subjective scores and outperform state-of-the-art metrics.

Future Work

A future work that deserves to be addressed is to seek the way to reuse the informations produced by the feature correspondence step within the coding process itself. For example, to reduce the search area when calculating the disparity vectors, or to extract the parameters of the 3D geometric model, thus achieving the complete

end-to-end 3D video system. Also, even though the proposed approach is providing satisfactory results, we believe that including perceptual features (spatial and/or temporal) is a promising step towards the improvement of the quality of experience. In addition, maintaining the temporal coherence between successive frames within a single view needs to be improved. The proposed approach tackled this problem by defining a group of picture and performing the correction group by group. However, the size of group is fixed. Future studies should focus on proposing more advanced strategies allowing to tune adaptively and automatically the size of each group. Also, the reduction of computational complexity remains an important issue. One solution to address this problem is to parallelize the color correction process by using a graphics processing pipeline or a GPU implementation. Another point is the fact that there is no ground truth allowing efficient evaluation. To address this, we plan to construct a benchmark in order to be able to evaluate and compare each method of the state-of-the-art. Finally, the impact of color variation is not fully investigated, the relation between color inconsistencies and binocular rivalry should be explored. As the latter can induce visual fatigue, the maximum tolerance of color difference between views need to be quantified.

We plan to extend the proposed asymmetric coding method to a higher number of views ($N > 2$). Furthermore, recently, different 3D visual attention models have been proposed. Accordingly, in contrast to use 2D saliency model, we expect to use a stereoscopic one, in order to enhance the performance and to be as close as possible to real 3D viewing.

Moreover, we will try to extend the metrics to 3D video by including temporal properties of the HVS. Also, the implementation of the metric in the frequency domain is planned, to take advantage of the representation. In addition, the value of α should be explored in depth through intense experiments on other databases. Finally, the asymmetric stereoscopic coding implemented through spatial filtering can cause visual fatigue. Consequently, we expect to include this factor in the proposed metric, which will lead to more reliable 3D quality of experience prediction.

Bibliography

- [1] <http://www.futuresource-consulting.com/2013-09-3D-Tracking.html>, April 2014. (Cited on page 2.)
- [2] W. Lotmar, “Theoretical eye model with aspherics,” *JOSA*, vol. 61, no. 11, pp. 1522-1529, 1971. (Cited on page 7.)
- [3] A. Chapanis and R. A. McCleary, “Interposition as a cue for the perception of relative distance,” *Journal of General Psychology*, vol. 48, no. 2, pp. 113-132, 1953. (Cited on page 10.)
- [4] P. J. Kellman and T. F. Shipley, “A theory of visual interpolation in object perception,” *Cognitive psychology*, vol. 23, no. 2, pp. 141-221, 1991. (Cited on page 10.)
- [5] G. Kanizsa and G. Kanizsa, *Organization in vision: Essays on Gestalt perception*, Praeger New York, 1979. (Cited on page 10.)
- [6] W. H. Ittelson, “Size as a cue to distance: Static localization,” *The American journal of psychology*, pp. 54-67, 1951. (Cited on page 10.)
- [7] K. A. Stevens, “Surface perception from local analysis of texture and contour,” Technical report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1980. (Cited on page 10.)
- [8] J. J. Gibson, *The perception of the visual world*, Houghton Mifflin, 1950. (Cited on page 10.)
- [9] R. A. Young, “Oh say, can you see? The physiology of vision,” in Proc. *SPIE Human Vision, Visual Processing and Digital Display*, vol. 1453, pp. 92-123, San Jose, CA, 1991. (Cited on page 6.)
- [10] A. Fernando, S. T. Worrall, and E. Ekmekcioglu, *3DTV: processing and transmission of 3D video signals*, John Wiley & Sons, 2013. (Cited on pages 7 and 8.)

- [11] J. E. Cutting and P. M. Vishton, *Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth*, in William Epstein and Sheena Rogers, pp. 69-117, Academic press, San Diego, CA, 1995. (Cited on pages 10 and 13.)
- [12] D. H. Kelly, "Visual contrast sensitivity, *Journal of modern optics*, vol. 24, no. 2, pp. 107-129, 1977. (Cited on page 10.)
- [13] McGill Site. Vision, 2013. http://thebrain.mcgill.ca/flash/index_d.html (Cited on page 7.)
- [14] E. J. Gibson, J. J. Gibson, O. W. Smith, and Howard Flock. Motion parallax as a determinant of perceived depth," *Journal of experimental psychology*, vol. 58, no. 1, pp. 40-41, 1959. (Cited on page 8.)
- [15] N. A Dodgson, "Variation and extrema of human interpupillary distance," in *IS&T/SPIE Electronic Imaging Proceedings*, vol. 5291, pp 36-46, 2004. (Cited on page 8.)
- [16] K. K. De Valois, *Seeing*, Academic Press, London, UK, 2000. (Cited on page 7.)
- [17] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion," *Vision research*, vol. 35, no. 3, pp. 389-412, 1995. (Cited on page 13.)
- [18] J. M. Hillis, M. O.Ernst, M. S. Banks, and M. S. Landy, "Combining sensory information: mandatory fusion within, but not between, senses," in *Science* vol. 298, no. 5598, pp. 1627-1630, 2002. (Cited on page 13.)
- [19] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, 2013. (Cited on pages 14 and 17.)
- [20] C. Koch and S. Ullman, "Shifts in selection in visual attention: Toward the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-27, 1985. (Cited on page 15.)
- [21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998. (Cited on page 15.)
- [22] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194-203, 2001. (Cited on page 16.)
- [23] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, 2008. (Cited on page 15.)

- [24] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008. (Cited on page 15.)
- [25] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205-231, 2005. (Cited on page 17.)
- [26] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, June 2007. (Cited on page 17.)
- [27] A. L. Yarbus, *Eye-Movements and Vision*, Plenum Press, 1967. (Cited on page 16.)
- [28] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 18, 2008. (Cited on page 16.)
- [29] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, pp. 1338-1352, 2010. (Cited on page 16.)
- [30] M. Cerf, E. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, 2010. (Cited on page 16.)
- [31] W. Lin, "Computational models for just-noticeable difference," Chapter in *Digital video image quality and perceptual coding*, Wu, H. R., & Rao, K. R. (Eds.), CRC press. 2005. (Cited on page 17.)
- [32] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," in *Proc. IEEE*, vol. 81, pp. 1385-1422, October 1993. (Cited on page 17.)
- [33] G. E. Legge, "Contrast masking in human-vision," *Journal of the Optical Society of America*, vol. 70, no. 12, pp. 1458-1471, 1980. (Cited on page 17.)
- [34] A. B. Watson, "Model of visual contrast gain control and pattern masking," *Journal of the Optical Society of America A - Optics Image Science and Vision*, vol. 14, no. 9, pp. 2379-2391, 1997. (Cited on page 17.)
- [35] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *Journal of Physiology*, vol. 197, pp. 551-566, 1968. (Cited on page 18.)
- [36] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images," *Journal of Visual Communication and Image Representation*, vol. 19, no. 1, pp. 30-41, Jan. 2008. (Cited on page 19.)

- [37] C. Chou, and C. Chen, "A perceptual optimized 3-D subband codec for video communication over wireless channels," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 2, pp. 143-156, Apr. 1996. (Cited on page 19.)
- [38] X. Yang, W. Lin, Z. Lu, E. 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, Aug. 2005. (Cited on page 19.)
- [39] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE Conf. Human Vision, Visual Processing, and Digital Display*, 1993. (Cited on page 19.)
- [40] Z. Wei, and K. N. Ngan, "Spatial-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 337-346, Mar. 2009. (Cited on page 19.)
- [41] Z. Liu, L. J. Karam, and A. B. Watson, "Jpeg2000 encoding with perceptual distortion control," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1763-1778, 2006. (Cited on page 19.)
- [42] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164-1175, August 1997. (Cited on page 19.)
- [43] Michael Nauge. *The exploration of statistical structures of the image for the prediction of visual saliency and perceived quality*. PhD thesis, University of Poitiers, Poitiers, December 2012. (Cited on pages 25 and 28.)
- [44] Wojciech Matusik Alexander Behrens Jinho Lee Hanspeter Pfster Anthony Vetro, Morgan McGuire. Multiview video test sequences from merl. Busan, Korea, ISO/IEC JTC 1/SC 29/WG 11 Doc. M12077, April 2005. (Cited on page 24.)
- [45] Norishige Fukushima Masayuki Tanimoto, Toshiaki Fujii. 1d parallel test sequences for mpeg-ftv. Archamps, France, ISO/IEC JTC 1/SC 29/WG 11 Doc. M15378, April 2008. (Cited on page 24.)
- [46] ISO/IEC JTC1/SC29/WG11. Requirements on multi-view video coding v.8. WG11 Doc. N9163,, Lausanne, Switzerland, July 2007. (Cited on pages 25 and 28.)
- [47] A. Smolic, "3D video and free viewpoint video—From capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958-1968, 2011. (Cited on pages 25 and 28.)

- [48] A. Vetro, S. Yea and A. Smolic, "Towards a 3D Video Format for Auto-Stereoscopic Displays," in *Proc. of the SPIE Conf. on Applications of Digital Image Processing XXXI*, San Diego, CA, USA. 2008 (Cited on page 29.)
- [49] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth Map Creation and Image Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 217-234, 2007. (Cited on page 29.)
- [50] Peter D. Symes, Digital video compression, McGraw-Hill Professional, 2004. (Cited on page 31.)
- [51] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, 2007. (Cited on pages 34, 35, and 36.)
- [52] G. J. Sullivan and T. Wiegand, "Video compression – from concepts to the H.264/AVC standard," *Proceedings of IEEE*, vol. 93, no. 1, pp. 18-31, Jan. 2005. (Cited on pages 34 and 36.)
- [53] P. Merkle, K Müller, T. Wiegand, "3D Video: Acquisition, Coding, and Display," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 946-950, May 2010. (Cited on page 23.)
- [54] P. Assunção, L. Pinto, and S. Faria, "3D Media Representation and Coding," in *3D Future Internet Media*. Springer New York, pp. 9-38, 2014. (Cited on page 26.)
- [55] Ismael Daribo, "Coding and rendering of 3D video sequences; and applications to Three-Dimensional Television (3DTV) and Free Viewpoint Television (FTV)," Ph.D. thesis, Telecom ParisTech, Nov. 2009. (Cited on page 27.)
- [56] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (ToF) cameras: a survey," *IEEE journal of Sensors*, vol. 11, no. 9, pp. 1917-1926, 2011. (Cited on page 28.)
- [57] I. E. Richardson, The H. 264 advanced video compression standard, John Wiley & Sons, 2011. (Cited on page 31.)
- [58] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. IEEE International Conference on Multimedia and Expo (ICME 2006)*, Toronto, Ontario, Canada, Jul. 2006. (Cited on pages 36 and 37.)
- [59] M. E. Lukacs, "Predictive coding of multi-viewpoint image sets," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, pp. 521-524, Tokyo, Japan, 1986. (Cited on page 38.)

- [60] ITU-T and ISO/IEC JTC 1, "Final Draft Amendment 3," Amendment 3 to ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video), MPEG document N1366, Sept. 1996. (Cited on page 38.)
- [61] A. Puri, R. V. Kollarits, and B. G. Haskell. "Stereoscopic video compression using temporal scalability," in *Proc. SPIE Conf. Visual Communications and Image Processing*, vol. 2501, pp. 745-756, 1995. (Cited on page 38.)
- [62] X. Chen and A. Luthra, "MPEG-2 multi-view profile and its application in 3DTV," in *Proc. SPIE IS&T Multimedia Hardware Architectures*, San Diego, USA, pp. 212-223, Feb. 1997. (Cited on page 38.)
- [63] J.-R. Ohm, "Stereo/Multiview Video Encoding Using the MPEG Family of Standards," in *Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems VI*, San Jose, CA, Jan. 1999. (Cited on page 38.)
- [64] D. Tian, P. Pandit, P. Yin, and C. Gomila, "Study of MVC coding tools," Joint Video Team, Doc. JVT-Y044, Shenzhen, China, Oct. 2007. (Cited on pages 40 and 41.)
- [65] S. A. Fezza, K. M. Faraoun, and S. Ouddane, "A Comparison of Prediction Structures for Multi-view Video Coding based on the H.264/AVC Standard," in *Proc. IEEE International Workshop on Systems Signal Processing and their Applications (WoSSPA'2011)*, pp. 111-114, Tipaza, Algeria, 2011. (Cited on page 40.)
- [66] V. D. Silva, "Improving Perceptual Quality of 3DTV Systems," PhD thesis, University of Surrey, UK, Sep. 2011. (Cited on pages 7 and 8.)
- [67] V. Bruce, P. R. Green, and M. A. Georgeson, "*Visual perception*," 4th Edition, Psychology Press, 4 edition, 2003. (Cited on page 9.)
- [68] D. Minoli, "*3DTV content capture, encoding and transmission: building the transport infrastructure for commercial services*," John Wiley & Sons, 2010. (Cited on page 9.)
- [69] W. J. M. Levelt, *On Binocular Rivalry*, Mouton, The Hague, Paris, 1968. (Cited on page 80.)
- [70] I. P. Howard and B. J. Rogers, *Binocular Fusion and Rivalry in Binocular Vision and Stereopsis*, New York, USA: Oxford Univ. Press, 1995. (Cited on page 80.)
- [71] B. Julesz, *Foundations of Cyclopean Perception*. Univ. Chicago Press, 1971. (Cited on page 80.)
- [72] S. B. Steinman, B. A. Steinman, and R. P. Garzia, *Foundations of Binocular Vision: A Clinical Perspective*, New York, USA: McGraw-Hill, 2000. (Cited on page 80.)

- [73] ITU-T Recommendation H.264: Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264, Apr. 2013. (Cited on pages 34, 38, and 80.)
- [74] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461-1473, Nov. 2007. (Cited on pages 37, 38, 40, 41, and 80.)
- [75] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012. (Cited on pages 37 and 80.)
- [76] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori, "Overview of the MVC+D 3D video coding standard," *Journal of Visual Communication and Image Representation*, vol. 25, no. 4, pp. 679-688, May. 2014. (Cited on page 80.)
- [77] A. Vetro, A. M. Tourapis, K. Müller, and T. Chen, "3D-TV Content Storage and Transmission," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 384-394, Jan. 2011. (Cited on pages 23, 41, and 79.)
- [78] A. Smolic, K. Müller, 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 *Proc. Picture Coding Symposium (PCS 2009)*, Chicago, IL, USA, May. 2009, pp. 1-4. (Cited on pages 26, 30, and 79.)
- [79] A. Beghdadi, M.-C. Larabi, A. Bouzerdoun, and K. M. Iftekharruddin, "A survey of perceptual image processing methods," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 811-831, Sept. 2013. (Cited on pages 13 and 80.)
- [80] H. R. Wu, A. Reibman, W. Lin, F. Pereira, and S. S. Hemami, "Perceptual visual signal compression and transmission," *Proceeding of IEEE*, vol. 101, no. 9, pp. 2025-2043, Sep. 2013. (Cited on page 80.)
- [81] P. Ndjiki-Nya, D. Doshkov, H. Kaprykowsky, F. Zhang, D. Bull, T. Wiegand, "Perception-oriented video coding based on image analysis and completion: A review," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 579-594, Jul. 2012. (Cited on page 80.)
- [82] M. Tanimoto, "FTV: Free-viewpoint Television," *Signal Process.: Image Commun.*, vol. 27, no. 7, pp. 555-570, Feb. 2012. (Cited on page 23.)
- [83] Y. Zhao, Z. Chen, C. Zhu, Y.-P. Tan, and L. Yu, "Binocular just-noticeable-difference model for stereoscopic images," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 19-22, Jan. 2011. (Cited on pages 19, 89, and 119.)

- [84] M. Kaaniche, R. Gaetano, M. Cagnazzo and B. Pesquet-Popescu, Disparity estimation techniques, in *Emerging technologies for 3D video: content creation, coding, transmission and rendering*, Wiley Eds, January 2013. (Cited on page 43.)
- [85] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7-42, Apr.-Jun. 2002. (Cited on pages 41, 58, 113, and 116.)
- [86] H. Hirschmüller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582-1599, Sep. 2009. (Cited on pages 41 and 58.)
- [87] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, Nov. 2004. (Cited on pages 43, 57, 59, and 62.)
- [88] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008. (Cited on pages 43 and 59.)
- [89] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of the British Machine Vision Conference (BMVC)*, Cardiff, UK, pp. 384-396, 2002. (Cited on pages 43 and 59.)
- [90] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. of the 7th European Conference on Computer Vision (ECCV)*, Copenhagen, Denmark, pp. 128-142, 2002. (Cited on pages 43 and 59.)
- [91] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in *Proc. of the Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, pp. 506-513, 2004. (Cited on pages 43 and 59.)
- [92] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT flow: Dense correspondence across different scenes," in *Proc. of the 10th European Conference on Computer Vision (ECCV)*, Marseille, France, (pp. 28-42) (2008) (Cited on pages 43 and 59.)
- [93] L. Juan and O. Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing*, vol. 3, no. 4, pp. 143-152, 2009 (Cited on pages 43 and 59.)
- [94] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transact. Patt. Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615-1630, 2005. (Cited on pages 43 and 59.)

- [95] M. A. Fischler and R. C. Bolles, “Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, Jun. 1981. (Cited on pages 57 and 59.)
- [96] A. Vetro, T. Wiegand, and G. J. Sullivan, “Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard,” *Proc. of the IEEE*, vol. 99, no. 4, pp. 626-642, Apr. 2011. (Cited on pages 38, 51, 80, and 116.)
- [97] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, “3D video services with the emerging MVC standard,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 786015, 13 pages, 2009. (Cited on pages 24 and 25.)
- [98] D. Marpe, H. Schwarz, and T. Wiegand, “Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 620-636, 2003. (Cited on pages 51 and 116.)
- [99] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y. K. Wang, Joint multiview video model (JMVM) 8.0, in *Joint Video Team (JVT), Doc. JVT-AA207*, Geneva, Switzerland, 2008. (Cited on pages 51, 69, 70, 95, and 116.)
- [100] M. Pölönen, J. Hakala, R. Bilcu, T. Järvenpää, J. Häkkinen, and M. Salmimaa, “Color asymmetry in 3D imaging: Influence on the viewing experience,” *3D Research*, vol. 3, no. 3, pp. 1-10, 2012. (Cited on page 46.)
- [101] J. Chen, J. Zhou, J. Sun, and A. C. Bovik, “Binocular mismatch induced by luminance discrepancies on stereoscopic images,” in *IEEE International Conference on Multimedia and Expo (ICME 2014)*, pp. 1-6, Jul. 2014. (Cited on pages 26 and 46.)
- [102] M. T. Lambooi, W. A. IJsselstein, and I. Heynderickx, “Visual discomfort in stereoscopic displays: a review,” in *Proc. of Stereoscopic Displays and Virtual Reality Systems XIV*, San Jose, CA, USA, 2007. (Cited on pages 26 and 46.)
- [103] A. Woods, T. Docherty, and R. Koch, “Image distortions in stereoscopic video systems,” in *Proc. Stereoscopic Displays and Applications IV*, San Jose, CA, USA, 1993. (Cited on page 46.)
- [104] M. Salmimaa, J. Hakala, M. Pölönen, T. Järvenpää, R. Bilcu, and J. Häkkinen, “Luminance Asymmetry in Stereoscopic Content: Binocular Rivalry or Luster,” in *Proc. SID Symposium Digest of Technical Papers*, pp. 801-804, 2014. (Cited on pages 46 and 89.)

- [105] S. Winkler and D. Min, "Stereo/multiview picture quality: Overview and recent advances," *Signal Processing: Image Communication*, vol. 28, no. 10, pp. 1358-1373, 2013. (Cited on page 47.)
- [106] Y. S. Kang and Y. S. Ho, "An efficient image rectification method for parallel multi-camera arrangement," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1041-1048, 2011. (Cited on page 48.)
- [107] J. G. Lou, H. Cai, and J. Li, "A real-time interactive multi-view video system," in *Proc. of the 13th annual ACM international conference on Multimedia*, Singapore, pp. 161-170, 2005. (Cited on page 48.)
- [108] F. Zilly, M. Müller, P. Eisert, and P. Kauff, "The Stereoscopic Analyzer-An image-based assistance tool for stereo shooting and 3D production," in *Proc. 17th IEEE International Conference on Image Processing (ICIP 2010)*, Hong Kong, pp. 4029-4032, 2010. (Cited on page 48.)
- [109] F. Pedersini, A. Sarti, and S. Tubaro, "Accurate and simple geometric calibration of multi-camera systems," *Signal Processing*, vol. 77, no. 3, pp. 309-334, 1999. (Cited on page 48.)
- [110] H. Nanda, and R. Cutler, "Practical calibrations for a real-time digital omnidirectional camera," in *Technical Sketches, Computer Vision and Pattern Recognition*, 2001. (Cited on page 48.)
- [111] N. S. Joshi, B. Wilburn, V. Vaish, M. Levoy, and M. Horowitz, "Automatic Color Calibration for Large Camera Arrays," Technical Report CS2005-0821, UCSD CSE, 2005. (Cited on page 48.)
- [112] N. S. Joshi, "Color Calibration for Arrays of Inexpensive Image Sensors," Master's thesis, Stanford University, 2004. (Cited on pages 48, 49, and 89.)
- [113] A. Ilie and G. Welch, "Ensuring color consistency across multiple cameras," in *Proc. of International Conference on Computer Vision (ICCV)*, Washington, DC, USA, pp. 1268-1275, 2005. (Cited on page 49.)
- [114] K. Li, Q. Dai, and W. Xu, "Collaborative color calibration for multi-camera systems," *Signal Processing: Image Communication*, vol. 26, no. 1, pp. 48-60, 2011. (Cited on pages 20 and 49.)
- [115] V. Slavkovikj, "Color calibration of a multi-camera array," Master's thesis, Gjøvik university college, Norway, 2011. (Cited on page 49.)
- [116] V. Slavkovikj, J. Y. Hardeberg, and A. Eichhorn, "Characterizing the response of charge-couple device digital color cameras," in *Proc. SPIE Sensors, Cameras, and Systems for Industrial and Scientific Applications XIII*, Burlingame, California, USA, 2012. (Cited on page 49.)

- [117] J. Zhong, B. Kleijn, and X. Hu, Camera Control in Multi-camera Systems for Video Quality Enhancement. *IEEE Sensors Journal*, vol. 14, no. 9, pp. 2955-2966, 2014. (Cited on page 49.)
- [118] J. -H. Kim, P. -L. Lai, J. Lopez, A. Ortega, Y. Su, P. Yin, and C. Gomila, "New Coding Tools for Illumination and Focus Mismatch Compensation in Multiview Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1519-1535, 2007. (Cited on pages 50 and 51.)
- [119] J. -H. Kim, P. -L. Lai, A. Ortega, Y. Su, P. Yin, and C. Gomila, Results of CE2 on multi-view video coding, in *Joint Video Team (JVT), Doc. JVT-T117*, Klagenfurt, Austria, 2006. (Cited on page 50.)
- [120] K. Kamikura, H. Watanabe, H. Jozawa, H. Kotera, and S. Ichinose, "Global brightness-variation compensation for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 8, pp. 988-1000, 1998. (Cited on page 50.)
- [121] Hur, J. H., Cho, S., & Lee, Y. L. (2007). Adaptive local illumination change compensation method for H.264-based multiview video coding. *IEEE Trans. Circuits Syst. Video Technol.*, 17(11), 1496-1505. (Cited on pages 51 and 52.)
- [122] Y. L. Lee, J. H. Hur, Y. K. Lee, S. H. Cho, H. J. Kwon, N. H. Hur, and J. W. Kim, Results of CE2 on multi-view video coding, in *Joint Video Team (JVT) Doc. JVT-T110*, Klagenfurt, Austria, 2006. (Cited on page 51.)
- [123] S. Park, and D. Sim, Core experiment results on CE2 for multi-view video coding, in *Joint Video Team (JVT) Doc. JVT-T113*, Klagenfurt, Austria, 2006. (Cited on page 51.)
- [124] J. M. Boyce, "Weighted prediction in the H.264/MPEG AVC video coding standard," in *Proc. of IEEE Int. Symp. on Circuits Syst. (ISCAS)*, Vancouver, Canada, pp. 789-792, 2004. (Cited on page 51.)
- [125] K. Sohn, Y. Kim, J. Seo, J. Yoon, and G. Kim, Results on CE2 using IBDE for multi-view video coding, in *Joint Video Team (JVT) Doc. JVT-T114*, Klagenfurt, Austria, 2006. (Cited on pages 51 and 99.)
- [126] X. Li, L. Jiang, S. Ma, D. Zhao, and W. Gao, "Template based illumination compensation algorithm for multiview video coding," in *Proc. SPIE Conf. Visual Communications and Image Processing (VCIP 2010)*, Vol. 7744, Huangshan, China, Jul. 2010. (Cited on page 52.)
- [127] K. Yamamoto, M. Kitahara, H. Kimata, T. Yendo, T. Fujii, M. Tanimoto, S. Shimizu, K. Kamikura, and Y. Yashima, "Multiview video coding using view interpolation and color correction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1436-1449, Nov. 2007. (Cited on pages 52, 53, and 57.)

- [128] U. Fecker, M. Barkowsky, and A. Kaup, "Histogram-based pre-filtering for luminance and chrominance compensation of multi-view video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 9, pp. 1258-1267, Sep. 2008. (Cited on pages 53, 63, 68, and 70.)
- [129] Y. Chen, C. Cai, and J. Liu, "YUV correction for multi-view video compression," in *Proc. Int. Conf. Pattern Recognition (ICPR 2006)*, Hong Kong, Aug. 2006, pp. 734-737. (Cited on pages 53 and 54.)
- [130] F. Pitié, A. C. Kokaram, and R. Dahyot, "N-dimensional probability density function transfer and its application to color transfer," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, pp. 1434-1439, 2005. (Cited on pages 53 and 54.)
- [131] F. Pitié, A. C. Kokaram, and R. Dahyot, "Automated colour grading using colour distribution transfer," *Computer Vision and Image Understanding*, vol. 107, no. 1, pp. 123-137, 2007. (Cited on pages 53 and 54.)
- [132] D. Corrigan, F. Pitié, G. Marcin, G. Kearney, V. Morris, A. Rankin, M. Linnane, M. O'Deas, C. Leez, and A. Kokaram, "A Video Database for the Development of Stereo-3D Post-Production Algorithms," *Journal of Virtual Reality and Broadcasting*, 10, 2013. (Cited on page 55.)
- [133] D. S. Lee, M. S. Ko, Y. H. Seo, D. W. Kim, and J. Yoo, "Illumination compensation for multi-view video based on layered histogram matching with depth information," *Optics Communications*, vol. 286, pp. 74-84, 2013. (Cited on pages 53 and 55.)
- [134] Y. Chen, K. K. Ma, and C. Cai, "Histogram-offset-based color correction for multi-view video coding," in *Proc. of the 17th IEEE International Conference on Image Processing (ICIP)*, pp. 977-980, 2010. (Cited on pages 53 and 55.)
- [135] C. Doutre and P. Nasiopoulos, "Color Correction Preprocessing For Multiview Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 9, pp. 1400-1406, Sept. 2009. (Cited on pages 53, 55, 56, 57, and 60.)
- [136] F. Shao, G. Y. Jiang, M. Yu, and Y. S. Ho, "Fast color correction for multi-view video by modeling spatio-temporal variation," *Journal of Visual Communication and Image Representation*, vol. 21, no. 5, pp. 392-403, 2010. (Cited on pages 53 and 56.)
- [137] J. I. Jung and Y. S. Ho, "Color correction algorithm based on camera characteristics for multi-view video coding," *Signal, Image and Video Processing*, vol. 8, no. 5, pp. 1-12, 2014. (Cited on pages 53 and 56.)
- [138] J. I. Jung and Y. S. Ho, "Geometric and colorimetric error compensation for multi-view images," *Journal of Visual Communication and Image Representation*, vol. 25, no. 4, pp. 698-708, 2014. (Cited on pages 53 and 56.)

- [139] M. P. Tehrani, A. Ishikawa, S. Sakazawa, and A. Koike, "Iterative colour correction of multicamera systems using corresponding feature points," *Journal of Visual Communication and Image Representation*, vol. 21, no. 5, pp. 377-391, 2010. (Cited on pages 53, 57, and 60.)
- [140] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Englewood Cliffs, NJ, Prentice-Hall, 2007. (Cited on pages 53 and 115.)
- [141] S. A. Fezza, M.-C., Larabi, and K. M. Faraoun, "Feature-based Color Correction of Multiview Video for Coding and Rendering Enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1486-1498, 2014. (Cited on pages 57 and 58.)
- [142] S. A. Fezza, M.-C., Larabi, and K. M. Faraoun, "Coding and Rendering Enhancement of Multi-View Video By Color Correction," in *Proc. IEEE European Workshop on Visual Information Processing (EUVIP'2013)*, Paris, France, 2013. (Cited on pages 57 and 58.)
- [143] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564-577, Apr. 2003. (Cited on page 61.)
- [144] M. S. Khalid, M. U. Ilyas, M. S. Sarfaraz, and M. A. Ajaz, "Bhattacharyya coefficient in correlation of gray-scale objects," *Journal of Multimedia*, vol. 1, no. 1, pp. 56-61, Apr. 2006. (Cited on page 61.)
- [145] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004. (Cited on pages 61, 108, 114, and 125.)
- [146] P. Harding and N. M. Robertson, "Visual Saliency from Image Features with Application to Compression," *Cognitive Computation*, vol. 5, no. 1, pp. 76-98, Mar. 2013. (Cited on page 62.)
- [147] M. Nauge, M.-C. Larabi, and C. Fernandez-Maloigne, "A statistical study of the correlation between interest points and gaze points," in *Proc. SPIE Conf. Human Vision and Electronic Imaging XVII*, Jan. 2012. (Cited on page 62.)
- [148] P. Harding and N. M. Robertson, "A Comparison of Feature Detectors with Passive and Task-Based Visual Saliency," in *Proc. 16th Scandinavian Conference (SCIA 2009)*, Oslo, Norway, pp. 716-725, Jun. 2009. (Cited on page 62.)
- [149] Y. Su, A. Vetro, and A. Smolic, "Common test conditions for multiview video coding," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Document JVT-U211*, Hangzhou, China, Oct. 20-27, 2006. (Cited on pages 41, 65, 69, and 101.)

- [150] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on 3D Video Coding Technology," Doc. N12036, Geneva, Switzerland, Nov. 2011. (Cited on page 73.)
- [151] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations," *Color Research and Application*, vol. 30, no. 1, pp. 21-30, Feb. 2005. (Cited on page 69.)
- [152] S. Westland, C. Ripamonti, and V. Cheung, *Computational Colour Science Using MATLAB*. 2nd Edition, The Wiley-IS&T Series in Imaging Science and Technology, 2012. (Cited on page 69.)
- [153] Y.-L. Lee, J.-H. Hur, Y.-K. Lee, K.-H. Han, S. Cho, N. Hur, J. Kim, J.-H. Kim, P.-L. Lai, A. Ortega, Y. Su, P. Yin, and C. Gomila, "CE11: Illumination compensation," Joint Video Team (JVT) Doc. JVT-U052, Hangzhou, China, 2006. (Cited on page 70.)
- [154] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in *ITU-T SG 16 Q. 6 Video Coding Experts Group (VCEG), Document VCEG-M33*, Austin, TX, Apr. 2-4, 2001. (Cited on page 70.)
- [155] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in *Proc. of SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, San Jose, CA, USA, Jan. 2004, pp. 93-104. (Cited on pages 28 and 75.)
- [156] ISO/IEC JTC1/SC29/WG11, "Report on Experimental Framework for 3D Video Coding," Doc. N11631, Guangzhou, China, Oct. 2010. (Cited on page 75.)
- [157] H. Li and M. Flierl, "SIFT-based modeling and coding of background scenes for multiview soccer video," In *Proc. IEEE Int. Conf. on Image Process.*, pp. 1221-1224, Orlando, FL, 2012. (Cited on page 78.)
- [158] C. Bal, A. K. Jain, and T. Q. Nguyen, "Detection and removal of binocular luster in compressed 3D images," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pp. 1345-1348, Prague, Czech Republic, May 2011. (Cited on page 46.)
- [159] SA. Fezza, M-C. Larabi, and K-M. Faraoun, "Asymmetric Coding of Stereoscopic 3D based on perceptual significance," in *Proc. IEEE International Conference on Image Processing (ICIP'2014)*, Paris, France, 2014. (Cited on pages 80 and 90.)
- [160] SA. Fezza, M-C. Larabi, and K-M. Faraoun, "Asymmetric Coding using Binocular Just Noticeable Difference and Depth Information for Stereoscopic 3D," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2014)*, Florence, Italy, 2014. (Cited on pages 80 and 90.)

- [161] SA. Fezza, M-C. Larabi, and K-M. Faraoun, "Stereoscopic image quality metric based on local entropy and binocular just noticeable difference," in *Proc. IEEE International Conference on Image Processing (ICIP'2014)*, Paris, France, 2014. (Cited on page 116.)
- [162] D. V. Meegan, L. B. Stelmach, and W. J. Tam, "Unequal weighting of monocular inputs in binocular combinations: implications for the compression of stereoscopic imagery," *Journal of Experimental Psychology: Applied*, vol. 7, no. 2, pp. 143-153, Jun. 2001. (Cited on pages 80, 83, 84, and 85.)
- [163] R. Blake, D. H. Westendorf, and R. Overton, "What is suppressed during binocular rivalry?," *Perception*, vol. 9, no. 2, p. 223-231, 1980. (Cited on page 80.)
- [164] L. B. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2, pp. 188-193, Mar. 2000. (Cited on pages 80, 81, 82, 84, and 85.)
- [165] L. B. Stelmach and W. J. Tam, "Stereoscopic image coding: Effect of disparate image-quality in left-and right-eye views," *Signal Process: Image Commun.*, vol. 14, no. 1, pp. 111-117, Nov. 1998. (Cited on pages 80, 83, and 85.)
- [166] P. Seuntjens, L. Meesters, and W. IJsselstein, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Appl. Perception*, vol. 3, no. 2, pp. 95-109, Apr. 2006. (Cited on pages 80, 83, 84, 85, 91, 113, and 116.)
- [167] I.-L. Jung, T. Chung, K. Song, and C.-S. Kim, "Efficient stereo video coding based on frame skipping for real-time mobile applications," *IEEE Trans. on Consumer Electronics*, vol. 54, no. 3, pp. 1259-1266, 2008. (Cited on page 82.)
- [168] G. Saygili, C. G. Gurler, and A. M. Tekalp, "Evaluation of Asymmetric Stereo Video Coding and Rate Scaling for Adaptive 3D Video Streaming," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 593-601, Jun. 2011. (Cited on pages 84 and 88.)
- [169] M. G. Perkins, "Data compression of stereopairs," *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684-696, Apr. 1992. (Cited on pages 84, 86, and 87.)
- [170] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video," in *Proc. 3DTV-Conference*, Tampere, Finland, Jun. 2010. (Cited on pages 80, 85, 88, 89, and 95.)
- [171] V. De Silva, H. Arachchi, E. Ekmekcioglu, A. Fernando, S. Dogan, A. Kondoz, and S. Savas, "Psycho-physical limits of interocular blur suppression and its

- application to asymmetric stereoscopic video delivery,” in *Proc. 19th International Packet Video Workshop (PV 2012)*, Munich, Germany, pp. 184-189, May. 2012. (Cited on pages 80, 85, and 88.)
- [172] P. Aflaki, M. M. Hannuksela, and M. Gabbouj, “Subjective quality assessment of asymmetric stereoscopic 3-D video,” *Journal of Signal, Image and Video Processing*, 2013. (Cited on pages 80, 85, and 91.)
- [173] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. B. Akar, R. Civanlar, and A. M. Tekalp, “Temporal and spatial scaling for stereoscopic video compression,” in *Proc. 14th European Signal Processing Conference (EUSIPCO 2006)*, Florence, Italy, 2006. (Cited on pages 82, 83, and 85.)
- [174] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, and J. Kim, “Asymmetric coding of stereoscopic video for transmission over T-DMB,” in *Proc. 3DTV Conference*, Kos Island, Greece, May. 2007. (Cited on pages 85, 86, and 87.)
- [175] Y. Chen, S. Liu, Y.-K. Wang, M. M. Hannuksela, H. Li, and M. Gabbouj, “Low-complexity asymmetric multiview video coding,” in *Proc. International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, Apr. 2008, pp. 773-776. (Cited on pages 85, 86, and 87.)
- [176] A. K. Jain, A. E. Robinson, and T. Q. Nguyen, “Comparing Perceived Quality and Fatigue for Two Methods of Mixed Resolution Stereoscopic Coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, Mar. 2014. (Cited on pages 85, 87, 90, and 95.)
- [177] A. E. Robinson, A. K. Jain, M. Scott, D. MacLeod, and T. Q. Nguyen, “Apparent sharpness of 3D video when one eye’s view is more blurry,” *i-Perception*, vol. 4, no. 6, pp. 456-467, 2013. (Cited on pages 85 and 87.)
- [178] H. Brust, A. Smolic, K. Müller, G. Tech., and T. Wiegand, “Mixed Resolution Coding of Stereoscopic Video for Mobile Devices,” in *Proc. 3DTV-Conference*, Potsdam, Germany, 2009. (Cited on page 85.)
- [179] H. Brust, G. Tech, K. Müller, and T. Wiegand, “Mixed resolution coding with inter view prediction for mobile 3DTV,” in *Proc. 3DTV-Conference*, Tampere, Finland, Jun. 2010. (Cited on pages 84, 86, and 87.)
- [180] Y. Chen, Y.-K. Wang, M. Hannuksela and M. Gabbouj, “Picture-Level Adaptive Filter for Asymmetric Stereoscopic Video,” in *Proc. of the IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, California, USA, Oct. 2008, pp. 1944-1947. (Cited on page 86.)
- [181] Y. Chen, Y.-K. Wang, M. Gabbouj, and M. Hannuksela, “Regionally Adaptive Filtering for Asymmetric Stereoscopic Video Coding,” in *Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS 2009)*, Taipei, Taiwan, May. 2009, pp. 2585-2588. (Cited on page 86.)

- [182] “MPEG-4 Video Verification Model Version 18.0,” ISO/IEC JTC1/SC29/WG11, MPEG Doc. N3908, Jan. 2001 (Cited on pages 87 and 101.)
- [183] A. Aksay, C. Bilen, and G. B. Akar, “Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images,” in *Proc. 13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, Sep. 2005. (Cited on page 88.)
- [184] F. Shao, G. Jiang, X. Wang, M. Yu, and K. Chen, “Stereoscopic video coding with asymmetric luminance and chrominance qualities,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 6, pp. 2460-2468, Nov. 2010. (Cited on page 88.)
- [185] S. Ince and J. Konrad, “Recovery of a missing color component in stereo images,” in *Proc. SPIE Conf. Stereoscopic Displays and Applications*, vol. 5664, San Jose, CA, USA, Jan. 2005. (Cited on page 88.)
- [186] M. M. Subedar and L. J. Karam, “Increased depth perception with sharpness enhancement for stereo video,” in *Proc. SPIE Stereoscopic Displays and Applications XXI*, vol. 7524, San Jose, CA, USA, Feb. 2010. (Cited on pages 89 and 122.)
- [187] R. T. Held, E. A. Cooper, and M. S. Banks, “Blur and disparity are complementary cues to depth,” *Current Biology*, vol. 22, no. 5, pp. 426-431, 2012. (Cited on page 90.)
- [188] C. Porac, S. Coren, et al., *Lateral preferences and human behavior*. Springer-Verlag New York, 1981. (Cited on page 90.)
- [189] A. Klaus, M. Sormann, and K. Karner, “Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure,” in *Proc. Int. Conf. Pattern Recognition (ICPR 2006)*, Hong Kong, Aug. 2006, pp. 15-18. (Cited on pages 91 and 117.)
- [190] Q. Yang, “Recursive Bilateral Filtering,” in *Proc. European Conference on Computer Vision (ECCV 2012)*, Florence, Italy, Oct. 2012, pp. 399-413. (Cited on page 92.)
- [191] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” *Advances in Neural Information Processing Systems*, vol. 19, pp. 525-552, 2007. (Cited on page 92.)
- [192] T. Judd, F. Durand, and A. Torralba, “A Benchmark of Computational Models of Saliency to Predict Human Fixations,” Tech. Rep. MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Laboratory, Jan. 2012. (Cited on page 92.)

-
- [193] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 477-500, 2001. (Cited on page 93.)
- [194] G. Ananger, and T.D.C. Little, "A survey of technologies for parsing and indexing digital video," *Visual Communication: Image Representation*, vol. 7, no. 1, pp. 28-43, 1996. (Cited on page 93.)
- [195] F. Idris, and S. Panchanathan, "Review of image and video indexing techniques," *Visual Communication: Image Representation*, vol. 8, no. 2, 146-166, 1997. (Cited on page 93.)
- [196] Middlebury stereo vision website, <http://vision.middlebury.edu/stereo>. (Cited on page 95.)
- [197] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in ITU-T SG 16 Q. 6 Video Coding Experts Group (VCEG), Document VCEG-M33, Austin, TX, Apr. 2-4, 2001. (Cited on pages 40, 98, and 102.)
- [198] ITU-R Rec. BT.500, "Methodology for the subjective assessment of the quality of television pictures", 46 pages, Geneva, Switzerland, 2012. (Cited on page 98.)
- [199] Motion Picture Association of America (MPAA): Theatrical market statistics. <http://www.mpa.org/policy/industry>, 2012. (Cited on page 104.)
- [200] P. Hanhart, F. D. Simone, M. Rerabek, and T. Ebrahimi, "3D Video Quality Assessment," Chapter in F. Dufaux, B. Pesquet-Popescu and M. Cagnazzo, (Eds.), *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*, Wiley, 377-391, 2013. (Cited on pages 104, 106, and 112.)
- [201] W. Lin, and C. C. J. Kuo, "Perceptual visual quality metrics: a survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297-312, May. 2011. (Cited on page 104.)
- [202] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 660-668, Sept. 2008. (Cited on page 104.)
- [203] ISO 3664:2000 Viewing conditions-graphic technology and photography. Technical report, ISO, Geneva, Switzerland, 2000. (Cited on page 105.)
- [204] ITU, "Methodology for the subjective assessment of the quality of television pictures," in Recommendation ITU-R BT.500-10, International Telecommunication Union., 2000. (Cited on page 105.)

- [205] ITU, "Subjective Assessment of Stereoscope Television Pictures, in Recommendation ITU-R BT.1438," International Telecommunication Union, 2000. (Cited on page 106.)
- [206] W. Chen, F. Jérôme, M. Barkowsky, P. L. Callet, "Methodologies for Assessing 3D QoE: Standards and Explorative Studies," *ZTE communications*, vol. 11, no. 1, pp. 2-10, 2013. (Cited on page 106.)
- [207] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA, USA: Morgan and Claypool, 2006. (Cited on pages 18 and 107.)
- [208] B. Girod, "What's wrong with mean-squared error?," In Andrew B. Watson, editor, *Digital images and human vision*, pp. 207-220. MIT press, London, England, 1993. (Cited on page 108.)
- [209] Z. Wang, Alan C Bovik, and L. Lu, "Why is image quality assessment so difficult?," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, Orlando, FL, USA, May. 2002. (Cited on page 108.)
- [210] Zhou Wang and Alan C Bovik. "Mean squared error: love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, 2009. (Cited on page 108.)
- [211] D. M. Chandler and S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284-2298, 2007. (Cited on pages 108 and 109.)
- [212] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, 2002. (Cited on pages 108 and 109.)
- [213] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, 312-322, 2004. (Cited on page 109.)
- [214] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. of the Second International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, USA, 2006. (Cited on pages 108 and 109.)
- [215] 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, 2010. (Cited on page 110.)
- [216] Y. Wang, "Survey of objective video quality measurements," *EMC Corporation Hopkinton, MA, 1748*, 2006. (Cited on page 110.)

- [217] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE International Conference on Image Processing (ICIP'2002)*, pp. 477-480, 2002. (Cited on page 110.)
- [218] X. Li, "Blind image quality assessment," in *Proc. IEEE International Conference on Image Processing (ICIP'2002)*, 2002. (Cited on page 110.)
- [219] H. R. Sheikh and A. C. Bovik, "No-reference quality assessment using natural scene statistics: JPEG2000," in *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1918-1927, 2005. (Cited on page 110.)
- [220] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *Proc. IEEE International Conference on Image Processing (ICIP'2005)*, pp. 141-144, 2005. (Cited on page 110.)
- [221] A. K. Moorthy, and A. C. Bovik, "A two-stage framework for blind image quality assessment," in *Proc. IEEE International Conference on Image Processing (ICIP'2010)*, pp. 2481-2484, 2010. (Cited on page 110.)
- [222] M. Carnec, P. Le Callet, D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE International Conference on Image Processing (ICIP'2005)*, pp. 185-188, 2005. (Cited on page 114.)
- [223] S. Wolf, and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2005)*, 2005. (Cited on page 110.)
- [224] S. Yang, "Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients," *Electronics Letters*, vol. 47, no. 6, pp. 382-383, Mar. 2011. (Cited on page 110.)
- [225] M. Tagliasacchi, G. Valenzise, M. Naccari, and S. Tubaro, "reduced-reference structural similarity approximation for videos corrupted by channel errors," *Multimedia Tools and Applications*, vol. 48, no. 3, pp. 471-492, Jul. 2010. (Cited on page 110.)
- [226] Z. Wang, and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. SPIE, Human Vision and Electronic Imaging*, 2005. (Cited on page 110.)
- [227] P. Campisi, P. L. Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proc. European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, Sep. 2007. (Cited on page 113.)
- [228] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *Electronics Letters*, vol. 44, no. 16, pp. 963-965, Jul. 2008. (Cited on page 113.)

- [229] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondo, "Quality analysis for 3D video using 2D video quality models," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 1969-1976, Nov. 2008. (Cited on page 113.)
- [230] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," in *Proc. SPIE Stereoscopic Displays and Applications XIX*, San Jose, CA, USA, Jan. 2008. (Cited on pages 113 and 125.)
- [231] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing*, vol. 2008, Jan. 2009. (Cited on pages 114 and 125.)
- [232] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2010)*, Scottsdale, AZ, USA, Jan. 2010. (Cited on pages 114 and 125.)
- [233] S. S. Sarikan, R. F. Olgun, and G. B. Akar, "Quality evaluation of stereoscopic videos using depth map segmentation," in *Proc. Int. Workshop QoMEX*, Mechelen, Belgium, Sep. 2011, pp. 67-71. (Cited on page 114.)
- [234] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo, "Objective quality assessment method of stereo images," in *Proc. IEEE 3DTV-Conference (3DTV-CON 2009)*, Potsdam, Germany, May 2009. (Cited on page 114.)
- [235] C. Hewage and M. Martini, "Reduced-reference quality metric for 3D depth map transmission," in *Proc. IEEE 3DTV-Conference (3DTV-CON 2010)*, Tampere, Finland, Jun. 2010. (Cited on pages 114, 115, and 125.)
- [236] Z. M. P. Sazzad, S. Yamanaka, Y. Kawayoke, and Y. Horita, "Stereoscopic image quality prediction," in *Proc. Int. Work. QoMEX*, San Diego, CA, USA, Jul. 2009. (Cited on pages 114 and 115.)
- [237] Z. M. P. Sazzad, R. Akhter, J. Baltes, and Y. Horita, "Objective no-reference stereoscopic image quality prediction based on 2D image features and relative disparity," *Advances in Multimedia*, ID 256130, Jan. 2012. (Cited on page 114.)
- [238] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 7, pp. 1332-1343, Nov. 2011. (Cited on page 114.)
- [239] R. Bensalma, M.-C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Systems and*

- Signal Processing*, Vol. 24, no. 2, pp 281-316, Jun. 2013. (Cited on pages 113 and 114.)
- [240] X. Wang, S. Kwong, and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," in *Proc. IEEE Visual Commun. Image Process.*, Tainan, Taiwan, Nov. 2011, pp. 1-4. (Cited on pages 114 and 115.)
- [241] A. Maalouf and M.-C. Larabi, "CYCLOP: A stereo color image quality assessment metric," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2011)*, May 2011, pp. 1161-1164. (Cited on pages 114 and 115.)
- [242] M.-J. Chen, C.-C. Su, D.-L. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, Vol. 28, no. 9, , pp. 1143-1155, Oct. 2013. (Cited on pages 114, 115, 124, 125, and 126.)
- [243] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-Reference Quality Assessment of Natural Stereopairs," *IEEE Transactions on Image Processing*, vol.22, no.9, pp.3379-3391, Sept. 2013. (Cited on pages 114, 115, 125, and 126.)
- [244] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual Full-Reference Quality Assessment of Stereoscopic Images by Considering Binocular Visual Characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940-1953, May. 2013. (Cited on pages 90, 114, and 115.)
- [245] S. Ryu, K. Sohn, "No-Reference Quality Assessment for Stereoscopic Images Based on Binocular Quality Perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, Apr. 2014. (Cited on pages 114, 115, and 121.)
- [246] J. Wang and Z. Wang, "Perceptual quality of asymmetrically distorted stereoscopic images: the role of image distortion types," in *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2014)*, Chandler, AZ, USA, Jan. 2014, 29-31. (Cited on pages 114 and 115.)
- [247] B. Julesz, *Foundations of Cyclopean Perception*. Univ. Chicago Press, 1971. (Cited on pages 115 and 116.)
- [248] W. J. M. Levelt, *On Binocular Rivalry*, Mouton, The Hague, Paris, 1968. (Cited on pages 115 and 116.)
- [249] I. P. Howard and B. J. Rogers, *Binocular Fusion and Rivalry in Binocular Vision and Stereopsis*, New York, USA: Oxford Univ. Press, 1995. (Cited on pages 115 and 116.)

- [250] R. Blake, D. H. Westendorf, and R. Overton, "What is suppressed during binocular rivalry?," *Perception*, vol. 9, no. 2, p. 223-231, 1980. (Cited on pages 115 and 116.)
- [251] S. B. Steinman, B. A. Steinman, and R. P. Garzia, *Foundations of Binocular Vision: A Clinical Perspective*, New York, USA: McGraw-Hill, 2000. (Cited on page 115.)
- [252] M. Shahid, A. Rossholm, B. Lövsström, and H. J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 40, no. 1, Aug. 2014. (Cited on page 110.)
- [253] Z. Wang, Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185-1198, May. 2011. (Cited on page 117.)
- [254] S. Gabarda and G. Cristobal, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am.*, vol. 24, no. 12, pp. B42-B51, 2007. (Cited on page 117.)
- [255] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517-526, Feb. 2012. (Cited on page 117.)
- [256] R. Soundararajan and A. C. Bovik, "Survey of information theory in visual quality assessment," *Signal, Image and Video Processing*, vol. 7, no. 3, pp. 391-401, May. 2013. (Cited on page 117.)
- [257] 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, Nov. 2010. (Cited on pages 19 and 119.)
- [258] Z. Wang, E.P. Simoncelli, A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems and Computers*, vol. 2, Nov. 2003, pp. 1398-1402. (Cited on pages 109, 121, and 125.)
- [259] S. Ryu and K. Sohn, "No-reference perceptual blur model based on inherent sharpness," in *Proc. IEEE International Conference on Image Processing (ICIP'14)*, Paris, France, Oct. 2014. (Cited on pages 110 and 122.)
- [260] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678-2683, Sep. 2011. (Cited on pages 17, 18, 110, and 122.)

-
- [261] C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 934-945, Mar. 2012. (Cited on pages 110 and 122.)
- [262] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006. (Cited on pages 112 and 124.)
- [263] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000. (Cited on page 112.)

Publications

International Journal

- “Feature-based Color Correction of Multiview Video for Coding and Rendering Enhancement,” *IEEE Trans. Circuits Syst. Video Technol.* (**T-CSVT**), vol. 24, no. 9, pp. 1486-1498, 2014.

International Conferences

- “Stereoscopic Image Quality Metric Based on Local Entropy and Binocular Just Noticeable Difference,” in *Proc. IEEE International Conference on Image Processing (ICIP’2014)*, Paris, France, 2014. **(recognized as the Top 10% papers.)**
- “Asymmetric Coding of Stereoscopic 3D based on perceptual significance,” in *Proc. IEEE International Conference on Image Processing (ICIP’2014)*, Paris, France, 2014. **(recognized as the Top 10% papers.)**
- “Asymmetric Coding using Binocular Just Noticeable Difference and Depth Information for Stereoscopic 3D,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’2014)*, Florence, Italy, 2014.
- “Coding and Rendering Enhancement of Multi-View Video By Color Correction,” in *Proc. IEEE European Workshop on Visual Information Processing (EUVIP’2013)*, Paris, France, 2013.
- “New Prediction Structure for Stereoscopic Video Coding Based on the H.264/AVC Standard,” in *Proc. International Conference on Digital Information and Communication Technology and its Applications (DICTAP’2011)*, edited by LNCS Springer, pp. 762-769, Dijon, France, 2011.
- “A Comparison of Prediction Structures for Multi-view Video Coding based on the H.264/AVC Standard,” in *Proc. IEEE International Workshop on Systems Signal Processing and their Applications (WoSSPA’2011)*, pp. 111-114, Tipaza, Algeria, 2011.

- “Stereoscopic Video Coding Based on the H.264/AVC Standard,” in *Proc. IEEE International Conference on Machine and Web Intelligence (ICMWI’2010)*, pp. 476- 478, Algiers, Algeria, 2010.

Tutorials

- “Understanding binocular perception in the context of asymmetric coding and evaluation of S-3D,” in *IEEE European Workshop on Visual Information Processing (EUVIP’2014)*, Paris, France, 2014.

Stereo and Multi-view Video Coding Extensions of the H.264/AVC Standard: from Coding to Quality of Experience

Abstract : Over the last few years, 3D video has been gaining interest with a fast pace. It can be regarded as a type of visual media that expand the capabilities of a 2D video by providing depth sensation of the observed scene. 3D video applications require Multi-View content (MVV) with a very large number of views, captured by several synchronized cameras from different viewpoints. As this huge amount of data is making its way to consumers, its transmission over existing bandwidth-limited infrastructure is a great challenge. Consequently, efficient compression is essential to store or transmit MVV streams. Based on these considerations, Multi-view Video Coding (MVC), an extension of the H.264/MPEG-4 AVC standard, has been developed to address the encoding of MVV content. MVC takes advantage of the redundancies existing between views to achieve high coding performance. However, one major problem with MVV systems is linked to illumination and color variations between views. These color mismatches may impair both coding performance and rendering quality. To fix this issue, we propose a preprocessing method based on an improvement of the histogram matching for correcting these color discrepancies between views. Furthermore, this thesis tackles the problem of stereoscopic video (*i.e.*, 2 views only) coding by relying on an asymmetric coding method that exploits the binocular suppression theory. For this point, based on recent findings on binocular perception, a fully automated model that dynamically determines the best level of asymmetry is proposed, thus offering optimal 3D visual experience. Finally, in 3D imaging systems, it is desirable to assess degradations and to predict the perceived quality of 3D content in line with the human visual system (HVS). Therefore, reliable quality evaluation metrics are really needed, in order to guarantee the final 3D viewing experience. To this end, this thesis presents three novel quality assessment metrics for stereoscopic images. Thanks to the integration of the perceptual binocular characteristics, these metrics handle effectively the asymmetric distortions of stereoscopic images.

In light of these developments, relying on the perceptual characteristics of the HVS, several methods have been presented in this thesis, thus covering the end-to-end 3D video system, from preprocessing to quality assessment, through coding, with the aim to provide an optimal quality of experience.

Keywords : 3D Video, Multi-view Video, Stereoscopic Video, Multi-view Video Coding, H.264/MPEG-4 AVC, Asymmetric Coding, Color Correction, View Rendering, Depth, Binocular Just Noticeable Difference, Quality of Experience.

Extension de H.264/AVC au Stéréoscopique et le Multi-Vues: du Codage à la Qualité de l'Expérience

Résumé : L'avènement de la 3D stéréoscopique a suscité un grand intérêt à la fois chez les professionnels et le grand public. Nul doute que cet intérêt est lié à l'effet immersif et l'impression de profondeur que procure la 3D. Au-delà des données stéréoscopiques, les vidéos multi-vues (*MVV : Multi-View Video*) permettent d'étendre les capacités de la 3D avec un visionnage sans lunette et un accès à différents points de vue de la scène. Ces séquences MVV sont constituées de plusieurs vidéos captées avec un ensemble de caméras. Cependant, cela se fait au prix d'une taille de données considérable endiguant leur transmission sur des infrastructures avec une bande passante limitée. Par conséquent, l'usage d'une solution de compression efficace devient crucial pour stocker ou transmettre les flux MVV. C'est pourquoi la norme H.264/MPEG-4 AVC a été étendue avec le profil Multi-view Video Coding (MVC). Ce dernier exploite la forte corrélation existant entre les vues en plus de la corrélation temporelle. Cependant, les systèmes d'acquisitions multi-vues peuvent générer une non-uniformité de luminance et/ou de chrominance entre les vues. Cette variation de couleur peut affecter négativement les performances de codage et la qualité du rendu visuel. Pour pallier à ce problème, nous proposons une méthode efficace de prétraitement basée sur une amélioration significative de la méthode *histogram matching*. En outre, cette thèse s'attèle au problème de la compression des vidéos stéréoscopiques en proposant une méthode de codage asymétrique exploitant la théorie de suppression binoculaire. En nous basant sur les récents résultats sur la perception binoculaire, nous proposons un modèle entièrement automatisé qui détermine dynamiquement le meilleur niveau d'asymétrie entre les vues, offrant ainsi une expérience visuelle 3D optimale. Enfin, dans les systèmes d'imagerie 3D, l'évaluation de la qualité perçue est une étape primordiale pour la validation des solutions développées. Même si l'évaluation subjective reste l'approche la plus fiable, il est indispensable de disposer de métriques de qualité se rapprochant du jugement humain. C'est ainsi que nous avons proposé trois nouvelles métriques de qualité dédiées aux images stéréoscopiques. Grâce à l'intégration des caractéristiques de la perception binoculaire, ces métriques évaluent efficacement les distorsions asymétriques dans les images stéréoscopiques.

À la lumière de ces développements et en s'appuyant sur les caractéristiques perceptuelles du SVH, plusieurs méthodes ont été présentées dans cette thèse, couvrant de bout en bout la chaîne d'exploitation des vidéos 3D, allant du prétraitement à l'évaluation de la qualité, en passant par le codage, et tout cela dans le but de fournir une qualité d'expérience optimale.

Mots clés : Vidéo 3D, Vidéo Multi-Vues, Vidéo Stéréoscopique, Codage de Vidéos Multi-Vues, H.264/MPEG-4 AVC, Codage Asymétrique, Correction de Couleur, Rendu de Vue, Profondeur, JND-Binoculaire, Qualité de l'Expérience.
