UNIVERSIDAD POLITÉCNICA DE MADRID

POLITÉCNICA

**ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN**

DOCTORAL THESIS

**Visual attention and perception models for
assessing quality in 2D and 3D stereoscopic
video**

JUAN PEDRO LÓPEZ VELASCO
Ingeniero de Telecomunicación

2016

SIGNALS, SYSTEMS AND RADIOCOMMUNICATIONS
DEPARTMENT
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

DOCTORAL THESIS

**Visual attention and perception models for
assessing quality in 2D and 3D stereoscopic
video**

Author:
Juan Pedro López Velasco
Telecommunications Engineer – PhD Candidate

Advisor:
José Manuel Menéndez García
Telecommunications Engineer - PhD

2016

DEPARTAMENTO DE SEÑALES, SISTEMAS Y
RADIOCOMUNICACIONES

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TESIS DOCTORAL

**Visual attention and perception models for
assessing quality in 2D and 3D stereoscopic
video**

Autor:

Juan Pedro López Velasco
Ingeniero de Telecomunicación

Director:

José Manuel Menéndez García
Dr. Ingeniero de Telecomunicación

2016

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE

TELECOMUNICACIÓN

DEPARTAMENTO DE SEÑALES, SISTEMAS Y RADIOCOMUNICACIONES

“Visual attention and perception models for assessing quality in 2D and 3D stereoscopic video”

Autor: Juan Pedro López Velasco

Director: José Manuel Menéndez García

El tribunal nombrado para juzgar la tesis arriba indicada el día 8 de febrero de 2016, compuesto de los siguientes doctores:

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Realizado el acto de lectura y defensa de la Tesis Doctoral, en Madrid, a de de 2016, acuerdan otorgarle la calificación de:

Calificación _____

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO

Resumen

La medida de calidad de vídeo sigue siendo necesaria para definir los criterios que caracterizan una señal que cumpla los requisitos de visionado impuestos por el usuario. Las nuevas tecnologías, como el vídeo 3D estereoscópico o formatos más allá de la alta definición, imponen nuevos criterios que deben ser analizadas para obtener la mayor satisfacción posible del usuario.

Entre los problemas detectados durante el desarrollo de esta tesis doctoral se han determinado fenómenos que afectan a distintas fases de la cadena de producción audiovisual y tipo de contenido variado. En primer lugar, el proceso de generación de contenidos debe encontrarse controlado mediante parámetros que eviten que se produzca el discomfort visual y, consecuentemente, fatiga visual, especialmente en lo relativo a contenidos de 3D estereoscópico, tanto de animación como de acción real. Por otro lado, la medida de calidad relativa a la fase de compresión de vídeo emplea métricas que en ocasiones no se encuentran adaptadas a la percepción del usuario. El empleo de modelos psicovisuales y diagramas de atención visual permitirían ponderar las áreas de la imagen de manera que se preste mayor importancia a los píxeles que el usuario enfocará con mayor probabilidad.

Estos dos bloques se relacionan a través de la definición del término *saliencia*. *Saliencia* es la capacidad del sistema visual para caracterizar una imagen visualizada ponderando las áreas que más atractivas resultan al ojo humano. La *saliencia* en generación de contenidos estereoscópicos se refiere principalmente a la profundidad simulada mediante la ilusión óptica, medida en términos de distancia del objeto virtual al ojo humano. Sin embargo, en vídeo bidimensional, la *saliencia* no se basa en la profundidad, sino en otros elementos adicionales, como el movimiento, el nivel de detalle, la posición de los píxeles o la aparición de caras, que serán los factores básicos que compondrán el modelo de atención visual desarrollado.

Con el objetivo de detectar las características de una secuencia de vídeo estereoscópico que, con mayor probabilidad, pueden generar discomfort visual, se consultó la extensa literatura relativa a este tema y se realizaron unas pruebas subjetivas preliminares con usuarios. De esta forma, se llegó a la conclusión de que se producía discomfort en los casos en que se producía un cambio abrupto en la distribución de profundidades simuladas de la imagen, aparte de otras degradaciones como la denominada “violación de ventana”. A través de nuevas pruebas subjetivas centradas en analizar estos efectos con diferentes distribuciones de profundidades, se trataron de concretar los parámetros que definían esta imagen. Los resultados de las pruebas demuestran que los cambios abruptos en imágenes se producen en entornos con movimientos y disparidades negativas elevadas que producen interferencias en los procesos de acomodación y vergencia del ojo humano, así como una necesidad en el aumento de los tiempos de enfoque del cristalino.

En la mejora de las métricas de calidad a través de modelos que se adaptan al sistema visual humano, se realizaron también pruebas subjetivas que ayudaron a determinar la importancia de cada uno de los factores a la hora de enmascarar una determinada degradación. Los resultados demuestran una ligera mejora en los resultados obtenidos al aplicar máscaras de

ponderación y atención visual, los cuales aproximan los parámetros de calidad objetiva a la respuesta del ojo humano.

Abstract

Video quality assessment is still a necessary tool for defining the criteria to characterize a signal with the viewing requirements imposed by the final user. New technologies, such as 3D stereoscopic video and formats of HD and beyond HD oblige to develop new analysis of video features for obtaining the highest user's satisfaction.

Among the problems detected during the process of this doctoral thesis, it has been determined that some phenomena affect to different phases in the audiovisual production chain, apart from the type of content. On first instance, the generation of contents process should be enough controlled through parameters that avoid the occurrence of visual discomfort in observer's eye, and consequently, visual fatigue. It is especially necessary controlling sequences of stereoscopic 3D, with both animation and live-action contents. On the other hand, video quality assessment, related to compression processes, should be improved because some objective metrics are adapted to user's perception. The use of psychovisual models and visual attention diagrams allow the weighting of image regions of interest, giving more importance to the areas which the user will focus most probably.

These two work fields are related together through the definition of the term *saliency*. *Saliency* is the capacity of human visual system for characterizing an image, highlighting the areas which result more attractive to the human eye. Saliency in generation of 3DTV contents refers mainly to the simulated depth of the optic illusion, i.e. the distance from the virtual object to the human eye. On the other hand, saliency is not based on virtual depth, but on other features, such as motion, level of detail, position of pixels in the frame or face detection, which are the basic features that are part of the developed visual attention model, as demonstrated with tests.

Extensive literature involving visual comfort assessment was looked up, and the development of new preliminary subjective assessment with users was performed, in order to detect the features that increase the probability of discomfort to occur. With this methodology, the conclusions drawn confirmed that one common source of visual discomfort was when an abrupt change of disparity happened in video transitions, apart from other degradations, such as window violation. New quality assessment was performed to quantify the distribution of disparities over different sequences. The results confirmed that abrupt changes in negative parallax environment produce accommodation-vergence mismatches derived from the increasing time for human crystalline to focus the virtual objects.

On the other side, for developing metrics that adapt to human visual system, additional subjective tests were developed to determine the importance of each factor, which masks a concrete distortion. Results demonstrated slight improvement after applying visual attention to objective metrics. This process of weighing pixels approximates the quality results to human eye's response.

Glossary

2D	Two Dimensions
3D	Three Dimensions
3DTV	Three-Dimensional Television
AC/A	Accommodation Convergence to Accommodation
ACR	Absolute Category Rating
AVM	Acommodation/Vergence Mismatches
BT	Broadcasting Television
COST	European Cooperation in Science and Technology
DOF	Depth Of Focus
DSSIM	Structural Dissimilarity
EU	European Union
FR	Full Reference
GATV	Grupo de Aplicación de Telecomunicaciones Visuales
HD	High Definition
HDTV	High Definition Television
HS	Hyperstereoscopy
HVS	Human Visual System
ITU	International Telecommunication Unit
ITU-R	International Telecommunication Union - Radiocommunication Standardization Sector
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
MOS	Mean Opinion Score
MSE	Mean squared error
NP	Negative Parallax
NR	No Reference
P-	Negative Parallax
P+	Positive Parallax
PCC	Pearson Pearson linear Correlation Coefficient
PP	Positive Parallax
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
ROI	Region of Interest
RTVE	Spanish Radio and Television Corporation
S3D	Stereoscopic Three-Dimensional
SAD	Sum of Absolute Differences
SI	Spatial Information
SNR	Signal-to-Noise Ratio
SPI	Spatial Parallax Information
SPIE	Society of Photo-Optical Instrumentation Engineers
SS	Single Stimulus

SSIM	Structural similarity
STD	Standard Deviation
SVT	Sveriges Television
SWV	Stereoscopic Window Violation
TI	Temporal Information
TPI	Temporal Parallax Information
TV	Television
VD	Visual Discomfort
VQA	Video Quality Assessment
VQEG	Video Quality Experts Group
WV	Window Violation
Y	Luminance
ZoC	Zone of Comfort
ZP	Zero Parallax

Table of Contents

RESUMEN	I
ABSTRACT	I
GLOSSARY	I
TABLE OF CONTENTS	3
LIST OF FIGURES	7
LIST OF TABLES	13
1 INTRODUCTION.....	17
2 PROBLEM DESCRIPTION	21
2.1 Sources of visual fatigue and importance of visual attention.....	22
2.2 3D stereoscopic video problems.....	23
2.3 Main sources of visual discomfort	23
2.4 Most common artifacts in encoded video sequences	25
2.4.1 Blurring and motion blur	26
2.4.2 Blocking artifact.....	26
2.4.3 Staircase effect and ringing.....	27
2.5 Non-correlation between objective metrics and subjective opinions in quality assessment	28
2.6 The problem with the lack of reference video	30
3 OBJECTIVES AND WORK DEVELOPMENT.....	35
3.1 Objectives.....	35
3.2 Work Development	36
3.3 Databases used for tests	37
3.3.1 Settings of video sequences used	40
3.3.2 Spatio-temporal information for video sequences	41
3.4 Video encoding evolution: from MPEG-2 to H.265.....	44
3.4.1 MPEG-1/2 standards	44
3.4.2 MPEG-4 Visual and H.264/AVC standards.....	45
3.4.3 H.265 HEVC (High Efficiency Video Coding)	46
3.4.4 Specifications in different codecs codification.....	47
3.5 Techniques for video quality degradation and distortion inclusion	48
4 VISUAL DISCOMFORT PREDICTION IN 3D STEREOSCOPIC VIDEO	53
4.1 Motivation for this work	53

4.1.1	Human Visual System	54
4.1.2	Visual Discomfort	58
4.2	State of the art for video quality assessment related to 3D stereoscopic video	69
4.3	Preliminary subjective studies for defining objectives	72
4.4	Model for predicting visual discomfort in 3D video sequences.....	76
4.4.1	Subjective assessment	79
4.4.2	Determination of depth maps and parallax histograms	82
4.4.3	Parameters of disparity complexity: Spatial Parallax Information (SPI) and Temporal Parallax Information (TPI)	84
4.4.4	Detection of Stereoscopic Window Violation (SWV)	86
4.4.5	Final Algorithm for Predicting Visual Discomfort	87
4.5	Results of Visual Discomfort Prediction	88
4.6	Conclusions to visual discomfort prediction	106
5	VISUAL ATTENTION MODEL FOR VIDEO QUALITY ASSESSMENT	111
5.1	Work motivation	111
5.2	State of the art for psychovisual models.....	113
5.3	Psychovisual Four-Component Model Based on Video Characterization.....	115
5.3.1	Introduction.....	115
5.4	Degradation of video sequences.....	118
5.4.1	Motion masks.....	121
5.4.2	Spatial complexity and high frequencies masks.....	122
5.4.3	Position of pixels in spatial sections.....	122
5.4.4	Face detection	123
5.4.5	Range of colors detection	123
5.5	Artifact mask generation based on the psychovisual model	123
5.6	Description of Subjective tests for psychovisual model.....	126
5.6.1	Methodology description	126
5.6.2	Video Database and Session Description	127
5.6.3	Subjective assessment for pixels weight definition	128
5.6.4	Masks for video degradation.....	131
5.7	Definition of the psychovisual model through four-components analysis	131
5.7.1	Psychovisual model definition.....	131
5.8	Use of visual attention model in objective quality metrics.....	135

5.8.1	Advanced Blurring FR objective metric	135
5.8.2	Other FR objective metric	136
5.9	<i>Results</i>	136
5.10	Conclusions to visual attention models	143
6	CONCLUSIONS	147
7	MERITS	151
7.1	Publications derived from the work of this doctoral thesis	151
7.1.1	3DTV Quality and Visual Discomfort	151
7.1.2	Video Quality Assessment and Visual Attention Models	151
7.1.3	Book chapters.....	152
7.2	Other publications.....	152
7.2.1	Peer-reviewed publications.....	152
7.2.2	Lecture Notes	152
7.3	Other merits: research projects	153
	ACKNOWLEDGEMENT	157
	REFERENCES	161
	ANNEXES	173
	Annex 1. Video Quality Metrics Definition	173
	Annex 2. Software for Video Quality Metrics Definition	175
	Annex 3. Information relative to subjective assessment with pairs of 3D stereoscopic sequences.....	177
	Annex 3.1. Depth maps in the transition (previous frame and next frame) and evolution of percentages of positive and negative parallax in the transition.....	177
	Annex 3.2. Temporal Information of sequences (frame by frame)	195
	Annex 3.3. Spatial Information of sequences (frame by frame).....	199
	Annex 3.4. Evolution of percentages of positive parallax (blue bar) and negative parallax (yellow bar) in the transition	203

List of Figures

Fig. 1. Broadcast production process.....	24
Fig. 2. Sharp frame from sequence “City” (left) and its blurred version (right)	26
Fig. 3. Gradient image in sharp (left) and blurred (right) frames from sequence “City”	26
Fig. 4. Detail of blocking artifact in sequence “Cathedral”	27
Fig. 5. Detail of staircase effect in diagonal edges from sequence “Barrier”	28
Fig. 6. Examples of ringing effect in sequences “Barrier” (left) and “Stewie” (right)	28
Fig. 7. Diagram of a Full-Reference system	30
Fig. 8. Diagram of a Reduced-Reference (RR) system	30
Fig. 9. Diagram of a No-Reference system.....	31
Fig. 10. Methodology followed for visual discomfort prediction	36
Fig. 11. Methodology followed for video quality assessment with visual attention models development	37
Fig. 12. Example of sequences used: “News Report”, “Rain Fruits”, “Fountain” and “ParkJoy”	38
Fig. 13. Video sequences resolution	40
Fig. 14. Example of the four parameters defined for each sequence in the diagram	43
Fig. 15. Spatio-temporal diagram for the sequences used	43
Fig. 16. Example of comparison of macroblock partitions in H.264/AVC (left, extracted from [RICHARDSON-1, 2004]) and HEVC (right, extracted from [SONNATI-1, 2015])	46
Fig. 17. Detail of frame from original not-encoded Crowd sequence (left) and its corresponding gradient image (right)	47
Fig. 18. Image “Crowd” (left) with its gradient representation (right), with MPEG-2 encoding	48
Fig. 19. H.264 encoded image with deblocking filter (left) and without deblocking filter (right) in homogeneous areas	48
Fig. 20. Example of cyclopean eye functioning.....	55
Fig. 21. Disparity interpretation over a stereoscopic screen	56
Fig. 22. Parallax possibilities: positive, zero or negative.....	57
Fig. 23. Crystalline accommodation for a far or near distance object.....	60
Fig. 24. The accommodation-vergence conflict when viewing a stereoscopic display	60
Fig. 25. Differences in vergence and accommodation process from real world to a 3D display (image taken from [HOFFMAN-1, 2008])	61

Fig. 26. Shibata's Zone of Comfort with Distance in diopters (D) and meters (m) (extracted from [SHIBATA-1, 2011])	62
Fig. 27. Response to a stimulus in accommodation and vergence in real world	63
Fig. 28. Response to a stimulus in accommodation and vergence in a 3D Display	63
Fig. 29. Area equivalent to Panum's fusional area (part of retina) where stereopsis occurs. "F" is the fixation point.	64
Fig. 30. Blur in real world versus image in 3D Display	65
Fig. 31. Keystone distortion (left) and depth plane curvature (right) caused by (weakly) convergent camera (taken from [BOEV-1, 2008]).....	66
Fig. 32. Cardboard effect (right) in comparison to its original depth map (left) (extracted from [BOEV-1, 2008])	66
Fig. 33. Sequence "Itaca 3D" with spinning logo and moving ship with Window Violation (W.V) on the left side of image	68
Fig. 34. Sequence "Itaca 3D" with Window Violation (W.V) on the right side of image	68
Fig. 35. Examples of left frames in sequence "Modernism"	73
Fig. 36. Examples of camera separation to create parallax variation.....	73
Fig. 37. Effect with anaglyph examples of parallax variation in sequence "Palco HD"	73
Fig. 38. Example of abrupt transition from "Positive Parallax" image (P.P) to "Negative Parallax" image (N.P) with their distribution of parallax histogram	74
Fig. 39. Results in Transitions between types of parallax (P.P: Positive Parallax, and N.P: Negative Parallax)	75
Fig. 40. Results in impact related to motion in stereoscopic sequences (low, medium and fast motion).....	75
Fig. 41. Results from sequences without Window Violation (W.V) or W.V. in lateral sides or down area	75
Fig. 42. Results derived from a sequence with parallax progressive or abrupt variations.....	75
Fig. 43. Conceptual depiction of spatial factors: (a) Max, average and range; (b) Screen; (c) Spatial complexity and (d) Depth position (image extracted from [CHOI-1, 2012]).....	76
Fig. 44. Scheme of objects with different parallax	77
Fig. 45. Transition from content in planes with positive parallax to planes with negative parallax.....	78
Fig. 46. Transition from content in planes with negative parallax to planes with positive parallax.....	78
Fig. 47. Image with content in different planes and motion	78
Fig. 48. Negative Parallax with fast motion	79

Fig. 49. Subjective assessment related to the appearance of visual discomfort for each sequence.	81
Fig.50. Depth map. (P refers to “parallax”)	84
Fig. 51. Example of parallax histogram corresponding to positive parallax image with hyperstereoscopy.....	84
Fig.52. Working scheme for detecting Window Violation with exaple of sequence "Itaca 3D"	86
Fig. 53. Level of parallax scale of to understand disparity diagrams, commonly called “depth maps”	88
Fig. 54. Parallax for transition in pixel [900:480] of sequence 12, marked in red	89
Fig. 55. Response to a stimulus (transition in sequence 1) in accommodation and vergence in a 3D Display.....	90
Fig. 56. Viewing near objects at different distances (Sequence 1).....	91
Fig. 57. Temporal Information frame by frame in transition of “Sequence 1”	91
Fig. 58. Scores from subjective assessment to "Sequence 1"	92
Fig. 59. Histograms of sequence 12, from frame in video12a (up) to frame in video 12b (down)	93
Fig. 60. Parallax for transition in pixel [982:563] of sequence 12, marked in red	93
Fig. 61. Temporal Information frame by frame of “Sequence 12”	94
Fig. 62. Scores from subjective assessment to "Sequence 12"	94
Fig. 63. Scores from subjective assessment to "Sequence 14”	95
Fig. 64. Disparity histogram of last frame in video 14a, where the transition of sequence 14 occurs	96
Fig. 65. Disparity histogram of first frame in video 14b, where the transition of sequence 14 occurs	96
Fig. 66. Scores from subjective assessment to "Sequence 16”	97
Fig. 67. Temporal Information frame by frame of “Sequence 16”	97
Fig. 68. Disparity histograms in transition of sequence 16.....	98
Fig. 69. Parallax for transition in pixel [965:542] of sequence 16	98
Fig. 70. Disparity histograms in transition of Sequence 2	99
Fig. 71. Parallax for transition in pixel [132:401] of sequence 2, marked in red	100
Fig. 72. Scores from subjective assessment to "Sequence 2”	100
Fig. 73. Scores from subjective assessment to "Sequence 4”	101
Fig. 74. Scores from subjective assessment to "Sequence 13”	102
Fig. 75. Disparity histograms for Sequence 13	103

Fig. 76. Parallax for transition in pixel [942:206] of sequence 13, marked in red	103
Fig. 77. Parallax evolution in complete "Palco HD" sequence in pixel [942:206]	104
Fig. 78. Temporal Information frame by frame of "Sequence 20"	105
Fig. 79. Parallax for transition in pixel [1044:607] of sequence 20, marked in red	105
Fig. 80. Scores from subjective assessment to "Sequence 20"	106
Fig. 81. Visual saliency analyzed by eye tracker (image taken from [WOODING-1, 2002]) .	111
Fig. 82. Example of high-detailed video sequence "Crowd" (left) with its Canny mask highlighting detailed areas in red (right)	116
Fig. 83. TI on different motion levels: sequence with fast motion (left) and low motion (right)	117
Fig. 84. Saliency diagram based in motion in sequence "Barrier" (frame 310)	117
Fig. 85. Segmentation in domains of a picture (left) and example in sequence "Boxers" (right)	118
Fig. 86. Face detected in "Umbrella" sequence (left) and its facial mask (right)	118
Fig. 87. Scheme of video sequence generation	119
Fig. 88. Artificially impaired video sequence "News Report" with high quality in faces detected and poor quality in other areas.	119
Fig. 89. Video sequence "News Report" with poor quality in faces and high quality in other areas, inverse of previous sequence.	120
Fig. 90. Diagram for feature mask sequence and its inverse mask	121
Fig. 91. "Barrier" sequence after applying motion mask	121
Fig. 92. Canny algorithm for spatial complexity mask	122
Fig. 93. Corner, center and lateral masks	122
Fig. 94. Distribution of sections to divide video frames	122
Fig. 95. Range of red colors mask in sequence "Barrier" (frame 125)	123
Fig. 96. Example of Sequence "Umbrella" with impairment located in faces ROI and its inverse	124
Fig. 97. Block diagram for impairment insertion process	125
Fig. 98. Detail of frame 125 of sequence "Umbrella" after applying different distortions: original, blurring (Gaussian filter), ringing (JPEG200 encoding), blocking (8x8 mosaic filter). .	125
Fig. 99. Motion Mask in sequence News Report (frame 0)	132
Fig. 100. Detail Mask in sequence News Report (frame 0)	132
Fig. 101. Faces Mask in sequence News Report (frame 0)	132
Fig. 102. Pixel Position Mask in sequence News Report (frame 0)	132

Fig. 103. Percentage of influence of each pixel characteristic to the psychovisual model ..	133
Fig. 104. Motion Mask in sequence News Report (frame 0)	134
Fig. 105. Detail Mask in sequence News Report (frame 0).....	134
Fig. 106. Faces Mask in sequence News Report (frame 0)	134
Fig. 107. Pixel Position Mask in sequence News Report (frame 0).....	134
Fig. 108 Complete psychovisual model for first frame from sequence "News Report" (frame 0).	134
Fig. 109. Content-weighted mask (right) corresponding to original frame (left) in sequence "Boxers" (frame 90)	137
Fig. 110. Content-weighted mask (right) corresponding to original frame (left) in sequence "Boxers" (frame 275)	137
Fig. 111. Content-weighted mask (right) corresponding to original frame (left) in sequence "Phone Call" (frame 131)	137
Fig. 112. Content-weighted mask (right) corresponding to original frame (left) in sequence "Tree Branches" (frame 16)	137
Fig. 113. PCC for Blur and Advanced Blur for sequence "Boxers"	141
Fig. 114. PCC for Blur and Advanced Blur for sequence "Hall"	141
Fig. 115. PCC for Blur and Advanced Blur for sequence "Phone Call"	141
Fig. 116 PCC for Blur and Advanced Blur for sequence "Tree Branches"	142
Fig. 117. Interface for 3D Analysis	175
Fig. 118. Interface for mask development.....	175
Fig. 119. Interface for TI/SI diagrams.....	176

List of Tables

Table 1. Classification of visual artifacts	25
Table 2. Sequences with same MSE value and affected by different distortions, from left to right: JPEG encoding, salt and pepper noise, defocus filter and blocking artifact.....	29
Table 3. Description of most representative sequences used.....	38
Table 4. Basic characteristics and settings of video sequences used	41
Table 5. Levels of the scale used in tests	79
Table 6. Basic characteristics and settings of device	80
Table 7. Video Sequences Used in Video Quality Assessment	82
Table 8. Depth maps for transition in sequence 1.....	89
Table 9. Depth maps for transition in sequence 12.....	92
Table 10. Depth maps for transition in sequence 14.....	94
Table 11. Depth maps for transition in sequence 16.....	96
Table 12. Depth maps for transition in sequence 2.....	99
Table 13. Depth maps for transition in sequence 4.....	101
Table 14. Depth maps for transition in sequence 13.....	102
Table 15. Depth maps for transition in sequence 20.....	104
Table 16. Results of different metrics for representative frames applied to different sequences in cases of artificially impaired mask. D means that distortion is located in that area, and “Inv” indicates that distortion is located in the inverse of that area.....	126
Table 17. Levels of the scale used in tests	126
Table 18. Basic characteristics and setting of devices	127
Table 19. Basic characteristics and settings of video sequences used	128
Table 20. Results from subjective tests with artificially impaired sequences and impairments/distortions located only in determined regions of interest corresponding to factors analyzed	129
Table 21. Results for impairment located in faces ROI for artificially distorted sequence “News Report”	130
Table 22. Results for impairment located in motion ROI’s for artificially distorted sequence “Barrier”	130
Table 23. Results for impairment located in position ROI’s for artificially distorted sequence “Crowd”	130

Table 24. Coefficients of influence to the psychovisual model depending on the ROI belonging..... 133

Table 25. Results for advanced blur metric applied to sequences in H.264 at different bitrates. In addition, Advanced MSE results are collected, these results correspond to applying the psychovisual model to MSE metric. 139

Table 26. Pearson linear Correlation Coefficient for Blur and Advanced Blur metrics 140

1. Introduction

1 Introduction

Quality of Experience (QoE) is defined as the degree of delight or annoyance of the user of an application or service [LE CALLET-1, 2012]. QoE results from the fulfilment of the user's expectations with respect to the utility and or enjoyment of the application or service in the light of their personality and current state.

Video quality measurement plays an important role for most image processing applications, especially related to internet and mobile video consumption, but it is still a difficult task because of the weaknesses of the error sensitivity based framework and the complication in finding models which adjust perfectly to viewer's vision. Video quality is also dependent of different artifacts and degradation derived from processes of data compression.

While the core technologies have been developing, there is still much to be done to optimize the broadcast systems to deliver the best possible visual output. Having perceptually acceptable and high quality contents on display is a challenging task.

Color pixel encodings have the objective of reducing the amount of information by restricting the range of color and light that can be represented and by optimizing the use of code values for storage or distribution purposes. Uncompressed videos require too high bit rates, which hinders the storage and transmission for digital broadcast through a limited bandwidth channel. In order to reduce the size of the video and fulfil the requirements of broadcasting limitations, video compression techniques are used. These techniques reduce the redundant information in the video signal, but also generate a collection of artifacts that deteriorates the quality of experience.

Estimation of quality for all visual content is the key factor in optimization and design. The vast majority of objective video metrics aim at close approximation of the quality as perceived by the user. The properties presented by an ideal quality metric would be the following:

1. Perceptual. This property means that the metrics are related to the human visual system (HVS) and its functioning.
2. Objectivity. This property means that the result provides a numerical representation of the quality as perceived by the user, and does not change depending on the environment and instant of calculation.
3. Reliability. This property means that the metric is able to predict the perceptual quality for wide variety of content, as perceived by a large amount of users.

However, not all the reduction in QoE is a consequence of video compression. Content generation and acquisition processes are a key factor for satisfying the user's requirements. With the advances in technologies, it is now possible for content creators to capture more

dynamic, colorful and realistic images with powerful production tools. The combination of these features offers more than improvement in image quality, but it is not enough if content generators do not control the limits and areas where comfort is assured for the observer. Stereoscopic artifacts not only produce visually unpleasant results, but are also known to cause visual discomfort and eye-strain, which decreases the quality of experience. Thus, it is important to quantify the probability of this effect to occur.

The first step towards objective quality estimation metric is to identify the artifacts and sources of visual discomfort, which could arise in various usage scenarios. This doctoral thesis aims to develop metrics with the mentioned properties, which improves the results by approximating to human eye's response. On the other hand, research works on quality of stereo images attempt to quantify the typical distortions that could occur in stereoscopic video sequences, and to design tools to reduce visual discomfort occurrence as possible.

2. Problem Description

2 Problem Description

Image quality assessment is a difficult process, which plays a major role in various processing applications [WANG-3, 2002]. Video analysis is even a more complex process, because video represents a sequence of static images in motion. Much research has been developed in this field for defining metrics and algorithms to predict estimated quality of video sequences. This research means an extension of the work developed in image quality assessment, but considering correlation between consecutive frames. Several factors derived from motion and the correlation among images mentioned, should be taken into account for ensuring the perceived quality. These factors, which could be seen as a disadvantage and a complication of simple metrics, are really an advantage, because it is common that the high amount of information play an important role in the development of more efficient algorithms.

Visual quality measurement is not always a precise science because of the influence of so many factors that modify the results obtained, especially focused on characteristics of the Human Visual System. For that reason, it is necessary that the final observers identify and assess the displayed video images with the objective of evaluating and comparing video communication systems.

Consequently, visual quality is mainly a subjective task, which has a high cost in time and effort that hinders the accuracy of the results, but they could be automatized with metrics that analyze a collection of factors affecting to HVS. Among these subjective factors, the observer's opinion depends on the task itself and the environment. Another important reason on why video quality assessment is a so difficult field is the weakness of error sensitivity derived from tests conditions used by most quality evaluation approaches.

Environmental factors affects to perception. For example, three different examples of environment can be analyzed (watching passively a movie, participating in a videoconference and analyzing surveillance video scenes); the characteristics of the video could be the same, but the demands of the observer for each case are completely different.

On the other hand, objective video quality assessment criteria gives accurate and repeatable results, but these results should reproduce the subjective experience of a human observer, which, as mentioned, is not so repeatable due to the surrounding conditions in visualization of video contents. [WANG-2, 2004].

A great effort is being made these recent years in the development of image quality algorithms, which correlate with perceived quality assessment. Due to its difficulty, this task has achieved only limited success.

Human eyes extract structural information from the viewing field and human visual system adapted for this purpose. Precisely for that reason, we need a new philosophy in designing

video quality assessment, which may be adapted to local content, measuring distortion as approximated as possible to perceived video quality.

In this section, the main sources of visual discomfort and video artifacts are analyzed as the main problem of video quality assessment. Additionally, the problem of the non-correlation with human visual system of existing video quality metrics is necessary to be analyzed.

The problems related to the lack of a reference to compare with are introduced as one additional topic that hinders the video quality assessment process. This lack justifies the necessity of developing no-reference metrics, which are also introduced in this section, as opposite to full-reference metrics, which have been traditionally used in this field.

Finally, the problems derived from new generation media, concretely stereoscopic three-dimensional video, are mentioned. The techniques used in quality assessment of this type of video were used as the basis of new theories for focusing the attention in determined regions with specific characteristics.

2.1 Sources of visual fatigue and importance of visual attention

In the literature, involving visual fatigue and discomfort when viewing video contents, it is usual to find different terms that are related to the Human Visual System (HVS) and psychophysical features. Among these common terms, fatigue and visual attention are used to determine the quality of the experience.

The word “fatigue” refers to *the reluctance to continue a task as a result of physical or mental exertion or a prolonged period of performing the same task* [BEIRNESS-1. 2005]. There are several factors, which induce the observer to present visual fatigue. These factors are significant in 3D stereoscopic video, but it could be present also in 2D video because of visual artifacts performance.

An important attribute of the human visual system is visual attention. This attribute has not been explored adequately in terms of image quality. Subjective results demonstrate that visual attention cannot be ignored for improving performance of video quality metrics. Attention models are necessary to be employed to find fixation points and regions of interest.

Salient objects and regions in natural scenes are commonly regarded as areas, which the HVS will typically focus on. Detecting these areas in region-based approaches is the key step in visual attention. Some factors, such as motion or detail level, are remarkable as an increase in user’s interest. These factors are quantified and ordered in visual attention models, which are considered to be the most suitable for predicting fixation points.

In 3D stereoscopic video, the saliency models are based not only in motion or detail level, but also in simulated depth. The most remarkable areas of each image will be the ones that are closer to the observer, i.e. the regions presenting the highest disparities in terms of negative parallax.

2.2 3D stereoscopic video problems

Stereoscopic three-dimensional contents offer an experience, which is more immersive than conventional 2D media. This next generation media have a collection of advantages, but also some factors, which uncontrolled, could generate problems to the observer. The viewing safety, especially in terms of visual comfort, should be ensured for increasing the quality of experience. Among these factors, it should be taking into account for this purpose disparity distribution, depth, perceptual and cognitive inconsistency, quick changes, excessive disparities, mismatch between left and right views and vergence conflicts [JUNG-3, 2015].

Human Visual System is sensitive to orderly regions for visual perception based on several factors, such as motion, regular structures and highlighted areas. HVS focuses on salient regions on an image, which is related to sensitive areas. To achieve this objective and obtain accurate sensitivity maps, pattern masking was used.

In 3D stereoscopic video, sensitivity maps are focused on saliency estimation, but other factors influence on visual perception. These factors are not specific for 3D video, as happened with saliency, so they should be taken into account when analyzing sensitive maps for 2D video.

The work in Quality of Experience for stereoscopic contents could be referred to different research fields: analysis of disparity, distribution of parallax, motion, encoding process, artifacts detection, packet loss in transmission, quality of the screen in fusion of views, detection of window violations or differences between left and right view. This study was centered in the content generation practices, i.e. the first step in the video production chain, for native stereoscopic video in both natural and synthetic capture.

The ability of human eye to focus objects located at different distances affects to the perceived quality. Abrupt changes cause human eye needs to change the focus position at different distances rapidly. This effect must be analyzed over the work collected in this thesis.

2.3 Main sources of visual discomfort

The next step towards objective quality estimation is to identify the main sources of visual discomfort in 2D video, which are related to visual artifacts that could arise when analyzing video contents. Most of these artifacts are derived from the necessity of compressing video information for transmission, but also there are factors that are a consequence of an uncontrolled content generation process.

Visual artifacts (also called artefacts) are defined as anomalies during visual representation of digital graphics, imagery or video. There are different kinds of artifacts depending on the source of production and the phase in the broadcasting chain in which they are caused. For example, a compression artifact is a noticeable distortion of media (including images, audio, and video) caused by the application of lossy data compression. Compression artifacts are very common in video coding and reduction of computational cost, and may be consequence of some processes to which video sequences are submitted, such as block-based coding, Gaussian blur, down-sampling or transmission errors.

The following stages of video broadcasting dataflow might be sources of visual artifacts, which also appear on scheme in Fig. 1 when describing the broadcast production chain:

- 1) **Content creation.** This phase includes both processes of natural or unnatural content creation. Additionally, the postproduction processes are included. In this phase, positioning of cameras and rendering parameters are the main sources of discomfort and artifacts generation. Settings in camera capture related to focus, color and brightness should be taken into account to avoid quality distortions.
- 2) **Compression or video coding/decoding processes.** Different coding schemes of video utilize temporal or spatial techniques with the purpose of minimizing the transmission cost, while reducing the redundancy.
- 3) **Transmission losses.** A common problem related to digital video, especially in wireless transmission, is packet losses. Some algorithms exist to evaluate the impact of this factor in the final quality perceived by the observer.
- 4) **Visualization on a display or screen.** The problems in this phase depend on each kind of displays or screens, which have their own artifacts and characteristics. The artifacts associated to visualization are dependent on the scene.

Fig. 1. Broadcast production process

In Table 1, the main artifacts related to each phase of the broadcast production process are collected. These artifacts affect to different characteristics of content nature, such as color, structure, motion or the possibility of 3D representation and binocularity. These settings allow structuring the distortions in four different groups, as defined in works by Boev *et al* [BOEV-1, 2008], to classify the artifacts from monoscopic to stereoscopic video.

Table 1. Classification of visual artifacts

	Capture	Representation - Conversion	Coding	Transmission - Error Resilience	Visualization
Structure	Blur by defocusing Barrel distortions Pincushion distortions Interlacing Temporal and spectral aliasing Down-sampling Noise introduction	Temporal and spatial aliasing Line replication	Blocking artifacts Blurring (Mosaic patterns) Staircase effect Ringing	Data loss Data distortion Jitter	Flickering Resolution limitations Aspect ratio distortions Display geometry distortions Spatial aliasing by subsampling on non-rectangular grid
Color	Chromatic aberration Vignetting-decreasing intensity		Cross-color artifacts Color bleeding	Color bleeding	Contrast range Color representation Baking and long-term use Viewing angle dependent color representation Rainbow artifact
Motion	Motion blur Temporal mismatch		Motion compensation artifacts <i>Mosquito</i> noise Judder	Loss distortion in motion Jitter	Smearing Blurring and judder
Binocular	Depth plane curvature Keystone-distortion Cardboard effect	Ghosting by distribution Perspective binocular rivalry ("WOW" artifacts)	Cross distortion Cardboard effect Depth bleeding, depth ringing	Data loss, one channel Data loss, propagating	Shear distribution Ghosting by crosstalk Angle dependent binocular aliasing Accommodation convergence rivalry Lattice artifacts Puppet theater effect Picket fence effect Image flipping (Pseudoscopic Image)

2.4 Most common artifacts in encoded video sequences

Once the sources of visual discomfort are defined, the first step towards objective quality assessment is to define concrete artifacts. This section aims to overview, describe and categorize the main artifacts, which occur during delivery video, especially related to the coding process.

For each artifact we explain the main effects and how to detect it. After that, we discuss the causes for occurrence. Where possible, a figure exemplifying the artifact is provided as well.

2.4.1 Blurring and motion blur

Blur is one of the most common artifacts detected in video sequences. Shapes of the objects and edges are observed as defocused and blurred. When this artifact is high, it may be difficult to distinguish foreground and background, as seen on Fig. 2.

Fig. 2. Sharp frame from sequence “City” (left) and its blurred version (right)

Different factors can be the main cause of its occurrence. In coding processes, blur appears by the loss of energy when quantifying at low levels. On the other hand, in capture it appears when motion is higher than the finite number of frames per second taken by a CCD sensor. In addition, defocusing of cameras in capture influences in the generation of this artifact, which may produce visual discomfort to spectators. The focus of the camera must lie within the depth of field to assure that an object is sharply mapped on the sensor.

As mentioned, in video coding blurring artifact appears when considerably reducing the encoding bitrate, presenting a defocus sensation on the image. The effect has a higher impact in highly detailed areas and high frequency components and, as a consequence, in edges. For example, when applying a Sobel filter to estimate the gradient in the previous sequences, it is necessary to obtain a homogenization of the objects’ contours, as shown in Fig. 3.

Fig. 3. Gradient image in sharp (left) and blurred (right) frames from sequence “City”

2.4.2 Blocking artifact

The blocking artifact (also named tiling) is a block boundary effect that appears in the coding process based on the usage of block-based schemes, like the Discrete Cosine Transform

(DCT). Blocking appears because of a severe decrease of bitrate, as happens in MPEG-2 or H.264 coding. When compression is high, the blocking artifact increases, making its effect more noticeable to the human eye. The blocking generates a sensation of mosaic in every single frame, derived from the homogenization of values in pixels contained in a macroblock. An example is shown in Fig. 4.

The artifact is defined as discontinuities at the borders of each block (or macroblock) in the reconstructed video frame. H.264 or MPEG-2 treat individually each block of pixels in the coding stage according to its content, and there is where the origin of blocking lies. These blocks have variable sizes. Blocking is visible when error introduced through coarse quantification is higher than the masking ability of the human visual system [LOPEZ-4, 2013].

Fig. 4. Detail of blocking artifact in sequence "Cathedral"

Blocking artifact is more noticeable in smoothly textured areas because the human eye is more sensitive to low spatial frequencies. The suppression of higher quantization coefficients and the consequent loss of information cause the blocking artifact.

Additionally, mosaic patterns detection is a variation of this artifact when reducing high frequency components in horizontal and vertical direction, revealing a similar distortion, which produces a *pixelation* effect in the image.

2.4.3 Staircase effect and ringing

Apart from blocking or blurring effects, discontinuities at the borders lead to a staircase effect, especially at diagonal edges, tilted by a certain angle. The reconstruction of the image in high frequency components during the decoding phase reveals the importance of edge information. Vertical and horizontal edges, with directions parallel to the framework, respond better to quantization than diagonal edges, which are presented as stairs as a result of the alignment of patterns derived from DCT transform (Fig. 5).

Fig. 5. Detail of staircase effect in diagonal edges from sequence “Barrier”

On the other hand, the ringing artifact is a consequence of Gibb’s phenomenon. Areas with high contrast are the source of wave shapes and shimmering parallel to the contours due to coarse quantization of high frequency components, as seen in examples from Fig. 6. These two artifacts often occur together.

Fig. 6. Examples of ringing effect in sequences “Barrier” (left) and “Stewie” (right)

2.5 Non-correlation between objective metrics and subjective opinions in quality assessment

While researchers have made significant progress in objective quality assessment by generating algorithms able to detect artifacts and distortion, the main problem is that pooling stage lacks theoretical principles and reliable computational models adapted to the human visual system. Consequently, viewing natural images requires optimal perceptual weights defined in psychovisual and visual attention models, which should be proportional to local information content [WANG-3, 2002].

Most of the objective video quality metrics treat the pixels equally, independently to content, as happens with MSE, SNR or PSNR. The description of those metrics is detailed in the annexes section.

MSE (Mean Squared Error) is an estimator that measures the average of the squares of the difference between the distorted image and the original frame pixel by pixel, calculated for every frame (see Annex 1 for definition).

SNR (Signal-to-Noise Ratio) is an estimator that measures the average of the squares of the difference, i.e. the MSE, compared to the values of the reference video sequence, calculated for every frame (see Annex 1 for definition).

Finally, PSNR (Peak Signal-to-Noise Ratio) is an estimator that measures the average of the squares of the difference, i.e. the MSE, compared to the highest possible value, which is the peak (maximum) value of the reference video sequence, calculated for every frame (see Annex 1 for definition).

In the following example, all distorted images have roughly the same mean squared error (MSE) values with respect to the original image, but very different quality.

Table 2. Sequences with same MSE value and affected by different distortions, from left to right: JPEG encoding, salt and pepper noise, defocus filter and blocking artifact.

MSE = 144	MSE = 144	MSE = 144	MSE = 144

As seen in the example of Table 2, these four images have the same value of MSE, but subjectively obtain different scores for the observers. This effect appears because the nature of the local content has high importance for the observer, and the treatment of pixels is not equal in every single frame.

Other objective metrics, such as SSIM (Structural Similarity), take advantage of the content itself and the inner structure of the image for obtaining an indicator of quality which is closer to the perceived quality detected by human eye [WANG-2, 2004]. Additionally, its inverse DSSIM (Structural Dissimilarity) is necessary in different contexts (see Annex 1 for definition).

Applying saliency and visual attention models to objective quality metrics, different algorithms detect the regions of interest for the user in each instant. Trying to assure the objective quality in the areas that the observer is most probably watching, the probability of increasing quality of experience is higher. For this reason, subjective results demonstrate that visual attention is an important tool to improve the performance of video quality metrics.

2.6 The problem with the lack of reference video

The usual metrics developed for video assessment, such as PSNR and MSE (see Annex 1 for definition), require a video reference as a way of comparison to detect the degradations on the image, which is called Full-Reference (FR) assessment. Typically, full-reference systems evaluate a video system or its features.

The problem of these algorithms is that the lack or the availability of a reference exists in some concrete environments, especially the ones that imply mobility, and the ones related to internet and streaming transmissions. That justifies the importance of the No-Reference (NR) assessment methods and the interest in working in this field. No-Reference metrics are usually the basis on the measurement of artifacts in sequences containing only the desired artifact signal or a combination of artifact signals, with description of techniques to detect blurring, blocking effect and ringing.

According to the presence of a reference video, different kinds of assessment may be defined. Depending on the availability of a reference video sequence, which allow the comparison of characteristics with the degraded sequence, it is possible to identify three groups:

- *Full Reference (FR) metrics.* This case implies the possibility of using the complete reference sequence to evaluate the degradation on an image (Fig. 7).

Fig. 7. Diagram of a Full-Reference system

- *Reduced Reference (RR) metrics.* This type of quality metrics are necessary when only some attributes or features of the reference sequences are available (Fig. 8). The image feature could be any information of the original video sequence that could help the metric to clarify the nature of the original image. For example the original resolution if the image is scaled, a segmentation image for object recognition or the mean and maximum of brightness and contrast to identify deviations.

Fig. 8. Diagram of a Reduced-Reference (RR) system

- *No Reference (NR) metrics.* In this case no references are available. There is no original image or any features of it, and the degraded image and its perception by the human visual system is the only tool available to draw conclusions about the

quality of the image. This kind of metric is more complicated, but of vital importance in environments in which references are difficult to provide, as is the case in mobile or internet multimedia services (Fig. 9).

Fig. 9. Diagram of a No-Reference system

The vast majority of published works are about Full-Reference (FR) metrics, i.e. when the ideal original signal without degradation is available. However, the lack of reference in the mobile video environment increases the difficulty. That is why No-Reference (NR) metrics, also known as blind algorithms, are necessary. NR metrics attempt to assess signal quality without using any information other than the distorted signal. There is little substantive work on this topic, which makes this work more challenging and interesting.

3. Objectives and Work development

3 Objectives and Work Development

The term quality of experience could be equivalent to the addition of other definitions such as visual comfort, fatigue reduction or elimination of visual artifacts. This large quantity of definitions shows the complexity of the term. Because of this complexity, it is necessary to concrete the objectives of this thesis and define the methodology to follow for work development.

3.1 Objectives

The main problems detected for this research, involving video quality assessment, were analyzed in previous chapter. After considering factors for 2D and 3D stereoscopic video, the objective defined for this doctoral thesis are the following.

For visual comfort assessment:

- Detecting the main sources of visual discomfort in 3D stereoscopic video after developing subjective assessment.
- Quantifying the cases of sequences where the probability of visual discomfort to occur is higher.
- Analyzing the factors of motion, distribution of parallax and abrupt disparity change in pairs of sequences for developing tools that correspond to human perception.
- Demonstrate with sequences that the results obtained in subjective assessment may be predicted with objective parameters and characteristics measurement.

For visual attention and saliency models:

- Improving objective quality metrics by applying visual attention models, which weight regions of interest to obtain results closer to human eye's response.
- Determining accurate visual attention models, particular for each sequence, which predict the most probable areas observed by the user.
- Weighting the *saliency* factors analyzed by the use of subjective assessment. These saliency factors are the following: motion, level of detail, face detection and position of pixel. In future works, other factors could be considered, such as color or textures.
- Demonstrating the improvement of the objective metrics for measuring quality and artifacts in the sequence when applying the developed visual attention model.

3.2 Work Development

The work development for obtaining the previous objectives is described in this section. The methodology for analyzing visual discomfort is schematized in Fig. 10. The phases for this process are the following:

- **Preliminary subjective assessment.** First results reveal the characteristics to analyze in the following phases, related to quality perceived by the users and their opinions when viewing 3D stereoscopic video in different conditions, and with a series of features.
- **Determination of visual discomfort sources.** The results after subjective assessment are needed to determine visual discomfort sources. These results will be the basis of new subjective assessment adjusted to the parameters for being analyzed.
- **Characterization of video sequences.** The new database based on parameters detected should be characterized analyzing factors as temporal information or distribution of disparities during each video sequence.
- **Statistics analysis, new subjective assessment, metrics development and drawing conclusions.** The characterization of video sequences is used to determine the features of a source of visual discomfort. The initial hypothesis are the core of results to be demonstrated with conclusions.

Fig. 10. Methodology followed for visual discomfort prediction

For the visual attention models applied to video quality assessment, the methodology will be different, but with much in common with the previous one. This methodology is schematized in Fig. 11 and the phases explained next.

- **Determining saliency factors: motion, face detection, level of detail and position of pixels.** The factors that influence saliency models are selected in the first phase.

Other factors, such as color or textures, are discarded for future work. Once these factors are selected, the sequences are artificially impaired to isolate each parameter to analyze their influence in the visual attention model.

- **Subjective assessment with artificially impaired sequences.** A database with artificially impaired sequence is used for subjective assessment to obtain the opinion of the user and weighting each of the previous factors, creating masks for each factor and artifact to analyze.
- **Weighting these factors in order of importance.** Results obtained in subjective assessment will reveal the order of importance of each for the final visual attention model.
- **Visual attention model generation.** The model is generated as a combination of masks for each factor. The visual attention model is individual for each frame of the video sequence.
- **Usage in objective metrics.** The visual attention model is applied to each objective metric to obtain the improvement of weighting the image pixels, depending on their belonging to a determined region of interest.

Fig. 11. Methodology followed for video quality assessment with visual attention models development

3.3 Databases used for tests

Aligned with the objectives, one of the first steps is the selection of contents used to test the metrics. This selection must be suitable to represent all the range of contents that a conventional television channel and streaming supplier normally broadcasts, including images of documentaries, news or talking head, cartoons, sports or music content. An example of each individual some sequences used is shown in Fig. 12.

Fig. 12. Example of sequences used: “News Report”, “Rain Fruits”, “Fountain” and “ParkJoy”

For this doctoral thesis, a collection of video sequences was used in different phases. Stereoscopic sequences were necessary for analyzing visual discomfort. Sequences with a variety of features were necessary for testing saliency model, including faces, motion and different levels of spatial detail. For this purpose, the Urvoy *et al* database [Urvoy et al. 2012] was used. This database fulfills the vast majority of these concrete requirements. These sequences are labeled as NAMA3DS1-COSPAD1.

Additionally, for 3D analysis, synthetic sequences were specifically generated using professional software for modeling. This kind of sequences is important because it allows the control of characteristics, such as distance to the camera or average parallax, i.e. the amount of stereoscopy.

Finally, the SVT database [HAGLUND-1, 2006] was selected because of their variety and combinations of temporal information (TI) and spatial information (SI) (more information in Section 3.3.2), and the resolutions availability from standard definition to 4K was decisive when selected this collection.

The spatio-temporal diagram for the sequences to assure the spectrum of complexities is collected next, as explained in Recommendation P.910 for video quality assessment [REC P.910].

Table 3 shows the sequences used for this research, including the name of the sequence, format, temporal and spatial information, length in frames, 3D stereoscopy and the inclusion or not of human faces.

Table 3. Description of most representative sequences used

SEQUENCE	IMAGE	FORMAT	Fram	SI	TI	Faces	3D	DATABASE
Barrier		1920x1080p25	400	69	6	No	Yes	NAMA3DS1-COSPAD1
Basket		1920x1080p25	400	68	14	Yes	Yes	NAMA3DS1-COSPAD1
Boxers		1920x1080p25	400	52	18	Yes	Yes	NAMA3DS1-COSPAD1
Crowd		1920x1080p25	250	100	29	Yes	No	SVT

Fountain		1920x1080p25	550	64	17	No	Yes	EBU
Hall		1920x1080p25	400	117	4	No	Yes	NAMA3DS1 -COSPAD1
IntoTree		1920x1080p25	250	47	12	No	No	SVT
ITACA 3D		1920x1080p25	250	27	15	No	Yes	SYNTHETIC
Laboratory		1920x1080p25	400	59	11	Yes	Yes	NAMA3DS1 -COSPAD1
Lupo		1920x1080p25	300	39	7	Yes	No	EBU
News Report		1920x1080p25	400	63	3	Yes	Yes	NAMA3DS1 -COSPAD1
OldTown		1920x1080p25	250	52	15	No	No	SVT
Palco HD		1920x1080p25	250	24	10	No	Yes	SYNTHETIC
ParkJoy		1920x1080p25	250	111	31	Yes	No	SVT
Phone Call		1920x1080p25	400	32	12	Yes	Yes	NAMA3DS1 -COSPAD1
Rain Fruits		1920x1080p25	300	71	19	No	No	EBU
Soccer		1920x1080p25	400	17	108	Yes	Yes	NAMA3DS1 -COSPAD1
Tree Branches		1920x1080p25	400	16	117	No	Yes	NAMA3DS1 -COSPAD1
Umbrella		1920x1080p25	325	87	16	Yes	Yes	NAMA3DS1 -COSPAD1

3.3.1 Settings of video sequences used

Depending on the characteristics and settings that define the video sequence, some features define the format. Among these features, resolution, frames per second or ratio aspects are included.

- **Resolution.** Resolution is the number of pixels of each individual frame of the video sequence in width and height. Depending on the dimensions of the video matrix, different formats can be defined and may be presented on a conventional monitor or television: Standard Definition SD, High Definition HD or Ultra-High Definition UHDTV. For SD: 720x576 for systems based on 625 lines/50 images/second, 720x484 for systems based on 525 lines/59,94 images/second, both according to Rec UIT-R BT-601 [REC BT.601]). For HD: 1280x720, 1920x1080, 2K for television (2048x1080), according to [REC BT.709]. For UHDTV1: also known 4K (3840 x 2160) and for UHDTV2: also known 8K (7680x4320), according to [REC BT.2020]. In informatics, monitors have different nomenclature and value. In Fig. 13, a scheme of basic television resolutions is collected.

Fig. 13. Video sequences resolution

- **Interlaced.** Two different formats are defined in this category, which is a consequence of old analog televisions and tends to progressive formats.
 - Progressive, when video sequences contain odd and even lines in each single frame, acquired in the same time instant.
 - Interlaced, when each single frame is composed of two fields acquired in different time instants, each field containing only the even or odd lines (leading to the odd field and the even field). The information of this frame is completed in next field, which contains the even or odd alternative lines to represent the whole frame.

- **Frames per second.** The number of frames per second defines the temporal information obtained from a video sequence. Human eye processes an average of 34 frames per second to have a complete sensation of continuous motion when observing video contents, between 25 and 40 depending on the person, which are the values, accepted for NTSC (30Hz) and PAL (25Hz) frequencies.
- **Aspect ratio.** The aspect ratio of an image describes the proportional relationship between its height and its width. The most common aspect ratios in television formats are 4:3 (1.33:1), the universal video format derived from the standard definition televisions, and 16:9 (1.77:1) universal for high-definition television and European digital television.
- **Pixels aspect ratio (PAR).** Pixel aspect ratio is a mathematical proportion that describes the shape of the pixels in a digital image, i.e. its width compared to its height. Most digital imaging, especially with high-definition formats, display an image as a grid of square pixels, i.e. PAR: 1.0. But there are other values: 1.33 for anamorphic formats or 1.067 for DV video capture.

For this work, sequences used correspond to format 1080p25 (which corresponds to 1920x1080 pixels of resolution, 25fps in progressive presentation) with these settings, collected in Table 4.

Table 4. Basic characteristics and settings of video sequences used

Settings	Value
Resolution	1920 x 1080
Frames per second	25 fps
Interlaced Mode	Progressive
Aspect Ratio	16:9
Pixel Aspect Ratio	1.0 (square pixels)

The coding scheme used for video sequences is H.264/AVC with profile 5.1, more information in [RICHARDSON-1, 2004], and bitrates between 75Mbps and 200Kbps, both with and without deblocking filter. Additionally, sequences encoded in MPEG-2 with bitrates between 100Mbps and 1Mbps were used in the database. These bitrates were, on first instance, selected for obtaining a wide range of qualities, which could be next concreted, discarding the highest and lowest amounts of information. For MPEG-2 video coding, High profile and High level were selected for obtaining best quality as possible.

3.3.2 Spatio-temporal information for video sequences

The human eye tends to detect the artifacts in a different way, depending of the motion of the frames, and the level of detail contained on a single frame. That is why the basis of the algorithm is fully related to the spatial and temporal complexity. The calculation of spatial complexity is based on measurement of variance for macroblocks. Moreover, the calculation

of the temporal complexity is based on measurement of the temporal variation of the image, in formal difference between frames, and longitude of the motion vectors.

When selecting test sequences, it is important to consider the relative spatial and temporal information (SI and TI). Generally, the compression complexity is related to this two types of information on a sequence. In subjective tests, for example, it is necessary to select only a small number of test sequences to avoid prolonging tests in time, so consequently, choosing sequences that span the maximum portion of the spatio-temporal dimension is very important to demonstrate the validity of the study. Choosing sequences with different combinations of SI and TI values would be desirable for completing correct tests.

3.3.2.1 Temporal Information (TI)

“TI” is a measure that generally indicates the amount of temporal changes of a video sequence. It is usually higher for high motion sequences. It is not meant to be a measure of entropy nor associated with the information defined in communication theory. The values of this parameter vary between 0 and 255 for 8 bits per pixel (bpp) sequences [REC P.910].

The temporal information parameter is based upon the depth difference function $D_n(i, j)$, which is the difference between the values of luminance of the pixels $F_n(i, j)$ when calculating SAD function at the same location in space at successive frames. Being $F_n(i, j)$ the value of luminance in each pixel from frame n, depth function is defined as:

$$D_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

The measure of temporal parallax information (TI) parameter is computed as the maximum over time ($\max_{time}$) of the standard deviation over space std_{space} of $D_n(i, j)$ over each i and j :

$$TI = \max_{time} \{std_{space}[D_n(i, j)]\}$$

3.3.2.2 Spatial information (SI)

“SI” is a measure that generally indicates the amount of spatial detail of a picture. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor associated with the information defined in communication theory. The values of this parameter vary between 0 and 255 for 8 bpp sequences [REC P.910].

The spatial parallax information is based on the depth variation in every single frame of the sequence. For each video frame (luminance plane) at time n, Sobel filter (F_n) in vertical and horizontal component is used to evaluate the spatial complexity and entropy of the video sequence obtaining the function $F_n(i, j)$ over luminance component for every single pixel. The use of the Sobel filter is defined in Recommendation ITU P.910.

The standard deviation over space (std_{space}) in parallax diagram of each frame is then computed, after all frames in the sequence are computed. The maximum value for a frame in

the complete time sequence is defined to represent the parallax spatial information content of the scene. In equation form, the SPI parameters can be defined as follows.

$$SI = \max_{time} \{std_{space}[P(F_n)]\}$$

3.3.2.3 Diagrams of spatio-temporal information for database used

Spatio-temporal diagrams are useful to define correct databases that include sequences with a variety of spatial and temporal information.

As seen in Fig. 14, the position of the points in spatial coordinates defines the values of TI and SI, with the spatial and temporal information of the video sequence.

Fig. 14. Example of the four parameters defined for each sequence in the diagram

For the sequences included in the database and characterized in Table 3, the TI-SI diagram is collected in Fig. 15. The distribution in the diagram indicates the validity of the database for the work, because it justifies the different cases of video contents to be found in every television channel, high and low values of TI combined with high and low values of SI.

Fig. 15. Spatio-temporal diagram for the sequences used

3.4 Video encoding evolution: from MPEG-2 to H.265

In this section, the encoding standards are briefly explained in order to understand the sources of video quality loss, after introducing the main techniques used for compressing the video signal.

3.4.1 MPEG-1/2 standards

MPEG-1, the first MPEG standard, was first developed for video storage and playback on Compact Disks including 70 minutes video duration with a transfer rate around 1.4 Mbps, competing with VHS formats.

This encoder started using block-based motion compensation, DCT and quantization, and it was already optimized for a video compression bitrate of 1.2 Mbps. Video CD format quality was not better enough than VHS tapes for commercial purposes. Nevertheless, MPEG-1 video was widely used during long time for storing compressed video files, only in progressive mode.

The evolution of MPEG-1 was the following standard MPEG-2 aimed to cover the deficiencies of previous standard and being able of support the potential market of digital broadcasting and compressed television. Based on the ideas of MPEG-1 but with efficient coding techniques of interlaced and progressive video, a more flexible syntax and a considerable improvements in coding efficiency. MPEG-2 was the first standard to introduce the concepts of Profiles and Levels (defined conformance points and performance limits) as a way of encouraging interoperability without restricting the flexibility of the standard.

On first instance MPEG-2 was a great success, being adopted for applications such as digital TV broadcasting in all its ways of transmission: cable, satellite and terrestrial, and for storage provided the video coding element of DVD-Video, which finally replaced VHS videotape. Nevertheless, researchers looked for the coding standard for next generation which provided better compression performance and more robust techniques for networked transmission [RICHARDSON-1, 2004].

Among the basic techniques developed for MPEG-2 standard: motion compensation, GOP structure and block-based DCT transform usage. Most of video codecs and standards comprise the following stages:

- **Transformation.** In order to exploit adequately psychovisual redundancies, the pictures are transformed to a different domain where frequencies ranges vary according to sensitivities of the human eye and human visual system. Two of the most common transforms, which achieve efficient results, are the discrete cosine transform (DCT), basis of the MPEG standards, and the wavelet transform, used in JPEG-2000. This procedure does not imply information loss, and it can be consider as reversible.
- **Quantization.** The numerical precision derived from the transformation step is applied to transform coefficients, reducing the number of bits in the stream in every single pixel of the frame. The degree of quantization for coefficients should affect as less as possible to the visibility of the resulting image to a human observer, avoiding annoying

distortions. High-frequency coefficients can be more coarsely quantized than low-frequency coefficients. As opposed to transformation process, quantization is the responsible for the *lossy* part of compression, and is irreversible.

- **Coding.** After quantization of data into a finite collection of values, the redundancy may be exploited in the codification process, in lossless way adding the quantized coefficients to the final bitstream. Entropy coding is often used for this process, being the most popular entropy coding schemes: Huffman coding and arithmetic coding [PLESS-1, 1998].

3.4.2 MPEG-4 Visual and H.264/AVC standards

After the development of MPEG-1 and MPEG-2, MPEG-4 Visual and Advanced Video Coding H.264/AVC are the standards selected for the coded representation of visual information. Each standard defines a coded representation system that describes visual data in a compressed form and a method of decoding the syntax to reconstruct visual information.

The Moving Picture Experts Group (MPEG) developed the standard MPEG-4 Visual (Part 2 of the MPEG-4 standards). The H.264 standardization effort was initiated by the Video Coding Experts Group (VCEG), working group of the International Telecommunication Union (ITU-T) that operates as an evolution of previous MPEG standards and has been responsible for a series of visual telecommunication standards. The final stages of developing the H.264 standard have been carried out by the Joint Video Team, a collaborative effort of both VCEG and MPEG, making it possible to publish the final standard under the joint auspices of ISO/IEC (as MPEG-4 Part 10) and ITU-T (as Recommendation H.264) in 2003 [REC H.264].

MPEG-4 Visual and H.264 are concerned with compression of visual data, but MPEG-4 Visual emphasizes flexibility on efficiency and reliability. MPEG-4 Visual provides a highly flexible variety of coding techniques, making it possible to deal with a wide range of types of visual data, video objects, still images and hybrids of natural and synthetic video information. MPEG-4 Visual is organized into profiles, depending on the toolset required for each application. Classes of profiles include simple profiles for codification of rectangular video frames, object-based profiles, still images and texture profiles, scalable profiles for encoding at different resolutions or quality levels and studio profiles high-quality professional applications and mastering.

On the other hand, and in contrast with MPEG-4 Visual, H.264 is specifically concentrated on obtaining the most efficient compression for video streams, obtaining significantly better compression than any previous standard. Moreover, transmission efficiency over a range of channels and networks applications of video compression. Three profiles are currently supported for different video communication applications. The *Baseline* profile may be particularly useful for videoconferencing applications, while the *Extended* profile adds extra tools useful for video streaming across networks. Finally, the *Main* profile includes tools that may be suitable for consumer applications, video broadcast and storage [RICHARDSON-1, 2004].

Among the advanced techniques developed for H.264 standard to obtain efficient codification results, in addition to MPEG-2 basic techniques: CABAC (Context-adaptive binary

arithmetic coding) and CAVLC (Context-Adaptive Variable Length Coding) models, variable macroblock size and deblocking filter usage.

3.4.3 H.265 HEVC (High Efficiency Video Coding)

The video encoding standard H.265/HEVC (High Efficiency Video Coding, [X265-1]) was recently developed by ISO/IEC MPEG Moving Picture Experts Group, also responsible of its predecessors MPEG-2 and H.264/AVC, and the ITU-T Video Coding Experts Group (VCEG), through a collaboration with JCT-VC Joint Collaborative Team on Video Coding. The standard is also known as ISO/IEC 23008-2 MPEG-H Part 2 and ITU-T H.265 [REC H.265].

HEVC offers high efficiency and quality compared to previous coding standards, and comprises data up to twice the compression of H.264/MPEG-4 AVC, i.e. a similar quality with half encoding bitrate.

The main advantage is the possibility of encoding resolution up to 8K, which is 8192x4320 pixels, four times the maximum resolution in High Definition. This is the main reason of the research in this field and the development of new techniques corresponding to cover the requirements of UHDTV (Ultra-High Definition Television), after the success of 4K televisions, with resolutions up to 4096x2160 pixels.

Among the main characteristics of video encoding with HEVC standard, it should be mentioned:

- Variable size transforms (from 4x4 to 32x32).
- Quad-tree structured prediction areas from 64x64 to 4x4 (see Fig. 16).
- Candidate-list-based motion vector prediction.
- A collection of different intra-frame prediction modes.
- Higher-accuracy filters for motion compensation.
- Optimization tools, such as deblocking, SAO (Sample Adaptive Offset) filtering or *adaptive* codification technique CABAC (Context Adaptive Binary Arithmetic Coder).

Fig. 16. Example of comparison of macroblock partitions in H.264/AVC (left, extracted from [RICHARDSON-1, 2004]) and HEVC (right, extracted from [SONNATI-1, 2015])

3.4.4 Specifications in different codecs codification

Among the main advances of H.264/AVC compared to MPEG-2 and H.264, an important point lies in the variability of macroblock size and the usage of deblocking filter; both techniques involve the reduction of one of most common artifacts, the blocking effect.

In MPEG-1 and MPEG-2, the macroblocks have fixed size of 8x8 pixels. In the case of the H.264, the size of the macroblock is dependant of the homogeneity of the area in which is located.

On the other side, the occurrence of a deblocking filter in this encoder improves the global quality at the expense of increasing the blur on an image. In H.264 encoding, a deblocking filter hides the effect of blocking in small heterogeneous areas by a process of slight low-pass filtering. On the other hand, in extended areas such as the clouds area, the artifact is still present and visible for the human eye Fig. 17.

A study for this thesis [LOPEZ-4, 2013] has been developed in order to compare the blocking effect in different environments, and consequently define the principles to detect this kind of artifact in video sequences.

The modulus of the gradient applying Sobel masks has been obtained for different sequences. In the original sequence, it is possible to see that although the area seem homogeneous in the clouding area of the sequence “Crowd”, but the pixels are variable, so the resulting image reflects the real heterogeneity of the values, as seen in Fig. 17.

Fig. 17. Detail of frame from original not-encoded Crowd sequence (left) and its corresponding gradient image (right)

When encoding with standards MPEG-2 or H.263, the grid is visible in little squares that reveal the presence of blocking artifact. (Fig. 18)

Fig. 18. Image “Crowd” (left) with its gradient representation (right), with MPEG-2 encoding

As a conclusion, in order to detect the effect of this artifact and quantify its value, it is necessary to concentrate in the big homogeneous areas, and analyze the macroblocks of biggest sizes, especially the ones of 16x16, which must be considered (Fig. 19).

Fig. 19. H.264 encoded image with deblocking filter (left) and without deblocking filter (right) in homogeneous areas

3.5 Techniques for video quality degradation and distortion inclusion

The database of sequences must be adequately distorted with known degradations, which allow the researcher to draw conclusions about the algorithms developed. Different techniques are used for this purpose, based on video characteristics, and must be properly analyzed.

- **Lossy compression.** The process of compressing a video signal with one of the different available standards to study, such as MPEG-2 or H.264, implies a consequent decrease in the perceived quality. Depending on the characteristics of the encoding process, different artefact are detected in the video sequence. The lower is the encoding bitrate, i.e. the compression is higher and the probability of obtaining worse quality is higher.

- **Motion.** Change in motion may cause distortions in perceived quality, especially when the framerate is lower than what human eye supports.
- **Filters appliance.** When applying Sobel or Gaussian filters, for example, artifacts such as blurring or blocking may be simulated. If a mosaic filter is applied with blocks of 8x8 pixels, this may simulate the result of a severe compression in MPEG-2 codification.
- **Down-sampling / Up-sampling.** The process of up-sampling the video frame to higher resolutions implies an effect similar to low-pass filters.
- **Packet losses.** Simulating packet losses with bursts of erroneous bytes is useful to detect the impact of lossy wireless transmissions in final perceived quality.
- **Artificially impaired sequences for calibrating metrics.** Different techniques may be applied for drawing conclusions about each kind of distortion or degradation. One of these techniques is explained in next section in order to isolate concrete distortions to analyze the impact of each artifact in the human eye, and to facilitate the development of accurate algorithms.

4. Visual discomfort prediction in 3D stereoscopic video

4 Visual discomfort prediction in 3D stereoscopic video

4.1 Motivation for this work

Since its origin in 1895, cinema has been three-dimensional in the sense that movies have been based on a view of the world captured by a one-lensed camera. In spite of the individual lens, a 3D picture has depth cues that produce saliency to imitate the real world. Filmmakers have made efforts to control this three-dimensional effect by adapting focal length, lighting, motion and other techniques [OKUN-1, 2010]. The evolution of technology in this field made possible the inception of stereoscopic films, whose popularity grew in the period from 1952 to 1954. In that period, Hollywood produced more than 65 stereoscopic feature films. Rapidly, spectators declined this system for reasons associated with misalignment of projectors in the cinema, and issues of visual discomfort, crosstalk derived from image separation techniques and the inclusion of excessive disparities to intensify the experience.

A new level of visuals has emerged with the recent resurgence of stereoscopic 3D. Last years, mass 3D cinema distribution was possible thanks to perfection of digital projection and a series of digital imaging tools, which gave filmmakers sufficient control for the stereo contents creation. Techniques for creating new media to engage the emotions and increasing the experience of an audience were studied and developed.

Stereoscopic three-dimensional contents and new media offer an experience that is more immersive than conventional 2D media. This next generation media have a collection of advantages. However, on the other side, viewing safety should be ensured for increasing the quality of experience, especially in terms of visual comfort.

The expansion of three-dimensional television (3DTV) on the consumer electronics market, for cinema, gaming and internet, was compared on first instance to the transition to color television from black and white. The behavior of users when visualizing 3D contents on television is different than when this kind of contents are projected on a big screen in theaters, due to the size of the image and the way itself of representing video contents. The consequence is a more effective reception of 3D contents in theaters than at home.

Because of the horizontal separation of our eyes, both eyes have their own point of view. For that reason, each eye receives a slightly different retinal image. The human visual system (HVS) provides the illusion of depth when visualizing stereoscopic 3D video contents on two-dimensional displays by the effect of projecting retinal disparities from both views, left and right. Stereopsis can be defined as the perception of depth based on the difference between these two retinal images. The distance between points in common from these images, normally called depth cues, is interpreted by the brain. As the brain is able to extract depth

information through this process from retinal disparity, left image and right image are fused simulating a volumetric image taken from the real world. [LAMBOOIJ-1, 2007]

Recommendation BT. 2021 [REC BT.2021] identifies three basic perceptual dimensions that collectively affect the quality of experience provided by a stereoscopic system:

- **Picture quality.** The first dimension refers to the technical parameters and viewing conditions that refers to the quality provided by the system as perceived by the user in the image itself. The processes of encoding and transmission affect picture quality.
- **Depth quality.** This second dimension is related to the capacity of creating a sensation of depth in the observer. Disparity information provides additional depth compared to 2D conventional video viewing.
- **Visual (dis)comfort.** Finally, this third dimension is associated with the depth quality but more concretely to the analysis of sources that produce a sensation discomfort, mostly related to disparity information.

Additionally, some researchers have considered another two dimensions related to the psychological impact of stereoscopic technologies:

- **Naturalness,** which is the perception of stereoscopic images as a truthful representation of reality.
- **Sense of presence,** which is related to the subjective experience as if the observer is located in one determined place or concrete simulated environment.

For obtaining the highest Quality of Experience according to these five dimensions, it is necessary to assure visual comfort limits. Excessive disparity or parallax causes visual discomfort because the conflict between binocular vergence and accommodation affects the sense of depth. The definition of a “comfort zone” minimizes these conflicts, which encloses the disparities in stereoscopic images for a correct perception of depth in the objects. Researchers have proposed several approaches, in order to define these limits.

In 3D stereoscopic video, sensitivity maps are focused on saliency estimation, but other factors influence on visual perception. These factors are not specific for 3D video, as happened with saliency, so they are important when analyzing sensitive maps for 2D video.

4.1.1 Human Visual System

Stereoscopic 3D video perception is based on the fact that two different video signals are captured in order to feed each of the viewer's eyes. There is a signal aimed to be received by the left eye and another one aimed to be received by the right eye. This system tries to recreate the experience of watching a real world scene, where each eye captures two different images. The difference between these two images depends on the position of the elements in the world related to the viewer's position. This means that the system is feeding the observer with a disparity depth cue.

Space perception is a process through which, humans and other organisms become aware of the relative positions of their own bodies and objects around them. Space perception provides cues, such as depth and distance, which are important for movement and orientation

to the environment. These factors influence on the eyes of the perceiving individual as a way of communicating and interacting with their surrounding universe.

Disparity is defined as the state of being different or dissimilar. In this work, the concept of disparity is related to the sensory information received by the human eye. This word, as a term, has been used to express what is known as parallax in photogrammetry by the computer vision community and shortly for retinal disparity by biological vision researchers.

Humans do not ordinarily perceive a binocular space (a separate visual world from each eye) but instead see a so-called Cyclopean space, as if the images from each eye fuse to produce a single visual field akin to that of Cyclops. Although binocular vision is based on perception of two independent views, the brain considers, as there were only one by an effect that is normally known as the cyclopean eye. A high correlation between these two views exists, which is the basis of stereoscopic perception.

The cyclopean eye is an imaginary eye situated midway between the two eyes. Using the cyclopean eye, crossed and uncrossed diplopia can be explored as the focus of space perception. Modeling the sight from this viewpoint is possible to have advantages in reducing the discomfort related to stereoscopic viewing. The cyclopean eye is used to determine the direction of point A and point B. Point A is stimulating the temporal retina of right eye. At the same time, the nasal retina of the left eye stimulates a retinal point to the right of the fovea [KOLB-1, 1995].

Fig. 20. Example of cyclopean eye functioning

The term stereopsis is also related to the concept of cyclopean vision. Stereopsis refers to the ability of the human brain for extracting depth from differences between left and right views information. Differences between points in images presented to the two eyes generate disparities. These disparities are used by the brain in order to infer relative depths between objects located at the scene [WARE-1, 1998]. Fig. 20 illustrates a simple stereoscopic display and two objects situated in points *a* and *b*. One is over the screen (*a*) and the other one is behind the screen plane (*b*). The views presented to both eyes show a difference in the relative distance between both objects. The brain fuses the information over the screen obtained in points *c* and *d* and interprets the result as a line behind the screen.

Fig. 21. Disparity interpretation over a stereoscopic screen

Disparities are the differences between the angles subtended between pairs of features ($\alpha - \beta$). There are a number of problems with screen-based stereo displays. If disparities are too large for human eye, the brain will not be able of fusing them and the result is the occurrence of an artifact called diplopia. The area in which the images can be fused is called Panum's fusional area. This area is remarkably small in the worst case and located at the surroundings of the screen plane.

Diplopia is an anomaly that occurs when objects are in our field of view but outside the binocular Panum's fusional area. As a consequence of this anomaly, the image appears as double in human brain. When viewing objects in the real world, this effect is normally not registered because the peripheral vision occurs in lower resolutions. On the other hand, in stereoscopic displays is a common effect if the size of the object is not proportional with the biological perception.

In natural view, parallax is created by disparities, what will determine the virtual perceived location of the objects in a scene. However, the experience of watching 3DTV is significantly different. Three different circumstances are possible to occur depending on the conditions presented by disparities: zero parallax, positive parallax and negative parallax, as shown in Fig. 21. When an object's left image is located at the same point as the right image, then the parallax is zero and the object is perceived as being on the screen plane, the same as if the contents presented to the observer were bidimensional. On the other side, if the left view of the object is seen on the screen at a position that lies to the left of its right image, then the parallax is positive and the object is seen as being beyond the screen. Finally, when the left image of the object is located to the right of the right image, then the parallax is negative and the object is perceived as being in front of the screen. As seen in Fig. 22, when viewing objects in negative parallax area, the eyes are obliged to cross their views to perceive depth on the screen. The difficulty of the eyes for changing the focus in this circumstance is one of the main sources of tiredness and discomfort, as will be explained with examples through the redaction of this doctoral thesis.

Fig. 22. Parallax possibilities: positive, zero or negative

In natural viewing, the point of view is free to change with the eye movement. However, in stereoscopic displays, the methodology is limited, as the point of view is prefixed by the fixed point of view of the camera lenses that have captured the scene, and so is the focus. Furthermore, in natural viewing, the eyes focus (accommodate) and converge to the same distance. However, when looking at a 3D object displayed on a screen, a viewer's eyes must focus on the screen while, at the same time, they converge on a point in space that may be located beyond the screen, on the screen, or in front of the screen. This is known as the *vergence-accommodation conflict*. This conflict limits the amount of parallax that a viewer can tolerate avoiding discomfort, also known as the Zone of Comfort [SHIBATA-1, 2011] (it will be studied in more detail in the next section).

The binocular vergence and the accommodation processes, produced by focusing over the crystalline lens, must be controlled for offering output stimuli that simulates a stable cyclopean 3-D experience. The vergence process can be defined as the movement of both eyes in different direction to locate the area of interest on the fovea; this first process is primarily driven by disparity. On the other hand, accommodation process is defined as the alteration of the crystalline lens to focus the area of interest on the fovea, this second process is primarily driven by the effect of blur. As each eye examines a real world scene, the brain through sent neural commands to the lens muscles to change focal point and minimize blur for the fixated part of the scene, but blur artifact is easy to appear in the final observation when viewing a 3D conventional display. These two processes constitute independent subsystems, which are synchronized when observing the natural world but, when producing an artificial illusion of depth, synchronicity is not assured. Thus conflicts arising from accommodation and vergence mismatches (AVM) may occur.

The ability of human eye to focus objects located at different distances affects to the perceived quality. Abrupt changes cause human eye needs to change the focus position at different distances rapidly. Some factors should be taking into account for assuring the quality of experience in 3DTV, such as disparity distribution, depth, perceptual and cognitive inconsistency, quick changes, excessive disparities, mismatch between left and right views and vergence conflicts. [JUNG-2, 2013]

HVS is sensitive to orderly regions for visual perception based on several factors, such as motion, regular structures and highlighted areas. HVS focuses on salient regions on an image, which is related to sensitive areas. To achieve this objective and obtain accurate sensitivity maps, pattern masking is usually used.

Studies using eye-tracking methods to analyze eye movement traces are useful to understand the human eye behavior when observing a static image [WOODING-1, 2002]. These traces are helpful as the basis of the attention models, considering after understanding large numbers of eye movement traces examined, adding motion and negative parallax disparities in the case of stereoscopic video.

4.1.2 Visual Discomfort

If the 3D process is not properly used, there could be consequences that can actually cause physical pain. Visual discomfort is an ambiguous concept that is defined as a diversity of subjective symptoms. These symptoms are related to the consequences for HVS after a wrong generation of the processes for viewing, in this case, a 3D stereoscopic illusion. Some issues must be respected before 3D can be implemented for experienced aesthetic sensations.

The concept of visual discomfort is also known as visual fatigue in terms of the tiredness produced when observing stereoscopic video. A prolonged observation of this type of video could conclude with a general headache or a certain kind of pain in shoulders and neck. Visual fatigue can be induced by different circumstances, including anomalies of vision related to convergence and accommodation dysfunctions. Visual fatigue is also related to display technology or image/video quality, which is dependent with the viewing distance and other contour conditions. A study developed by Berezin [BEREZIN-1, 2010] in Russia concluded that 30% of the viewers participating in this research reported eye tiredness after watching stereoscopic 3D movies in cinemas. Among the most common effects, these viewers reported altered vision, confusion, nausea, lightheadedness, and even convulsions. For that reason, some researchers do not recommended S3D video for people with special physical conditions, such as pregnancy or under the effects of alcohol.

Visual fatigue is detected and diagnosed by the presence of signs and symptoms. Signs are considered as objective indications related to physiological reaction by the spectators and symptoms are subjective indications related to their physical state and perceived mental health [URVOY-2, 2013]. In most cases, visual fatigue is caused by an excessive visual effort. This is not serious for human body, because visual fatigue disappears after a rest period, but this effect should be considered by filmmakers and content generators to intensify the experience without putting in danger or risking the security of spectators' vision.

For the assessment of visual fatigue, signs and symptoms should be analyzed in terms of nature, intensity and temporal characteristics. In addition, *asthenopia* and *eye-strain* should be considered as health issues derived from pain concentrated around the eyes area and general headache.

In this section, the most probable factors that affect to visual discomfort when viewing 3D video in stereoscopic displays are analyzed.

- **Excessive binocular parallax**

As demonstrated in subjective disparity transition models analyzed after developing subjective assessment [LOPEZ-1, 2013], one important source of visual discomfort is the transition between images with different levels of disparity. This factor is especially accused when viewing objects with negative parallax. If these objects are immersed in an environment of motion and changes this effect is even more intensified, because it obliges the human eye to change the depth of focus continuously.

Excessive parallax is commonly perceived as one of the main sources of discomfort, and generally negative parallax brings more discomfort than the positive one, because is where disparity skewness features take effect [ZHANG-1, 2015].

When viewing stereoscopic contents, excessive binocular disparity may be a source of blurring or double vision, and it is part of the accommodation and vergence process, inducing unnatural oscillations. Human brain may be affected by these phenomena with a consequent visual fatigue and activation of sensory functions and problems with eye movement functions. Research developed by Yamanoue et al [YAMANOUE-2, 2012] examine characteristics of visual fatigue and used magnetic resonances in order to investigate the effect on human brain of excessive binocular disparity.

- **Accommodation and vergence mismatches (AVM)**

The sensation of depth (and consequently the distance towards objects) in the real world is estimated by human visual system mostly from the linkage between accommodation and vergence. Stereoscopic displays in contrast lose this linkage because the environment is artificially generated. According to Kooi *et al* [KOOI-1, 2004], Vergence-accommodation conflicts are one of the main sources of visual discomfort. Stimuli in the real world present a collection of depth cues for specifying the 3D layout. On the other hand, stimuli presented on 3D displays present simulated cues for representing the specified depth. These cues, which are defined as screen cues, are probable to cause perceived distortions, derived from a wrong determination of parallax and motion, generating sometimes a series of artifacts.

The eye is also a single-lens system, with is not rigid and deform by the effect of ciliary muscles to focus the object, which is located at single plane. A sharp image is obtained in the retina for only one position, which is adjusted by the optical lens according to that exact position [BLANCHE-1, 2012]. The basis of the accommodation cue is adjusting this lens to focus objects located at different distances, which is a complicated process as seen in Fig. 23, where a far distance object requires different position on the lens in comparison to a near distance object. Accommodation is not specific of stereoscopic vision, but when viewing a 3D display the eyes accommodate to the position of the screen, which is the emitter plane. Unlike viewing in real world where distance from observer to objects is constantly changing, in stereoscopic displays this distance remains constant.

Fig. 23. Crystalline accommodation for a far or near distance object

When viewing an object in stereoscopic displays the eyes accommodate to the screen but when rotating to fix the apparent object a mismatch between them occurs. This effect is the accommodation-vergence conflict Fig. 24.

Fig. 24. The accommodation-vergence conflict when viewing a stereoscopic display

The vergence changes and is correctly reproduced, but the accommodation or depth of focus is fixed at the distance of the screen (Fig. 25). There is a certain degree of mismatch tolerable for the brain, especially when, as in a theater, the screen is further from the observer. The accommodation is less variable at large distances and the eye lens changes in a slower manner. For that reason, the experience of viewing 3D content is more effective in large theaters than on televisions.

Fig. 25. Differences in vergence and accommodation process from real world to a 3D display (image taken from [HOFFMAN-1, 2008])

For determining potential effects and knowing how to reduce them or eliminate them, it would be useful to establish guidelines for filmmakers to determine range of disparities in order to avoid visual discomfort in observers [SHIBATA-1, 2011].

The Zone of Comfort (ZoC) was introduced by Percival [PERCIVAL, 1892], when, based on experiments with spectacles, he suggested the limits to vergence-accommodation postures that could be achieved without causing discomfort. This discomfort rises from the unnatural way in which the eyes have to act when the focus does not match an object's spatial position. Some examples like using lenses or watching stereoscopic images increase the possibilities of this conflict to happen.

Recent studies developed by Shibata et al [SHIBATA-1, 2011] conclude that the ZoC may differ from Percival's ZoC when the experiments are based on stereoscopic vision rather than based on vision through spectacles. The vergence-accommodation relation in stereo vision is in permanent change, as opposed to a lens or spectacles system where it is a fixed value [BANKS-1, 2012]. Shibata's ZoC for different accommodation distances in stereo vision, both in diopters, the optic measurement, and meters is presented in Fig. 26. Vergence-Accommodation distance in meters helps to understand what happens in a real deterministic situation where the viewer is positioned at a precise distance from the screen where the stereo image is produced. According to this diagram, images with positive parallax have little or none capability to induce discomfort, while negative parallax is most likely to cause discomfort if this parameter is not controlled in production.

Fig. 26. Shibata's Zone of Comfort with Distance in diopters (D) and meters (m) (extracted from [SHIBATA-1, 2011])

The variation of time of exposure is necessary to be analyzed in order to determine its effects of visual discomfort, thus, relating it to the ZoC [LI-1, 2011]. In order to adapt a stereoscopic ZoC to 3D stereoscopic video, it is necessary to take into account motion and time of exposure to a stereo scene. The ZoC will be further reduced when these elements appear. The time to converge and accommodate in this case is relevant, thus there is a need to adapt the concept of ZoC.

A subjective evaluation of visual discomfort is necessary for this purpose as developed in study by Cho et al [CHO-1, 2012], concluding that parallax limits and regions of comfort must be assured to the observer, according to viewing conditions: screen size, disparity and viewing time.

It is assumed by most researchers that symptoms are caused by differences between original stimuli and conflicts between the vergence and accommodation process. Most viewers report fatigue and visual discomfort when viewing stereoscopic displays than when viewing conventional displays.

Some researchers have used the definition of the AC/A ratio (Accommodation Convergence to Accommodation ratio) for determining a correct depth perception based on binocular fusion, which is different for each individual because this ratio is related to interpupillary distance [BAHN-1, 2002]. This ratio measures the change of convergence due to accommodation per change in accommodation, for example, the magnitude of the crosslink-interaction. People with extremely high AC/A ratios have trouble with depth perception.

As mentioned before, the accommodation and vergence response is the same when viewing an object in real world (Fig. 27). Schor [SCHOR-1, 1992] studied this difference. The response to a continuous stimulus is the time that the stimulus takes to stabilize after a short period (between a half and a second depending on the distance to the stimulus). The vergence distance and the focal distance are the same in this case. At steady state, input and output amplitudes are equivalent and the HVS will comfortably process the input stimulus.

Fig. 27. Response to a stimulus in accommodation and vergence in real world

On the other side, when the step stimulus is constant and the object is represented beyond the screen, the focal distance (accommodation distance) to the screen is different to the vergence distance to the simulated projection to the object (Fig. 28). The processes of vergence and accommodation occur in a different way. Accommodation remains at the 3D screen depth, but the vergence response to a virtual depth beyond the display with a determined disparity, which provoke the stereopsis.

Fig. 28. Response to a stimulus in accommodation and vergence in a 3D Display

- **Panum's fusional area and DOF revisited**

Panum's fusional area corresponds to a region of the retina, and is where binocular vision really occurs. This space narrows at the fixation point (F) and expands in the periphery both sides of the horopter in consistence with cell receptors in the ganglion.

Stereopsis is produced by binocular disparity only in cases when retinal disparity is not excessive. Retinal images are not fused by the brain if binocular disparity does not exceed a certain limit, and then the resultant perception of depth through stereopsis occurs. On the other hand, if retinal disparity is excessive, binocular fusion is not appropriately produced and images on retinal areas fall in different relative positions with the consequence of diplopia or double images.

The boundaries of this area are at both sides of the horopter. The size of Panum's area is approximately 10 arc minutes in adults. In children, the area is proportional to peak contrast sensitivity and the size would be approximately 200 arc minutes [ASLIN-1, 1993]. Nevertheless, the size of this area depends on different visual parameters such as size and duration of the stimuli. The horopter could be defined as the exact point where all the points of the binocular field are seen individually.

Fig. 29. Area equivalent to Panum's fusional area (part of retina) where stereopsis occurs. "F" is the fixation point.

Because of the different viewpoint derived from eye's horizontal separation, each eye receives a different retinal image, which is, as mentioned before, the basis of depth perception. This effect happens because the brain extracts depth information from cues in left and right view. The classical notation of Panum's fusional area is considered an absolute limit from disparities that can be fused by the brain without diplopia or any similar defect. The limits of Panum's fusional area are affected by a collection of factors, which include spatial frequency, stimuli size and duration, and motion. The limits of disparity change depending on the size of the objects and disparity distribution. Both factors affect to visual comfort in terms of relation between range of disparity and the amount of pixels, i.e. the size of that disparity [SU-1, 2015].

After the comparison between natural viewing conditions and stereoscopic display viewing conditions, objects further from the focusing point are perceived more blurred. The brain represent the objects at increasing distance from the focus to decrease the effects of diplopia while this objects appear out of the limits of Panum's area, decreasing the spatial information in high frequencies.

On the other hand, when viewing objects on a stereoscopic display, the objects are located around the focusing point, which is over the screen itself. Light emitted by effect of the display separate virtual object from display increases: Consequently, the light is diffused and HVS perceives the object as more defocused (Fig. 30, left). Viewing objects in the real world (left image) generates blur gradient in the retinal image, depending on the changes in focal distance derived from scene depth. Otherwise, when viewing scenes in 3D displays, the focal distance is constant (Fig. 30, right). For that reason, no blurring appears in the retinal image, specifying flatness in the image.

Fig. 30. Blur in real world versus image in 3D Display

- **3D Artifacts**

The codification of 3D video sources is a generation of common artifacts among 2D video encoding process, such as *blocking*, *mosaic patterns*, *staircase effect*, *ringing*, *color bleeding*, *noise*, as well as other artefacts specific of 3D video. [BOEV-1, 2008]

This section aims to do an overview, description and categorization of main stereoscopic artefacts that occur during the process of delivery of stereoscopic video. In order to guarantee a correct transmission of large amounts of 3D content for digital television content, some artifacts are necessary to be taken into account.

- Keystone distortion and depth plane curvature

This artifact occurs when increasing the distance between the object and the camera. The consequence is a convergence decrease and then, the viewer will perceive the objects in the center of the screen as if they were located closer than objects in the corners. When this visual artifact appears derived from vertical plane distortions, this effect is called keystone distortion. On the other hand, when this effect occurs for the horizontal plane, a wrong representation of the reality is perceived consequently to the gradual variation of depth and a magnification of the objects. An example appears in next figure (Fig. 31).

Fig. 31. Keystone distortion (left) and depth plane curvature (right) caused by (weakly) convergent camera (taken from [BOEV-1, 2008])

○ Cardboard effect

The cardboard effect is an artifact that affects the perception of depth. The viewer perceives the distance and size of the object with a certain mismatch between them. The scene is perceived as cut into discrete layers of depth planes, as if they had no volume at all. The object appears unnatural because the information is limited. An example of depth maps is shown in Fig. 32.

Fig. 32. Cardboard effect (right) in comparison to its original depth map (left) (extracted from [BOEV-1, 2008])

○ Ghosting by disocclusion

The disocclusion artifacts appear when depth is not represented in a proper way. A mismatch in the combination between the image and depth information, occurs when looking from different angles since there is only one plane for the depth available. For

3D images and video content, this effect provokes a false contour around the objects, similar to ringing in the 2D objects.

- Crosstalk

Crosstalk is the result from an imperfect separation between left and right views. This artifact is similar to ghosting, but crosstalk is a consequence of electrical or optical mixing. Researchers as Siegel [SIEGEL-1, 2001] consider crosstalk as one of the main display-related perceptual factors degrading image quality and visual comfort. However, crosstalk may also have some beneficial effects on image quality and visual comfort, but this artifact must be reduced. The optimal amount of crosstalk is still an issue of debate; the amount of induced depth should be a balance between annoying degrees of blur, perceived banding and clear transitions between views.

As it happened with two-dimension conventional video, different artifacts can be identified over the dataflow stages which affects the 3D vision, considering the adaptation to the different visual subsystems of human eye, including structural, color, motion or binocular vision itself.

- Picket-Fence Effect and Image Flipping

These effects are typical when observing contents on an autostereoscopic display, especially the ones that work with parallax barrier techniques for spatial multiplexing in order to show the each view for each eye of the observer.

Both artifacts are perceived if the observers make a lateral head movement in front of the display. Then, as the gaps between the pixels are predominantly visible for some observation angles with a consequent of experienced unnatural disparities compared to the continuous parallax observed in the real world [MEESTERS-1, 2004].

- Shear distortions

Shear distortion commonly appears when viewing stereoscopic displays that allow only one correct point of view from the observer's perspective [MEESTERS-1, 2004]. If the spectator changes his viewing position, objects in the image seem to follow the user and the perspective is represented as distorted. Tracking users positions according to screen distance could be a solution for this type of artifact, but this alternative is not easy to implement in most of the cases.

- Window Violations

In the process of producing S3D video and stereoscopic films is important considering the manner in which the 3D effects achieve the disparity of the scene and the appropriate parameters of stereo cameras. Although camera are correctly adjusted, special care must be taken when filming, especially in screen borders. Conflict between screen borders and the zero-parallax plane may cause conflicts. The

effect of objects located close to the screen borders with part of them outside the limits of the frame may cause visual discomfort derived from an effect called stereo window violation (SWV) [TSENG-1, 2012].

An example of this effect is shown in Fig. 33 and Fig. 34. As part of the object is out of the limits of the screen, the observer finds conflicts for brain fusion of both views, left and right. This effect may cause discomfort in HVS and should be avoided when shooting S3D contents or making S3D CG movies to reduce the discomfort effect or correct camera parameters automatically. Window violation can occur in laterals: left side or right side of the picture, and up or down of the picture.

Fig. 33. Sequence "Itaca 3D" with spinning logo and moving ship with Window Violation (W.V) on the left side of image

Fig. 34. Sequence "Itaca 3D" with Window Violation (W.V) on the right side of image

4.2 State of the art for video quality assessment related to 3D stereoscopic video

As mentioned in previous chapters, Qualinet White Paper [LE CALLET-1, 2012] defines Quality of Experience (QoE) as the user's degree of delight or annoyance derived from an application or service. For that reason, it is necessary to know that video quality is influenced by different factors, such as content, network, device, application, user expectations and context, and further more when 3D video is considered. This work is specifically focused on the content generation.

The work in Quality of Experience for stereoscopic contents could be referred to different research fields: analysis of disparity, distribution of parallax, motion, encoding process, artifacts detection, packet loss in transmission, quality of the screen in fusion of views, detection of window violations or mismatches between left and right views. This thesis is centered in the content generation practices, i.e. the first step in the video production chain, for native stereoscopic video in both natural and synthetic capture.

As mentioned before, the success of 3D cinema and the high amount of sales in stereoscopic televisions nowadays have increased the necessity of developing new video quality studies for 3DTV. For that reason, extensive literature has been found in this research field. The objectives derived from these studies need to obtain and describe guidelines about the use of technical parameters to improve the experience by end users. This research is helping stereoscopic filmmakers, video editors and stereoscopic supervisors to generate appropriate contents through the sensation of perceived depth. These types of contents must avoid the usual sources of visual discomfort for human eye, such as excessive negative parallax, abrupt changes or fast motion. Among this extensive list of topics, one aspect selected for being analyzed in QoE is how 3D stereoscopic video affects to HVS, which is the main purpose for this doctoral thesis.

In addition to the standard recommendations for general subjective assessment, *Recommendation ITU-R BT. 2021* [REC BT. 2021] offers guidelines to develop subjective test experiments in a 3D environment. This Recommendation presents different methods and procedures for the assessment of the three primary dimensions described previously which affect to Quality Experience: picture quality, depth quality and visual comfort.

New topics emerge when dealing with stereo vision quality which differ from 2D vision, based on the optic effect of the stereo images combination, such as visual discomfort or perceptual inconsistencies between depth cues [RODRIGO-1, 2013].

In 3D media, new factors are concerned in order to assess quality, artifacts that commonly appear on 2D video are inherit, but also these new factors related to the optic effect of stereoscopy, such as visual discomfort or perceptual inconsistencies. When evaluating Visual Quality Assessment (VQA) over 3D video, these other factors must be taken into account. Extensive work is being developed in this field. Researchers have defined metrics and algorithms to predict the quality of a video sequence, also for 2D video. In 3D stereoscopic

video, research is oriented to implement tools for decreasing visual discomfort. For that reason, much work has also been developed in this field relating depth and motion.

Much work related to 3D quality assessment is based on a simplistic approach of extending 2D quality metrics to 3D by combining evaluation done separately on left and right views [WINKLER-2, 2013]. This asymmetric assessment is interesting for a first approach, but the interaction between both views is crucial to obtain accuracy. Other recent methods for 3D quality assessment take into account additional parameters. Lambooij *et al* [LAMBOOIJ-2, 2011] propose a model that is expressed as a weighted sum of image quality and perceived depth, combining discomfort produced by stereoscopic depth with image artifacts such as blur or noise produced by the content generation and transmission processes. Studies developed by Hewage *et al* [HEWAGE-1, 2013] are based on artifacts introduced by 3D video streaming affecting to the reconstructed 3D video at the receiver end, leading to inferior quality and user experience

3DTV artifacts have been widely analyzed. There are different approaches to describe and to quantitatively measure the effect of cardboard effect and puppet theatre, such as the work by Yamanoue *et al* [YAMANOUE-1, 2006], however this research is rather preliminary. An extensive compilation of artifacts appearing during the delivery of 3D contents in mobile devices was collected by Boev *et al* [BOEV-1, 2008] and Meesters *et al* [MEESTERS-1, 2004] offered an interesting survey of perceptual evaluations and requirements of three-dimensional television.

Other artifacts such as stereo window violation (SWV) and temporal continuity of the disparity (TCD) have been widely studied in research by Tseng *et al* [TSENG-1, 2012]. Guidelines to create comfortable and faster stereoscopic films are offered in this work, which includes assessment of visual discomfort based on depth maps and 3D cinematography principles. In addition, studies by Chen *et al* described the performance of window violation when predicting visual discomfort [CHEN-1, 2011].

Extensive literature confirms that visual comfort is one of the most important indexes in the measurement of image quality and the viewing experience [REN-1, 2015]. For that reason, numerous studies have been developed in this research field. Eye-tracking analysis, which compared the response of viewing S3D contents compared to watching 2D content, showed that time and video content have an influence on video fatigue accumulation and visual functions [IATSUN-1, 2015].

Tam *et al* [TAM-1, 2011] developed an overview that describes the main topics relevant to comfort when viewing stereoscopic television. This research tried to analyze with subjective tests focused on the accommodation-vergence conflict, parallax distribution, binocular mismatches, depth, and cognitive inconsistencies. Speranza *et al* [SPERANZA-1, 2006] reported in their research that depth and motion are closely related to visual discomfort.

An interesting subjective evaluation of visual discomfort is developed in [CHO-2, 2012], where parallax limits and regions of comfort, dependent on the screen size, disparity and viewing time, are obtained.

Factors such as strength of the excessive disparity range, complexity of the background objects, variation of the motion-depth or contrast of the objects in the scene, are determined

as a possible source of visual fatigue, as demonstrated in research developed by So et al [SO-1, 2015]. This research follows the work by Nojiri et al [NOJIRI-2, 2004] that demonstrated that visual discomfort is caused by the difficulty of fusing left and right images because of excessive binocular parallax and its temporal changes. This study concluded how visual comfort is affected by the statistics of disparity distribution and the consequent temporal parallax changes occurred by motion.

Most visual discomfort models are based on analyzing statistics of the computed depth maps [YANO-1, 2002], [NOJIRI-3, 2003], which is assumed as one of the most important sources of visual discomfort. The model proposed by Choi *et al* [CHOI-1, 2012] predicts a visual fatigue score based on linear combination of selected factors after developing subjective assessment with subjects of proved validity. On the other hand, model proposed by Kim et al [KIM-1, 2011] detects impairments caused by camera misalignment or other inappropriate shooting parameters that induces excessive binocular disparities. Other models include physiological optic descriptors, which analyze mismatches between accommodation and vergence, such as studies by Park *et al* included in [PARK-1, 2014] and [PARK-1, 2015]. The model for predicting Quality of Experience proposed by Chen *et al* [CHEN-2, 2012] includes more viewing characteristics beyond the ones affecting to HVS. This effect confirms the versatility of studies concerning to video quality assessment.

Much work has also been developed in this field related to depth and motion, such as [JUNG-2, 2013]. Filtering is also a technique used to reduce visual discomfort on screens, as demonstrated in research by Bosc et al [BOSC-1, 2011] and Hanhart et al [HANHART-1, 2012].

Visual saliency computation influenced by depth perception and visual comfort is the basis of other research works. The model proposed for assessing visual discomfort by Jiang et al [JIANG-2, 2015] is based on saliency. This saliency model is composed of three elements: 2D image saliency, depth saliency and visual comfort based saliency, taking into account factors as color, texture and spatial compactness, which improve the results [JIANG-1, 2015]. Another model proposed by [IATSUN-1, 2015] is based on saliency video characteristics: motion activity, disparity range and changes for predicting visual fatigue.

Jung et al [JUNG-1, 2012] offered a preliminary visual comfort model for detecting salient object's motion features in depth of field. These same researchers in following studies [JUNG-3, 2015] more accurately analyzed saliency and its relationship with the experience of the observer. These recent studies were based in the analysis of saliency maps for describing the information of visual attention in 3D stereoscopic video. Saliency described through these saliency maps was compared to the concept of sensitivity, which is common to both 2D and 3D video. Thus, a consequent pattern masking emerged from these sensitivity maps, confirming that depth must be taken into account as an important factor in visual comfort models for HVS.

Other researchers have based their studies in analyzing saliency. The multi-stimuli stereoscopic images database MSID, which was developed for research in this field by Zhang *et al* [ZHANG-1, 2015], analyzed human eye's behavior derived from observing images with multiple salient objects. The number of salient objects, and their corresponding shapes, sizes and distribution are subject of analysis to evaluate cases with potential correlations with visual discomfort. Four types of salient objects were identified in this work considering disparity

jump features, parallax distribution and the distribution of salient objects itself. These three factors improve the prediction performance of visual discomfort.

Lambooij et al [LAMBOOIJ-3, 2013] developed an extensive experiment for analyzing subtitles included in 3D video contents. Some viewers experience visual discomfort that can be induced by certain video characteristics related to subtitles, such as fast motion, large changes in disparity and depth discontinuities. This study concluded that subtitles affect negatively the experience on how viewers watch stereoscopic films in terms of visual comfort. Subtitles position is a common case of objects located at a constant plane. Filmmakers decide the depth of this constant plane, which could be located at the zero plane, i.e. over the screen plane, or if they must be located at depths closer to the viewer to avoid collision conflicts.

4.3 Preliminary subjective studies for defining objectives

Preliminary tests were run over a set of 3D video sequences to understand and analyze different features that generate visual discomfort or quality reduction. These tests were necessary to determine the following studies and focus on the main sources of visual discomfort.

The observers participating in the research were asked to rank the sequences taking into account their 3D quality. Results were compared to the objective data obtained through the usage of tools specifically developed for this purpose. The objective is deciding which features would be a possible source of visual discomfort and how to modify these features to obtain satisfactory 3D experiences. All tests were carried out on a 46" 3D LCD monitor model JVC GD-463D10 with passive glasses.

A group of 16 subjects took part in the tests evaluating their stereoscopic quality in MOS scale, taking into account that Winkler research considers enough 15 subjects for subjective assessment [WINKLER-2, 2009]. The test consisted in observing each sequence and giving them a score in MOS scale from 5 to 1. The MOS scale considers the meaning of each note as follows: 5 (perfect), 4 (fair), 3 (annoying), 2 (very annoying) and 1 (impossible to communicate). The experiment followed the ACR (Absolute category rating) explained in Rec. P.910 [REC P.910].

The session had duration of approximately 20 minutes and each user did the tests individually. Before the tests, some sequences were used to prove the observers normal stereopsis, as recommended in Recommendation BT. 2021 [REC BT.2021].

Sequences used for the assessment included a 3 minutes sequence called "Modernism" (Fig.35), created by Mediapro company, in which different scenes appeared with different levels of motion and depth. Sequences "Rain Fruits" and "Fountain" from EBU were also used. For more information, the reader is recommended to consult the material section.

Fig.35 Examples of left frames in sequence "Modernism"

Additionally, synthetic test sequences "Palco HD" (Fig.37) and "Itaca 3D" (Fig. 33) were used, which included parallax and object distance variation. These sequences were created specifically for these tests using the software tool *Autodesk Maya* to develop an environment with virtual pairs of cameras (Fig.36). The purpose is enabling an easy methodology for modifying depth in the objects, generating high negative parallax images by changing the interpupillary distance simulated by the cameras.

Fig.36. Examples of camera separation to create parallax variation

Fig.37. Effect with anaglyph examples of parallax variation in sequence "Palco HD"

All these sequences are high definition resolution (1920x1080) with no compression, available in side-by-side formats, which is half-effective HD resolution. Sequences generated by EBU are also available in side-by-side with complete HD resolution for both left and right views, which is interesting to test the necessity of higher resolution.

The sequences used were distributed among six different groups, each one related to one experiment to analyze. The user gave a score to every sequence (or pair of sequences), and had free space to write a general comment on each one, in order to draw conclusions about their opinions. These six groups of sequences for the test are the following:

- Pair of sequences with transitions from different types of parallax (Fig.38), negative and positive, to detect the impact over abrupt stereoscopic changes. A "Positive parallax" sequence (P.P) is considered when it has not remarkable negative parallax and pixels with positive parallax represent more than 25% of an image. On the other

side, “Negative Parallax” sequences (N.P.) are sequences whose images possess more than 15% of pixels in negative parallax (assuming an environment of positive parallax). See Fig. 39 for results.

Fig.38. Example of abrupt transition from "Positive Parallax" image (P.P) to "Negative Parallax" image (N.P) with their distribution of parallax histogram

- Same static images and video sequences with different values of parallax, from zero and positive parallax, to negative out of ZoC values.
- Negative parallax sequence with different levels of motion: low, medium and high motion. Test statistics are collected in Fig. 40.
- Sequences with window violation (W.V) produced in different sides of the image, in lateral or top/bottom regions of the image. Examples are shown in Fig. 33 and Fig. 34. See subjective results in Fig. 41.
- Negative parallax variations to compare zones inside and outside of Shibata's ZoC.
- Long sequence with soft variation of parallax, at the end the sequence starts from the beginning producing an abrupt parallax change. Results from this experiment appear in Fig. 42.

Results of these tests are collected in next figures Fig. 39-Fig. 42.

Fig. 39. Results in Transitions between types of parallax (P.P: Positive Parallax, and N.P: Negative Parallax)

Fig. 40. Results in impact related to motion in stereoscopic sequences (low, medium and fast motion)

Fig. 41. Results from sequences without Window Violation (W.V) or W.V. in lateral sides or down area

Fig. 42. Results derived from a sequence with parallax progressive or abrupt variations

Analyzing the graphs presented and observers' opinions obtained in the study, it is possible to conclude the following statements related to the preliminary subjective assessment.

- Derived from variation of parallax in static images:
 1. Negative parallax with objects situated in the edges of the picture frame, generates disparities in visualization of objects closer to the observer, by the effect of a window violation.
 2. Sequences with objects out of the Zone of Comfort (ZoC), generate visual discomfort and fatigue in the observers eye.
 3. The variance in the parallax histogram, especially in negative parallax areas but also in positive areas, has a high influence in 3D experience and, as a consequence, in quality in general.
- Derived from variation of parallax in motion sequences:
 1. When a scene change occurs, the abrupt variation from high-percentage positive parallax pixels to sudden emergence of negative parallax objects is very annoying for

human eye, if it has not enough time for accommodation and focus on the image. As seen on abrupt changes graph Fig. 42.

2. In fast motion video sequences, even positive or negative parallax could be a cause of visual discomfort (see Fig. 40), but it has been concluded after several experiments that negative parallax is especially annoying for human eye.
3. In cases of low motion sequences, the time of convergence and accommodation is higher, but the high variations are still considered.
4. Images with motion in objects with different values of negative parallax are difficult to focus simultaneously, as demonstrated empirically in sequence "Itaca 3D" sequence, with a logo turning in circles, while a starship is getting close to observer. Observers' commentaries about this sequence revealed their difficulty to focus both objects, because of the difference in negative parallax.

4.4 Model for predicting visual discomfort in 3D video sequences

Viewer's satisfaction depends on the level of similarity with the natural content when viewing stereo video contents. The difference between natural watching directly from real world and its reflection in stereo performance can produce fatigue and visual discomfort. [].

Several factors have been identifies as a source of visual fatigue during the viewing of stereo contents and should be taking into account to assess visual comfort of stereoscopic video contents, including range, average and maximum negative disparity, distribution of positive and negative parallax, spatial complexity, depth position, temporal complexity, motion and variation between consecutive frames [CHOI-1, 2012].

Fig. 43. Conceptual depiction of spatial factors: (a) Max, average and range; (b) Screen; (c) Spatial complexity and (d) Depth position (image extracted from [CHOI-1, 2012])

Among this collection of factors, four different cases of study have been analyzed in order to establish recommendations about disparity settings for reducing visual discomfort when observing stereoscopic contents.

- **Average and range of disparity:** this parameter corresponds to the average and absolute distribution of the depth distribution taking into account both positive and negative parallax. Works defined by Nojiri et al. [NOJIRI-2, 2004] and [YANO-1, 2002] demonstrate that visual comfort increases when the average approximates to the screen value, i.e. the parallax zero.

- **Maximum negative disparity:** this parameter assess the maximum value, which an user can observe in front of the screen with the highest simulated proximity to observer's eyes. Negative parallax has large influence, which is why this factor is important for describing the probability that visual discomfort happens. Additionally, distribution and percentage of negative disparities on a video sequence is important for complementing the function of this parameter because distribution of disparity should be enough for being appreciated by the observer.
- **Ratio of disparity summations.** This factor is associated to a predefined threshold that determines the percentage of pixels whose disparity is in the surroundings of the screen level. The sum of disparities inside of the threshold is closely related to visual comfort, and disparities out of this threshold raises the probability of happening visual discomfort.
- **Spatial complexity.** This factor is closely related to range and average of disparity. High ranges could difficulty the focusing effect over the whole image. Human eye has difficulties to focus objects in different depths when observing stereoscopic contents.
- **Depth position.** This parameter is related to the comfort zone, and the proximity of objects to the screen plane.

The work by Hoffman et al [HOFFMAN-1, 2008] analyzed the influence of pixels located at different depths with the usage of a volumetric display, considering three cases: pixels in far image planes (called in this thesis: "back planes"), mid planes ("parallax zero plane" and surroundings) and near planes ("front planes"). This display was useful for evaluating focus cues on perceptual artifacts, fusion problems, accommodation-vergence mismatches and the consequent visual fatigue (Fig. 43).

Considering the factors previously analyzed, four cases are identified in which the probability of discomfort happening is high.

Fig. 44. Scheme of objects with different parallax

The screen plane as a plane with parallax zero, situated between the back (maximum positive disparity) and the front plane (maximum negative disparity). The transition from viewing an object in back plane to viewing another object in the front plane should be considered as a case of study to be quantified. This case of study is explained in Fig. 44.

Fig. 45. Transition from content in planes with positive parallax to planes with negative parallax

The opposite case (Fig. 45), i.e. the transition from an object situated in the front plane to another situated in the back plane is an interesting case of study, which generates difficulties in the time for the human eye to focus both objects.

Fig. 46. Transition from content in planes with negative parallax to planes with positive parallax

The third case of study (Fig. 47) is related to the spatial disparity complexity and range of depths, in an environment of low temporal complexity. When viewing simultaneously two objects in motion situated in the front and back planes of the image, the probability of happening visual discomfort rises.

Fig. 47. Image with content in different planes and motion

Finally, the last case of study (Fig. 48) is derived from the viewing of objects in the front plane in motion with high levels of temporal complexity, i.e. fast motion. The time of focusing should be enough for the user to avoid visual discomfort. The situation reflects a transition among negative disparities similar to the second case, but with a determined concrete object.

Fig. 48. Negative Parallax with fast motion

4.4.1 Subjective assessment

Subjective assessment was developed in order to analyze different cases of transitions in sequences with different types of depth maps and disparity diagrams. This variation includes sequences with high amount of pixels with negative parallax, even in hyperstereoscopy areas, and sequences with disparities in positive parallax area. Results were collected in published work [LOPEZ-3, 2015].

Methodologies for the assessment of picture quality were developed following procedures included in the Recommendation ITU-T P.910 [REC P.910] and BT. 2021 [REC BT.2021]. The method used was single-stimulus (SS), since the objective was to evaluate the quality perceived by the observer, while reducing the duration of each individual session. The test sequences were presented only once in the test session, while in the beginning, dummy sequences were presented as a reference for the user, just to make the procedure understandable. Observers taking part in the tests were asked to evaluate the transition between pairs of videos in each sequence. For expressing their disagreement if detecting visual discomfort while watching the change the MOS scale has been used for considering how comfortable the eye focus process was. This scale is described in Table 5.

Table 5. Levels of the scale used in tests

MOS Scale	Annoyance derived from transition	Quality of Experience
5	Very comfortable	Excellent Experience
4	Comfortable	Good Experience
3	Mildly uncomfortable	No visual discomfort
2	Uncomfortable	Visual discomfort
1	Extremely uncomfortable	High visual discomfort

The observers were placed at 2.5 meters from the screen. It is commonly known that the recommended optimum distance for HDTV systems varies depending on the desired angle of vision. ITU-T P910 [REC P.910] specifies a wide range of distances ranging from 1 to 8 times the

height of the screen. A typical recommendation of 1.6 times the screen diagonal offers a 30° angle. Others prefer an even higher angle of 40° to increase immersive experience. This means a distance of 1.2 times the screen diagonal. In tests, viewers were standing at 2.5 meters from the screen, 1.5 times the screen's diagonal (65"). Description of the device used in the tests is included in Table 6.

Table 6. Basic characteristics and settings of device

Device	Settings
Resolution	3840 x 2160p
Brightness	450 cd/m ²
Size	65" Widescreen
Technology	LED Ultra HD
3D	Passive glasses

72 observers took part in this testing and fulfilled the whole 20 sequences. It is recommended between 4 and 40 observers, but 72 will offer more accuracy to the study. They were previously explained the tests with video sequences as examples. In addition, their capacity to see 3D contents was analyzed with sequences to demonstrate their validity for tests. Most participants were male, not only scholars but also faculty staff, mainly led by curiosity to complete the test. A vast majority (90%) was considered non-professional observers. Professional observers, used to this kind of assessment, carried out only small percentage of the tests. People less than 20 years old and between 20 and 30 years old represent the highest percentage in the sample. Only a small group of people was over 30 years old. Only 3% of the observers were considered outliers. Their opinions were discarded because their scores were out of the correlation margin.

Video sequences used for subjective assessment were mainly based on stereoscopic database developed by Urvoy *et al* [URVOY-1, 2012]. In addition, synthetic sequences specifically created for this work with specific parallax conditions were used for this study. Professional software for 3D modelling allows obtaining effects of hyperstereoscopy. Finally, other sequences used came from different sources. The encoding process is not considered in the results, only the type of contents appeared on the video sequence assuming an acceptable quality in acquisition.

The advantage of using virtual cameras for creating synthetic sequences is that it helps to control disparity modifying the distance between left and right cameras, and position of the objects closer or further from the screen.

A session of twenty video sequences was presented to the observers in the subjective assessment tests. Each sequence was composed of two videos, and the users were asked to assess the visual effect provoked by the transition between them.

As mentioned, the method used for the tests was single-stimulus (SS). The observers, in groups of four or five people, are presented each video of duration between 10 and 30 seconds only once.

Next, the pairs of videos are collected in Table 7, describing the reason why they were selected. Most of the cases, each transition represents a variation in type of disparity (mostly in negative parallax environment) in the same or in a different area of the image, with a higher or lower temporal entropy, corresponding to different motion levels.

The selection of videos generating transitions depends on the distribution of parallax and disparity related to position, amount of pixels and type of parallax. Hyperstereoscopy was considered when there are groups of pixels with negative parallax of at least 10 pixels of disparity.

According to the evaluation of observers, the most significant results obtained from video pairs are analyzed in this section. The sequences in which observers have detected visual discomfort in the change from first video to the second are collected in Fig. 49. Sequences marked in red are the ones with a higher probability of generation visual discomfort, while sequences marked in yellow have also a high probability of producing visual discomfort but in a lower manner.

Fig. 49. Subjective assessment related to the appearance of visual discomfort for each sequence.

Temporal Information is considered to assess the quality of the transition and the probability of visual discomfort to occur. In static images, the visual discomfort is possible but

in an environment of motion, the probability is higher even with better conditions of disparity. For that reason the TI value is in last second of the video sequence, in this case the last 25 frames for a 25fps sequence, in order to consider the real previous time to the transition. The value is selected considering that the accommodation and vergence response to step stimuli have duration between 0.5 and 1 second. The surrounding frames to the transition should represent an environment of medium or fast motion to determine a high probability of visual discomfort to occur.

Table 7. Video Sequences Used in Video Quality Assessment

S	Video 1	P+ (%)	P- (%)	HS	Video 2	P+ (%)	P- (%)	HS	TI1	SI1	TI2	SI2
1	Ladder	27	7	Y	Plunger	30	10	Y	20	50	19	62
2	Boxers	25	15	Y	Dance	24	15	N	19	50	16	44
3	Barrier	27	10	Y	Basket	36	8	N	21	59	41	71
4	Hall	29	15	Y	Lab	33	12	Y	5	82	12	53
5	Line	28	9	N	Frontline	29	5	Y	6	40	13	44
6	PhoneCall	32	8	Y	Soccer	28.1	15	Y	13	36	38	89
7	Earth	34	9.7	N	Explosion	32	6	N	11	44	22	20
8	Space	33	15	Y	Moon	25	13	Y	15	32	18	57
9	Shoot	31	5.5	N	Laser	31	5	Y	64	19	63	19
10	Frontline	29	5.3	Y	Line	28	9	N	13	44	6	40
11	Soccer	28.1	15	Y	TreeBranch	24.5	18	N	38	89	14	101
12	Angel	22.4	14.7	Y	Ladder	27.7	7.1	N	7	73	20	50
13	PalcoHd1	12	9.5	Y	TreeBranch	24	18	N	17	60	14	101
14	Spaceship	24	17	N	Astronaut	25	7	N	14	73	9	96
15	Muscles	27	14	N	Aphrodite	25	17	N	2	62	8	80
16	Station	24	13	Y	Itaca3d	25	5	Y	32	60	6	53
17	PalcoHd2	32	10	Y	PhoneCall	32	8	Y	17	60	13	36
18	Chairs	23	10	N	Muscles	27	14	N	7	76	2	62
19	Hanging	30	6	N	Venus	23	15	N	13	76	10	90
20	Itaca3d	25	5	Y	TreeBranch	24.5	18	N	6	53	14	101

Where P+ (%) is the distribution of average positive parallax in the video sequence; P- (%) the distribution of negative parallax in the video sequence; HS when hyperstereoscopy happens in the sequence: Yes (Y) or No (N); TI means Temporal Information; SI means Spatial Information; S is the sequence title.

4.4.2 Determination of depth maps and parallax histograms

The determination of an image parallax range is obtained through an associated depth map. There are several ways to obtain a depth map from a stereo image, depending on computing complexity and accuracy restrictions. As a rule of thumb it can be stated that complexity is proportional to accuracy, thus, low complexity algorithms such as Sum of Absolute Differences (SAD) can perform well under certain circumstances, as stated in [LECLERQ-1, 2002] and

[LECLERQ-2, 2002]. Scharstein and Szeliski [SHARSTEIN-1, 2002] described the procedures to define the Middlebury Database [MIDDLEBURY DATABASE], which offers the most complete stereoscopic algorithms benchmark to date, and software to evaluate new algorithms to predict image parallax.

SAD based algorithms are among the least complex, but more often used. Census' based algorithms, first introduced in [ZABIH-1, 1994], are common in real-time hardware based systems [WOODFILL-1, 1997] [JIN-1, 2010] [CHANG-1, 2010]. These algorithms are demonstrated to work better in homogeneous zones of the image. Its complexity increases when used in software-based systems because of its bit based nature. The mixture of both types of both algorithms improve the results as demonstrated model described by Mei *et al* [MEI-1, 2011], which is ranked in second place in the Scharstein and Szeliski list, and by Chambon *et al* [CHAMBON-1, 2011] based on dense stereo matching.

The computation of depth maps from stereoscopic images for the system is performed with a SAD based algorithm. As it has been stated earlier, other algorithms are more powerfmiddleber computational demand. We need to obtain general depth characteristics of a scene and its evolution, though pixel depth accuracy in the whole image is not necessary. SAD based algorithms work well enough to fulfill our goal and are less computational demanding.

The weakest detections with SAD algorithms occur in homogeneous zones, where the capability to discern between possible pair candidates is low. In order to alleviate these probable errors, the system performs a difference between both views in order to calculate depths only over those pixels that will differ from one image to another, reducing homogeneous zones and, therefore, noise in resulting depth maps. Discarded pixels are not taken into account for statistical calculations. For that reason, the histogram contains less depth information, but more accurate. The purpose of this research is predicting visual discomfort and depth maps are a tool for obtaining this, not the purpose itself.

Fig.50 shows the original depth map (left) and the filtered depth map (right). In the original depth map there are several errors in the background zone, where the sky is homogeneous. In the left image this homogeneous area is discarded by the usage of a mask which determines the difference between left and right views. Therefore, this area is not taken into account to classify general depth in the image parallax histogram.

Fig.50. Depth map. (P refers to “parallax”)

Fig. 51 shows the histogram calculated for the previous depth map. All the elements in the scene have positive parallax. There is a very small amount of negative parallax pixels, which represent noise (bad information) that result from the calculation of the depth map algorithm.

Fig. 51. Example of parallax histogram corresponding to positive parallax image with hyperstereoscopy

4.4.3 Parameters of disparity complexity: Spatial Parallax Information (SPI) and Temporal Parallax Information (TPI)

Recommendations such as P.910 [REC P.910] suggest parameters TI (temporal information) and SI (spatial information) for characterizing video sequences for quality assessment. These two parameters are necessary to justify the usage of a dataset of video sequences for being the core of an objective or a subjective assessment research, but they are insufficient when the environment is 3DTV stereoscopic video. 2D methods have severe limitations when analyzing stereoscopy. For this purpose, we suggest two additional parameters based on disparity called SPI (spatial parallax information) and TPI (temporal parallax information) to characterize 3DTV video sequences [LOPEZ-2, 2015].

Recommendation BT.2021 [REC BT.2021] offers specific requirements for stereoscopic quality tests. These requirements are specifically focused on analyzing the observer's correct vision, when involved in subjective assessment, but it lacks of information about characterization of video sequences used in the studies. On the other hand VQEG research work is currently working on defining new models for 3D Video Quality Assessment [HUYNH-THU-1, 2010] [VQEG-1, 2015]. Other proposals of new requirements for subjective assessment includes depth rendering abilities and limitations of conventional 2D recommendations discussed by Chen et al [CHEN-3, 2010], which are important for understanding new characteristics of 3DTV video sequences. Depth rendering and analysis of different multiview displays are part of the new considerations to have in this special type of video assesement, which are very useful for the evolution of complexity in stereoscopic sequences.

Algorithms such as Sum of Absolute Differences (SAD) perform well calculating parallax diagrams and depth maps under most circumstances, as stated by Scharstein et al [SCHARSTEIN-1, 2002], with low computing charge. Parallax is necessary to define the stereoscopic complexity of each video sequence and their evolution through the sequence. This depth evolution represents a fundamental part of the algorithm for predicting visual discomfort. From parallax estimation, both parameters Spatial Parallax Information (SPI) and Temporal Parallax Information (TPI) are obtained, offering extra information regarding the sequence disparity complexity. By calculating SPI and TPI parameters, new three-dimensional complexity relative to spatial and temporal information with the stereoscopic component can be found, and this information is found to be an important component for achieving an accurate prediction of visual discomfort.

- **Spatial Parallax Information (SPI)**

The spatial parallax information is based on the depth variation in every single frame of the sequence. Parallax diagrams for each video frame (luminance plane) at time n (F_n) are processed. SAD algorithm is used to calculate parallax diagrams to evaluate depth variation in the sequence as explained before, obtaining the parallax function $P[F_n(i, j)]$ over luminance component for every single pixel.

The standard deviation over space (std_{space}) in parallax diagram of each frame is then computed, after all frames in the sequence are computed. The maximum value for a frame in the complete time sequence is defined to represent the parallax spatial information content of the scene. In equation form, the SPI parameters can be defined as follows.

$$SPI = \max_{time} \{std_{space}[P(F_n)]\}$$

- **Temporal Parallax Information (TPI)**

The temporal parallax information parameter is based upon the depth difference function $D_n(i, j)$, which is the difference between the values of parallax of the pixels $P[F_n(i, j)]$ when calculating SAD function at the same location in space at successive frames. Being $F_n(i, j)$ the value of luminance in each pixel from frame n , depth function is defined as:

$$D_n(i, j) = P[F_n(i, j)] - P[F_{n-1}(i, j)]$$

The measure of temporal parallax information (TPI) parameter is computed as the maximum over time ($\max_{time}$) of the standard deviation over space std_{space} of $D_n(i, j)$ over each i and j .

$$TPI = \max_{time} \{std_{space}[D_n(i, j)]\}$$

More parallax variation in adjacent frames will result in higher values of TPI. This value is calculated in previous frames to transition for determining the complexity of human eye's focusing process.

4.4.4 Detection of Stereoscopic Window Violation (SWV)

Window Violation is considered one of the main phenomena related to visual comfort in stereoscopic video. This effect occurs when depth cues are located in the extremes of the frame. The cue is found by human eye only in one of the left and right views, and the brain is not able to fuse that part of the image in a correct manner.

Window violation occurs when an object with negative parallax does not fit the screen and, therefore, is cut by the screen edges. Having negative parallax, it is supposed to be out of the screen, which means that screen edges should not be able to hidden its view. This generates an incoherent depth cue situation.

In order to measure this feature, the algorithm will examine the depth map's limits looking for negative parallaxes, which will compute as a factor of visual annoyance. Due to the difficulties of obtaining disparities near the edges, a group 10 pixel rows or columns closest to the edges are analyzed to decide whether a window violation is detected. Fig.52 represents how the algorithm is computed.

Fig.52. Working scheme for detecting Window Violation with exaple of sequence "Itaca 3D"

4.4.5 Final Algorithm for Predicting Visual Discomfort

After developing subjective assessment, the parameters to generate an automatic algorithm for testing transitions in pairs of video sequences have been obtained. These parameters are related to temporal information and parallax variation.

$$p(v.d.) = f(TI_1, TI_2) + f(TPI_{1-2}, SPI_{1-2}) + f(p_{1-2})$$

where $p(v.d.)$ is the probability function of visual discomfort, $f(TI_1, TI_2)$ is a function of correspondence between the Temporal Information of both sequences, $f(TPI_{1-2}, SPI_{1-2})$ is a function of parallax complexity evolution for each sequence and the transition between them, and $f(p_{1-2})$ the variation of parallax between last frame of the last sequence and the first of the next one.

The probability of occurring visual discomfort depends on the temporal information (TI) in the first video and in the second $f(TI_1, TI_2)$. The higher the speed is, the higher is the probability of visual discomfort. As demonstrated in sequences with higher visual discomfort detected, the TI (temporal information) variable was over 10. This value was taken into account as a higher probability of occurring visual discomfort.

TI is calculated according to the depth function for each single pixel of luminance from one frame to the next one considered in the last 20 frames before the transition. Formulas follow the indications included in ITU-T P.910 [REC P.910].

$$D_n(i, j) = f_n(i, j) - f_{n-1}(i, j)$$

$$TI_n = \max_{time} f_n(i, j) - f_{n-1}(i, j)$$

$$TI = \max_{time} \{std[D_n(i, j)]\}, n = 25(25fps)$$

Disparity is the second parameter for the function $f(TPI_{1-2}, SPI_{1-2})$. Abrupt changes in disparity, as happens in pair of sequences 16 increase the probability of visual discomfort, especially in a negative parallax environment (i.e. hyperstereoscopy), with values of change over 40 pixels.

Finally, the third parameter refers to the area of the image, which suffers a higher change of parallax. The function $f(\rho_{1-2})$ measures the variation of parallax between each individual pixel from the first video sequence compared to this pixel in the second one. In formula (4), the probability of occurring visual discomfort is dependent of the variability of parallax in each individual pixel.

$$\rho_{1-2} = \sum_{j=0}^H \sum_{i=0}^W |\rho_1(i, j) - \rho_2(i, j)| \quad (4)$$

Where $\rho_1(i, j)$ is the parallax of each pixel of the first frame, $\rho_2(i, j)$ the parallax of the second frame, and $\rho_{1-2}(i, j)$ the difference of parallax between first and second frame for each pixel.

The definition of these three parameters present an improvement in the final algorithm, related to the results obtained in first subjective tests enables the correct environment for detecting the most probable discomfort occurring.

Tools developed for these three factors were implemented and results for relating the scores obtained in subjective assessment with the final algorithm results are obtained. Different tools to measure visual discomfort justified the occurrence of visual discomfort or not in the tested sequences.

4.5 Results of Visual Discomfort Prediction

Visual discomfort is considered for users scoring 1 or 2 value of MOS scale, and the rest of scores are considered of producing no visual discomfort.

The sequences, in which more than 40% of the sample complained about the content, were considered as conflictive, as occurred in sequences: 4, 12, 13, 16, 17 and 20. In addition, other sequences whose result is over 20% of the sample have been also considered, as happens with sequences: 3, 11 or 18. These sequences are analyzed in next section, to decide which characteristics of these videos affects to human eye, producing visual discomfort.

It is necessary to understand the scale in Fig. 53 for the interpretation of depth maps or diagrams of disparity. The shades of grey represent the levels of parallax from -100 to 100, the darker shades would be the points when hyperstereoscopy occurs and the lighter gray would be the areas of positive parallax.

All depth maps in this section are normalized between -100 and 100 pixels of parallax, which is the highest distribution for this collection of sequences.

Fig. 53. Level of parallax scale of to understand disparity diagrams, commonly called “depth maps”

- **SEQUENCE 1: Transition from “Ladder” to “Plunger”**

Different accommodation and vergence responses are induced when viewing S3D on a stereoscopic display in contrast to viewing an object in the real world. The difference between accommodation and vergence induces HVS to accommodation-vergence mismatches.

Depth maps for frames in which the transition in sequence 1 occurs are included in Table 8. The center of the images present the highest negative disparities and consequently the nearest to the viewer objects. The transition is from near object to a further object, which is also close to the viewer’s HVS.

Table 8. Depth maps for transition in sequence 1

Original Frames	Depth Maps

As mentioned before, when viewing S3D the accommodation input stimulus is the distance from the viewer’s HVS to the screen. HVS accommodates to focus the image located at the fixation point, which will always lie on the screen, into focus. The fixation point of the image for sequence 1 is [900:480], considering this point as a representative pixel which the user is most probably observing (Fig. 54).

Because of the difference between the input depth stimuli for accommodation and vergence, there exists conflict between the associated output responses. Hence, the steady-state amplitude responses of accommodation and vergence are not equal to the input depth stimulus. The transition in sequence 1 have variations in negative parallax from -30 to -15, which is a high variation over 10 points.

Fig. 54. Parallax for transition in pixel [900:480] of sequence 12, marked in red

Following the procedure of measuring the response to a step stimulus (Fig. 54), the simulation of results for the stimulus corresponding to the transition of sequence 1 is included in Fig. 55. The time for focusing the first step stimulus when the second stimulus occurs is a change in vergence response. On the other hand, the accommodation response remains constant because the distance to 3D display is the same in both instants. If these two images were static the time for vergence, which is between 0.5 and 1 second would be superposed generating a conflict with the accommodation response. These stimuli are part of a sequence in motion instead of step stimuli as reflected in the TI diagram (Fig. 57) with variances over 5 points. As this effect occurs, the complexity for HVS to focus is higher and the vergence response would generate a mismatch with the accommodation response. HVS attempts to force the responses to lie inside the Panum's fusional area and DOF driven by the accommodation process and by fusional vergence. The accommodation-vergence mismatch inevitably occurs, which can lead to visual discomfort.

Fig. 55. Response to a stimulus (transition in sequence 1) in accommodation and vergence in a 3D Display

When the stimuli, which are part of the transition, is relative to near objects located at different distances to the observer, the viewing process consist in a fast crystalline accommodation. The transition from points out of the ZoC and with a difference of more than ten pixels of disparity is complicated for the eye. The fast change in the transition provokes a fast accommodation process in the crystalline. During accommodation, the anterior chamber depth decreases for focusing the near objects as shown in Fig. 56, and the crystalline lens consequently increases in thickness. Additionally, the lens decreases in diameter with a reduction in the radius of curvature of the posterior lens surface. Studies by Kasthurirangan *et al* reported ciliary muscle movement with accommodation and age, especially in humans [KASTHURIRANGAN-1, 2015]. The thickness of the crystalline increases and becomes more rounded with age, which increases the probability of visual discomfort to occur.

Fig. 56. Viewing near objects at different distances (Sequence 1)

The time of the stimulus should have enough duration for the human brain to be detected, with variation between consecutive frames of at least 5 points. The accommodation and vergence conflict may conclude with visual discomfort occurrence because of the temporal variation between frames. In static images, the probability of visual discomfort is lower because the time for focusing at fixation point would be enough.

Fig. 57. Temporal Information frame by frame in transition of "Sequence 1"

Less than 20% of asked people detected discomfort when viewing the transition included in this sequence (Fig. 58). Although the scores obtained in subjective tests revealed that no discomfort should occur in the transition of this sequence, the truth is that according to statistics and graphical results, the sequence should represent a clear case of visual discomfort. The reason of this effect is that in Quality of Experience other factors influence the opinions. Some users found special affinity with the content in the sequence and that factor affected to their final opinion.

Fig. 58. Scores from subjective assessment to "Sequence 1"

• **SEQUENCE 12: Transition "Angel" to "Ladder"**

Transition presented in sequence 12 corresponds to the opposite case as transition in sequence 1. The objects in image before transition are located near the viewer's eye (parallax: -10 in central area), but the objects in subsequent image are located considerably closer to viewer's HVS (parallax in central area: -30).

As can be seen in disparity diagrams of videos "Angel" and "Ladder" (Table 9), the negative parallax is in the center area of the picture. The abrupt change from negative parallax to a bigger amount of negative parallax pixels produces a visual discomfort in the observer. 40% of the people give a score that manifests visual discomfort, derived from a change in disparity that affects to the same area in the video sequence. Motion in the first video is medium, but faster in the second, what affects to the final score.

Table 9. Depth maps for transition in sequence 12

Original Frames	Depth Maps

Histograms revealed that the transition from negative parallax to a disparity in a range even more negative produces discomfort to HVS. More than 20 pixels of disparity in transition, difficulty for the eye to focus, whose response to the second stimulus oblige the crystalline to change in a quick manner.

Fig. 59. Histograms of sequence 12, from frame in video12a (up) to frame in video 12b (down)

Analyzing the saliency of the distribution of both frames, the region of pixels with most probable visual attention is the area located at center of the picture. Considering the change in one of these pixels [982, 563] the variation of disparity is more than 15 pixels length in the transition. This is high variation of parallax, which is translated into a source of probable visual discomfort.

Fig. 60. Parallax for transition in pixel [982:563] of sequence 12, marked in red

The TI diagram (Fig. 61) presents temporal variation of more than 10 in the surrounding frames of the transition. This variation is high enough to consider the environment of medium motion, which is translated in forcing the eye to focus through a series of stimuli.

Fig. 61. Temporal Information frame by frame of "Sequence 12"

Unlike the previous example of transition in sequence 1, about 40% of asked people admitted presenting discomfort when viewing the transition included in this sequence (Fig. 62). The disparity diagrams analysis and the subjective response coincide in the occurrence of visual discomfort in this sequence.

Fig. 62. Scores from subjective assessment to "Sequence 12"

- **SEQUENCE 14: Transition from "Spaceship" to "Astronaut"**

This sequence was well scored by the observers. Fluctuation from negative parallax in right side of first video to negative/positive combination of parallax in the second one makes the human eye perceive no visual discomfort. Depth maps are collected in (Table 10), showing that the most salient disparities are located at right side of the first video, while in the second video disparities are more expanded and located at central area.

Table 10. Depth maps for transition in sequence 14

Original Frames	Depth Maps
-----------------	------------

For this transition, less than 20% of people asked rated this sequence as a source of visual discomfort (Fig. 63). Although the scores obtained in subjective tests revealed that no discomfort should occur in the transition of this sequence, the truth is that according to statistics and graphical results, the sequence should represent a clear case of visual discomfort. The reason of this effect is that in Quality of Experience other factors influence the opinions. Some users found special affinity with the content in the sequence and that factor affected to their final opinion.

Fig. 63. Scores from subjective assessment to "Sequence 14"

The analysis of depth maps in combination with their respective histograms revealed that the transition occurs from disparities inside the ZoC. The last frame of the first video in the transition possesses disparities with values over -10, which is a low negative parallax (Fig. 64). On the other hand, the first frame of the second video presents positive parallax in the same region of the image (Fig. 65). Although the abrupt change occurs in negative parallax areas, the transition occurs inside of Shibata's ZoC. Consequently, discomfort decreases even when image is presented in front of the screen.

Fig. 64. Disparity histogram of last frame in video 14a, where the transition of sequence 14 occurs

Fig. 65. Disparity histogram of first frame in video 14b, where the transition of sequence 14 occurs

• **SEQUENCE 16: Transition “Station” to “Itaca3d”**

The first video of this sequence is a long travelling with high motion and negative disparities in both sides of the picture (Table 11). On the other hand, hyperstereoscopy derived from a synthetic modelling, with controlled variation of negative disparities, affects the second video.

Table 11. Depth maps for transition in sequence 16

Original Frames	Depth Maps

More than 80% of observers in the survey presented visual discomfort after viewing this transition, as shown in (Fig. 66). Motion and high negative parallax in the transition makes the

human eye to assess this sequence as the worst scored in the tests. The transition in this sequence exceeded the limits of disparity, and the environment difficulties the accommodation-vergence linkage.

Fig. 66. Scores from subjective assessment to "Sequence 16"

A collection of factors influence the results obtained after viewing this sequence. On first instance, the temporal information reveals that fast motion is detected in the first sequence. High values of TI (Fig. 67) oblige the human eye to focus more rapidly than necessary, especially considering an environment of negative parallax. As shown in Station's depth map (Table 11), laterals of the video frame are located at areas in front of the screen, producing a box effect in spectator's eye.

Fig. 67. Temporal Information frame by frame of "Sequence 16"

Additionally, the high variance in disparity histograms of both sequences (Fig. 68) reveals an abrupt transition with a consequent accommodation-vergence mismatch.

Fig. 68. Disparity histograms in transition of sequence 16.

In central areas of images (Fig. 69), the strength and size of the transition is so high that crystalline have not enough time for adapting. The consequence is a high probability of visual discomfort to occur.

Fig. 69. Parallax for transition in pixel [965:542] of sequence 16

- **SEQUENCE 2: Transition from “Boxers” to “Dance”**

Moderated negative parallax is included in both videos. In first video, the main disparity is found on the left side, changing in the second video to significant negative parallax in the right side of the picture (Table 12).

Table 12. Depth maps for transition in sequence 2

Original Frames	Depth Maps

Depth maps show that most of the pixels are inside the ZoC, surrounding the zero parallax plane in both sequences (Fig. 70). On the other hand, a significant volume of pixels is located at areas of negative parallax out of the ZoC (values under -10). Visual discomfort could probably occur, but the location of negative parallax in different areas of the picture decreased this probability.

Fig. 70. Disparity histograms in transition of Sequence 2

Considering the salient objects of the first sequence, which would be the human figures, the visual attention is found in the lateral and central area of the first video frame. On the other side, the objects nearer the viewer's eye are located at the right side. Unlike the transition occurs in negative parallax areas and the modulus of this transition is high, the areas of visual attention between both frames are different. This effect is translated as a decrease in probability of visual discomfort.

Fig. 71. Parallax for transition in pixel [132:401] of sequence 2, marked in red

As predicted analyzing the depth maps and disparity histograms, the observers ranked this sequence as a comfortable transition. Less than 10% of observers confessed having presented some discomfort after viewing this sequence.

Fig. 72. Scores from subjective assessment to "Sequence 2"

- **SEQUENCE 4: Transition "Hall" to "Laboratory"**

The transition in this pair of videos consisted in a transition of low negative parallax in both videos. This negative parallax is located at central area of the picture for first video and distributed all along the picture in the second one. Video "Laboratory" presents high window violation in the down side of the picture, what affects to scores from observers (Table 13).

Table 13. Depth maps for transition in sequence 4

Original Frames	Depth Maps

Observers ranked this sequence as source of visual discomfort. Concretely, more than 40% of observers detected annoyance in the transition (Fig. 73).

Fig. 73. Scores from subjective assessment to "Sequence 4"

• **SEQUENCE 13: Transition from “Palco HD” to “Branches”**

Sequence 13 is interesting to analyze because the transition occurs from a synthetic sequence with control and progressive parallax evolution to a sequence with uniform distribution of disparity beyond the screen. Depth maps are collected in Table 14, revealing areas of high negative parallax in the first frame.

Table 14. Depth maps for transition in sequence 13

Original Frames	Depth Maps

Analyzing the subjective results, Sequence 13 is considered by most observers as a high source of visual discomfort. More than 60% of the observers asked revealed their disconformity after viewing this sequence, as shown in (Fig. 74).

Fig. 74. Scores from subjective assessment to "Sequence 13"

The reasons why this sequence was ranked as mentioned are included next. The variation of values of parallax shown in disparity histograms (Fig. 75) revealed an accommodation-vergence mismatch in the transition. A high amount of pixels near the observer's eye difficulties the focus to pixels with positive parallax beyond the screen. The variation of more than 80 pixels in the transition justified by itself the occurrence of visual discomfort.

Fig. 75. Disparity histograms for Sequence 13

More concretely, the response to the stimuli of a pixel located at areas of visual attention (pixel [942:206]) has been analyzed (Fig. 76), where a value of -40 in disparity was measured. The modulus of disparity from this variation is high enough to generate visual discomfort. Areas under -10 of disparity start being difficult to focus for human eye because of the crossed-eyes vision.

Fig. 76. Parallax for transition in pixel [942:206] of sequence 13, marked in red

Sequence “Palco HD” included in this transition is highlighted because of the progressive evolution of parallax in the objects. The separation of cameras, over the levels of average interpupillary distance, simulated the approximation of the letters. Asked observers revealed that no visual discomfort appeared in the progressive variation of parallax to high negative values. On the other hand, when abrupt change from these high negative values to a parallax inside the ZoC occurs, the sensation is very annoying for their eyes. The evolution of disparity frame by frame in pixel [942:206] of this sequence is collected in (Fig. 77).

Fig. 77. Parallax evolution in complete "Palco HD" sequence in pixel [942:206]

• **SEQUENCE 20: Transition from “Palco HD” to “Branches”**

The transition produced in sequence 20 is characterized by the abrupt change from an image with salient objects located at different distances from viewer’s eye to a sequence with a homogeneous distribution of positive parallax with low motion. The depth maps of these two videos are collected in Table 15.

Table 15. Depth maps for transition in sequence 20

Original Frames	Depth Maps

Itaca 3D possess salient objects located at different depths. The objects have independent motion, what difficulties the focus of the crystalline on different parts of the frame simultaneously. The abrupt transition from this situation to a frame with different distribution of parallax increases the probability of visual discomfort to occur.

The sequence presents enough motion for considering that the eye is changing focus continuously in the previous instants to the abrupt transition. The TI diagram evaluates motion, which is collected in (Fig. 78), evaluates motion.

Fig. 78. Temporal Information frame by frame of "Sequence 20"

Supposing that the visual attention is focused on the spaceship front, which is the closest cue to the viewer, the evolution of parallax collected in (Fig. 79) is measured in pixel [1044:607]. As seen on the graph, the change between both frames in this ROI is more than 25 pixels of disparity, which is translated in a considerably fast transition for the crystalline to modify its focus. This effect increases the probability of the eye to suffer from visual discomfort.

Fig. 79. Parallax for transition in pixel [1044:607] of sequence 20, marked in red

Unlike the transition occurs from negative parallax to positive parallax, the subjective scores revealed a high probability of visual discomfort (Fig. 80). The difficulty of human eye to focus in transitions with high modulus variation is translated as a source of discomfort.

Fig. 80. Scores from subjective assessment to "Sequence 20"

4.6 Conclusions to visual discomfort prediction

The work developed in this section analyzed the probability of visual discomfort to occur in relation to stereoscopic display technology and 3D video content generation.

The development of multiple kinds of measurements, both objective and subjective, is necessary for adequately characterizing and understanding visual fatigue and visual discomfort.

Preliminary subjective tests revealed that abrupt changes of disparity from one frame to the following one increase this probability. A progressive change of disparity, even when this change is produced in areas of hyperstereoscopy, does not necessarily induce to visual discomfort, if human eye have time to adapt to each new stimulus. For that reason, new subjective tests were developed for implementing tools that predict automatically the cases in which visual discomfort most probably occurs.

Excessive binocular disparity detected in the areas of visual attention, combined with fast motion, difficulties the crystalline to focus. When viewing near objects, the HVS must focus faster. When this happens, viewing objects in real world, then consistency between accommodation and vergence exists. On the other hand, accommodation and vergence mismatches occur when viewing an object on a 3D display, especially in 3DTV where the distance to the screen is lower than in a big theatre.

Changes in accommodation generate conflicts in vergence during the transition from one image to another, which induce to visual discomfort.

Depth and motion are main factors in perceived quality of experience. Information provided by depth maps and estimated temporal information is useful to avoid effects that can cause visual discomfort and fatigue in observers when contemplating 3d stereoscopic contents.

Subjective assessment allowed us to isolate the main features to be detected, in order to perform an algorithm, which could translate the user's opinion to an automatic objective system.

As demonstrated after testing with both subjective and objective assessment, three parameters must be taken into account to obtain adequate results in the final algorithm:

- Temporal information of both sequences implied in the transition.
- Difference of disparity in same areas of both frames
- Parallax distribution of both frames.

Information obtained from depth maps, its associated histograms and parallax distribution of each image are the main features needed to use for developing the final quality measurement algorithm.

The presence of objects with negative parallax on a static image is not necessarily enough for visual discomfort to occur. Nevertheless, when motion is detected in the video sequence that contains that image, quantifying the probability of observer's annoyance is required. In graphics comparing parallax and motion evolution, presented in the tests section, is remarkable the relation between both parameters in the final experience of users. Previous Zone of Comfort (Zoc) studies demonstrate that motion and time of visualization affect greatly the limits, diminishing its range significantly. Parallaxes getting near the ZoC edges (especially negative parallax) have been proved to be undesirable when fast motion or high parallax variance appears.

Other artifacts, such as stereoscopic window violation, increase the probability of v.d. to occur, and should be quantified when determining the quality of a 3D video sequence.

Tests developed through this work show good results when applying the techniques to video sequences that contain effects that could be considered annoying for human eye. Algorithms were validated with a collection of sequences with positive results. Tools developed for this purpose revealed the reasons why visual discomfort occurs or not in the collection of sequences analyzed.

These results permitted extracting probabilities of visual discomfort and visual fatigue in order to reach consensus between 3D perception and annoyance over observer's eye. Obtained results recommend special care when creating stereoscopic video, especially in terms of excessive variation of disparities between consecutive frames and fast motion. Nevertheless, in this context, the content provider and the user have the final decision to accept or not a particular content that could offer a strong experience, but not paid with vision problems.

5. Visual Attention Model for Video Quality Assessment

5 Visual Attention Model for Video Quality Assessment

5.1 Work motivation

Video Quality Assessment needs to correspond to human perception. Pixel-based metrics (PSNR or MSE) fail in many circumstances for not taking into account the spatio-temporal property of human's visual perception. For that reason, a new pixel-weighted method is proposed in this doctoral thesis in order to improve video quality metrics with the objective of assessing artifacts such as blurring. The method applies a psychovisual model based on four video characteristics: motion, level of detail, position of the pixel in the frame and the appearance of human faces, which approximate the quality to the human eye's response. Subjective tests have been developed to adjust the psychovisual model and to demonstrate the improvement of the algorithms.

A perceptual phenomenon that has been studied extensively since the 1960s is visual masking. This phenomenon refers to the inability of the HVS to detect one stimulus, the target, in the presence of another, which acts as a mask to the first stimulus. The visibility of the target depends on many factors, in particular frequency, orientation, contrast of both mask and the target or the content itself [BEGHDADI-1, 2013]. For example, if there is a human face in a frame, the interest in the stimulus will be masking the rest of the picture.

Fig. 81. Visual saliency analyzed by eye tracker (image taken from [WOODING-1, 2002])

First, information on what observer really visualizes on the screen must be analyzed. Researchers use their own visual systems to detect emerging patterns and trends in graphical representations of datasets, determining where best to employ further statistical analysis. The spatial relationship between different fixations is not a construction of the analysis used; rather, it is their real spatial separation relative to the stimulus used. This makes the analysis of fixation data by fixation maps both powerful and more straightforward. Visualization is only one application of fixation map analysis and is the least quantitative of the applications described here. However, it does provide a useful and intuitive means of describing and

communicating the underlying patterns of otherwise overwhelming data sets [WOODING-1, 2002].

Finding the observer's regions of interest is a key factor to weight the value derived from a quality metric. The graphical representations of the data described are an immediate and powerful demonstration of a key feature of eye movement data: that the fixations are not spread evenly or randomly over the stimulus, but cluster into regions of interest according to the features of the stimulus (Fig. 81).

The terms saliency, attention, and gaze are often used interchangeably. Attention is a general concept covering all factors that influence selection mechanisms. On the other hand, saliency characterizes some concrete parts of a scene, being these parts objects or regions relative to their neighboring parts. First works in cognitive psychology developed by Treisman et al [TREISMAN-1, 1980] stated which visual features are important and how they are combined to direct human attention. A collection of paradigms were analyzed and tested including visual search, texture segregation, identification and localization, separable dimensions, such as shape and color and local elements or parts of figures such as lines or curves to be integrated into more complex regions. Later, different models were proposed by Koch et al [KOCH-1, 1987], combining these features and introducing the concept of a saliency map, defined as a topographic map that represents conspicuousness of scene locations.

Borji et al [BORJI-1, 2013] summarize a collection of thirteen different features for visual attention analysis in research. For selecting determined features, different factors were taken into account. According to these research distribution factors in order from f1 to f13 are: Bottom-up (f1), Top-down (f2), Spatial and Spatio-temporal (f3), Static (f4), Dynamic (f5), Synthetic (f6) and Natural (f7) stimuli, Task-type (f8), Space-based(+)/Object-based(-) (f9), Features (f10), Model type (f11), Measures (f12), and Used dataset (f13).

The attention model proposed by Ma et al. [MA-2, 2010] for video summarization incorporated classical bottom-up framework extracting semantic cues such as face, speech and camera motion. Each feature generate attention maps separately with fusion schemes to be employed to obtain a comprehensive attention map used for ranking the regions of video content. Perceptual image segmentation involves extraction and grouping pixels with similar information to facilitate automatic image processing [CHEN-4, 2005]. Image segmentation allows obtaining a compact representation from an image. The main attention models developed are based on two factors that participate in the segmentation:

- Spatial entropy, based on diagrams of gradients and objects segmentation, this kind of image complexity offers an idea about the level of detail in the picture.
- Temporal entropy: it is based on optical flow and analysis of motion vectors. This parameter shows the level of motion in the video sequence.

The third factor that affects to the other two previous is the position of the pixel in the image. As proposed by [NOJIRI-2, 2004], the division of the picture in nine quadrants is necessary, because human eye tends to observe on first instance the center, so this area will weight higher than the rest. This factor affects to all of them, and it is related to the size of the screen.

Finally, a fourth factor is extracted from eye tracking studies and visual saliency maps analyzed by Wooding [WOODING-2, 2002], which are related to the content itself. Human faces appearance focus the attention of the spectator from all the other factors, as demonstrated in subjective studies developed for this work. The quality in pixels, which form human faces, should be higher than in the rest of frame, because of the importance of the content for the observer. This effect may be seen in example that appears in Fig. 81.

For future work, an additional image characteristic that would be interesting for analyzing is pixels color. The contrast derived from the color combination of pixels implies a different response to the human. Moreover, some colors highlight distortions in a higher level, influencing in the result of some algorithms.

5.2 State of the art for psychovisual models

Visual attention is a mechanism that filters out redundant visual information and detects the most relevant parts of human visual field [LE MEUR-1, 2006]. For this purpose, psychovisual models are necessary to complement the metric.

Pixel-based metrics, such as Root Mean Square Error (RMSE) or its derivatives (PSNR or SNR), are dominant in practice. However, these metrics do not take into account the spatio-temporal property of human's visual perception, which is the reason why it fails under many circumstances [XIAO-1, 2000]. Under the assumption that human visual perception is highly adapted for extracting structural information from a scene [WANG-3, 2002], quality metrics should offer a closer response to visual perception.

Apart from pixel-based metrics, some QA indices that presume to consider human visual sensitivity, such as mean structural similarity (MSSIM) with structural sensitivity, visual information fidelity (VIF) with statistical sensitivity, analyze the differences between reference and distortion frames. But, they ignore the role of human visual attention (HVA). Ma et al. [MA-2, 2012] proposed a model by adding weights related to saliency features to original IQA (Image Quality Assessment) or VQA (Video Quality Assessment) criteria to approach more closely to human visual system, but it is not focused on the distortions.

In this work, we present an innovative metric based on blurring artifact measurement focused on the artifact itself and the masking process of the different visual features. For classification of the sequence, techniques involved in this metric include detection of motion areas and isolation of edges to create user visual attention and psychovisual models. For example, Zhai proposes a psychovisual metric based on the free-energy principle [ZHAI-1, 2012]. Other examples of combination between VQA and prediction based on visual attention are included in works by Wu et al [WU-1, 2013] and Watson et al [WATSON-1, 2001].

The work in developing metrics for blurring detection is extensive, both with full-reference and no-reference algorithms, as happens in works by [BOVIK-1] search for automatic prediction systems. Other no-reference and blurring research is collected in works by Hu et al [HU-1, 2006], Li et al [LI-3, 2011], Serir et al [SERIR-1, 2011], Kerouh et al [KEROUH-1, 2012], Oelbaum et al [OELBAUM-1, 2009] and Yamada et al [YAMADA-1, 2012].

Visual attention is a main feature of the HVS, but few studies have been done on using it applied to image quality assessment. The use of the visual attention information was investigated by Nisanni et al [NINASSI-1, 2007], concluding that applying the visual attention to image quality assessment is not trivial and, as HVS feature, influence to visual perception and quality of experience. Eye-tracking was used for this analysis, but for this thesis, we propose a method for automatically developing visual attention models focused on quality itself and concrete artifacts. The state of the art for visual attention is extensive based on complex processes such as object recognition or scene interpretation.

On the other hand, the HVS-based methods are based on evaluating the subjective impression on observer's eye. Diagrams of saliency and evaluation of sharpness on edges are helpful to create visual models that focus on user visual attention. These models are useful to weight the error at each pixel by the visual saliency of every pixel and obtain results closer to a human eye's response. Models of visual attention are developed by analyzing two factors of the video sequence: objects motion, which generates a parameter called temporal complexity, and spatial complexity, related to the number of edges and variety of textures in the pixels. The studies developed by Wooding [WOODING-1, 2002] on static images, analyzing with the help of eye tracking, are crucial for the thesis defended by this paper. Wooding describes how visual attention works for pictures, analyzing the objects most probably observed. Other research derived from visual attention is collected in different works by You [YOU-1, 2009], Borji [BORJI-1, 2013] which offers a complete state of the art in visual attention modelling and the saliency model by Itti [ITTI-1, 1998]. Models such as the one developed by Le Meur *et al* [LE MEUR-2, 2007] try to predict visual fixations in order to produce rapid scene analysis. Finally, works by Boujut *et al* [BOUJUT-1, 2012] offer good estimation for a no-reference metric based on semantic saliency maps.

Other researchers have previously developed content-weighted video quality assessment. Li et al [LI-2, 2010] proposed an image model based on three components to define three kinds of regions: texture, edge and smooth regions. This metric is used for static images, but can also be applied to video frame-by-frame. In this paper, we take into account the motion regions to expand the idea to a more extensive environment.

Nojiri et al [NOJIRI-1, 2006] treat another interesting topic about image quality for 3DTV sequences, dividing the picture in quadrants and defining a pondered weight for each quadrant, depending on the importance which human's eye gives to that area of the picture. Central quadrant will be observed firstly most probably than the image corners.

Finally, other interesting studies are based on the evaluating of motion quality, using motion as an emerging parameter for visual attention, such as the ones by Seshadrinathan *et al* [SESHADRINATHAN-1, 2010] and Engelke *et al* [ENGELKE-1, 2010]. Motion is a key factor on visual attention, because motion calls human eye's attention further more than static object.

5.3 Psychovisual Four-Component Model Based on Video Characterization

5.3.1 Introduction

In this section, the five saliency features to be analyzed for the final visual attention model are explained. Borji et al research [BORJI-1, 2013] and their distribution in factors from f1 to f13 were analyzed for detecting appropriate features that could be adaptable to the model. For the selection of features it was necessary to select factors that could be easily implemented by software for generating an automatic model for each frame of a video sequence.

According to context-aware saliency detection model proposed by Goferman et al [GOFERMAN-1, 2012], image regions of interest are detected based on four principles of human attention supported by psychological evidence [TREISMAN-1, 1980] [WOLFE-1, 1994] [KOCH-2, 1999]:

1. Low-level characteristics affecting to each individual pixel, such as color and contrast
2. Global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm.
3. Visual organization rules which state that visual forms may possess one or several centers of gravity about which the form is organized, and
4. High-level factors, such as human faces or concrete objects recognition. This factor could be content dependent, but human faces generate specific patterns in human retina that increase the probability of be perceived related to psychological and cognitive features.

Humans are capable of attending to multiple (up to five) regions of interest simultaneously. For that reason, the model should be a combination of low-level features, with spatio-temporal features and high-level factors such as human faces. Thus, five characteristics of pixels in the video sequence that has an influence in degradation and human perception:

- Spatial entropy or spatial complexity (level of detail). Simple texture analysis of the image can provide a useful cue towards rapid scene identification. Simple texture discrimination explains early scene identification [WALKER-1, 2002] [ITTI-1, 1998]. The search of objects and complex textures among homogeneous areas highlight the high frequencies and edges, with attention results related to high contrast and search of high-level features through low-level spatial features. These results are obtained by understanding neural mechanisms of selective visual attention [DESIMONE-1, 1995].
- Position of each pixel in the frame. An eye focusing density map predicts positions that people likely pay attention to base on the stochastic saliency map and eye movement patterns an eye movement pattern controls the degree attention through a Hidden Markov Model [PANG-1, 2008] [VERGHESE-1, 2001]. This feature could be englobed in the visual organization rules, with Gaussian models that increase the probability from the center to the corners of the frame. This factor is dependent on the screen size. The model derived from this factor was simplified to a generic television between 32" and 46" by the one used by Nojiri et al. [NOJIRI-2, 2004].

- Motion derived from objects in the scene (not considered the one produced by camera capture). Human beings see stable and moving components in a moving scene effortlessly. An object tracked by the camera is seen as moving even if it is stationary on the frames. We assume that visual attention is attracted by motion contrast and we define it as the motion of regions against background [MARAT-1, 2009].
- Human faces detection, which are context-aware high-level features but very important for psychology of visual perception.
- Color and contrast, which are low-level features affecting to different regions of the image. An analysis of the colors and experiences, which are more visual and the ones that mask distortion with higher probability, could require an expansive analysis that is out of the objectives of this doctoral thesis.

The first four characteristics are explained in detail through next subsections. Color and contrast are selected for future work for its special properties.

5.3.1.1 Spatial entropy (level of detail)

Spatial complexity becomes a key factor while coding the frames of the video sequence. When the image presents a great amount of high frequencies, this means that the level of detail is high, showing multiple edges in the objects on a fixed image. In consequence, pixels must be treated differently than if the opposite happens.

After detecting the details and edges of objects derived from high frequencies, the pixels are grouped to obtain the regions of interest where the observer is most probably watching. The results of highlighted areas in sequence “Crowd” are collected in Fig. 82.

Fig. 82. Example of high-detailed video sequence “Crowd” (left) with its Canny mask highlighting detailed areas in red (right)

5.3.1.2 Temporal entropy (motion)

Temporal entropy (also called) complexity is used in the algorithm to gauge metrics influence as it has a lot of influence on the perception of distortion. There are different methods of calculation, but one of the most common is the temporal information (TI) of the sequence. TI is obtained by comparing the values of luminance in pixels at the same position in consecutive frames, following the next equation:

$$TI(i, j, t) = Y(i, j, t) - Y(i, j, t - 1), \quad \forall (i, j) \in SROI$$

where Y represents the luminance image, and SROI the region of interest.

The graphics on Fig. 83 represent the TI evolution of the sequence. The vertical axis reveals the temporal complexity of the sequence.

Fig. 83. TI on different motion levels: sequence with fast motion (left) and low motion (right)

The TI and optical flow based on allows the calculation of motion in the image, as can be seen on example in Fig. 84. The objects in motion are captured by human eye previously to the rest of the objects in the picture. In that image motion areas catch the attention of the observer, who tends to observe the barrier and the car, also the tree branches are moving with the wind.

Fig. 84. Saliency diagram based in motion in sequence “Barrier” (frame 310)

5.3.1.3 Image pixels distribution

The image distribution is taken from studies by Nojiri, previously mentioned. In Nojiri et al. works [NOJIRI-1, 2006], the screen is divided in nine domains for analyzing 3DTV stereoscopic parallax distribution, assuming that attention would tend to be focused towards the center of the screen. For that reason, the central domain was also defined larger. The same distribution was exported to 2-D distribution, in order to obtain the domains in which visual fixation is most probable, as shown in Fig. 85.

Fig. 85. Segmentation in domains of a picture (left) and example in sequence “Boxers” (right).

5.3.1.4 Human faces content

When analyzing video quality on a video sequence, pixels in face regions must be treated separately because the visual quality is increased in areas where visual impact is higher. Observer tends to be more exigent in areas where visual attention is focused, concretely in human faces regions.

Fig. 86. Face detected in “Umbrella” sequence (left) and its facial mask (right)

5.4 Degradation of video sequences

The sensitivity of the human eye has been fully explored and investigated via qualitative experiments. Development of databases containing degraded video sequences is necessary for adapting quality systems to HVS is necessary for these purposes. For subjective tests, algorithms based on detecting four types of ROI's were developed: selected position of pixels, motion, spatial complexity and face detection, allowing the isolation of specific artifacts, for creating a psychovisual model adapted to human eye.

Most studies introduce impairments in the image by encoding the video sequence at different bitrates. This methodology works when assessing the quality of an encoder but the perceived impairment could be hardly distinguished from the impairment hidden by the effect of motion, position of the pixel or high frequencies. Coding introduces impairment in the image in the same way, but in order to analyze individually each of these three factors, it is necessary to generate video sequences that have previously been artificially impaired [LOPEZ-5, 2014].

To this effect, the methodology must be based on generating masks for each video sequence to cover the pixels where motion, spatial complexity position or faces are detected, and setting a comparison to the sequence with opposite situation (creating impairments only over the masks first and then creating impairments only where there is no mask). In the first case, the pixels included in the mask keep the original image and the rest of the image is covered by the impaired image with a sequence containing the artifact to be evaluated, e.g. blurring, ringing or blocking or, in general, by a sequence encoded at low bitrates, following the procedure in next figure (Fig. 87).

Fig. 87. Scheme of video sequence generation

Two examples of the usage of this methodology for artificially impairing video sequences are included next. In Fig. 88, the faces ROI's belong to a sharp and high quality image, but the rest of the picture belongs to the video sequence after a severe degradation progress when applying low H.264 bitrates to those areas.

Fig. 88. Artificially impaired video sequence "News Report" with high quality in faces detected and poor quality in other areas.

On the other hand, in Fig. 89, the low quality areas correspond to the regions where a face is detected, while the rest of the picture containing sharp images extracted from high-bitrate-encoded video sequence.

Fig. 89. Video sequence “News Report” with poor quality in faces and high quality in other areas, inverse of previous sequence.

Consequently, five types of binary masks are detecting the regions of interest to be quantified, using the methodology explained in previous example. These masks will be used in future subjective assessment [LOPEZ-6, 2015] (Fig. 90). The four kinds of masks are:

- **Motion mask.** This kind of mask and its inverse is based on temporary information algorithms and object detection which allow analyzing if motion affects to vision or if, on the other hand, it provokes hidden impairments. Optical flow and TI (temporal information) tools were developed to create these masks.
- **Spatial complexity mask.** Based on the detection of pixels in edges and high frequencies region, this mask is generated by the usage of Canny algorithm [CANNY-1, 1986].
- **Position masks.** The frame is divided into different sections to analyze the human vision, which is usually focused on central areas. Nojiri distribution [NOJIRI-1, 2006] was used for the pixel-regions divisions.
- **Face detection.** After studies of eye movement and fixation maps included in [WOODING-1, 2002], it is noticeable that visual impact is important when a human face is detected in the picture. Pixels included on an area where a human face is detected belong to the main regions of interest and should be analyzed as an important area for quality assessment. Haar [VIOLA-1, 2001] algorithm is used for face detection.
- **Color masks.** The content itself is primary for understanding human perception [Fig. Color]. Consequently, human eye is not equally sensitive to different colors, because of the effect of photoreceptors [WINKLER-3, 1999]. Color, and additionally bright and contrast, must be analyzed to determine the colors which make artifacts more visual for human eye. This mask is selected for future work, not this present research work.

Fig. 90. Diagram for feature mask sequence and its inverse mask

These four characteristics are analyzed in subjective assessment in order to demonstrate that human visual system is more sensitive to some concrete phenomena. Objective metrics should be adapted to perceptual quality, and pixels weighted to areas of interest, which are most probably visualized by human eye. For this purpose a software was developed for this doctoral thesis, whose intercafe can be seen in Annex 2.

5.4.1 Motion masks

For motion detection, temporal information in consecutive frames is scrutinized. The pixels with variation are analyzed in order to conclude subjectively how pixels with motion weight to different algorithms for quality assessment.

$$Pix_i(i, j) \in .Mask, \text{ if } [F_i(i, j) - F_{frame+1}(i, j)] > 0$$

The example shows how pixels in the motion region of interest are detected over the barrier, which is the only element within motion, and it is independent from other effects, such as edges and detailed information.

Fig. 91. "Barrier" sequence after applying motion mask

5.4.2 Spatial complexity and high frequencies masks

Textures, edges and objects in motion are the source of hiding impairments in cases of blocking algorithms but the opposite happens with blurring algorithms. Canny algorithm is used to create a binary mask, which separates the homogenous areas from the high-frequencies area. Canny is an improved method for edge detection which generates a binary mask with techniques based on Sobel matrixes, that is the reason why it was selected instead of Sobel gradient itself, and the quality of its performance in real time.

Fig. 92. Canny algorithm for spatial complexity mask

5.4.3 Position of pixels in spatial sections

The image is divided into nine sections (Fig. 94), as indicated in research by Nojiri et al [NOJIRI-1, 2006]. The idea of these sequences is to analyze the focus on different sequences, independently of their content. Then, it is possible to analyze if there is a high influence on the fact that a pixel is located on a corner, or on a lateral or central area (Fig. 93).

Fig. 93. Corner, center and lateral masks

SECTION 1 (Up Left)	SECTION 2 (Up Center)	SECTION 3 (Up Right)
SECTION 4 (Center Left)	SECTION 5 (Center)	SECTION 6 (Center Right)
SECTION 7 (Down Left)	SECTION 8 (Down Center)	SECTION 9 (Down Right)

Fig. 94. Distribution of sections to divide video frames

5.4.4 Face detection

Haar algorithm based on a boosted cascade of simple features [VIOLA-1, 2001] is used for face detection. Pixels belonging to a facial region of interest are important because are part of the visual impact derived from user experience.

When analyzing video quality on a video sequence, pixels in face regions must be treated separately because the visual quality is increased in areas where visual impact is higher. Observer tends to be more exigent in areas where visual attention is focused, concretely in human faces regions.

A software specifically used for this purpose was developed containing the necessary tools for analyzing video sequences in each of these five image features, using C# programming language environment. Libraries included in Emgu CV [EMGU-1, 2013] and Open CV platform [BRADSKI-1, 2008] were used for the process, importing Haar cascade techniques for facial detection included in “haarcascade_frontalface_alt2.xml” and “haarcascade_profileface.xml” files.

5.4.5 Range of colors detection

Through the optics of the eye, two different types of photoreceptors, rods and cones, exist. Cones are responsible of the photopic vision and the distinguish colors, with different normalized spectral sensitivities to spectrum wavelengths for the three cone types: L-cones, M-cones, and S-cones [WINKLER-3, 1999].

Some artifacts allow the visualization of artifacts more easily by human eye. For that reason, range of colors should be analyzed to determine the weight of this factor to visual algorithms. Three ranges of colors are developed for masks: red, blue and green. The mask contains pixels corresponding to the determined color and the ones with a similarity related to a threshold, as seen in Fig. 95.

Fig. 95. Range of red colors mask in sequence “Barrier” (frame 125)

5.5 Artifact mask generation based on the psychovisual model

Most studies introduce impairments in the image by encoding the video sequence at different bitrates. This methodology works when assessing the quality of an encoder, but the perceived impairment could be hardly distinguished from the impairment hidden by the effect

of motion, position of the pixel or high frequencies. Coding introduces impairment in the image in the same way, but in order to analyze individually each of these factors, it is necessary to generate video sequences that have previously been artificially impaired.

To this effect, the methodology must be based on generating masks for each video sequence to cover the pixels where motion is detected, and setting a comparison to the sequence with opposite situation. Just the same occurs in the other two cases. The pixels included in the mask showed the original image and the rest of the image was covered by the impaired image with a sequence containing the artifact to be evaluated, for example, blurring, ringing or blocking, or in general by a sequence encoded at low bitrates, following the procedure in.

An example of how each sequence is created is collected in Fig. 96. The sequence combines two different sequences, one with high quality and the other with poor quality, which contains the impairment. Each of the sequences is located in areas where the image feature to analyze is defined.

Fig. 96. Example of Sequence “Umbrella” with impairment located in faces ROI and its inverse

The first step is defining a collection of distortions to analyze in video quality assessment for simulating each kind of artifact. From the list of artifacts introduced previously, these ones are selected for distortion simulated. The result of the usage of these techniques is the “Impaired video sequence” which appeared in scheme in Fig. 97.

- Gaussian low-pass filters for blurring simulation.
- JPEG2000 encoder for ringing simulation.
- An 8x8 mosaic filter for MPEG-2 blocking artifact simulation.
- Other encoders at low bitrates for general distortion simulations.

These techniques connect the block diagram of defining an impaired video sequence, included in Fig. 97, with the complete scheme for generating artificially impaired sequence with distortion located in a ROI or its corresponding inverse ROI.

Fig. 97. Block diagram for impairment insertion process

Examples of distortions are collected and explained in Fig. 98. Distortion is applied to the whole sequence equally. This sequence will be discriminated by the mask to locate distortion in only a part of pixels of the whole frame.

Fig. 98. Detail of frame 125 of sequence “Umbrella” after applying different distortions: original, blurring (Gaussian filter), ringing (JPEG200 encoding), blocking (8x8 mosaic filter).

For defining the coefficients of influence for each video characteristic, three sequences containing different content and artificially impaired were mainly used. In Table 16 the characteristics of this sequences are collected for each of the degradation, when applied the distorted image only to one on the six regions of interest for each phenomena (details or spatial entropy; motion or temporal entropy; center, lateral or corner position and faces presence), conserving the rest of the picture sharped. On the other hand, for each of these six phenomena the inverse version was also developed, i.e. sharp pixels where the phenomena is located and the rest of the picture distorted. Subjective evaluation for each of these sequences are named MOS (Mean Opinion Score). PSNR or MSE indicates the results of this full-reference metrics, and the conventional blurring metric is collected. Finally, the results for the sharp image (coded at 75Mbps in H.264) and the distorted sequence (coded at 500Kbps in H.264) are

collected in the first column to describe the sequences used in the artificially impairment introduction process.

Table 16. Results of different metrics for representative frames applied to different sequences in cases of artificially impaired mask. D means that distortion is located in that area, and "Inv" indicates that distortion is located in the inverse of that area

Seq.	FR Metric	H.264		Details		Motion		Center		Lateral		Corner		Faces	
		75Mbps	500Kbps	D.	Inv.	D.	Inv.	D.	Inv.	D.	Inv.	D.	Inv.	D.	Inv.
Barrier	PSNR	49.82	33.19	37.81	35.59	39.85	34.24	38.29	34.79	45.26	33.46	47.04	33.37	-	-
	Blur	0.27	8.36	3.05	4.37	1.97	6.24	2.34	5.82	0.44	7.86	0.37	7.94	-	-
	MSE	0.51	3.34	1.11	2.23	0.359	2.98	0.927	2.41	0.34	3.00	0.20	3.13	-	-
	MOS	4.77	1.33	2.89	3.22	3.11	3.89	2.00	3.89	3.56	2.11	4.22	1.44	-	-
Crowd	PSNR	34.33	25.34	34.15	28.35	26.30	32.37	30.74	26.82	33.87	26.00	35.95	25.88	46.57	25.38
	Blur	3.44	22.55	5.67	10.51	18.11	2.87	6.27	15.33	2.60	19.44	0.95	22.47	0.16	22.33
	MSE	3.55	8.76	7.32	4.99	6.13	2.38	2.30	6.21	1.21	7.30	0.64	7.87	0.06	8.45
	MOS	4.68	1.22	2.56	3.00	2.56	1.89	1.44	2.44	3.78	1.33	4.11	1.22	2.89	4.22
News Report	PSNR	47.93	37.58	47.85	37.52	45.97	35.82	41.48	39.86	45.35	38.38	46.52	37.25	46.82	34.52
	Blur	0.44	3.63	0.88	2.69	0.68	4.85	1.52	2.06	0.51	3.08	0.37	4.12	0.38	5.17
	MSE	0.67	1.93	0.95	1.45	0.22	2.19	0.62	1.31	0.30	1.62	0.21	2.17	0.10	2.30
	MOS	4.81	1.54	3.22	3.22	2.00	3.44	1.67	3.89	3.78	1.44	4.56	1.44	1.33	3.78

5.6 Description of Subjective tests for psychovisual model

5.6.1 Methodology description

Methodologies for the assessment of picture quality were developed following procedures included in the Recommendation ITU-T P.910 [REC P.910]. The method used was single-stimulus (SS), since the objective was to evaluate the quality perceived by the observer, while reducing the duration of each individual session. The test sequences were presented only once in the test session, while in the beginning, dummy sequences were presented as a reference for the user, just to make the procedure understandable. First, sequences with the highest quality were presented to the observer and rated as 5 (excellent). Then, sequences with the poorest quality were visualized and rated as 1 (bad). Observers taking part in the tests were asked to evaluate the video quality in the rest of the sequences, 30 sequences. (Table 17)

Table 17. Levels of the scale used in tests

MOS Scale	Perceived Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The observers were placed at 2 meters from the screen. It is commonly known that the recommended optimum distance for HDTV systems varies depending on the desired angle of

vision. ITU-T P.910 specifies a wide range of distances ranging from 1 to 8 times the height of the screen. A typical recommendation of 1.6 times the screen diagonal offers a 30° angle. Others prefer an even higher angle of 40° to increase immersive experience. This means a distance of 1.2 times the screen diagonal. In developed tests, the viewers were standing at 1.7 meters from the screen, 1.5 times the screen's diagonal (46"). (Table 18)

Table 18. Basic characteristics and setting of devices

Device	JVC GD-463D10 3D
Resolution	1920 x 1080p
Brightness	450 cd/m ²
Size	46" Widescreen
Technology	LED HD

Ten observers took part in this testing and fulfilled the complete set of thirty sequences. According to ITU-T P.910, between four and forty observers are recommended for this type of tests. Explanations about the test were previously provided to the observers, using video sequences as examples. Most participants were male, not only scholars but also faculty staff, professional and non-professional observers, i.e. people not familiar with quality assessment. A vast majority (70%) were considered non-professional observers. None of the observers was considered as outlier.

5.6.2 Video Database and Session Description

The video sequences used for the work development were mainly based on the stereoscopic database provided by the University of Nantes [URVOY-1, 2012]. Furthermore, the selected sequences containing faces were important for the database used for this work. A session of thirty video sequences was presented to the observers in the subjective assessment tests.

As mentioned, the method used for the tests was single-stimulus (SS). The observers, in groups of four or five people, are presented each video of duration between 10 and 30 seconds only once.

Two databases were used as sources of the artificially impaired video sequences collection, because of their variety and combinations of TI (temporal information) and SI (spatial information): SVT [HAGLUND-1, 2006] and NAMA3DS1-COSPAD1 [URVOY-1, 2012]. For each sequence, 24 different combinations of masks were developed. One motion mask, one spatial complexity mask, nine position masks, one for each of Nojiri's sections [NOJIRI-1, 2006], one based on face detection, and also all their corresponding inverses. The spatio-temporal diagram for the sequences to assure the spectrum of complexities is collected next, as explain in [NOJIRI-1, 2006].

For this work, sequences used correspond to format 1080p25, with these settings collected in Table 19. Coding of video sequences was done in MPEG-2 (High profile, high level) and H.264 (High profile 5.1), at bitrates between 500Kbps and 100Mbps.

Table 19. Basic characteristics and settings of video sequences used

Settings	Values
Resolution	1920 x 1080
Frames per second	25 fps
Interlaced Mode	Progressive
Aspect Ratio	16:9
Pixel Aspect Ratio	1.0 (square pixels)
Coding	MPEG-2 (Bitrates: 500kbps-100Mbps, High profile, high level) H.264 (Bitrates: 500Kbps-75Mbps, profile 5.1)

5.6.3 Subjective assessment for pixels weight definition

After introducing the video sequences used for the study, results based on subjective tests are included and analyzed in this section, in Table 20.

Table 20. Results from subjective tests with artificially impaired sequences and impairments/distortions located only in determined regions of interest corresponding to factors analyzed

Where is the Impairment located in Video Sequences	Sequence	MOS	Variance	CI (Confident Interval)
Impairment in motion areas	Crowd	2,56	0,28	0,18
	News Report	2,00	0,50	0,33
	Barrier	3,11	0,36	0,24
No impairment in motion areas (inverse of the previous one)	Crowd	1,89	0,36	0,24
	News Report	3,44	0,28	0,18
	Barrier	3,89	0,61	0,40
Impairment in details	Crowd	2,56	0,78	0,51
	News Report	3,22	1,44	0,94
	Barrier	2,89	0,61	0,40
No impairment in details (inverse of the previous one)	Crowd	3,00	0,50	0,33
	News Report	3,22	0,69	0,45
	Barrier	3,22	1,94	1,27
Impairment in faces	Crowd	2,89	0,86	0,56
	News Report	1,33	0,25	0,16
	Barrier	N.A.	N.A.	N.A.
No impairment in faces (inverse of the previous one)	Crowd	4,22	0,44	0,29
	News Report	3,78	0,94	0,62
	Barrier	N.A.	N.A.	N.A.
Impairment in selected corner	Crowd	4,11	0,36	0,24
	News Report	4,56	0,28	0,18
	Barrier	4,22	0,44	0,29
No impairment in selected corner (inverse of the previous one)	Crowd	1,22	0,19	0,13
	News Report	1,44	0,28	0,18
	Barrier	1,56	0,28	0,18
Impairment in one lateral	Crowd	3,78	0,19	0,13
	News Report	3,78	0,44	0,29
	Barrier	3,56	0,78	0,51
No impairment in that lateral (inverse of the previous one)	Crowd	1,33	0,25	0,16
	News Report	1,44	0,28	0,18
	Barrier	2,11	0,36	0,24
Impairment in center	Crowd	1,44	0,28	0,18
	News Report	1,67	0,50	0,33
	Barrier	2,00	1,25	0,82
No impairment in center (inverse of the previous one)	Crowd	2,44	0,53	0,34
	News Report	4,33	0,25	0,16
	Barrier	3,89	0,61	0,40

For these three examples, the combination of sequences includes a video sequence encoded at 75Mbps in H.264, which is a high quality, and poor quality derives from a sequence encoded at 500Kbps. All sequences are in 1920x1080 progressive formats, at 25fps. “D” indicates that the distortion is located in a determined ROI and “Inv.” indicates the inverse sequence of previous case of study.

The first example refers to sequence “News Report” which represents two talking heads from a news channel. When distortion is located in the areas corresponding to human faces, the subjective MOS values are lower (1.33) compare to those cases when located in the rest of the picture, and faces appear sharp (3.78). This effect is completely opposite to PSNR (46.82 vs. 34.52) or MSE’s behavior (0.10 vs. 2.30) as seen in Table 21.

Table 21. Results for impairment located in faces ROI for artificially distorted sequence “News Report”

Sequence	FR Metric	H.264		Impairment located in Faces ROI.	
		75Mbps	500Kbps	D.	Inv.
News Report	PSNR	47.93	37.58	46.82	34.52
	Blur	0.44	3.63	0.38	5.17
	MSE	0.67	1.93	0.10	2.30
	MOS	4.81	1.54	1.33	3.78

A similar situation occurs when analyzing motion in “Barrier sequence”, as seen in values collected in Table 22.

Table 22. Results for impairment located in motion ROI’s for artificially distorted sequence “Barrier”

Sequence	FR Metric	H.264		Impairment located in Motion ROI.	
		75Mbps	500Kbps	D.	Inv.
Barrier	PSNR	49.82	33.19	39.85	34.24
	Blur	0.27	8.36	1.97	6.24
	MSE	0.51	3.34	0.359	2.98
	MOS	4.77	1.33	3.11	3.89

Finally, Table 23 contains the results when comparing distortions located in a corner, a lateral or the center area in sequence “Crowd”. For the observers, a high distortion located in a corner is insignificant. On the other hand, when impairment is located in central area, opinion scores decrease to 1.44. The PSNR and MSE reveals the distortion related to the size of the impaired area, while the influence in the human eye is related to the position of that impaired area.

Table 23. Results for impairment located in position ROI’s for artificially distorted sequence “Crowd”

Seq.	FR Metric	H.264		Impairment located in Position ROI’s					
		75 Mbps	500 Kbps	Center		Lateral		Corner	
				D.	Inv.	D.	Inv.	D.	Inv.
Crowd	PSNR	34.33	25.34	30.74	26.82	33.87	26.00	35.95	25.88
	Blur	3.44	22.55	6.27	15.33	2.60	19.44	0.95	22.47
	MSE	3.55	8.76	2.30	6.21	1.21	7.30	0.64	7.87
	MOS	4.68	1.22	1.44	2.44	3.78	1.33	4.11	1.22

5.6.4 Masks for video degradation

Most studies introduce impairments in the image by encoding the video sequence at different bitrates. This methodology works when assessing the quality of an encoder but the perceived impairment could be hardly distinguished from the impairment hidden by the effect of motion, position of the pixel or high frequencies. Coding introduces impairment in the image in the same way, but in order to analyze individually each of these three factors, it is necessary to generate video sequences that have previously been artificially impaired.

To this effect, the methodology must be based on generating masks for each video sequence to cover the pixels where motion is detected, and setting a comparison to the sequence with opposite situation. Just the same occurs in the other two cases. The pixels included in the mask showed the original image and the rest of the image was covered by the impaired image with a sequence containing the artifact to be evaluated, for example, blurring, ringing or blocking, or in general by a sequence encoded at low bitrates, following the procedure in Fig. 90.

5.7 Definition of the psychovisual model through four-components analysis

5.7.1 Psychovisual model definition

For each of the phenomena, a mask for each frame is implemented. The level of importance of the mask is indicated in black and white images. Black pixels are not taking into account, while white pixels are primordial for the final algorithm. Motion, detail and facial mask are binary, as seen in Fig. 99 to Fig. 101. On the other hand, position mask is constant for all the frames of the whole sequence, but it offers a shade of grey to determine the level of importance of the lateral areas, as seen in Fig. 102. The four masks are previously combined for the final psychovisual model in order of influence to the algorithm. Each frame of the sequence possesses a different and unique psychovisual model mask.

Fig. 99. Motion Mask in sequence News Report (frame 0)

Fig. 100. Detail Mask in sequence News Report (frame 0)

Fig. 101. Faces Mask in sequence News Report (frame 0)

Fig. 102. Pixel Position Mask in sequence News Report (frame 0)

After analyzing the subjective results, coefficients were assigned to the four different masks related to each of the image characteristic selected. The values were defined ordering the level of importance derived from observers' opinions. The most important effect highlighted in video sequences is face detection. Sequences with a lower PSNR obtained better scores than sequences with higher PSNR when low quality was concentrated in pixels contained in faces ROI's. Sequence "News Report" obtained a score of 1.33 while when distortion was introduced in the rest of the image leaving the faces without distortion the score was 3.78 (table 4). Considering the relative space occupied by the faces the other objective results are as expected. PSNR, Blur and MSE show a huge drop in quality. This contradiction validates the hypothesis of the relevance of the faces.

The same conclusions can be extracted from the region analysis. "Barrier" sequence obtains a score of 2.00 when distorting the center of the frame while it gets a score of 3.89 when distorting the rest of the frame.

Again, the center occupies just a third of the total frame, so PSNR, Blur and MSE indicate that the image is more damaged when distorting the outer zone of the frame. This does not occur when the distortion is applied to the laterals and, much less, when applied to the corners. Thus, the importance of the center is higher than the laterals, which at their time are more important than the corners.

Following the same pattern it has been decided to consider motion distortion to affect more the result than detail distortion, both considered to be more important than lateral or corner distortions.

At the end of the analysis, it has been considered the following chain of influence: Faces > Central > Motion > Detail > Lateral > Corner

Again, based on the results obtained from the quality tests, a weight for the final distortion value of each pixel has been assigned to each element of distortion. As a start point, the greatest influence (faces) has been assigned with a weight value of one, and the following elements have been assigned with decreasing values. These values have been normalized to obtain the final distortion coefficients.

The coefficients are justified with the results obtained from subjective tests, whose complete results table are collected in Annex 1. The focus is on the cases where contradictions exist between the MOS scores and classical metrics, such as PSNR or MSE, and psychovisual models are especially developed for introducing visual attention factors to correct the values of distortion metrics.

Fig. 103. Percentage of influence of each pixel characteristic to the psychovisual model

The distribution of importance of each factor related to subjective tests results are collected in Fig. 103, and the influence to obtain the final algorithm and mask is collected in Table 24.

Table 24. Coefficients of influence to the psychovisual model depending on the ROI belonging

Characteristics		Pixels ROI Belonging	Partial coefficient	Normalized coefficient	Percentage of influence
Faces		Pixels $\in$ (ROI Faces)	1	0,29	28,6%
Position	Center	Pixels $\in$ (ROI Center)	0,9	0,26	25,7%
	Lateral	Pixels $\in$ (ROI Lateral)	0,4	0,11	11,4%
	Corner	Pixels $\in$ (ROI Corner)	0,1	0,03	2,9%
Motion		Pixels $\in$ (ROI Motion)	0,6	0,17	17,1%
Detail		Pixels $\in$ (ROI Detail)	0,5	0,14	14,3%
Total			3,5	1,00	100,0%

Figures 103 to 106 include the computed weighted pixels of this four types of masks.

Fig. 104. Motion Mask in sequence News Report (frame 0)

Fig. 105. Detail Mask in sequence News Report (frame 0)

Fig. 106. Faces Mask in sequence News Report (frame 0)

Fig. 107. Pixel Position Mask in sequence News Report (frame 0)

The resultant model of visual attention as a combination of the developed masks processed for each effect is included in Fig. 108.

Fig. 108 Complete psychovisual model for first frame from sequence "News Report" (frame 0).

5.8 Use of visual attention model in objective quality metrics

5.8.1 Advanced Blurring FR objective metric

As explained in previous chapters, blurring is an artifact that appears in a video sequence by the energy change of each individual encoded frame, which results from the quantization process [MA-1, 2012]. Due to the reduction of high frequency in horizontal and vertical direction this distortion can occur, generating mosaic patterns [BOEV-1, 2008].

Typically, the bandwidth limitation is the main cause of high-frequency detail and sharpness loss, homogenization, and blurring artifacts. The artifact of blurring appears on a video sequence when decreasing the encoding bitrate, especially in transform-based coding. Because of this decrease, blurring implies a loss of energy in edges of the objects. The degradation has an impact in highly detailed areas, i.e. the high frequency components, which are located in the edges of the objects.

The next formula allows calculating the blurring metric result when comparing the decrease of energy of the luminance function between the distorted and the reference sequences. The result is normalized to the size of the frame.

$$Blur = \frac{1}{W \cdot H} \sum_{j=0}^{W-1} \sum_{i=0}^{H-1} |GE_{ref}(f(i, j)) - GE_{cod}(f(i, j))| \quad (3)$$

where GE means gradient entropy, $f(i, j)$ is the function to be treated, in this case luminance of the pixel with coordinates (i, j) , ref is the video sequence and cod is the distorted video sequence. W, H are the width and height of the frame.

The advanced blurring metric corresponds to the appliance of the coefficient related to every pixel of the psychovisual model when comparing the energy of the gradient modulus between distorted and original video sequences. The formula appears next indicated as follows:

$$\overline{Blur} = \sum_{j=0}^{W-1} \sum_{i=0}^{H-1} psy(i, j) \cdot |GE_{ref}(f(i, j)) - GE_{cod}(f(i, j))|$$

where $psy(i, j)$ represents the normalized coefficient of the psychovisual model for pixel with coordinates (i, j) , GE means gradient entropy, and $f(i, j)$ is the luminance function.

The coefficients obtained after subjective results (and collected in previous section) are normalized to the size of the frame and to the percentage of influence of each characteristic of the image, as shown in formula (5).

$$psy(i, j) = \frac{coef_{MOT}(i, j) + coef_{DET}(i, j) + coef_{POS}(i, j)}{W \cdot H \cdot \sum_{c=0}^3 coef_{MAX}(c)}$$

where $psy(i, j)$ represents the normalized coefficient of the psychovisual model for pixel with coordinates (i, j) , $coef_{MOT}(i, j)$ is coefficient related to motion, $coef_{DET}(i, j)$ is the coefficient related to detail or spatial entropy, $coef_{POS}(i, j)$ luminance image, and

$coef_{MAX}(c)$ is the maximum value of the coefficient for each characteristic, to normalized the result. W, H are the width and height of the frame.

5.8.2 Other FR objective metric

The experiment developed for the concrete artifact of blurring has been repeated for other quality metrics of global assessment of quality images. In this case, the visual attention model was applied to MSE (Mean Squared Error), which is defined in Annex 1. The results of the application of this “Advanced MSE” metric are included in next section.

$$\overline{MSE(l)} = \frac{1}{WH} \sum_{j=0}^{H-1} \sum_{i=0}^{W-1} psy(i, j) \cdot \left(f^l_{DIST}(i, j) - f^l_{REF}(i, j) \right)^2$$

where:

- ◆ $psy(i, j)$: Value of weighting of the pixel (i, j) from the psychovisual or visual attention model. The coefficient for each pixel is normalized to a value between zero and one.
- ◆ $f^l_{REF}(i, j)$: Luminance of the frame l in reference video sequence corresponding to pixel (i, j) .
- ◆ $f^l_{DIST}(i, j)$: Luminance of the frame l in distorted video sequence corresponding to pixel (i, j) .
- ◆ l : Number of the analyzed frames.
- ◆ W : Width of the video frame, in number of columns.
- ◆ H : Height of the video frame, in number of rows.

5.9 Results

For validation of the results derived from the studies of psychovisual modelling applied to blurring full-reference metric, different sequences than those studied for calibration of the metric were used. The first results offer a general vision that could be improved in the future with more sequences and a finer calibration to the psychovisual model. This fact could allow the application of this psychovisual model to other metrics, i.e. blocking or blurring.

Examples of the developed successful content-weighted psychovisual model are shown in Fig. 109 to Fig. 112. The most important part of the picture is the central area, and the zones where human faces are located as expected. Motion is a key factor in sequences such as “Boxers”, and detail or spatial is distributed through the picture as happens in “Tree Branches”, but the visual attention is specially focused on central areas where branches are located.

Fig. 109. Content-weighted mask (right) corresponding to original frame (left) in sequence “Boxers” (frame 90)

Fig. 110. Content-weighted mask (right) corresponding to original frame (left) in sequence “Boxers” (frame 275)

Fig. 111. Content-weighted mask (right) corresponding to original frame (left) in sequence “Phone Call” (frame 131)

Fig. 112. Content-weighted mask (right) corresponding to original frame (left) in sequence “Tree Branches” (frame 16)

It is especially interesting to study the effect of weighted blur when comparing results between “Boxers” and “Phone Call”. Table 25 shows how these two sequences have similar

results in PSNR and MSE when the first one is coded at 1Mbps and the second one at 500Kbps. These results indicate similar level of overall distortion but, in fact, “Phone Call” at 500 Kbps is perceived with lower quality than “Boxers” at 1Mbps. Weighted blurring does indicate a worst quality. “Phone Call” has most of its relevant information and a face in the center of the frame, so compression will affect more this central zone, although global distortion is the same as when “Boxers” is less compressed. Weighted blur takes into account the central distortion and the face distortion to achieve a more accurate quality measure in this case. The reverse happens when both sequences are coded at 1Mbps. PSNR and MSE indicate a greater quality in “Phone Call” while weighted blur shows a similar, more accurate, value of quality (2.00 and 1.99 respectively).

“Tree branches” has also a good response to weighted blur. In this case, the image has most of its detail and motion in the perimeter. There are no faces to take into account this time. Distortion is severe in this sequence, but PSNR falls and blur and MSE grow to quickly compare to the observed quality reduction. This time weighted blur grows slowly and is more related to the quality perceived. A decrease between 6Mbps and 4Mbps is not perceived as a big quality loss, mainly because the distortion does not affect the center of the frame as hard as it affects the perimeter. Weighted blur does consider this condition to match better the subjective results.

Table 25. Results for advanced blur metric applied to sequences in H.264 at different bitrates. In addition, Advanced MSE results are collected, these results correspond to applying the psychovisual model to MSE metric.

Sequences	FR Metric	H.264			
		6Mbps	4Mbps	1Mbps	500Kbps
Boxers	PSNR	43.60	42.73	39.29	35.30
	MOS	4.77	4.11	2.44	1.33
	Blur	0.65	0.92	3.04	6.88
	Adv. Blur	1.34	1.48	2.00	2.66
	MSE	1.13	1.25	1.79	2.68
	Adv. MSE	1.15	1.28	1.87	2.83
Hall	PSNR	39.97	36.56	30.98	27.72
	MOS	4.88	4.11	2.44	1.33
	Blur	0.79	3.28	14.18	27.23
	Adv. Blur	2.24	3.49	6.88	9.67
	MSE	1.84	2.67	4.85	6.84
	Adv. MSE	1.92	2.88	5.36	7.72
Phone Call	PSNR	42.07	41.49	40.49	39.15
	MOS	4.88	4.00	2.44	1.33
	Blur	1.95	2.26	3.46	4.49
	Adv. Blur	1.64	1.78	1.99	2.17
	MSE	1.37	1.44	1.58	1.79
	Adv. MSE	1.41	1.50	1.66	1.89
Tree Branches	PSNR	29.16	26.00	19.19	18.11
	MOS	4.88	3.77	2.55	1.55
	Blur	11.92	17.36	22.38	20.12
	Adv. Blur	6.15	8.03	9.79	12.09
	MSE	5.74	7.76	14.11	16.02
	Adv. MSE	6.24	8.53	16.27	18.31

The results of Blur and the proposed Advanced Blur metric are compared next with the Pearson Correlation. The Pearson linear Correlation Coefficient (PCC) is used to describe the

accuracy of the estimation between the metric and the MOS (Mean Opinion Score), which is the subjective result. Comparing both metrics the improvement of the psychovisual model is obtained in the parameter $\Delta(\text{Adv.Blur-Blur})$. The improvement is positive for the collection of sequences and it is over 13% for “Tree Branches” sequence, because of the distribution of quality on the image.

“Tree Branches” presents a uniform distribution of high frequencies affected by blur distortion. The conventional blur metric treats equally the pixels of the image, even in the corner areas, while the Advanced Blur metric weights the pixels. This process orders the areas that are more probably viewed by any observer, considering medium-level features that are not considered in the classic metric. This medium-level correspond to the Gaussian distribution of viewpoints of the observer from the center to the margins of the frame.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The result is negative because the variables have contrary evolution, they are inverse. The proximity to 1 (or -1 in this case) are interpreted as a more linear function.

Table 26. Pearson linear Correlation Coefficient for Blur and Advanced Blur metrics

Sequence	Value	6Mbps	4Mbps	1Mbps	500Kbps	PCC	$\Delta(\text{Adv.Blur-Blur})$
Boxers	Blur	0,650	0,920	3,040	6,880	-0,953	2,97%
	Adv Blur	1,340	1,480	2,000	2,660	-0,983	
	MOS	4,778	4,111	2,444	1,333		
Hall	Blur	0,790	3,280	14,180	27,230	-0,982	1,40%
	Adv Blur	2,440	3,490	6,880	9,670	-0,996	
	MOS	4,889	4,111	2,667	1,556		
Phone Call	Blur	1,950	2,260	3,460	4,490	-0,990	0,94%
	Adv Blur	1,640	1,780	1,990	2,170	-0,999	
	MOS	4,889	4,000	2,444	1,333		
Tree Branches	Blur	11,920	17,360	22,380	20,120	-0,863	13,24%
	Adv Blur	6,150	8,030	9,790	12,090	-0,996	
	MOS	4,889	3,778	2,556	1,550		

From Fig. 113 to Fig. 116, the linearity of the Pearson for the four sequences is included, demonstrating the validity of the advanced metric.

Fig. 113. PCC for Blur and Advanced Blur for sequence "Boxers"

Fig. 114. PCC for Blur and Advanced Blur for sequence "Hall"

Fig. 115. PCC for Blur and Advanced Blur for sequence "Phone Call"

Fig. 116 PCC for Blur and Advanced Blur for sequence "Tree Branches"

5.10 Conclusions to visual attention models

Most of the previously published video quality metrics are based on objective parameters, such as PSNR, MSE or SNR, which compares the original image to the degrading one, establishing a relationship between them. This metrics are often not correlated to the human visual system because they are not adapted to what the human eye most probably sees.

The visual attention systems and saliency diagram, based on the factors analyzed in the state of the art, improves the response of metrics, in the case of blurring. Previous studies using eye-tracking with static images highlights the importance for the human eye of details and the area in which high frequencies, especially when the size of screen is taken into account.

The state of the art for visual attention is extensively based on complex processes, such as object recognition or scene interpretation. Despite visual attention is a main feature of the HVS, only few studies have been done on using it applied to image quality assessment, and even less studies have applied this research field to video quality assessment. Studies such as the one developed by You et al. [YOU-1, 2009] use saliency models and semantic image analysis by human faces detection, improving the results obtained with PSNR or SSIM metrics but miss the motion factor, reducing the results to static images. Following this path, something similar happens to the model developed by Ma et al. [MA-1, 2010], which improves the results obtained by models that ignore the role of human visual attention.

An example of the use of the visual attention information was developed by Nisanni et al [NINASSI-1, 2007], concluding that applying the visual attention to image quality assessment is not trivial and, as HVS feature, influences to visual perception and quality of experience. This work used eye-tracking technics for this analysis, generating visual attention models based on subjective assessment. For that reason, a limitation in this work exists because the model is not adapted to any environment and depends on subjective results. For the work of this doctoral thesis, we propose a method for automatically developing visual attention models focused on quality itself and concrete artifacts. The novelty resides in the application of saliency maps to summarize videos for video quality assessment, adapted to each individual frame of the sequence.

Another important novelty in this doctoral thesis is the creation of a methodology to adapt the visual attention models to concrete artifacts, such as blurring or blocking. Isolating these features is the only way of understand them individually and quantify them in an environment where some factors mask them to human eye. Pixel-based metrics could detect a distortion that is not visible to human eye, which is more probably concentrated in the other side of the image. Our visual attention model filters the objective quality approaching the response of HVS.

Characterization of video sequences to particular artifacts and visual attention models is a key factor when developing a useful database for assessing quality.

The developed video database and used methodologies allow the research in the field of video quality assessment and the definition of psychovisual models adapted to the HSV. The

database, with sequences artificially impaired, performed successfully in subjective tests in order to decide the human eye's response to impairments. Due to the difficulty of isolating impairment and characterize the video sequence, it is impossible to completely distinguish between the effects of two different artifacts. As demonstrated, weighting pixels for improving a quality algorithm allow the approximation of results to the HVS. The improvement of video quality metrics by the use of visual attention models with a model that weights the pixels through different individual saliency maps and fuses them into a more complex attention maps is a novelty in this doctoral thesis. The influence of these regions of interest through masks, for each of the considered effects (motion, spatial complexity, face detection or position of the pixel) helps predicting human eye's final response, which is the one that eventually decides the QoE.

The blurring metric, with the correction derived by the spatial and temporal entropy quadrants, obtain better results than without taking into account the image characteristics. This means than human eye is more sensitive to artifacts in the central areas of the picture. In addition, motion and level of detail detected in the video frame are an important factor to generate an attention model, which helps improving the result of a metric as blurring.

This doctoral thesis demonstrates that visual attention influences perceived quality, and consequently the quality of experience. The use of visual attention models is necessary for analyzing video quality assessment. However, it is important creating models that adapts automatically to the content of each video frame for every sequence, which is the limitation of the use of eye-tracking techniques that the methodology proposed does cover

The calibration of the visual attention model obtained with sequences "News Report", "Crowd" and "Barrier" is exportable to other video sequences. The results after applying the model to MSE or Blurring metrics (Advance Blurring and Advance MSE) modify the value of these parameters. Sequences "Tree-branches", "Phone-call" or "Boxers" present an improvement from their respective conventional metrics (Blurring) as demonstrated with the Pearson lineal Correlation Coefficient, which offers an increase up to 13% for the Advance Blur metric, because the high-level features in the model modify the results of metrics where only low-level features are taken into account.

Future work could improve the results of blurring and blocking metrics enlarging the visual model to other metrics and using more low-level features, such as contrast or color, and high-level features determined by content-aware characteristics.

6. Conclusions

6 Conclusions

The audiovisual production chain presents different phases to produce video contents that satisfy user's requirements. Estimation of quality for visual contents is necessary for new technologies and its assessment is still a complicated task. A collection of problems affect to different phases of the chain. This doctoral thesis focuses on two phases and concrete cases to assure the quality of experience to the user: generation of contents in 3D stereoscopic video and video encoding for 2D conventional video.

Among the problems detected from the generation of contents in 3D stereoscopic video, the fact that 3D contents result more immersive than conventional 2D media implies a collection of circumstances that increase the probability of visual discomfort, and consequently visual fatigue, to occur.

The first hypothesis in 3D research field points to the analysis of disparity and distribution of parallax on a motion environment, apart from the artifacts and common degradations that the encoding process generates. Preliminary subjective tests determined that abrupt changes in disparity and parallaxes is one of the main sources of visual discomfort. This point was the basis for designing new tests with characterized sequences, in order to analyze the cases that provoke visual discomfort.

Pairs of sequences with different combinations of parallax and temporal information revealed that motion and high negative parallaxes combined increase the probability of discomfort to occur in human eye. Four cases of changes between these pairs of sequences were identified. Tests revealed that accommodation and vergence mismatches provoke discomfort. Moreover, this conflict happens because the HVS has not enough time to focus, because of the rapid changes of crystalline.

Disparities out of the Zone of Comfort, especially negative parallaxes and hyperstereoscopy, increase the probability of affecting to quality of experience. On the other hand, positive parallax itself does not affect to visual discomfort. The results highlighted that abrupt changes of disparities, especially related to negative parallax environment and derived from accommodation-vergence mismatches, which increase the time for human crystalline to focus the virtual objects.

The novelty of this doctoral thesis resides in the proposition of techniques to estimate the objective quality of stereoscopic 3D video content, which have as inputs the motion and depth map features of 3D video. The results conclude that information provided by depth maps and estimated motion vectors is useful to model the visual discomfort and fatigue experienced by observers when contemplating 3D stereoscopic contents. The addition of temporal information and motion is also an important novelty. Other models offer good results in 3D image, such as the ones proposed by Jung et al. [JUNG-3, 2015], but ignore the importance of

visual discomfort as a result of motion in a video sequence. Analyzing the visual discomfort in static images is a limitation of the state-of-the-art that the work of this thesis exceeds.

For future work, tools for automatization of process should be developed, which indicate to the filmmaker the cases when the probability of visual discomfort to happen is higher. The inclusion of objective quality metrics would improve the performance of 3DTV contents, in order to develop better guidelines for the user and the creator.

Literature analyzed through studies that used eye-tracking with static images revealed that observers focus on determined regions of an image. Central areas of an image, textures, human faces and, in video sequences, motion, are factors that induce the observer to focus the view over selected parts of the frame.

The problem of the vast majority of metrics, which do not correspond to the human visual system's response, is solved with the development of visual attention metrics. Subjective assessment allowed weighting different characteristics of video sequences, such as motion, level of detail, and position of pixels or human faces detection. The development of individual masks for each factor allows the creation of a visual attention model that weight the pixels, giving importance to the regions of interest that the user most probably will observe.

This process of weighing pixels approximates the quality results to human eye's response, as demonstrated with quality results, and it represents important novelty in the research of video quality assessment involving visual attention models.

Results obtained with the inclusion of visual attention masks to objective assessment, demonstrated slight improvement, close to the HVS's response. The process of weighting pixels and the use of this models in quality assessment is an innovation that improves the results from the state-of-the-art, and open an important research field for quality assessment. The linearity of Pearson linear Correlation Coefficient demonstrate the validity of the advanced metric, with positive results up to 13% of increase for some sequence, because of the use of high-level parameters that are not taken into account in conventional pixel-based metrics.

For future work, it is necessary to analyze the effect of domain analysis, i.e. the spatial and temporal entropy to the results of other metrics. Other metrics should be considered, such as the ones that assess quality for measuring blocking or ringing artifacts. Additionally, other low-level factors, such as color and luminance will be included for future versions of the visual attention model, and other high-level features such as concrete object detection. Furthermore, the adaptation of the model to the size screen could improve the result in determined environments. The improvement of the visual attention will allow obtaining results even closer to the human eye's response.

The inclusion of the visual attention in no-reference objective metrics will solve the problem of environments in which exists a lack of reference, such as the ones related to internet, streaming services or transmission through mobile devices.

7. Merits

7 Merits

7.1 Publications derived from the work of this doctoral thesis

Most of the thesis is based on the following peer-reviewed publications and achievements.

7.1.1 3DTV Quality and Visual Discomfort

- López, J. P., Rodrigo, J. A., Jiménez, D., & Menéndez, J. M. (2013). Stereoscopic 3D video quality assessment based on depth maps and video motion. *EURASIP Journal on Image and Video Processing*, 2013(1), 1-14. December 2013. Impact Factor: 0.74.
- López, J. P., Rodrigo, J. A., Jimenez, D., & Menendez, J. M. (2015, January). Subjective quality assessment in stereoscopic video based on analyzing parallax and disparity. In *Consumer Electronics (ICCE), 2015 IEEE International Conference on* (pp. 164-167). IEEE. ISBN: 978-1-4799-7543-3. Las Vegas (United States). January 9-12, 2015.
- López, J. P., Rodrigo, J. A., Jimenez, D., & Menendez, J. M. (2015, January). Proposal for characterization of 3DTV video sequences describing parallax information. In *Consumer Electronics (ICCE), 2015 IEEE International Conference on* (pp. 353-355). IEEE. ISBN: 978-1-4799-7543-3. Las Vegas (United States). January 9-12, 2015.
- Rodrigo, J. A., López, J. P., Jiménez Bermejo, D., & Menendez Garcia, J. M. (2013). Automatic 3DTV Quality Assessment Based On Depth Perception Analysis. *Nem Summit 2013 Proceedings*, 69-74. Nantes (France), 28th-30th October 2013.

7.1.2 Video Quality Assessment and Visual Attention Models

- López, J. P., Jimenez, D., Cerezo, A., & Menéndez, J. M. (2013, May). No-reference algorithms for video quality assessment based on artifact evaluation in MPEG-2 and H. 264 encoding standards. In *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on* (pp. 1336-1339). IEEE. May 2013.
- López, J. P., Rodrigo, J. A., Jimenez, D., & Menendez, J. M. (2014, November). Insertion of Impairments in Test Video Sequences for Quality Assessment Based on Psychovisual Characteristics. In *Artificial Intelligence, Modelling and Simulation (AIMS), 2014 2nd International Conference on* (pp. 193-198). IEEE. November 2014.
- López, J. P., Rodrigo, J. A., Jimenez, D., & Menendez, J. M. (2015, June). Definition of masks related to psychovisual features for Video Quality Assessment. In *Consumer Electronics (ISCE), 2015 IEEE International Symposium on* (pp. 1-2). IEEE. June 2015.
- López, J. P., Slanina, M., Arnaiz, L., & Menéndez, J. M. (2013, July). Subjective quality assessment in scalable video for measuring impact over device adaptation. In *EUROCON, 2013 IEEE* (pp. 162-169). IEEE. July 2013.
- López, J.P., Jiménez, D., Díaz, M., & Menéndez, J.M. Metrics for the objective quality assessment in high definition digital video". *IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*. 2008.

- López, J.P., Díaz, M., Jiménez, D., & Menéndez, J. M. Tiling effect in quality assessment in high definition digital television. *12th IEEE International Symposium on Consumer Electronics- ISCE2008*, ISBN: 978-1-4244-2422-1, Vilamoura, April 2008.

7.1.3 Book chapters

- López, J.P. Video Quality Assessment. Video Compression, Ed. InTech, ISBN: 978-953-51-0422-3, March 2012.

7.2 Other publications

7.2.1 Peer-reviewed publications

- López, J.P., Sánchez, F.A., González-Miranda, S., Valhondo, J., Jiménez, D., & Bourg, L. "Design of a platform for assessing multimedia advertising content consumption with mobile devices validation". 19th IEEE International Symposium on Consumer Electronics- ISCE2015, Madrid, June 2015.
- López, J.P., Ballesteros, J.P., & Menéndez, J.M. "Mobile application for syntax analysis in Spanish Language oriented to secondary education". 19th IEEE International Symposium on Consumer Electronics- ISCE2015, Madrid, June 2015.
- López, J.P., Ballesteros, J.P., Cerezo, A., & Menéndez, J.M. "Usage of Mobile Devices as Collaborative Tools for Education and Preparation of Official Exams". 19th IEEE International Symposium on Consumer Electronics- ISCE2015, Madrid, June 2015.
- Ballesteros, J. P., & López, J. P. "Child Soldier: ejemplo de cómo llevar las técnicas del teatro contemporáneo al aula de secundaria". *Didáctica. Lengua y Literatura* (Universidad Complutense de Madrid), Volumen 26, p. 57-79. 2th December 2014.
- Jimenez, D., Diaz, M., Martin, C. A., Lopez, J. P., & Menendez, J. M. (2008, April). Media campus: A new interactive video experience for end users. In *Consumer Electronics, 2008. ISCE 2008. IEEE International Symposium on* (pp. 1-4). IEEE.
- López, J. P., Ballesteros, J. P., Jiménez Bermejo, D., & Menendez Garcia, J. M. (2013). Impact of new technologies and social networks on a secondary education theatre project. *Nem Summit 2013 Proceedings*, 111-116.

7.2.2 Lecture Notes

- Menéndez, J. M., López, J. P., Cerezo, A., & Jiménez Bermejo, D. (2010). Televisión 3D: Aspectos básicos y tendencias. *Revista BIT-Boletín Informativo de Telecomunicación*, (183), 56-59. 2010.

7.3 Other merits: research projects

In parallel to the thesis development, some research projects were carried out. Apart from allowing the funding of activities, influenced the proposed framework for the development of research development.

- ACTIVA. Ministerio de Industria, Turismo y Comercio (FIT-330300-2007-42).
- BUSCAMEDIA: hacia una adaptación semántica de medios digitales multirred-multiterminal. [2009-2012].
- CIUDAD2020: HACIA UN NUEVO MODELO DE CIUDAD INTELIGENTE SOSTENIBLE. [2011-2014].
- COST Action IC1105: 3D-ConTourNet 3D Content Creation, Coding and Transmission over Future Media Networks.
- EPSIS. ENTRETENIMIENTO Y PUBLICIDAD SEGMENTADA EN ENTORNOS INMERSIVOS. Ministerio de Economía y Competitividad (IPT-2011-1393-430000) [2011-2013].
- FURIA 2009. Futura red integrada audiovisual. Ministerio de Industria, Turismo y Comercio (TSI-020301-2009-33).
- HBB4ALL Hybrid Broadcast Broadband TV For All [2013-2016]: European project for the deployment of trials for the validation of access services based on Connected TV.
- HORFI-Radar MIMO de banda ultra ancha. TEC2012-38402-C04-01 HORFI.
- ICT 2020. Ministerio de Industria, Turismo y Comercio (TSI-020302-2011-23).
- IMMERSIVE TV: UNA APROXIMACIÓN A LOS MEDIOS INMERSIVOS. Ministerio de Industria, Turismo y Comercio (TSI-020302-2010-61).
- ITACA 3D. Plataforma de creación, producción y distribución de video estereoscópico de entretenimiento para la visualización de televisión en 3D a través de broadcast. Ministerio de Industria, Turismo y Comercio (TSI-020110-2009-396).
- MELISMAS - Generación automática de mensajes en lengua de signos para aplicaciones sanitarias. Ministerio de Economía y Competitividad (RTC-2014-2762-1).
- Palco HD Convergencia de plataformas digitales hacia la Alta Definición y medidas de calidad asociadas. Ministerio de Industria, Turismo y Comercio (FIT-330300-2007-44).
- PALCO HD2. Ministerio de Industria, Turismo y Comercio (TSI-020110-2009-29) [2009-2011].
- PLEASE Plataforma de alta eficiencia avanzada para distribución de contenidos [2014-15].
- PRO-TVD-CM PRO-TVD-CM: Proyecto Integral de Investigación en Televisión Digital (S0505/TIC-0398).
- S3D: Proyecto para el desarrollo de un equipo servidor-editor de vídeo 3D realizado en colaboración con las empresas Overon y Aicox.
- SIRENA: Sistemas y tecnologías 3D Media sobre Internet del Futuro y REdes de difusión de Nueva generación. Ministerio de Economía y Competitividad (IPT-2011-1269-430000).

ACKNOWLEDGEMENT

Acknowledgement

The work developed in this Doctoral Thesis was performed in the framework of a collection of projects: TSI-020302-2010-61 IMMERSIVE TV, TSI-020110-2009-396 ITACA 3D and TSI-020110-2009-29 PALCO HD 2, which are partially funded by the Spanish Ministry of Industry, Tourism and Commerce; IPT-2011-1269-430000 SIRENA, which is partially funded by the Spanish Ministry of Economy and Competitiveness; and TEC2012-38402-C04-01 HORFI and CENIT BUSCAMEDIA, which are partially funded by the Spanish Ministry of Science and Innovation.

The author would like to acknowledge the work of 3D-ConTourNet network, a COST Action supported by the EU Framework Programme Horizon 2020, which undertakes coordinated research collaboration at European level in 3D multimedia creation, encoding, delivery and reception of services and applications over future networking technologies.

Additionally, the author would like to acknowledge the Spanish broadcaster RTVE, the manufacturer company JVC for lending the television used in subjective assessment, people from Mediapro for the cession of 3D stereoscopic video contents and University of Nantes for the cession of the stereoscopic database. We would also like to acknowledge all the people who offered their precious time to contribute to the subjective tests.

Last but not least important, I would like to thank my advisor José Manuel Menéndez García for all his efforts and time spent to finish this Doctoral Thesis. Finally, special thanks to all the people in GATV group for their support and collaboration in this work.

References

References

[ASLIN-1, 1993] ASLIN, Richard N. Perception of Visual Direction in Human Infants. *Visual Perception and Cognition in Infancy*, 1993, vol. 23, p. 91.

[BAHN-1, 2002] BAHN, J. K., SON, J. Y., CHOI, Y. J., KHAM, K., & CHUNG, C. S. Effects of interpupillary distance and AC/A ratio on binocular fusion and depth perception. *Proceedings of the 9th International Display Workshops (IDW'02)*. 2002. p. 1303-1306.

[BANKS-1, 2012] BANKS, Martin S., READ, J. C., ALLISON, R. S., & WATT, S. J. Stereoscopy and the human visual system. *SMPTE motion imaging journal*, 2012, vol. 121, no 4, p. 24-43.

[BEGHDADI-1, 2013] BEGHDADI, Azeddine, LARABI, M. C., BOUZERDOUM, A., & IFTEKHARUDDIN, K. M. A survey of perceptual image processing methods. *Signal Processing: Image Communication*, 2013, vol. 28, no 8, p. 811-831.

[BEIRNESS-1, 2005] BEIRNESS, Douglas J.; SIMPSON, Herb M.; DESMOND, Katherine. *The road safety monitor 2004: Drowsy driving*. 2005.

[BEREZIN-1, 2010] BEREZIN, O. Digital cinema in Russia: Is 3D still a driver for the development of the cinema market. *3D Media*, 2010, vol. 2010.

[BLANCHE-1, 2012] BLANCHE, P.A. Holographic stereography and a new photorefractive material may help create a 3-D display that reproduces all human visual cues. *SID Information Display*. March, 2012.

[BOEV-1, 2008] BOEV, Atanas; HOLLOSI, Danilo; GOTCHEV, Atanas. Classification of stereoscopic artefacts. *MOBILE3DTV Project*, 2008, vol. 216503.

[BORJI-1, 2013] BORJI, Ali; ITTI, Laurent. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013, vol. 35, no 1, p. 185-207.

[BOSC-1, 2011] BOSC, Emilie, Pepion, R., Le Callet, P., Köppel, M., Ndjiki-Nya, P., Pressigout, M., & Morin, L. Towards a new quality metric for 3-D synthesized view assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 2011, vol. 5, no 7, p. 1332-1343.

[BOUJUT-1, 2012] BOUJUT, Hugo, BENOIS-PINEAU, J., AHMED, T., HADAR, O., & BONNET, P. No-Reference Video quality assessment of H. 264 video streams based on semantic saliency maps. *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012. p. 82930T-82930T-9.

[BOVIK-1] BOVIK, A.C., "Automatic Prediction of Perceptual Image and Video Quality," *Proceedings of the IEEE*, vol.101, no.9, pp.2008,2024, Sept. 2013

[BRADSKI-1, 2008] BRADSKY, G., & KAEHLER, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. "O'Reilly Media, Inc."

[CANNY-1, 1986] CANNY, John. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1986, no 6, p. 679-698.

[CHAMBON-1, 2011] CHAMBON, Sylvie; CROUZIL, Alain. Combination of Correlation Measures for Dense Stereo Matching. *VISAPP*. 2011. p. 598-603.

[CHANG-1, 2010] CHANG, Nelson Yen-Chung, Tsai, T. H., Hsu, B. H., Chen, Y. C., & Chang, T. S.. Algorithm and architecture of disparity estimation with mini-census adaptive support weight. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2010, vol. 20, no 6, p. 792-805.

[CHEN-1, 2011] CHEN, Wei, Fournier, J., Barkowsky, M., & Le Callet, P.. New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone. *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011. p.78631O-78631O-13.

[CHEN-2, 2012] CHEN, Wei, Fournier, J., Barkowsky, M., & Le Callet, P.. Quality of experience model for 3DTV. En *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012. p.82881P-82881P-9.

[CHEN-3, 2010] CHEN, Wei, Fournier, J., Barkowsky, M., & Le Callet, P.. New requirements of subjective video quality assessment methodologies for 3DTV. *Video Processing and Quality Metrics 2010 (VPQM)*. 2010.

[CHEN-4, 2005] CHEN, Junqing, Pappas, T. N., Mojsilović, A., & Rogowitz, B. E.. Adaptive perceptual color-texture image segmentation. *Image Processing, IEEE Transactions on*, 2005, vol. 14, no 10, p. 1524-1536.

[CHO-1, 2012] CHO, Sang-Hyun; KANG, Hang-Bong. Subjective evaluation of visual discomfort caused from stereoscopic 3D video using perceptual importance map. *TENCON 2012-2012 IEEE Region 10 Conference*. IEEE, 2012. p. 1-6.

[CHOI-1, 2012] CHOI, Jaeseob, Kim, D., Choi, S., & Sohn, K.. Visual fatigue modeling and analysis for stereoscopic video. *Optical Engineering*, 2012, vol. 51, no 1, p. 017206-1-017206-11.

[CHOI-2, 2012] CHOI, Ji-Hoon, Yun, M. J., Kim, A. R., & Kim, J. O. . Visual comfort measurement for 2D/3D converted stereo video sequence. En *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2012. IEEE, 2012. p. 1-4.

[DESIMONE-1, 1995] DESIMONE, Robert; DUNCAN, John. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 1995, vol. 18, no 1, p. 193-222.

[EMGU-1, 2013] Emgu, C. V. (2013). Emgu CV: OpenCV in .NET (C#, VB, C++ and more). Online: <http://www.emgu.com>

[ENGELKE-1, 2010] ENGELKE, Ulrich, Barkowsky, M., Le Callet, P., & Zepernick, H. J.. Modelling saliency awareness for objective video quality assessment. *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*. IEEE, 2010. p. 212-217.

[GOFERMAN-1, 2012] GOFERMAN, Stas; ZELNIK-MANOR, Lihj; TAL, Ayellet. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012, vol. 34, no 10, p. 1915-1926.

[HAGLUND-1, 2006] HAGLUND, Lars. The SVT high definition multi format test set. *Swedish Television Stockholm*, 2006. Available online at: <ftp://vqeg.its.bldrdoc.gov>.

[HANHART-1, 2012] HANHART, Philippe; DE SIMONE, Francesca; EBRAHIMI, Touradj. Quality assessment of asymmetric stereo pair formed from decoded and synthesized views. *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012. p. 236-241.

[HEWAGE-1, 2013] HEWAGE, Chaminda TER; MARTINI, Maria G. Quality of experience for 3D video streaming. *Communications Magazine, IEEE*, 2013, vol. 51, no 5, p. 101-107.

[HOFFMAN-1, 2008] HOFFMAN, David M., Girshick, A. R., Akeley, K., & Banks, M. S.. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 2008, vol. 8, no 3, p. 33.

[HU-1, 2006] HU, Haibo; DE HANN, G. Low cost robust blur estimator. *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006. p. 617-620.

[HUYNH-THU-1, 2010] HUYNH-THU, Quan; LE CALLET, Patrick; BARKOWSKY, Marcus. Video quality assessment: From 2D to 3D—Challenges and future trends. *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010. p. 4025-4028.

[IATSUN-1, 2015] IATSUN, Iana; LARABI, Mohamed-Chaker; FERNANDEZ-MALOIGNE, Christine. Investigation and modeling of visual fatigue caused by S3D content using eye-tracking. *Displays*, 2015, vol. 39, p. 11-25.

[ITTI-1, 1998] ITTI, Laurent; KOCH, Christof; NIEBUR, Ernst. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998, no 11, p. 1254-1259.

[JIANG-1, 2015] JIANG, Qiuping, Shao, F., Jiang, G., Yu, M., Peng, Z., & Yu, C.. A depth perception and visual comfort guided computational model for stereoscopic 3D visual saliency. *Signal Processing: Image Communication*, 2015.

[JIANG-2, 2015] JIANG, Qiuping; SHAO, Feng. 3D Visual Comfort Assessment via Sparse Coding. En *Image and Graphics*. Springer International Publishing, 2015. p. 10-20.

[JIN-1, 2010] JIN, Seunghun, Cho, J., Pham, X. D., Lee, K. M., Park, S. K., Kim, M., & Jeon, J. W.. FPGA design and implementation of a real-time stereo vision system. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2010, vol. 20, no 1, p. 15-26.

[JUNG-1, 2012] JUNG, Yong Ju, Lee, S. I., Sohn, H., Park, H. W., & Ro, Y. M.. Visual comfort assessment metric based on salient object motion information in stereoscopic video. *Journal of Electronic Imaging*, 2012, vol. 21, no 1, p. 011008-1-011008-16.

[JUNG-2, 2013] JUNG, Yong Ju, Sohn, H., Lee, S. I., Speranza, F., & Ro, Y. M.. Visual importance and discomfort region-selective low-pass filtering for reducing visual discomfort in stereoscopic displays. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2013, vol. 23, no 8, p. 1408-1421.

[JUNG-3, 2015] JUNG, Cheolkon, Cao, L., Liu, H., & Kim, J.. Visual comfort enhancement in stereoscopic 3D images using saliency-adaptive nonlinear disparity mapping. *Displays*, 2015.

[KEROUH-1, 2012] KEROUH, Fatma; SERIR, Amina. A no-reference blur image quality measure based on wavelet transform. En *Digital Information Processing and Communications (ICDIPC), 2012 Second International Conference on*. IEEE, 2012. p. 127-132.

[KIM-1, 2011] KIM, Donghyun; SOHN, Kwanghoon. Visual fatigue prediction for stereoscopic image. *IEEE transactions on circuits and systems for video technology*, 2011, vol. 21, no 2, p. 231-236.

[KOCH-1, 1987] KOCH, Christof; ULLMAN, Shimon. Shifts in selective visual attention: towards the underlying neural circuitry. *Matters of intelligence*. Springer Netherlands, 1987. p. 115-141.

[KOCH-2, 1999] KOCH, Christof; POGGIO, Tomaso. Predicting the visual world: silence is golden. *nature neuroscience*, 1999, vol. 2, p. 9-10.

[KOLB-1, 1995] KOLB, Helga; FERNANDEZ, Eduardo; NELSON, Ralph. Visual Acuity--Webvision: The Organization of the Retina and Visual System. 1995.

[KOOI-1, 2004] KOOI, Frank L.; TOET, Alexander. Visual comfort of binocular and 3D displays. *Displays*, 2004, vol. 25, no 2, p. 99-108.

[LAMBOOIJ-1, 2007] LAMBOOIJ, Marc TM; IJSSELSTEIJN, Wijnand A.; HEYNDERICKX, Ingrid. Visual discomfort in stereoscopic displays: a review. *Electronic Imaging 2007. International Society for Optics and Photonics*, 2007. p.64900I-64900I-13.

[LAMBOOIJ-2, 2011] LAMBOOIJ, Marc, IJsselsteijn, W., Bouwhuis, D. G., & Heynderickx, I.. Evaluation of stereoscopic images: beyond 2D quality. *Broadcasting, IEEE Transactions on*, 2011, vol. 57, no 2, p. 432-444.

[LAMBOOIJ-3, 2013] LAMBOOIJ, M., Murdoch, M. J., IJsselsteijn, W. A., & Heynderickx, I.. The impact of video characteristics and subtitles on visual comfort of 3D TV. *Displays*, 2013, vol. 34, no 1, p. 8-16.

[LE CALLET-1, 2012] LE CALLET, Patrick, Möller, S., & Perkis, A.. Qualinet white paper on definitions of quality of experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012.

[LE MEUR-1, 2006] LE MEUR, Olivier, Le Callet, P., Barba, D., & Thoreau, D.. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2006, vol. 28, no 5, p. 802-817..

[LE MEUR-2, 2007] LE MEUR, Olivier; LE CALLET, Patrick; BARBA, Dominique. Predicting visual fixations on video based on low-level visual features. *Vision research*, 2007, vol. 47, no 19, p. 2483-2498.

[LECLERCQ-1, 2002] LECLERCQ, Philippe; MORRIS, John. Assessing stereo algorithm accuracy. *Proceedings of Image and Vision Computing*, 2002, vol. 2, p. 89-93.

[LECLERCQ-2, 2002] LECLERCQ, Philippe; MORRIS, John. Robustness to noise of stereo matching. *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on*. IEEE, 2003. p. 606-611.

- [LI-1, 2011] LI, Jing; BARKOWSKY, Marcus; LE CALLET, Patrick. The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos. *Quality of multimedia experience (QoMEX), 2011 third international workshop on*. IEEE, 2011. p. 155-160.
- [LI-2, 2010] LI, Chaofeng; BOVIK, Alan Conrad. Content-weighted video quality assessment using a three-component image model. *Journal of Electronic Imaging*, 2010, vol. 19, no 1, p. 011003-011003-9.
- [LI-3, 2011] LI, Cong, Yuan, W., Bovik, A. C., & Wu, X.. No-reference blur index using blur comparisons. *Electronics letters*, 2011, vol. 47, no 17, p. 962-963.
- [LOPEZ-1, 2013] LOPEZ, J. P., RODRIGO, J. A., JIMENEZ, D., & MENENDEZ, J. M.. Stereoscopic 3D video quality assessment based on depth maps and video motion. *EURASIP Journal on Image and Video Processing*, 2013, vol. 2013, no 1, p. 1-14.
- [LOPEZ-2, 2015] LOPEZ, J. P., RODRIGO, J. A., JIMENEZ, D., & MENENDEZ, J. M. Proposal for characterization of 3DTV video sequences describing parallax information. *Proceedings of Consumer Electronics (ICCE), 2015 IEEE International Conference on*. IEEE, 2015. p. 353-355.
- [LOPEZ-3, 2015] LOPEZ, J. P., RODRIGO, J. A., JIMENEZ, D., & MENENDEZ, J. M. Subjective quality assessment in stereoscopic video based on analyzing parallax and disparity. *Consumer Electronics (ICCE), 2015 IEEE International Conference on*. IEEE, 2015. p. 164-167.
- [LOPEZ-4, 2013] LOPEZ, J. P., JIMENEZ, D., CERESO, A. & MENENDEZ, J. M. No-reference algorithms for video quality assessment based on artifact evaluation in MPEG-2 and H. 264 encoding standards. *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*. IEEE, 2013. p. 1336-1339.
- [LOPEZ-5, 2014] LOPEZ, J. P., RODRIGO, J. A., JIMENEZ, D., & MENENDEZ, J. M. Insertion of Impairments in Test Video Sequences for Quality Assessment Based on Psychovisual Characteristics. *Artificial Intelligence, Modelling and Simulation (AIMS), 2014 2nd International Conference on*. IEEE, 2014. p. 193-198.
- [LOPEZ-6, 2015] LOPEZ, J. P., RODRIGO, J. A., JIMENEZ, D., & MENENDEZ, J. M. Definition of masks related to psychovisual features for Video Quality Assessment. *Consumer Electronics (ISCE), 2015 IEEE International Symposium on*. IEEE, 2015. p. 1-2.
- [MA-1, 2012] MA, Lin; LI, Songnan; NGAN, King Ngi. Reduced-reference video quality assessment of compressed video sequences. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2012, vol. 22, no 10, p. 1441-1456.
- [MA-2, 2010] MA, Qi; ZHANG, Liming; WANG, Bin. New strategy for image and video quality assessment. *Journal of Electronic Imaging*, 2010, vol. 19, no 1, p. 011019-011019-14.
- [MARAT-1, 2009] MARAT, Sophie, Phuoc, T. H., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 2009, vol. 82, no 3, p. 231-243.
- [MEESTERS-1, 2004] MEESTERS, Lydia MJ, IJsselsteijn, W. A., & Seuntjens, P. J. A survey of perceptual evaluations and requirements of three-dimensional TV. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2004, vol. 14, no 3, p. 381-391.

[MEI-1, 2011] MEI, Xing, Sun, X., Zhou, M., Jiao, S., Wang, H., & Zhang, X.. On building an accurate stereo matching system on graphics hardware. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011. p. 467-474.

[MIDDLEBURY DATABASE] The Middlebury Database – Stereo Evaluation (version 2). <http://vision.middlebury.edu/stereo/eval/>

[NINASSI-1, 2007] NINASSI, Alexandre, Meur, O. L., Callet, P. L., & Barbba, D. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. *Image Processing, 2007. ICIP 2007. IEEE International Conference on*. IEEE, 2007. p. II-169-II-172.

[NOJIRI-1, 2006] NOJIRI, Y., Yamanoue, H., Ide, S., Yano, S., & Okana, F. Parallax distribution and visual comfort on stereoscopic HDTV. *Proc. IBC*. 2006. (No. 3, pp. 373-380).

[NOJIRI-2, 2004] NOJIRI, Yuji, Yamanoue, H., Hanazato, A., Emoto, M., & Okano, F. Visual comfort/discomfort and visual fatigue caused by stereoscopic HDTV viewing. *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004. p. 303-313.

[NOJIRI-3, 2003] NOJIRI, Yuji, Yamanoue, H., Hanazato, A., & Okano, F. Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV. *Electronic Imaging 2003*. International Society for Optics and Photonics, 2003. p. 195-205.

[OELBAUM-1, 2009] OELBAUM, Tobias; KEIMEL, Christian; DIEPOLD, Klaus. Rule-based no-reference video quality evaluation using additionally coded videos. *Selected Topics in Signal Processing, IEEE Journal of*, 2009, vol. 3, no 2, p. 294-303.

[OKUN-1, 2010] OKUN, Jeffrey A.; ZWERMAN, Susan. The VES handbook of visual effects: industry standard VFX practices and procedures. Taylor & Francis, 2010.

[PANG-1, 2008] PANG, Derek, Kimura, A., Takeuchi, T., Yamato, J., & Kashino, K. A stochastic model of selective visual attention with a dynamic Bayesian network. *Multimedia and expo, 2008 IEEE international conference on*. IEEE, 2008. p. 1073-1076.

[PARK-1, 2014] PARK, Jincheol; LEE, Sanghoon; BOVIK, Alan C. 3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation. *Selected Topics in Signal Processing, IEEE Journal of*, 2014, vol. 8, no 3, p. 415-427.

[PARK-2, 2015] PARK, Jincheol, Oh, H., Lee, S., & Bovik, A. C. 3D Visual Discomfort Predictor: Analysis of Disparity and Neural Statistics. *Image Processing, IEEE Transactions on*, 2015, vol. 24, no 3, p. 1101-1114.

[PERCIVAL-1, 1892] PERCIVAL, Archibald Stanley. The relation of convergence to accommodation and its practical bearing. *Ophthalm. Rev*, 1892, vol. 11, p. 313-328.

[PLESS-1, 1998] PLESS, Vera; BRUALDI, Richard A.; HUFFMAN, William Cary. *Handbook of coding theory*. Elsevier Science Inc., 1998.

[REC BT.1438] REC, I. T. U. R. BT. 1438. Subjective assessment of stereoscopic television pictures, 2000.

[REC BT.2021] ITU-R RECOMMENDATION, BT.2021. Subjective methods for the assessment of stereoscopic 3DTV systems. 2015 (last version). BT Series, Broadcasting service (television).

[REC BT.500] RECOMMENDATION, I. T. U. R. BT. 500-11, "Methodology for the subjective assessment of the quality of television pictures," ITU, 2002.

[REC P.910] ITU-T RECOMMENDATION, P.910. Subjective video quality assessment methods for multimedia applications. 1999 (first version), 2011 (last version). *Recommendations of the ITU (Telecommunication Standardization Sector)*.

[REC BT.601] REC, I. T. U. R. BT 601: Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios. *ITU-R Rec. BT*, 1995, vol. 656.

[REC BT.709] Recommendation, I. T. U. R. B. T. (1990). 709. *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange*, now ITU-R BT, 709.

[REC BT.2020] Recommendation, D. N. (2012). Parameter Values for UHD TV Systems for Production and International Programme Exchange, ITU-R BT. 2020. *International Telecommunications Union*.

[REC H.264] ITU-T RECOMMENDATION, H. 264 "Advanced Video Coding for Generic Audiovisual Services". *ISO/IEC*, 2003, vol. 14496.

[REC H.265] REC, I. T. U. T. H. 265 and ISO/IEC 23008-2: High efficiency video coding. *ITU-T and ISO/IEC JTC*, vol. 1.

[REN-1, 2015] REN, Hui, Su, Z., Lv, C., & Zou, F.. Effect of region contrast on visual comfort of stereoscopic images. *Electronics Letters*, 2015, vol. 51, no 13, p. 983-985.

[RICHARDSON-1, 2004] RICHARDSON, Iain E. H. *264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.

[RODRIGO-1, 2011] RODRIGO, Juan Antonio; JIMENEZ, David; MENÉNDEZ, José Manuel. Real-time 3-D HDTV depth cue conflict optimization. *Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference on*. IEEE, 2011. p. 5-9.

[SERIR-1, 2011] SERIR, Amina. No-reference blurred image quality assessment. *Visual Information Processing (EUVIP), 2011 3rd European Workshop on*. IEEE, 2011. p. 168-173.

[SESHADRINATHAN-1, 2010] SESHADRINATHAN, Kalpana; BOVIK, Alan Conrad. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 2010, vol. 19, no 2, p. 335-350.

[SHARSTEIN-1, 2002] SCHARSTEIN, Daniel; SZELISKI, Richard. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 2002, vol. 47, no 1-3, p. 7-42.

[SHIBATA-1, 2011] SHIBATA, Takashi, Kim, J., Hoffman, D. M., & Banks, M. S. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of vision*, 2011, vol. 11, no 8, p. 11.

[SIEGEL-1, 2001] SIEGEL, Mel. Perceptions of crosstalk and the possibility of a zoneless autostereoscopic display. *Photonics West 2001-Electronic Imaging*. International Society for Optics and Photonics, 2001. p. 34-41.

[SO-1, 2016] SO, Gil-Ja; KIM, Sang-Hyun; KIM, Jeong-Yeop. Evaluation Model of the Visual Fatigue on the 3D Stereoscopic Video. *International Journal of Computer Theory and Engineering*, Vol. 8, No. 4, August 2016

[SONNATI-1, 2015] Sonnati, F. H265 – part II: Considerations on quality and “state-of-the-art”. Video Encoding & Streaming Technologies. <https://sonnati.wordpress.com/>

[SPERANZA-1, 2006] SPERANZA, Filippo, Tam, W. J., Renaud, R., & Hur, N. Effect of disparity and motion on visual comfort of stereoscopic images. *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006. p.60550B-60550B-10.

[TAM-1, 2011] TAM, Wa James, Speranza, F., Yano, S., Shimono, K., & Ono, H. Stereoscopic 3D-TV: visual comfort. *Broadcasting, IEEE Transactions on*, 2011, vol. 57, no 2, p. 335-346.

[TREISMAN-1, 1980] TREISMAN, Anne M.; GELADE, Garry. A feature-integration theory of attention. *Cognitive psychology*, 1980, vol. 12, no 1, p. 97-136.

[TSENG-1, 2012] TSENG, Kun-Lung, Huang, W. J., Luo, A. C., Huang, W. H., Yeh, Y. C., & Chen, W. C. Automatically optimizing stereo camera system based on 3d cinematography principles. *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2012. IEEE, 2012. p. 1-4.

[URVOY-1, 2012] URVOY, Matthieu, Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P.... & Garcia, N.. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on. IEEE, 2012. p. 109-114.

[URVOY-2, 2013] URVOY, M., LI, J. & LE CALLET, P. Visual Comfort and Fatigue in Stereoscopy. *3D Video: From Capture to Diffusion*. 2013. Ed. Wiley.

[VIOLA-1, 2001] VIOLA, Paul; JONES, Michael. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001. p. I-511-I-518 vol. 1.

[VERGHESE-1, 2001] VERGHESE, Preeti. Visual search and attention: A signal detection theory approach. *Neuron*, 2001, vol. 31, no 4, p. 523-535.

[VQEG-1, 2015] VQEG Research group. Establishing a Ground Truth for Quality of Experience in 3D for assessment methodologies in 3D Video Quality Assessment. 2015. Available at: <http://www.its.bldrdoc.gov/vqeg/projects/3dtv/3dtv.aspx>

[WALKER-1, 2002] WALKER, Laura L.; MALIK, Jitendra. When is scene recognition just texture recognition. *Journal of Vision*, 2002, vol. 2, no 7, p. 255.

[WANG-1, 2004] WANG, Zhou, Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 2004, vol. 13, no 4, p. 600-612.

[WANG-2, 2004] WANG, Zhou, Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. The SSIM index for image quality assessment. *MATLAB implementation available online from: http://www.cns.nyu.edu/lcv/ssim*, 2003, vol. 23, p. 66.

[WANG-3, 2002] WANG, Zhou; BOVIK, Alan C.; LU, Ligang. Why is image quality assessment so difficult? *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002. p. IV-3313-IV-3316.

[WANG-4, 2006] WANG, Zhou; BOVIK, A. C. Modern image quality assessment (synthesis lectures on image, video, and multimedia processing). *San Rafael, CA: Morgan Claypool*, 2006.

[WARE-1, 1998] WARE, Colin; GOBRECHT, Cyril; PATON, Mark Andrew. Dynamic adjustment of stereo display parameters. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 1998, vol. 28, no 1, p. 56-65.

[WATSON-1, 2001] WATSON, Andrew B.; HU, James; MCGOWAN, John F. Digital video quality metric based on human vision. *Journal of Electronic imaging*, 2001, vol. 10, no 1, p. 20-29.

[WINKLER-1, 2005] WINKLER, Stefan. *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005. ISBN-13: 978-0470024041

[WINKLER-2, 2009] WINKLER, Stefan. On the properties of subjective ratings in video quality experiments. En *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009. p. 139-144.

[WINKLER-3, 1999] WINKLER, Stefan. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, 1999, vol. 78, no 2, p. 231-252.

[WOLFE-1, 1994] WOLFE, Jeremy M. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1994, vol. 1, no 2, p. 202-238.

[WOODFILL-1, 1997] WOODFILL, John; VON HERZEN, Brian. Real-time stereo vision on the PARTS reconfigurable computer. *Field-Programmable Custom Computing Machines, 1997. Proceedings, The 5th Annual IEEE Symposium on*. IEEE, 1997. p. 201-210.

[WOODING-1, 2002] WOODING, David S. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, 2002, vol. 34, no 4, p. 518-528.

[WOODING-2, 2002] WOODING, David S. Fixation maps: quantifying eye-movement traces. *Proceedings of the 2002 symposium on Eye tracking research & applications*. ACM, 2002. p. 31-36.

[WU-1, 2013] WU, Jinjian, Lin, W., Shi, G., & Liu, A. Perceptual quality metric with internal generative mechanism. *Image Processing, IEEE Transactions on*, 2013, vol. 22, no 1, p. 43-54.

[X265-1] x265 HEVC Encoder / H.265 Video Codec. <http://x265.org/>

[XIAO-1, 2000] XIAO, Feng, et al. DCT-based video quality evaluation. *Final Project for EE392J*, 2000, vol. 769.

[YAMADA-1, 2012] YAMADA, Toru; NISHITANI, Takao. No-reference quality estimation for compressed videos based on inter-frame activity difference. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2012, vol. 95, no 8, p. 1240-1246.

[YAMANOUE-1, 2006] YAMANOUE, Hirokazu; OKUI, Makoto; OKANO, Fumio. Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2006, vol. 16, no 6, p. 744-752.

[YANO-1, 2002] YANO, Sumio, Ide, S., Mitsuhashi, T., & Thwaites, H. A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays*, 2002, vol. 23, no 4, p. 191-201.

[YOU-1, 2009] YOU, Junyong, Perkis, A., Hannuksela, M. M., & Gabbouj, M. Perceptual quality assessment based on visual attention analysis. *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009. p. 561-564.

[ZABIH-1, 1994] ZABIH, Ramin; WOODFILL, John. Non-parametric local transforms for computing visual correspondence. *Computer Vision—ECCV'94*. Springer Berlin Heidelberg, 1994. p. 151-158.

[ZHAI-1, 2012] ZHAI, Guangtao, Wu, X., Yang, X., Lin, W., & Zhang, W. A psychovisual quality metric in free-energy principle. *Image Processing, IEEE Transactions on*, 2012, vol. 21, no 1, p. 41-52.

[ZHANG-1, 2015] ZHANG, Xiaonan, Zhou, J., Chen, J., Guo, X., Zhang, Y., & Gu, X. Visual comfort assessment of stereoscopic images with multiple salient objects. *Broadband Multimedia Systems and Broadcasting (BMSB), 2015 IEEE International Symposium on*. IEEE, 2015. p. 1-6.

Annexes

Annexes

Annex 1. Video Quality Metrics Definition

MSE: Mean Square Error

$$MSE(l) = \frac{1}{W \times H} \sum_{j=0}^{H-1} \sum_{i=0}^{W-1} \left(f^l_{DIST}(i, j) - f^l_{REF}(i, j) \right)^2$$

- ◆ $f^l_{REF}(i, j)$: luminance of the frame l in reference video sequence corresponding to pixel (i, j) .
- ◆ $f^l_{DIST}(i, j)$: luminance of the frame l in distorted video sequence corresponding to pixel (i, j) .
- ◆ l : number of the frame analyzed.
- ◆ W : width of the video frame in number of columns.
- ◆ H : height of the video frame in number of rows.

SNR: Signal-to-Noise Ratio

$$SNR(l) = 10 \log \frac{\sum_{j=0}^{H-1} \sum_{i=0}^{W-1} \left(f^l_{REF}(i, j) \right)^2}{\sum_{j=0}^{H-1} \sum_{i=0}^{W-1} \left(f^l_{DIST}(i, j) - f^l_{REF}(i, j) \right)^2} [dB]$$

- ◆ $f^l_{REF}(i, j)$: luminance of the frame l in reference video sequence corresponding to pixel (i, j) .
- ◆ $f^l_{DIST}(i, j)$: luminance of the frame l in distorted video sequence corresponding to pixel (i, j) .
- ◆ l : number of the frame analyzed.
- ◆ W : width of the video frame in number of columns.
- ◆ H : height of the video frame in number of rows.

PSNR: Peak Signal-to-Noise Ratio

$$PSNR(l) = 10 \log \frac{\sum_{j=0}^{H-1} \sum_{i=0}^{W-1} P^2}{\sum_{j=0}^{H-1} \sum_{i=0}^{W-1} \left(f^l_{DIST}(i, j) - f^l_{REF}(i, j) \right)^2}$$

- ◆ $f^l_{REF}(i, j)$: luminance of the frame l in reference video sequence corresponding to pixel (i, j) .
- ◆ $f^l_{DIST}(i, j)$: luminance of the frame l in distorted video sequence corresponding to pixel (i, j) .
- ◆ l : number of the frame analyzed.
- ◆ W : width of the video frame in number of columns.
- ◆ H : height of the video frame in number of rows.
- ◆ P : Peak or maximum value of the color component used as a reference to the difference of pixels. For example, considering a resolution of 8 bits per pixel, the peak value of luminance is $2^8=255$.

SSIM: Structural Similarity

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

- ◆ μ_x : The average of x component for the reference video sequence.
- ◆ μ_y : The average of y component for the distorted video sequence.
- ◆ σ_x^2 : The variance of x component for the reference video sequence.
- ◆ σ_y^2 : The variance of y component for the distorted video sequence.
- ◆ σ_{xy} : The covariance between components x and y.
- ◆ σ_{xy} : The covariance between components x and y.
- ◆ $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$: two constants to stabilize the division with weak denominator.
- ◆ L : Dynamic range of the pixel-values.

DSSIM: Structural Dissimilarity

$$DSSIM(x, y) = \frac{1 - SSIM(x, y)}{2}$$

Annex 2. Software for Video Quality Metrics Definition

For the present doctoral thesis, a software in Visual C# (.NET) was developed to cover all the tests described. The software offers different interfaces for each concrete development, as seen in next figures.

Fig. 117. Interface for 3D Analysis

Fig. 118. Interface for mask development

Fig. 119. Interface for TI/SI diagrams

Annex 3. Information relative to subjective assessment with pairs of 3D stereoscopic sequences

Annex 3.1. Depth maps in the transition (previous frame and next frame) and evolution of percentages of positive and negative parallax in the transition

Seq	Original image, depth map of the original image and disparities histogram	
1a
1a
1b
1b

2a
2a
2b
2b
3a

3a
3b
4a
4a

4b
4b
5a
5a
5b

5b
6a	<div> </div> <div> </div>
6a
6b	<div> </div> <div> </div>
6b

7a
7a
7b
7b
8a

8a
8b
8b
9a
9a

9b
9b
10a
10a
10b

10b
11a	<div> </div> <div> </div>
11a
11b	<div> </div> <div> </div>
11b

12a
12a
12b
12b
13a

13a	
13b	
13b	
14a	
14a	

14b
14b
15a
15a
15b

15b
16a
16a
16b
16b

17a
17a
17b
17b
18a

18a
18b
18b
19a
19a

19b
19b
20a
20a
20b

Annex 3.2. Temporal Information of sequences (frame by frame)

Annex 3.3. Spatial Information of sequences (frame by frame)

Annex 3.4. Evolution of percentages of positive parallax (blue bar) and negative parallax (yellow bar) in the transition