Dissertação

Mestrado em Engenharia Eletrotécnica

Subjective assessment of 3D still images using attention models

Auridélia Moura de Arruda

Leiria, Setembro de 2015

Dissertação

Mestrado em Engenharia Eletrotécnica

Subjective assessment of 3D still images using attention models

Auridélia Moura de Arruda

Dissertação de Mestrado realizada sob a orientação do Doutor Pedro António Amado Assunção, Professor Coordenador da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

Leiria, Setembro de 2015

This page was intentionally left blank

Muda de vida se tu não vives satisfeito
Muda de vida, estás sempre a tempo de mudar
Muda de vida, não deves viver contrafeito
Muda de vida se há vida em ti a latejar
("Muda de vida" – António Variações)

Abelha fazendo o mel vale o tempo que não voou.
("Amor de índio" – Milton Nascimento)

Quanto a vós, não vos façais chamar de 'rabi', pois um só é vosso Mestre e todos vós sois irmãos.
(Mt 23, 10)

This page was intentionally left blank

Dedicated to Maria Amélia de Arruda (in memorian).

This page was intentionally left blank

Acknowledgments

To God, the Triune, Creator, Saviour and Sanctifier, and Virgin Mary, the venerable track, cosy cradle, source of strength and peace.

To my parents, Mariedna and João Alberto, main counsellors and disciplinarians, source of all that makes me truly human. Best representativeness of love. “To educate is not to cut the wings, but to guide the flight”. To my brother, João Alberto, and my sister, Maridélia, essentials in every growth of mine. To her and her husband, Igor, who gave me a couple of beloved nephews, Catherine and Dimitri, certainties of a blessed future.

To my dear colleague João F. M. Carreira, who kindly volunteered to assist me with everything about he was available to, at every time I needed during the execution of this work, especially in the very early parts.

To Prof. Dr. Pedro A. A. Assunção, by his dedication to this beautiful career and to those who benefit from it. Also to Instituto de Telecomunicações for providing the logistics, technical equipment and scientific support for this research.

To my colleague and friends Nelson C. Francisco, Mrs Lúcia P. Carreira and Mr. Isidro de J. Francisco, who so lovingly treat me, even since before arriving in Portugal.

To my friends Larissa Fumiga Leite, Eliamara da Silva, Diana Rodrigues, Sandra Santos, Marycota Machado, Bruno Rebelo, Sofia Gualdino, Silene and Sylvain Marcelino; the Silva, the Ferreira and the Affonso families; the spectacular CHIC Woman’s Health group and friends; my great job leaders Rita Lucas and José Leal, the fantastics YBRIK team and KIRBY® Company; the colleagues Gilberto Jorge, Marlene Machado, Luis Pinto, Luis Lucas, Ricardo Monteiro, Filipe Xavier; the invaluable IPL employee Lucinda Carreira; Prof. Dr. Sérgio M. M. de Faria; and other friends, colleagues and teachers who contributed in each aspect of the implementation of this work. Since the beginning, far or near.

This page was intentionally left blank

Resumo

O sistema visual humano (HVS) conjugado com o processo de atenção visual tipifica rapidamente partes notáveis em cenas visuais, que subjetivamente definem regiões de interesse (ROI). Os mecanismos subjetivos dominam os movimentos dos olhos nos dois primeiros segundos de visualização e, tendo em consideração a relação entre os movimentos dos olhos e a atenção visual, o registo dos movimentos oculares permite obter dados experimentais para definir modelos de atenção 2D: os movimentos oculares registrados e processados são expressos como um Mapa de Densidade da Fixação (FDM). No domínio 3D, é necessário considerar parâmetros adicionais de dimensão visual, como sejam as diferentes profundidades dos objetos da cena, na definição de ROI com base na atenção visual

O principal objetivo deste trabalho foi verificar o impacto de ROIs definidas com modelos de atenção visual, na percepção da qualidade subjetiva de imagens 3D estáticas. Foi usada a base de dados de imagens e dados de fixação do olhar 3DGaze, obtida a partir de um teste de rastreamento ocular descrito em [1], criada para a avaliação de modelos de atenção visual de imagens estereoscópicas 3D fixas. Esta base de dados foi escolhido para ser usada devido à informação estar disponível, com imagens originais de vários tamanhos HD, todas em formato PNG e com conteúdo natural. Foram geradas máscaras binárias para cada FDM e a cada imagem correspondente foi adicionado ruído considerando as fronteiras das ROIs para diferenciar a intensidade de ruído. Usando as posições dos pixels da máscara binária e da imagem, o ruído foi adicionado à imagem nos pixels posicionados dentro ou fora da ROI, conforme os casos. No testes subjetivos os observadores registaram a sua apreciação sobre a qualidade das imagens, verificando assim a diferente importância que a localização do ruído na ROI tem na avaliação subjetiva da qualidade das imagens. As imagens foram classificadas em imagens com ruído adicionado dentro ou fora da ROI, de acordo com o tipo de ruído (*Gaussian*, *Speckle*), valores dos parâmetros do ruído (nível de intensidade) e a vista da imagem ruidosa (esquerda ou direita).

Verificou-se que o ruído *Gaussian* teve menos impacto na qualidade do que o *Speckle*, bem como o nível de intensidade, que se torna mais importante quando o ruído é introduzido na vista direita e quando o está dentro da ROI. Isso é justificado devido aos visualizadores terem fixado seus olhos sobre aquela região durante mais tempo, tornando-se maior a distorção percecionada. Como o espectro de análise foi pequeno, as informações acerca do olho dominante parecem ser inconclusivas. A alteração dos parâmetros é sugerida para trabalhos futuros, de tal modo que haja mais certeza sobre os resultados alcançados, sem interferências entre análises. Como este trabalho foi limitado aos dados contidos nas imagens disponíveis (interesse visual do tipo *bottom-up*), alguns conceitos relacionados ao contexto da visualização (interesse visual do tipo *top-down*), como a raridade ou surpresa, podem naturalmente ser incluídos em trabalhos futuros, bem como a característica de se olhar principalmente para rostos humanos ou coisas humanóides.

Palavras-Chave: Modelo de Atenção, bottom-up, profundidade, eye tracking, FDM, ROI, still images, VA

This page was intentionally left blank

Abstract

The subjective process associated with image quality evaluation is endorsed by human psychophysical and physiological measurements. In the human visual system (HVS), the visual attention (VA) is a crucial element, which quickly identifies the notable regions of the images, subjectively linked to Regions of Interest (ROI). These are represented by a binary mask, indicating whether a pixel in the corresponding image belongs to the ROI. Subjective mechanisms dominate eye movements in the first two seconds of viewing and, due to the high relation between the eye actions and the VA, eye-tracking tests are used to validate 2D attention models: eye movements are recorded and processed to generate a Fixation Density Map (FDM). In the 3D domain, an essential factor for VA among the additional parameters of visual dimension is the scene depth.

The main objective of this work was to study the impact of a particular ROI on the subjective quality perception of 3D still images, considering different types and level of noise (or distortion) in and out of the ROI. The 3DGaze images and eye movement database, obtained from an eye tracking experiment described in [1] and specifically created for performance evaluation of stereoscopic 3D attention models was used. Besides the full public availability, this database has original images in various HD sizes, all in PNG format and with natural content.

Binary masks were generated for each FDM and different types and intensities of noise was added to each corresponding image according to the ROIs. By using the generated binary masks and image pixels positions in the ROI, the noise was added in the image regions located inside or outside the ROI. Subjective testing with users observing the images and scoring their quality was done to verify the importance of these regions in the subjective quality evaluation. The images were classified according to whether the noise was added inside or outside the ROI, the noise type (Gaussian, Speckle), parameter values (intensity level) and the noisy image view (left or right).

The results have shown that Gaussian noise has less impact on the quality than Speckle, with higher intensity level and also when the noise is added to the right view and inside the ROI. This is justified due to the fact that viewers fix their eyes over the ROI during more time, thus perceiving higher distortion. As the amount of different image content was small, the information about the dominant eye appears to be inconclusive. Changing some parameters is suggested for future works, in such a way that there is more certainty on the results without interferences between analyses. As this work was limited to data contained in the image (bottom-up visual interest), some concepts related to the visualization context (top-down visual interest), such as rarity or surprise, may naturally be included in future works, as well as the characteristic of looking primarily for human faces or humanoid things.

Key-Words: Attention model, bottom-up, depth, eye tracking, FDM , ROI , still images, VA.

This page was intentionally left blank

Figures

This page was intentionally left blank

Acronyms

ACR
Absolute Category Rating, 17, 18, 19, 22, 23, 27, 28, 52

DCR
Degradation Category Rating, 22, 23

DCT
Discrete Cosine Transform, 10, 11

DMOS
Differential or Degraded Mean Opinion Score, 16, 23

DS
Double Stimulus, 19, 22, 26, 27, 54

DSCQS
Double Stimulus Continuous Quality Scale, 17, 22, 24, 25, 26, 27, 28, 29

DSCS
Double Stimulus Comparison Scale, 20, 21, 27

DSIS
Double Stimulus Impairment Scale, 17, 22, 23, 24, 25, 26, 27, 28

DWT
Discrete Wavelet Transform, 3

FDM
Fixation Density Map, iii, iv, vi, 5, 13, 37, 38, 41, 43, 44, 54, 57, 61, 73

FR
Fully Referenced, 10, 11, 13

HDTV
High Definition Television, 10, 11

HVS
Human Visual System, iii, vi, 1, 2, 4, 5, 6, 7, 10, 13, 15, 22, 32, 33, 35, 54

ITU
International Telecommunications Union, 9, 11, 15, 17, 20, 21, 52

ITU-R
Radiocommunication Sector of the International Telecommunications Union, 17, 19, 21, 26

This page was intentionally left blank

Summary

This page was intentionally left blank

1. Introduction

For quality optimization it is important to precisely define the concept of quality, to ensure modelling efficiency and user satisfaction when such models are used to guarantee quality of experience (QoE). Most video quality metrics are chosen according to studies in medical areas addressing the Human Visual System (HVS) and its descriptions: based on human perception of the scenes, trying to model the human sensory system, but just in parts, because the recreated 3D visuo-sensory experience is not completely acessible. Measuring quality perception is complex, then, due to the need of selecting those parameters that better represent the entire HVS, combinations of algorithm-extracted features that summarize the perceptually most relevant stimuli through its individualities are used, covering the whole context of video consumption experience [2] [3] [4].

Choosing the features to be extracted needs their perceptual weighting and interdependencies modelling and the effects between them being masked. Also several factors can affect the results, such as distortions introduced in each algorithm processing step, influencing the viewer experience and turning it difficult to obtain a completely accurate calculus of video visual fidelity. For 3D video, in general only the left and right image views and the associated depth or disparity pixels map are processed and approximations validate them, turning it even less faithful to reality (though generally reasonable). 3D quality measurement is even more useful than 2D to reach better performance and user acceptance in emerging services and applications (considering human capabilities to distinguish and classify images) [3] [4].

Quality assessment studies have vast relevance on multimedia systems, where subjective evaluation protocols and processes and objective metrics for multi modal content are crucial: the interaction of audio and speech with video and haptic data, influence of user background, attention, motivation, task, and subjective as well as objective metrics for interactive applications [5] [6].

Objective quality quantitative measurements are usually applied on different stages in the transmission chain, such as compression and processing systems, aiming overcome huge obstacles faced in subjective quality measurement, such as complexity and cost. Furthermore, to monitor hundreds of video channels, for long periods of time, objective video quality metrics are the only practicable solution, with algorithms calculating

values that represent the combinations of different factors on the network [7] [8] [9] [10].

Beyond the term “quality” being defined (as also as the calculus), it is also essential to clarify the vocabulary applied for the material in use. The frames (or set of them) used as reference video source are also called Source Reference Circuit (SRC), unprocessed, unimpaired or original images, and those to be compared with the SRC are commonly a distorted version of them, being often referred as Processed Video Sequence (PVS). Most perceptual image quality assessment approaches proposed in the literature attempt to weight different aspects of the error signal (difference between PVS and SRC) according to their visibility, as concluded by psychophysical measurements in humans or physiological measurements in animals, what not necessarily is the basis, but it is what inevitably determines the success of every method.

Visual attention (VA) is the biological vision ability to rapidly detect the interesting parts of a given scene. It is usually used in a preprocessing step in computer vision systems, by selecting features that represent conspicuous parts of the scene (available sensory information), reducing the computation cost of high level tasks (such as segmentation and object recognition). Psychophysical studies show that it plays a fundamental role in human vision. At HVS, VA guides eye movements to place the fovea (a high central resolution part of the retina) on the interesting part of the scene, where is then processed [11].

Image processing may be done from cleaning up to alter, from defining a Region Of Interest (ROI) to filtering and spatial manipulation. A VA portion of an image usually subjectively associated to a ROI. Specifying a ROI allows for the definition of shaped regions within a given image, often called subimages defined by a placeholder that bounds a location within an image. The ROI representation can be static (defined during encoding), or dynamic (set by the user during a progressive transmission) and it need to have priority on transmition. Every point of an image is either inside or outside of a given ROI, which may be represented as an amount of points and lines, or as single pixels width (square, the simplest two dimensional ROI contour, with a selected region with four equal sides, useful to return single value data about a region of the image, like intensity, colour data, etc; or rectangle, the most popular two dimensional ROI contour, with more flexibility in defining the required region), returning one-dimensional information. However, setting the corners is not enough to fully define a ROI, as it contains several other attributes [12] [13].

The regions that composes the ROI can be geographic in nature (such as polygons encompassing contiguous pixels: with irregular shapes, rectangles or squares, circles, annuluses, polygons, rotated rectangle, annulus arc, or even with freehand shapes), or defined by a range of intensities (not necessarily contiguous pixels: the intensity contours following the outer shape of the ROI, starting at the first defined point, for example). A geographic ROI representation is defined by creating a *binary mask*, a binary image with the same size of the SRC (pixels set to 0 or 1, relative to it defining the ROI or not) and that can be associated with a particular image, without having an associated image, or it may also not require an input image [12] [13] [14].

ROI application is useful when certain parts of the image are more important or should have higher encoding quality than the background or less interesting regions. Dynamic ROI definition is useful in inspection routines, where the user can pan and zoom a source image to locate features, in image processing routines or being defined around the alphanumeric characters, to make it much simpler and faster for an Optical Character Recognition (OCR) routine to decode. Fields of computer vision that can benefit from this task are industrial quality control, surveillance or autonomous mobile systems [12] [13].

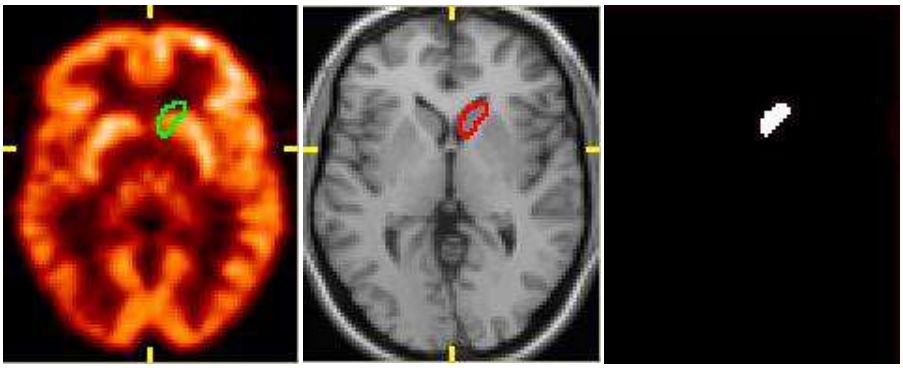


Figure 1.1: ROI application on neuroimaging: MRI, PET and the obtained mask [15].

The JPEG-2000 standard does ROI tagging: standard partitions of the original image (independently compressed blocks, treated like distinct images), have their DC levels passed through a Discrete Wavelet Transform (DWT). ROI is applied in the edge regions: they are encoded separately by different bit planes, and at the decoder, to avoid artifacts along the borders, the map of disparities are encoded in parts, using a ROI. Drawing a ROI is also useful on medical neuroimage processes to examine the morphological properties of an anatomic structure and to extract information for a specific structure from a corresponding functional data set, through defining a specific

shape. In Figure 1.1 there is a ROI drawn on a Magnetic Resonance Imaging[1] (MRI) image, the ROI is superimposed onto a Positron Emission Tomography (PET) image (not used at all in the drawing process, as the ROI define pixels from a homogenous region, named caudate, difficult to discern in the PET image) and its mask (pixels each have a value of 1 or 0, consonant belonging to the caudate or background) [11] [15].

The visual importance is a top-down visual representation (related to the context), specified by an object-level map in which the objects are rated according to a subjective relevance (content usability, user behaviour and needs, appropriateness, context, object recognition, artistic intent), by comparison of image objects (including the subjective background). Still supplementary subjective measurements are necessary, since quantifying visual quality via objective criteria gives accurate and repeatable results, but none objective system completely reproduces the subjective experience [8] [16] [17].

Approaches based on visual salience are used to detect the ROI from a perceptual viewpoint not related to the user opinion (bottom-up). Visual gaze patterns are, at least for early gaze locations, driven by bottom-up, signal-based attributes (saliency-driven, fast and independent of particular viewing tasks) and saliency (standing out in terms of one or more attributes, commonly contrast or colour) is an aspect which correlates with those locations. Visual saliency aims to compute a saliency map showing points that usualy are local maxima or minima of the image informative function. In video data this also includes the temporal characteristics. The link between visual salience and visual importance is strongest in the first two seconds of the observation interval, when top-down mechanisms dominate eye movements (saliency maps created from early gaze samples have more reliable values, as minus less bottom-up mechanisms are considered to be at work in eyes movements, being stronger the prediction for the early fixations or primary ROI) [15] [18].

The recognition of the main subjects in a scene (the vast majority of early gaze position samples, the gist of the scene) might be achieved in HVS via neurons interactions, providing psychophysical evidence that lends support to a gist-based strategy and a possible role for the feedback connections prevalent in mammalian visual systems. Due to the strong relation between eye movements and VA, the validation of VA computational objective models is done through eye tracking experiments: eye movements are recorded, post-processed and are represented as successions of fixations

[1] MRI is a medical imaging technique used in radiology to investigate the anatomy and function of the body health and disease.

and saccades or as a Fixation Density Map (FDM), which identifies the local fixed by the viewers [15] [18].

Algorithms that search for visual degradations often rely on HVS parts. Human perception motivates handling horizontal or vertical structures in an image differently from diagonal edges, and errors at the edges are perceived to be less disturbing than errors in the middle of the picture. The vision modelling approach of picture metrics applies thus the possibility to base the complete design of an algorithm on the human brain simulation, aiming to process the visual data by simulating the visual pathway of the eye-brain system [19] [10] [20].

1.1. Context and motivation

In the computational model of stereoscopic visual saliency portraied in [21], the authors have further analysed each parameter, trying not to use generalizations and to be accurate about factors that are customarily largely overlooked in similar models. The images and FDMs generated and employed on that work were choosen to be used in this work due to their characteristics and the information known about them, besides their full public availability (as 3DGaze database, depicted in 9.1). The images were of different sizes (the same height, 1080, but different widths: 1278, 1191, 1282, 1279, 1228, 1274, 1286, 1194, 1247, 1192, 1920), all in PNG[2] format, with natural content, synthetic stimuli and a background and some identical objects deliberately displayed at different depth planes. The limitation to models that compute saliency maps representing the level of bottom-up visual interest applied on that work was also valid here. Notwithstanding being considered bottom-up models, some top-down concepts, such as rarity or surprise may naturally be included, as well as the human characteristic of looking primarily for human faces or humanoid things.

This work aimed to examine the quality of images from their respective presented ROIs (defined through binary masks generated from the FDMs). Noise was added to these regions or outside them, so that viewers witnessed their appreciation of the variation in

[2] PNG (Portable Network Graphics) is a patent-free file format for image lossless compression aimed to replace the Unisys© GIF (Graphics Interchange Format). An image in a PNG file can be 10 to 30% more compressed than in GIF format and its image interlacing is faster in developing. It is not possible to make one colour transparent, but to control the degree of transparency (opacity) and there is the possibility of correcting gamma, tuning the image in colour brightness along with display requirements. PNG images can be saved using true colour in addition to the palette and gray-scale formats provided by the GIF, but it can not contain multiple images, not supporting animation, thus [63].

the images quality, thus being noticed the importance of these regions in the subjective analysis of their images.

1.1.1. Factors that affect 3D quality perception

Video systems consist on many components, from the signal acquisition (capturing real frames or generating a virtual scene) to its displaying. This includes converters, codecs, networks, the environment hardware, the signal processing and transmission path, each of them affecting the signal quality perception. Despite the higher complexity, huge bitrate and propensity to perceptual degradations in comparison to analogic, digital signals have higher robustness to noise and interference, efficiency on the encoded signal regeneration, privacy on the information transit, accuracy, and compatibility with digital transmission and media (uniformity in the format of video, audio or data services) [5] [19] [7].

A complex interaction between the HVS components involving the eye and the brain generates the Human perception of a visual scene, which is influenced by spatial fidelity (distinct distortions on parts of the scene) and temporal fidelity (motion' natural-ness and smoothness). Temporal HVS mechanisms can be the source of performance loss, like the number of temporal channels and the models simplicity on describing the HVS motion processing. To quantify errors visibility, some metrics simulate functional properties of early stages of the HVS based on linear or quasilinear operators (using restricted and simplistic stimuli), relying on a large number of assumptions and generalizations. Psychophysical experiments are usually conducted by characterizing some phenomena as superpositions of a few different relatively simple patterns (such as spots, bars, or sinusoidal gratings), much simpler and with less interactions than those in real world images. Thus, inaccurate modelling of the HVS may occur, as some neurons activities which play important role in motion perception are commonly ignored. A limited number of experiments are applied to build models that aim to predict the visual quality of complex-structured natural images [8] [19] [4].

Images digital representation can be converted into physical properties (for example, brightness measured in candela per square meter), but optics and physical structures of the eye and retina usually are not modelled, because the image scanning done by the eyes make it difficult to predict which part of the image is centred on the fovea. Absolute brightness is not very meaningful, as in relation to colour it is not linearly

detected by the HVS (according to its sensitivity peaks in the green-yellow colours, green is brighter, much more clearly seen than blue, while the saturated blue is too dark, almost without contributing to the perceived brightness), so the wide used term Luminance indicates the brightness properly adjusted to what the HVS really sees [19] [7] [10] [22].

Important image effects need to be selected according to the system purpose, as the others are difficult to quantify, not well understood or not essential to the application. It is usual that objective metrics quantize the image dealing with sharpness and colourfulness, attractive characteristics to the viewers (together with well-lit and high contrast), in opposition to pictures with low contrasts, blurred or dark. Even though there are spatiotemporal artefacts, some quality metrics ignore temporal distortions in videos like ghosting (static pixels surrounding moving objects move in the PVS due to temporal low-pass filtering), jitter (variable delay), motion compensation mismatch (background pixels that are static in the SRC but move with the objects in the PVS due to block motion estimation), smearing (such as blur, caused by the relative motion between camera and scene objects), mosquito noise, rods (elongated artefacts like light-rods produced by cameras, which appear in film because of an optical illusion/collusion and are typically traces of a flying insect's wingbeats) and stationary area fluctuations (visual appearance of motion created from temporal frequencies in the PVS not present in the SRC) [7] [23].

The 3D content capture can be done by using single or multi camera techniques[3], holographic devices, pattern projection techniques or time-of-flight techniques and its representation can be dense depth, surface-based, point-based, volumetric, texture mapping, pseudo-3D, light field or object-based. Some of them are internationally standardized, as also as the coding, compression and streaming processes [24].

An ideal camera aperture is described as a point, without lenses to focus light (i.e. no geometric distortions or blurring of unfocused objects) and real cameras characteristics have only discrete image coordinates. The internal and external camera parameters, the 3D scene structure and the relation between them also influence the signal quality. To achieve better ratios on coding or compressing, similarities between video frames or

[3] An example of multi-camera application was developed by Immersive Media: Telemmersion System® provides an end-to-end solution for full motion interactive 360° videos on the workflow stages from capture till distribution, including edition; the Dodeca® 2360 camera system (used on the online mapping industry by Google© on its Street View) captures high-resolution video from every direction simultaneously, supports various video formats and is highly portable [61].

image blocks can be used, but only considering the pixels located at the same correspondent position in the video frames (the mismatch between stereo correspondences), what may conduce to flickering depth perception or visual fatigue: a vertical disparity correction thus calibrates the left and right image planes of the video sensors, rectifying the stereoscopic videos [25].

Besides the shooting conditions (such as lens focal length and camera separation or convergence angle), 3D effects also depend on the system resolution. Beyond those that occur in monoscopic television systems (resolution, picture motion, depth, sharpness), there are also the puppet theatre effect (distorted objects angular retinal size and perceived distance and people looking like animated puppets), the depth resolution (spatial resolution in depth direction), the depth motion (related to the movement smooth play in depth direction) and the cardboard effect (objects seen disjoined or flat as if the scene were divided into discrete depth planes, due to a crude quantization of depth values). Puppet theatre effect is avoided or reduced through orthostereoscopic parallel shooting and display conditions (simulating human viewing angles, enlargement and convergence) or auto-stereoscopic displays that enable a large volume of depth, allowing larger images to be presented at a greater distance behind the screen. Cardboard effect can be avoided or reduced with camera parameters adjustments such that the thickness of objects can be perceived [26] [20].

In auto-stereoscopic displays (lenticular, barrier, etc), the major peculiar factors are frame effect (pictures appear unnatural when objects approach the screen frame: a larger screen is useful then, as viewers become less conscious of the frame) and inconsistency between accommodation (focus point fixed on the screen and pictures displayed within this range) and convergence (image defocusing controlled through the gaze point, minimum value for depth of field). Discrepancies between focus and convergence cause sickness after a few minutes. In auto-stereoscopic displays, 3D effect must be thought since the images creation (sets of images interlaced together, resulting on blurred and double image for a common screen), the lenses must direct the light so that each eye of the viewer sees only one image (and the brain overlaps it to create depth effect, interpreting the 3D scene), the equipment is expensive and depends on the physical position of the audience in few specific spots to have the sense of depth, not seeing both views unmerged. Cameras that track the viewer position avoid sweet spots and the restriction of having a single viewer per evaluation [27] [28].

The distance betweeen each eye is nearly the same among each layer in a glass, thus the eye wear based equipment (stereoscopic) are other specific outfit to display and view 3D video: passive and active glasses synchronized with the monitor. Passive glasses (anaglyph, with polarized lenses) do not use batteries or any electronics, becoming lighter, more comfortable and cheap. The cheapest ones, anaglyph (shutter, with liquid crystal lenses), use a simple technique of generating 3D: two coloured layers are observed through the glasses filter lenses (each eye sees one image layer, usually cyan for the right and red for the left). Glasses with polarized lenses use the light polarization that reaches each eye, the display has a specific configuration (overlapping images subject to mutually orthogonal polarizations filters) and linearly polarized filters are needed to the lenses, orthogonally and with the same orientation of the display filters. Both systems allow multiple simultaneous viewers. The crystal liquid or active glasses need a display with a 120 frames/sec rate: associated with the playback equipment (commonly a computer), the are synchronized glasses can work via wireless technology and, while the display merges the frames, the issuer ensures that each sunglass lens is only open to receive information during the display of one view frame. To avoid the flickering effect since each lens operates at half of the visualization frequency, the display reproduce frames at a fairly high frequency, leading to the equipment high cost [29].

Samsung© frequently use shutter technology in 3D televisions and the 3D passive glasses with polarized lenses of their UE40ES6300 have are said to have a crosstalking problem (images for the left and right eyes overlapping) due to the user position, what leads to the single viewer problem. The distance from the screen is also a problem, as the display is most comfortable at around a metre, with 3D visible from 90 cm to 120 cm back from the screen. A digital palindromic (or forward and back sequence) display method is applied to medical purposes, improving visual detection of low-contrast luminal morphological feature for coronary artery diseases analysis. The stabilization of a coronary segment is reached by digitally shifting each image and the sequence so that a point of interest remains in one position with respect to the observer [30] [31].

Seeking to standardize subjective assessments environment, the ITU defined parameters of general viewing conditions in both lab and home environments, arranged into ratios of luminance, observation angle, background chromaticity and other room illumination and monitor relevant characteristics (like resolution threshold with luminance and contrast, also strongly influenced by the environment luminance). It is also recommended the importance

on results stability of the input signal quality, its source type, still pictures picture by picture adjustment, the downstream processing, the impairments recording accumulated along the chain, and that, for both SDTV and HDTV, the viewing distance and the screen sizes should be selected to satisfy the ratio of viewing distance and picture height (the preferred viewing distance) [17] [8].

Quality metrics classification is addressed on [7] with focus on distortions, expatiating about data metrics (look at the signal fidelity and not to the data), reference information metrics (based on the SRCamount of available information) and picture metrics (treat the data as its visual information):

Data metrics algorithms are designed to describe the data reliability, but not the content. They are fast and straight forward to execute, by simply analysing pixels and their spatial relation or interpreting images and its differences. A widely used signal, the data metric Mean Squared Error (MSE), sums the SRC and an error signal (as its visibility relates directly to the loss of perceptual quality), quantifying the error signal strength (averaging the subtractions of pixels squared intensity. Images with the same MSE may have diverse types of errors, clearly visible or not). [7] [10] [20].

Reference information metrics are non-intrusive (non-referenced, NR) or intrusive (fully, FR, or reduced referenced, RR). FR methods often operate on an error image and do not react well to global shifts in brightness, contrast or colour, can include HVS models and require entirely available and calibrated signals (usually uncompressed SRC) and a strict spatiotemporal alignment to match each pixel with its counterpart in the other image. NR approaches are applied when the SRC is not on hand, focusing on artefacts due to a given transmission scheme. Since there is not a pristine reference 3D stimulus as a baseline for comparison, 3D-based algorithms are NR in nature, without the HVS aspect of learning from experience or alignment issues. Video content or distortions assumptions turn NR measurements unfeasible and may confuse actual content and distortions (e.g. blockiness estimating, the flashiest artefact of block-DCT, Discrete Cosine Transform, based compression methods such as H.26x, MPEG, Moving Picture Experts Group, standard). Used when the SRC is only partially available, RR methods has relevant controllable amount of images descriptors (extracted features, like amount of motion or spatial detail) [19] [9] [4] [7] [5] [16].

As only the extracted features need to be aligned in RR methods, alignment requests are naturally less rigid than for FR metrics (the most inflexible), while for NR methods the

signal quality can be individualized (the SRC is inaccessible in broadcasting and on real-time traffic: low-complexity measurements are needed, without knowing the traffic effect). Thus, methods are out-of-service or in-service, consonant with the usability covering adaptive streaming solutions: without or strictly with time constraints [3] [9].

RR quality models are applied for standardization by the HDTV VQEG (both in laboratory and in operational conditions), bearing in mind the access to the source video bandwidth, withdrawing all NR models. About FR assessment for standard definition television (SDTV), there are the reports in [32] and in [33]. Most existing approaches are FR, such as the subjective methods described in [17], applied for the general picture quality of stereoscopic systems as well as sharpness and depth (not looking at the SRC data features, but only its existence or not, and the categorical judgement method may be used, e.g., to identify the stereoscopic systems merits). Some algorithms are especially useful for NR applications, explicitly designed to detect compression or coding artefacts in a specific type of codec, e.g. blockiness in DCT-based algorithms or ringing in wavelet base [10] [4] [34] [35] [27].

Picture metrics can be classified into vision modelling and engineering approaches. The engineering approach is based on the extraction and analysis of the strength of certain features or artefacts in the video: structural elements (such as contours) or specific distortions (introduced by a processing step, compression technology or transmission link). Psychophysical effects can also be considered, but image analysis rather than fundamental vision modelling is the conceptual basis for their design [7].

The attributes of the evaluation methods classification, some variables or extra components could be added to the ITU recommendations, like psychoperceptual quality evaluation (analysing the relation between physical stimuli and sensory experience, including test sequences, procedures and classification: the overall quality or a certain attribute judged), user-centred quality evaluation or group of factors (relating the evaluation to the system or service potential use), user-centred design (including human and ergonomics subjects/factors, participatory design and design to user experience) and multimedia quality (combining produced and perceived qualities by categorizing multimedia technical factors into levels of abstraction: network, media and content). User-centred methods look at the evaluators of the essential system features as potential users, the potential context of use, evaluation tasks connected to the viewing purpose, or aims to understand the quality including ergonomic measures [29].

Together with the above mentioned viewing environment conditions (viewing distance, ambient lighting, pixel position in the image) and structural information, the quality perception (focusing on fidelity measure, instead of perceived quality) is also affected by the image appeal, the delivery and audio visual quality, the users' preference, experience (the recency effect[4] and the fact that trained viewers are able to spot certain changes more easily), state of mind, cognitive understanding and interaction with the scene (eye movements, instructions received, prior information about the content, attention, fixation, etc.). Idiosynchrasies of viewer attention may lead to a continuous failure risk on predicting the focused points, since the must for quality can vary with the concentration demanded on part(s) of the scene and the significantly reduced human sensitivity outside the focused ROIs. Thus, when challenges to ROIs based assessments like understanding and modelling attention are ignored, distortions are equally weighted over the entire frame [7] [36] [4].

A VA and delivery quality concepts are applied to 3D TV on the Sony© HX850 2012 TV, that selects the correction system along with the type of content (viewers are free to explore the image and sharper objects are valued, like those protruding from the screen and background scenes), while the earlier Sony© HX823 2011 TV slightly blurs the image background, persuading viewers to concentrate on the foreground action. The video purpose is also under debate, in line, for instance, with the pleasure the images give and the artistic choices. When the focus is not on the object chosen to be highlighted in the foreground, eye strain and fatigue may occur, leading to an eventual need of adjustments to sharpen up the various planes [37].

Encoders and compressors can enable simultaneous transmission of multiple frames in just one stream at a reduced bit rate by sampling rate conversion techniques, spatial or domain transformations and digitalizing images, reducing temporal redundancies. Defects produced due to compression depend on the scene and its variations, tending to quality fluctuations during a long program. Irreparable damages to the SRC (like the introduction of nonlinear distortions) may occur and the impact of network losses on video quality directly affects the encoded bit stream and some metrics are based on parameters that can be extracted from the bit stream with no one or only partial decoding. In audio-visual systems, synchronization between different media is done by delaying one of the signals in

[4] The recency effect is related to the major influence of recently-viewed rather than older material in human opinion of a visual sequence: the last 10 to 15s contain the most memorable images of a sequence [8].

relation to the other (the linear combination – signals product – is a effective model of audio-visual quality) [16] [38].

Most works focus on FR metrics for TV/broadcast applications, many remains to be done in the areas of NR and RR quality assessment. Also in quality evaluation of low-bitrate video and transmission error artefacts, in which the development of reliable metrics have many issues still to be solved [7].

Many issues regarding the impact of 3D TV source on HVS are still under discussion, such as the possibility of such technology causing problems on children with epilepsy. It is known that visual patterns can provoke seizures in people susceptible to them, but researchers found no link between this specific technology and seizures. Some scientists believe that video content, not technology is more likely to bring on seizures [39].

1.2. Objectives

The main objective of this work was to check the impact of a relevant ROI representation on the perceived quality of 3D still images. It was used the eye movement images database obtained from an eye tracking experiment described in [1] that handle subjective stimuli (created for the performance evaluation of 3D VA model in [21]), employing the generated and publicly available stereoscopic images of natural content, depicted on the Annex 9.1. Binary masks were generated for each FDM and to each corresponding image was added noise considering the 0 and 255 values as determining the ROI borders. Outputs were classified into images with noise added inside or outside the ROI, according to the noise type (Gaussian or Speckle), the noise parameters values (noise intensity) and to the noisy image view (left or right).

1.3. Outline

This document is structured as follows:

In Chapter 1 there is a brief introduction to subjects covered in this dissertation, encompassing the context and motivation for this work as well as its objectives and this Outline. Concepts and overviews of topics that support the contents are treated here.

Chapter 2 presents a review of more specific literature regarding the various methodologies of video subjective evaluation.

In Chapter 3 it is shortly covered some content on attention models for 3D video and some papers over that topic are reviewed.

The metodology of the work done is depicted on Chapter 4. The objective measurements and subjective validation of the attention models are detailed.

The Chapters 5 and 6 comprise a succint discussion over the results and the conclusions reached from this study.

The References are listed on the Chapter 7 and the Appendix (Chapter 8) comprises the Instructions given to the participants of the subject evaluation procedure.

Finally, the Chapter 9 contains the Annex, with the description of the 3D Gaze eye tracking employed in this work and the display specifications.

2. Literature Review

2.1. Video subjective evaluation methodologies

Subjective assessments, the most reliable method to assess the perceived video quality since the HVS is involved, are slow (in both setup and execution times), demand operational costs and are severely limited in the scope of the video material that can be evaluated. Subjective measurements can be divided in categories, according to the analysis point of view, the way they are implemented and the expected results for each specific test scenario. Then, it is predominantly exploited the correlation between the output of each algorithm and human subjective scores to validate practical objective methods. These experiments involve people with normal vision (according to medical definitions), in a controlled environment, and defined by several standards. There are different correlation parameters to validate the effectiveness of each algorithm such as, for example, between the objective quality evaluation algorithm and human perception of quality about a publicly available dataset of 3D images or videos [7] [9] [5] [16].

The ITU considers that there are two classes of subjective assessments. One of them establishes the system performance under optimum conditions (quality assessment). In this case, an objective metric is used to dynamically monitor and adjust image quality and a subjective metric is applied to determine a quality value for the sequence. The other class considers subjective assessments which establish the ability of systems to retain quality under non-optimum conditions that relate to transmission or emission (impairment assessments). Subjective metrics are used to optimize algorithms and parameter settings of image processing systems, as a result of a comparison between two sequences (a SRC and a PVS), customarily evaluated in a random order. ITU also recommends that the exhaustive descriptions of test configurations, materials, observers, and methods should be provided in all test reports, given the importance of establishing the basis of subjective assessments [17].

Test sequences of different content besides the one to be evaluated must be used, and the whole testing procedure should not exceed 30 minutes. According to [40], the content and duration of the sequences submitted for testing should be representative of

the content and duration of the potential use of the system. The selection of evaluators also does need to obey stipulated rules. The sample of participants must be representative of the system potential users. It is recommended the participation of at least 15 non-expert evaluators, who could be potential users of the system. For stereoscopic television it is required to have normal "stereopsis" observers, hence all participants must be tested about their vision. They also should receive instructions about the assessment method to be used, the types of defects likely to occur, the sequences rating scales, the sequences to be used and the whole timing [17] [27].

Using the Mean Opinion Score (MOS) subjective judgement categorization, each pair of images or short video sequences is assigned a quality score that is normalized, averaged and each standard deviation is calculated, resulting on numerical values from 1 to 5. To obtain a truthful result, many people must participate in the subjective evaluation, since single opinions may widely vary. Quantitative performance measures using objective video quality metrics usualy relate to aspects of their ability to estimate subjective scores. Such performance measures include: prediction accuracy (ability to predict with little error); monotonicity (amount of agreement between predictions and relative magnitudes of subjective quality ratings); and consistency (maintenance degree of the prediction accuracy over the test sequences range, response robustness to impairments). These attributes can be determined by different metrics based on the relationship between the Differential (or Degraded) MOS (DMOS) and predicted DMOS (DMOSp) for subsets within each video quality measurement. The DMOS rates the impairment of the PVS in relation to the SRC, normalizing the result to a a 5-grade scale [32] [41].

Evaluation metrics employ those procedures, such as the correlation coefficient between subjective and objective scores after variance-weighted and nonlinear regression analysis metrics (providing a prediction accuracy evaluation[5]), a Spearman rank-order correlation coefficient between both scores (DMOSp and DMOS, providing a prediction monotonicity measure) and the outlier ratio (a measurement of prediction consistency[6]). [32] [41].

[5] This prediction is done by combining both Pearson linear correlation coefficient between DOSp and DOS plus a test of significance of the difference and between DMOSp and DMOS [41].

[6] I. e., the percentage of the predictions outside the range of 2 times of the standard deviations evaluates an objective model's ability to provide consistently accurate predictions for every video sequence and not fail excessively for a subset of sequences. This prediction consistency is the number of predictions outlier

Due to the absence of subjective assessment methods specifically designed and standardized to 3D content, 2D ones are used as basis (consonant the ITU in [17], divided along with the test material presentation): Double Stimulus Impairment Scale (DSIS, the EBU - European Broadcasting Union - method), Double Stimulus Continuous Quality Scale (DSCQS), Single Stimulus Continuous Quality Evaluation (SSCQE), Single Stimulus (SS) methods and Stimulus Comparison (SC) methods. In [40], other methods are described: the Absolute Category Rating (ACR), the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE, also addressed in [17]) and Subjective Assessment Methodology for Video Quality (SAMVIQ).

Five-grade scale	
Quality	**Impairment**
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

Figure 2.1: ITU-R quality and impairment scales [17].

For SS, SC and DSIS, the commonly used rating scales such as verbal or numerical categories are given in Figure 2.1. The subjects evaluate the perceived image quality, the impairment, or the relation between two stereoscopic images by placing the presented stimuli in one of these categories [26].

In Single Stimulus **(**SS) methods, home viewing conditions are replicated: the observer does not have a SRC to compare with the presented sequence to give his judgement, but watch each piece of video displayed and then sort it according to a provided scale. A test sequence (a single image or a sequence of images) is presented at a time, being each one independent from the others and a quality index for the entire presentation is provided by the assessor. In the Single Stimulus with Multiple Repetitions, SSMR, the test material is randomly presented three times, allowing stabilizing the observer's opinion, being the first rating discarded from further analysis. The test material might include only test sequences or both the test sequences and their corresponding SRC, presented as a freestanding stimulus for rating like any other test stimulus. The content to be analysed is selected and then the test images are prepared in accordance with

points (with an error greater than some threshold as a fraction of the total of points: the smaller this fraction, the more consistent are the predictions) after the nonlinear mapping [32] [41].

design options or ranges of factors to be evaluated. When more than one factor is considered, each image may represent one level of a single factor or one level of every examined factor, in which case it is possible to detect interactions among factors (non-additive effects). Both situations allow establishing a relation between the results and specific factors [41] [17].

SS methods can also be classified according to the evaluation procedure. The evaluator perception score is obtained by answers to the panel, as, for example, the identification of a certain impairment factor in the scene. The analysis of results is based on central tendency and dispersion of the response speed and accuracy characteristics. Rating of the test material can be done either in a post-presentation manner, or in a continuous rating. By employing continuous quality evaluation, longer sequences can be presented and evaluated. These are more representative of realistic video content and error statistics [38] [41].

Four types of SS methods have been used in television assessments: adjectival categorical judgement, numerical categorical judgement, non-categorical judgement and performance methods. Some of them are also used in SC methods.

The ACR method is useful for evaluating the end result of a service quality, considering that the content presentation is similar to the common use of the systems. It is obtained through the direct observation of the sample (test sequences are presented consecutively, one at a time, and independently on a category scale), without a SRC, to view the absolute quality values of the sample. The test sequences are observed one after another, and judged after each presentation, according to a five-point discrete scale, or more detailed (from 9 to 11 points) if necessary, under the subjective rating system MOS [29] [9] [42].

When large distortions occur and they shall be compared to near lossless scenarios, a single stimulus test saves a lot of time. However, small differences between sequences cannot be evaluated with ACR at all because the viewer cannot remember other sequences so precisely and the voting scale is too coarse. Hence, this method is not suitable for very small distortions. In the case of transmission distortions or strong coding artefacts however, this is a very good method to assess many sequences in a short time [10] [41].

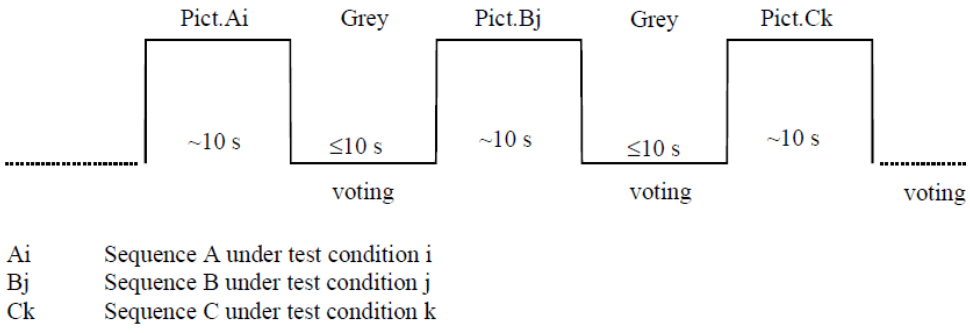


Figure 2.1: Stimulus presentation in the ACR method [40].

The time pattern for the stimulus presentation is illustrated on Figure 2.1. If a constant voting time is used (e.g., multiple viewers simultaneously), then it should be 10s or less. The presentation time may be determined according to the content of the test material. If higher discriminative power is required, a nine-level scale may be used. For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test. The ACR scale with hidden reference removal method (ACR-HRR) is an example of applying post presentation assessment, where the SRC is included in the session without the knowledge of the viewers. It is made a subtraction operation between their opinion of the SRC and of each PVS, resulting in a difference MOS [40] [10] [41].

The SSCQE technique considers scene-dependent and time-varying impairments, quality variation related to scene content (impairments may be very short-lived), the ineptitude of conventional ITU-R methods alone to assess these materials, and disablility of the Double Stimulus (DS) methods of laboratory testing on replicating the single stimulus home viewing condition. The material is viewed once, without a SRC, in the context of 3DTV, to assess humans' sense of presence, depth, and naturalness. A single quality measurement is obtained from the relation between the continuous assessment of a coded sequence and an overall single quality rating of the same segment. This method is proposed for long stereoscopic sequences (from 60s to 20min), covering the selection of meaningful test material (often limited to 10s) and mimicking home viewing conditions, since artefacts are dependent on spatial and temporal content. It is possible to increase the sampling rate of the subjective quality ratings (viewers can dynamically rate a sequence using a slider mechanism with an associated quality scale: useful for tracking rapid changes in quality and, thus, more useful for evaluating real-time quality monitoring systems) [16] [17] [41] [26] [38] [43].

ITU recommends that an electronic recording handset connected to a computer should be the recording device used to the continuous overall quality assessment. The sequences may have 5 min at least, each one with different quality parameters under evaluation. The instantaneous quality is assessed in real time according to the variable position of a hand-held slider. The viewer impressions are transmitted by moving the cursor proportionally to the perceived sensation. Image quality is indicated in a range from excellent (slider at the top of the grading scale) to bad (slider at the bottom of the scale) values, which are recorded at two samples per second rate and the data obtained allow the collection of histograms. The quality ratings have regular time intervals, capturing, thus, the perceived time variations in quality [17] [44] [38] [26] [43].
The procedure described in the ITU recommendations are adapted to the equipment in use, as in [43], where adjustments are done in computer-based testing. The slider for the SSCQE test was not a stand-alone hardware device, but a graphical on-screen slider steered by moving the mouse up and down (vertical mouse movements translated directly into slider shifts), aiming to give viewers a good tactile feeling of the quality scale, considering the additional advantage of the familiarity with handling a computer mouse. A break time is suggested to help reduce fatigue, as well as a random order of the test sequences at the clip level to minimize contextual effects [43].

In Stimulus comparison (SC) methods, images or sequences of images created in the same manner as in SS methods are displayed simultaneously and the viewer score is an index of the relation between them. Two equally adjusted displays or the major proportions of a divided screen can be used. The test session initiates with the observers' opinion stabilization, which is not useful data to the test results. Three types of SC methods are used in television evaluations: adjectival categorical judgement, non-categorical judgement and performance methods [17] [38].

Comparison scale	
-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
1	Slightly better
2	Better
3	Much better

Figure 2.2: Comparison table in DSCS Methods [17].

The Adjectival Categorical Judgement methods, also known as Double Stimulus Comparison Scale (DSCS), can be applied as SC or as SS method. There is a judgements distribution, depending on the information required or sought, across scale categories for each attribute detection (e.g. to establish the impairment threshold), in SS cases, or, in SC cases, the relation between members of a pair to one of a set of categories defined in semantic terms, expressing the existence and direction of perceived differences. In operational monitoring, half of those grades can be used, and, in special cases, scales that assess text legibility, reading effort or image usefulness. The ITU-R scales are given in Figure 2.2 [17] [38] [44].

The Single Stimulus Numerical Categorical Scale (SSNCS) describes fast and easily automated judgement categories, not limited in value by adjectives, alowing the linear scale to be used for quite different ranges, with a quantity of points dependent on the conditions and range of perceptual attributes. 5 points may be enough, or half points on a scale of 10 may be useful. The number of trials per condition depends entirely on the assessment purpose, but the ITU defines that at least 3 are needed for statistical control [45].

In non-categorical judgments methods, the ratio of perceived quality is given by a number, a point or line in a vertical scale (a continuous variant of the categorical method, or numerical, a discrete method) labelled with two boundaries at the ends of this scale (e.g., same-different ends or the ends of a categorical scale). As in DSCS, they can be applied in the context of both SC and SS methods. Observers attribute a value to each image or image sequence shown, in SS cases, or to the relation between the elements of an assessment pair, in SC cases [38].

Using a continuous scale, the evaluator assigns each image or image sequence (SS) or each connection (SC) to a point on a line drawn between the labels. The scale may include additional reference labels at intermediate points. The distance from one end of the scale line is taken as the index for each condition (SS) or as the value for each condition pair (SC) [17].

In discrete scale form, the evaluator assigns each image or image sequence (SS) or each connection (SC) a number that reflects its judged quality level in a specified dimension (e.g., image sharpness or the difference in quality). The range of numbers used may be limited, either previously defined or not. The assigned number may describe the relation in absolute terms (without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation) or in terms of a standard pair [17].

Both forms, continuous and discrete, result in a distribution of values for each condition (SS) or pair of conditions (SC). The method of analysis depends on the nature of the judgement and the information required (e.g. ranks, central tendency, differences between values) [38] [17].

The performance measure of externally driven tasks (finding targeted information, reading texts, identifying objects, etc.) may be used as an image index, because those tasks express some aspects of the HVS. Thus, in SS cases, performance methods result in distributions of accuracy or speed scores for each condition (indices of the relation between the pair frames). The analysis concentrates upon establishing relations among conditions in the scores central tendency (and dispersion) and often uses variance analysis or a similar technique. In some cases, performance measures can also be derived from SC procedures. In the forced-choice method, the pair is prepared such that the PVS contains a particular level of an attribute (e.g. impairment) while the other frame contains either a different level or none of the attribute. The observer decides which element contains the greater or lesser level or nothing of the attribute [17].

Double Stimulus (DS) methods are especially useful when it is not possible to provide test conditions that exhibit the full range of quality parameters. The SRC and PVS are presented side-by-side rather than random time order and shown simultaneously on the same splited monitor or on two monitors equally adjusted. Scoring is done through comparison. Like in SS methods, this can be done either by post-presentation or continuous assessment, and sequences can either be presented once or multiple times. The differences between them are evaluated and the fidelity of the video is judged through a lever sensor mechanism. Perfect fidelity is coded as 100, at a maximum length, while at the bottom the scale is 0. The data set can generate a number of statistic graphs, but the attention must be shifted between both presentations. Examples of these methods are the DSIS, the DSCQS and the SDSCE [44] [38] [17] [41] [29].

Degradation Category Rating (DCR), or DSIS, is used when there are high quality samples and the ACR method is inappropriate to discover quality variations. In this case the DCR is able to measure the degree of perceived impairment rather than the perceived quality and the stability of the results is greater for small impairments than for large ones. It was proposed to identify defects in transmission paths (system robustness measure) at very low bit rate and it is used on the evaluation of new systems or of the fidelity of visual information affected by time-varying degradation, such as the effect of sparse impairments (like transmission errors) [16] [17] [40].

The test sequences are presented in pairs: the first stimulus presented in each pair is the SRC and the second is the outcome when the SRC pass through one of the systems under test. During the test session, a pseudo-random sequence of pictures and impairments (which range is chosen so that all grades are used by the majority of the observers) should be presented, and the same test picture or sequences should never be presented on two successive occasions with the same or different levels of impairment. The test results are the DMOS between the PVS and the SRC. Like ACR, in DCR the presentation time may be influenced by the content of the test material. A number of replications is obtained for the DCR method by repeating the same test conditions at different points of time along the test. A five-level rating scale is used for the impairment rating and the numerical value paired with each grade on these scales indicates the mapping between category rating and MOS, these values are not presented to the viewer. There are also the numerical categorical rating (in which both category and numerical values are presented) and the non-categorical judgement (in which only a numerical value is used for assessment) scales [42] [40] [9] [17].

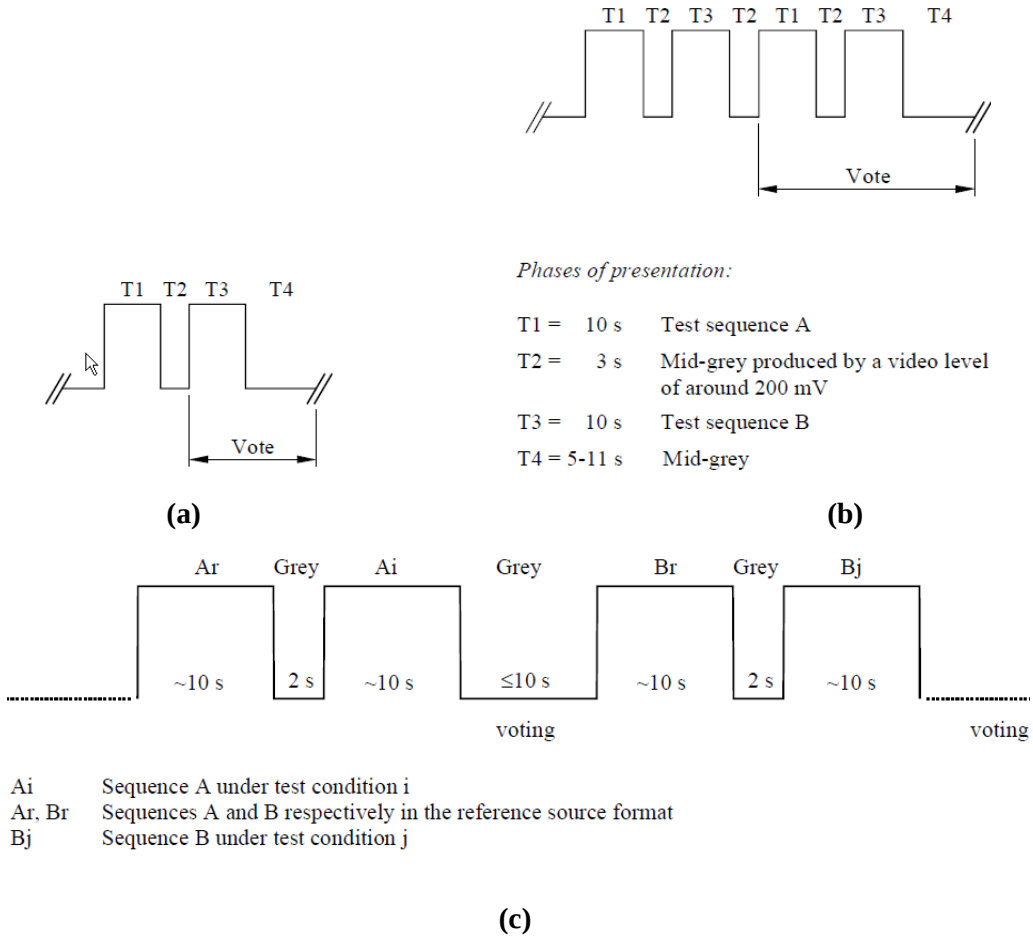


Figure 2.3: DSIS method structure: (a) Variant I, (b) Variant II (c) pattern for the stimulus presentation [40] [17].

The sequences are presented in a Double Stimulus, as shown in Figure 2.3. One out of two variants can be used, Variant I or II. The first gives a unique presentation of SRC and a single image to be evaluated and the viewer can switch between sequences until being satisfied with a mental measure of the quality associated with each signal. Typically this is chosen by doing that two or three times for periods of up to 10s. In variant II both presentations are repeated before the trial, one or more times for the same time duration. The test scheme set provides a recovery and rest time during the transition (grey image, between the SRC and the test picture), and a final time for the trial. The observers should look at the picture throughout the duration of T1 and T3. Extending the periods T1 and T3 beyond 10s does not improve the ability of the viewer to grade the sequences. The evaluation is done during T3 and T4, according to the artefact impact. For still pictures, a sequence of 3 to 4s with five repetitions (voting during the last two) may be appropriate. For moving pictures with time-varying artefacts it is recommended a 10s sequence with two repetitions (voting during the second). When practical considerations limit the available sequences duration to less than 10s, the display time can be extended to 10s, by using them as compositions segments [38] [17] [29] [3].

The judgement of real service situations is not representative with sequences limited to 10s. Digital artefacts are strongly dependent upon the spatiotemporal content of the SRC and for compression schemes it also concerns the error resilience behaviour of digital transmission schemes. In [45], when there is no SRC, a preference for SSNCS to graphic and ratio scales is verified, in terms of sensitivity and stability. SSCQE is useful for viewing conditions closely to real situations (quality measured continuously, with the material viewed once, without a SRC). When the drawback of a method is linked to the existence of context-related artefacts there is a problem to choose representative sequences (or at least to assess their representativeness) and, when fidelity has to be evaluated, reference conditions must be introduced. In this case, a continuous evaluation from the SSCQE should be developed, considering also that human memory can significantly influence test results (the recency effect), by making slight deviations concerning the way of presenting the images to the subjects and concerning the rating scale. Short segments are used when the trial is done after the presentation and long segments are used during its implementation [17] [8] [38] [16].

DSCQS is a cyclic method such as DSIS, within each pair of sequences, one is the SRC and the other is the PVS, the same sequence, modified by a system or process under

test. In sessions which last up to half an hour, a randomly arranged series of picture pairs (internally random) with random impairments cover all required combinations, including the SRC, preventing pre judgements. At the end of the sessions, scores are converted to a normalized range and the mean scores for each test condition and picture are calculated. The end result indicates the relative quality of both sequences (quality score). The presentations structure is the same as in Variant II of DSIS, showed in Figure 2.3(b) [16] [17] [8].

The Subjective Assessment Methodology for VIdeo Quality (SAMVIQ) is a multi-stimuli continuous quality scale method with three versions of the SRC: an explicit, a processed and a hidden (which participants must find). A possible lower anchor sequence with intentionally very bad quality in each set allows for a calibration between different sets with the SRC and also to compare different subjects, but it limits the number of conditions in a set. The PVS of one SRC can be shown in two or three sets, thus each of them containing seven evaluated PVS and the lower anchor [29] [46] [10].
Each version of a sequence is displayed and rated, similarly as in DSCQS method: the different versions are selected randomly through a computer graphic interface, moving a slider on a continuous scale graded from 0 to 100 annotated by 5 explicit quality items linearly arranged (excellent, good, fair, poor, bad). A 10 to 15s maximum viewing duration gets a stabilized reliable quality score. The proprietary decoder-players, or a screen copy of their output, should be used to maintain the appropriate display performance. Once all the test set votes are adjusted, the next ones are available and can be from a different SRC. From one scene to another, the corresponding access from an identical button is randomized. It is always possible a comparison to other PVS of this set or to the SRC, permiting very small distortions in both situations and a high degree of resolution in the grades. Thus, using an implicit comparison process, it is possible to combine quality evaluation capabilities and also to discriminate similar levels of quality. This is a proper method for multimedia context, since it can combine different elements of image processing (codec type, image format, bit-rate, temporal updating, zooming), extensible to cover full format television environment [29] [10] [47] [46].
The stability of this protocol allows conducting experiments in a reliable way. Variation of criticality during an audio visual content scene is limited as homogeneous contents are chosen under the same rules implicitly used by methods providing a global score (e.g. SS methods): it is then functionally similar to a SS method with random access,

but the explicit reference can always be viewed. Various stimuli can be accessed all at once, at any time, reviewed and modified as desired (commonly less than half of the sequences are reviewed and stimuli with 15s at most are enough). Thus, it is a time consuming method (the time for voting is not fixed), so a limited number of tests can be done [47] [46] [29] [10].

SAMVIQ is a methodology very similar to a widely used test for audio samples, MUSHRA (Multiple Stimuli with Hidden Reference and Anchor[7]), recommended in [48], in which a continuous scale is applied on the judgement of the quality of each item from a set of PVS from the same SRC. In [49], it is proposed the Subjective Evaluation of Stereo VIdeo Quality (SESVIQ), an interactive method to evaluate perceptual quality for asymmetric coding of 3D video as a stereo extension of SAMVIQ where subjective evaluation is presented for low and high bit rates, verifying that spatial scaling is always inferior at high bitrates, whereas subjective results depend on video characteristics for low bitrate scenarios. So, SESVIQ test methodology gives lower standard deviations than DSCQS, especially for low bitrate scenarios [10] [49].

Comparison between subjective methods for quality assessment

Several methods were developed for the subjective test of video sequences, especially by ITU-R and ITU-T. They differ significantly and should be carefully chosen to accomplish the desired task, according to the distortions severity, evaluation type, estimated suitability for different distortion classes and relative time necessary for each sequence judgement. Some test characteristics must also be discussed: DS or SS, using a discrete or a continuous quality rating scale and single or continuous voting [10].

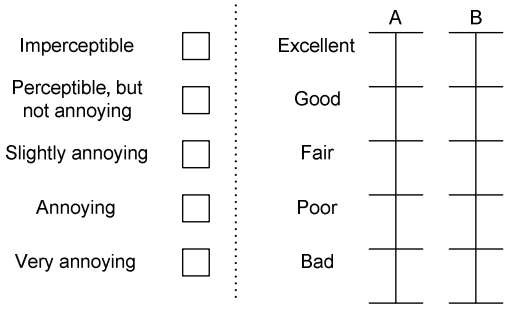


Figure 2.4: Rating scales of DSCQS and DSIS [16].

[7] MUSHRA is a subjective quality evaluation method for lossy audio compression algorithms to assess intermediate impairments, that serves a similar purpose of MOS, with fewer participants and all codecs presented at the same time, on the same samples. A 0-100 scale turns possible very small differences to be rated. There are a labeled reference and its hidden version, the test samples and one or more anchors to normalize the scale (minor artefacts must not be rated as having very bad quality).

SSCQE method can be used to assess widely time-varying quality of long video sequences in a way DSCQS cannot. The accuracy of SSCQE compared to DSCQS is questioned in [44], because since only the quality of a single video stream (SS with no immediate SRC) is seen and rated, contextual effects may be present. In [17] it is said that, for the DSIS and SC methods, contextual effects are also evident and the strongest effect is found for the Variant II. In ACR, SSCQE and SDSCE methods, individual scores might drift along the test, badly impacting the reliability as well as the accuracy (due to differences in each viewer time to react to quality changes). DSCQS has the same structure as DSIS, but not so dependent on classifications (due to the test sequences order): a 5 points scale from bad to excellent, as it can be seen on Figure 2.4 [16].

SSCQE differs from ACR in terms of the assessment process and the scale used. There is no continuous sequential presentation of items in SAMVIQ as in DSCQS, reducing errors due to lack of concentration, thus with higher results reliability. In [3] a comparison is done emphasizing that unlike DSCQS, in DSIS the evaluators are aware of the presentation sequence and each sequence is showed only once [38] [47].

Memory-based biases can exist in longer single rating DSCQS sessions of digitally-coded video, but such effects are not considerable for a 10s video. Consequently, the quality histogram could be calibrated in SSCQE process using DSCQS on representative 10s samples extracted from the histogram data. Thus, this continuous scale is used at DSCQS, but there is a single grade at the end of a short presentation, while at SSCQE the grades are continuous along the demonstration time [17] [38].

In Figure 2.5 some parameter differences of the methods ACR, SSCQE, SDSCE, DSIS, DSCQS and SAMVIQ are depicted. SSCQE with hidden reference removal (that replicates DS testing results) and multiple randomized viewer orderings (at least two) can produce quality estimates comparable to DSCQS and DSCS. For this translation it can be compared the last SSCQE time sample of the scene with the DSCQS or DSCS value. This reassuring result shows that viewers perform essentially the same error pooling function (i.e., the judgment process where perceived errors distributed in space and time are mapped to overall estimates of perceived quality) in SSCQE, DSCQS, and DSCS tests [44].

Parameter	ACR	SSCQE	SDSCE	DSIS	DSCQS	SAMVIQ
Comparison	Single Stimulus	Single Stimulus	Double Stimulus	Double Stimulus	Double Stimulus	Multi Stimulus
Explicit reference	No	No	Yes	Yes	No	Yes
Hidden reference	No	No	No	No	Yes	Yes
High anchor	No	No	No	No	Yes	Hidden reference
Low anchor	No	No	No	No	Yes	Yes
Scale	5 point discrete scale (higher if required)	5 point continuous scale	5 point continuous scale	5 grade impairment scale	5 point continuous scale	5 point continuous scale
Sequence length	10s	Long stimuli (>60s) up to 20min	10s	10s	10s	Maximum 15s
Picture format	All	All	All	All	All	All
Two simultaneous stimuli	No	No	Yes	No	No	No
Presentation of test material	Once	Once	Once	I: Once II: Twice in succession	Twice in succession	Several concurrent (multi-stimuli)
Voting	Only test sequence	Test sequences	Difference between test sequence and reference shown simultaneously	Only test sequence	Test sequence and reference	Test sequences and reference
Possibility to change the vote before proceeding	No	No	No	No	No	Yes
Continuous quality evaluation	No	Yes (moving slider in a continuous way)	Yes (moving slider in a continuous way)	No	No	No
Minimum accepted votes	15	15	15	15	15	15
Assessors per display	One or more	One or more	One or more	One or more	One or more	One
Display	Mainly TV	Mainly TV	Mainly TV	Mainly TV	Mainly TV	PC*
Rating moment	Retrospective	Continuous	Continuous	Retrospective	Retrospective	Retrospective (rating can be adapted several times)

***SAMVIQ can also be applied to standard television displays, rather than solely PC displays.**

Figure 2.5: Comparison between methods.

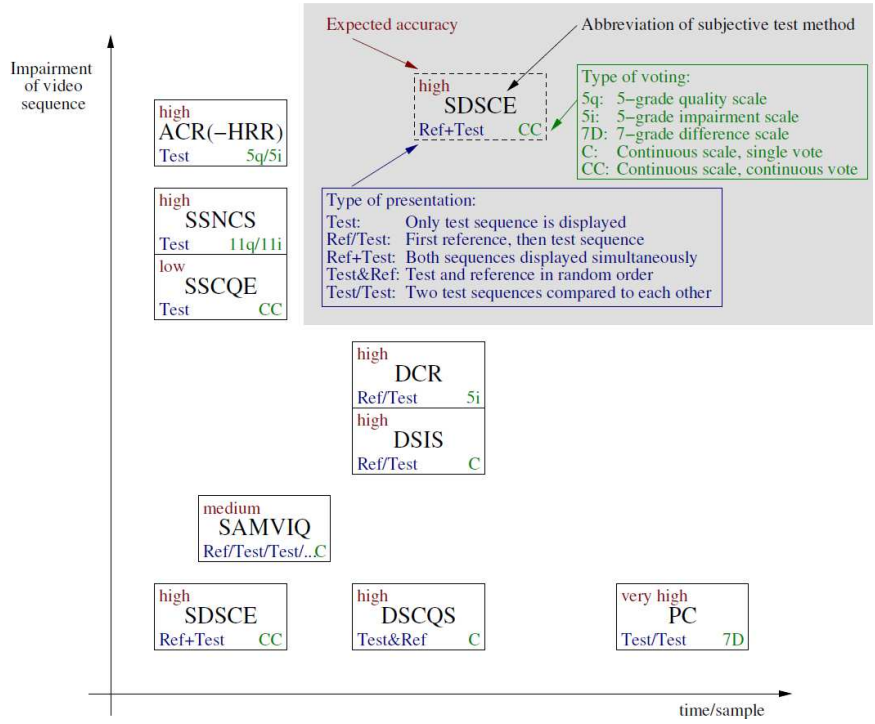


Figure 2.6: Commonly used subjective test methods [10].

Properly designed SSCQE testing (suggested 9 to 15s) may be an effective substitute for more complicated DSCQS testing, because it provides less viewer fatigue and faster testing (or more clips rated for the same amount of viewing time spent). In Figure 2.6 it can be seen a graphic comparing some of those methods impairment versus their time per sample relation [44].

Pair Comparison (PC) tests are the most time consuming but also the most accurate tests very well suited for small impairments, but each rating only compares two conditions each other, then for large differences (as multimedia material) is not very helpful. DSCQS can also be used for small impairments, being much faster but usually less accurate than SC. Larger impairments can be tested with DSCQS, but the comparison to the SRCwould not help the viewer anymore. SAMVIQ allows evaluating very good video quality and very poor quality together without losing the discrimination accuracy when comparing two video sequences with similar distortions [10].

This page was intentionally left blank

3. Attention models for 3D video

Mostly computational models of attention follows the general concept of Visual Attention (VA): founded on prominent cues (saliency-based), a parallel stage with preattentive cues preceeds a serial multi-resolution feature integration principle stage (with typical camera images as input) to predict salient 2D images areas. Consistently with psychophysical studies, the saliency map is frequently used, with parameters like constrast and colour, as representative of the image information that stand out on human appreciation of the reality. However, to consider the 3D realm it is necessary to extend the efforts in 2D visual modeling about what affects the human viewing behavior, including parameters that are representative of the additional visual dimension, among which the most fundamental source for VA is the scene depth. Thus, several approaches are possible to find relationships between parameters and analysis of each one in particular. As a confirmation of the outputs, subjective tests identify the viewers' visual emphasis and quality analysis as the parameters are changed.

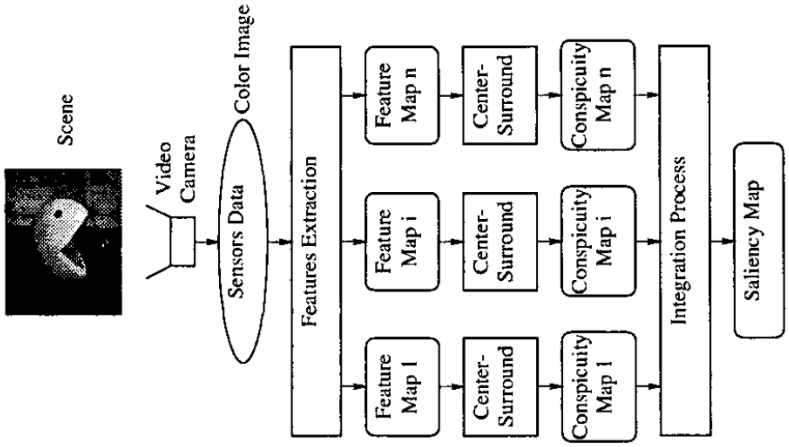


Figure 3.1: A 2D computational model of attention [11].

As it can be seen in Figure 3.1, on a 2D saliency-based computational VA model the scene features are extracted (usually intensity, colour and intensity gradient components), leading to a multi-feature version of the scene (converting each feature map into its conspicuity map) and a integration (in a competitive purely data-driven way: it is a Data Metric) of the conspicuity maps (weighted so that there is few strong peaks of activity and none maps with several comparable peak responses).

Since saliency maps are effective at predicting main subjects but less effective with objects of secondary importance and the unimportant ones, a timing data, considering a multi-stage predictor might be more effective than single-stage attempts. This is in line

with the user opinion related to each object in the scene: main subjects being determined first and this knowledge of them being then used to guide subsequent predictions [18].

The multi-resolution view 2D VA computational model considers the variety of objects sizes to be detected, using a multi-scale conspicuity operator. One of the biologically plausible methods to compute the conspicuity maps is the Centre Surround Mechanism (considering differences of scene parts according to a specific feature), which has a high computational cost due to the application of variable size centre-surround filter on fixed images.

The most noticeable differences between 3D models rely on the 2D model basis, the applied meaning of depth and how it is computed (e.g., equal to disparity, as epipolar geometry assures[8], or to the distance between the camera and the scene object, as in [11] and [50]) and the validating process (subjective tests performed to correlate with the algorithm output, and how the sequences quality values are determined). Every computational model of 3D VA contains a stage in which 2D visual features are extracted and used to compute 2D saliency maps. Consistent with the way depth information is used, a model can be classified as Depth-weighting, Depth-saliency or, the most unusual, Stereo vision.

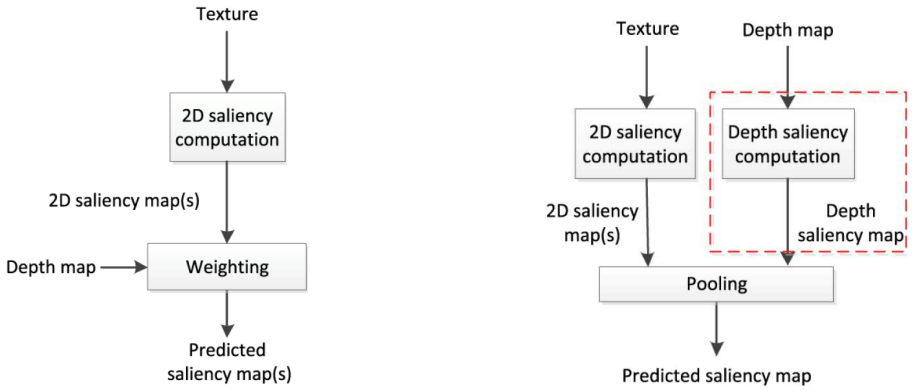


Figure 3.2: Respective schemes of the depth-weighting and depth-saliency VA models [21].

[8]Stereo depth estimation recovers depth from images, and epipolar geometry (the geometry of stereo vision) deals with limits of geometric relations among 3D points and their projections on the 2D space (one *epipolar* ray in each image, as a two cameras scheme is mostly used, due to HVS affinities), relating cameras (slightly shifted each other) and estimating disparity. Cameras are assumed as pinholes, as those relations depends on their parameters (when they are known, coordinates can be normalized and one image turned into global reference) but not on the 3D structure. Thus, depth is the inverse of disparity when epipolar lines are scanlines, cameras has the image planes parallel to each other and to the baseline, their centres are at same height and focal lengths are the same. Constraints are the uniqueness (for any point in a image, at most one matching point in another image), ordering (analogous points in the same order) and smoothness (disparity values changing, for the most part, slowly) [62].

On depth-weighting models, 2D saliency is weighted by depth information, both corresponding to the same (pixel, target, depth plane, ROI) location. This is relatively easy to adopt existing 2D models as there is not a depth saliency map creation, as it occurs in depth-saliency models (through the extraction of depth features from the depth maps), and this is the main difference between them, as it can be seen on Figure 3.2. The computationaly less complex depth-weighting models might fail to detect salient areas caused only by depth features. Differently, Stereo vision models consider mechanisms for stereoscopic perception in the HVS. Some of these methods are depicted below, and the model described in [11] is mathematically explained as an example [21].

Stereoscopic VA model for 3D video [50]

The bottom-up stereoscopic VA depth-weighting model proposed in [50] considers the fact that, due to user interactive functionalities and stereoscopic perception provided by 3D video systems (focusing on intensity, orientation and colour), humans are more interested in regions that pop-out or with small depth value. It is based on multiple perceptual stimuli including depth information, luminance, colour, orientation and motion contrast with a depth based dynamic fusion integrating them, maintaining high robustness and efficiently simulating stereoscopic VA for HVS, modeling it by three attributes with low-level features, including depth, image saliency and motion saliency [50].

A computational model of depth-based attention [51]

This depth-weighting computational model for attention integrates disparity, image flow and motion cues, so that the attention stays on the closest moving object (a moving viewer may mask out different moving objects in real scenes). The preattentive cues employed on early modules are independent from each other. The stereo disparity module applies a phase-based algorithm considering disparity as the spatial shift of the phase (convolution argument of a complex filter with the images) difference. Disparity selection is composed of histogramming and disparity prediction and its back projection produces a target mask (slicing up the input image corresponding part and the scene consonant relative-depth), thus relative depth is found from the dense disparity map. The image flow module stabilizes the attentional performance by applying the stereo algorithm not to a stereo image pair but to consecutive image frames, obtaining

horizontal image flow data, which is used in one dimension along the horizontal direction (to share identical input with the depth module). Depth and image flow cues are combined to deal with complex scenes with multiple target candidates. The motion detection module computes (with a moving target relatively small compared to the background) an affine fit between consecutive images and, alongside that, exploits the brightness constancy restriction (database limitation imposed): considering the background small variations in depth and its distance relative to the motion, the background cancels in the residual image and moving objects appear. Thus, using multiple cues, relative depth as a target selection criterion and simple computations, this model provides expected results for a control scheme for target selection based on nearness and motion [51].

Computing VA from scene depth [11]

This multi-resolution view, bottom-up, task-independent, saliency-based computational model of VA integrates depth as a component for pure image vision. Comparing synthetic and real range images show a selected depth-related features ranking: the distance from camera to the scene objects (here called Depth); the scene objects geometry (provided by an intrinsic surface feature, here called Mean curvature); and depth changes in the scene, like angles and corners (detected by the Depth gradient vector). The 2D model is extended to the scene depth component: depth-related conspicuity maps are created and a multi-scale conspicuity operator selects an integration module that combines and adds it to the original features map [11].

The conspicuity maps are created by calculating the difference between fine and coarse scales from a group created using Gaussian pyramids, which gradually lowpass filter and subsample the feature map, extracting local activities for each feature type. The centre is a pixel at scale $c \in \{c_1, c_2, \dots, c_N\}$ and the surround is the corresponding pixel at scale $s = c + \delta$. $\delta \in \{c_1, c_2, \dots, c_{N-1}\}$. Thus, for each pyramid P (consequently, for each feature), there are $N \times (N-1)$ maps $F(c,s) = |P(c) - P(s)|$ forming a unique conspicuity map, weighted considering the maximum activity of the related map, M, and the average of local maxima, $\bar{m}$.

$$\textbf{Eq. 3.1: } \mathbf{C_j} = \sum_{\mathbf{j=1}}^{\mathbf{N\times(N-1)}} \mathbf{w_j} \times \mathbf{F_j(c,s)}, \qquad \mathbf{w} = (\mathbf{M} - \bar{\mathbf{m}})^{\mathbf{2}}$$

where N and n are the feature numbers for depth and for colour and w and c are the relative weight and conspicuity map [11].

Thus, the depth target mask (the depth conspicuity map) is computed based on histogramming and peaks observed on the histogram of the disparity map are considered as conspicuous locations. Their weighted sum (seeing the maximum activity of the related map, M, and the average of local maxima, $\overline{m}$, considering both depth and colour) leads to the saliency map [11]. As the depth enhanced model detects noticeable locations, it is significant (and equal to colour features, in the saliency map) the contribution of depth features to the VA, besides the depth or 3D vision intrinsic components of biological vision, at an early stage in the HVS [11].

Computational model of stereoscopic 3D visual saliency [21]

This is a computational model of VA for stereoscopic 3D still images which, as various similar models, through extensions of the effort made in 2D visual modeling, adds as a visual dimension input the prediction of salient areas of 2D images and depth information affecting the human viewing behaviour [21]. A depth-saliency model is chosen to be applied, after proposal and examination of this way of integrating depth information in the modeling of 3D VA and depth-weightning one [21].

The eye movement data obtained from an eye-tracking experiment using synthetic stimuli is used on measuring depth saliency [21]. The measure of depth saliency is derived from the eye movement data obtained from a publicy available eye-tracking database with stereoscopic images of natural content using synthetic stimuli [21].

Across different types of scene, there is a variation on the depth saliency map performance and on its added value to a 2D model. At this work, depth saliency map is defined as a distribution of probability of the fixation points as a function of depth features. It is assumed that the depth map computation needs to be the first step of modeling 3D VA and that the disparity map (which usually represents depth information in stereoscopic 3D display systems, showing the parallax of each pixel between the image views) alone do not exactly correspond to the depth (the same disparity value also may meet with differents perceived depth values). A transformation from a disparity map (measured in unit of pixels for display systems) is therefore added to a depth map (perceived depth in unit of length), considering also the viewing distance between observer and screen plane, the interocular distance (set to 6.3 cm) and the width and horizontal resolution of the screen (set according to the eye-tracking experiment setup). For depth-weighting models, the resulting depth map is adopted as depth data for the weighting or the saliency maps depth-based merging.

It is said that this theme approach is advantageous because each scene can have a precise depth map (controling background objects depth). The influences of 2D features can be limited on viewing behavior (as all objects are uniformly located, with constant shape, size and distance from the center of the screen, so that the stimuli minimizes bottom-up VA features) and for binocular[9] depth cues when coming from depth cues other than disparity (it is not easy to quantitatively measure for monocular ones, e.g., perspective, occlusion, blur). The only depth feature exploited is the contrast, a saliency cue that directs the attention about 3D still images. More features (surface curvature, depth gradient, orientation contrast, etc) would lead to the need of a more complex pooling strategy and a potential normalization step for each feature dimension.

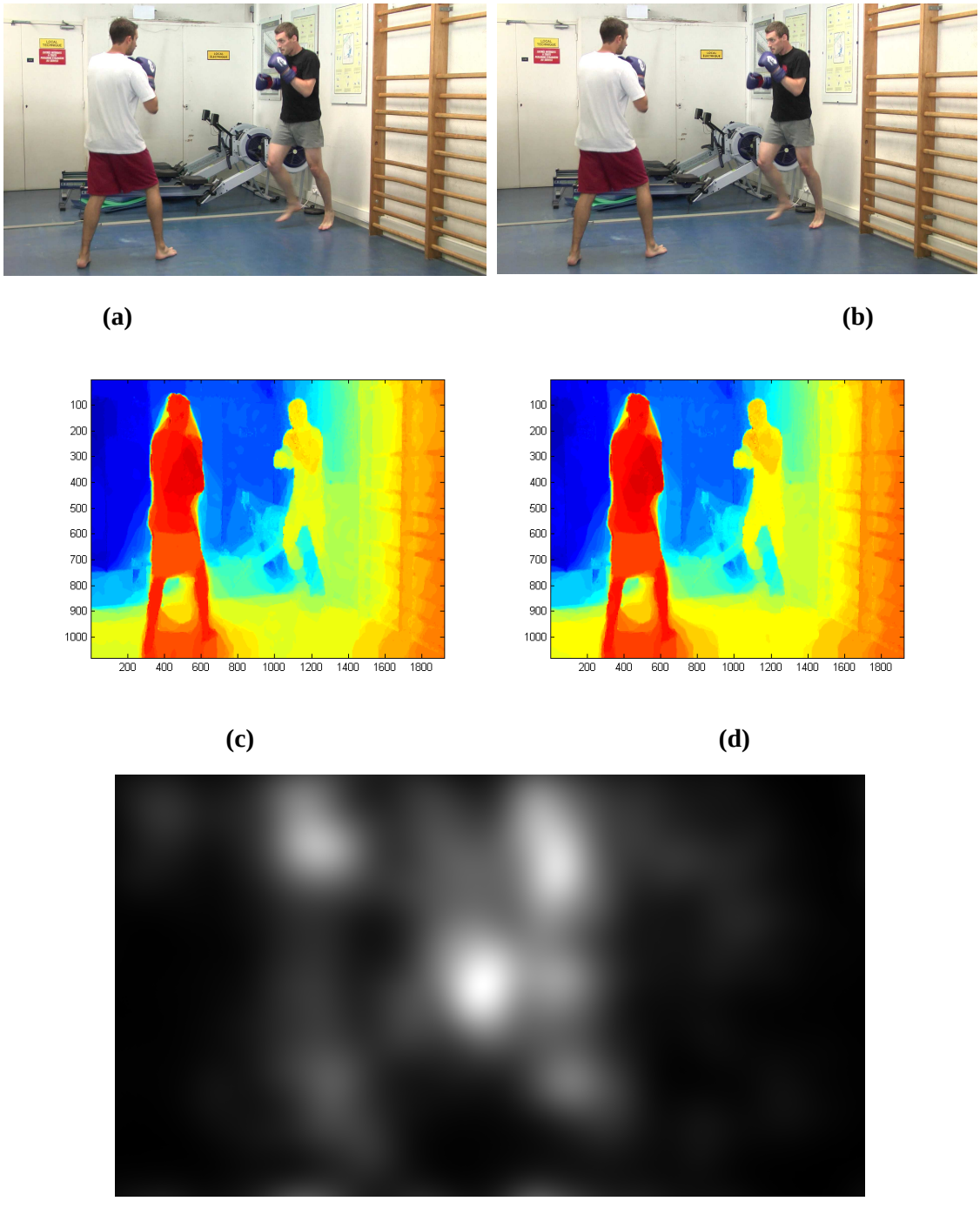
For the 2D saliency prediction, the authors performed the bottom-up visual attention models in [52], [53] and [54] (with quite different mechanisms), all of them also applied in the final performance evaluation. One view was chosen (the left one) because, besides the reduction of computational complexity, the images are subjectively similar (their 2D features differences have insignificant influence on VA).

Getting a precise depth map for natural content 3D images is not cheap and eye movements are affected by natural stimuli abundant features, increasing the difficulty on evaluating people's viewing behavior about depth information. Hence Bayes' theory shapes the correlation among depth features and depth saliency level, computing depth saliency map from an eye-tracking experiment result using synthetic stimuli.

A depth-weighting and a depth-saliency models are compared: a depth saliency map is important in 3D VA modeling, without a strong conclusion about those models (although the depth-based attention format has better performance for the purposes) besides the idea of a model combining both: depth information as an additional visual dimension (from which features are extracted to create saliency maps) and depth used as weighting data (relating the attention distribution and the distance between observer and each scene object). There is still no consensus on the 2D and depth information sources of saliency interaction and their effects on the saliency distribution, thus it was adopted a linear pooling strategy equally weighting their contributions. As there was no standardized methodologies for the conduction of eye-tracking experiments for 3D images, for the performance evaluation of stereoscopic 3D VA still images models, the

[9] Binocular vision cues (from two eyes), such as stereopsis and parallax, depend on eyes accurate alignment of the eyes and appropriate unification of the two images by the brain. Some people has only monocular (one-eye depth perception) skills, doing fine in situations where it is not required depth perception.

publicly available 3DGaze eye movement database was obtained from an eye-tracking experiment using synthetic stimuli with a background and some identical objects deliberately displayed at different depth plane. The database also contains, besides the free-task viewing eye-tracking data, 18 stereoscopic 3D images with natural content and the associated disparity and depth maps and FDM. Three objects parameters vary among scenes: number, size, and distance from the screen center [1].



(a) **(b)**

(c) **(d)**

(e)

Figure 3.3: Respectively Boxing image: (a) left view, (b) right view, (c) Disparity map, (d) Depth map, and (e) FDM.

As an example, in Figure 3.3: there are, respectively, images and its relative partial results obtained in [21]: on (a) and (b) the Boxing image left and right views; on (c) and (d) the Disparity and Depth Images resultant of the calculus about Boxing image; and on (e) the FDM obtained throught the eye tracking applied to the same Boxing image.

Since there are no specific measures to compare the similarity between FDMs and saliency maps created by 3D models, similarity measures used to judge saliency maps for 2D content where applied to each view, revealing a great influence of the depth contrast on the 3D VA distribution. A single 3D gaze point is produced by a not-straightforward triangulation of the two 2D gaze points from both eyes that relies on the system calibration. While 3D cases require a volumetric calibration, for 2D stimuli the calibration points can be shown on the screen (it is clear when the viewer looks accurately at a point, as 2D coordinates are known and the 2D gaze can be exactly tracked on the screen plane). Also the presentation time of each stimulus is a key in an eye-tracking experiment, being set to 15 s (long in comparison to 2D eye-tracking cases). The impact of that time on the resulting ground truth FDM were based on the FDMs obtained with different viewing durations (from 1 to all the 15s), having the model presented a better performance than 2D models or the Depth Saliency map itself. The models performance does not decrease with a longer presentation time, while it might be affected for a less than 10 seconds of presentation to create the ground truth FDMs.

Saliency, attention and visual search: an information theoretic approach [53]

This proposal for saliency computation is built entirely on computational constraints but results in a structure similar to that appearing in the visual cortex. Various visual search behaviors are emergent properties of the model and therefore basic principles of coding and information transmission.

The AIM model is based on a premise that localized saliency computation serves to maximize information sampled from one’s environment. The source code used is pubicly available and its default parameters are used in [21], with the rescaling factor was set to 0.25 (input image was rescaled to a quarter of its original size before being processed) to speed up the computation.

Saliency detection: a spectral residual approach [54]

The first step towards object recognition is object detection, extracting an object from its background through specific categories of objects. This simple method for visual saliency detection, independent of features, categories or other forms of prior knowledge of the objects, analises the log-spectrum of an input image, extracts the spectral residual in spectral domain, and, in a fast manner, constructs the corresponding saliency map in spatial domain. This model was tested on both natural pictures and artificial images such as psychological patterns. The Fourier spectrum is calculated based on luminance only and the image spectral residual is analised. The source code used is pubicly available and the default parameters are applied when this model is used in [21].

A model of saliency-based VA for rapid scene analysis [52]

This VA system is anchored in the early primate visual system behavior and neuronal architecture, where neurons are represented as capacitance. Parallel low-level visual multiscale image features extraction leads to the conspicuity maps of colour, intensity and orientation contrasts, which are combined into a single bottom-up topographical saliency map, modeled as a dynamical neural network. The features types applied are essential to the system efficiency and, as its maps are diverse, as also as the dynamic ranges and extraction mechanisms, protruding objects within maps can be masked by noise or less prominent objects existent in a larger number of maps.

The saliency map feeds synaptic interactions, ensuring that all the locations are removed, except the most active, thus portraying the conspicuity at every visual field location, in a decreasing scalar quantity of saliency (the value on the map refers to the saliency intensity on the spot, where it is the focus of attention), using feedback to search the next most prominent locations. The saliency spatial distribution guides the locations selection.

Saliency Toolbox 2.3 [55]

This collection of Matlab© functions and scripts calculates the 2D saliency map of an image and is also useful to determine the extent of a proto-object and for serially scanning a image with a focus of attention. It requires any computer and operating system that runs Matlab© release 13 or later with the image processing toolbox. This is not a feature richness system and does not have a fast processing speed, but it can be

used tois useful for computing the saliency map or attending to salient proto-objects in an image in a transparent and platform independent way. The Matlab© source code, saliencytoolbox, is publicly available and the saliency maps were obtained on [21] by performing the ’batch- Saliency’ command with default parameters.

4. Experimental evaluation

4.1. Objective Measurements

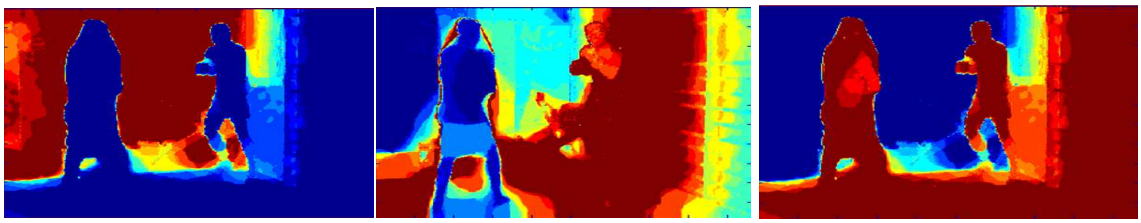
This chapter describes the objective measurements done in this work to reach the results to be discussed. Images ROIs were determined from each ones' FDM from [1], creating binary masks. Noise was then added (Gaussian or a Speckle) to the images area correspondent to its binary masks. 3D sequences generated were a combination of a noisy and a brightness view (considering that their arrangement may have influenced on the work purpose, according to the viewer dominant eye). Analysis and test sequences were classified according to the parameters to be analysed: the noise being inside or outside the ROI, the type of noise, its intensity and the 3D view where it is contained. Only some sequences were selected for being used on the still images subjective evaluation.

In this work, it was considered that every pixel of the RGB colour model computes an 8 bit luminance gray tone value between 0 and 255 (from the less to the most saturated degree) via the formula:

Eq. 4.1: Luminance = 0.3 * R + 0.59 * G + 0.11 * B

In images with all saturated colours (value 255), the luminance value is proportional to the weighted RGB components sum (notice that 0.3 + 0.59 + 0.11 = 1.0). Therefore, the distribution of light and dark tones range in coloured images was mapped by gray tone range in luminance values.

(a)



(b) **(c)** **(d)**

Figure 4.1: (a) 3DGaze database Boxing image left side brightness; (b), (c) and (d) respectively, its B, G and R components.

The 3DGaze database was used in this experimental study. These images were arrays of luminances degrees of each chrominance, as it can be seen on Figure 4.1. The original images were distinguished through new names given to each one of them.

Image	Name	Views	Resolution
01	Art	1_R	1278X1080
		1_L	
02	Baby	2_R	1191x1080
		2_L	
03	Books	3_R	1282x1080
		3_L	
04	Dolls	4_R	1279x1080
		4_L	
05	Laundry	5_R	1228x1080
		5_L	
06	Merchan	6_R	1274x1080
		6_L	
07	Objects	7_R	1286x1080
		7_L	
08	Plastic	8_R	1194x1080
		8_L	
09	Things	9_R	1247x1080
		9_L	

Image	Name	Views	Resolution
10	Rocks	10_R	1192x1080
		10_L	
11	Boxing	11_R	1920x1080
		11_L	
12	Hall	12_R	1920x1080
		12_L	
13	Lab	13_R	1920x1080
		13_L	
14	Report	14_R	1920x1080
		14_L	
15	Phone	15_R	1920x1080
		15_L	
16	Football	16_R	1920x1080
		16_L	
17	Tree	17_R	1920x1080
		17_L	
18	Umbrella	18_R	1920x1080
		18_L	

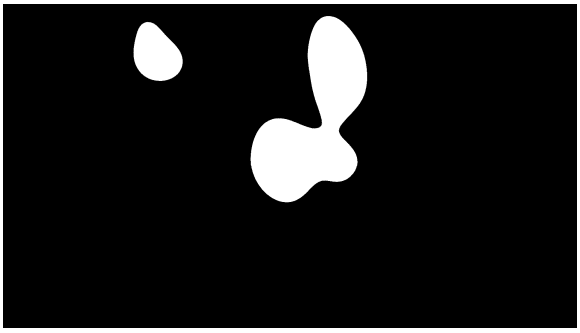
Figure 4.2: Images names and respective original sizes.

At Figure 4.2 it can be seen the names given to each image, its views and original resolutions. To convert the images (originally on the PNG format) into still images (YUV format), a Matlab© script was generate to add a gray pixels column to the original PNG images Baby, Dolls and Things, such as their resolution width and height values became both odd[10]. Those images needed in that process to be changed from arrays of uint8 values into arrays of double values.

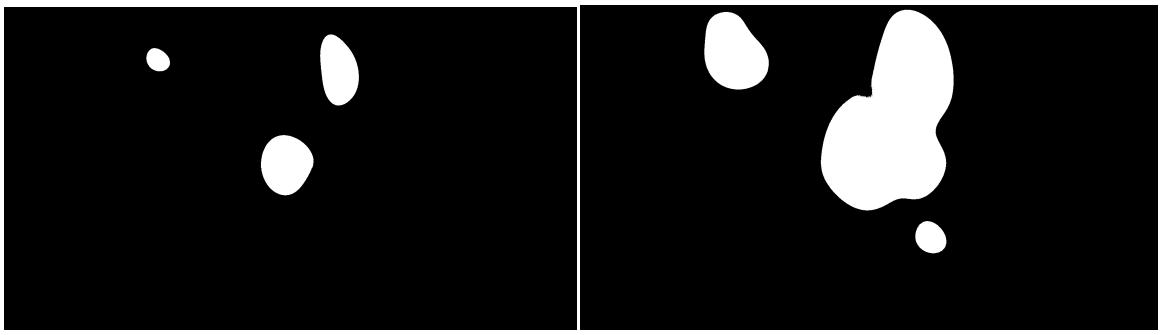
[10] Width and height values not divisible by 2 are not supported in YUV format.

Figure 4.3: Fixation Density Map (FDM) of the Boxing image.

Figure 4.3 is an example of the FDMs created on [1] using all gaze points recorded by the eyetracker from both eyes, considering the eye movements' position and duration. The left gaze points map was created by directly using the gaze positions coordinates, while the right one was created by adding a displacement on each right-eye gaze point coordinates. A disparity map indicates each gazed point displacements and the gaze points maps were summed and the noise was added (the eye tracker and the visual accuracies both decrease accordingly) [21].



(a)

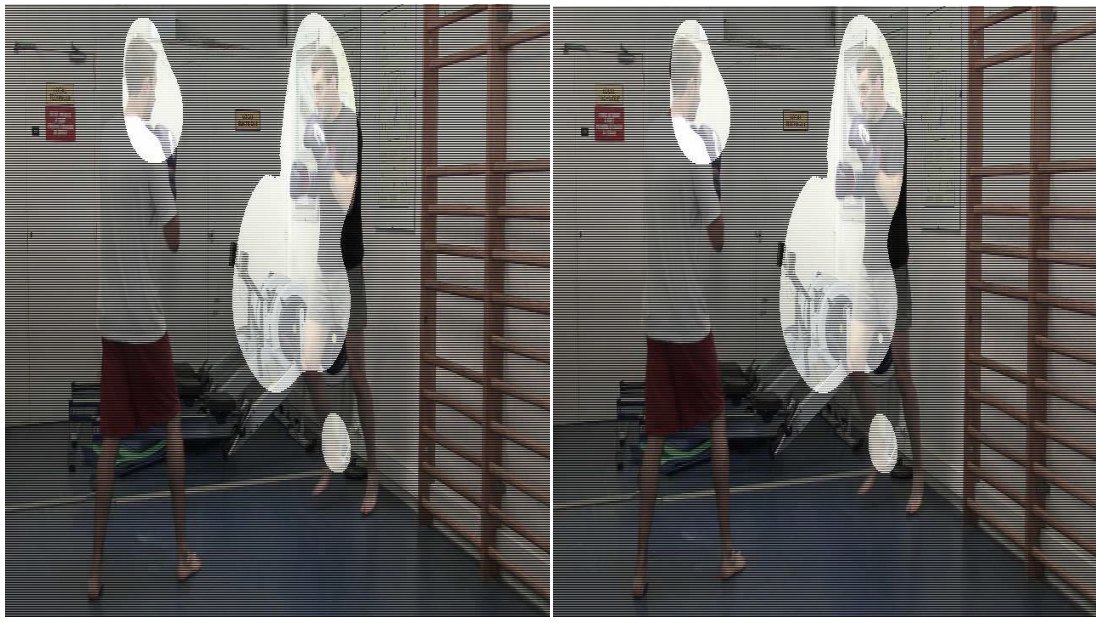


(b) **(c)**

Figure 4.4: The masks of left Boxing image: Mask128, Mask170 and Mask100, respectively.

In order to add noise specifically to the most fixed regions, a *binary mask*, with the same size of the SRC relates to that original image when a constant value is subtracted from each FDM pixel, being the noise added considering the 0 and 255 values as determining the ROI borders (generating the geographic in nature ROIs). Three ceiling pixels values (100, 128 and 170) were chosen to be used as the constant values.

The mask chosen to be used in this work was the Mask100, because of its greater coverage area, as it can be seen on the comparison done in the Figure 4.4 between the three constant values chosen (128, 170 and 100, respectively), such that the analysis of the added noise was made easier.



(a) **(b)**

Figure 4.5: The left and right Boxing images, each one overlapped by its correspondent Mask100.

In Figure 4.5 it can be seen an overlapping of the original Boxing image views and each correspondent FDM Mask100. Each 3D sequence generated was chosen to be a combination of a noisy and a SRC brightness view, considering that their adopted order may influence on the purpose of the work, according to the viewer dominant eye.

NOISE	VARIABLE1	VARIABLE2	Level
GAUSSIAN	0.01	0.002	1
	0	0.002	2
		0.003	3
SPECKLE	0.01	--------	1
	0.02	--------	2
	0.06	--------	3

Figure 4.6: Noises parameters: Variable 1 are the Gaussian mean and the Speckle multiplicative values; Variable 2 ir the Gaussian variance value.

The noises applied were the Gaussian and the Speckle due to a facility that the Matlab© provided: its function *imnoise* adds noise of a given type to an intensity image. Some parameters were also possibly selected there (every number was normalized, corresponding to operations with images with intensities ranging from 0 to 1), being the mean and variance the Gaussian parameters (the default values were zero mean noise with a 0.01 variance, applied when no parameters were mentioned). The Speckle noise added a multiplicative noise to the image I, considering the equation J = I+n*I, where n was an uniformly distributed random noise with mean 0 and variance v (the default for v was 0.04). The noise types' parameters were specified as if the image were of class double in the range [0, 1]: if the input image was uint8 or uint16, it was converted by the *imnoise* function to double, the noise was added accordingly and then the noisy image was converted back to the same class as the input. Figure 4.6 shows the parameters (variables) randomly selected for each noise. The Levels were related to the noise intensity [56].

(a)

(b)

Figure 4.7: Examples of NonROI and ROI noisy images

The noise was added inside or outside the images ROI (ROI or NonROI), through a Matlab© algorithm. An example is exposed on Figure 4.7, where there are a Boxing image, left view NonROI Gaussian Mask100 noise, with mean and variance values respectively of 0.01 and 0.002 and a Boxing image, right view ROI Speckle Mask100 noise, with a multiplicative value of 0.06.

The open source ANY2YUV[11] is a small console tool for image conversion into YUV colour space and it was applied to convert the images, originally in PNG format, into YUV images**.** Each still image was then reduced to a 960x1080 size (with the half width and the same height of the largest SRC still images), to be concatenated, performing the total of 384 3D sequences (24 test and 360 analysis sequences) created. The ffmpeg multimedia framework was used on that rescaling process. The ChromaLuma algorithm was created to add the SRC chrominance to the images to be evaluated. It followed the command line [57]:

ChromaLuma –isColour 1 –y 1080 –x 960 –l imageToBeEvaluated.yuv –c SRC.yuv –out newimage.yuv

where ChromaLuma was the command line to call the executable algorithm file; isColour parameter determined if the input file to be read had only luminance values or if it had also chrominances (0 or 1, respectively); y and x parameters defined the files

[11] This Project was rewritten from the ground and has reincarnated with a new name, YUVIT, in 2012. But the version used in this work is the original one, released in 2006 on Sourceforge, named ANY2YUV, created by Alexander Shashkevych, with the source code licensed under LGPLv3.

dimensions; l and c called the input still images to be used respectively as luminance and chrominance image sources; and newimage.yuv was the name of the final sequence created.

Then, the LeFrames algorithm was applied to each pair of still images to be concatenated, through the command line:

LeFrames –isColour 0 –n 10 –g 10 –y 1080 –x 960 –l LeftView.yuv –r RightView.yuv –out outputname.yuv

Where LeFrames was the command line to call the executable algorithm file; isColour parameter determined if the input file to be read had only luminance values or if it had also chrominances (0 and 1 values, respectively); n and g set how many times the 3D still images and the grey sequences would be printed, respectively; y and x parameters defined the files dimensions; l and r called the input still images to be used respectively as left and right images in the 3D concatenation; and outputname.yuv was the name of the final sequence created.

Still Image	Noise	View	ROI	Level	Test
Art	Gaussian	R	Y	1	1
				2	2
				3	3
			N	1	4
				2	5
				3	6
		L	Y	1	7
				2	8
				3	9
			N	1	10
				2	11
				3	12
	Speckle	R	Y	1	13
				2	14
				3	15
			N	1	16
				2	17
				3	18
		L	Y	1	19
				2	20
				3	21
			N	1	22
				2	23
				3	24

Still Image	Noise	View	ROI	Level	Test
Baby	Gaussian	R	Y	1	25
				2	26
				3	27
			N	1	28
				2	29
				3	30
		L	Y	1	31
				2	32
				3	33
			N	1	34
				2	35
				3	36
	Speckle	R	Y	1	37
				2	38
				3	39
			N	1	40
				2	41
				3	42
		L	Y	1	43
				2	44
				3	45
			N	1	46
				2	47
				3	48

Still Image	Noise	View	ROI	Level	Test
Books	Gaussian	R	Y	1	49
				2	50
				3	51
			N	1	52
				2	53
				3	54
		L	Y	1	55
				2	56
				3	57
			N	1	58
				2	59
				3	60
	Speckle	R	Y	1	61
				2	62
				3	63
			N	1	64
				2	65
				3	66
		L	Y	1	67
				2	68
				3	69
			N	1	70
				2	71
				3	72

Still Image	Noise	View	ROI	Level	Test
Dolls	Gaussian	R	Y	1	73
				2	74
				3	75
			N	1	76
				2	77
				3	78
		L	Y	1	79
				2	80
				3	81
			N	1	82
				2	83
				3	84
	Speckle	R	Y	1	85
				2	86
				3	87
			N	1	88
				2	89
				3	90
		L	Y	1	91
				2	92
				3	93
			N	1	94
				2	95
				3	96

Still Image	Noise	View	ROI	Level	Test
Laundry	Gaussian	R	Y	1	97
				2	98
				3	99
			N	1	100
				2	101
				3	102
		L	Y	1	103
				2	104
				3	105
			N	1	106
				2	107
				3	108
	Speckle	R	Y	1	109
				2	110
				3	111
			N	1	112
				2	113
				3	114
		L	Y	1	115
				2	116
				3	117
			N	1	118
				2	119
				3	120

Still Image	Noise	View	ROI	Level	Test
Objects	Gaussian	R	Y	1	121
				2	122
				3	123
			N	1	124
				2	125
				3	126
		L	Y	1	127
				2	128
				3	129
			N	1	130
				2	131
				3	132
	Speckle	R	Y	1	133
				2	134
				3	135
			N	1	136
				2	137
				3	138
		L	Y	1	139
				2	140
				3	141
			N	1	142
				2	143
				3	144

Still Image	Noise	View	ROI	Level	Test
Plastic	Gaussian	R	Y	1	145
				2	146
				3	147
			N	1	148
				2	149
				3	150
		L	Y	1	151
				2	152
				3	153
			N	1	154
				2	155
				3	156
	Speckle	R	Y	1	157
				2	158
				3	159
			N	1	160
				2	161
				3	162
		L	Y	1	163
				2	164
				3	165
			N	1	166
				2	167
				3	168

Still Image	Noise	View	ROI	Level	Test
Rocks	Gaussian	R	Y	1	169
				2	170
				3	171
			N	1	172
				2	173
				3	174
		L	Y	1	175
				2	176
				3	177
			N	1	178
				2	179
				3	180
	Speckle	R	Y	1	181
				2	182
				3	183
			N	1	184
				2	185
				3	186
		L	Y	1	187
				2	188
				3	189
			N	1	190
				2	191
				3	192

Still Image	Noise	View	ROI	Level	Test
Boxing	Gaussian	R	Y	1	193
				2	194
				3	195
			N	1	196
				2	197
				3	198
		L	Y	1	199
				2	200
				3	201
			N	1	202
				2	203
				3	204
	Speckle	R	Y	1	205
				2	206
				3	207
			N	1	208
				2	209
				3	210
		L	Y	1	211
				2	212
				3	213
			N	1	214
				2	215
				3	216

Still Image	Noise	View	ROI	Level	Test
Hall	Gaussian	R	Y	1	217
				2	218
				3	219
			N	1	220
				2	221
				3	222
		L	Y	1	223
				2	224
				3	225
			N	1	226
				2	227
				3	228
	Speckle	R	Y	1	229
				2	230
				3	231
			N	1	232
				2	233
				3	234
		L	Y	1	235
				2	236
				3	237
			N	1	238
				2	239
				3	240

Still Image	Noise	View	ROI	Level	Test
Lab	Gaussian	R	Y	1	241
				2	242
				3	243
			N	1	244
				2	245
				3	246
		L	Y	1	247
				2	248
				3	249
			N	1	250
				2	251
				3	252
	Speckle	R	Y	1	253
				2	254
				3	255
			N	1	256
				2	257
				3	258
		L	Y	1	259
				2	260
				3	261
			N	1	262
				2	263
				3	264

Still Image	Noise	View	ROI	Level	Test
Report	Gaussian	R	Y	1	265
				2	266
				3	267
			N	1	268
				2	269
				3	270
		L	Y	1	271
				2	272
				3	273
			N	1	274
				2	275
				3	276
	Speckle	R	Y	1	277
				2	278
				3	279
			N	1	280
				2	281
				3	282
		L	Y	1	283
				2	284
				3	285
			N	1	286
				2	287
				3	288

Still Image	Noise	View	ROI	Level	Test
Football	Gaussian	R	Y	1	289
				2	290
				3	291
			N	1	292
				2	293
				3	294
		L	Y	1	295
				2	296
				3	297
			N	1	298
				2	299
				3	300
	Speckle	R	Y	1	301
				2	302
				3	303
			N	1	304
				2	305
				3	306
		L	Y	1	307
				2	308
				3	309
			N	1	310
				2	311
				3	312

Still Image	Noise	View	ROI	Level	Test
Tree	Gaussian	R	Y	1	313
				2	314
				3	315
			N	1	316
				2	317
				3	318
		L	Y	1	319
				2	320
				3	321
			N	1	322
				2	323
				3	324
	Speckle	R	Y	1	325
				2	326
				3	327
			N	1	328
				2	329
				3	330
		L	Y	1	331
				2	332
				3	333
			N	1	334
				2	335
				3	336

Still Image	Noise	View	ROI	Level	Test
Umbrella	Gaussian	R	Y	1	337
				2	338
				3	339
			N	1	340
				2	341
				3	342
		L	Y	1	343
				2	344
				3	345
			N	1	346
				2	347
				3	348
	Speckle	R	Y	1	349
				2	350
				3	351
			N	1	352
				2	353
				3	354
		L	Y	1	355
				2	356
				3	357
			N	1	358
				2	359
				3	360

Still Image	Noise	View	ROI	Level	Test
Phone	Gaussian	R	Y	1	1T
				2	2T
				3	3T
			N	1	4T
				2	5T
				3	6T
		L	Y	1	7T
				2	8T
				3	9T
			N	1	10T
				2	11T
				3	12T
	Speckle	R	Y	1	13T
				2	14T
				3	15T
			N	1	16T
				2	17T
				3	18T
		L	Y	1	19T
				2	20T
				3	21T
			N	1	22T
				2	23T
				3	24T

Figure 4.8: Still images generated.

Figure 4.8 shows the analysis and test (the Phone sequences, from 1T to 24T) sequences generated, classified according respectively to: the applied noise type (Gaussian or Speckle); the noisy image view (left or right, L or R); being a ROI or a NonROI image (ROI: yes or no, Y or N); and the noise parameters values (intensity level). Due to the time limit and the purposes of this work, only some of these sequences were selected for being used on the still images subjective evaluation: the sequences 11, 23, 29, 35, 63, 66, 73, 75, 297, 309, 197, 203, 242, 245, 343 and 345. The images were, finally, joined together to be used in the subjective tests and had the time extension of the sum of each input file amalgamated.

4.2. Subjective validation of attention models

This chapter describes the subjective measurements done in this work to reach the purpose of examining the still images quality according primarily to the defined ROI.

To conduct appropriate subjective assessments, it is first necessary to select from the different options available those that best suit the purposes and circumstances of the

assessment problem. Thus, ITU recommends some conditions of tests conducted at a laboratory environment or home situation, concerning size, contrast, resolution and brightness of the screen, ambient lightning and observation angle, but limited to methods description (choosing a suitable method also depends on the purposes). Hence, more complete evaluation procedures of specific applications needed to be reported, such as considering the input signal quality and the source type, important features to obtain stable results. ITU recommends that a session does not take more than roughly half an hour, including explanations (at the beginning of each session, about the type of assessment, the grading scale, the sequence and timing) and preliminaries. The test sequences can begin with a few pictures indicative of the impairments range, which, together with their type to be assessed, should be illustrated on pictures other than those used in the analysis sequences (however of comparable sensitivity). The judgements of these pictures would not be taken into account in the final results [38] [17].

Eye dominance is a main subject concerning 3D video asymetric schemes and the overall quality is subjectively evaluated in [58] for representative types of asymmetry: Slightly blurred, Highly blurred, Slightly compressed, Highly compressed, Synthetized, Interpolated from low resolution, and Interpolated from very low resolution. It is verifyied that the dominant eye is determinant on quality appreciation only on Slightly assymetric video compression, changing the quality MOS by 16% at most. In [59], claims about the visual or oculomotor significance of a dominance eye are exposed: the eye dominance being related to handedness of hemispheric dominance; the existence of a single dominant eye for each person; the fact that for a given test there is a dominant eye; and the sighting dominant eye serving as the egocentric visual reference point. It is concluded that none of the common explicit or implicit claims has been supported by empirical evidence, and the conclusions reached in the 18th century first half still seem applicable. Therefore it is suggested that the sighting-dominant eye is used for monocular tasks and has no unique functional role in vision [58] [59].

To make a performance evaluation of quality metrics easier to be done it is important that the test materials databases are publicly available, also being considered that video require larger storage and bandwidth than still images. This work has employed, thus, the 3DGaze database, the general viewing conditions for subjective assessment of the quality of television pictures given on [17] and the ACR method for subjective evaluation. The system used was the autostereoscopic Toshiba© 55ZL2 LED smart TV, which specifications are depicted in the Annex 9.

The type of changes between both human eyes may differ depending on which level of stereoacuity is to be detected. The term stereopsis (or stereoscopic depth) is commonly referred specifically to the unique impression of depth associated with binocular vision. Hence, in this work a stereopsis test was then performed to verify each viewer perception of depth and the 3-dimensional structure, using slightly different YUV files (converted from original PNG images, of the same 3DGaze database, that was not used in the analysis or test steps) shown to each eye, such that a 3D image was perceived in case stereovision was there. Then a test was done to determine which one was the dominant eye for each normal vision participant.

Figure 4.9: Still images for tests, used on the example shown to the evaluators (4T and 13T).

Each participant received instructions on the evaluation method, the types of defects likely to occur, the rating scales, the sequences and the sequence-assessment duration. This content was written in Portuguese, printed and shown to them as described in the Appendix 8.1. At Figure 4.9 there are examples of still images that were used for tests. For the analysis itself, it was a must to choose only some of the groups of images due to the time it would be needed to be spent on each subjective test.

1A	Art	100NonROI01_L_Gaussian_M0_V002.yuv	brightness_01_R.yuv	11
2A	Baby	brightness_02_L.yuv	100NonROI02_R_Gaussian_M0_V002.yuv	29
3A	Books	brightness_03_L.yuv	100ROI03_R_Speckle_M06.yuv	63
4A	Dolls	brightness_04_L.yuv	100ROI04_R_Gaussian_M01_V002.yuv	73
5A	Boxing	brightness_11_L.yuv	100NonROI11_R_Gaussian_M0_V002.yuv	197
6A	Lab	brightness_13_L.yuv	100ROI13_R_Gaussian_M0_V002.yuv	242
7A	Football	100ROI16_L_Gaussian_M0_V003.yuv	brightness_16_R.yuv	297
8A	Umbrella	100ROI18_L_Gaussian_M01_V002.yuv	brightness_18_R.yuv	343
9A	Art	100NonROI01_L_Speckle_M02.yuv	brightness_01_R.yuv	23
10A	Baby	100NonROI02_L_Gaussian_M0_V002.yuv	brightness_02_R.yuv	35
11A	Books	brightness_03_L.yuv	100NonROI03_R_Speckle_M06.yuv	66
12A	Dolls	brightness_04_L.yuv	100ROI04_R_Gaussian_M0_V003.yuv	75
13A	Boxing	100NonROI11_L_Gaussian_M0_V002.yuv	brightness_11_R.yuv	203
14A	Lab	brightness_13_L.yuv	100NonROI13_R_Gaussian_M0_V002.yuv	245
15A	Football	100ROI16_L_Speckle_M06.yuv	brightness_16_R.yuv	309
16A	Umbrella	100ROI18_L_Gaussian_M0_V003.yuv	brightness_18_R.yuv	345

Figure 4.10: Ordered list of subjective analysis sequences applied.

The sequences applied in this work originally have different resolutions, according to the Figure 4.2, but all the 3D videos generated have a 1920x1080 resolution and all of them were generated in a video format colour space YUV with a 4:2:0 subsampling. From the total of 384 sequences (24 test sequences and 360 analysis sequences) generated, 4 of them were randomly selected to be applied to the tests step (4T, 9T, 13T and 24T) and 16 to be applied to the analysis step, those chosen based on the parameters combinations. As it can be seen on the Figure 4.10, the analysis sequences were chosen to evaluate the impact of each sequence characteristic on the viewer perception. For a posterior comparison, just one parameter differs between each two sequences: the type of noise (if it is Gaussian or Speckle), the noisy view (right or left – to verify the impact of the dominant eye), the noisy region (ROI or NonROI, to verify the location influence) or the noise intensity level. As a result, there were 4 classes of sequences: Noise, Level, ROI and View. Each class with two sequences presented: a quantity chosen due to the subjective tests duration, relative specially to the factors that affect HVS quality when viewing 3D. It can be seen that the two sequences to be compared between each other were not exposed one right after the other: the sequences 1A to 8A are to be compared respectively with the sequences 9A to 16A.

Each sequence was composed by a selected still image repeated 10 times (to reach a 10 seconds exposition time), followed by a 10 seconds grey still image (given as a time to the evaluator judge according to the discrete 5 points scale), being the two existing views of each sequence concatenated to rearrange the images as 3D. These sequences were exposed to the viewers in a 1 frame per second rate (i.e., reaching the exposition time of 20 seconds), preceeded by a full screen gray frame with an image of a number correspondent to the sequence to be analysed (for organization purposes of the analysis), so that the entire analysis video amounted a total of 06 minutes and 08 seconds.

The time of presentation was choosen based on: the fact that, for still pictures on DS methods, it is said on [17] that a 3 to 4 seconds sequence and five repetitions (voting during the last two) may be appropriate; that the models performance do not decrease with a longer presentation time (while less than 10 seconds of presentation to create the ground truth FDMs might affect it); and that the scenes low complexity allow a shorter observation duration, being the viewing time of natural content images in eye-tracking experiments generally set to 5 seconds or more, as it is concluded on [21]. It is also said that human memory effects can distort quality ratings if noticeable impairments occur in

approximately the last 10 to 15 seconds of the sequence (the judgement of real service situations is not representative if video sequences under evaluation are limited to 10 seconds). Nevertheless, on this work there were not videos, but still images sequences, which also were not from real service situations, but collected from a previously existent database. So, the 10 seconds limit were believed not to be applied in this situation [21] [17].

This work was applied to 3D cases, in which there must be considered a period of adaptation besides the factors that influence human perception. The environment configuration was in accordance with the Toshiba© 55ZL2 LED smart TV specifications and the display configuration was done based on [60]: the viewer distance from the TV was the recommended nearly 2,2m; the side-by-side 3D format was selected to be displayed, according to the images views disposal; the 3D viewing position test feature was used to adjust the viewers positions according to a frame face tracking by the built-in camera under the TV screen. However, it was a home-likely environment, not a lab-likely: for example, the lightning was not always the same between viewers.

The ffplay media player, based on the ffmpeg multimedia framework libraries, was applied to the execution of the video displaying: first the tests sequence, then, after clarifying any doubts, the analysis sequence.

This page was intentionally left blank

5. Results - discussion

The impact of adding noise to ROI defined by FDM in 3D images was studied to evaluate whether the ROI is subjectively identified as a remarkable part of the scene, influencing the quality perception. That is, if the same image with noise in the ROI were subjectively uncomfortable (perception of lower quality) to the viewer. The parameters analysed were the noise inside or outside the ROI, the type of noise, its intensity and the 3D view where it is contained (respectively ROI, Noise, Level and View). The objective measurements in this work used geographic in nature ROI (not a range of intensities, but through the use of a binary mask) and the ROI size was not used as a parameter (the same binary mask was applied during all the analysis), as also as the tendency of humans fixation on humanoid images (specially on human faces), and there was not a distinction between input images according to their content.

Autostereoscopic viewing has the benefit that no special eye-wear is needed and the possibility of many viewers concomitantly. However, since there is some parallax, the need of high-resolution devices (larger screen makes the user less aware of the frame, minimizing the frame effect) and, to be as a lab, the environment must be very well-controlled. However, the expected discomfort due to the sweet spot need was not noticed on this work. The inconsistency between acommodation and convergence, which may cause sickness after a few minutes (eye fatigue) in this work only happened with some women at the end of the analysis.

35 subjects, ranging in age from 20 to 41 years, participated in the eye tracking experiment. All subjects were naive in regard to the experiments and they had either normal or corrected-to-normal visual acuity.

Number / Name	Mean	Round mean	Std Dev	Confid interval	Mode	LE Mean	LE Round mean	Std Dev	LE Mode	LE Confid interval	RE Mean	RE Round mean	Std Dev	RE Mod	RE Confid interval
1A	3.54	4	0.95	0.83	4	3.69	4	1.03	4	0.90	3.45	3	0.91	4	0.80
2A	3.57	4	0.85	0.75	4	3.54	4	0.88	4	0.77	3.59	4	0.85	4	0.75
3A	4.00	4	0.80	0.71	4	3.85	4	0.99	3	0.87	4.09	4	0.68	4	0.60
4A	3.91	4	0.89	0.78	3	3.92	4	0.86	3	0.76	3.91	4	0.92	3	0.81
5A	2.66	3	0.84	0.73	2	2.69	3	0.95	2	0.83	2.64	3	0.79	2	0.69
6A	3.11	3	0.90	0.79	3	3.15	3	0.99	3	0.87	3.09	3	0.87	3	0.76
7A	3.31	3	0.99	0.87	3	3.77	4	0.93	4	0.81	3.05	3	0.95	3	0.83
8A	3.40	3	1.03	0.91	4	4.00	4	0.71	4	0.62	3.05	3	1.05	4	0.92
9A	2.94	3	1.16	1.02	4	3.00	3	1.22	3	1.07	2.91	3	1.15	4	1.01
10A	2.49	2	0.98	0.86	2	2.85	3	0.80	3	0.70	2.27	2	1.03	2	0.90
11A	3.91	4	0.82	0.72	4	4.08	4	0.76	4	0.67	3.82	4	0.85	4	0.75
12A	4.03	4	0.79	0.69	4	4.23	4	0.83	4	0.73	3.91	4	0.75	4	0.66
13A	2.09	2	0.82	0.72	2	2.31	2	0.95	3	0.83	1.95	2	0.72	2	0.63
14A	3.17	3	0.82	0.72	3	3.46	3	0.88	3	0.77	3.00	3	0.76	3	0.66
15A	2.06	2	1.06	0.93	1	2.54	3	1.13	2	0.99	1.77	2	0.92	1	0.81
16A	2.97	3	0.98	0.86	3	3.46	3	1.05	3	0.92	2.68	3	0.84	2	0.74
Dom. eye					R (23)										
Age	29.91	30	7.44	6.52	20	30.77	31	7.07	35	6.20	29.41	29	7.77	22	6.81
Sex					M (27)										

Figure 5.1: Analysis sequences subjective results.

In Figure 5.1 there are the subjective tests results: Mean, Round mean, Standard deviation and Confidence interval for every viewers judgements, besides for specifically each eye-dominant viewer (LE, left eye and RE, right eye), as also as the global Modes for each sequence evaluated. The outcomes are analysed according to the framework: a comparison between corresponding images pairs (Art 1A and 9A, Baby 2A and 10A, Books 3A and 11A, Dolls 4A and 12A, Boxing 5A and 13A, Lab 6A and 14A, Football 7A and 15A, Umbrella 8A and 16A) determines the parameters variation to be evaluated (Noise, Views, ROI, Level).

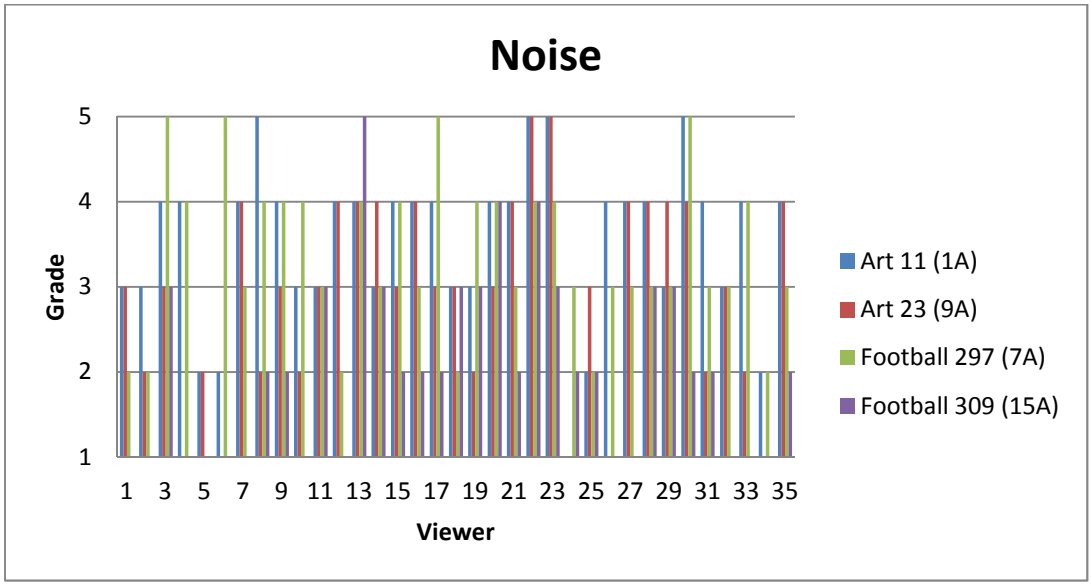


Figure 5.2: Noise parameter results.

In Figure 5.2 there is a graphic with the subjective tests results about the Noise parameter. It can be noticed that the Gaussian noise had less impact on quality, as it was

observed a superior MOS. The fact that the Art sequences were of a NonROI type while the Football sequences were ROI did not have influence in the results, as the comparison was done between each Art sequence or each Football sequence.

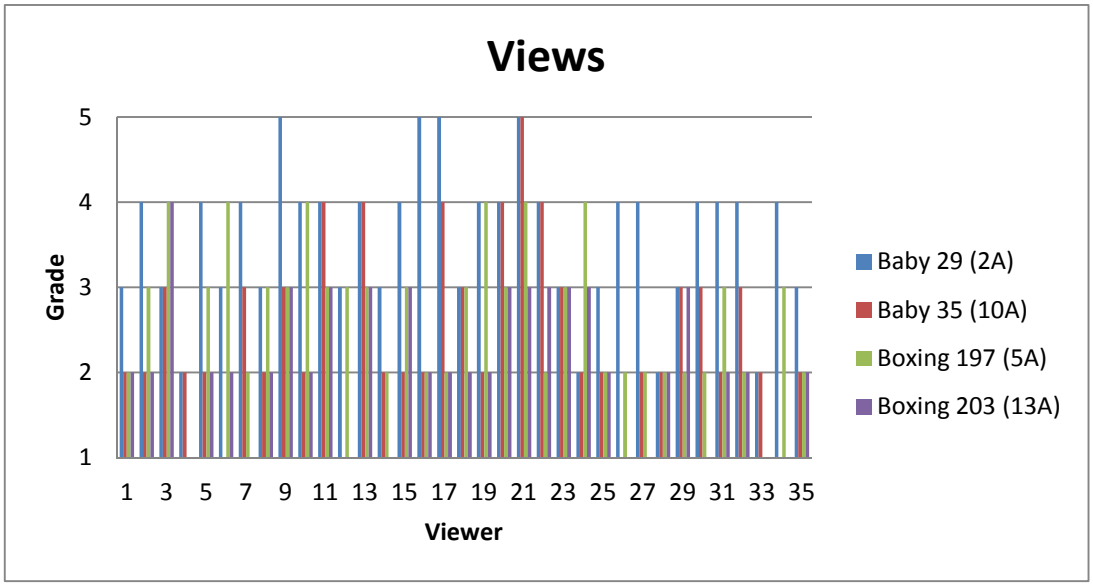


Figure 5.3: Views parameter results.

In Figure 5.3 there is a graphic with the subjective tests results for the Views parameter. It can be seen that the quality was mostly choosen as worse when the noise was on the right view than when the noise was on the left view. It seems to be probably due to the viewers eye dominance. Although, between the Boxing sequences the quality variation was so small that it is inconclusive.

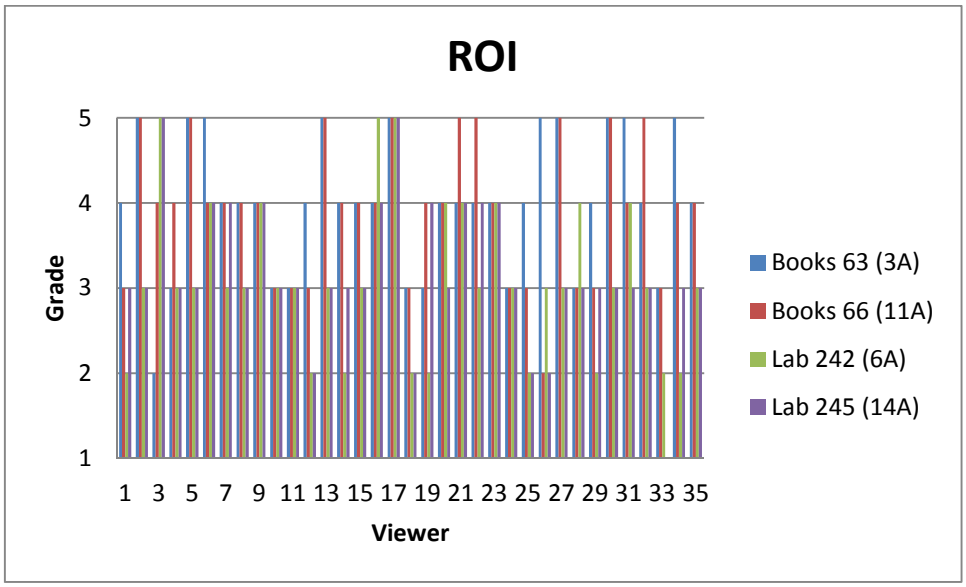


Figure 5.4: ROI parameter results.

In Figure 5.4 there is a graphic with the subjective tests results for the ROI parameter. It could be noticed a inferior MOS when the noise was inside the ROI, what is justified

due to the viewers had fixed their eyes over that region during more time, becaming higher the distortion.

It was observed that viewers with the right eye dominance had the tendency of judging the sequences with the noise inside the ROI as having better quality, in opposition of the opinion of most people with the left eye dominance. But, as the analysis spectrum was small, the dominant eye information seem to be unconclusive.

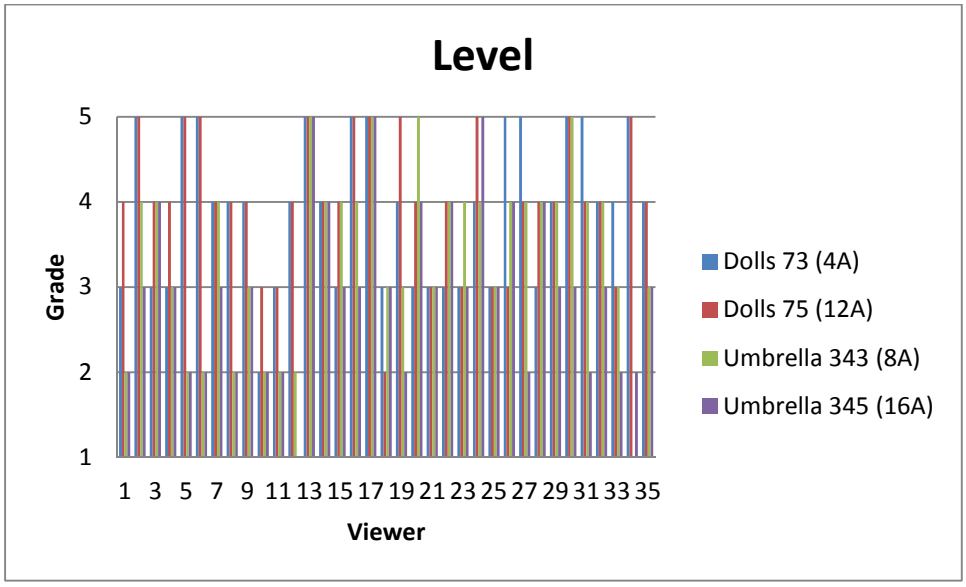


Figure 5.5: Level parameter results.

In Figure 5.5 there is a graphic with the subjective tests results about Level parameter. It was verified that the quality was judge as worse when the Gaussian noise applied had a higher intensity Level and viewers with the left eye dominance gave in every situation higher scores than viewers with the right eft eye dominance.

6. Conclusion

The Gaussian noise had less impact on quality than Speckle. Also the quality was mostly choosen as worse when the Gaussian noise applied had a higher intensity Level, when the noise was on the right view and when the noise was inside the ROI, what is justified due to the viewers had fixed their eyes over that region during more time, becaming higher the distortion.

It was observed that viewers with the right eye dominance had the tendency of judging the sequences with the noise inside the ROI as having better quality, in opposition of the opinion of most people with the left eye dominance. But, as the analysis spectrum was small, the dominant eye information seem to be unconclusive.

It is suggested for future works to apply other intensity levels to the same Gaussian and Speckle noises or applying other noises, such as Salt and Pepper or Poisson, both available on Matlab. Changing the parameters is also suggested for future works, in such a way that there is more certainty on results achieved with one do not interfere in the analysis done with another (such as the eye dominance interferrying in the viewer not even noticing if there is any noise, making it impossible to know the opinion about the others parameters). The combination of the application of more noise levels and less parameters to be evaluated (without changing the noisy view side, for example) could also be applied in future works, leading to more optimized results. Instead of using a binary mask, the objective measurements of future works could use a range of intensities, such as adding noise on the region determined by the FDM. The ROI size could also be used as a parameter (different binary masks being applied during the analysis) in future works. Also different subjective methods could be used, such as a top-down performance SS or SC method.

A distinction between input images according to their content could be a parameter for future works, as also as the study of the difference between only-luminance and with-chrominance situations, between still images or videos, a Lab or a Home environment, an autostereoscopic display or a non-aurostereoscopic display.

The TV used was the first model commercialized in Portugal with a Native 4K resolution of 3840x2160 pixels and this characteristic could not me exploited. A database (still images and FDMs) generated to this situation is suggested for future works. 3D stereoscopic interlaced images could be used in evaluation, in opposition of

the applied side-by-side positioned images. Moreover, a study about if the viewer angle position influence on the image perception is suggested for future works, as also as the expansion of the analysis to other areas, such as asking if animals would also look firstly for animals faces, considering the human characteristic of looking primarily for human faces or humanoid things. As top down mechanisms are strongly dependent also on the semantic information and this work was limited to bottom-up visual interest (applying the usual limitation to models that compute saliency maps representing the level of bottom-up visual interest), some top-down concepts, such as rarity or surprise may naturally be included on future works.

Furthermore, more complete evaluation procedures of specific applications can be reported in future works, such as considering the input signal quality and the source type, important features to obtain stable results.

7. References

[1] J. Wang, M. P. Da Silva, P. Le Callet e V. Ricordel, “IRCCyN/ IVC 3DGaze database,” Institut de Recherche en Communications et Cybernétique de Nantes, 2011. [Online]. Available: http://130.66.64.103/spip.php?article1103&lang=.

[2] European Cooperation in Science and Technology, “COST Action IC1003 - European Network on Quality of Experience in Multimedia Systems and Services (QUALINET),” MAY 2011. [Online]. Available: http://www.qualinet.eu/.

[3] S. Chikkerur, V. Sundaram, L. J. Karam e M. Reisslein, “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison,” em *IEEE TRANSACTIONS ON BROADCASTING*, 2011.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh e E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” em *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 2004.

[5] A. Mittal, A. K. Moorthy, J. Ghosh e A. C. Bovik, “Algorithmic Assessment of 3D Quality of Experience for Images and Videos,” em *IEEE Digital Signal Processing and Signal Processing Education Meeting*, Arizona, 2011.

[6] T. M. O'Neil, “Quality of experience and Quality of Service - for IP Video Conferencing,” Milpitas, 2003.

[7] S. Winkler, “Video Quality and Beyond,” em *Proc. European Signal Processing Conference*, Poznan, Poland, 2007.

[8] I. E. G. Richardson, H. 264 and MPEG-4 Video Compression, UK, Aberdeen: John Wiley & Sons Ltd, 2003, pp. 20 - 24.

[9] F. De Rango, M. Tropea, P. Fazio e S. Marano, “Overview on VoIP: Subjective and Objective Measurement Methods,” em *IJCSNS International Journal of Computer Science and Network Security*, Italy, 2006.

[10] M. Barkowsky, *Subjective and Objective Video Quality Measurement in Low-Bitrate Multimedia Scenarios,* Erlangen: The Technical Faculty of the Friedrich Alexander University Erlangen-Nuremberg, 2009.

[11] N. Ouerhani e H. Hiigli, “Computing Visual Attention from Scene Depth,” em *Proceedings of the International Conference on Pattern Recognition*, 2000.

[12] L. S. Kreulich, *Sistema de Compressão Otimizado para Informação de Textura e Disparidade,* Rio de Janeiro, Rio de Janeiro: Departamento de Ciência e Tecnologia do Instituto Militar de

Engenharia, Ministério da Defesa - Exército Brasileiro, 2011.

[13] C. G. Relf, *Image Acquisition and Processing with LabVIEW™,* P. A. Laplante, Ed., Pennsylvania: Pennsylvania Institute of Technology, 2003.

[14] The MathWorks Inc ©, "Specifying a Region of Interest (ROI) - MATLAB & Simulink," 2014. [Online]. Available: http://www.mathworks.com/help/images/specifying-a-region-of-interest-roi.html.

[15] T. Oakes, "Introduction to Drawing Regions of Interest (ROIs)," 2006. [Online]. Available: http://brainimaging.waisman.wisc.edu/~oakes/spam/BrainMaker_Intro.html. [Acesso em 2014].

[16] R. N. d. Fonseca e M. A. Ramírez, "Avaliação da Qualidade de Vídeo em Televisão Digital," *Revista de Radiodifusão,* 2008.

[17] International Telecommunications Union, *Recommendation ITU-R BT.500-12 - Methodology for the subjective assessment of the quality of television pictures,* Geneva: Radiocommunication Sector of the International Telecommunications Union, 2009.

[18] J. Wang, D. M. Chandler e P. L. Callet, "Quantifying the Relationship between Visual Salience and Visual Importance," em *Proceedings of the XV Spie Human and Electronic imaging (HVEI)*, United States of America, 2010.

[19] K. Seshadrinathan e A. C. Bovik, "A Structural Similarity Metric for Video Based on Motion Models," em *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, Hawaii, 2007.

[20] Z. Wang e A. C. Bovik, "Mean Squared Error: Love it or Leave it? - a new look at signal fidelity measures," em *IEEE Signal Processing Magazine*, 2009.

[21] J. Wang, M. P. Da Silva, P. Le Callet e V. Ricordel, "Computational Model of Stereoscopic 3D Visual Saliency," em *IEEE Transactions On Image Processing*, 2012.

[22] W. Fulton, "More details about Luminance in Histograms," 1997-2010. [Online]. Available: http://www.scantips.com/lumin.html. [Acesso em 2014].

[23] A. C. Bovik e K. Seshadrinathan, "Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos," em *IEEE Transactions on Image Processing*, 2010.

[24] L. Onural, "The 3DTV Toolbox - The results of the 3DTV NoE," em *Workshop pf 3DTV Broadcasting*, Geneva, 2009.

[25] S.-Y. Park, C.-I. Kim e E.-K. Jung, "Vertical disparity correction of stereoscopic video using fast feature window matching," em *International Conference on Consumer Electronics (ICCE)*, U.S.A., 2012.

[26] L. M. J. Meesters, W. A. IJsselsteijn e P. J. H. Seuntiëns, “A Survey of Perceptual Evaluations and Requirements of Three-Dimensional TV,” em *IEEE Transactions on Circuits and Systems for Video Technology*, 2004.

[27] International Telecommunications Union, *Recommendation ITU-R BT.1438 - Subjective assessment of stereoscopic television pictures,* Geneva: Radiocommunication Sector of International Telecommunications Union, 2000.

[28] J. Strickland, “HowStuffWorks "Lenticular Displays",” June 2009. [Online]. Available: http://electronics.howstuffworks.com/3d-tv5.htm.

[29] L. Pinto, *Descodificador de Vídeo 3D com Robustez a Erros e Perdas de Pacotes,* Leiria: Instituto Politécnico de Leiria, Escola Superior de Tecnologia e Gestão, 2011.

[30] P. Anzil e P.-J. Alzieu, “Review: The Samsung ES6300 does a great job in 3D mode with practically no crosstalk...,” Dec 2013. [Online]. Available: http://www.digitalversus.com/tv-television/samsung-ue40es6300-p13200/test.html.

[31] J. S. Whiting, K. N. Mahrer, M. P. Eckstein e N. L. Eigler, “Improving detection of coronary morphological features from digital angiograms. Effect of stenosis stabilized display,” *Circulation, Journal of the american heart association,* vol. 89, p. 2700 – 2709, 1994.

[32] Video Quality Experts Group, *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment,* 2000.

[33] Video Quality Experts Group, *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II,* 2003.

[34] Video Quality Experts Group, *Report on the validation of Video Quality Models for High Definition Video Content,* 2010.

[35] M. Carnec, P. Le Callet e D. Barba, “An Image Quality Assessment Method Based on Perception of Structural Information,” em *IEEE International Conference on Image Processing*, Barcelona, 2003.

[36] W. Lin, “Determine Visual Just-noticeable Difference (JND) for Multimedia Applications,” em *IEEE Multimedia Communications Technical Committee (IEEE COMSOC MMTC) E-Letter*, Singapore, 2009.

[37] A. Vincent e P.-P. Garcia, “Sony: New 3D Mode Reduces Eye Fatigue,” Mar 2012. [Online]. Available: www.digitalversus.com/tv-television/sony-new-3d-mode-reduces-eye-fatigue-n23563.html.

[38] J. F. Rehme, *Avaliação da qualidade de vídeo trafegando sobre redes IP,* Curitiba, Paraná: Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, 2007.

[39] N. Boone, “3D TV appears not to cause problems for children with epilepsy,” May 2011. [Online]. Available: http://www2.scnow.com/lifestyles/2011/dec/05/3d-tv-appears-not-cause-problems-children-epilepsy-ar-2805620/.

[40] International Telecommunications Union, *Recommendation ITU-T P.910 - Subjective video quality assessment methods for multimedia applications,* Geneva: Telecommunication Standardization Sector of International Telecommunications Union, 2008.

[41] S. S. Vorren, *Subjective quality evaluation of the effect of packet loss in High-Definition Video,* Norwegian University of Science and Technology, Department of Electronics and Telecommunications, 2006.

[42] 3D@Home Consortium and MPEG Industry Forum 3DTV Working Group, “Glossary for Video & Perceptual Quality of Stereoscopic Video,” http://www.3dathome.org/files/ST1-01-01_Glossary.pdf, 2010.

[43] S. Winkler e F. Dufaux, “Video quality evaluation for mobile applications,” em *Proceedings of the Visual Communications and Image Processing Conference, SPIE VCIP 03*, Lugano, Switzerland, 2003.

[44] S. Wolf e M. H. Pinson, “Comparing subjective video quality testing methodologies,” em *Proceedings of the Visual Communications and Image Processing Conference, SPIE VCIP 03*, Lugano, Switzerland, 2003.

[45] International Telecommunications Union, *Recommendation ITU-R BT.1082 - Studies toward the unification of picture assessment methodology,* Geneva: Radiocommunication Sector of the International Telecommunications Union, 1990.

[46] International Telecommunications Union, *Recommendation ITU-R BT.1788 - Methodology for the subjective assessment of video quality in multimedia application,* Geneva: Radiocommunication Sector of the International Telecommunications Union, 2007.

[47] A. Benoit, P. Le Callet, P. Campisi e R. Cousseau, “Quality Assessment of Stereoscopic Images,” em *EURASIP Journal on Image and Video Processing, special issue on 3D Image and Video Processing*, 2008.

[48] International Telecommunications Union, *Recommendation ITU-R BS.1534-1 - Method for the subjective assessment of intermediate quality level of coding systems,* Geneva: Radiocommunication Sector of International Telecommunications Union, 2003.

[49] N. Ozbek, G. Ertran e O. Karakus, “Interactive Quality Assessment for Asymmetric Coding of 3D Video,” em *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, 2011.

[50] K. Chen, M. Yu, G. Jiang e Y. Zhang, “Stereoscopic visual attention model for 3d video,” em *Advances in Multimedia Modeling*, vol. 5916, L. Zhang, Q. Tian, S. Boll, Z. Zhang e P. C. Yi-Ping,

Eds., Chongqing, Springer Berlin Heidelberg, 2010, pp. 314-324.

[51] A. Maki, P. Nordlund e J.-O. EMundh, “A Computational model of Depth-based attention,” em *IEEE Proceedings of 13th International Conference on Pattern Recognition (ICPR'96)*, Vienna, 1996.

[52] L. Itti, E. Niebur e C. Koch, “A model of saliency-based visual attention for rapid scene analysis,” em *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998..

[53] N. D. B. Bruce e J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” em *Journal of Vision* , Canada, 2009.

[54] X. Hou e L. Zhang, “Saliency detection: a spectral residual approach,” em *IEEE Conference on Computer Vision and Pattern Recognition,* , United States of America, 2007.

[55] D. W. a. C. Koch, “Modeling attention to salient proto-objects,” em *Neural Networks 19*, 2006.

[56] The MathWorks Inc ©, 2015. [Online]. Available: http://www.mathworks.com/help/images/ref/imnoise.html.

[57] A. Shashkevych, “Any To YUV - Welcome,” February 2012. [Online]. Available: http://any2yuv.sourceforge.net/HomePage. [Acesso em June 2015].

[58] A. Banitalebi-Dehkordi, M. T. Pourazad e P. Nasiopoulos, “Effect of eye dominance on the perception of stereoscopic 3D video,” em *IEEE International Conference on Image Processing (ICIP)*, France, 2014.

[59] A. P. Mapp, H. Ono e R. Barbeito, “What does the dominant eye dominate? A brief and somewhat contentious review,” *Perception & Psychophysics,* vol. 65, pp. 310-317, Feb 2003.

[60] Toshiba Corporation, 2015. [Online]. Available: http://www.toshiba.eu/discontinued-products/55zl2/.

[61] Immersive Media, “Immersive Media | Any Device. Any Place. Any Time,” Jun 2014. [Online]. Available: http://immersivemedia.com/.

[62] D. Hoiem, *Epipolar Geometry and Stereo Vision,* Illinois: University of Illinois - CS 543 / ECE549, 2012.

[63] G. Roelofs, “libpng (the free reference library for reading and writing PNGs),” 2014. [Online]. Available: http://www.libpng.org/pub/png/. [Acesso em 2014].

This page was intentionally left blank

8. Appendix

8.1. Instructions to the participants

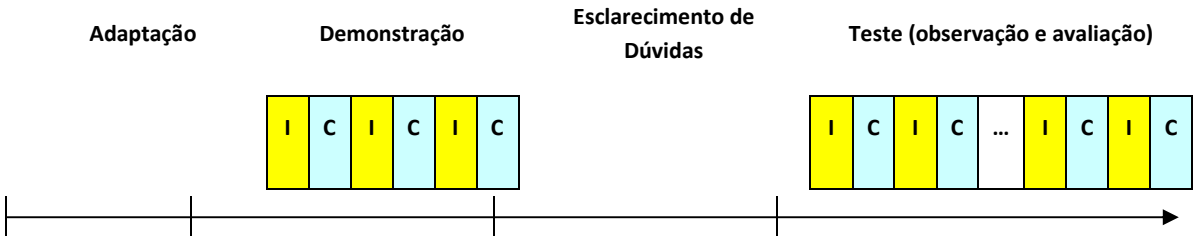
Firstly a stereopsis test is done, as also as a test to determine which eye is the dominant one for each normal vision participant. Then each viewer receives instructions about the assessment method to be used, the types of defects likely to occur, the sequences rating scales, the sequences to be used and the timing sequence-assessment. This content is written in Portuguese, printed and shown to them as follow.

TESTES DE AVALIAÇÃO SUBJETIVA DE IMAGEM 3D

Obrigada pela disponibilidade para efetuar testes subjetivos.
Nestes testes você irá ver diversas imagens 3D durante cerca de 10s cada.
No final da visualização de cada imagem, deverá avaliá-la relativamente a um único aspecto: ***A percepção da qualidade da imagem tendo em consideração a presença de distorções visíveis***
A qualidade da imagem deverá ser avaliada atribuindo uma classificação dentro da seguinte escala:

5 – Muito Boa
4 – Boa
3 – Razoável
2 – Má
1 – Muito Má

Deve observar atentamente cada imagem de modo a que, no final do período de observação, a sua opinião corresponda a um destes níveis de comparação.
Após a visualização, deverá decidir sobre a classificação a atribuir (entre 1 e 5), em um tempo de 10 segundos.
A sessão de avaliação é constituída pelas seguintes fases:



No período de Adaptação será feito um teste para identificar o olho dominante. A seguir, serão apresentadas imagens como aquelas que serão avaliadas. Este período tem por objetivo familiarizar o seu sistema de percepção 3D com a tecnologia e os conteúdos que lhe serão apresentados.

Segue-se um passo de Demonstração, em que serão visualizados exemplos de sequências cujo impacto perceptual se pretende avaliar (tempo: I). Após cada visualização deverá ser efetuada a respectiva classificação (tempo: C). A classificação será atribuída usando a tabela que consta em folha à parte.

Na etapa de Esclarecimento de Dúvidas você poderá colocar todas as questões que considerar pertinentes.

O período de Teste consiste em visualizar diversas sequências (I) seguidas de avaliação e classificação (C). Este período terá uma duração máxima de 30 min.

Obrigada pela sua colaboração!

TABELA PARA CLASSIFICAÇÃO DE CADA IMAGEM VISUALIZADA

Escala de percepção da qualidade da imagem:

5 – Muito Boa

4 – Boa

3 – Razoável

2 – Má

1 – Muito Má

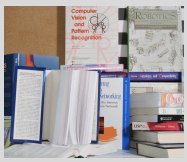
SEQUÊNCIA	Classificação				
	1	2	3	4	5
01					
02					
03					
04					
05					
06					
07					
08					
09					
10					
11					
12					
13					
14					
15					
16					

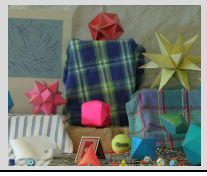
This page was intentionally left blank

9. Annex

9.1. 3D Gaze: an eye tracking on 3D images Database

The 3DGaze image database contains 18 stereoscopic images and the associated FDM, disparity map, depth map, and the raw eye tracking data. Among the 18 stereoscopic 3D image contents with no impairment provided, 10 are from the Middlebury College database and 8 from the University of Nantes. There is only one version of each image content. The database is provided free of charge, was created to measure how different features (2D or 3D ones) affect the VA distribution and can be used to evaluate the performance of 3D VA computational models [1].

SRC	Name	Source name	Preview	Description	Resolution
01	Art	Middlebury College 2005 Art		A painting artist desk	1278x1080
02	Baby	Middlebury College 2006 Baby 1		A baby on a block and in front of a map of the world.	1191x1080
03	Books	Middlebury College 2005 Books		Some books on a desk.	1282x1080
04	Dolls	Middlebury College 2005 Dolls		Some dolls on a desk.	1279x1080
05	Laundry	Middlebury College 2005 Laundry		Some cleaning stuff on a desk.	1228x1080

06	Merchan	Middlebury College 2006 Midd2		Some Middlebury merchandising.	1274x1080
07	Objects	Middlebury College 2005 Moebius		Some coloured and shaped different objects.	1286x1080
08	Plastic	Middlebury College 2006 Plastic		Some plastic objects.	1194x1080
09	Things	Middlebury College 2005 Reindeer		Some objects at different plans.	1247x1080
10	Rocks	Middlebury College 2006 Rocks2		Some Rocks.	1192x1080
11	Boxing	IRCCyN NAMA3DS1 Boxers (frame 275)		Two boxers training.	1920x1080
12	Hall	IRCCyN NAMA3DS1 Hall (frame 336)		An indoor hall between two buildings.	1920x1080
13	Lab	IRCCyN NAMA3DS1 Lab (frame 279)		Two girls are working in a chemical laboratory.	1920x1080
14	Report	IRCCyN NAMA3DS1 PersonReport (frame 175)		Two men reporting.	1920x1080
15	Phone	IRCCyN NAMA3DS1 PhoneCall (frame 245)		A man speaking on phone.	1920x1080
16	Football	IRCCyN NAMA3DS1 Soccer (frame 321)		Two men playing football.	1920x1080

17	Tree	IRCCyN NAMA3DS1 TreeBranches (frame 300)		Tree branches slow moving.	1920x1080
18	Umbrella	IRCCyN NAMA3DS1 Umbrella (frame 111)		A man using an umbrella.	1920x1080

Figure 9.1: Source description [1].

The still images employed on the database creation are depicted on Figure 9.1. For each one, there are the original SRC name, the new name given to be used in this work, the reference of the source from which it was extracted, a preview, a brief description of what is the scene about and its resolution.

9.2. Display specifications

The Toshiba© 55ZL2 is a 55" (140 cm) LED Smart TV, with a glasses-free 3D panel; quad full HD (4x Full HD) resolution, Intellligent 3D+, 3D Resolution+; 800 AMR (Active Motion & Resolution); WiFi® built-in; Toshiba© Places online portal; personal TV with 4 user profile and personalized user settings; Audyssey® EQ sound technology with integrated soundbar; and colour specifications: black aluminium with chrome applications, chrome stand [60].

It has a tracking system whereby a camera built into the TV scans the room for faces and adjust its viewing sweetspots (util the number of nine possibilities) accordingly, using the Toshiba©'s Cevo Engine processing.

PICTURE

Diagonal Screen Size (cm): 139

Diagonal Screen Size (inch): 55

Visible Area (H x V): 1209.6 x 680.4

Screen Format: 16:9

Panel Resolution: 3840 x 2160

Brightness (cd/m²): 450

Dynamic Contrast Ratio: 9,000,000:1

Response Time (G to G) (ms): 5

Viewing Angle (º): 178

LED TV

LED Edge

Pro LED 32

Panel Technology: QFHD (Quad Full High Definition)

CEVO ENGINE

3D TV: HD 3D-TV

Glasses-free 3D

2D - 3D Conversion

2D - 3D Conversion incl. Depth Control

AMR (Active Motion & Resolution)/Frame Refresh Rate: 800

Resolution+
Network Resolution+
3D Resolution+
Intelligent 3D+
24p Mode
Real Digital Picture
Digital Noise Reduction
MPEG Noise Reduction
3D Digital Comb Filter
Active Backlight Control
Ambient Light Sensor
3D Colour Management
Colour Temperature selectable
Manual Picture Size select
Exact Scan Mode

PICTURE MODE
AutoView
Dynamic
Game
Movie: Hollywood 1 & 2
PC
Standard
Store

EXPERT SETTINGS
Hollywood Day/Night/Pro
3D Colour Management
Greyscale Settings
Gamma Settings
RGB Filter
Integrated Universal Test Pattern
601/709 Colour Decoding Selection
Copy to All Inputs

HDMI™ Information
Auto Calibration

AUDIO
NICAM Stereo
Audyssey® EQ
Virtual Surround Sound
Dolby® Technology: Dolby® Digital Plus
Dolby® Volume
Sound Navi
Number of Subwoofers (built-in): 1
Sound Output (RMS) in W: 2 x 10

TUNING
TV Standards: PAL I/BG/DK; SECAM BG/DK/L, NTSC BG 4.43
Number of Channels: ATV (100) R DTV (9999)
Analogue
DVB-T
DVB-T2
DVB-C/DVB-C (HD)
DVB-S/DVB-S2 (HD)
H.264
DVB Common Interface+ (CI+)
NTSC Video-Playback
Auto Set-up

INTERACTIVE FEATURES
Digital Text
Text Page Memory: 500
Electronic Programme Guide 8 Day
Now and Next Information

Favourite Channel Memory

USB FEATURES
Picture (USB): JPEG
Audio (USB): MP3, MP4
Video (USB): AVCHD, AVCHD Lite, mts, m2ts, H.264, MPEG2PS, mpg, mpeg, MPEG2, MP4, mp4, m4v

SMART TV
PVR Record (to external USB storage device)
WiFi® (Built-in)
WiFi® Protected Setup (WPS)
DLNA
DLNA DMP
DLNA DMR
Toshiba Media Controller compatible
Compatible with Windows® 7
BBC iPlayer (UK only)
Toshiba Places
Camera integrated: Yes (for 3D Control)
Toshiba AppsConnect

PERSONAL TV
4 User Modes
Personal Picture Setting
Personal Volume Setting
Personal Favourite List
Personal Recording Management

OTHER FEATURES
Freeze Screen
Timer
Panel Lock
Auto Format
4:3 stretch
No Signal Off

HOTEL FEATURES
Clone Mode

CONNECTIONS
HDMI™ (back): 4
MHL
2160p, 1080p, 1080i, 720p, 720i, 576p, 576i, 480p, 480i
24Hz (24p)
HDMI™ Audio Return Channel
HDMI™ Audio Content Enhancement
HDMI™-CEC
INSTAPORT™
LAN
Component Video
3.5mm Component Adapter
Composite Video
3.5mm Composite Adapter
SCART
3.5mm Scart Adapter
Analogue audio (Cinch)
Audio (mini jack)
PC Input
USB
Number of USB: 2
Digital Audio Out
Headphone
Subwoofer

STORE MODE

Digital POP Demo

Resolution+ demo

POWER SUPPLY

Line Voltage (Volt/ Hz): 220-240/50-60

Power Consumption EN62087 - Home Mode (W): 236

Power Consumption EN60065 - Maximum power (W): 297

Power Consumption - Stand-by (W): 0.2/0.0 with AC adapter

Yearly Energy Consumption (average kWh): 345

ECO

EU Eco Label

Energie Label: D/power switch exist

Presence of Lead (Below RoHS Directive limit including exemptions)

Mercury Content (mg/kg): < 0,1

DIMENSIONS

Width (mm): 1253

Height without Pedestal Stand (mm): 755

Height with Ped.Stand (mm): 832

Depth without Ped.Stand (mm) - Top section: 40

Depth without Ped.Stand (mm) - Bottom section: 58

Depth with Ped.Stand (mm): 357

Weight without Pedestal Stand (kg): 27.8

Weight with Ped.Stand (kg): 30.5

Packed width (mm): 1632

Packed depth (mm): 254

Packed height (mm): 839

Packed weight (kg): 30.5

CABINET

Swivel stand

Wall Mountable (VESA): 400 x 400 (M8)

ACCESSORIES (INCLUDED IN CARTON BOX)

Remote Control (Type): CT90394

Remote Control Battery: R03(AAA) x2

TV Stand Integrated

Instruction Manual: Quick Guide D, E, F, I, SP, NL, PL, P, GR, FIN, DK, NO, S, RU, TR, HU, CZ

3.5mm Component Adapter

3.5mm Composite Adapter

3.5mm Scart Adapter